UNIVERSITY OF SOUTHAMPTON

The Role of Open Data in the Digital Economy: A Data Science and Economic Perspective.

by

Benito Alán Ponce Rodríguez

A thesis submitted for the degree of
Doctor of Philosophy

in the
Faculty of Physical Science and Engineering
School of Electronics and Computer Science

Monday 10th June, 2019

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCE AND ENGINEERING
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

THE ROLE OF OPEN DATA IN THE DIGITAL ECONOMY: A DATA SCIENCE AND ECONOMIC PERSPECTIVE.

by Benito Alán Ponce Rodríguez

This thesis explores the economic effect of open data released mainly by governments (OGD) on entrepreneurship. First, we use optimization techniques to develop a formal theoretical model that allows us to study the relationship between (OGD) and the choice of individuals to become entrepreneurs or employees. In order to test this hypothesis, we adopt an interdisciplinary approach developing an empirical model using econometric and data science techniques. We collect data from macroeconomic indicators and we create a sample composed of 135 countries from 2013 to 2016 to examine the empirical connection between entrepreneurship and (OGD). Our model shows that changes in (OGD) are correlated with positive changes in the global entrepreneurship and development index (GEDI), this effect is more remarkable in high-income countries. Our estimates suggest that a 1% increase in the index of (OGD) leads to increases of 0.11% in the GEDI index. Second, this research focuses on exploring how companies in Europe are using (OGD) in order to innovate and create new products and services. This work also exposes the business models adopted by these entrepreneurs. For this purpose, we use the data collected by the Open Data Incubator for Europe (ODINE) project from 2015 to 2017 and we implement a text mining approach due to the unstructured of this data. Our results show that the adoption of (OGD) by entrepreneurs is more perceptible in Southern and Northern than Western and Eastern regions of Europe and mainly in high-income countries. Another finding is that the companies that are transforming (OGD) into business proposals are mainly related to the technology sector. Besides, we identify a wide range of companies offering value proposition based on (OGD) as a core idea, some of these companies are not only looking for an economic impact but also they are focusing on the social and/or environmental field. Finally, this investigation presents a list of risks and challenges described by entrepreneurs who are using this digital asset as part of their production process. We found that data quality is the main concern for entrepreneurs. An argument is based on entrepreneurs description stating data quality plays a crucial role in terms of (OGD) demand, this means that if the quality is not improved over time stakeholders could stop using this digital asset.

Contents

List of Figures

Acronyms

OD	Open Data
LOD	Linked Open Data
OGD	Open Government Data
OGP	Open Government Partnership
AI	Artificial Intelligence
DS	Data Science
ML	Machine Learning
TWB	The World Bank
GEM	Global Entrepreneurship Monitor
GEDI	Global Entrepreneurship Index
EF	Economic Freedom
GCI	Global Competitiveness Index
GII	Global Innovation Index
PCT	Patent Cooperation Treaty
GODI	Global Open Data Index
ODB	Open Data Barometer
ODINE	Open Data Incubator for Europe
IP	Intellectual Property
GEDI	The Global Entrepreneurship Index
WBDB	World Bank Doing Business
ODB	Open Data Barometer
GODI	Global Open Data Index
EFI	Economic Freedom Index
GCR	Global Competitiveness Report
GII	Global Innovation Index
WEF	The Word Economic Forum
ODINE	Open Data Incubator for Europe
PCA	Principal Component Analysis
NLP	Natural Language Processing
POS	Part of Speech
ODBM	Open Data Business Models
PPP	Premium Product Services

FPS	Freemium Product Services
OSA	Open Source Alike
IRB	Infraestructural Razor and Blades
DOP	Demand Oriented Platform
SOP	Supply Oriented Platform
FAB	Free as Branded
WLD	White-Label Development
ODA	Open Data Applications
DSR	Design Science Research

Declaration of Authorship

I, Benito Alán Ponce Rodríguez, declare that this thesis entitled The Role of Open Data in the Digital Economy: A Data Science and Economic Perspective. and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as: Ponce et al. 2016.

Signed: Benito Alán Ponce Rodriguez

Date: 23/10/2020

Acknowledgements

I would like to thank my family for their never ending support. To my father because he always showed me the right way to do things and the importance of education. To my mother for her unconditional love and support during all my stages of my life. To my brothers Gustavo y Flavio for being always an example of life and guidance for me. In particular, my brother Raul for his sacrifices and because he is always taking care of the family in good and bad moments. He has been an amazing brother that with his example, dedication, and support he encouraged me to start this amazing journey called Ph.D. To my son Alan Daniel for being my engine and motivation to be a better person every day. I am deeply indebted to all of them. I would also like to thank both of my supervisors. In particular, Francesco Rentocchini for all the feedback and support not only during the process of this work but also for other academic projects. Finally, I thank the Consejo Nacional de Ciencia y Tecnología (CONACYT) for its financial support throughout my Master and PhD. I am also grateful to all faculty and staff at the University of Southampton's Web Science Institute. Especially, Prof Leslie Carr for his guidance and advice. Elena Simperl, Luis Ibáñez, Chris Phethean, Manuel Leon, Rob Blair for giving me the opportunity to work and learn from them on different projects. AAnd to all my friends that I meet during my days in the UK, they made my stay more pleasant. In particular, Miguel Nuñez, Angeles Camacho, Norberto Ramirez, Joel Mejia, and all the amazing Mexican crew that makes you feel at home.

Chapter 1

Introduction

"Data is the new oil".

.

— Clive Humby

1.1 Abstract

This chapter presents the context of the open data movement and the contribution from the public sector to this movement. Moreover, this chapter provides a bibliometric analysis of entrepreneurship as a discipline exploring the historical scientific production, the most productive countries studying this area, the inter and multidisciplinary fields that collaborate with this discipline and the type of methodological research (qualitative, quantitative, or mixed) implemented in this area. Lastly, this chapter presents the interdisciplinary research framework based on economics and data science implementing economy theory and machine learning adopted in this work.

Keywords: Open (Government) Data, Entrepreneurship, Data Science.

.

1.2 Background and Overview

The topic of open data is an area of study that has attracted the attention not only of academia but also the public and private sectors. This could be attributed to the potential shared value in terms of social and economic benefits. The former refers to accountability, transparency, and empowering citizens that access to public information could offer to society. The latter is associated with new business models, patents, and job creation which are triggers of economic growth. An example is the implementation of innovative technology that new startups and established companies are generating with the data released and licensed as open. Another example is the development of new business models that entrepreneurs are adopting in order to monetize, disseminate, and scale their ideas. Lastly, the positive externalities in terms of the creation of the new jobs that this open data ecosystem is creating.

This thesis centers on the economic side implementing an interdisciplinary perspective of the use of open (government) data for entrepreneurial purposes. The aim of this chapter is to provide an introduction and overview of our main concepts. On one hand, this chapter explains the notion of open data. In particular, how the information generated by the public sector is one of the main contributors to the open data movement and it is known as open government data. The approach selected to analyse the field of open (government) data is based on a literature review methodology. The contribution of this analysis is to have a solid understanding of the concept of open (government) data because in the next chapters of this thesis we use this domain knowledge generated to study and measure the relationship and effect of open (government) data on entrepreneurship at the country level. Moreover, how entrepreneurs are transforming these public datasets on business ideas and innovative entrepreneurial models in Europe. Besides, we study what are the main risks and challenges when entrepreneurs are using open (government) data as a digital asset.

On the other hand, we present the entrepreneurship concept implementing a bibliometric analysis which is a statistical exploration of citation and written publication. We adopted this methodology because this research field is a more extensive topic in terms of time and it has been studied from different angles and disciplines. Specifically, we are interested in analyzing the literature and contribution of fields such as business, economics, and government policy to entrepreneurship. The contribution of this analysis is to explore the composition of the entrepreneurship field from an academic perspective. We are interested in analyzing and discovering aspects such as what are the main research topics, its interdisciplinary or multidisciplinary interaction with other fields, and what are the countries that are developing an entrepreneurial research agenda. All these topics are related to our research purposes that are described in detail in the following chapters. In the next section, we explain the scope of this research work.

1.3 Scope of this Thesis

The research framework of this thesis is based on an interdisciplinary approach promoted by web science (WS) due to it encourages the combination and integration of approaches, techniques, methods and theories from a number of fields that can study the Web as a socio-technological phenomenon. This research adopts a qualitative approach supported by economics and data science (DS) as disciplines, adopting an economic theory and machine learning techniques. In economic theory, there is an interest in finding relationships between different variables. Econometrics is the technique that measures this relationship based on data and implementing statistical techniques to analyse, interpret and explore outcomes among diverse factors (Verbeek, 2017). In addition, Data Science is an interdisciplinary field focused on statistical methodologies based on data, supported by Computer Science disciplines such as Machine Learning (ML) and Artificial Intelligence (AI) (Phethean, Simperl, Tiropanis, Tinati, & Hall, 2016).

The aim of machine learning is to find patterns, perform predictions, classifications, and clustering. These tasks are possible using and learning from data through the implementation of different types of computational libraries and development of customized code referred as algorithms. Machine learning is divided into two areas of study cataloged as Supervised and Unsupervised Learning. The former is performed using label training data to learn and implement the function from X (usually referred as independiente or input variables) to Y (dependent or output variable) and it is expressed as

$$Y = f(x) \tag{1.1}$$

Supervised algorithms are helping to solve problems that involve prediction. Regression analysis is one of such technique that is used to predict the outcome of a given sample containing real training values in the dependent variable. Classification is another type of supervised algorithm for prediction using categorical data on the output variable. Conversely, Unsupervised Learning implements algorithms which only contains the input variable (X). This type of algorithm uses unlabeled training data and the implementation of this algorithm is to solve problems related to association, clustering, and dimension reduction.In the next section, this work explains the definition and interdisciplinary scope of open (government) data and entrepreneurship.

1.4 Open Data

Open Data (OD) refers to information that has been created or collected by governmental entities, business, academics and individuals which owns the Intellectual Properties (IP) rights,

but which is then published online for other organizations to use and share freely (Wainwright, Huber, & Rentocchini, 2014). As a complement of this definition, Tim Berners-Lee[1] proposes a 5 stars scheme that explains OD from a technical perspective. This categorization is explained in the following way: it is assigned 1 star to data that should be available on the Web under an open license (this data could be in any format e.g. PDF, JPG). Data that is assigned to 2 stars means that this information is available on the Web as structured data (e.g. Excel instead of images). 3 stars are given to data in a non-proprietary open format (e.g. CSV instead of Excel). Data labeled as 4 stars is categorized to denote things using URIs so that people can locate your information (e.g. RDF structure). And 5 stars are assigned to data linked that provide context (e.g. Web Data).

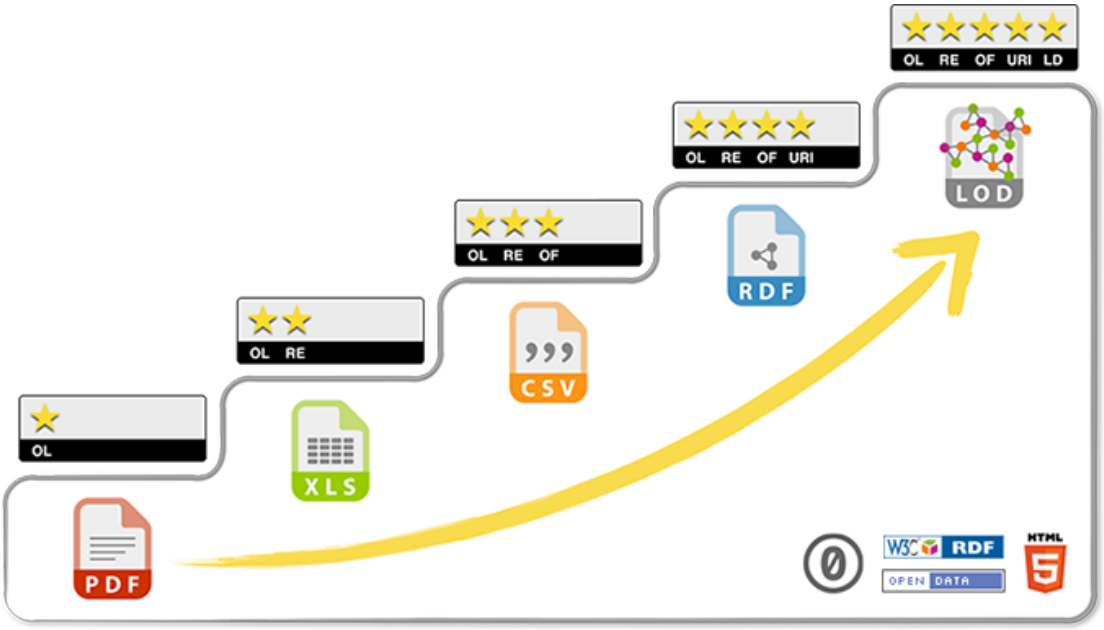


FIGURE 1.1: Composition of the 5 star of Open data

However, an important point for of any type of open data (released by public, private or academic sector) is that in order to avoid legal uncertainty and promote its adoption and utilization, this data should have an explicit license that describes its terms of use (Filippi & Maurel, 2014; Morando, 2013; The World Bank Group, 2013).

The adoption and implementation of open data is a socio-technological phenomenon that has been studied by different disciplines trying to understand and estimate its dimensions and barriers. For instance, the technical outlook is associated with the relevance of improving data interoperability, quality, accessibility, usability, accuracy, platforms, and infrastructure needed in order to release open data. The social stance refers to the empowerment that data offers to society. For example, the potential benefits that the information released by governments could produce through transparency and accountability on citizens. The economic point of view is related to possible impacts to the economy that open data could

[1] https://5stardata.info/en/

offer through the creation of new business, products and services as well as employment. This perspective also includes the crucial role that innovation plays as driver of economic growth in the private and public sector using open data (A Zuiderwijk et al., 2014). The political perspective covers the strategies, policies, and impacts of the data released by the state. The data published and freely accessible by public entities is referred as Open Government Data (OGD). This particular kind of data plays an important role in the open data movement because it is considered as one of their main supporters through initiatives such as the Public Sector Information (PSI)[2] and Open Government Partnership (OGP)[3] due to these political actions are looking for increasing efficiency, promoting transparency, empowering citizens and driving a knowledge-based economy through the release of data generated by public sectors. Examples of these datasets are census, weather, healthcare, educational information, geographical, business, patents, and transportation.

1.4.1 Public Sector Information (PSI) and Open Government Data (OGD)

Governments at the federal, regional, and municipal level are one of the main contributors to the open data movement because of the number of datasets that generate and release as a public good the governmental entities that are part of these levels. One of the origins of this symbiotic relationship was the development of Public Sector Information (PSI) directives in different nations. One remarkable example of these directives was the European legislation on the re-use of public sector information created in 2003 (European Commission, 2003) in which member states agreed to ensure the re-use for commercial or non-commercial purposes of information created for public bodies and these datasets shall be made publicly available through electronic means. Another notable example of this relationship, that was a core component in the evolution and transformation of PSI directives into more structural OGD policies, is the political global movement Open Government Partnership (Harrison & Sayogo, 2014). This is an initiative involving 75 countries that promotes the implementation of open data policies as a tool to foster transparency, accountability, fight against corruption and empowerment of citizens. (Ubaldi, 2013) claims that the creation of open data policies are essential for the publication, infrastructure required, legal certainty and political sustainability of open government data (OGD). The author also argues that open data policies should disseminate the economic and social value of OGD in order to stimulate the use and reuse of it on society. (Thorhildur Jetzek, 2013) argue that the release of OGD is relevant because there are datasets collected by different sectors and for specific purposes (i.e. transportation, pollution, agriculture, education, health, census, etc). Authors also claim that OGD is as driver for innovation and business opportunities for society. Finally, they argue that the infrastructure of these data sets were paid by taxpayers; therefore, this information is considered a public good.

[2] `http://www.oecd.org/sti/ieconomy/oecdrecommendationonpublicsectorinformationpsi.htm`
[3] `https://www.opengovpartnership.org/`

This work focuses on the economic and business perspective explaining how open (government) data is used for the identification of new business opportunities, strategic planning and the evaluation of investment projects by entrepreneurs. For this reason, it is important to have an understanding of the evolution and the inter and multidisciplinary scope of entrepreneurship.

In the next section, we examine the field of entrepreneurship through a bibliometric analysis in order to explore its scientific production, main countries that contributes to its study, and its connections with other scientific fields.

1.5 Entrepreneurship Landscape

1.5.1 Bibliometrics Analysis on Entrepreneurship

Entrepreneurship is a very extensive field that interacts with many disciplines. A bibliometric analysis is referred a quantitative and statistical exploration of citation and written publication (books and articles) that could be implemented to any scientific field (Osareh, 1996; Pato & Teixeira, 2016) . Citations are one of the indicators to measure activity in the scientific community (Garfield, Malin, & Small, 1983; Garfield, Sher, & Torpie, 1964; Small, 1973). The scope of this bibliometric study is a combination of both macro and micro analysis in order to explore the composition of entrepreneurship as an academic field. In the macro analysis scope, we start describing the overall structure of scientific production, countries, sources, and disciplines that are related to entrepreneurship. In the micro analysis, we focus on the relationship of entrepreneurship with other specific disciplines such as business, economics, and governmental policy, the goal is to identify the factors that drive entrepreneurship from these disciplines.

We conducted our bibliometric study using the academic databases engines Web of Knowledge and Scopus, searching for the term "Entrepreneurship" without any restriction of time or topics. Then, we used the open source software R (Foundation, 2013) and the bibliometrix library (Aria & Cuccurullo, 2017) to merge and analyze our data. Our results show 13,824 documents covering the period of time from 1970 until May 2017 (see Appendix section A.1). This scientific information collected is from different disciplines, subjects, and countries and we describe it in the next section.

1.5.1.1 Bibliometrics Macro-Analysis

Entrepreneurship is a topic that has attracted the attention of many researchers publishing in diverse journals and books around the word in the last decades, with a remarkable increase in the last 10 years. Figure 1.2 illustrates the historical scientific production per year from 1970 to 2016. (For the purpose of this figure, we are showing only completed years).

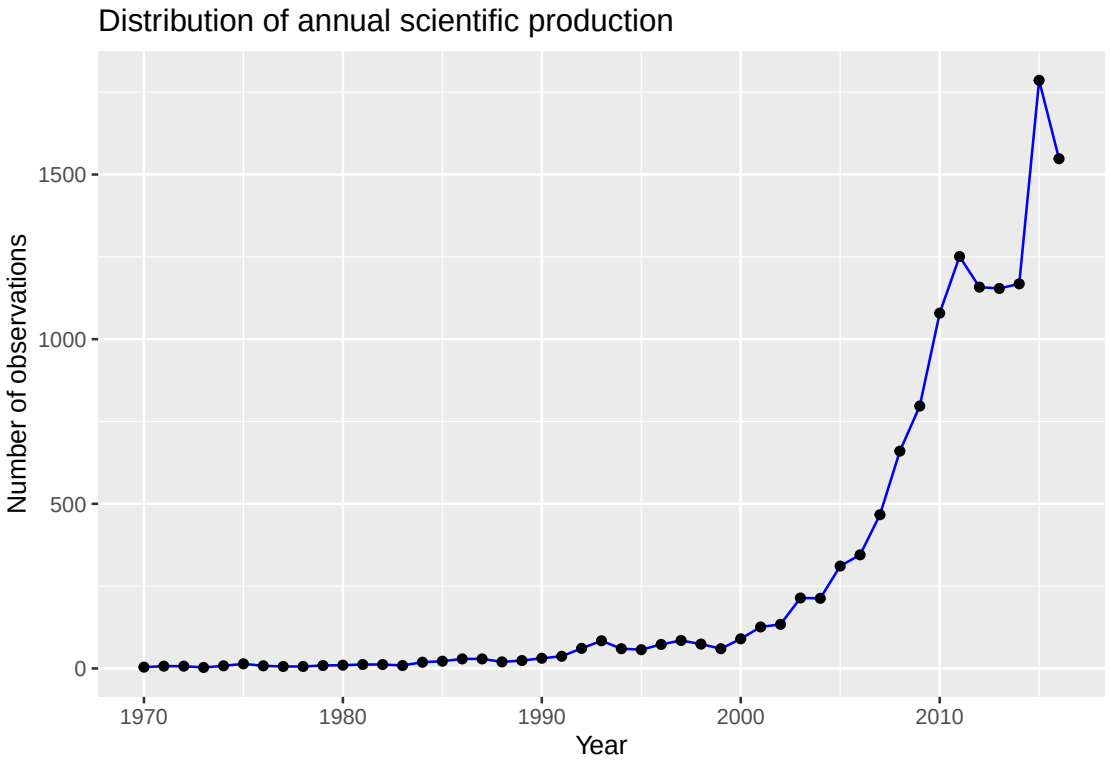


FIGURE 1.2: Number of articles published per year

Results of the analysis show that the total of articles published on entrepreneurship until May 2017 are 13,824. Filtering the top 10 countries in terms of academic productivity about this topic, the most proficiency economy is the United States, followed by the United Kingdom, China, Spain, and Canada. Figure 1.3 shows the full list of countries researching the topic of entrepreneurship around the world.

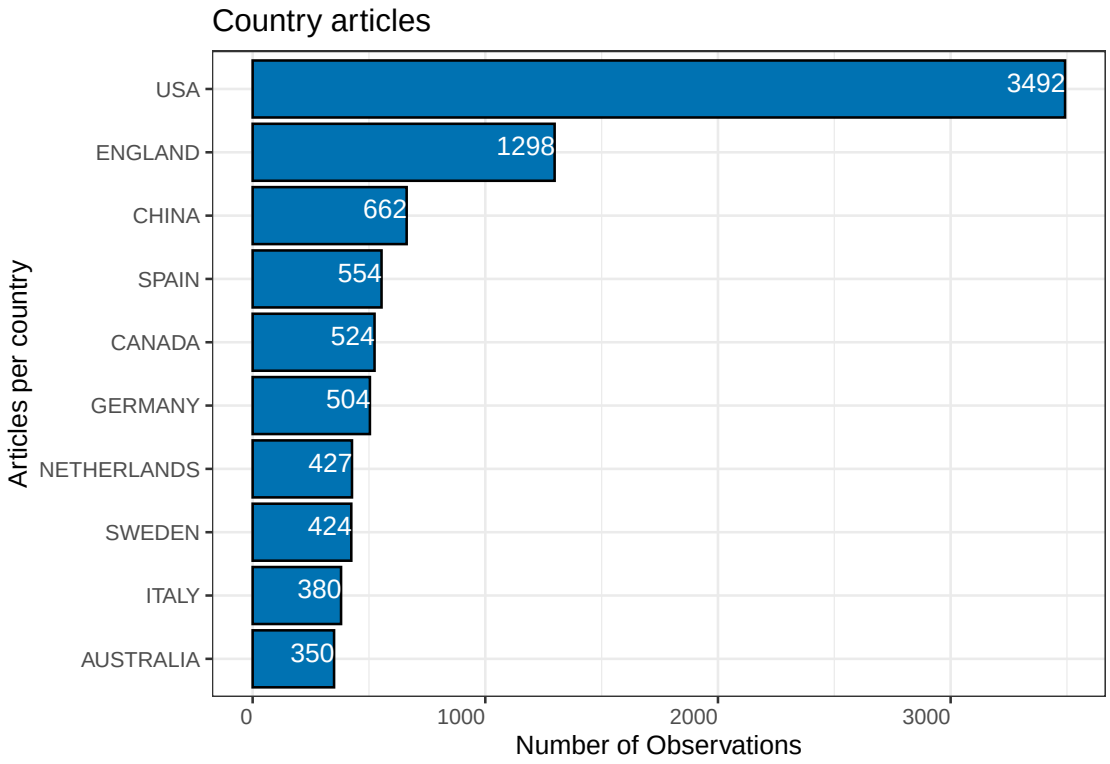


FIGURE 1.3: Number of articles published per year

1.5.1.2 Bibliometrics Micro-analysis

According to the data analysed, entrepreneurship is a field that has been studied using multi and interdisciplinary perspectives. The main fields related to entrepreneurship using an interdisciplinary approach are business and management, the data shows 2,863 academic documents between these fields. Using a multidisciplinary approach between business and economics, there are 334 academic documents. The next figure shows the top 10 areas that have a close relationship (inter and multidisciplinary approach) to entrepreneurship.

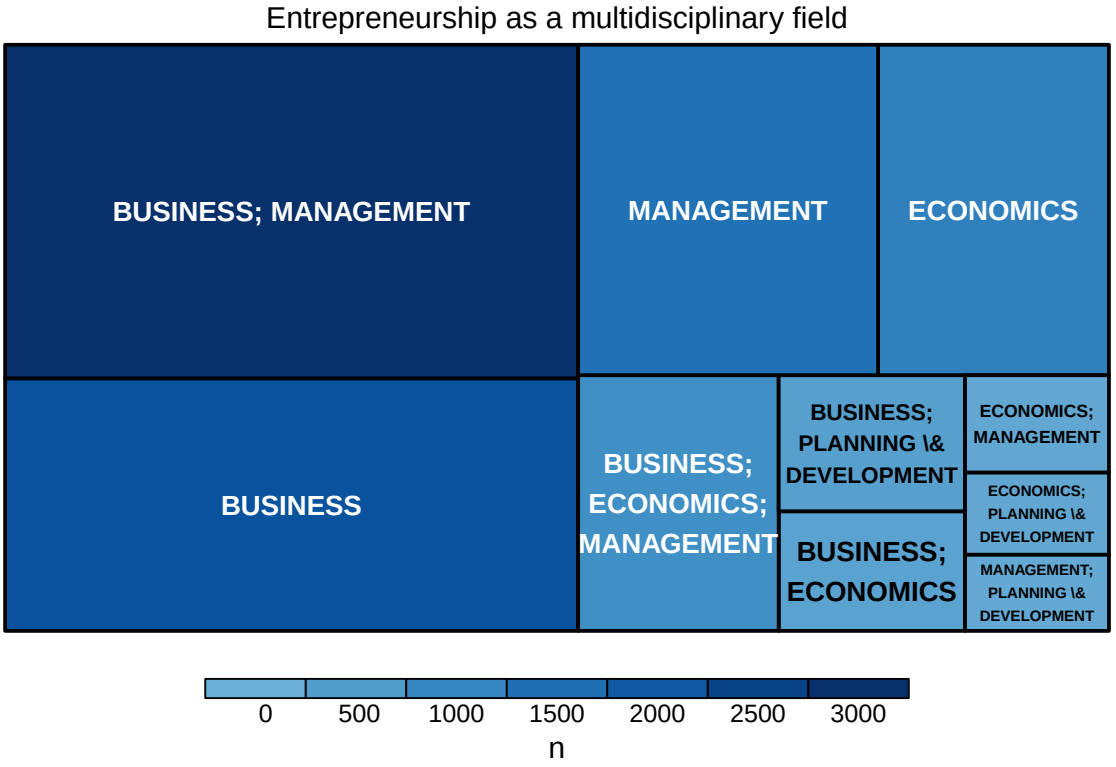


FIGURE 1.4: Main disciplines related to entrepreneurship

Another aspect of this analysis is to explore the relationship of entrepreneurship and other fields such as economics and business. The purpose is to discover what the main sources of publication and what the topics and research line are. Results of the analysis show that the entrepreneurship journal Small Business Economics has a bigger number of academic publications.This journal covers a wide range of topics such as self-employment, family firms, small and medium-sized firms, and new venture creation. Moreover, this journal has a special interest in the economic and social perspective about entrepreneurs' characteristics, occupational choice, new ventures and innovation, firms life courses and performance; as well as the role played by institutions and public policies within local, regional, national and international contexts. Results show that other main sources of publications are Journal of Business Venturing (accepting academic proposals from different disciplines such as economics, psychology, sociology, anthropology, geography, history, and so on) and Entrepreneurship Theory and Practice (this journal focuses on national and international studies of enterprise creation, public policymakers, research methods, and venture financing). An important point to mention about this particular analysis is that not all of the academic engines databases include this information; therefore; this field contains a lot of NA. Figure 1.5 illustrates the top 10 sources of publication about entrepreneurship.

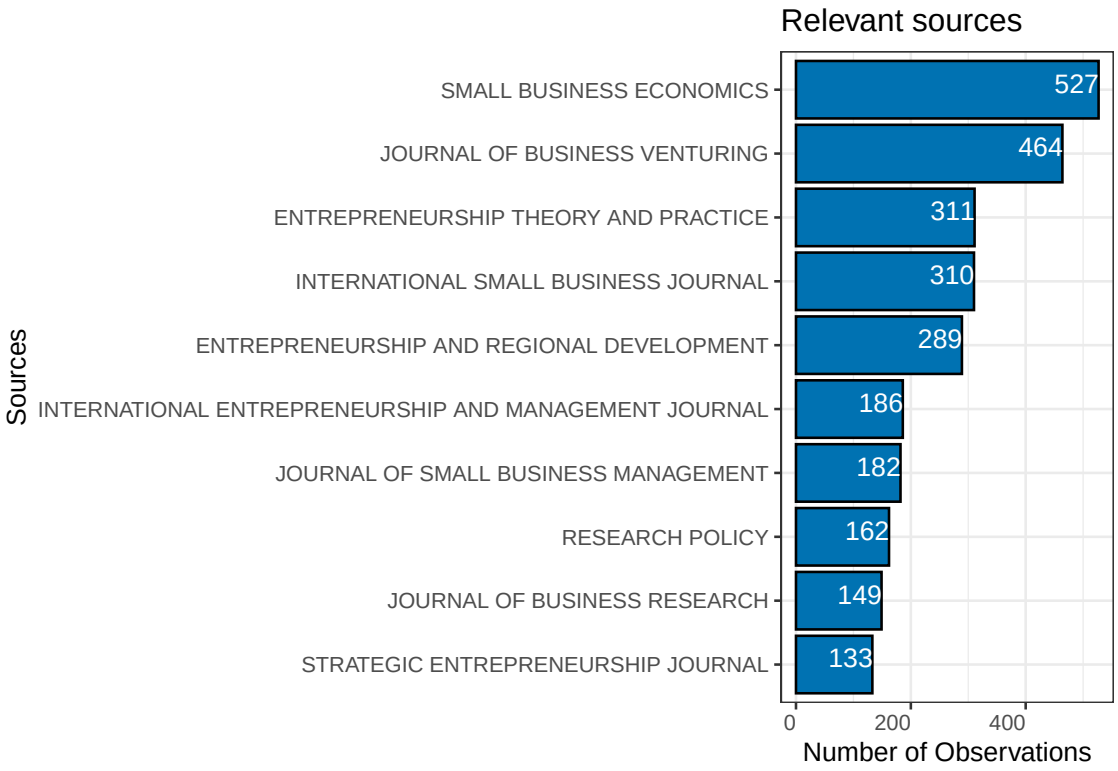


FIGURE 1.5: Illustrates the main journals publishing about entrepreneurship

Our next step in our bibliometric analysis is to explore the type of methodology (quantitative or qualitative) employed by the authors in their research and described in the abstract section. The aim is to discover and the type of approaches used in the field of entrepreneurship. Results show that there are 1105 articles published since 1994 using the word "Qualitative". Contrarily, we found 395 publications using the word "Quantitative" in the abstract section. There are 694 documents combining both approaches. Figure 1.6 illustrate the number of articles using the word "Quantitative" or "Quantitative" or "Quantitative and Quantitative" (Mixed) in the abstract section.

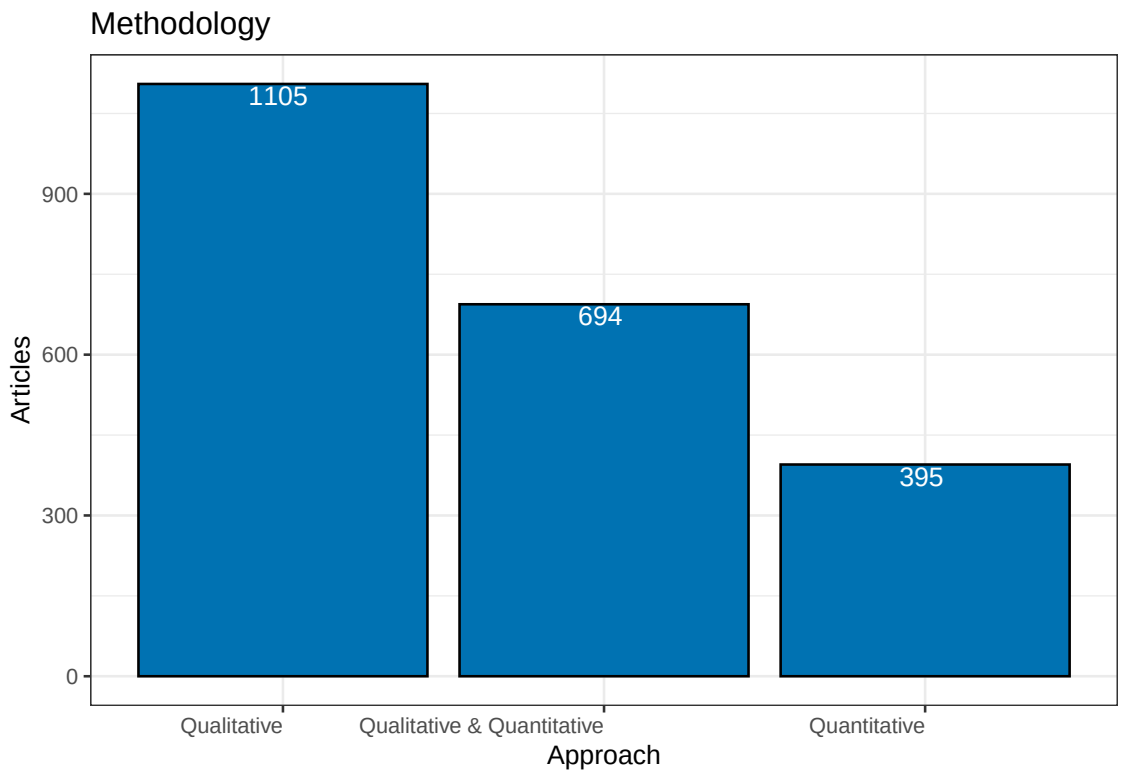


FIGURE 1.6: Methodologies implemented on entrepreneurship research

In the next section, an outline is provided describing the following chapters' aims, contribution, and methodologies of this thesis.

1.6 Outline of this Thesis

Chapter 1, presents the definition and context of the open data movement and the relevance of the public sector to this movement. Moreover, this chapter provides a bibliometric analysis of entrepreneurship as a discipline exploring the historical scientific production, the most productive countries studying this area, the inter and multidisciplinary fields that collaborate with this discipline and the type of methodological research (qualitative, quantitative, or mixed) implemented in this area. Lastly, this chapter presents the interdisciplinary research framework based on economics and data science implementing economy theory and machine learning adopted in this work.

Chapter 2 provides a cross-sectional regression analysis in order to systematically test the relationship and effect of open (government) data on entrepreneurship at the country level. This analysis is composed of a theoretical framework to justify and strengthen our variable selection using economic and business theory. Moreover, this chapter includes an empirical model measuring the association and impact of open (government) data on entrepreneurship using multiple linear regression.

Chapter 3 focus the analysis providing evidence of the relationship between open (government) data and entrepreneurship, describing the commercial opportunities perceived by entrepreneurs to enhance or produce new products, services, or innovative business models in Europe. Furthermore, this chapter exposes where these companies are located and the economic sectors which these companies belong. This chapter also explores each element of the Business Model Canvas framework (key partners, key activities, value propositions, channels, customer segment, cost structure and revenue stream) adopted by these entrepreneurs in order to commercialize and scale their products or services.

Chapter 4 explores the risks and challenges that entrepreneurs are facing when they are using open (government) data in Europe. Moreover, this chapter provides a comparison between the risk and challenges described in the literature review by academics and government authorities with entrepreneurs that are using open (government) data as part of their production process.

Lastly, Chapter 5 provides a summary and final thoughts about our main findings for each chapter.

Chapter 2

Entrepreneurship and Open (Government) Data

"Data is the lowest level of abstraction from which information and then knowledge are derived.".

.

— Open Data Institute

2.1 Abstract

The recent adoption of Open Data (OD) policies by Governments and the increased release of Open Government Data (OGD) around the world is a phenomenon being studied by different disciplines trying to understand and estimate its potential benefits, limitations, and dimensions. The purpose of this research is to analyze the relationship and effect of open (government) data on entrepreneurship at the country level. The framework of this research is based on a theoretical model explaining how different levels of open (government) data could affect the decision of individuals of becoming an entrepreneur. This work also tests this relationship and effect, developing an empirical model through a selection of macroeconomic variables, this selection process is supported by a literature review approach based on entrepreneurship theory. The sample of this analysis is a panel data composed of 137 economies that includes indicators from The Global Entrepreneurship Index (GEDI), Open Data Barometer (ODB), Global Open Data Index (GODI), Economic Freedom Index (EFI), The Global Competitiveness Report (GCR), and the Global Innovation Index (GII) from 2013 to 2016. A multiple linear regression analysis adopting an econometric and machine learning approach is used in order to produce a more robust analysis and estimate the relationship between open data and an index of entrepreneurship at the country level. Our estimates

suggests that open (government) data has a positive and statistically significant impact on entrepreneurship and its potential benefits are that open (government) data gives access to information and the identification of new business opportunities, more effective strategic planning, and a more efficient evaluation of investment projects. All these concepts are closely related with the formation of new entrepreneurs and new business.

Keywords: Open Government Data, Entrepreneurship, Econometrics, Data Science.

.

2.2 Introduction

Policy makers and researchers worldwide are paying increasing attention to the potential role of the open data in boosting entrepreneurship and innovation (Cabinet Office, 2012; Lee, Almirall, & Wareham, 2015; Open Data Institute, 2015). Following this interest, governments at different levels in the US, Europe, and increasingly emerging economies are explicitly supporting open data through the introduction of new legislation, which requires that their departments publish data which can be used by private sector SMEs and corporations to develop new innovative products or services (Chattapadhyay, 2013; Dos Santos Brito, Da Silva Costa, Garcia, & De Lemos Meira, 2014; OECD, 2016; Tinati, Carr, Halford, & Pope, 2012).

Researchers of the open data movement claim that open data reduces costs, it enables organisations to retain more monetary value from new innovations and it provides access to previously unavailable data, enabling new business opportunities to be developed (mainly in the form of innovative digital services), which in turn creates new social and economic value (Cappgiemini Consulting, 2013; Open Data Institute, Lateral Economics, 2016; Stott, 2014; Takagi, 2014). Following these expectations, the potential global economic impact of open data has been estimated as being worth $3.5 trillion dollars annually (Manyika et al., 2013).

Despite government involvement and the anticipated impacts in different dimensions (political-social-economic-technological), open data still offers several challenges and barriers that need to be studied. One of this is that many governments are opening and releasing their data which are referenced as Open Government Data (OGD) from different agencies at different levels (federal, provincial and/or municipal); nevertheless, from a user perspective, there is a gap in understanding the value of open (government) data and how to exploit it (e.g. by combining and transforming it in order to be redeployable in business activities) (Kaasenbrood, Zuiderwijk, Janssen, De Jong, & Bharosa, 2015). Moreover, there is limited research and evidence on the impact of open (government) data on entrepreneurship. Although some qualitative studies provide evidence of how open data is related to economic impact, this evidence is far from being conclusive (Davies, 2013; Stott, 2014; Anneke Zuiderwijk & Janssen, 2014).

The aim of this research is to analyse and provide quantitative evidence for the association and impact of open (government) data on entrepreneurship.This study provides an analysis and discussion that open (government) data allows economic agents such as households and firms access to information and encouragement to make rational decisions. open (government) data is central for the decision making-process of households that seek to allocate efficiently the family's resources, such as the decision to allocate resources for consumption, or the formation of human capital, and the resources to be allocated for retirement. Open (government) data is also an important tool for the basic decisions of firms such as the level of production, the optimal mix of inputs, and for the identification of profitable and

non-profitable investments. Therefore, this analysis aims to answer the following research question:

Is there a relationship and effect in the publication of Open (Government) Data on Entrepreneurship at the country level?

The significance of this question is based on the fact that the discovery and exploitation of entrepreneurial opportunities is at the core of the creation of new companies, which are the engine of economic growth (Mueller, 2007; Shane, 2000). Furthermore, open (government) data tends to provide the information needed for the identification of new business opportunities, strategic planning and the evaluation of investment projects (Bonina, 2013). As individuals and companies require access to information to make optimal decisions related to all aspects of entrepreneurship, such as the identification of new business opportunities and the development of new products/services (Ivanova & Gibcus, 2003) , open (government) data is likely to provide pivotal information which can avoid suboptimal decisions and lead to the discovery and exploitation of new opportunities (Casson, 2005).

The contribution of this research lies in the quantitative evidence showing the relationship between open (government) data and entrepreneurship. Notably, this research collected a cross-country panel dataset using macroeconomic variables from 2013 to 2016 that study the determinants of Entrepreneurship on 137 economies. The main interest of this research is to test whether an increase in the publication of open (government) data has an effect on entrepreneurial activities at the country level. An analysis implementing multiple linear regression models adopting an econometric and data science approach in order to systematically test and estimate the relationship of entrepreneurship (as an endogenous variable) and open government data (as one of the exogenous control variables). This analysis is developed using different data sources. For instance, we collect data from The Global Entrepreneurship Index (GEDI), Open Data Barometer (ODB), Global Open Data Index (GODI) Economic Freedom Index (EFI) The Global Competitiveness Report (GCR) and the Global Innovation Index (GII) .

This chapter is organised as follows: Section 1 describes the aim, purpose, content, justification, model and methodology used in this research. Section 2 discusses the theoretical background of entrepreneurship and open (government) data as well as other variables of control in order to establish its empirical connections, for this purposes we implemented a literature review approach based on entrepreneurship theory. Section 3 provides a theoretical model linking the mechanisms that explain how changes in the availability of open (government) data could affect the decision of becoming an entrepreneur. Section 4 the methodology, data, and models are presented analysing the relationship between the dependent and independent variables described in the literature review section. Results and discussion are presented in section 5. Finally, section 6 proposes conclusions and key findings about our hypotheses and future work.

2.3 Literature Review

There has been research conducted separately from entrepreneurship and open (government) data; nevertheless, the focus of this section is to find literature and evidence joining these fields analyzing its relationship and impact in the promotion of a new entrepreneur culture on the digital economy.

2.3.1 Entrepreneurship

Despite entrepreneurship is a field that has been extensively studied, there is not a universal definition on it due to its multi and interdisciplinary scope (Bull & Willard, 1993; Lumpkin & Dess, 1996; OECD, 1998; Van Praag, 1999); therefore, the concept of entrepreneurship tends to be explained based on the research perspective. For instance, business, economic, sociology (Aldrich, 2000; Gartner, 2001; Shane, 2000; Verheul, Wennekers, Audretsch, & Thurik, 2002).

However, (Hebert & Link, 1989), offer a concept that describes an entrepreneur as: *"someone who specializes in taking responsibility for and making judgmental decisions that affect the location, form, and the use of goods, resources or institutions"*. Furthermore, (Audretsch, Thurik, Verheul, & Wennekers, 2006) claims that entrepreneurship requires an interdisciplinary approach due to it spanning a broad range of fields as well as management, finance, sociology, economics and its study covers different aspect such as individual or group culture, geography location and particular periods of time.

As stated by The Eclectic Theory of Entrepreneurship proposed by (Verheul et al., 2002) there are several determinants that influence entrepreneurship. One of these is the role that governments plays directly or indirectly on it. A direct intervention is the creation of specific rules and legal frameworks that support and promote policies that encourage entrepreneurial activities in specific sectors. An indirect intervention is through generic campaigns sponsoring individuals or collective groups.

Another determinant is composed by the level of entrepreneurship in a country. This could be measured by micro, meso and macro scope. Analysis at the micro level is focused on individual choices as well as people motivation to become self-employed. Studies at meso level are related to specific segment of markets. Finally, the macro level analysis attempts to combine both micro and macro approaches applying them on variables such as technological or economic environments.

A third element that is mentioned by the authors is the combined field approach due to entrepreneurship has a multidisciplinary procedure that cannot be limited to one discipline. Authors argue that in order to understand and measure entrepreneurship, researchers should have an holistic perspective including fields such as psychology to understand what are the

traits and characteristics of individuals who want to be an entrepreneur. Sociology in order to understand the atmosphere and variables of entrepreneurs in a collective way, economy to stimulate the impact of entrepreneurship and politics to promote policies that encourage society to adopt an entrepreneurial culture and so on.

A final determinant stated by the authors is the demand and supply of entrepreneurs. On one hand, authors suggest that the supply is influenced by the size of the population, and individual or cultural aspects towards entrepreneurship. On the other hand, the demand side is composed of the perceptions and business opportunities for entrepreneurship. According to the authors, the demand for entrepreneurship is strongly related to technological development and government regulation.

2.3.2 Open Government Data (OGD)

The adoption of open data covers a wide spectrum of fields such as government, business, journalism, academic or social media data (Gurin, 2014). An example of this relationship in the public sector is the symbiotic adoption of open data in The Open Government Partnership (OGP) initiative.[1] The OGP is a global movement involving 75 countries that promotes the implementation of open data policies as a tool to fosters transparency, accountability, it can also help to fight against corruption and empowers citizens (Sandoval-Almazan, Gil-Garcia, Luna-Reyes, Luna, & Rojas-Romero, 2012). (Ubaldi, 2013) claims that the creation of open data policies are crucial for the publication, required infrastructure, legal certainty and political sustainability of releasing Open Government Data (OGD). The author also argues that open data policies should disseminate the economic and social values of OGD in order to stimulate the use and reuse of it on society. (Thorhildur Jetzek, 2013) argue that the release of OGD is relevant because there are datasets collected by different sectors and for specific purposes (i.e. transportation, pollution, agriculture, education, health, census, etc) that potentially have a shared value (economic and social). Authors also claim that OGD is a driver for innovation and business opportunities for society. Finally, they argue that the infrastructure of these data sets was paid by taxpayers; therefore, this information is considered a public good.

Researchers propose that the release of open (government) data is seen as a trigger for business opportunities and it has started an entrepreneur movement. (Verheul et al., 2002) claim that high level of entrepreneurial activity is related to innovation, competition, economic growth and job creation. For instance, the McKinsey Global Institute estimates that the potential economic benefits of open data are at least $3 trillion a year globally (Manyika et al., 2013). According to The World Bank, there are many economic benefits related to open data from governments such as economic growth, business and job creation (Stott, 2014). (Dawes, 2012) claims that government is the dominant supplier of open data. Evidence found by (Lakomaa & Kallberg, 2013) suggests that more access from government to

[1] https://www.opengovpartnership.org/

data, the bigger the chances to perceive or promote additional business ideas. (Barne, 2014; Lab, 2014) claim that an open data entrepreneurial culture has recently emerged and that about 500 companies were created using open (government) data in the United States in areas such as education, energy efficiency, and health services. (Bonina, 2013) exposes that open data promotes an ecosystem in which governments are one of the main actors through policies generation and information providers. The formation of entrepreneurial ecosystems has positive externalities in the economy. The Open Data Institute (Open Data Institute, 2015) in the United Kingdom has a startup program that supports business models using open government data and other sources of open data such as business or academia as part of its services. In Sweden, (Lakomaa & Kallberg, 2013) find that technological entrepreneurs consider open data as core for their business plan and that more access to other data set could promote additional business ideas. The Open Data Barometer (Davies, 2013) reports that the entrepreneurial open data use is the second indicator of impact on its study around the world.

Organisations and consortiums are also providing evidence of the relationship between entrepreneurship and the use of open (government) data. For instance, The Center for Open Data Enterprise[2] which is an nonprofit organization based in Washington D.C, and that is in charge of the development and maintenance of the Open Data Impact Map, launched in 2016 . This map displays information of 1700 organisations composed of companies, NGOs, academic institutions and developer groups from 96 countries around the world that are using open (government) data for economic and social purposes. Then, filtering the information displayed in this map regarding the economic side of open (government) data, there are 1209 companies[3] from different continents using open (government) data as an asset of their business and accessing public information from diverse domains such as transportation, weather, health care, education, geoSpatial, housing, agriculture, and energy.

Another organization created is The Open Data Incubator for Europe (ODINE)[4] which is a project supporting entrepreneurs that are using open data (from any source such as: government, corporate, academic) as part of their business proposals. This project is a consortium composed of 7 partners from different sectors such as private, academic, nonprofit and journalist composed of The University of Southampton, Fraunhofer Institute, Open Data Institute, Guardian, Open Knowledge Foundation, and Telefonica Open Future, and it is funded by the European Commission through the Horizon 2020 initiative. This program started in 2015 and the consortium received more than 1100 business proposals from all countries that are part of the European Union to participate in their incubation program. Companies that applied to the ODINE program are offering solutions based on open data from different domains such as finance and insurance activities, transportation and storage, legal, professional scientific and technical activities, human health and social work activities, real estate, energy, and public administration and defence.

[2] `http://www.opendataenterprise.org/`
[3] `https://opendataimpactmap.org/map`
[4] `https://opendataincubator.eu/`

2.3.3 Government Policy

The role of governments designing public policies plays a vital role in order to foster an entrepreneurship and open data culture, infrastructure and ecosystem. (Audretsch et al., 2006) points out that governments serve as an authority in order to create the legal framework and economic conditions that should regulate the harmony among all actors involved in an entrepreneurial development.

One of its functions is to act as a regulatory supervisor of the supply and demand market in case it turns into dysfunctional market, information discrepancy, or unfair competition. Other examples of government mediation are distribution of income and performance of the economy. Authors claim that in the entrepreneurial sphere, the government intervention has a significant impact on the small and medium-sized enterprises (SMEs) through the development of policies such as interest rates and taxation, deregulation and simplification, administrative burdens, finance, internationalisation, labor training and information. These policies play a crucial role in this sector since SMEs are an important group of job creation and economic growth. In addition, this sector is characterized as a weak and volatile sector that requires government support.

(Sepulveda & Mendez, 2005) claims that the government is motivated by positive externalities such as production, employment, market stimulation, and economic growth to impose regulation on entrepreneurship. According to (Lundstrom & Stevenson 2006), entrepreneurship policy is referred as the creation of a favourable ecosystem to motivate individuals to enter the entrepreneurial process. This process is composed of 5 stages: 1).- Awareness (raise the interest or promoting opportunities to become a entrepreneur) 2).- Pre-startup (supporting and promoting certain types of opportunities) 3).- Startup (reducing regulatory and procedural barriers and promoting finance, market plan, and training opportunities) 4).- Post-start-up (measuring early success and failure and promoting seed financing, networking, regulatory burdens, technology transfer and information access) 5).- Maintenance and expansion (growth financing, labour regulations, tax burdens, internationalisation).

In addition,, the role of governments in the context of open data is more than open and release information in a readable machine format (Stott, 2014). The engagement and participation of all actors are important pieces of any open data ecosystem and governments play a significant role in these tasks for several reasons. First, governments act as one of the main suppliers in this open data environment, an important engagement is the possibility to release data that entrepreneurs can request based on demand and this could be done through the national open data portal created by each government. Furthermore, governments should guarantee in some way the continuity of the information released in order to promote confidence to investors in this sector. Second, governments should act as leaderships promoting policies and stating the economic and social benefits that could incite other governments to open their data. Third, governments should be a trigger agent in the creation of an open data ecosystem through creating policies that motivate the participation

of data users, developers and data-driven business. Lastly, governments should implement a national agenda promoting and using their own open data and it should done involving regional and city level institutions.

As an example of this connection between government and open data agendas is The Open Government Partnership (OGP) which is a global initiative that involved 75 countries developing policies allowing to release information in an open format and at no cost in order to promote transparency, accountability, fight against corruption and empowerment of citizens. (Sandoval-Almazan et al., 2012) argue that this empowerment allows individuals to have a more active role in society . For instance, activists might support their facts, journalists link more information, and individuals take more informed daily life decisions.

2.3.4 Economic Growth -Gross Domestic Product per Capita (GDP PPP)

Literature agrees that economic growth is referred to as an increase in the capacity to produce goods and services, and it is measured over certain periods of time that could be affected by different agents. For instance, (Gylfason & Zoega, 2006) observe that natural resources are an important source of national wealth. (Benhabib & Spiegel, 1994) claim that the human capital is an important determinant of the rate of growth of a country through the skilled and trained people because they can adopt, implement, and produce technological development in a better way and therefore generate economic growth. Furthermore,(Becker, Murphy, & Tamura, 1990) state that the growth of countries is achieved more rapidly when there is an accumulation of human capital composed by scientific knowledge and labor force.

Entrepreneurship is considered another agent that affects economic growth. (Schumpeter, 1934) considers entrepreneurs as the main cause of economic growth. In addition, (Gwartney, Lawson, & Holcombe, 1999) claim that entrepreneurship creates an atmosphere in which entrepreneurs and innovation are constantly increasing productivity. The author suggests that its incorporation to the economic framework not only promotes development but also entrepreneurship, thus helping to encourage the creation of economic policies that generate growth. (Carree & Thurik, 2003) observe various impacts of entrepreneurship and economic growth. One of these is related to the small and medium-size enterprises (SMEs) due to the crucial role this sector plays in the economy of a country acting as an agent of change and stimulating innovation and new firms and jobs. Another impact is throughout the adoption and implementation of new technologies because they are able to minimize the dimension of scale economies in different sectors.

According to The Endogenous Growth Theory proposed by (Zilibotti, Aghion, Howitt, & Garcia-Penalosa, 1999), economic growth could be affected by government policies, technological knowledge, human capital and innovation. This theory proposes that in order to maintain long-term growth, there must be continual developments such as new goods, services, markets and process generated by technological knowledge. Besides, (Solow, 1957;

Swan, 1956), from an neoclassical economic perspective, argue that technological progress is a cornerstone of economic growth. Finally, (De Loo & Soete, 1999) are in agreement with (Aghion & Howitt, 1990; Grossman & Helpman, 1991; Romer, 1990) that argue that better spending and allocation of resources on the development of new technologies leads to a continuous expansion of the economic growth.

2.3.5 Competitiveness

Literature points out another determinant that drives entrepreneurship which is the level of competitiveness in a country. This factor attracted the researcher's attention due to the globalization of economies and market competition among them (Porter, 1986). Furthermore, the author also claims that competitiveness is referred as national productivity (Porter, 1998). Another definition made by The World Economic Forum (WEF), is that competitiveness refers to *"the set of institutions, policies, and factors that determine the level of productivity of an economy"* (Klaus Schwab, 2017). Furthermore, the WEF claims that competitiveness can be measured through a set of 12 pillars composed of 3 groups called factors, efficiency and innovation-driven. The first group is related to the basic requirements of institutions, infrastructure, macroeconomic stability, health and primary education. The second group contains the sources of efficiency higher education, goods market efficiency, labour market efficiency, financial market development, technological readiness, market size and business sophistication. The third group includes innovation and business sophistication factors. (Korez-Vide & Tominc 2016) argue that competitiveness and entrepreneurship are factors of economic growth quantified by gdp per capita growth. The authors claim that there is a positive relationship between competitiveness and growth rates in some countries in the EU, this association was statistical positive with efficiency driver pillar.

(Ferreira, Fayolle, Fernandes, & Raposo, 2017) based on Schumpeterian and Kirznerian theories explain that entrepreneurship is considered as an engine for economic growth and national competitiveness. The first theory considers entrepreneurship (based on innovation) as triggers in the economy through the introduction of new competitors, productivity generators, job creation and national competitiveness. Moreover, Schumpeterian theory also argue that entrepreneurship tends to promote innovative initiatives that foster new forms of production and organisation, new products, technologies, markets and production resources (Schumpeter, 1934, 1939, 1942). In the second theory in which entrepreneurship is viewed based on opportunity (Kirzner, 1973) explain that individuals are dynamic agents playing an important role in the balance of markets, performing activities that are crucial to competitiveness and this is intrinsic to the process of entrepreneurship.

(Rocha, 2004) agrees that entrepreneurship is positive associate to economic development and one important piece of this development is related to enhancing global competitiveness. Moreover, (Reynolds et al., 2005) claims that developed countries show higher levels of competitiveness than developing countries; however, advanced economies show lower levels

of a certain type of Entrepreneurship indicator i.e. TEA (Total Early-Stage Entrepreneurship Activity) than developing economies. This phenomenon could be explained due to the opportunity cost that individuals have in the former countries. In other words, in developed countries, there are substantial number of employment opportunities against the risk to become an entrepreneur. (Gonzalez-Pernia, Pe na-Legazkue, & Vendrell-Herrero, 2012) found that the most innovative and entrepreneurial areas are also the show high levels of productivity.

2.3.6 Innovation

Several authors have highlighted the importance between entrepreneurship and innovation. For instance, (Slappendel, Carol, 1996) claims that the attention of newness is crucial to the idea of innovation because it helps to separate innovation from change. The point of newness plays an important role linking innovation and entrepreneurship due to previous research indicates its focal part in new venture creation and management (Gartner, 1988; Lumpkin & Dess, 1996; Stevenson & Jarillo, 1990; Vesper, 1988). (Johnson, 2001) argue that innovation is a crucial element of entrepreneurship and one of the core factors of business success. The author also claim that Innovation implies newness. Furthemore, he explains that the use and advance of technology is changing not only business process and customer requirements but also global competition

(Kotabe & Scott Swan, 1995) explain the difficulties to understand innovation due to the absence of a measure. In 2007 was released The Global Innovation Index (GII), the purpose of this index is to find metrics and methods that measure Innovation offering additional dimensions such as the Research and Development (R&D) sub pillar of the Human Capital indicator (Wipo, 2017)(Wipo 2017). According to the GII, there are several reasons to collect and disseminate this information. One of this is governments are settling innovation as one of their main points of their growth strategies. Second, innovation is perceived as a multidimensional value that could impact different domains such as economic, social, academic and technical. Third, innovation is happening around the world and analysing develop and developing economies helps to understand how influential ideas are inspiring people to embrace entrepreneurship and innovation

Innovation can be referred as creative thinking (Jon-Arild Johannessen, Bjørn Olsen, & G.T. Lumpkin, 2001). According to Schumpeter's innovation theory there are different types of innovation: the introduction or a change in an existing product; the application of a new methodology in production or sales of a product; opening a new market; obtaining of new sources of supply raw material; or changing an industrial structure. Moreover, the author also argues that innovation is crucial for economic development through generating eruptions of creative destruction and that entrepreneurs play a crucial role in this process. Besides, he proposes that innovation involves 4 dimensions: invention, innovation, diffusion and imitation (Schumpeter, 1934). (Zhao, 2005) claims that entrepreneurship and innovation

are symbiotic elements and they are not limited to the initial stages of an organization; rather, they should be part of a holistic process in an organization.

Open data is considered as a transmission channel for the development and dissemination of innovation in public and private sector (Bedini et al., 2014). For instance, the development of new products or services, as well as novel production methods by companies that can enhance user experience based on open data. Innovation is connected to the public sector in the sense of reducing bureaucracy and promoting citizen engagement and participation through open (government) data (Chan, 2013).

A concept that has been attracting the attention of public, private and academic sectors is Open Innovation (OI) which refers to *"a paradigm that assumes that firms can and should use external ideas as well as internal ideas"* (Chesbrough, Vanhaverbeke, & West, 2008). Open innovation in combination of the increased release of open data by governments, have been stimulating the creation of innovative business models based on digital transformations with profit-oriented focus (Zimmermann & Pucihar, 2015).

The intersection of open data and open innovation offers several proposals. One of these is that the adoption of open data in the private sector is principally represented by SMEs which are more flexibles to embrace new paradigms of innovation (Vrande, Jong, Vanhaverbeke, & Rochemont, 2009). In addition, this sector is mainly related to the supply of services that represent a potential area for the implementation of open service innovation (Pooran Wynarczyk, Panagiotis Piperopoulos, & Maura McAdam, 2013). Moreover, open data publishers could adopt an inbound approach (accepting external inputs from users) in order to improve the quality of the data released which is one of the main critiques of the open data movement from the demand side (Huber, Rentocchini, & Wainwright, 2016). Finally, the cultural acquisition and mixture of openness associated with data, software, services, governmental policies and innovation is another area that is still under research and that needs to be assessed in order to estimate its impacts.

2.3.7 Economic Freedom

According to (Miller & Kim, 2013) economic freedom refers to the basic right that individuals or firms have to legally to choose what to produce and how to produce it in a free and open market. They argue that in a free economic society the government decision-making is distinguished by openness, transparency and inclusive social policy. The authors also argue that economic freedom is composed by several elements. One is business freedom that is considered as the individual right to legally create and run firms without interference from the state. Trade freedom is another element and it refers to globally interact as a buyer or seller thorough importing and exporting goods and services in an open economy.

Investment freedom refers to an economic environment that promotes entrepreneurial risks generating benefits such as new firms and job creation. Besides, transparency and equality are

part of the investment framework that could encourage competition and innovation through entrepreneurs; on top of that, openness helps to eliminate unnecessary restriction due to the fact that analyzing market restrictions could promote more entrepreneurial activity. Property rights is an additional element of economic freedom that gives reliance to individuals through a legal framework system to venture into business world due to their physical and intellectual innovations being safe for expropriation. Independence, transparency and effectiveness should be components of this legal system in order to provide certainty to citizens.

Researchers also state that in an economic freedom market the government policy is present in business freedom through licensing, decision-making and bureaucracy, which affects the creation of new business. Besides, the government intervention in the creation of business is made through the price-setting process and taxation burden. With respect to trade freedom, the government plays a regulatory role in which export taxes or trade quote try to keep the balance for all actors. Finally, the government supports the legal system that protects citizens and their creations and possessions. (Kreft & Sobel, 2005; Miller & Kim, 2013) claim that policies addressed to economic freedom are the basis on strong economies. Furthermore, The author justifies his findings with previous work made by (Cole, 2003; Gwartney et al., 1999; Gwartney, Lawson, Park, & Skipton, 2005; Ken Farr, Lord, & Wolfenbarger, 1998; Powell, 2002) who found that economies with higher economic freedom rates not only have large per-capita income, but also a higher rate of economic growth.

2.3.8 Summary

Entrepreneurship as a multidisciplinary activity with a strong impact in the economy that could be influenced by government intervention, and the supply and demand of the market economy. Entrepreneurship is also influenced by the state of technology and development of the economy and by geographic factors. In addition, open data is considered by literature as a socio-technological movement that offers shared social and economic values from several domains. For instance, the information from the public sector is one of these domains and it plays a crucial role on this movement. It is usually referred as Open Government Data (OGD) and offers the possibility to attract entrepreneurs, foster business opportunities, and create new businesses in an ecosystem composed of private sector, academia, and developers. Open (government) data also promotes the development or implementation of innovative business models such freemium, premium, or supply-oriented platforms.

Different areas of study have been interested in studying the relationship and impact of government policy, innovation, competitiveness, economic freedom and growth on entrepreneurship. One of the main findings in the literature is related to government policy due to its regulatory function on the supply and the demand of goods and services, distribution of income, its intervention in the development of SMEs, and its role developing and promoting an entrepreneurial culture. Another relevant finding is how the application and the variation

of different tax policies proposed through government policies could affect the creation of new entrepreneurs, jobs and business.

An additional outcome proposed by the literature is that economic freedom involves the legal framework to protect property rights which is considered as a key element in an entrepreneurial environment. Besides, the business freedom to legally create and develop firms without government dictation and the possibility to globally interact with other buyers or sellers in open markets under the cover of a legal framework. Literature also states that a government policy promotes investment freedom, encouraging individuals to take the risk of becoming an entrepreneur and create jobs. Finally, economic literature agrees that entrepreneurship is one of the main factors of a long-run growth due to the fact that it promotes an environment of productivity. Moreover, economic theory also agrees that the economic growth depends on variables such as technological knowledge, human capital and innovation in order to create and develop new good and services.

In spite of the large literature on entrepreneurship and on the recent literature on open (government) data, it is considered that the analysis of the role of OGD in determining entrepreneurship has received little attention. For instance, the question whether open government data affects positively or negatively the formation of new entrepreneurs is still an open question and there few analysis on this issue. Another issue of interest is to capture the relationship among open government data and other economic agents in order to understand its connections and determinants of impact of open data on entrepreneurship. Lastly, it is considered important to estimate whether open (government) data has sizable effects on entrepreneurship or the effects are marginal. To the best of my knowledge, there is no research estimating such effects.

Other important questions that have not been deeply analyzed are related to the opportunity cost to use open government data and become an entrepreneur. Regarding entrepreneurs already in the market, there is scarce literature or quantitative evidence about what is the value proposition that entrepreneurs are offering when they are using open (government) data in their business models. Furthermore, what are the risks and challenges in using open (government) data. These questions, in particular, could be relevant to understand the possible impact of open (government) data on the economy, through job creation, innovation and the efficiency in the allocation of resources, which are considered core topics in economics. If open (government) data has sizable effects on entrepreneurship then governments could have incentives to allocate more resources for the promotion of it.

In the next section, we will address one of these questions developing a theoretical model that test the perception of business opportunities and cost-benefit analysis to become an entrepreneur using open (government) data.

2.4 Theoretical Model

In this section, we develop a model of a rational choice between offering labor services in the market or becoming an entrepreneur. Hence, this mode considers the opportunity cost to become an entrepreneur as an important determinant of choosing to become (or not) an entrepreneur; I start the analysis assumingan individual with the choice of her careerd The decision is between becoming an entrepreneur or offering labor services in the labor market. If the individual decides to offer his labor services then he(she) can earn a competitive salary . If the individual decides to become an entrepreneur then he(she) produces a good with technology showing constant returns to scale.

$$y = f(k, L, O_d) = \alpha_0 ln(k) + \alpha_1 ln(O_d L) \tag{2.1}$$

In condition (2.1), the production of good y depends on capital K, labor L, and Open (government) data O_d. Open (government) data is a productive input since it provides information that makes labor more productive. The term $O_d L$ is interpreted as the effective level of labor. Open (government) data is a productive public good determined by the information made available by individuals and the government, hence, from the perspective of the individual, open (government) data is an exogenous variable. (the firm does not control the level of O_d)[5]

The individual sells the good y in a competitive market at a price P. The cost of production (or cost from inputs) is given by $C = wL + rK$ where w is a competitive salary for labor services L and r is a competitive rental cost of capital. The firm's fixed costs (those costs that are not dependent of the level of production of the firm) are given by Φ

The profit π from entrepreneurship is given by

$$\pi = pf(K, L, O_d) - wL - rK - \Phi \tag{2.2}$$

Hence an individual decides to become an entrepreneur if

$$\pi(k, L, O_d) \geq w_{oc} \tag{2.3}$$

Condition (2.3) says that if the return to become an entrepreneur is higher than the individual's opportunity cost in the labor market, then the individual decides to become an entrepreneur. If $\pi(K, L, O_d) < w_{oc}$ the individual decides to sell his labor services in the labor market at an opportunity cost of w_{oc} (the individual does not become an entrepreneur).

[5] In the economics literature a public good satisfies the properties of non exclusion and non rivalness. Non exclusion means that once the good is provided (by the private or the public sector) then no one can be excluded from its consumption. Non rivalness means that each individual consumes the good with the same properties, that is to say, the consumption of an individual i precludes the consumption of individual j. For instance a radio signal satisfies these properties. If one individual wants to connect to the signal of radio he(she) can do it (the property of no exclusion is satisfied). And if the quality of the signal of the radio is the same for different individuals connecting to the signal then the property of no rivalness is satisfied.

We can state condition (2.3) as a difference of returns of becoming an entrepreneur or offer the individual's labor services by the term X as shown below

$$X = \pi(K, L, O_d) - w_{oc}$$

In this section we seek to show that if the relative profitability between becoming an entrepreneur and selling labor services at a wage w_{oc} that is X, increases as there is an increase in open (government) data. Advancing the main result of this section, I prove that higher levels of open (government) data increase the likelihood of becoming an entrepreneur because higher levels of open (government) data increase the profitability of becoming an entrepreneur.

2.4.1 Open Data and the Decision to Become Entrepreneur

In this section, I develop a comparative analysis of how changes in the available level of open (government) data affects the decision of becoming an entrepreneur. To do so, we first analyze the impact of changes in open (government) data in the profitability of entrepreneurship.

For the analysis of this section, I start by studying the optimal choices on L^* and K^* of the firm. We use differential calculus to obtain the optimal level of L^* and K^* as shown below:

The problem of the firm is

$$Max_{k,l}\pi = pf(K, L, O_d) - rk - wL - \Phi \tag{2.4}$$

Where $y = f(k, L, O_d) = \alpha_0 ln(k) + \alpha_1 ln(O_d L) and \alpha_0 + \alpha_1 = 1$

Following the literatura, it is assumed that prices of inputs and the fixed costs of production are given for the firm (the firm's actions in the market are small such that the decisions of the firm do not affect the prices of labor, capital and the good provided by the firm). This assumption is equivalent to assume that markets are competitive. Using the technique of mathematical optimization I use differential calculus to solve for the optimal levels of capital and labor that determines the level of output of the firm. Once capital, labor and the production decision of the firm is reached then the firm sells its production at given prices and the profitability of the firm, π s determined. Hence:

$$Max_{k,L}\pi = p\{\alpha_0 ln(k) + \alpha_1 ln(O_d L)\} - rk - wL - \Phi$$

The first order conditions are:

$$\frac{\partial \pi}{\partial k} = P\frac{\partial f(k, L, O_d)}{\partial k} - r = 0 \Rightarrow k^* = k^*(P, r, \alpha_0) \tag{2.5}$$

$$\frac{\partial \pi}{\partial L} = P\frac{\partial f(k, L, O_d)}{\partial L} - w = 0 \Rightarrow L^* = L^*(P, r, \alpha_0) \tag{2.6}$$

Conditions (2.5) and (2.6) represent the optimal demand functions of the inputs to determine the level of production that maximizes the firm's profit. These functions are given by $k^*(P, r, \alpha_0)$ and $L^*(P, r, \alpha_0)$ and represent the optimal choices with respect capital and labor of the firm. Using the parametric form of the production function then conditions (2.5) and (2.6) become

$$\frac{\partial \pi}{\partial k} = P\frac{\alpha_0}{k^*} - r = 0 \Rightarrow k^* = k^*(P, r, \alpha_0) \Rightarrow k^* = P\frac{\alpha_0}{r} \tag{2.7}$$

$$\frac{\partial \pi}{\partial k} = P\frac{\alpha_1}{L^*} - w = 0 \Rightarrow L^* = L^*(P, w, \alpha_1) \Rightarrow k^* = P\frac{\alpha_1}{w} \tag{2.8}$$

Since capital and labor are productive (that is production increases when capital and/or labor increases) then the profit of the firm increases as capital and labor increase. Equation (2.7) says that the firm's demand of capital increases (and therefore the profitability of the firm π increases) if the price P of the good increases and it falls (and therefore the profitability of the firm π falls) if the price of capital r increases. Moreover, if capital becomes more productive (this is the effect of technological change), that is there is an increase in α_0 then the demand of capital increases. Equation (2.8) says that the firm's demand of labor increases (and therefore the profitability of the firm π increases) if the price P of the good increases and it falls (and therefore the profitability of the firm π falls) if the price of labor w increases. Moreover, if labor becomes more productive (this is the effect of technological change), that is there is an increase in α_1 then the demand of labor increases.

Note that the firm does not choose the optimal level of open data because open data is a public good provided by the government and therefore the amount of open (government) data is exogenous for the firm.

To analyse the impact of open (government) data in the profitability of the firm (and hence in the decision of becoming an entrepreneur or not) we use the optimal choices of capital and labor in the profitability function and state the following:

$$Max_{k,L}\pi = P\{\alpha_0 ln(k^*(P, r, \alpha_0)) + \alpha_1 ln(O_d L^*(P, w, \alpha_1))\} - rk^*(P, r, \alpha_0) - wL^*(P, w, \alpha_1) - \Phi \tag{2.9}$$

Now we are interested in analysing how changes in open (government) data affect the profitability of the firm at the economic equilibrium (this is why we need to use the optimal responses of the firm $k^* = k^*(P, r, \alpha_0)$ and $L^* = L^*(P, w, \alpha_1)$ in the profitability function). To do so, we need to calculate the derivative $\frac{\partial \pi}{\partial O_d}$ from (3.10). Hence

$$\frac{\partial \pi}{\partial O_d} = P\frac{\alpha_1 L^*(P, w, \alpha_1)}{O_d L^*(P, w, \alpha_1)} = P\frac{\alpha_1}{O_d} > 0 \tag{2.10}$$

$$\frac{\partial^2 \pi}{\partial^2 O_d} = -P\frac{\alpha_1}{(O_d)^2} < 0 \tag{2.11}$$

Conditions (2.10) and (2.11) imply that the relationship between open (government) data and the profitability of entrepreneurship is characterized by figure 2.1 Now we return to the

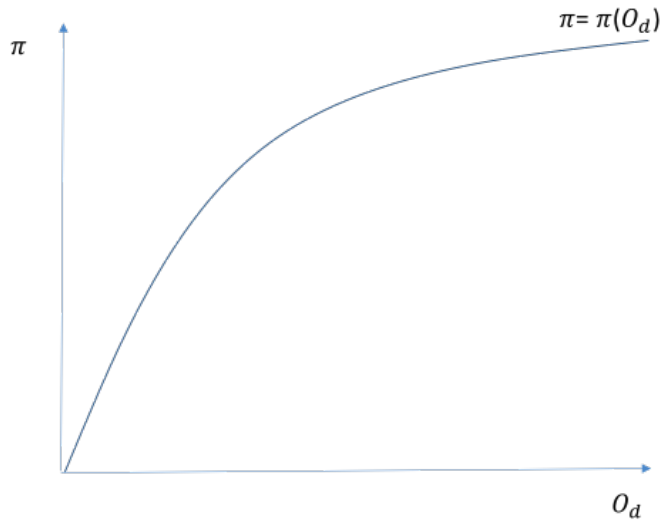


FIGURE 2.1: Profitability of entrepreneurship and open (government) data

decision of becoming entrepreneur

$$\pi(k, L, O_d) \geq w_{oc} \tag{2.12}$$

We can state condition (2.12) as a difference of returns of becoming an entrepreneur or offer the individual's labor services by the term X as shown below

$$X = \pi(k, L, O_d) - w_{oc} \tag{2.13}$$

If we derive (2.13) by changes in open (government) data we obtain

$$\frac{\partial X}{\partial O_d} = \frac{\partial \pi}{\partial O_d} > 0 \tag{2.14}$$

We have shown in condition (2.10) that $\frac{\partial \pi}{\partial O_d} > 0$ (that is an increase in open (government) data makes the activity of entrepreneurship more profitable). Therefore the relative profitability between becoming an entrepreneur relative selling labor services at a wage w_{oc} increases as there is an increase in open (government) data. This result shows that increases in open (government) data increases the profitability (and the likelihood) of becoming an entrepreneur. This outcome is explained by the fact that having more and better information might help individuals to recognize more profitable investment projects and to take more rational (efficient) decisions that affects the development of the firm. As a result, more open (government) data increases the rate of return of entrepreneurs and therefore more individuals would choose to become an entrepreneur.

2.5 Empirical Model

The purpose of this section is to develop a multiple regression analysis to test whether open (government) data is systematically linked to entrepreneurship and estimate the marginal effect of entrepreneurship as an endogenous variable and open (government) data with other control variables as exogenous indicators. In the following sections, we describe the process to test this relationship.

2.5.0.1 Interdisciplinary Approach

In order to develop our models, we adopted an interdisciplinary approach implementing econometrics and data science techniques. On one hand, In economic theory, there is an interest on find relationship between different variables and quantities and econometrics is the technique that measures this relationship based on data and implementing statistical techniques to analyze, interpret and explore outcomes among diverse factors (Verbeek, 2017). On the other hand, data science is an interdisciplinary field focused on statistical methodologies based on data, supported by computer science disciplines such as machine learning and artificial intelligence (Phethean et al. 2016). Machine learning's aim is to find patterns, perform predictions, classification and cluster data. These tasks are possible using and learning from data through the implementation of different types of algorithms.

Machine learning is classified as supervised learning which is performed using label training data to learn and implement the function from (X) (usually referred as independent or input variables) to Y (dependent or output variable) and it is expressed as

$$Y = f(x) \tag{2.15}$$

Supervised algorithms are helping to solve problems that involve prediction. Regression is one of them that predict the outcome of a given sample containing real training values in the dependent variable. Classification is another type of supervised algorithm for prediction using categorical data on the output variable.

Other types of machine learning algorithms are unsupervised learning which only contains the input variable (X). This type of algorithm uses unlabeled training data and the implementation of this algorithm is to solve problems related to association, clustering, and dimension reduction.

2.5.0.2 Intersection of Machine Learning and Econometrics

In this research, the intersection point of these fields lies using economic theory and data science techniques in order to estimate and predict the effect of open (government) data

and other control variables on entrepreneurship. On one hand, economic theory plays an important role in shaping the context and structure of the variables that are the determinants to promote or inhibit entrepreneurship at the country level. The selection and interaction of these variables are a fundamental piece for the development of our statistical models. On the other hand, machine learning provides us with sophisticated techniques and algorithms that help us deal with data that has very high dimensionality and tuning modeling that can improve the quality of prediction.

Multiple linear regression (MLR) is the statistical tool implemented to systematically test and predict the relationship between entrepreneurship defined as dependent variable (also also known in the econometric and data science literature as exogenous or response variable) and open (government) data, economic freedom, innovation adoption, economic development and competitiveness indicators denoted as independents variables (also referred in the literature as explanatory, features or predictors variables). Multiple Linear Regression is suitable when you have to deal with more than one independent variable to estimate the effect in the response variable (Rao & Toutenburg, 1995).

A challenge to deal with our response variable is that we have several independent variables of control that according to the economic theory described in the literature review and our bibliometric analysis are relevant determinants of entrepreneurship; therefore, we need to define a methodology to analyze and solve issues such as multicollinearity and high dimension reduction. In the next section, we describe the workflow that aims to solve these issues.

2.5.1 Processes and Research Workflow

An important piece of any data science approach is defining the workflow of our problem, describing every step involved in the transformation of the data and looking for research reproducibility (Davenport & Patil, 2012). Figure 2.2 Illustrates the pipeline developed for our research

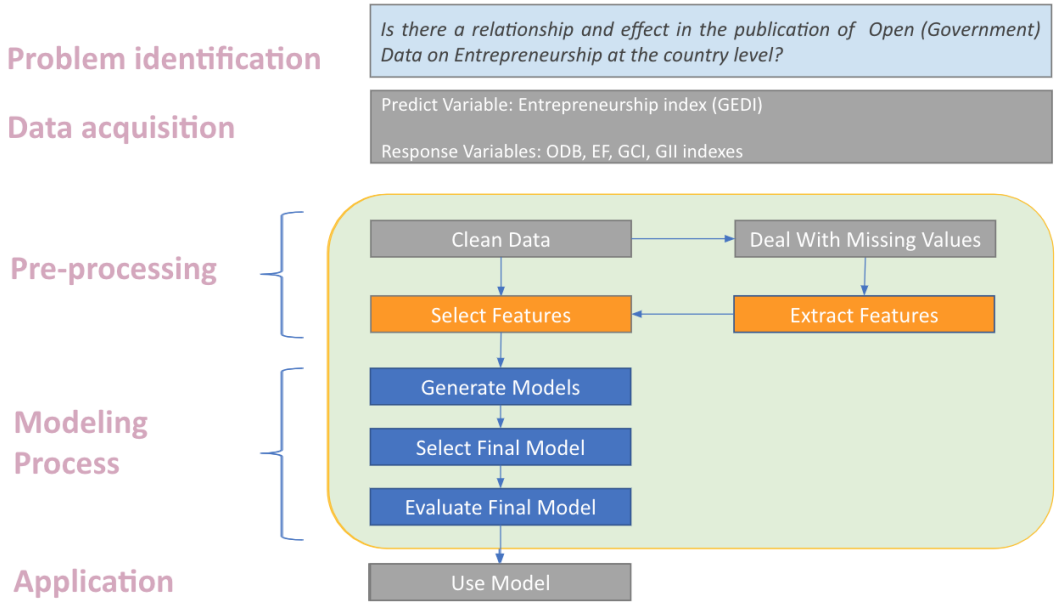


FIGURE 2.2: Shows the pipeline developed for our research

Problem Identification and Hypothesis

Literature agrees about the crucial role that governments play as a trigger agent in an open data ecosystem since it could promote the legal and economic framework to develop a cultural entrepreneurial culture. Besides, it has an important function of engagement, participation and elimination of barriers through its authority and leadership. Lastly, as stated before, governments are considered as one of the main suppliers in the open data movement.

The formation of new business requires the growth of entrepreneurship. Open (government) data provides the information needed for the identification of new business opportunities, strategic planning and the evaluation of investment projects. All these concepts are closely related with the formation of new entrepreneurs and new business. Individuals and firms require access to information to make rational decisions. For instance, information empower citizens to exercise their rights and they could have a more active and inclusive participation on society. Furthermore, economists argue that firms select capital, labor and raw materials to produce goods and services that maximize the firm's' profits. Without access to information, individuals and firms might take suboptimal decisions that interfere with their aims. Therefore, the main focus of this research is analyzing the relationship and effect that open (government) data plays on entrepreneurship and how this relationship is affected by other economic elements such as business, trade, and investment freedom, property rights and taxation burden at the country level. This leads to our hypotheses.

here a positive relationship and effect in the publication of open (government) data on entrepreneurship at the country level..

.

The next stage in our research workflow is the data collection. This process is described in detail the next section.

Data Acquisition

The data collection process is divided into two main phases due to the composition of our dependent and independent variables; therefore, the information is collected from different sources. In the first stage, we collect data related to our endogenous variable. This information is obtained from the Global Entrepreneurship and Development Institute (GEDI)[6] which is an indicator measuring entrepreneurship determinants at national and regional levels around the world. The GEDI index is especially suitable for our study because this indicator incorporates a policy development (focus that is missing in other entrepreneurship measures (Acs, Szerb, & Lloyd, 2017; Szerb, Aidis, & Acs, 2013). This index is developed by a consortium based on Washington D. C. and founded by academic institutions such as George Mason University, University of Pécs and Imperial College London. Furthermore, this indicator is funded by the European Union, World Bank, and other institutions interested in entrepreneurship determinants at the country level. A more detailed explanation of the composition of this index (pillars and sub-pillars) and the date frame time collected for this study is provided in the next section.

During the second phase, we collect data related to our exogenous variables from different sources such as The Open Data Barometer (ODB)[7] which is an index developed by the World Wide Web Foundation. The aim of this indicator is to measure and compare the adoption and impact of open data initiatives around the world. This information is collected and analysed by researchers, civilians and government representatives that are involved in the open data movement. This index is funded by the Open Data for Development (OD4D) a programme supported by different international institutions such as the World Bank, United Kingdom's Department for International Development (DFID) and the International Development Research Centre (IDRC) and Global Affairs in Canada.

The Economic Freedom Index (EF)[8] produced by The Heritage Foundation is an indicator that focuses on four pillars (rule of law, government size, regulatory efficiency, and market openness) in which governmental entities regularly use it as policy control. This measure is useful for our study because it reflects the economic and entrepreneurial ecosystem at the country level through the twelve sub-pillars that emanates from the main four elements that composed this index. Some references in case (Diaz-Casero, Diaz-Aunion, Sanchez-Escobedo, Coduras, & Hernandez-Mogollon, 2012).

[6] https://thegedi.org/research/gedi-index/
[7] https://opendatabarometer.org/barometer/
[8] https://www.heritage.org/index/about

The Global Competitiveness Report (GCR)[9] published by the World Economic Forum. This index refers to competitiveness as a collection of several factors such as institutions, infrastructure, policies, education and other determinants (which will be explained in the next section) that allows measuring the productivity at the country level. This index is composed of three main pillars and twelve sub-pillars that creates a set of variables . We consider relevant to our study this index because as stated in by the literature the elements of these pillars (e.g. strong of institutions, policy development, human capital) are part of the determinants of entrepreneurship in any economy (Korez-Vide & Tominc, 2016).

The Global Innovation Index (GII)[10] co-published by Cornell University, INSEAD, and the World Intellectual Property Organization (WIPO) is an index composed of two pillars and seven sub-pillars that measures innovation performance at the country level. The information provided in this index is relevant to our analysis because innovation is considered by the literature as a trigger for entrepreneurship (Crumpton, 2012). Moreover, this index collects data from around the world about indicators such as political environment, business and market sophistication intangible assets, creative good and services which are suitable for our study of the effect of open data on entrepreneurship.Figure 2.3 Displays our data selection and composition.

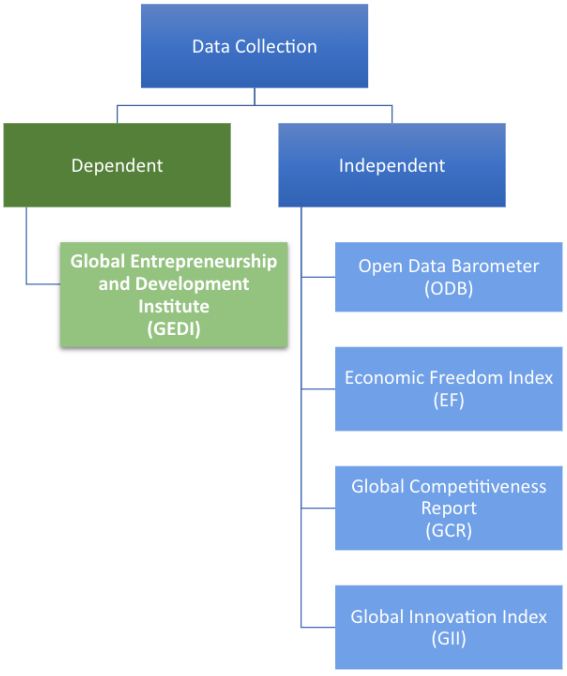


FIGURE 2.3: Displays our data selection and composition.

In the next section, we explain the composition and the methodological collection process for each index.

[9] https://www.weforum.org/reports/the-global-competitiveness-report-2017-2018
[10] https://www.globalinnovationindex.org/Home

Dependent Variable

Entrepreneurship

We collect data from the Global Entrepreneurship and Development Institute (GEDI) which is an indicator measuring entrepreneurship factors at national and regional levels around the world. The sample collected involves 137 economies from 2013 to 2016. This index is composed of 3 main components labeled as entrepreneurial attitudes, abilities, and aspirations. Each of these 3 components contains a set of different entrepreneurial features. The index also shows how nations are performing in terms of these components and ranking among them.

Index	Entrepreneurial Attitudes	Entrepreneurial Abilities	Entrepreneurial Aspirations
The Global Entrepreneurship and Develop Index	1.-Opportunity recognition 2.-Start-Up skills 3.-Risk acceptance 4.-Networking 5.- Cultural support	6.-Opportunity start-up 7.-Technological absorption 8.-Human capital 9.-Competition	10.-Product innovation 11.-Process innovation 12.-High growth 13.-Internationalization 14.-Risk capital

FIGURE 2.4: GEDI Index and its 14 pillars composition.

The pillar of attitudes which refers to the set of beliefs and perceptions that individuals and society have about identifying entrepreneurship opportunities, having the abilities or expertise to start a business, connecting with other entrepreneurs, and identify risks during all this process. This pillar is composed of 5 sub-pillars categorize as 1).- Opportunity recognition, 2).- Start-Up skills, 3)Risk acceptance, 4).- Networking and 5).- Cultural support. The first sub-pillar refers to how individuals recognize business opportunities and how entrepreneurs perceive the institutional environment to realize these ideas. The second sub pillar, quantify the skills and education that individuals should have to start a business. The third sub-pillar measures the risk that potential entrepreneurs have to take to start a business and the role that institutions play in this decision process and risk analysis. Networking is another indicator quantified in this pillar and it is related to how entrepreneurs create connections and how they located at the country level. The fifth attitude measure is the cultural support which refers to how intrinsic values in family, society or government affect the decision to become an entrepreneur.

Entrepreneurial abilities relate to the entrepreneurs' features (e.g. age, education, gender) and business characteristics (e.g. industrial sector, demographic, legal structure). The structure of this pillar is constituted of 4 sub-pillars (following the previous sub-pillars numbers) 6).- Opportunity start-up, 7).- Technological absorption, 8).- Human capital, 9).- Competition. Opportunity startup quantify the people's intention to start a business but they face administrative or bureaucratic constraints. Technological absorption is a metric to quantify businesses that are in technology sectors due to this area is considered as an angular field on economic development. The Human capital indicator measures individual characteristics such as the education, labor experience, healthy workforce. According to entrepreneurial literature, education level plays a crucial role because it is argued that individuals with high education are more capable to begin, manage, and/or growth business. Competition measures how companies create or develop a differentiator in terms of a business's product and market.

Entrepreneurial aspiration reveals the purposes or ideas to create new products or services or innovate them and grow as a company. This indicator also measures the chance to enter in new markets or expand their presence in them and the possibility to have external funds such as venture capital. This entrepreneurial feature has 5 sub-pillars composed of 10).-Product innovation, 11).- Process innovation, 12).- High growth, 13) Internationalization, 13).- Risk capital. The first sub-pillar of this component that quantifies the potential at the country level to create new products or innovate them. It is claimed by entrepreneurial authors that the development of new or innovative products is an important indicator for each economy that can be reflected through patents. The process innovation sub-pillar quantifies the level of technology applied or developed as a part of the production process in a company. It is argued in the GEDI report that there is a difference across economies because developed countries tend to create new technology and implemented in their business and developing economies tend to buy or copy it. The sub-pillar high growth point out the companies' projects and strategies to hire more people and future plans to expand as a company in more than 50% during the next 5 years. The internationalization sub-pillar is considered as a proxy measure of companies growth because it demands more companies capabilities in terms of infrastructure and human capital. This metric is related to the level of companies that has the potential to export at regional and country level. Risk capital is an important indicator that quantifies the availability of risk finance which is an important and additional source of financial support for entrepreneurs.

Finally, the GEDI indicator is developed by a consortium based on Washington D. C. and founded by academic institutions such as George Mason University, University of Pécs and Imperial College London. Furthermore, this indicator is funded by the European Union, World Bank, and other institutions interested in entrepreneurship determinants at the country level. The GEDI index is especially suitable for our study because this indicator incorporates a policy development focus that is missing in other Entrepreneurship measures (Acs et al., 2017; Szerb et al., 2013).

Independent Variables

Open Data

Concerning our independent variables, the sample is composed of The Open Data Barometer (ODB) which is an index developed by the World Wide Web Foundation. The aim of this indicator is to measure and compare the adoption and impact of open data initiatives around the world. This measure is composed of 115 economies during the period 2013 to 2016. The structure of this indicator has 3 main components labeled as readiness, implementation and impact and it has 10 sub-pillars that helps to quantify the impact of open data initiatives around the world.

Index	Readiness	Implementation	Impact
The Open Data Barometer	1.Policy and data management approaches 2.Government action at the national and subnational level 3.-Civil rights and the role of citizens 4.-Business and Entrepreneurship	6.Open Government Data availability 7.Open Government Data quality	10.Transparency and accountability 11.Environmental impacts 12.Contribution to economy and support to startups

FIGURE 2.5: ODB survey composition.

The component of Readiness measures how qualified are government designing and adopting open data initiatives and what kind of policies are around open data and government actions, civil rights, business and entrepreneurship. This component has 4 sub-pillars named as 1).- Policy and data management approaches. 2).- Government action at the national and subnational level. 3).- Civil rights and the role of citizens. 4).- Business and Entrepreneurship. The first sub-pillar captures the regulatory framework implemented by governments in order to guarantee the sustainability of open data policies. The second sub-pillar refers to the involvement not only of federal public bodies but also on the state and municipal level in order to permeate this open data initiative. The third sub-pillar point out the civil rights and commitment to participate and collaborate on the development of open data policies. The forth sub-pillar involves the sector private involvement through the business and entrepreneurs participation in consuming or producing open data.

The Implementation component measure not only the level of government data published but also the degree of accessibility, openness and timely. This component is composed of 6)

Open Government Availability. 7).- Open Government Quality. The former measure whether the data is available from government in any form, free of charge and containing a license. The latter whether is provided in a machine-readable format, and whether the data is up to date and containing metadata.

The Impact indicator quantify whether the data released by governments have a practical benefit to society. This indicator has the following structure 8).- Transparency and accountability. 9).- Environmental impacts. 10).- Contribution to economy and support to startups. The first sub-pillar quantify the how open data has an effect in terms of governmental transparency and accountability at the country level. The second sub-pillar refers to how the data released by governments (e.g. carbon emissions, emission of pollutant) has an impact on the environment. The third sub-pillar measure the contribution of open data to the economy and startups development.

All this information is collected and analysed by researchers, civilians and government representatives that are involved in the Open Data movement. This index is funded by the Open Data for Development (OD4D) a programme supported by different international institutions such as the World Bank, United Kingdom's Department for International Development (DFID) and the International Development Research Centre (IDRC) and Global Affairs in Canada.

Economic Freedom

The Economic Freedom Index (EFI) produced by The Heritage Foundation is an indicator that focuses on 4 pillars categorized as rule of law, government size, regulatory efficiency, and market openness which are determinants that governmental entities regularly use it as policy control. This index has 12 sub-pillars which are indicators that measure individual autonomy to choose, acquire and use economic resources and goods. The data collection involved 137 economies during the period 2013 to 2016.

Index	Rule of Law	Government Size	Regulatory Efficiency	Open Markets
The Economic Freedom Index	1.Property rights 2.Government integrity 3. Judicial effectiveness	4.Government spending 5. Tax burden 6. Fiscal health	7. Business freedom 8. Labor freedom 9. Monetary freedom	10.Trade freedom 11.Investment freedom 12. Financial freedom

FIGURE 2.6: Economic Freedom pillar composition.

Property rights which is a sub-pillar of the component Rule of Law is a metric that evaluates the legal framework in the country level that provides juridical certainty provided by governments (e.g. private property, copyrights, expropriation of property). Government integrity is another sub-pillar that is part of the Rule of Law component due to corruption is an important concern and factor that affects public entities, policy development and decision-making process globally. Furthermore, it is argued by the literature that corruption and lack of government integrity tend to increase bureaucracy, production costs and uncertainty in economic relationships. Judicial effectiveness which is an additional component of the Rule of Law is a measure of the efficiency of the legal framework and judicial system in order to guarantee the laws are applicable and respectable some examples of this metric are judicial independence and quality of the judicial process.

Government size component is composed of Government spending which quantifies the government consumption by the state. This indicator varies across countries because of their size and location. However, research evidence on this topic reveals that unnecessary spending and/or a waste of resources causes a government deficit and it is a serious economic issue. Tax burden is another component of Government Size that captures the rates imposed on individuals and corporations and the general taxation level as a percentage gross domestic product (GDP). Fiscal health is an indicator that estimates government management and effective spending because an inefficient process in these areas are related to macroeconomic and social problems.

Regulatory Efficiency is the third component of the Economic Freedom Index. Business freedom is a sub-pillar that quantify the regulatory and infrastructure frameworks that inhibit or promote business operations. Some elements of this indicator are starting a business process (time, cost, administrative bureaucracy), closing a business, getting licenses. The next sub-pillar is Labor freedom which captures the extent to which the regulatory labor framework (e.g. wages, hiring process, labor laws), labor force participation and employment opportunities varies across countries. Price stability is an important indicator of an economy because it is intimately related to inflation, these (price and inflation) are crucial elements of the Monetary freedom indicator.

Open Markets component contains Trade freedom sub-pillar that captures the imports and export of good and services and their regulatory restrictions. Investment freedom is an index that measures that in a free society, economic agents (e.g. individuals and firms) would not have restrictions to invest or allocate resources in any legal activity. Financial freedom measure the level of government control, banking efficiency, and financial development in an economy. It is argued by the literate that independent central banks play an important role in terms of regulatory financial frameworks. Furthermore, the position of banks promoting credits and is also crucial for economic mobility.

Finally, this index is useful for our study because it reflects the economic and entrepreneurial ecosystem at the country level through the twelve sub-pillars that emanates from the main four elements that composed this index (Dıaz-Casero et al., 2012).

Global Competitiveness Report

The Global Competitiveness Report (GCR) published by the World Economic Forum is an index that measures and compares competitiveness at the country level. We collected data from 2013 to 2016 in 137 economies. Furthemore, the structure of this indicator is composed of 3 pillars that determine competitiveness as the level of productivity of an economy through quantifying factors such as institutions, policies, infrastructure, education, and other determinants displayed in figure 2.7. We considered relevant to our study this index because as stated in by the literature the elements of these pillars (e.g. strong of institutions, policy development, human capital) are part of the determinants of entrepreneurship in any economy (Korez-Vide & Tominc, 2016).

Index	Basic Requirement Factor Driven	Efficiency Enhancers	Innovation and Sophistication Factors
The Global Competitiveness Index	1.Institutions 2.Infrastructure 3. Macroeconomic environment 4. Health and primary education	5. Higher education and training 6. Goods market efficiency 7. Labor market efficiency 8. Financial market development 9. Technological readiness 10. Market size	11. Business sophistication 12. Innovation

FIGURE 2.7: Global Competitiveness Index pillar composition.

The institutional sub-pillar measure the judicial and administrative framework in which economic agents (households, firms, and governments) interact. Furthermore, this framework is considered as a crucial factor in the quality of public bodies at the country level. Infrastructure is another sub-pillar considered as a factor driven by competitiveness. This agent plays an important role because it provides support for economic development through the mobility of good, services and human capital. Examples of necessary infrastructure are ports, air

transport, roads, telecommunication, electricity, and water networks. Another determinant of competitiveness is the macroeconomic solidity that a country shows for global markets. High rates of government deficits, inflation and lack of credits are perceived as insecure signals for investors, venture capitals and entrepreneurs. The welfare of the working class and basic education also are determinants considered as competitiveness for any country.

Regarding the Efficiency Enhancers component, the first sub-pillar labeled as Higher education and training is captured through secondary and tertiary enrollment rates. This indicator is an important determinant because globalize competence require economies based on knowledge development produced by education and training. Goods market and efficiency sub-pillar refers to the capability of not only producing but also balancing services and products according to the country level demand and supply conditions. The Labor market efficiency indicator captures the allocation and labour mobility of individuals. The Labor market efficiency indicator captures the allocation and labour mobility of individuals. The labour efficiency is captured through the relationship between employers and employees and the incentives that this connection produces in terms of better salaries, productivity, reduce unemployment. The Financial market development refers to an efficient allocation of resources, in particular to the business development because it is crucial for productivity and economic growth. Moreover, this indicator also measures the regulatory framework that allows for capital availability, financial loans, securities exchanges, and investor protections. The Technological readiness pillar measures how technology is adopted for industrial processes in order to improve productivity at the country level. In particular how companies have access to and use of information and telecommunication infrastructure. Market size measures at the country level the capacity to consolidate the domestic market and the opportunity to penetrate foreign markets which both indicators are considered by the literature as externalities of productivity.

The component of Innovation and Sophistication Factors is composed of Business sophistication that measures the quality of business networks and the quality of individual firms' operations and strategies. The indicator quantifies the cluster formed in different entrepreneurial sectors and it is composed of the quantity, quality, and interaction of domestic suppliers. The Innovation sub-pillars is an indicator that contains information about investment in research and development (R&D) at the country level, one of its dimensions is the private sector in terms of the application of new technologies. Another dimension is the academic sector and how they are developing new research collaborations and knowledge that will be the basis for new technologies or patents.

Global Innovation Index

The Global Innovation Index (GII) co-published by Cornell University, INSEAD, and the World Intellectual Property Organization (WIPO) is an index structured of 2 main components and

7 sub-pillars that measures innovation performance at the country level. Our sample is composed of 127 economies during the period 2013 to 2016.

Index	Innovation-Input	Innovation-Output
The Global Innovation Index	1.Institutions 2.Human capital and research 3. Infrastructure 4. Market sophistication 5. Business sophistication	6. Knowledge and technology output 7. Creative outputs

FIGURE 2.8: Global Innovation Index input and output composition.

The first component labeled as Innovation-Input is related to national economy indicators which are considered determinants of innovation. This component is composed of 5 sub-pillars identified as 1) Institutions 2) Human capital and research 3) Infrastructure 4) Market sophistication 5) Business sophistication. The Institutions sub-pillars measure how public entities are stables in terms of legal, political and social frameworks (e.g. avoiding social disturbs, political violence, and terrorism). The Institutions sub-pillars measure how public entities are stables in terms of legal, political and entrepreneurial frameworks (e.g. avoiding social disturbs, political violence, and bureaucracy). This sub-pillar is composed of data that captures the political, regulatory, and business environment in the country level. The Human capital sub-pillar measure how governments around the world invest in improving the level and quality of education through measuring aspects such as school life expectancy, pupil-teacher ratio, tertiary enrolment and mobility, graduates in science and engineering. Furthermore, this sub-pillar also capture the ratio of researchers and the proportion of expenditure (% GDP) in R&D per country. The Infrastructure sub-pillar is composed of several elements such as government online services and participation, fixed telephone subscriptions, mobile subscriptions, Internet bandwidth, the percentage of households with a computer and percentage of households with Internet access. Furthermore, this index also captures other types of infrastructure such as electricity production, transportation (ports, airports, railroads, highways) in order to measure logistic performance. The next sub-pillar named Market sophistication measures at the country level various aspects such a how easy is to get a credit, the domestic credit for the private sector and microfinance institutions gross loan portfolio. This indicator also captures information about protection for investors in terms of regulatory frameworks, market capitalization, the value of stock traded, and Venture capital

deals. Moreover, it takes into account additional indicator such as trade, competition, and market scale. The Business sophistication indicator is composed of different types of measures such as knowledge workers which are related to the type of job that demands skilled workers. In addition, firms training, Gross expenditure on R&D performed and financed by businesses enterprise, Females employed with advanced degrees. Other indicators of this sub-pillars are Innovation linkages which refer to the connection and alliances between the private and academic sectors, this section also considers quantify the number of patents. Knowledge absorption is another measure of this indicator and it is related to intellectual production, the number of high-tech imports and research talent in the business enterprise sector.

The second component named as Innovation-Output. This indicator is composed of 2 sub-pillars identified as 6) Knowledge and technology output 7) Creative outputs. The former is a sub-pillar that measures resident patents applications per country, the international patents are measured implementing the Patent Cooperation Treaty (PCT). Scientific, technical publication and citable documents using the H index are other measures of this sub-pillar quantifies knowledge output. Knowledge impact is also part of this sub-pillar and it quantifies elements such as new business density, total computing software spending, the implementation of quality certifications (e.g. ISO 9001) and high and medium tech manufactures output. Knowledge diffusion measure is composed of intellectual property receipts, high tech, and ICT services exports. The latter sub-pillar identified as Creative outputs is measured through intangible assets such as trademark applications, industrial designs, ICTs and business model creation, ICTs and organizational model creation (e.g., virtual teams, remote working, telecommuting) within companies). Another element of this sub-pillar is creative good and services indicator which is structured with these indicators creative goods and service exports (e.g. Advertising, market research, and public opinion polling services), national feature films produced, global entertainment and media market. Online creativity is a measure that is composed of the following data generic top-level domains (gTLDs) (e.g. (com, info, net, and org). Country-code top-level domains (ccTLDs) (e.g. us. uk. mx) which are a classification given by the Assigned Numbers Authority (IANA) for use on the Internet. Other indicators of this sub-pillar are Wikipedia monthly edits and Video uploads on YouTube.

Finally, all the information provided in this index is relevant to our analysis because innovation is considered by the literature as a trigger for entrepreneurship (Crumpton, 2012). A summary and conceptual map including all variables, pillars and subcategories can be found in the Appendix section A.2 as dependent and independent variables.

The final dataset has 474 variables (including our response indicator and control variables) and 548 observations during the period 2013 to 2016. This information also contains categorical values such as -region and income group -. The purpose to include these labels is to implement a clustering analysis per region and/or income group in our research and compare

the effect of open data. In the next section, we describe the stages of the pre-processing of our data sample and we will explain the importance of these tasks.

Pre-processing

Once the phase of data acquisition is completed, the next stage is data preprocessing which is an important process because we can identify elements such as missing values, outliers or others noise elements that can affect the quality of our model (Kotsiantis, Kanellopoulos, & Pintelas, 2006). This stage involves the following tasks: cleaning, extracting and selecting features that we will use in our model.

Cleaning Data

During the preprocessing stage, we cleaned the data (we checked for consistency in terms of integers because some values had commas instead of numbers e.g. 1,2 to 1.2) and we explore missing values for each attribute. These tasks are crucial because they help us to better understand the structure of the data sample and we can identify issues such as how many values are missed for each indicator. Missing values can affect the performance of our models creating incorrect results. In our dataset, we found that our missing values follow the pattern labeled as Missingness At Random (MAR) which is referred as *"the probability of a variable is missing depends only on available information"* (Gelman & Hill, 2006). A summary of missing values per each indicator is presented in Appendix A.3

In order to solve our missing values, we imputed data using the mice package in R (Buuren, Buuren, & Groothuis-Oudshoorn, 2011). This library allows us to correct the problem of missing values through a predictive mean matching method. The method included in this library generates multiple imputations for multivariate missing data for continuous, binary, unordered and ordered categorical data. An additional step in the cleaning process was to rename the variables adding a prefix that indicates the name of each index. This step was done in order to manage and recognize each variable according to their index, this is useful for clustering purposes.

Extract Features

After we cleaned and dealt with missing values in our dataset, the next step is extracting features that could be relevant to our dependent variable. A feature refers to an attribute of data that represent individual data objects (Dong & Liu, 2018). Feature extraction can be explained as the process of dimensionality reduction from a large number of attributes without losing meaningful information by removing redundant data. This reduction process helps to identify features that could be are more conducive to our analysis (Liu & Motoda, 1998).

An important consideration in this feature extraction process is that we have a big set of control variables (474 features) to analyze and these attributes are related to our target variable (entrepreneurship) at different levels. In order to explore and understand these relationships, we decided to implement the Principal Component Analysis (PCA) algorithm. This is a machine learning algorithm which aim is to reduce dimensionality, extract relevant information from a multivariate dataset and generate a new set of variables called components. This method is very useful when there is redundancy in the data (correlation between variables). Moreover, this technique helps to scale (normalize) the data in order to ensure us that we are comparing same dimensions. The PCA algorithm implements a linear combination using a weighted average of a set of variables identifying directions also referred as index or components.

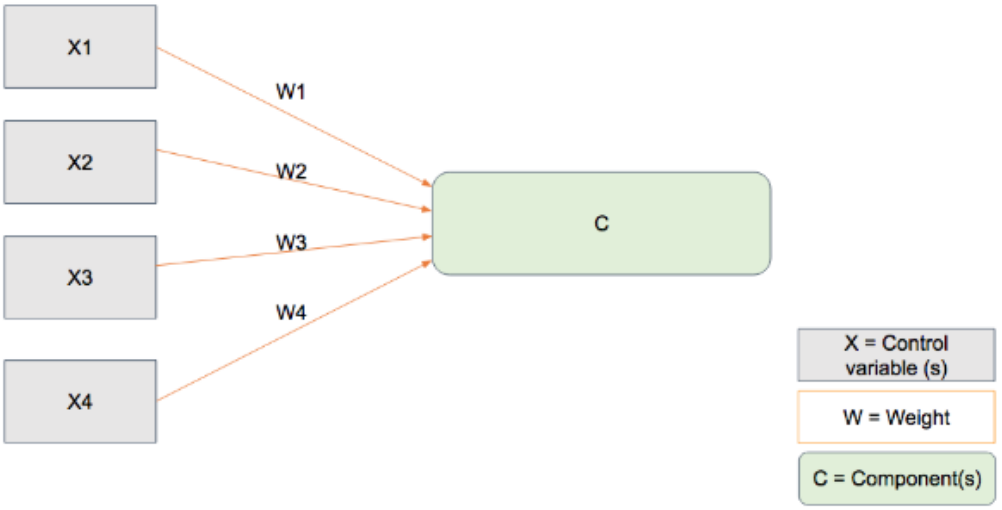


FIGURE 2.9: Illustrates the creation process of an index.

Our aim implementing PCA into our dataset is to extract the main components that are related to entrepreneurship. In order to do so, we need to define in the PCA algorithm our target variable (entrepreneurship) as our supplementary variable and our control variables as active variables. The purpose of these declarations is to establish the relationship between dependent and independent variables. Then, we run our PCA in our selected dataset, the data standardization process is made automatically for the PCA algorithm.

Our results show that Dimension 1 (Dim1) is the main component because it contains 38.6% of the variance of our independent indicators. The amount of variance suggested in this component is measured by the term eigenvalue. The PCA algorithm tends to display larger eigenvalues for the first dimensions and smaller for the following components; therefore, the first dimensions contains the most relevant variances to our dependent variable. Figure 2.10 illustrates this result and the indicators that compose the Dim1 according to their correlation and p.value are described in the Appendix section A.4.2

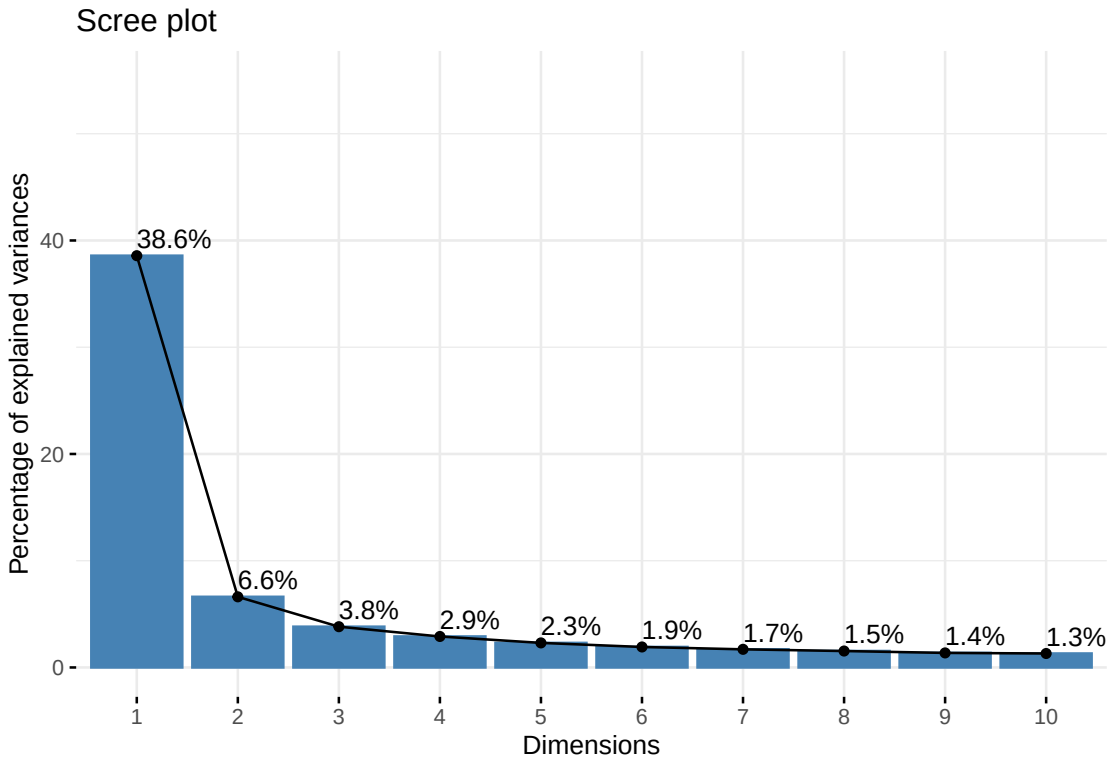


FIGURE 2.10: Displays the number of dimensions created using PCA

Despite the fact that we were able to extract the indicators that are related to entrepreneurship, just the main 4 components together (Dim 1, Dim 2, Dim3 and Dim4) represent the 51% of the variance, this suggests that our components still have high dimensionality.

In the next section, we are going to reproduce this PCA process but this time using a subset of variables based on the literature review, the aim is to narrow and improve the dimensionality in order to analyze its relationship and effect with entrepreneurship.

Select Features

In order to narrow our analysis finding a group of indicators that are relevant to entrepreneurship, we created a subset (from our main dataset) of control variables based on the literature review (This step in the data science approach is considered as the use of domain knowledge or domain expertise). Then, we created 6 indexes that we labelled as open data, competitiveness, innovation, economic freedom, open government partnership, and Income group. Our objective is to systematically test these indices adding open data as a supplementary component in order to analyze the effect of all these variables on entrepreneurship. Appendix A.4.3 shows the composition of our dependent and selected variables of control.

Once our data selection was made based on these 6 indexes, our sample is composed by 97 variables and 548 observations. The next step is to define in the PCA algorithm our target variable (entrepreneurship) as our supplementary variable and our control variables as

active variables. The purpose of these declarations is to establish the relationship between dependent and independent variables. Then, we run our PCA in our selected dataset, the data standardization process is made automatically for the PCA algorithm. Our results show again that Dimension 1 (Dim1) is the main component. It contains 44.3% of the variance of our variables; therefore, this approach implementing a subset of variables selected based on the literature review improved our previous results. Figure 2.11 illustrates this result and the composition of the Dim1 is described in the Appendix section A.5. amd A.5.1

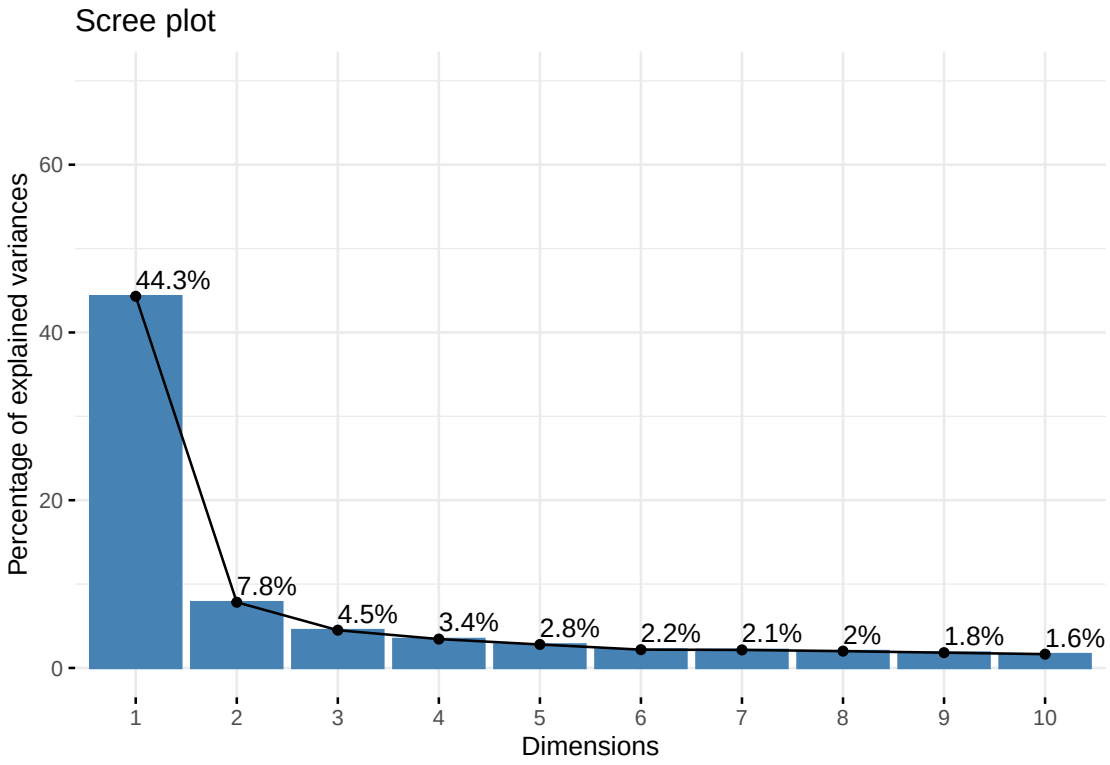


FIGURE 2.11: Shows the number of dimensions from subset

Once we have this list of variables, the next step is to test and analyze the relation of those control variables on entrepreneurship using a multiple regression technique. In the next section, we will start to develop our testing and analysis.

Modelling Process

During the modeling process stage, we developed and compared several models using our dataset generated in our previous stages. We used descriptive multilinear regression techniques such as stepwise forward regression, stepwise backward regression, and stepwise regression in order to select the most appropriate model in terms of statistical parameters and knowledge domain.

Generate Models

The following step in our research pipeline is to systematically test the relation of our selected independent variables on our target or dependent variable For this purpose, we implemented a feature selection approach adopted in econometrics (Kapetanios, Marcellino, & Papailias, 2014) and machine learning fields (Guyon & Elisseeff, 2003; Hall & Smith, 1999). Feature selection or variable selection refers to the process of selecting relevant features for using in model construction (James et al., 2013a). The variable selection procedure was performed using different machine learning algorithms called stepwise forward regression, stepwise backward regression and stepwise regression in order to generate different models and analyse its results.

The stepwise forward regression algorithm starts with a constant term; then, the algorithm begins including variables iteratively in the equation based on p values until all variables in the model are tested. The stepwise regression starts implementing a full equation of all variables; then, it drops variables based on their p values until there is no variable to analyse. Stepwise regression builds a regression model from a group of variables by adding and removing predictors based on p values, in a stepwise manner until there is no variable left to enter or remove any more (Chatterjee & Hadi, 2015). The detailed results containing the R-Square, Adj. R-Squared, Mallow's Cp, Akaike's information criterion (AIC) and root mean square error (RMSE) of the implementation of these 3 algorithms (stepwise forward regression, stepwise backward regression, and stepwise regression) are illustrated in Appendix section A.6.

In the next section, we analyze these results based on the variables predicted for these 3 algorithms and the level of collinearity among variables.

Selection of the Final Model

The results generated using stepwise forward regression and stepwise backward regression are a list of 43 indicators in which the variables related to the Innovation index are the predominant component, followed by competitiveness, economic freedom, income group and open Data. Figure 2.12 illustrates the number of observations selected per these algorithms

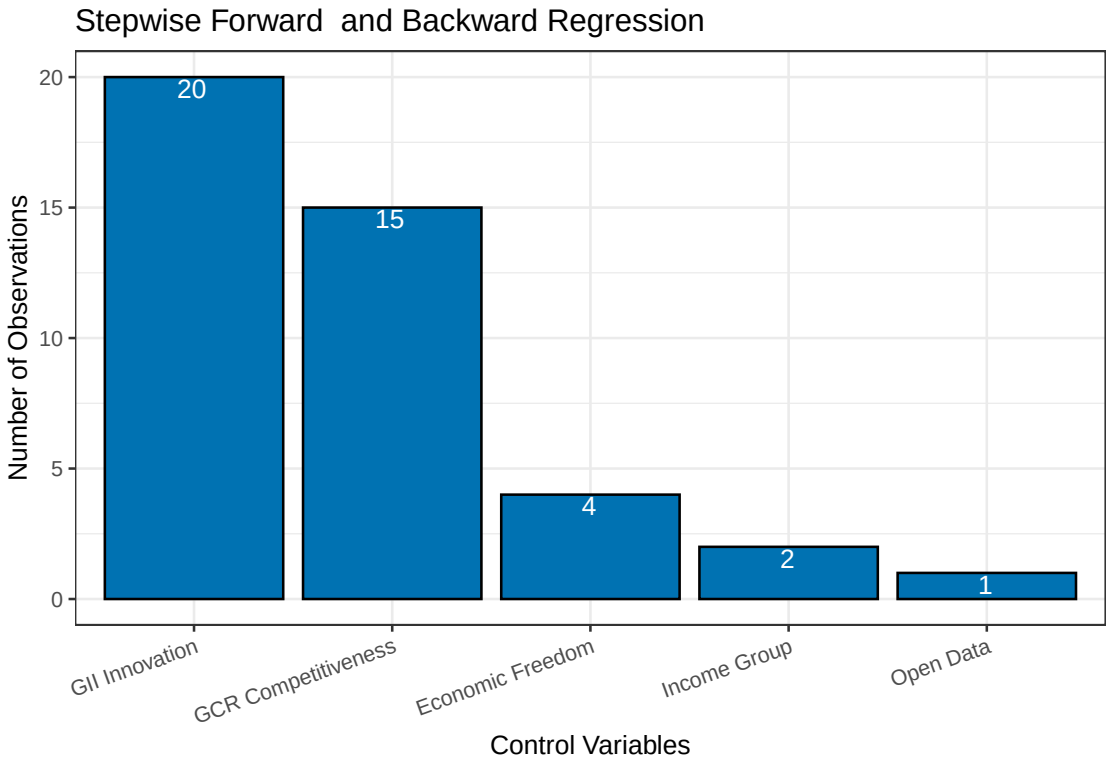


FIGURE 2.12: Illustrates the number of observations selected

However, after analyzing these independent variables through a collinearly diagnostic, we find that 21 variables have a high level of variance inflation factor (VIF) which measures the amount of multicollinearity in a set of variables. Multicollinearity refers to a high level of association among the independent variables. Figure 2.13 shows the variables that contain a VIF value higher than 6 in our set of independent variables.

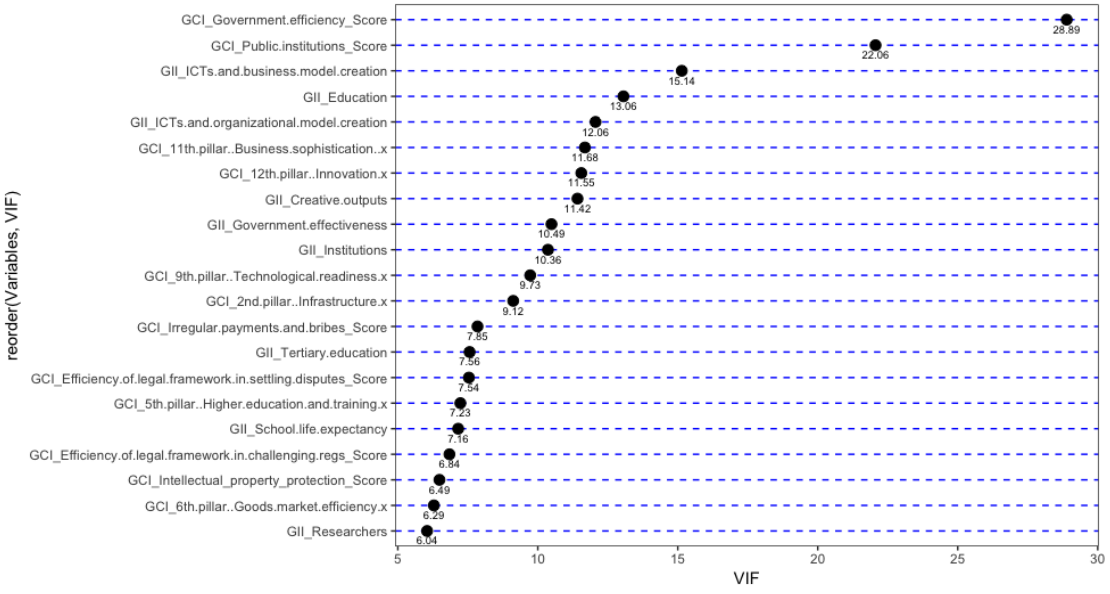


FIGURE 2.13: Displays VIF values using stepwise forward and backward regression.

In contrast, the result using the Stepwise Regression algorithm shows a list of 32 indicators in which variables composed for the innovation index is also the main component, followed for competitiveness, and then, economic freedom, open data, and income group.Figure 2.14 illustrate the Stepwise results.

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	GCI_9th.pillar..Technological.readiness.x	addition	0.567	0.566	405.7130	4241.2193	11.5553
2	IG_High_income	addition	0.630	0.628	270.0650	4157.4133	10.6949
3	GII_Research.and.development._R_D	addition	0.670	0.668	184.4800	4096.8726	10.1110
4	GCI_5th.pillar..Higher.education.and.training.x	addition	0.684	0.682	154.1420	4073.8487	9.8918
5	ODB_Score.Scaled	addition	0.694	0.692	133.9710	4058.0077	9.7411
6	GCI_Efficiency.of.legal.framework.in.challenging.regs_Score	addition	0.704	0.701	115.3790	4042.9005	9.5991
7	GII_Infrastructure	addition	0.709	0.705	106.7970	4035.8972	9.5294
8	GCI_9th.pillar..Technological.readiness.x	removal	0.708	0.705	105.2720	4034.3043	9.5241
9	GII_Creative.goods.and.services	addition	0.713	0.709	96.9690	4027.4007	9.4558
10	GCI_2nd.pillar..Infrastructure.x	addition	0.717	0.712	91.3030	4022.6802	9.4067
11	EF_Business.Freedom	addition	0.721	0.716	84.0540	4016.4612	9.3451
12	GCI_Transparency.of.government.policymaking_Score	addition	0.724	0.719	79.0040	4012.1127	9.2998
13	GCI_6th.pillar..Goods.market.efficiency.x	addition	0.727	0.721	74.3690	4008.0690	9.2572
14	GII_Ease.of.starting.a.business	addition	0.731	0.725	68.2210	4002.5555	9.2026
15	GCI_Intellectual_property_protection_Score	addition	0.734	0.727	63.8230	3998.5829	9.1611
16	GII_Institutions	addition	0.736	0.730	59.8890	3994.9850	9.1230
17	GII_ICTs.and.business.model.creation	addition	0.739	0.732	55.6230	3991.0106	9.0819
18	GII_Intangible.assets	addition	0.743	0.735	49.0560	3984.7354	9.0221
19	GII_Education	addition	0.746	0.738	43.9730	3979.7978	8.9736
20	GII_Human.capital.and.research	addition	0.749	0.741	39.6780	3975.5578	8.9311
21	GCI_10th.pillar..Market.size.x	addition	0.751	0.742	37.7170	3973.5950	8.9073
22	GII_Tertiary.enrolment	addition	0.753	0.744	35.5490	3971.3929	8.8816
23	GII_Expenditure.on.education	addition	0.755	0.745	33.3810	3969.1592	8.8558
24	GCI_Efficiency.of.legal.framework.in.settling.disputes_Score	addition	0.757	0.746	31.5130	3967.2005	8.8323
25	GII_Political.stability.and.absence.of.violence.terrorism	addition	0.758	0.748	29.7850	3965.3575	8.8098
26	GII_Researchers	addition	0.760	0.749	28.6510	3964.1064	8.7921
27	GII_Research.and.development._R_D	removal	0.759	0.749	27.4400	3962.9263	8.7902
28	GCI_Public.institutions_Score	addition	0.761	0.750	26.3510	3961.7084	8.7728
29	GCI_Public.trust.in.politicians_Score	addition	0.762	0.751	25.1080	3960.3084	8.7540
30	GII_Market.sophistication	addition	0.764	0.752	23.9050	3958.9291	8.7355
31	GII_Human.capital.and.research	removal	0.764	0.752	22.2570	3957.3012	8.7301
32	GII_School.life.expectancy	addition	0.765	0.753	21.2970	3956.1630	8.7134

FIGURE 2.14: Shows variables generated using stepwise regression

Once we had evaluated the variance inflation factor for these variables, we found that the stepwise regression models shows less indicators with high level of VIF. Figure 2.15 illustrates the VIF diagnostic used in the Stepwise regression model

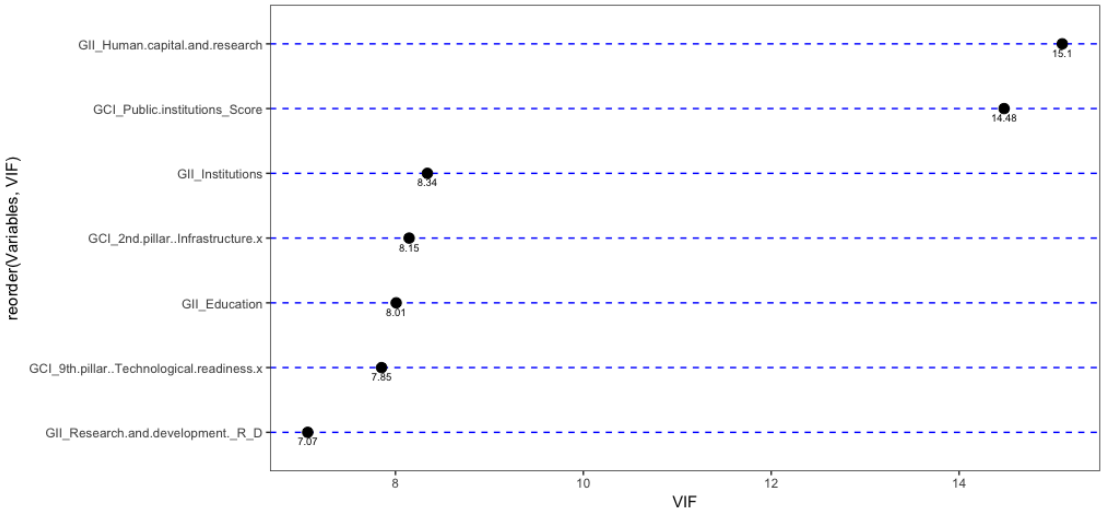


FIGURE 2.15: Displays VIF values using stepwise.

Although the previous algorithms were able to propose a list of potential variables, we found a high level of multicollinearity among some variables. The topic of multicollinearity is an ongoing research area in feature engineering due to its implications using big datasets (Garg & Tai, 2013; George, Osinga, Lavie, & Scott, 2016; John Lu, 2010). For this reason, we include in the next section a domain knowledge approach based on the theoretical background introduced in the literature, selecting variables proposed by these algorithms used in this section and removing multicollinearity.

Evaluation of the Final Model

In this section, we analyze and evaluate a model using entrepreneurship (E) as a dependent variable and the independent variables are composed of a set of indicators from different indexes such as global innovation (GII), competitiveness report (GCI), economic freedom (EF), income group (IG), and open data (ODB) based on our previous work. The model to be estimated is characterized by the linear model shown below:

$$E = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + ... + \beta_8 X_8 + \epsilon$$

Where:

E is the dependent variable and $\beta_1 + \beta_2 + ... + \beta 18$ are the marginal effects of variables $X1, X2, ..., X18$ on the dependent variable E

- E = GEDI
- X1 = IG_High_income
- X2 = ODB_Score.Scaled
- X3 = GCI_Efficiency.of.legal.framework.in.challenging.regs_Score
- X4 = GII_Infrastructure
- X5 = GII_Creative.goods.and.services
- X6 = EF_Business.Freedom
- X7 = GCI_Transparency.of.government.policymaking_Score
- X8 = GCI_6th.pillar..Goods.market.efficiency.x
- X9 = GCI_Intellectual_property_protection_Score
- X10 = GII_ICTs.and.business.model.creation
- X11 = GII_Intangible.assets
- X12 = GII_Education
- X13 = GCI_10th.pillar..Market.size.x
- X14 = GII_Tertiary.enrolment
- X15 = GII_Expenditure.on.education
- X16 = GCI_Efficiency.of.legal.framework.in.settling.disputes_Score
- X17 = GII_Researchers
- X18 = GII_Market.sophistication

We implemented the function lm() which is used to fit linear models using the open source software R (Chambers & Hastie, 1992; Wilkinson & Rogers, 1973)(Chambers & Hastie 1992) (Chambers & Hastie 1992; Wilkinson & Rogers 1973). Furthermore, we split the dataset into training and testing dataset (85% and 25% respectively) ordering the dataset randomly. The results of this model are presented in the figure 2.16

```
Call:
lm(formula = GEDIScore ~ IG_High_income + ODB_Score.Scaled +
    GCI_Efficiency.of.legal.framework.in.challenging.regs_Score +
    GII_Infrastructure + GII_Creative.goods.and.services + EF_Business.Freedom +
    GCI_Transparency.of.government.policymaking_Score + GCI_6th.pillar..Goods.market.efficiency.x +
    GCI_Intellectual_property_protection_Score + GII_ICTs.and.business.model.creation +
    GII_Intangible.assets + GII_Education + GCI_10th.pillar..Market.size.x +
    GII_Tertiary.enrolment + GII_Expenditure.on.education + GCI_Efficiency.of.legal.framework.in.settling.disputes_Score +
    GII_Researchers + GII_Market.sophistication, data = gedi_model_df)

Residuals:
    Min      1Q  Median      3Q     Max
-54.344  -5.190   0.486   4.653  44.745

Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         -19.67564    4.38501  -4.487 8.87e-06 ***
IG_High_income                                       10.55814    1.25779   8.394 4.31e-16 ***
ODB_Score.Scaled                                      0.11559    0.02681   4.311 1.94e-05 ***
GCI_Efficiency.of.legal.framework.in.challenging.regs_Score   1.82330    0.79585   2.291 0.022355 *
GII_Infrastructure                                    0.21258    0.05130   4.144 3.98e-05 ***
GII_Creative.goods.and.services                      -0.09858    0.02992  -3.295 0.001052 **
EF_Business.Freedom                                   0.09149    0.03874   2.362 0.018550 *
GCI_Transparency.of.government.policymaking_Score    -1.22427    0.72416  -1.691 0.091501 .
GCI_6th.pillar..Goods.market.efficiency.x             5.97472    1.49954   3.984 7.72e-05 ***
GCI_Intellectual_property_protection_Score           -1.80524    0.62191  -2.903 0.003853 **
GII_ICTs.and.business.model.creation                 -0.28950    0.06278  -4.611 5.02e-06 ***
GII_Intangible.assets                                 0.18720    0.05059   3.700 0.000238 ***
GII_Education                                        -0.17736    0.05052  -3.511 0.000484 ***
GCI_10th.pillar..Market.size.x                        1.56732    0.40830   3.839 0.000139 ***
GII_Tertiary.enrolment                                0.11446    0.02453   4.666 3.89e-06 ***
GII_Expenditure.on.education                          0.08456    0.03346   2.527 0.011791 *
GCI_Efficiency.of.legal.framework.in.settling.disputes_Score   1.77977    0.73424   2.424 0.015686 *
GII_Researchers                                       0.12762    0.02657   4.803 2.04e-06 ***
GII_Market.sophistication                             0.16692    0.04773   3.497 0.000510 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.077 on 529 degrees of freedom
Multiple R-squared:  0.741,     Adjusted R-squared:  0.7322
F-statistic:  84.1 on 18 and 529 DF,  p-value: < 2.2e-16
```

FIGURE 2.16: Summarizes model results.

The structure of this result is explained as follows. Call refers to the composition of the linear model developed in which GEDIScore represent the dependent/target variable, the other indicators are the response/independent variables and the definition of the data used (gedi_model_df). Residuals are the difference between the observed data (GEDIScore) and the response values that the model predicted. This section describes 5 main points (Min, 1Q, Median, 3Q, Max). In the case of the Median, a value of zero (0) represent symmetrical distribution and it is considered as an estimation of how well the data fit into the model. Our model shows a Median of 0.18.

In a linear model, the Coefficients, $\beta_1 + \beta_2 + ... + \beta 18$ show the composition of variables that represent the rate of change of our dependent variable (y) as a function of changes in the independent variables (x). Coefficients also illustrate the intercept and slope terms. The coefficient-estimates illustrate the effect that each variable has on entrepreneurship.

For instance, the slope term in our model is showing that for every unit increase of open data the effect on entrepreneurship is increased by 0.11. The standard error measures the variation of the Coefficient-Estimates from the actual average value of our control variables. For example, the effect of open data on entrepreneurship can vary 0.02. The Standard Error is also used to compute confidence intervals and to statistically test the hypothesis of the existence of a relationship between open data and entrepreneurship. The Coefficient-t value is calculated dividing the Coefficients by its Standard Error. If these values are far away from zero indicating then this indicates that we can reject the null hypothesis which states that the marginal effect of a variable X is equal to zero. If we reject the null hypothesis then the marginal effect of X on entrepreneurship statistically different from zero. In our case, the value that open data shows is 4.03, hence we find a relationship that is statistically speaking different from zero between open data and entrepreneurship. The Coefficient Pr($>$t) which refers the probability of observing values equal or larger than t. it is also known as p-value. Although there is some controversy about its interpretation (Kuffner & Walker, 2017), it is generally accepted that a p-value of 5% or less could be used as an indicator to reject the null hypothesis. In our model, the p.value is 0.0000653 (6.53e-05) which means that we reject the null hypothesis, concluding that statistically speaking there is a positive relationship between open data and entrepreneurship. That is to say, increases in open data are correlated with increases in the index of entrepreneurship.

The Residual standard error section illustrates the quality of a linear regression. Linear models might not able to capture the universality of factors that could affect our target variable; therefore, we can not perfectly predict it. For this reason, the linear models contain an error term. The Residual standard error is the average amount that our control variables will deflect from the true regression line. The degree of freedom refers to the observations that were involved in the estimation of parameters.

The Multiple R-squared provides a value that illustrates how well our selected variables fit the observed data. The range value of this measure lies between 0 and 1 (a 0 means an absence of variance in the independent variables and a number equal to 1 represent a perfect fit of our estimated values of entrepreneurship with the observed values of entrepreneurship. An important point of this indicator is that when you include an additional control variable the Multiple R-squared value tends to increase because the new independent variable will explain some proportion of the variance. The Adjusted R-squared controls this increase and include penalties for the number of independent variables. The inclusion of more independent variables implies the loss of degrees of freedom because an additional parameter needs to be estimated. This changes the estimates of the t values of the inference and hypothesis testing. Our model shows a Multiple R-squared value of 0.72. This should be interpreted in the following manner: the estimated model explains as much as 72% of the variance found in our target variable. The Adjusted R-squared value is 0.71 which shows a symmetry to the Multiple R-squared value.

The F-statistic value indicates the presence or absence of a relationship between the target and all of the independent variables considered in the model. While the t value is a test of significance of a particular independent variable the F indicator is used to test the model as a whole (it tests the significance of all the independent variables considered in the model). The null hypothesis is that all values of $\beta_1 = \beta_2 = ... = \beta 18 = 0$ which means the model can not explain the observed variations of entrepreneurship. The alternative hypothesis is $\beta_1 \neq 0, \beta_2 = 0, \neq ... = \beta 18 \neq 0$ meaning that the changes in all the independent variables considered in the model are correlated with changes in entrepreneurship.. In our model, the F-statistic value is 66.76 that means that variations in all of the independent variables are statistically correlated with changes in entrepreneurship independent.

As a complement of our results description, we use the R package "olsrr" (Hebbali, 2017) in order to analyze and evaluate visually the behavior and structure of our model. For instance, the Residual vs Predicted values plots display nonlinear, outliers and irregular variances. The QQ plot exposes violation of normality assumption. The Residual Fit compares side by side the Fit-Mean and Residuals charts. These graphs illustrate how much variation in the data is explained by the fit and how much remains in the residuals. For unsuitable models, the spread of the residual is often greater than the spread of the centered fit. In our case, there is a symmetry between both results. The plot showing the Observed vs Predicted for GEDIScore evaluate how the model fit. An example of a good model is in which one that points tend to fit the diagonal line displayed in the graph. In our case, the R square is .72 and tend to fit the line. The Cook's Distance (Cook's D) Chart represents the influential outliers in our model. This distance is calculated using each observation's leverage and residual values. Our model shows some outliers that potentially could have a negative effect on our results. Finally, the histogram show a relatively fair symmetric distribution. Figures 2.17, 2.18 and 2.19 contains all the plots described.

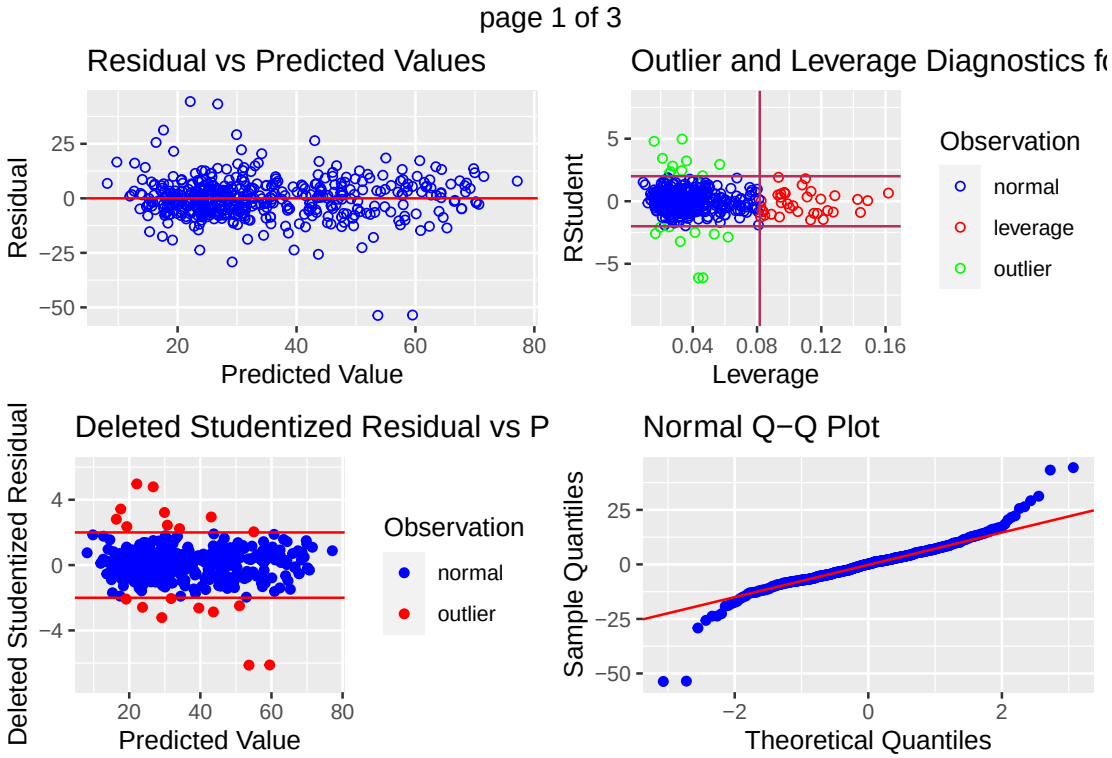


FIGURE 2.17: Shows the diagnostic of our model

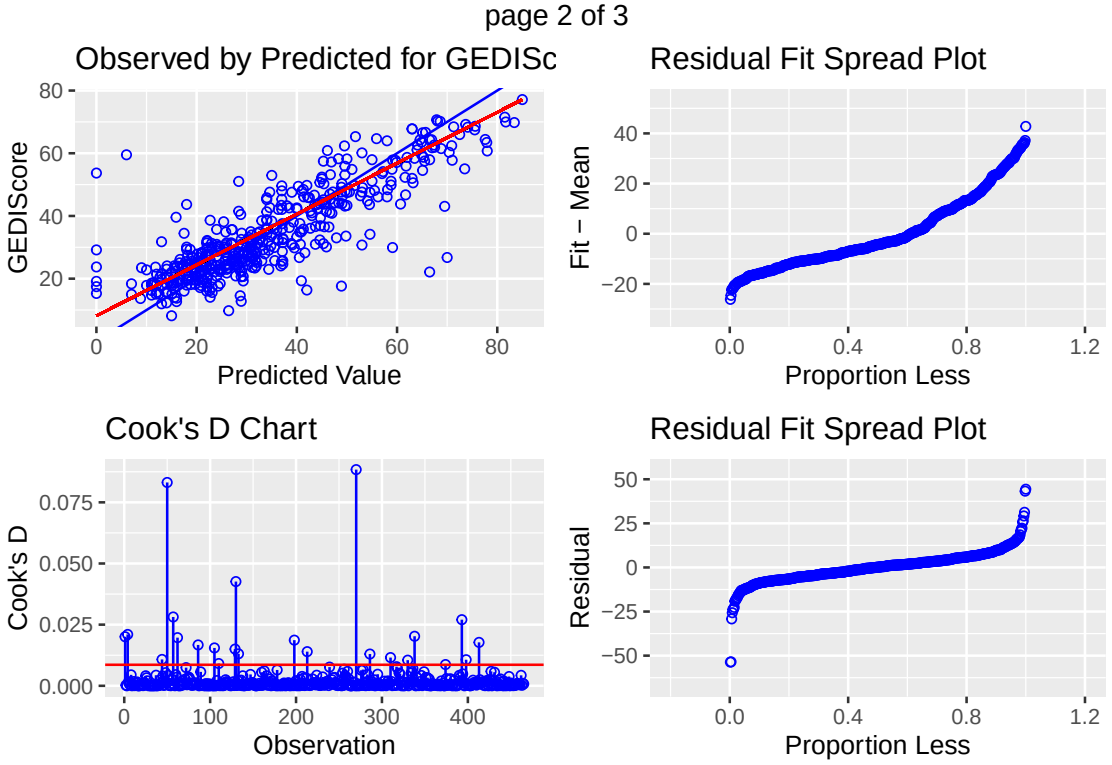


FIGURE 2.18: Shows the diagnostic of our model

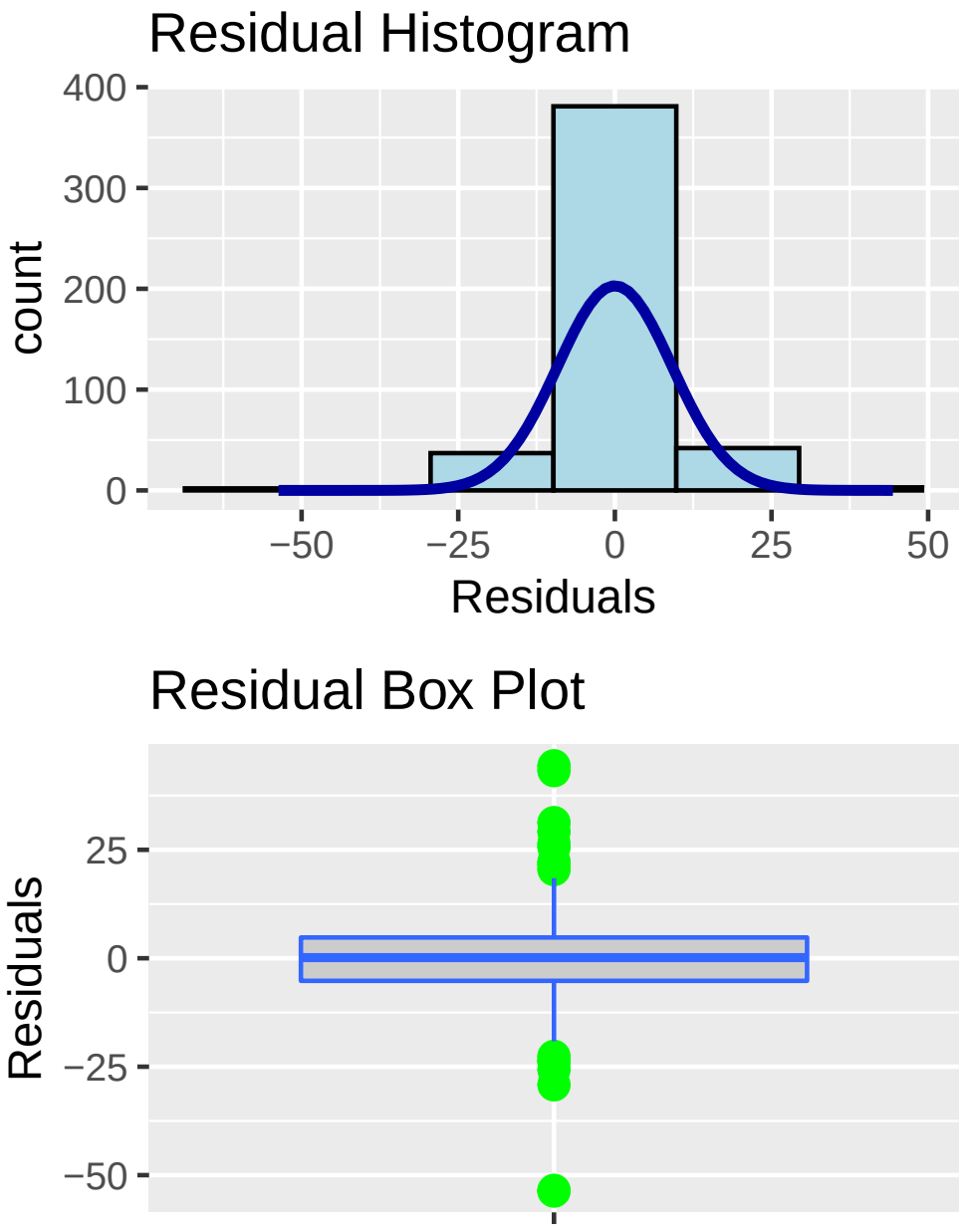


FIGURE 2.19: Shows the diagnostic of our model

An additional evaluation is provided of how the independent variables tend to adjust to the dependent variable. This assessment is illustrated visually showing the residuals versus each term in a mean function and versus fitted values. Also computes a curvature test for each of the plots by adding a quadratic term and testing the quadratic to be zero. This is Tukey's test (analysis of variance) when plotting against fitted values (Fox & Weisberg, 2011). Figures 2.20, 2.21, 2.22 show the results of these tests

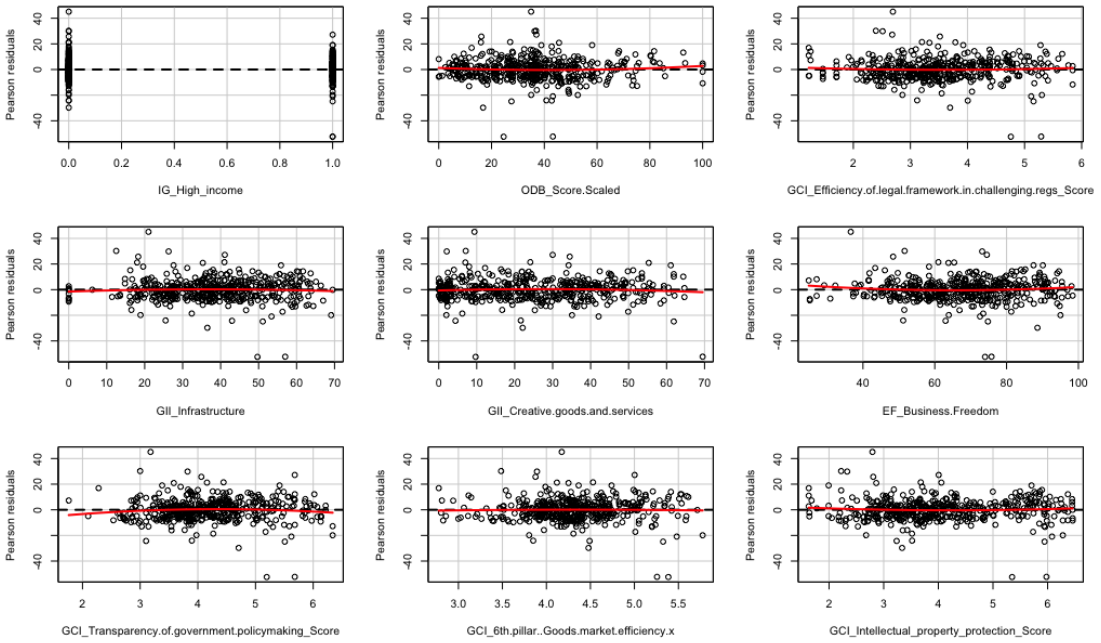


FIGURE 2.20: Shows the analysis of variance.

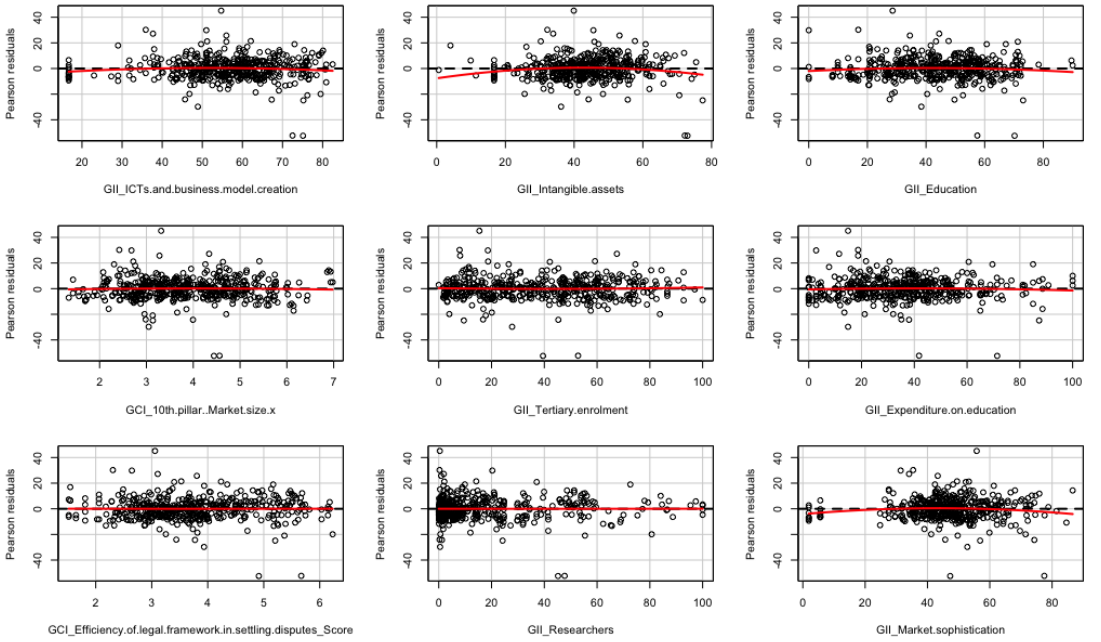


FIGURE 2.21: Shows the analysis of variance.

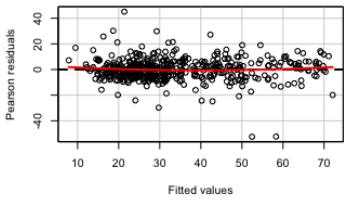


FIGURE 2.22: Shows the analysis of variance.

In the next section, we develop a battery of robustness checks test in order to inspect for issues related to regression analysis modeling.

Robustness Checks

An important part of any statistical model is robustness checking. In this context, we examine our model in order to avoid issues such as multicollinearity, heteroskedasticity, and endogeneity. Multicollinearity is found when two or more variables or features in a multiple regression model are associated (Blalock, 1963; Farrar & Glauber, 1967; Katrutsa & Strijov, 2017). Heteroskedasticity is found in a linear model when there is a variability of a feature that is not equal over the range of values (Godfrey, 1978; Hassan, Hossny, Nahavandi, & Creighton, 2012; Zarembka, 1990). Endogeneity is when there is a high association between your X variable and the error term (Bettin, Lucchetti, & Zazzaro, 2012; L'Heureux, Grolinger, Elyamany, & Capretz, 2017; Nakamura & Nakamura, 1998). In addition to these robustness tests for our main model, we also created other models desegregating the composite GEDI index.

We check our GEDI model for multicollinearity issues using the R package "performance" (Lüdecke, Makowski, & Waggoner, 2019). We analyze the variance inflation factor (VIF) which is a measure to quantify the level of multicollinearity of our independent variables. A VIF less than 5 shows a low correlation of that predictor with other predictors. A register between 5 and 10 indicates a moderate correlation, while VIF values larger than 10 are a sign for high, not tolerable correlation of our independent variables. A register between 5 and 10 indicates a moderate correlation, while VIF values larger than 10 are a sign for high, not tolerable correlation of our independent variables (James et al., 2013b). Figure 2.24 displays the degree of multicollinearity of our independent variables.

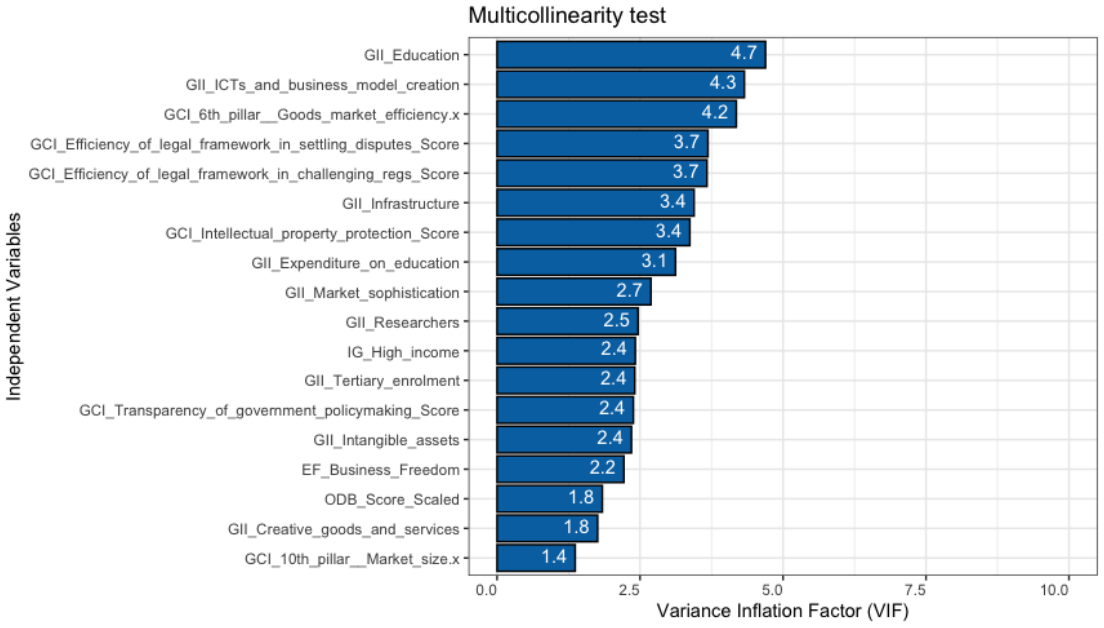


FIGURE 2.23: Shows the multicollinearity test.

In order to test our model for heteroskedasticity in our model, we use the R packages -car- (Fox & Weisberg, 2011) and -sandwich- (Zeileis, 2004). In particular, we analyze

the presence of heteroscedasticity issues through the implementation of the Performs the Breusch-Pagan test using the function bptest. Furthermore, we also implement the vcovHC function which calculates a heteroskedasticity-consistent estimation of the covariance matrix of the coefficient estimates in regression models. Our results show that Ho is not rejected because the p-value is higher than the significance level (0.2968$>$ 0.05); therefore, there is evidence that the variance of the residuals is homoscedastic since Ho is not rejected. Moreover, our results using the vcovHC function show that the robust standard errors are smaller compared to our GEDI model illustrated in figure 2.16 and, since the coefficients are the same, the t -statistics are higher and the p -values are smaller. Figure 2.25 and figure 2.26 shows the result of the heteroskedasticity test and Robust (HC1) standard errors respectively.

Breusch-Pagan heteroskedasticity test

statistic	p.value	parameter	method
20.66253	0.2967795	18	studentized Breusch-Pagan test

FIGURE 2.24: shows the result of the heteroskedasticity test.

MODEL INFO:
Observations: 548
Dependent Variable: GEDIScore
Type: OLS linear regression

MODEL FIT:
$F(18,529) = 84.1027926$, $p = 0.0000000$
$R^2 = 0.7410481$
Adj. $R^2 = 0.7322369$

Standard errors: Robust, *type* = HC1

	Est.	S.E.	t val.	p
(Intercept)	-19.6756386	4.5936392	-4.2832356	0.0000219
IG_High_income	10.5581359	1.0813754	9.7636173	0.0000000
ODB_Score_Scaled	0.1155891	0.0239951	4.8171882	0.0000019
GCI_Efficiency_of_legal_framework_in_challenging_regs_Score	1.8233027	0.6592450	2.7657439	0.0058778
GII_Infrastructure	0.2125817	0.0436809	4.8666934	0.0000015
GII_Creative_goods_and_services	-0.0985769	0.0363384	-2.7127449	0.0068902
EF_Business_Freedom	0.0914869	0.0414075	2.2094298	0.0275724
GCI_Transparency_of_government_policymaking_Score	-1.2242718	0.7161285	-1.7095699	0.0879319
GCI_6th_pillar__Goods_market_efficiency.x	5.9747221	1.5940131	3.7482265	0.0001977
GCI_Intellectual_property_protection_Score	-1.8052424	0.5891097	-3.0643571	0.0022926
GII_ICTs_and_business_model_creation	-0.2895045	0.0705042	-4.1061997	0.0000466
GII_Intangible_assets	0.1872018	0.0603947	3.0996410	0.0020407
GII_Education	-0.1773633	0.0473712	-3.7441192	0.0002009
GCI_10th_pillar__Market_size.x	1.5673217	0.3847633	4.0734695	0.0000534
GII_Tertiary_enrolment	0.1144569	0.0236009	4.8496754	0.0000016
GII_Expenditure_on_education	0.0845614	0.0293449	2.8816412	0.0041168
GCI_Efficiency_of_legal_framework_in_settling_disputes_Score	1.7797670	0.6337177	2.8084541	0.0051620
GII_Researchers	0.1276174	0.0234127	5.4507712	0.0000001
GII_Market_sophistication	0.1669158	0.0573127	2.9123685	0.0037385

FIGURE 2.25: shows the result of the Robust (HC1) standard errors test.

For the endogeneity problem, we use an instrumental variable (IV) by including an additional source of data collected by the Open Government Partnership (OGP) which is a global initiative promoting transparency, empowering citizens, and driving a knowledge-based economy through the release of data generated by public sectors. An important point of this initiative

is that countries that are part of this movement generate commitments through specific action plans that they need to deliver and these actions are monitored and auditable by independent researchers, non-governmental organizations, and civil participants. A specific action plan that several governments are adopting is developing an open data agenda in which they are getting commitments about the release of governmental datasets that are considered crucial for transparency, empower society, and economic benefits i.e. tenders, governmental spending budget, pollution, among others.

We use OGP as an IV because it helps us to explain the adoption of open government data in some countries; however, this variable is not affecting our target variable GEDI, giving a reasonable exclusion restriction for our model. We use the package -AER- (Kleiber & Zeileis, 2008) in order to test our model including the OGP variable. In particular, we implement the Instrumental-Variable Regression (ivreg) function which implements the two-stage least squares that is able to handle endogenous explanatory variables.

Our results show that the open data variable is still statistically and positively associated with our target variable. Furthermore, the diagnostic test section illustrates that the null hypothesis is rejected meaning that our instruments are suitable for our model. The Wu-Hausman test describes whether OLS estimates are significantly different from the instrumental variable (IV) estimates. This means that if the p-value is small (less than 0.05), reject the null hypothesis. Regarding the Sargan test, it identifies if you have overidentification restrictions which means that you have more than one instrument per endogenous variable. Our results show that this is not the case. Figure 2.27 illustrates the implementation of the ivreg function.

```
Call:
ivreg(formula = GEDIScore ~ ODB_Score_Scaled + IG_High_income +
    GII_Infrastructure + GII_Infrastructure + GCI_6th_pillar__Goods_market_efficiency.x +
    GII_Infrastructure + GII_Infrastructure + EF_Business_Freedom +
    GCI_6th_pillar__Goods_market_efficiency.x + GCI_Intellectual_property_protection_Score +
    GII_ICTs_and_business_model_creation + GII_Intangible_assets +
    GCI_10th_pillar__Market_size.x + GII_Researchers + GII_Market_sophistication |
    OGP, data = gedi_model_df)

Residuals:
     Min       1Q   Median       3Q      Max
-44.4615  -9.8428   0.9642   9.9439  45.2954

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.8134     3.7281   2.364   0.0184 *
ODB_Score_Scaled  0.7253     0.1027   7.059 5.11e-12 ***

Diagnostic tests:
                                                                  df1 df2 statistic  p-value
Weak instruments (ODB_Score_Scaled)                                 1 546    44.767 5.51e-11 ***
Weak instruments (IG_High_income)                                   1 546     5.576 0.018554 *
Weak instruments (GII_Infrastructure)                               1 546    51.665 2.17e-12 ***
Weak instruments (GCI_6th_pillar__Goods_market_efficiency.x)        1 546    10.108 0.001560 **
Weak instruments (EF_Business_Freedom)                              1 546    24.577 9.54e-07 ***
Weak instruments (GCI_Intellectual_property_protection_Score)       1 546    10.232 0.001460 **
Weak instruments (GII_ICTs_and_business_model_creation)             1 546    14.923 0.000125 ***
Weak instruments (GII_Intangible_assets)                            1 546    10.726 0.001124 **
Weak instruments (GCI_10th_pillar__Market_size.x)                   1 546    21.198 5.16e-06 ***
Weak instruments (GII_Researchers)                                  1 546    17.206 3.89e-05 ***
Weak instruments (GII_Market_sophistication)                        1 546     7.525 0.006286 **
Wu-Hausman                                                          1 535     4.180 0.041395 *
Sargan                                                            -10  NA        NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.64 on 546 degrees of freedom
Multiple R-Squared: 0.3048,     Adjusted R-squared: 0.3035
Wald test: 49.83 on 1 and 546 DF,  p-value: 5.111e-12
```

FIGURE 2.26: illustrates our model using IV.

In addition to these tests (multicollinearity, heteroskedasticity, and endogeneity), we developed other models in order to desegregate our dependent variable (GEDI). First, we develop a model using the attitudes sub-index (ATT) which reflects the people's stance toward entrepreneurship. This sub-index is composed of the following indicators: opportunity recognition, startup skills, risk perception, networking, and cultural support of entrepreneurs. Moreover, this indicator is collected at the country level and it is used to compare and contrast with other economies for research and public policies purposes. Figure 2.28 displays our model using the attitudes sub-index (ATT).

```
Call:
lm(formula = ATT ~ IG_High_income + ODB_Score_Scaled + GCI_Efficiency_of_legal_framework_in_challenging_regs_Score +
    GII_Infrastructure + GCI_Transparency_of_government_policymaking_Score +
    GII_Researchers, data = gedi_model_df)

Residuals:
    Min      1Q  Median      3Q     Max
-27.911  -7.739  -1.088   6.385  46.052

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                              6.67616    2.97703   2.243 0.025330 *
IG_High_income                                           7.27448    1.45917   4.985 8.34e-07 ***
ODB_Score_Scaled                                         0.17378    0.03044   5.708 1.88e-08 ***
GCI_Efficiency_of_legal_framework_in_challenging_regs_Score  2.55073    0.69020   3.696 0.000242 ***
GII_Infrastructure                                       0.11774    0.04837   2.434 0.015240 *
GCI_Transparency_of_government_policymaking_Score        1.50098    0.80335   1.868 0.062246 .
GII_Researchers                                          0.12699    0.03086   4.116 4.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.25 on 541 degrees of freedom
Multiple R-squared:  0.5143,    Adjusted R-squared:  0.5089
F-statistic: 95.47 on 6 and 541 DF,  p-value: < 2.2e-16
```

FIGURE 2.27: illustrates the ATT model.

Our results in this model show that open (government) data is still positive and statistically significant associated with entrepreneurship attitudes to a confidence level of 99%. This statistical association infers that in 99% of cases in which there is a change in the variable of open (government) data, it induces a change in the attitude variable (ATT). The model reveals that the marginal effect of the open (government) data in our target variable (ATT) is 0.17; therefore, an increase of 1% in the open (government) data rate, it induces a 17% increase of ATT.

For robustness check, we run a multicollinearity test to our model. Our results show that our independent variables have a variance inflation factor (VIF) less than 5 which means that there is a low correlation among our predicted variables. Figure 2.29 shows the level of VIF for our independent variables.

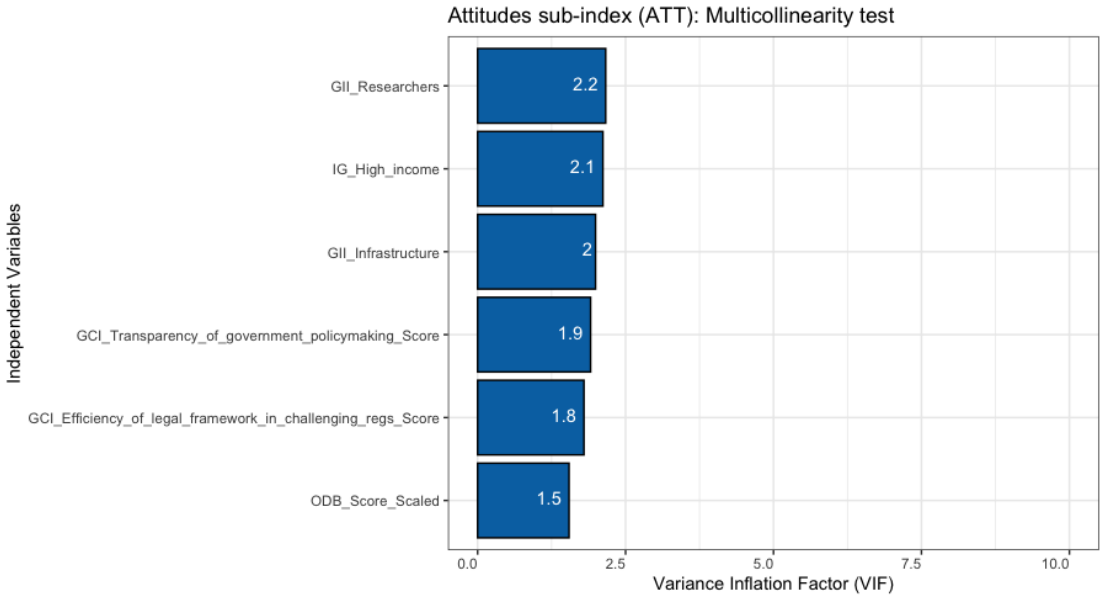


FIGURE 2.28: shows the variance inflation factor.

We also run a test to check heteroskedasticity issues for our model. Similar to the GEDI model, we perform the Breusch-Pagan test to the ATT. Our results reveal that Ho is not rejected because the p-value is higher than the significant level ($0.8063 > 0.05$). Furthermore, we implement the heteroskedasticity-consistent covariance matrix estimation (vcovHC) which is a function in R software to test heteroskedasticity issues. Our result using this function shows that the robust standard errors are smaller compared to the ATT model displayed in figure XX; therefore, the t-statistics are higher and p-values smaller. However, it is important to mention that in this ATT model, we reduce the number of predictors since some of them were not statistically significant. Figure 2.30 and figure 2.31 display the implementation and results of the Breusch-Pagan (bptest) and heteroskedasticity-consistent covariance matrix estimation (vcovHC) tests respectively.

Breusch-Pagan heteroskedasticity test

statistic	p.value	parameter	method
3.019855	0.8063507	6	studentized Breusch-Pagan test

FIGURE 2.29: shows the result of the heteroskedasticity test using the bptest function.

MODEL INFO:
Observations: 548
Dependent Variable: ATT
Type: OLS linear regression

MODEL FIT:
$F(6,541) = 95.4733974$, $p = 0.0000000$
$R^2 = 0.5142931$
Adj. $R^2 = 0.5089063$

Standard errors: Robust, *type* = HC1

	Est.	S.E.	t val.	p
(Intercept)	6.6761585	3.2282991	2.0680112	0.0391137
IG_High_income	7.2744755	1.3473774	5.3989887	0.0000001
ODB_Score_Scaled	0.1737798	0.0308454	5.6339026	0.0000000
GCI_Efficiency_of_legal_framework_in_challenging_regs_Score	2.5507348	0.6117157	4.1698046	0.0000355
GII_Infrastructure	0.1177424	0.0524194	2.2461614	0.0250964
GCI_Transparency_of_government_policymaking_Score	1.5009780	0.7995686	1.8772349	0.0610238
GII_Researchers	0.1269932	0.0341944	3.7138602	0.0002253

FIGURE 2.30: shows the result of the heteroskedasticity test using the vcovHC function.

In order to continue desegregating the GEDI index, we use one of the five pillars of the ATT sub-index which is the opportunity perception indicator. This variable measures how people recognize and explore novel business opportunities at the country level. we consider analyzing this indicator since we argue that open (government) data is used for the identification of new business opportunities, strategic planning, and the evaluation of investment projects by entrepreneurs. Figure 2.32 illustrates the result of our model using the opportunity perception indicator.

```
Call:
lm(formula = Opportunity_Perception ~ IG_High_income + ODB_Score_Scaled +
    EF_Business_Freedom + GCI_Transparency_of_government_policymaking_Score +
    GII_Intangible_assets + GII_Education + GII_Tertiary_enrolment +
    GII_Expenditure_on_education + GCI_Efficiency_of_legal_framework_in_settling_disputes_Score,
    data = gedi_model_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44629 -0.11147 -0.01644  0.10743  0.67106

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     -0.2103099  0.0506038  -4.156 3.77e-05 ***
IG_High_income                                   0.0725511  0.0218861   3.315 0.000978 ***
ODB_Score_Scaled                                 0.0011381  0.0004565   2.493 0.012957 *
EF_Business_Freedom                              0.0026779  0.0006812   3.931 9.57e-05 ***
GCI_Transparency_of_government_policymaking_Score 0.0362783  0.0120740   3.005 0.002783 **
GII_Intangible_assets                            0.0019283  0.0007006   2.752 0.006121 **
GII_Education                                   -0.0029661  0.0008541  -3.473 0.000556 ***
GII_Tertiary_enrolment                           0.0008605  0.0004223   2.038 0.042084 *
GII_Expenditure_on_education                     0.0021372  0.0005894   3.626 0.000315 ***
GCI_Efficiency_of_legal_framework_in_settling_disputes_Score 0.0450770  0.0093937   4.799 2.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.17 on 538 degrees of freedom
Multiple R-squared:  0.4234,    Adjusted R-squared:  0.4138
F-statistic:  43.9 on 9 and 538 DF,  p-value: < 2.2e-16
```

FIGURE 2.31: shows the effect of the opportunity perception indicator.

Our results show that open (government) data has also a positive and statistically significant association with the perception of entrepreneurial opportunities to a confidence level of 99%. This result could be interpreted as 99% of the cases in which there is a variation of open (government) data, it induces a change in the entrepreneurial opportunity perception. The marginal effect of open (government) data in our dependent variable (opportunity perception) is 0.012; therefore, an increase of 1% in the open (government) data rate, it induces a 1.2% in entrepreneurial opportunity perception. In this model is also important to mention that there are fewer predictors variables than in the GEDI model because some of them were not statistically significant

We also run a battery of robustness checks for this model. First, we run a multicollinearity test. Our results are consistent with our previous multicollinearity tests showing a variance inflator factor (VIF) less than 5 among our predictors. Figure 2.33 illustrates the multicollinearity test applied to the opportunity perception model.

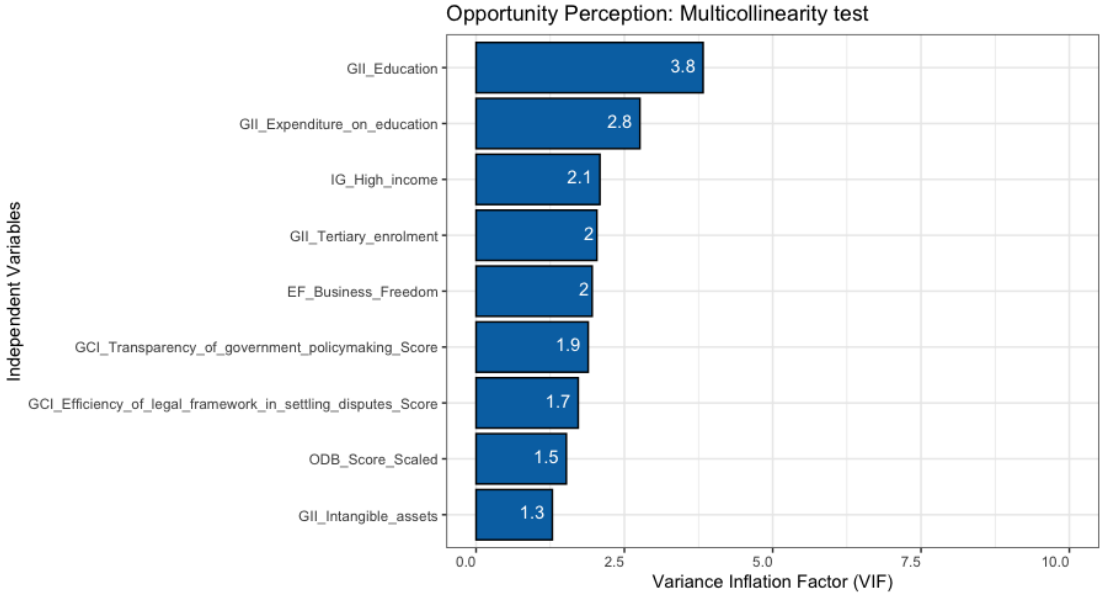


FIGURE 2.32: illustrates the VIF per predictor.

As we performed in our previous models, we run an additional robustness check testing heteroskedasticity issues. First, we runt the Breusch-Pagan test to the opportunity perception model. Consistent with our results in our previous models (GEDI and ATT), this result reports that Ho is not rejected because the p-value is higher than the significant level ($0.03131 > 0.05$). Moreover, the application of the vcovHC functions revels is similar to our previous models in terms that the t-statistics are higher and p-values smaller. Figure 2.34 and figure 2.35 present the outcomes of the Breusch-Pagan and heteroskedasticity-consistent covariance matrix estimation tests respectively.

Breusch-Pagan heteroskedasticity test

statistic	p.value	parameter	method
18.35116	0.0313113	9	studentized Breusch-Pagan test

FIGURE 2.33: shows the result of the bptest function of our opportunity perception model.

MODEL INFO:
Observations: 548
Dependent Variable: Opportunity_Perception
Type: OLS linear regression

MODEL FIT:
$F(9,538) = 43.9005692$, $p = 0.0000000$
$R^2 = 0.4234305$
Adj. $R^2 = 0.4137852$

Standard errors: Robust, *type* = HC1

	Est.	S.E.	t val.	p
(Intercept)	-0.2103099	0.0496565	-4.2352912	0.0000268
IG_High_income	0.0725511	0.0218570	3.3193475	0.0009634
ODB_Score_Scaled	0.0011381	0.0004359	2.6107312	0.0092868
EF_Business_Freedom	0.0026779	0.0006784	3.9470334	0.0000896
GCI_Transparency_of_government_policymaking_Score	0.0362783	0.0131456	2.7597433	0.0059819
GII_Intangible_assets	0.0019283	0.0006693	2.8808520	0.0041242
GII_Education	-0.0029661	0.0009599	-3.0899931	0.0021051
GII_Tertiary_enrolment	0.0008605	0.0004666	1.8443718	0.0656785
GII_Expenditure_on_education	0.0021372	0.0006486	3.2953052	0.0010479
GCI_Efficiency_of_legal_framework_in_settling_disputes_Score	0.0450770	0.0107522	4.1923511	0.0000323

FIGURE 2.34: shows the result of the heteroskedasticity test for this model.

In the next section, we will provide an explanation of our GEDI model and the implication of the relationship between the dependent/target and independent/predicted variables.

2.6 Results and Discussion

This section describes the results of our linear model developed in which the dependent variable is the score of The Global Entrepreneurship and Development Index (GEDI) and the variables of control are composed of the Open Data (OD), Innovation (GII), Competitiveness (GCI) and Economic Freedom Index (EF).

The main goal of this model is to test our hypothesis: is there an effect in the publication of open government data on entrepreneurship at the country level?. Figure 2.23 shows the results of our model

```
Call:
lm(formula = GEDIScore ~ IG_High_income + ODB_Score.Scaled +
    GCI_Efficiency.of.legal.framework.in.challenging.regs_Score +
    GII_Infrastructure + GII_Creative.goods.and.services + EF_Business.Freedom +
    GCI_Transparency.of.government.policymaking_Score + GCI_6th.pillar..Goods.market.efficiency.x +
    GCI_Intellectual_property_protection_Score + GII_ICTs.and.business.model.creation +
    GII_Intangible.assets + GII_Education + GCI_10th.pillar..Market.size.x +
    GII_Tertiary.enrolment + GII_Expenditure.on.education + GCI_Efficiency.of.legal.framework.in.settling.disputes_Score +
    GII_Researchers + GII_Market.sophistication, data = gedi_model_df)

Residuals:
    Min      1Q  Median      3Q     Max
-54.344  -5.190   0.486   4.653  44.745

Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                             -19.67564    4.38501  -4.487 8.87e-06 ***
IG_High_income                                           10.55814    1.25779   8.394 4.31e-16 ***
ODB_Score.Scaled                                          0.11559    0.02681   4.311 1.94e-05 ***
GCI_Efficiency.of.legal.framework.in.challenging.regs_Score  1.82330    0.79585   2.291 0.022355 *
GII_Infrastructure                                        0.21258    0.05130   4.144 3.98e-05 ***
GII_Creative.goods.and.services                          -0.09858    0.02992  -3.295 0.001052 **
EF_Business.Freedom                                       0.09149    0.03874   2.362 0.018550 *
GCI_Transparency.of.government.policymaking_Score        -1.22427    0.72416  -1.691 0.091501 .
GCI_6th.pillar..Goods.market.efficiency.x                 5.97472    1.49954   3.984 7.72e-05 ***
GCI_Intellectual_property_protection_Score               -1.80524    0.62191  -2.903 0.003853 **
GII_ICTs.and.business.model.creation                     -0.28950    0.06278  -4.611 5.02e-06 ***
GII_Intangible.assets                                     0.18720    0.05059   3.700 0.000238 ***
GII_Education                                            -0.17736    0.05052  -3.511 0.000484 ***
GCI_10th.pillar..Market.size.x                            1.56732    0.40830   3.839 0.000139 ***
GII_Tertiary.enrolment                                    0.11446    0.02453   4.666 3.89e-06 ***
GII_Expenditure.on.education                              0.08456    0.03346   2.527 0.011791 *
GCI_Efficiency.of.legal.framework.in.settling.disputes_Score  1.77977    0.73424   2.424 0.015686 *
GII_Researchers                                           0.12762    0.02657   4.803 2.04e-06 ***
GII_Market.sophistication                                 0.16692    0.04773   3.497 0.000510 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.077 on 529 degrees of freedom
Multiple R-squared:  0.741,     Adjusted R-squared:  0.7322
F-statistic:  84.1 on 18 and 529 DF,  p-value: < 2.2e-16
```

FIGURE 2.35: Model results.

Our result shows that open (government) data has a positive and statistically significant relationship with the GEDI index to a confidence level of 99%. This result infers that in 99% of cases in which there is a change in the variable open data, it induces a change the GEDI variable. The model shows that the marginal effect of the variable open government data in GEDI is 0.11. Therefore, an increase of 1% in the open (government) data rate induces an 11% increase in the GEDI index. The significance of this relationship could be represented in that the impact of increasing the GEDI index in a country could be captured by individuals and firms in the consolidation of business ideas, due to this indicator is composed

of several elements that encourage entrepreneurship such as opportunity perception, process innovation, risk capital, technology innovation, internationalization, product innovation, and opportunity startup.

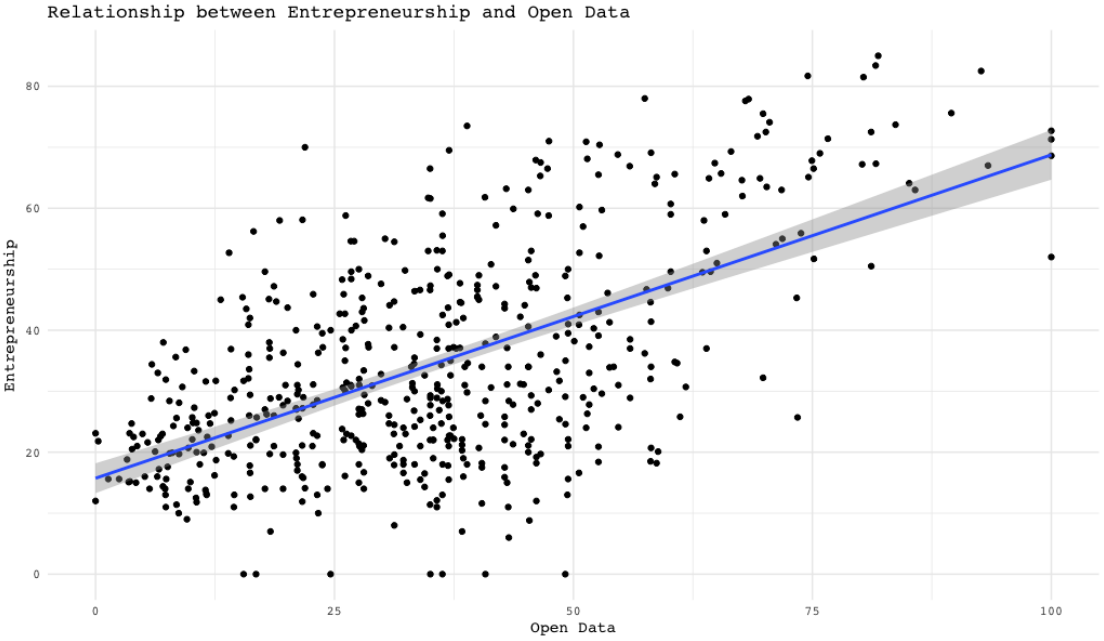


FIGURE 2.36: Shows the relationship between GEDI and OGD.

Our results also show that the relationship between entrepreneurship and open (government) data tends to have better effects in high-income countries. For instance, the indicator High-Income also have a positive and statistically significant effect on entrepreneurship. The marginal effect of the variable High-Income in GEDI is 10.42. This result could be explained since these developed countries tend to have a better regulatory framework, better infrastructure, a well functioning legal framework and policies in order to support entrepreneurship as economic engine. Figure 2.25 shows the relationship between entrepreneurship and open (government) data in filtered by income group

FIGURE 2.37: Shows the relationship between GEDI and OGD per Income group.

Our model also shows a positive and statistically significant relationship among some of the components of the Competitiveness Index (GCI) and entrepreneurship (GEDI). The components that show an association to the GEDI indicator are related to the Institutions pillar of the GCI. These indicators are labelled as the efficiency of a legal framework, intellectual property protection and transparency of government and policy-making. The relevance of these indicators to Entrepreneurship is that the institutional performance of a country depends on the efficiency and the execution of both public and private stakeholders. These indicators also constituted a legal and administrative framework in which economic agents (individuals, firms, and governments) have influence in the quality of institutions. This ecosystem composed of institutional indicators is one of the drivers that could promote investment decisions, job creation, and economic growth.

Additional components of the Competitiveness Index (GCI) such as goods market efficiency and market size pillars also shows a positive and statistically significant relationship with Entrepreneurship. The relevance of this relationship could be explained given their particular supply-and-demand conditions across countries. In a free market, the national and foreign competition plays an important role in promoting market efficiency, and thus business productivity. This market condition triggers to firms to be more innovation and to implement new technologies in order to be more competitive. An external variable from the Economic Freedom index (EF) labelled as Business freedom also shows a positive and statistically significant relationship with Entrepreneurship. The relevance of this variable is that it measures the regulatory and infrastructure environments constrain the efficient operation of businesses. Figure 2.26 illustrates the relationship between the GEDI and GCI and EF indexes.

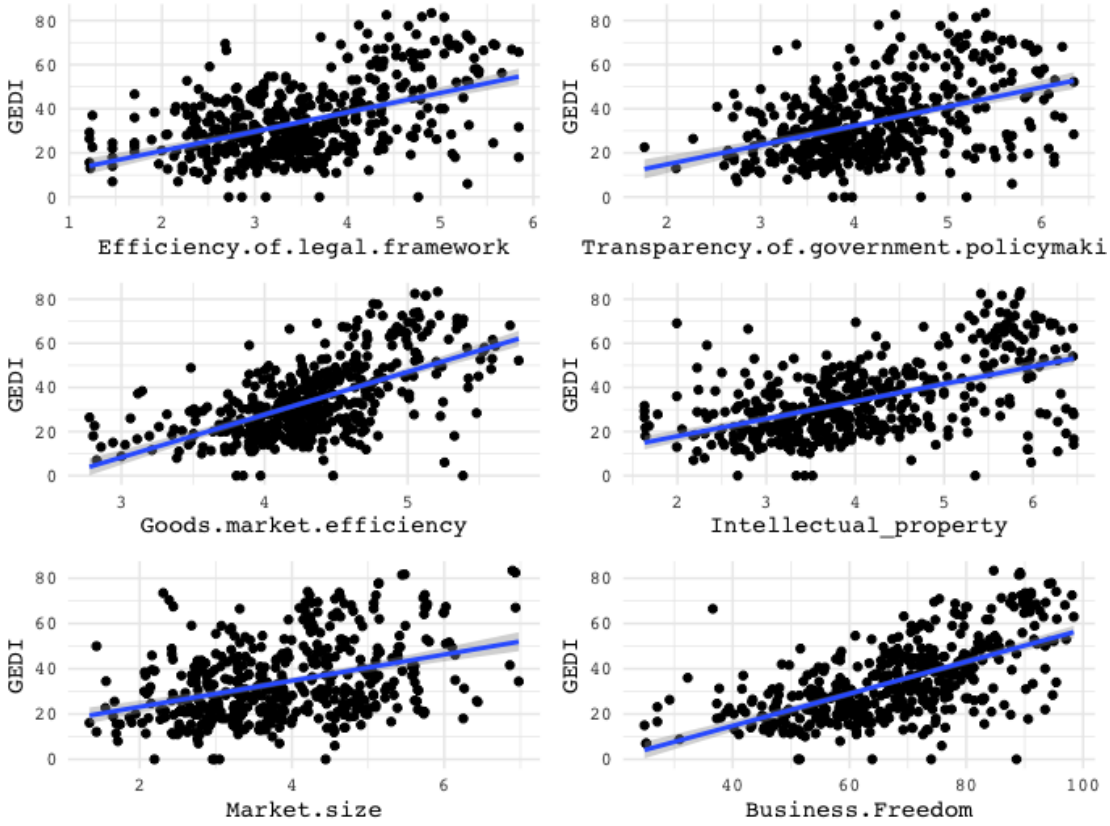


FIGURE 2.38: Illustrates the relationship between the GEDI, GCI and EF indexes.

The indicators composed of the Innovation index also show a positive and statistically significant relationship to entrepreneurship (GEDI). Expenditure on education, researchers, and tertiary enrollment are some of the indicators grouped in the education pillar. The association between these indicators and entrepreneurship is dictated by the rate of education at the country level. The education pillar and its components are considered as one of the main factors of the capacity to entrepreneur and innovate in a country. Infrastructure is another important indicator that shows a positive and statistically significant relationship to entrepreneurship. This association is crucial in order to create the connectivity and linkages that could promote entrepreneurial opportunities. For example, in the last decade, the impact of broadband as an infrastructure has been considered as one of the main topics for policy-making. Another important factor linked to the demand side of entrepreneurship is technological development. Furthermore, the investment and implementation of information and communication technologies (ICT) in the business model and organizational process are highly associated with innovation. In our model, the variables related to creativity and technology (Intangible assets, creative goods and services, and ICT and business models creation) show a positive and statistically significant relationship to Entrepreneurship (GEDI). Figure 2.27 illustrate the relationship between the GEDI, GII sub-pillars.

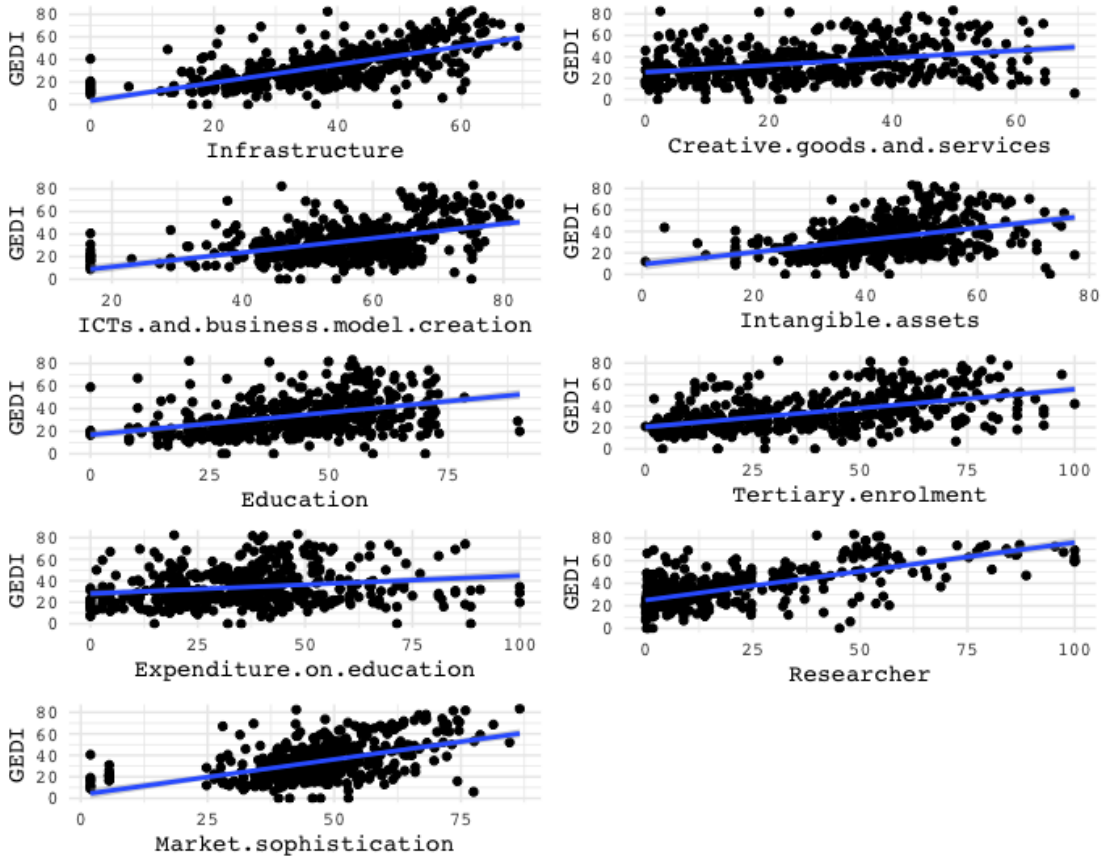


FIGURE 2.39: Shows the relationship between GEDI and GII.

In the Appendix section 1.6, it is present a full description of our model and its respective visualizations.

Although we use in our model a wide range of indexes and indicators, that confirm the relationship between entrepreneurship and open (government) data, there are still some limitations in terms of longitudinal data that we will describe in the next section.

2.6.1 Limitations

Despite the conception and philosophical idea of open (government) data has been present in the last decades (previously referred to as Public Sector Information -PSI-) there were no consistent initiatives or policies in order to collect and publish this information from governments.

A more formal efforts collecting and measuring the use of open data around the world were found in 2013 by the Open Government Data index (GODI) a recompilation of information supported by the Open Knowledge Foundation covering 60 economies and measuring the specific datasets released as open data by governments. During the same year, the Open

Data Barometer (ODB) founded by the Word Wide Web Foundation released their first study measuring the readiness, implementation, and impact of open data covering 77 economies.

This study is composed of the Open Data Barometer (ODB) dataset which monitors the evolution and performance of different economies in the adoption, use, and impact of Open Government Data. We created a panel data and developed a descriptive regression analysis revealing a positive and statistically significant relationship between the effect of open (government) data and entrepreneurship ant the country level. In order to create a more deep analysis trying to explain causalities between these variables, we must have longitudinal data. However, at the moment of this research, this data is not available yet.

In the next section, we will summarize the main points of this chapter and we will propose our next directions in terms of research.

2.7 Conclusions

The contribution of this research is to systematically analyze and test the relationship and effect of open (government) data on entrepreneurship at the country level. In this research we argue that open (government) data allows the access of crucial information that households, firms, and the government uses to allocate resources, information is key to make more rational and efficient decisions which makes entrepreneurship more attractive (by having an impact on the rate of return of entrepreneurship) and affects the economic development of a country.

In order to test our hypothesis, we conduct a literature review about the relationship of open data, competitiveness, economic freedom, and innovation as determinants of entrepreneurship. We also develop a formal theoretical model that studies the decision of becoming an entrepreneur or an employee. These alternatives have benefits and costs and the alternative of becoming an employee obtaining a salary is used as the opportunity cost to become an entrepreneur. Our model shows that open (government) data might increase the rate of return of becoming an entrepreneur and therefore increases in open (government) data can be associated with more individuals having access to critical information that might make the decision of becoming an entrepreneur more profitable. For this reason, more open (government) data might induce more individuals to choose to be an entrepreneur. Next, we develop an empirical model using econometric and data science techniques to examine the empirical connection between entrepreneurship and open (government) data. We also test, as control variables, other determinants detailed in the literature review.

Concerning our empirical model we develop a linear regression analysis of our dependent variable (entrepreneurship) and independent variables (open data, competitiveness, economic freedom, and innovation). As we mentioned before, we adopt an interdisciplinary approach combining econometrics and data science techniques, creating a research workflow to collect data and evaluate our econometric models. The first step in this pipeline was to ask the following question: *is there an effect of the publication of open (government) data on entrepreneurship at the country level?*

The next stage was data acquisition from a set of different indicators such as The Global Entrepreneurship and Development Institute (GEDI) created by The GEDI Institute, we define this indicator as our endogenous variable. As independent variables, our dataset is composed of the Open Data Barometer (ODB) generated by The World Wide Web Foundation. The Economic Freedom Index (EF) produced by The Heritage Foundation. The Global Competitiveness Report (GCR) published by the World Economic Forum and the Global Innovation Index (GII) co-published by Cornell University.

Our sample is composed by 137 economies generating 474 variables and 528 observations from 2013 to 2016. Because of this wide number of variables, we developed a series of pre-processing tasks such as data cleaning (e.g. changing the comma to a period in some

variables) and dealing with missing values (performing imputations). Furthermore, we extracted variables implementing machine learning algorithms such as principal component analysis (PCA) and selecting relevant features using algorithms such as forward, backward and stepwise regression. Once we reduce redundant data and analyze multicollinearity effects in our dataset, we proceed to generate and analyze our model.

The main findings of this research are the following. First, the model shows that changes in open (government) data are correlated with positive changes in the global entrepreneurship and development institute (GEDI) index. Our estimates suggest that a 1% increase in the index of open (government) data, increases 0.11% the global entrepreneurship and development index. This result means that a rise in the index of entrepreneurship is associated with more opportunity perception, product, process and technological innovation; therefore, open (government) data could provide the information needed for the identification of new business opportunities, strategic planning and the evaluation of investment projects. Moreover, open (government) data could give more access to information and this helps that entrepreneurs can reach rational decisions when they have access to information about the needs of supply and demand in markets. Without the access to this information, entrepreneurs might end up choosing dominated alternatives which, in turn, affects the efficiency in the allocation of resources.

Second, we find a positive and higher relationship between open (government) data and entrepreneurship in high Income countries. That is to say, the positive marginal effect of open (government) data on entrepreneurship is higher in high income countries compared with the marginal effect of open (government) data on entrepreneurship in low income countries. This result is consistent with the information found in the literature in which entrepreneurship tends to have a high level of entrepreneur policies, culture and adoption on developed economies. This variation is also connected to additional factors such as the level of better quality of education, better infrastructure, better institutions, and more innovation in developed economies relative to underdeveloped countries. In the case of open (government) data, these factors also tends to affect its adoption and implementation. This means, the potential benefits of open (government) data are related to a shared value (social-economic). On one hand, the advantages are associated to government transparency, accountability, fight against corruption and empower citizens. On the other hand, the economic benefits are associated to innovation developing new products or services, and business model, job creation and technological development and economic growth.

Third, there is a positive and statistically significant relationship between the pillar of institutions and regulatory framework from the Competitiveness index that contains elements such as the efficiency of a legal framework, transparency of government policymaking, intellectual property, goods market size, and efficiency and the GEDI index. These correlations could explain that the level of efficiency in which these institutions across countries perform, depends on the execution of public and private stakeholders. For instance, the indicator about efficiency of legal framework in settling disputes refers to the certainty and confidence

of entrepreneurs that governments will support through a legal process any issue that entrepreneurs will face in order to continue work. The transparency policy-making process is another important point that could promote or inhibit the entrepreneurial ecosystem. An example of promotion is when governments funds programs to stimulate entrepreneurial activities in specific sectors. The access to governmental data is crucial in terms of business opportunities, benefits, financial aid, and obligations clearly stated to entrepreneurs to access these programs. In addition, the lack of transparency could inhibit entrepreneurial activities due to the uncertainty of the benefits and penalties.The relationship between market efficiency and entrepreneurship could be related to the balance of production of goods and services in a supply and demand market. For instance, in a market with demanding customers, entrepreneurs need to provide innovative products and services in order to be more competitive and survive. Another component of these indicators is intellectual property and it is a crucial element in an entrepreneurial ecosystem because it allows the defense of the rights of economic agents under the legal framework. This could promote the creation of innovation by entrepreneurs. An additional component is business freedom, this variable is from the Economic Freedom Index. This indicator shows a positive and statistically significant on entrepreneurship and measures the regulatory and infrastructure environments constrain the efficient operation of businesses.

Four, our results also show a positive and statistically significant relationship among our dependent variable entrepreneurship and different indicators such as education, expenditure on education, tertiary enrolment and researchers which are part of the pillar of human capital and research, from the Innovation index. The level of education is an important factor for entrepreneurship and Innovation. For instance, entrepreneur's performance can be enhanced through education implementing their knowledge in order to achieve some level of business survival, investments, and firm growth. Other important components of education are the expenditure on education and tertiary enrolment because they are good proxies for education coverage at the country level. Researchers are also a crucial component in an entrepreneurial ecosystem because they are an important part of the innovative process, performing investigation improving designs and discovering new solutions that in turn can be new products or services. An additional indicator is the infrastructure pillar that shows a positive and statistically significant relationship with entrepreneurship. The relevance of this indicator is that is a proxy indicator for the use of ICT access and use which is considered as an enabler of entrepreneurship creating the connectivity and connections that could encourage business opportunities. Examples of these nascent opportunities are the businesses created around the world using the broadband as infrastructure in different sectors i.e. financial services, telecommunications, education. The Innovation index also offers a pillar called creative outputs which refers to the measure of innovative activities at the country level. In our case, the indicators that belong to this pillar and show a positive and statistically significant relationship with entrepreneurship are ICT and business model creation, Intangible assets and creative goods and services. The first refers to the use of ICT as a baseline to set up business model creation. These are the kind of companies that are offering products,

services and customer values through the implementation of a technical solution i.e. software development. In our context, these are known as open data companies which are using data freely available in order to implement innovative business models. Examples of these revenues strategies which are part of their business models based on ICT are Freemium, Premium, or Open Source, Software as a Service (SaaS). Creative goods and services refer to goods that have intrinsic value. Some examples of intangible assets from different industries can be software applications, databases, patents, franchise rights, goodwill, and non-compete agreements among others.

To sum up, there is a positive and statistically significant relationship between entrepreneurship and open (government) data, this correlation stands out in developed countries. Moreover, entrepreneurship also shows a positive and statistically significant relationship with other factors related to education, infrastructure, institutions, etc.

In the next chapters, we will illustrate the relationship between open (government) data and entrepreneurship presenting evidence of open data business models adopted by entrepreneurs when they are using open (government) data as an asset. We will also present the challenges and limitations that entrepreneurs are facing when they are implementing open (government) data and their part of their business process.

Chapter 3

Open Data Companies and Business Models

"Data by itself has no value.".

.

— U Fayyad

3.1 Abstract

The economic benefits of releasing open (government) data are related to the development of new products and services, job creation and innovation. However, the identification of business roles for companies that are using open (government) data as a digital asset is a topic under exploration. The purpose of this research is to analyze how entrepreneurs are transforming open (government) data into business ideas in Europe. The analysis of this research is using information collected by the Open Data Incubator for Europe (ODINE) which contains more than 1173 business proposals of entrepreneurs using open (government) data as part of their production process. The methodology implemented in this research is based on a text mining approach because these business proposals are composed of descriptions about the company profile, structure of the business model adopted, the core idea about how they are using open data as part of their business proposals, description of the economic, social or environmental impact implementing Open Data and team compositions. The contribution of this work is extracting and measuring these valuable textual descriptions (unstructured data) using Machine Learning algorithms through the implementation of Natural Language Processing (NLP) techniques such as word count, biagrams, and word associations. Results show that open (government) data is considered as a digital asset by entrepreneurs and perceived as a business opportunity to enhance or produce new products

or services and the chance to implement innovative business models. This implementation of open (government) data by entrepreneurs in their business process is more perceptible in southern and northern regions than western and eastern of Europe. Furthermore, the use of open (government) data by entrepreneurs in Europe is mainly in high-income countries rather than upper middle and lower income. Another important finding is that most of the companies that submitted business proposals to the ODINE program are related to the technology sector. This means that most of the entrepreneurs that applied to ODINE program stated that they companies belong to the "Information and Communication" area. These companies are considered data-driven companies because they are implementing approaches such as data science, software development, cloud computing, linked data, semantic web, knowledge management, and information systems in order to extract and transform open data into products and services. Regarding companies profile, the data shows that companies are relatively young because they were created between 2011 and 2017.

Keywords: Open Government Data, Entrepreneurship, Business Models.

.

3.2 Introduction

A firm may be able to exploit an innovative opportunity with the data or organise to capture its value, if it executes an appropriate business model (Chesbrough & Rosenbloom, 2002). The identification of business models for open data can act as a stimulus not only for entrepreneurship, but also for the opening of data by private companies, (Immonen, Palviainen, & Ovaska, 2014).

The term 'business model' has no single agreed definition but can be described as the set of (operationalised) assumptions an organisation holds about its key activities, revenue streams, customers, costs and value proposition (Ovans, 2015). The topic of business models is relatively new since it appears in the academic field for the first time in 1957 (Bellman, Clark, Malcolm, Craft, & Ricciardi, 1957) and its expansion might be attributed to the Internet/Web revolution (Osterwalder, Pigneur, & Tucci, 2005). Furthermore, the term is associated with different domains such as E-Business, Information Systems, Strategy and Management (Pateli, 2003). Nevertheless, the term of business models is usually used in ambiguous ways, confusing its definition and its elements. Some authors use the concept of business models interchangeably when they are explaining the business process in a company. Others are using the business strategy definition in order to explain and illustrate the entire business model (key activities, revenue streams, customers, costs and value proposition). Another situation is confusing enterprise models (mission, vision) to business models.

In the open data movement, a similar situation happens to the identification of business models adopted by companies that are using data as an asset. The recent increase of releasing open data by governments around the world is perceived by entrepreneurs as business opportunities and the chance to transform it into business models. Theoretical work has been performed with the aim of identifying business models appropriate for use with Open Data, where the raw resource is shared and has neither rarity nor inimitability (Ferro & Osella, 2013; Zeleti, Ojo, & Curry, 2016). Some case studies have also been studied (Zimmermann & Pucihar, 2015). However, the adoption and implementation of Open Data Business Model (ODBM) is still incipient, blurred and under research development.

The aim of this research is to analyse the business opportunities perceptions collected by the Open Data Incubator for Europe (ODINE)[1] which is a project funded by the European Commision through the Horizon 2020 innovation program in order to foster digital business cases based open (government) data. Advocates of the open data movement claim that it potentially has the power to transform and scale business structures due to the global release of open data from different sectors such as politics, finance, information and communication, agriculture, health, consultancy, education, or real estate (just to mention some of them) engaging individuals to innovate and create economic value. Nevertheless, these benefits are still incipient or not clear due to the recent growth of companies that are using open

[1] `https://opendataincubator.eu/`

(government) data as an asset in their business processes and the scarcity of evidence about the composition of these business model based on it; therefore, this leads our research questions

What are the components of the business models adopted for companies that are using open (government) data ?

We are also interested in researching:

What are the company profiles, business ideas, impacts (in terms of economic, social, or environmental effects), and the team compositions (human capital) that entrepreneurs are describing in their business proposals?

The contribution of this research lies in exploring the business model composition adopted by entrepreneurs when they are using open (government) data as a digital asset. The relevance of this analysis is because there is a gap in the literature, explaining how business ideas using open (government) data are transformed into a product or service and what is the business model adopted to monetize this idea. This gap is explained due to the complexity of data collection from the private sector. This means entrepreneurs and companies are not always open or allowed to explain their opportunity recognition using and transforming open (government) data into a product or service.

The methodology is based on a descriptive content analysis and text mining approach extracting the Business Model Canvas information stated by entrepreneurs that applied to the ODINE program. Our goal is to extract and illustrate the key partners and activities, value proposition, customer relationships, channels, segments, cost structure, and the revenue stream of some companies (not all entrepreneurs opened their business models due to legal or confidential restrictions). Furthermore, this research explores the company's profiles, business idea, impacts (in terms of economic, social, or environmental effects) and the team composition that these entrepreneurs described in their business proposals. This information is mined from the business proposals submitted by entrepreneurs to the ODINE consortium (See appendix 1.8.- ODINE application template) from the following questions:

What is the core idea that entrepreneurs are describing when they are using open data as part of their production process?

What is the economic, social or environmental impact that entrepreneurs are proposing with open data?

What is the team composition of these companies?

This work is composed of 6 sections. Section 1 explains the aim, purpose, content, justification, model and methodology used in this investigation. Section 2 presents the theoretical background of the business models concept, definition, history, and implementation. Furthermore, this section also covers the business models adopted by companies that are using open

(government) data as part of their production business. Section 3 describes the methodology implemented in order to examine unstructured data (textual descriptions on pdf format) collected by ODINE, it is proposed a text mining approach to extract and analyze the information related to company profile, business idea, what is the problem solving proposed by companies in their application, how they are describing the monetization strategy and classifying the economic, social or environmental impact of their business proposal. Section 4 proposes a descriptive analysis reporting the result of our research question. The results and discussion are presented in section 5. Finally, section 6 introduces the conclusions and key findings of our hypotheses and future work.

3.3 Background and Literature Review

Analyzing the structure and composition of digital business developed by entrepreneurs, as well as the business models implemented by them, have increasingly attracted the attention of researchers in different domains during the last decades. An important factor of this tendency is the expansion and evolution of the Internet/Web technologies and the subsequent concepts such as E-Commerce and E-Business that were created based on these technologies. However, the concept of a business model is sometimes used ambiguously or misunderstood. The aim of this section is to provide a context of the concept of digital business transformation. Furthermore, clarifying the definition and scope of the business models. Finally, we research the state of the art of the relationship among digital businesses, business models, and the open data movement.

3.3.1 Digital Business

Economic theory has studied how the development and implementation of technology have become an engine of economic growth (Solow, 1956). In particular, companies have adopted information technology for digital business transformation (DBT). For instance, improving the decision-making process (i.e. collecting and analyzing enterprise information) developing digital strategies (i.e. increasing productivity and reducing costs), and enhancing business infrastructure (i.e. intranets, websites, social media). According to (LeHong, 2019) the digitalization of business offers *"the creation of new value chains and business opportunities that traditional businesses cannot offer"*., (Vial, 2019) argue that the digital transformation is referred to as *"a process that aims to improve an entity by triggering significant changes to its properties through combinations of information, computing, communication, and connectivity technologies"*. Furthermore, the author claims that business digital transformation promotes the development of innovative strategies and enhances operational performance. Besides, (Carlo, Lyytinen, & Boland, 2012; Karimi & Walter, 2015; Selander & Jarvenpaa, 2016) argue that business digital transformation is not only the use and implementation of technology but also the definition of a business strategy, process reengineering, and cultural inclusion.

According to (Briel, Davidsson, & Recker, 2018; Huang, Henfridsson, Liu, & Newell, 2017; Lyytinen, Sørensen, & Tilson, 2017; Srinivasan & Venkatraman, 2018), digital transformation is considered as a trigger for entrepreneurship and innovation, involving cross-sectoral industries (i.e. healthcare, manufacturing, education, defense) and generating economic ecosystems. Furthermore, (Brynjolfsson, 2011; Katz, Koutroumpis, & Fernando, 2014; Kenney & Zysman, 2016) claim that digital transformation has implications promoting entrepreneurship activities at the country and regional level. Akter 2020 argues that business digital transformation spurs the development of new business models and strategies. For instance, reaching and connecting with existing or new customers through the use of online platforms

or mobile services. Another example is improving and customizing user experiences and needs through the use of disruptive technology such as artificial intelligence which in order to work it requires vast amounts of data. In this context, (Smith, Ofe, & Sandberg, 2016) suggests the importance of open (government) data as an enabler acting as a digital service innovation not only for social but also for economic growth

(Laudien, Bouncken, & Pesch, 2018) claims that digital transformation is considered as one of the main topics and challenges not only for the private sector but also for the public sector. (Mergel, Kattel, Lember, & McBride, 2018) states that public servants have been adopting the use of technology in their public policies in order to improve attention, services, and government efficiency. The author suggests that these policies have focused on user demands being one of these the creation and release of open data for social and economic gains. (Goldfarb, Greenstein, & Tucker, 2015; Martin, 2018; Varian, 2018) state that some governments had to create or modify some regulations such as property rights, data privacy, taxation due to business digitalization.

Based on these arguments, we found that the disruptive change that digital transformation is offering to companies is an ongoing research topic. However, the use of the business model term is blurred and used in different ways. In the next section, we analyze the state of the art of the business model concept in order to clarify its scope.

3.3.2 Business Models

(Pateli, 2003) claims that the concept of business model is conceived in different domains such as E-Business, Information Systems, Strategy and Management. From a semantic point of view, the definition of business could be understood as "the activity of providing goods and services involving financial, commercial and industrial aspects". On the other hand, the meaning of model could be referred to as "a simplified description and representation of a complex entity or process". However, in the literature, we can find a lack of understanding and ambiguity about the business model meaning (Linder & Cantrell, 2000). This could be explained as the term is usually referred to the way a company does business e.g. (Galper, 2001; Gebauer & Ginsburg, 2003) or others perspectives that just focus on the model aspect (Gordijn & Akkermans, 2003; Osterwalder & Pigneur, 2004).

Other ambiguities or interchangeable uses of the business models term are related to business process modeling, business strategy, and enterprise models concepts. The former refers to set of activities that are related to generate a particular product or service through the implementation of a specific process (Williams, 1967). According to (Gordijn & Akkermans, 2003) the blurring aspect of these two terms is related to the "business modeling" concept. Regarding the difference between business strategy and business model, some authors considered the business models as an abstraction of a firm's strategy that could be applied to many firms (Peter B. Seddon, Geoffrey P. Lewis, Phil Freeman, & Graeme Shanks, 2004).

However, (Magretta, 2002) claims that these differences are less notable because some authors tend to use it interchangeably. The author also explains that these two terms are linked but business models describe a composition of elements (strategy, business operation, infrastructure, customer value) that fit together, while the scope of strategy is associated with planning and competition. Concerning the terms enterprise and business models, the difference lies in that the first one is related to the enterprise engineering, tasks, process and activities (Bernus, 2001; Wortmann, Hegge, & Goossenaerts, 2001) and the approach of business models is basically centred on value creation and customers.

(Osterwalder et al., 2005) offer a clarification of the term business model. These authors developed a schema based on layers to classify how different authors define the concept of business models. While, the first tier refers to how some authors focused on the business model concept based just on an abstract concept and other as metamodels. For instance, (Magretta, 2002; Timmers, 1998), argue that some companies describe their the business ideas as the whole business model. On the other hand, (Afuah, 2002; Amit & Zott, 2001; Applegate, 2001; Chesbrough & Rosenbloom, 2002; Gordijn & Akkermans, 2003; Hamel & Ruben, 2000; Linder & Cantrell, 2000; Mahadevan, 2000; Osterwalder et al., 2005; Petrovic, Kittl, & Teksten, 2001; St"ahler, 2002; Weill & Vitale, 2002) explain the business model using meta-models which refers to different isolated elements that compose a business model. In the second layer, (Osterwalder et al., 2005) explain how the literature classify business models through a similar type of business ideas sharing common features. This kind of business model taxonomy could be set to particular industries. For instance Mobile Games, Software (MacInnes, Moneta, Caraballo, & Sarni, 2002) Development (Shubar & Lechner, 2004). In the third group, (Osterwalder et al., 2005) describe how different authors use a concrete business model based on specific companies in order to explain their concept and elements For example, Xerox (Chesbrough & Rosenbloom, 2002), Dell (Kraemer, Dedrick, & Yamashiro, 2000), and online business models such as eBay, Amazon (Krueger, Swatman, & Beek, 2004).

As we can see the term business model is a broad topic and still under research due to the adoption of new technologies. In the next section, we will cover the literature about companies using open (government) data in order to define their business models.

3.3.3 Open Data Business Models

The recent adoption of policies that allows the release of open data by governments around the world can be perceived by entrepreneurs as business opportunities to develop or innovate new products and services. Nevertheless, the literature on this topic is still under development not only about the use of open data by entrepreneurs but also the implementation of a business model composed of open (government) data.

Some authors argue that the implementation of open (government) data for commercial purposes promises to produce economic value through developing or innovating products and services. For instance, (Magalhaes, Roseira, & Manley, 2014) analyse 500 companies in the US that use open (government) data as part of their production process. These authors propose a scheme to classify these firms according to its business model. The methodology described in order to create this taxonomy is based on a content analysis which extracts inference from text and then implementing a clustering approach. Furthermore, the authors -based on this taxonomy- analysed the value proposition for each category. Then, they describe a framework which explains the value creation process that these companies offer to the market when they are implementing open (government) data. This taxonomy is composed of enablers, facilitators and integrators. The former is referred as companies are providing technologies for the use and implementation of the open data ecosystem. For instance, hosting, cloud computing, or data management software. Companies using the facilitators business models are those that support or facilitate the access and exchange of data between suppliers (governments) and claimants (entrepreneurs, developers, activists, society in general). For example, API, web applications, merging different sources of databases. Integrators are companies on which the business model is based on include or combine open (government) data in their production process in order to enhance or innovate their products or services.

As an example of the abstract categorization of the conceptualization of business models proposed by (Osterwalder et al., 2005), we find that (Ferro & Osella, 2013) propose 8 archetypes that they labeled as business models for companies that use public datasets. The classification of these 8 archetypes is composed of Premium Product Service (PPP) which implies a payment in exchange for a high intrinsic value, this could be done by granting access to additional features of the product or service provided. The Freemium Product Service (FPS) model which essentially offers basic features of the product without any cost, the customer has the option of switching to the Premium model. The Open Source Alike (OSA) refers to a free release of the data in order to attract customers or promote their philosophical enterprise culture (similar to the Open Source movement), this archetype is adopted by companies as a cross-subsidization in which the operational costs are covered by other business lines. Infrastructural Razor And Blades (IRB) is a model in which companies are operating as intermediaries that enable and promote the access of public datasets. This model operates based on a low price or free trial (razor) that potentially incentive future purchases or services (blades) i.e. API and cloud computing services. Regarding the Demand-Oriented Platform (DOP), the enablers (entities that facilitate the access of public datasets) provides to end users (developers, entrepreneurs, society in general) access through API once the datasets are categorized and tuned in terms of formats allowing the democratization of datasets, the implementation of this model could be a mix of freemium and premium strategies. Conversely, the Supply-Oriented Platform (SOP) is based on intermediary stakeholder playing an infrastructural role. In other words, suppliers (in this context governments) are

the customers or the part that assumes the cost of publishing and releasing datasets in instead of end users. The archetype based on Free as Branded (FaB) advertising is based on persuading customers towards a brand, company. In the context of open data, the authors argue services proposed using this model tends to have a positive externality instead of a direct revenue impact. The last archetype proposed by the author is called White-Level Development (WLD) which refers to the chance to act as outsources for entities that do not have the knowledge, expertise or resources to adopt or implement services such as data retrieval, software development, service maintenance and so on.

According to (Yu, 2016), more research is required to understand the value proposition and value creation of business models based on open (government) data. In other words, value proposition and value creation are considered key elements in a business model because the former is the nucleus of a firm, while value creation is the outcome expected by customers, for this reason, it is envisioned that a successful company design its business model based on a value-centric scheme see (Ferro & Osella, 2013; Kaasenbrood et al., 2015; Zeleti et al., 2016). According to his work, a value proposition defines what types of value and how particular value can be delivered to customers and society in general. The author also argues that value creation is the implementation of actions to produce value and benefits. Moreover, Based on a Design Science Research (DSR) and Action Research (AR) approach, the author proposes a value-centric framework for Open Data Applications (ODA) composed of different elements such as stakeholders (individuals, business companies, NGOs) systems (cloud, web, mobile, and network infrastructure), services (developed by the public sector in order to access datasets, policy-making), value (benefits for the stakeholders through the implementation of systems and services elements), objectives (tasks that are expected to achieve and measure), performances (KPI indicators based on objectives), resources (financial, human, technical and others to be required to operate), costs (expenditures that need to be done in order to start and implement activities), functions (related to processes and activities required to performs the firm's tasks), and strategies (action plans in order to achieve goals).

(Zeleti & Ojo, 2017) argue that some research has been done on the topic of Open Data Business Models (ODBMs) but the value orientation of these models are still blurred and therefore the value creation is not clearly understood. Furthermore, the authors claim that in the open data sphere the business model concept is used interchangeably with revenue models, pricing strategies, distribution models, marketing techniques, and architectural models. Based on qualitative approach and following a Design Science Research (DSR) methodology, the authors propose a 6 value (6-V) framework that describes a business model. This 6-V schema is composed of Value-proposition which stipulates the benefits that the firm offer in their services, products, or distribution channel. The Value-adding process includes human, technical, financial, and partnership elements performing operational activities, strategic planning, and knowledge management. The Value-in-return obtained through revenues or commissions. The Value-capture is generated by retaining some percentages in operations.

The Value-manager refers to the role that human capital plays in firms. The Value-network refers to the ecosystem generated by all stakeholders (customers, suppliers, partner business, etc) involved. This framework 6-V was developed finding patterns and clustering business models based on their value orientation. According to the author analysis, the result of this categorization defines 5 business models freemium, premium, cost saving, indirect benefits and parts of tools. However, these categories could be summarized in 2 main business model freemium and premium because the other models could be considered as subcategories of these two

3.3.4 Summary

According to the literature review, the business model concept has been used interchangeably or in an ambiguous way. Furthermore, the analysis of a firm's business models can be a complex task because companies have the option to mix or select specific components of the business model in order to make a difference to their competitors or create a unique brand. However, as is mentioned in the literature, a core aspect of a business model based on open government data are the value proposition and value creation.

In the next section, we will explain the source and collection process of the data used in this chapter. Moreover, in the next sections, we will extract and analyse the components of the business models adopted for companies that are using open (government) data.

3.4 Data

The data that is part of this chapter was collected by the Open Data Incubator for Europe (ODINE) program which is a project funded by the European Commision through the research and innovation program Horizon 2020. The aim of this project is to attract and fund entrepreneurs and companies that are using open government) data as part of their business ideas or production process. The project is an industry-focused network of open data startups and SMEs around Europe.

The consortium in charge of the creation, organization, dissemination, and implementation of this project was composed of 7 entities from the academic and private sectors listed as follows the University of Southampton, University, Fraunhofer Institute for Intelligent Analysis and Information Systems, Open Data Institute, Open Knowledge Foundation, Telefonica Open Future and the Guardian. The structure of the project is constituted for 5 pillars. The first one is referred as "Competitive Call" aiming to recruit, select and support the most innovative business ideas that are using or producing open data. The second pillar is "Data and Computing Services" in which the consortium offers infrastructure (cloud-based Data-As-a-Service) to SMEs in order to facilitate the implementation of open data business ideas. "Business Incubation" is the third pillar that supports open data entrepreneurs through mentorship, training, networking in the data-driven ecosystem. The fourth pillar is "Engagement and Dissemination" aiming the promotion and orientation about the use of open data as an asset to value-generation. The fifth pillar is related to the "Exploitation and Sustainability" of the successful applications that will be incorporated in the incubator program.

The ODINE's project ran from 2015 to 2017 promoting and disseminating 8 open calls around Europe. The consortium received 1173 business proposals submitted by entrepreneurs using open data as part of their production process from different sectors such as agriculture, education, consultancy, finance, information and communication, real estate, transportation, among others. The process to apply to the incubator program was submitting an application form, which describes a business proposal in pdf format that answers a series of questions (see Appendix A.8. Application template). These business ideas were analysed by two independent external reviewers that evaluated the feasibility of the proposal based on the following criteria: Idea, Impact and Team. The first criteria evaluate the strength or novelty of the idea, the usability, creation or contribution that the business proposal generates to the open data ecosystem. In other words, entrepreneurs or companies that are using open data as an asset should demonstrate that they can explain the core idea of their business proposal and demonstrate a clear differentiation with their competitors. The second criteria is related to the Impact which should describe the value proposition, business scalability, size of the target market and the identification of potential impact in the economic, social or environmental sector. In this section, applicants should explain what is their revenue strategy and the benefits (value proposition) that they are offering to their customers, they

also should describe the target market and explain the potential impact in the different sectors such as economic (job creation, innovation, saving costs, improving decision making), social (empowering less privileged groups or promoting culture) or environmental (entrepreneurs or companies focused on reducing carbon emissions, encouraging reuse, etc). The third criteria focus on Team composition and Budget structure defining the skills that human capital has in order to materialize the business proposal and defining the sources of funding and a likelihood of success. An application template containing all questions about Idea, Impact, Team and Budget is found in the appendix sections A.8.1, A.8.2 and A.8.3

The evaluation process of the business proposals was performed in 3 stages called eligibility check, external review, and interviews. In the first phase, an internal team of the consortium validates that each applicant meet the requirements in terms of country eligibility, signed a declaration of honour that does not have a conflict of interest and submit a business proposal filling the following considerations: No more than 4 pages long and all questions about the idea, impact, team and budget answered. The second stage was in charge of two external evaluators[2] who reviewed the applications assigned by the consortium. Evaluators assigned a score of Excellent (4), Good (3), Average (2), Poor (1) for each section (idea, impact, team and budget) and each application (business proposal submitted). Then, evaluators were also asked to select the best 3 promising applications through a scoring system composed of Invite to Interview Yes (3), Maybe (2), No (1). At the end of this phase, a list of companies were invited to the next stage. The Interview phase was in charge of an internal team by members of the consortium and two members of the external evaluator's team. The structure of the interview was composed of 5 minutes for a company presentation and 25 minutes for answering questions made by internal and external evaluators. Finally, there are other 30 minutes for analysis and deliberation of the company profile and business proposal but this process only involves internal and external evaluators. An important point to mention is that the external evaluators were selected from candidates suggested by all members of the consortium in order to create a more impartial and transparent group of evaluators. This selection process was conducted balancing nationality, gender, expertise and domain knowledge.

The next section describes the methodology implemented in order to extract and analyse the data collected by ODINE.

[2] `https://opendataincubator.eu/resources/odine-external-evaluators/`

3.5 Methodology

In order to analyse and extract insights from the data collected by ODINE's project, we implement a text mining approach due to the data collected by the consortium during the time life of the project was through template in which applicants described and submitted their business proposal as a pdf format. This type of data is categorized in the data science domain as unstructured data.

The main goal of text ming is to analyse, quantify and extract useful insights from unstructured (textual) information. One of the main challenges dealing with text data is that it is sparse (the number of cells in a table that are empty) and high dimensional. Another is the semantical meaning of the text analysis. In other words, representing the text data as named entities (cities, people, organizations) tends to show interesting patterns; however, it depends of the kind of data and research scope. (Kwartler, 2017) proposes that analyzing text data and implementing text ming is branched into 2 broad types "bag of words" and "syntactic parsing". The former treats every word as a unique feature of the document. This means that this approach discards the sequence of the words in all sentences of a text and handle it as a bag of words. Conversely, syntactic parsing analysis is based on word syntax which specifies a set of rules that define the elements of a sentence. This approach tends to recognize grammatical features of the words such as nouns, articles, verbs and adjectives because it uses part of speech (POS) tagging techniques to make a sentence. However, certain types of syntactic structures tends to be more complex and requires more computational infrastructure. Figure 3.1 illustrate the difference between these two approaches.

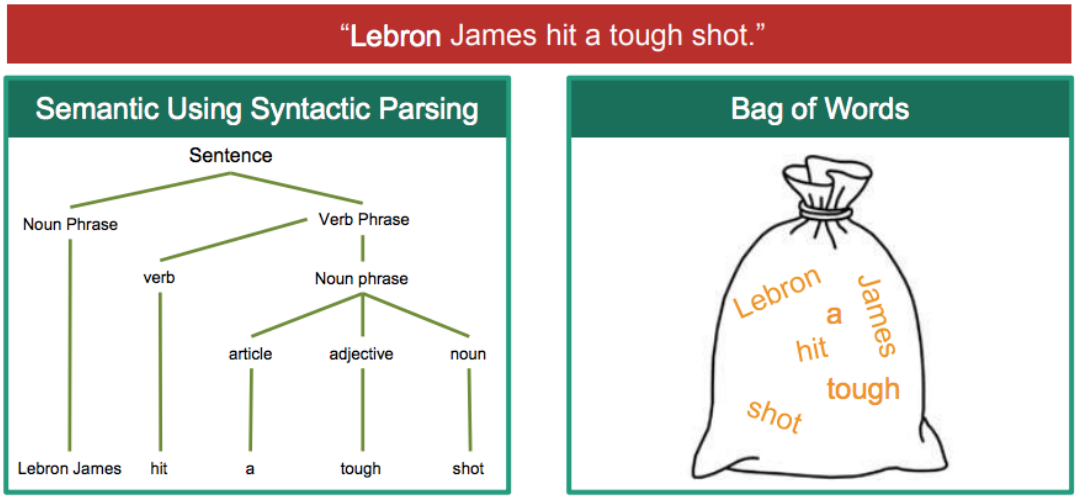


FIGURE 3.1: Illustrates the text mining approach.

In this research, we implemented a bag of words approach because our goal is to analyse the text contained in the business proposals through a document classification technique. In order to do so, the next section describes the research workflow proposed.

3.5.1 Research Workflow

The text mining research flow involves several stages that involves different subroutines such as the identification of the problem to solve, data collection, steps to extract, clean and manipulated the text data and the development of models to extract valuable insights. Figure 3.2 illustrates the research flow developed.

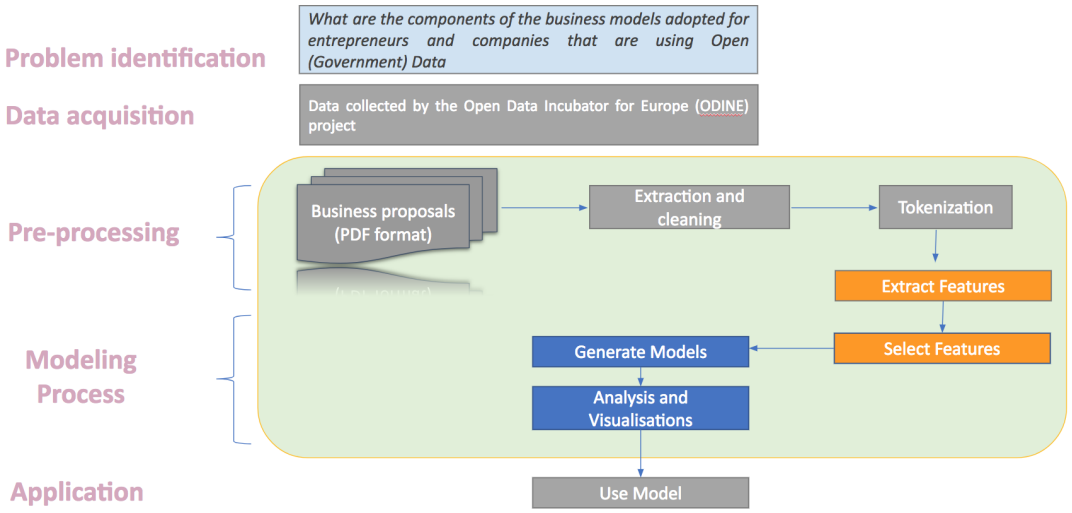


FIGURE 3.2: Shows the text mining research workflow.

3.5.1.1 Problem Identification

The purpose of this chapter is to understand the entrepreneurial ideas, the economic, social or environmental impact and the team composition that individuals and companies are describing when they are using open data as part of their business proposals. All these components are concentrated in 3 main pillars (Idea, Impact, and Team, see appendix A.8.- ODINE application template) contains information that will help us to answer our research question: *What are the components of the business models adopted for companies that are using open (government) data ?*

3.5.1.2 Data Acquisition

We analysed the data collected by the ODINE project which are 1173 business proposals submitted by entrepreneurs from different sectors. These applications have 22 questions divided into 3 sections (idea, impact, and team) that ask to describe the core idea, impact and team composition of their business proposal. Entrepreneurs that applied to the ODINE programm should downloaded this template from the ODINE's website and answered these questions. Then, they should submit it using the Easychair platform in a PDF format.

3.5.1.3 Pre-processing

The first step was to collect all the applications submitted and organise this information for each round (ODINE run 8 open call during their lifetime). Then, we merge and complement this data with additional information such as the application's country, region, and the industrial sector according to the Statistical Classification of Economic Activities[3] established by the European community. All this additional information was collected when entrepreneurs submitted their application through the EaseChair platform. The final result is a dataset composed of 1173 business proposals from different regions and industrial sectors across Europe. One of the main issues dealing with unstructured data (in this case, text) is to analyse and extract valuable insights from it due to its sparsity and high dimensionality.

Extraction and Cleaning

Once we have all the data in a PDF format, the next stage is extracting insight from them. As the previous chapter, we use the open source software "R" and several of their libraries in order to convert and extract from PDF files to a corpus data (character) format. For this purpose, we use the pdftool library (Ooms, 2017) which their purpose is to extract text and metadata from PDF files. Once we have the text data (corpus), the next step is cleaning it following these preprocessing tasks removing punctuation, tolower, stripping extra whitespace, removing numbers, and removing "stopwords" (articles or common words that do not provide additional information). For these tasks, we use the tm (Feinerer, Hornik, Software, & GPL Ghostscript), 2015) and tidytext (Silge & Robinson, 2016) R packages.

Tokenization and Text Mining

In text mining, a token is referred to a unit of text. In other words, tokenization is the process to split textual information into individual words or terms (Vijayarani & Janani, 2016). The main reason to apply this process is that text data is only a set of characters (words, punctuation, numbers, alpha-numerics, etc); therefore, these require being segmented and clustered in order to extract insights from them.

Descriptive Analysis and Visualisations

Once the data is already clean and in a format that we can manipulate, we start providing an overview of the open calls made by the ODINE project across Europe. We develop a descriptive analysis showing a).- number of countries covered per these open calls detailed by region (Southern, Northern, Western, Eastern) and income group (High, Middle, Lower), and b).-number of applications submitted per country and industrial sector. This information

[3] `https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN`

gives an overview of the number of companies per country and sector that are considering Open Data as a digital asset in their production process. Furthermore, we provide information about when these companies were created in order to explore for how long have been in the market; however, it is important to mention that not all companies provided this information because of this data was not explicitly asked to provide it during the application's submission process. Second, we provide a text mining analysis of the company profile to explore the description made entrepreneurs about their company and the type of services that they are providing. Third, we adopt a Use Cases approach to explore 5 of the 57 companies that were selected by the ODINE consortium as part of their incubation program. We fully examined these 5 business proposals using the Business Model Canvas design aiming to provide a context of the key partners and activities, the value proposition that they are offering, customer relationships, channels segments, cost structure and revenue streams. Finally, we analyzed (using text mining) 1173 business proposals that applied to the ODINE program but they were not selected to participate in the incubation program. However, these companies that submitted their business proposals generated valuable information about their business idea, the economic, social or environmental impact that they were looking for and the team composition.

In the next section, we will present the results and discussion of our analysis.

3.6 Results and Discussion

This section shows the results of our text mining extraction and it offers a descriptive analysis of this data, the structure is as follows. First, we provide an overview of the open calls made by the ODINE project across Europe about the number of companies per country and sector that are considering open (government) data as a digital asset in their production process. Second, we provide a text mining analysis of the company profile to explore the description made entrepreneurs about their company and the type of services that they are providing. Third, we adopt a Use Cases approach to describe 5 companies that completed successfully all stages of the selection process and were elected by the ODINE consortium as part of their incubation program. Four, we explore 20 business proposals -which also were selected by the ODINE consortium- using the business model canvas design aiming to provide a context of the key partners and activities, the value proposition that they are offering, customer relationships, channels segments, cost structure, and revenue streams. Finally, we analyzed through a text mining approach 1173 business proposals submitted to the ODINE program. However, these proposals were not selected to participate in the full incubation program due to budget restrictions. Nevertheless, these business proposals generated valuable information about their business idea, the economic, social or environmental impact that they were looking for and the team composition of these companies when they are using open data as a digital asset; therefore, all this information is related to our research question.

3.6.1 ONDINE's Overview

ONDINE's project ran from 2015 to 2017, during this period the project disseminate and ran 8 open calls around Europe. Each call had a duration of 2 months in order to submit a business proposal based on open data. Each round was composed of 5 steps: 1).-Application reception, 2).-Eligibility check, 3).-Review, 4).-Interview and final selection, 5).-Negotiation. The result of the data analysed indicate that after a slow start (the first call only attracted 68 applications) the number of submissions was stabilized between 122 and 152; however, in the last call the project received more than 200 business proposals. Figure 3.3 shows the number of applications per round.

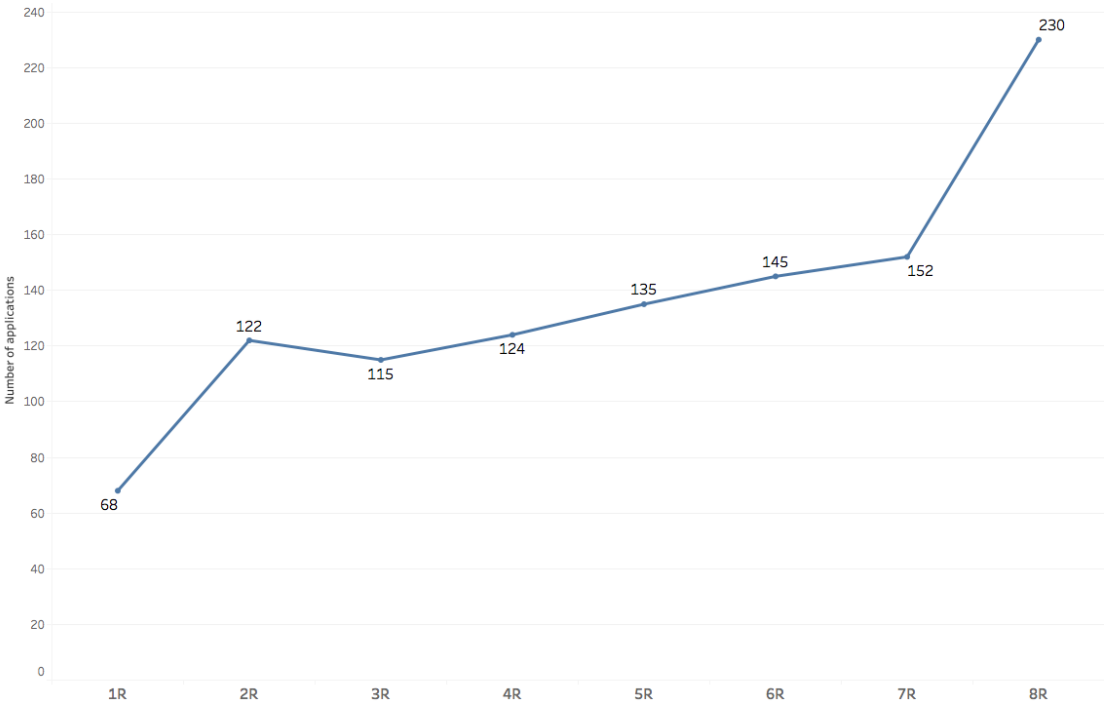


FIGURE 3.3: Shows the ONDINE's open calls.

One of the main goals of the dissemination phase was to trigger awareness about the open data movement and its potential economic and social impacts. Another was the propagation of information and engagement activities for data entrepreneurs and venture capitalists. ODINE's intention was to attract public attention to promote open (government) data as an asset for business proposition and innovation. Furthermore, members of the consortium also wanted to promote the project around countries that are part of the European Union in order to cover as many entrepreneurs that are been using open (government) data in their business as possible. Figure 3.4 illustrates the European coverage and the number of submissions per country

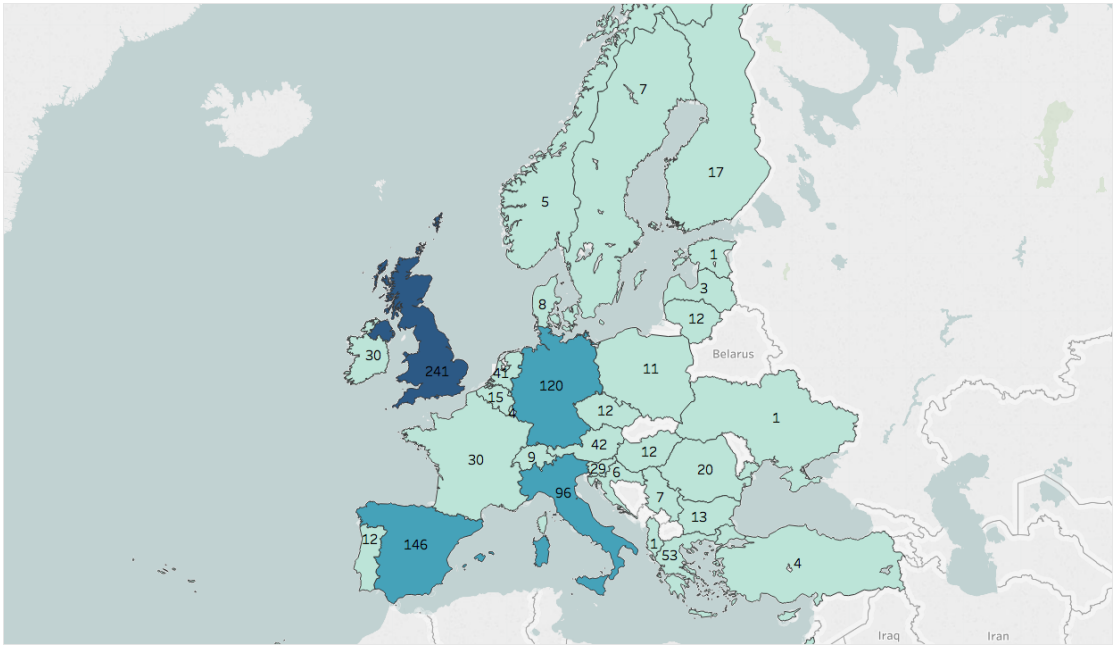


FIGURE 3.4: Illustrates ONDINE's coverage in Europe.

Results show that the distribution of submissions was concentrated mainly in the South and North of European countries due to entrepreneurs from these regions presented more than 600 business proposals. Countries from the Western region submitted 269 applications and only 70 business proposals were received from Eastern countries. Figure 3.5 illustrates the number of applications per region

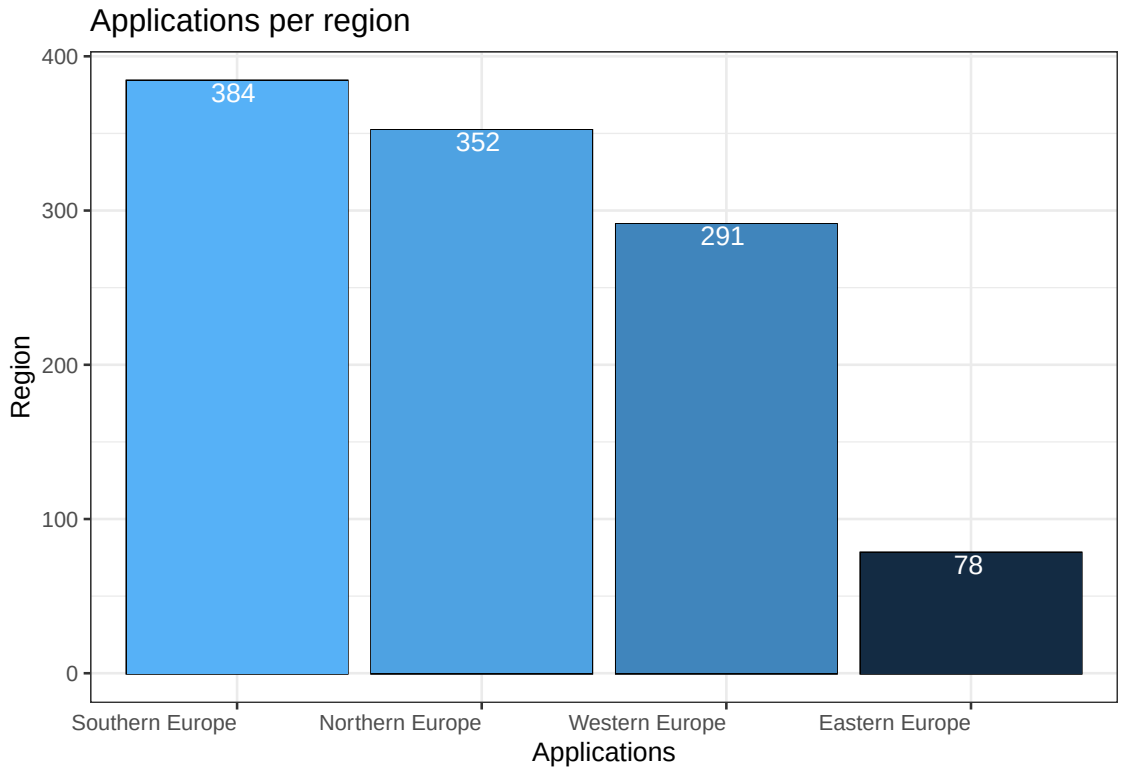


FIGURE 3.5: Shows the number of applications per region

The main countries that applied to the ONDINE's program are considered by the World Bank as high-income economies. This information could act as evidence of our previous assumptions (chapter 2: Entrepreneurship and Open (Government) Data) about the determinants in the adoption and use of open data as a digital asset at the country level are more notorious in high-level economies. These determinants are related to the trust on their institutions by society, the type of infrastructure (in this case, internet adoption, and bandwidth costs), the level of education and digital skills just to mention some of them. Figure 3.6 display the number of applications according to its income group.

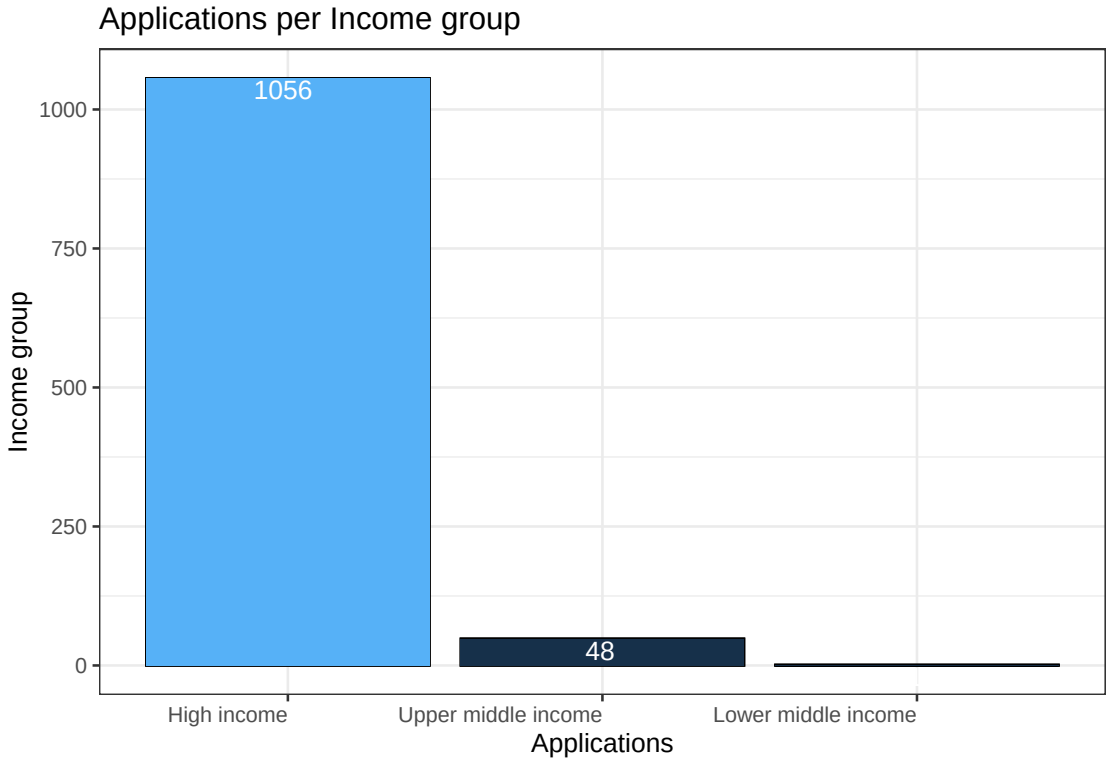


FIGURE 3.6: Displays applications per income group

Regarding the distribution of applications per countries, the United Kingdom is the economy with a larger number of submissions (with 241 applications) followed by Spain (146), Germany (120), Italy (96), Greece (53), Austria (42), Netherlands (41), Ireland (31) and France (30) which are the top 10 countries in terms of business proposal submitted to ODINE. Figure 3.7 shows the number of applications submitted per country. These top ten countries also have been in the first positions during the last 4 years in the Global Open Data (GODI)[4] and Open Data Barometer (ODB)[5] indexes which potentially could imply a link between the actions and open data policies promoted by governments at the national level and the business opportunities perceived by entrepreneurs.

[4] https://index.okfn.org/place/

[5] https://opendatabarometer.org/?_year=2016&indicator=ODB&region=:EU

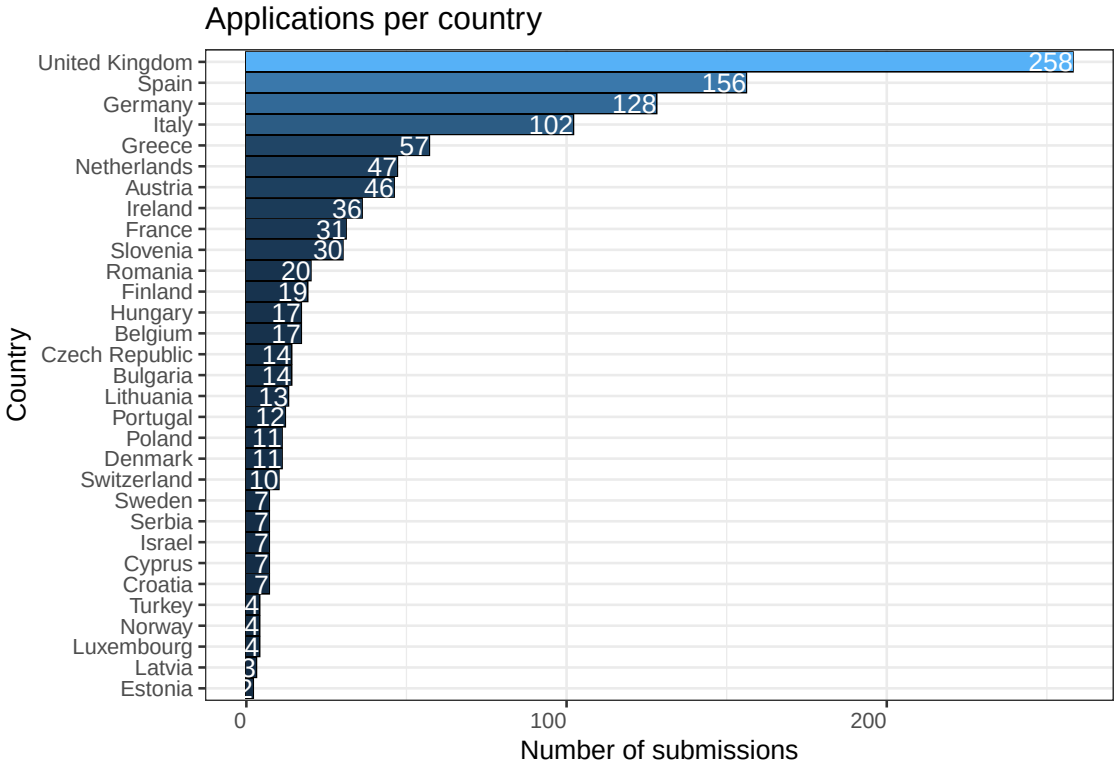


FIGURE 3.7: Shows the number of applications per country

Another important aspect that reflects these results is that the predominant sector from which ODINE received more applications is the Information and Communication sector (with 606), An important consideration is that according to Statistical Classification of Economic Activities in the european community, this sector is related to the production, processing, and distribution of data, communication, information technology and other information services activities. The next sector is related to the Professional, Scientific and Technical activities (208), Other services (63), Agriculture (50), Education (26), Health (15), Finance (11), Real Estate (7), Entertainment (6) and Transportation (6). Figure 3.8 illustrates the number of applications submitted per economic sector.

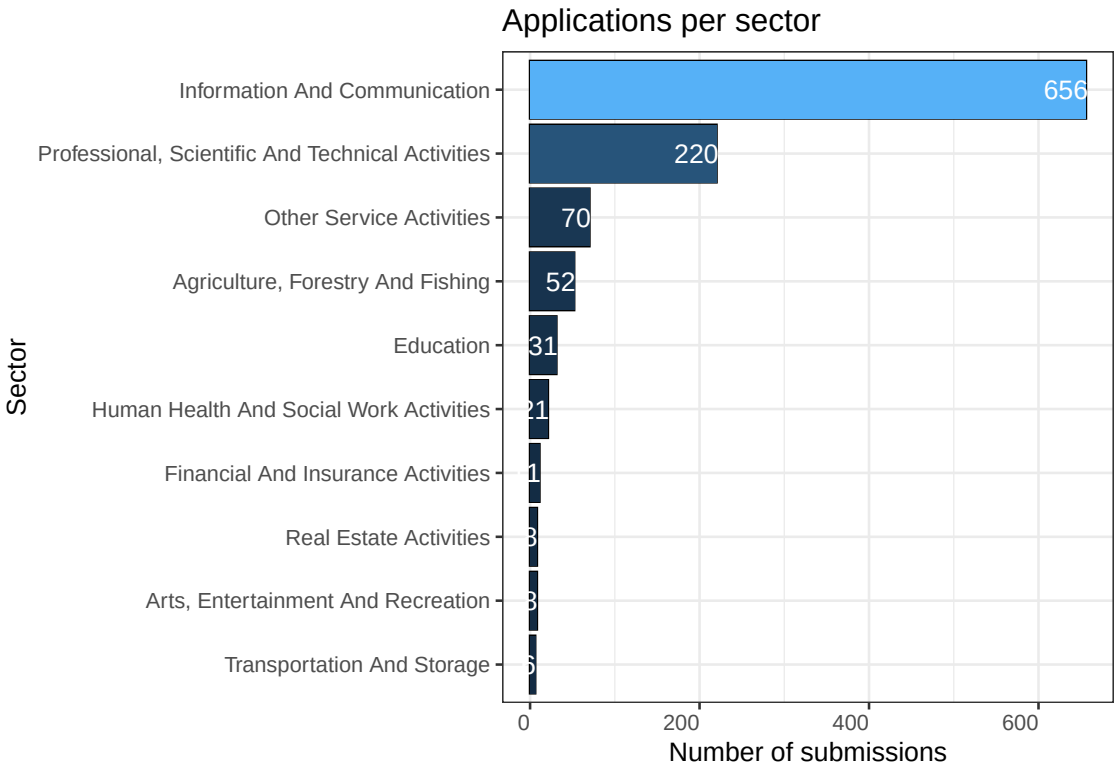


FIGURE 3.8: Shows the number of applications per country

Finally, we extracted the information about when these companies were created in order to explore for how long have been in the market; however, it is important to mention that not all companies provide this information because of this data was not explicitly asked to provide during the submission process. Our results show that more than 50% of the companies started operations between 2012 and 2017. Figure 3.9 illustrates some of the companies that applied to the ODINE program and that provided information about when they starting operation.

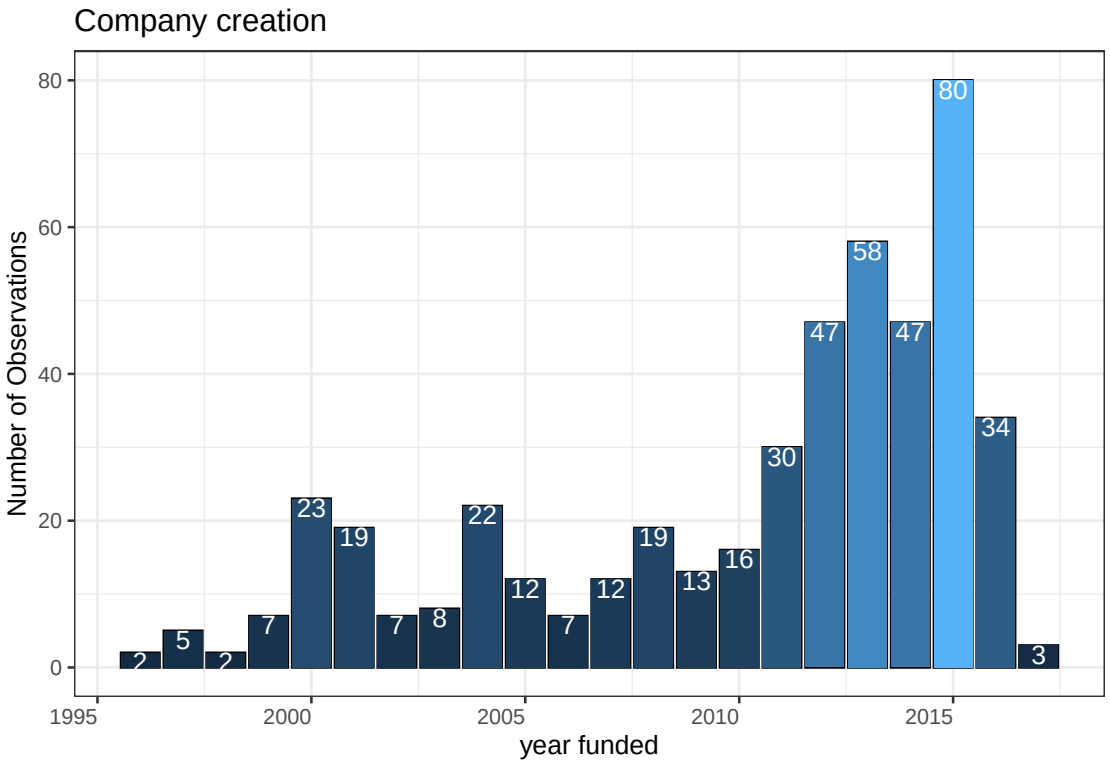


FIGURE 3.9: Shows the year that some companies started to operate

Once we have this overview of the ODINE's data, we move to analyze the companies' profile described on these business proposals. Then, we describe 5 companies that were selected by the ODINE consortium using a use case approach. Later, we examine the business model canvas reported by 5 companies that were selected by the ODINE consortium to move from the application process to the incubation stage. Finally, we analyze the description made by entrepreneurs of the business idea, impact and team composition in their business proposal when they are using open data in their production process.

3.6.2 Company Profile

We extracted data related to the companies profile which is a general description of what the company is doing and what type of services they are offering. We start grouping our data by each round and then apply text mining scripts concatenating a series of two adjacent tokens referred as bigrams (Tan, Wang, & Lee, 2002) in order to extract a more meaningful text and have a better idea about the company profile. Our results show that most of the extracted words suggest that companies are related to the data-driven domain, performing activities such as artificial intelligence, semantic web, linked data, knowledge management, software development, cloud computing, data science, and data analytics. An interesting point about the profile of these companies is that one of the main challenges of unlocking the value to any data is based on the capability to extract insights from it. For these companies, open

data is the raw material that needs to be transformed into information and knowledge as the baseline to its products or services.

This result is consistent and strongly associated with the outcome of companies sector, which illustrate that most of these companies belong to the "Information and Communication" field[6], (this sector is related to the production, processing and distribution of data, communication, information technology and other information services activities) followed by "Professional, Scientific, and Technical Activities"[7] (this sector includes specialized professional, scientific and technical activities that require a high degree of knowledge, training, and specialization skills) These two sectors together account for 80% of the applications submitted to the ODINE programme. Although most of the companies are included in these two sectors, all business proposals submitted are implementing the use of ICT techniques and skills to retrieve and process data. Figure 3.10 extracts the bigram grouped by round.

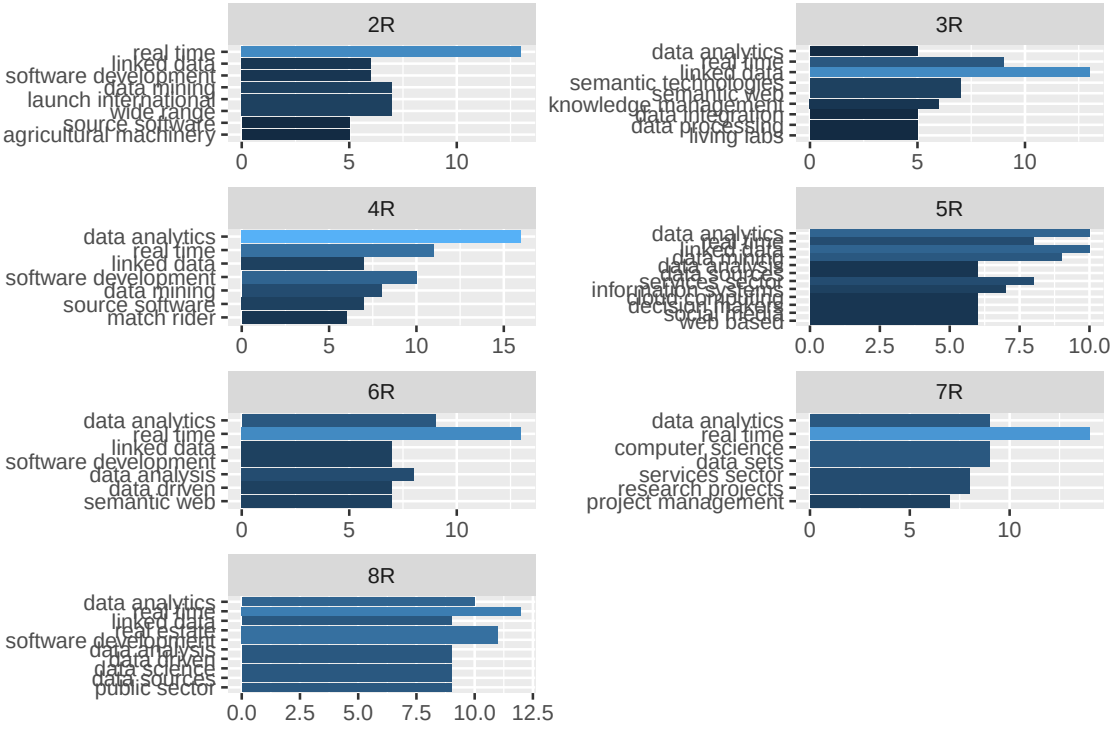


FIGURE 3.10: Bigram grouped by round

Now, that we have a more clear picture of these data-driven companies about how they are transforming and using open (government) data, we move to analyse deeper the business model composition describes as what are the key partners and activities, the value propositions that are they offering, customer relationships, channels, segments, cost structure

[6] https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=DSP_NOM_DTL_VIEW&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=18514214&IntKey=18514214&StrLayoutCode=HIERARCHIC&IntCurrentPage=1

[7] https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_LINEAR&StrNom=CL_NACE2&StrLanguageCode=EN&IntCurrentPage=26

and revenue streams. All these elements were described using the business model Canvas framework and taking 5 uses cases of companies selected by the ODINE team.

3.6.3 ODINE: 5 Use Cases

3.6.3.1 5 Selected companies and their Business Model Canvas

After a dissemination process of 8 open calls from 2015 to 2017 through newspaper, social media, webinars, promotion in technological and data forums in countries that are part of the European Union. The ODINE's team in conjunction with external evaluators selected 57 business ideas based on their value proposition, market size, and revenue forecast. The information about the business models is publicly available on the ODINE website only for 20 companies[8]. In this section, we describe in detail 5 business ideas proposed by these entrepreneurs selected by ODINE using a use case approach and adopting the Business Model Canvas framework in order to generate a structure and context of their key partners and activities, the value propositions that are they offering, customer relationships, channels, segments, cost structure and revenue streams. Furthermore, we analyze all these 20 companies using a text mining approach in order to find additional insights. We considered important to develop this approach because is not clearly stated in the literature how entrepreneurs are using open (government) data in order to extract insights from data and create products and services and sometimes this process is perceived as a black box.

Use Case 1: OpenCorporates

Opencorporates[9] is a company based in the United Kingdom and it is the largest open data publisher of companies information around the world. This company submitted a business idea through an ODINE call and it was accepted by the consortium. This company submitted a business idea through an ODINE call and it was accepted by the consortium. Their proposal is based on their business diversification and scalability, developing a new product called Opengazettes[10]. They are planning on include government gazettes about critical company-related notices to their database in order to improve insights into activities of companies in the EU. Their key partners are governments because they are the main source of the gazettes publications. Data extraction, quality assurance, and integration to their database are some of their key activities. The human capital in terms of technical knowledge and sales expertise are considered as their key resources. Their value proposition is based on collecting, curing and integrating statutory publications that contain crucial company information such as formation or change in share capital, or environmental and legal regulations. This information is made by governments and companies that are published

[8] https://opendataincubator.eu/resources/
[9] https://opencorporates.com/
[10] http://opengazettes.com/

in most EU countries but this data are little known and difficult to consolidate and track. Regarding their customer relationships, they have a dedicated sales account manager in order to reach their customers they are proposing to develop alerts based on services tracking. Furthermore, they are planning to develop this service in order to be implemented in a CRM platform. Their customer segment is composed of professional services business (legal professional management, consultants, industry-specific services professionals) no only for European business but also for any company in the world which deals with EU companies. The cost structure described by these entrepreneurs are related to salaries for developers, data quality team, sales staff and marketing plus technical infrastructure (computers, cloud services, electricity). Their revenue stream is based on a share-alike strategy mixing the freemium and premium service. The former offers a "search service" to the gazettes data collected and supporting user based on data contribution and/or creation. The premium offers an API service improving searching and retrieving services.

Use Case 2: GreenSpin

GreenSpin[11] is another selected company by the ODINE consortium. The business proposal of this company is based on offering open geographical data, developing analytics and decision support tools that can help customers from agriculture and related sectors. Their key partners are providers of geospatial technologies, institutions conducting research about agriculture, multipliers and distributions from the agricultural sector, and farm management systems. The key activities are collecting and building scalable data streams, consulting and innovation support, and perform geographical data analytics. Their key resources are human capital expertise, cloud technologies, and transformation, extraction and results based on data analytics. The value proposition of this company is offering support to analyse and maximize through data analytics their production process. Moreover, the company also provides consultancy advice to improve efficiency and simplify the implementation of precision farming operations. Their customer relationship is made by personal assistance and digital services such as automated map and data services. The channels used by the company are social media, service delivery, personal and online support. The customer segment is composed of the business chain around agriculture (seeds, chemicals, machines), large farms and public authorities. The cost structure of this company are salaries (developers, marketing and sales) technological infrastructure and investments in R&D. Their revenue streams is composed of consulting fees, licenses (platform, data, and map services) and project-based payment for application development.

Use Case 3: OpenLaws

Another company selected by the ODINE consortium is OpenLaws[12] aiming to connect open legal data sources from the EU and member states. The key partners are the University of

[11] https://www.greenspin.de/
[12] https://openlaws.com/home

Amsterdam, Salzburg University of Applied Sciences, University of Sussex, London School of Economics, Alpenite srl, national governments databases and other entities such as European Commision and The Council of Bars and Law Societies of Europe. The key activities of this company are the legal content integration, the development of technological infrastructure to analyse this information and the marketing and selling process. Their key resources are the human capital and legal expertise, IT infrastructure, the collection and processing of data and metadata, the development and application of Intellectual Property legislation and trademarks to operate across EU countries. Their value proposition is offering free legal search for laws and judicial decisions in Europe through their platform. They also offer solutions collecting, curing and aggregating legal data and specialized service to business, legal experts, scholars and public bodies. The company reports that the customer relationship is mainly implemented using online and social media channels, trying to reach a broad spectrum of clients. The customer segment is composed of citizens, public, private and academic sectors. The costs structure defined in their business model is software development, hosting and storage costs. The revenue stream is a mix of freemium and premium services in which the company offer free access to basic functionalities and Open Data content and a charge will be imposed for advanced features and content.

Use Case 4: CommoPrice

The next company selected to describe its buses model is CommoPrice[13]. This business proposal is based on a web portal publishing commodity prices based on Open Data covering several domains such as agriculture, fertilizers, metals, chemicals, plastics, energy. The key partners for this company are data providers from global and regional markets, organizations, institutions and companies. The key activities are related to the data collection, integration, refining, and broadcast, the development and maintenance of the platform and the customer care. The key resources are scrappers, algorithms treating the data and human capital with knowledge in data structure and commodity expertise. The value proposition of this company is to collect and provide accurate commodity price references through a web platform helping purchasing teams track efficiently commodity prices. The customer relationship is made mainly online (email, web demo), showing specific data requests or product feedback. Their customer relationship is implemented through their communication channels such as SEO, press, social media, newsletter and professional networks. The customer segment is composed of industrial companies, agrobusiness companies, retailers, SMEs, consulting and auditing firms. The cost structure reported by this company is mainly in the human capital (80%) between developers and business solver, the other 20% is reported in general and administrative costs. The revenue stream is based on a mix of freemium and premium strategies in which they leave some basic features for free and additional services are charged.

Use Case 5: Viomedo

[13] `https://commoprices.com/en`

Viomedo[14] is another company selected by ODINE. This business idea is related to a platform that connects patients with clinical trials opportunities. The platform aggregates 2,000 clinical trials that are open to patients and their doctors in order to consider participating in a clinical trial as a therapeutic option. The key partners of this company are patients support groups, researchers, doctors, and publishers. The key activities are related to analysis of trial protocols, translation of protocol into patient-centric language, training of sites, developing the platform, and enterprise sales to acquire new sponsors. The key resources are software to match patients and distribute this information, know-how in patient recruitment and Viomedo brand. The value proposition is giving access to information and connecting patients to innovative treatments, getting benefits from better care inside the trial and contributing to medical progress. Furthermore, reducing time to market for new therapies and trial operations costs. The customer relationship is connecting patients through their platform and sponsors having access using an account management. The communication channels are through patients support group, online marketing, patient databases, publishers, conference and trade shows. The customer segment is composed of patients that require high medical need (e.g. oncology, rare diseases), rapid development (neural diseases) or bad usability (chronic diseases) and sponsors. The cost structure is divided into personal, marketing, IT infrastructure, software and administrative costs. The revenue stream is free to access the platform to patients and charging fees to the sponsor per each trial published.

Once we analyzed these 5 companies selected by the ODINE team to their incubation programme and we have provided a context and description of the composition of their business model (key partners and activities, the value propositions that are they offering, customer relationships, channels, segments, cost structure and revenue streams). In the next section, we proceed to systematically analyze using a text mining approach the data available of other 20 companies selected by the ODINE consortium in which describe their business model canvas. The goal of adopting this approach is to find patterns in their business model descriptions

3.6.4 ODINE: 20 Selected Companies

3.6.4.1 20 Selected companies and their BMC using Text Mining

Key Partners

We start analyzing who are the key partners or suppliers for these companies. Entrepreneurs are describing in their business models that although the data generated and publicly available by governments is one of their key partners, the private and academic sectors are also playing an important role as partners. This means that entrepreneurs are combining data from private and/or academic sector in order to enhance their value proposition and create a differentiator in their products or services. Figure 3.11 illustrates the type of sector in which

[14] https://www.viomedo.de/

entrepreneurs described who are the key partners and/or suppliers. For example, the private sector is composed of companies, (also mentioned as a business, or firms), consultancies and manufactures that provide some kind of service or asset that it is part of the business process. The public sector is composed of governments, regional entities, or public agencies across the European Union that act as a source of datasets for the business proposal. The academic sector is composed of research centers, universities, or institutions that collaborate in some way with entrepreneurs in order to develop their idea.

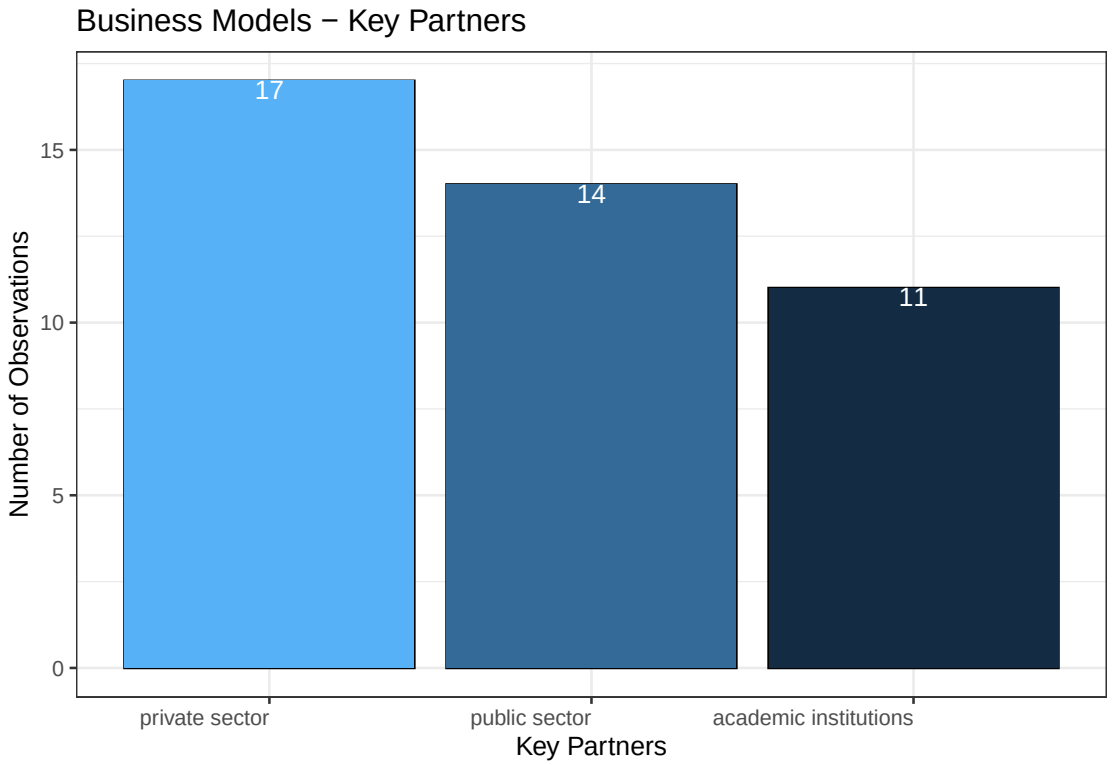


FIGURE 3.11: Shows the sectors that entrepreneurs describe as key partners

Key Activities

Regarding the key activities that entrepreneurs described in their business models, we find several categories. One of these key activities is related to the data pipeline processing that involves the collection, curation, transformation, integration, and quality assurance of the data. Another is related to the development of platforms and interfaces to will work as the channel of content integration in order to attract and connect users and promote their services. An additional category is about the process to monetize their idea performing tasks such as marketing, sales, customer care and management which are basics task to keep the sustainability and promote scalability of any businesses. Figure 3.12 shows the type and frequency of words that entrepreneurs described in their business plan as key activities.

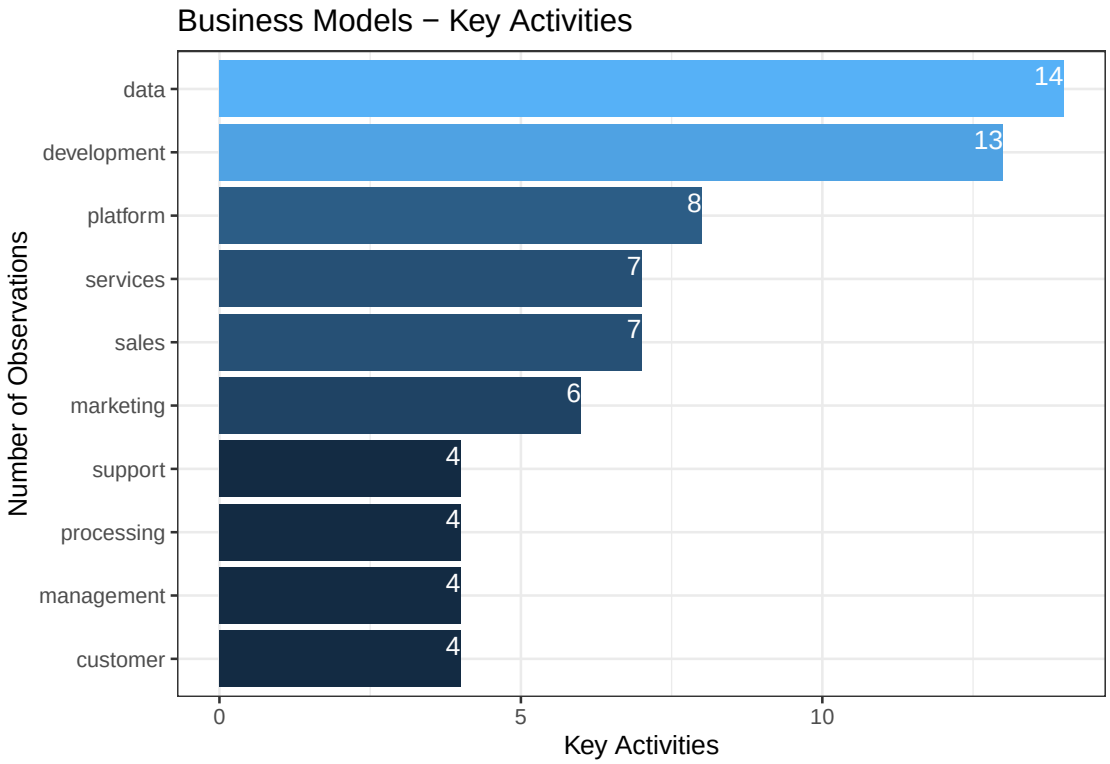


FIGURE 3.12: Displays the main words that entrepreneurs describe as key activities

Value Proposition

The value proposition refers to how companies are solving a customer's problem or need. Furthermore, the value proposition is also related to what kind of products or services companies are offering. Analysing the descriptions made by these entrepreneurs about their value proposition, we found that they are offering support and technical infrastructure implementing data science and analytics techniques (collect, clean, transform, integrate and enhance information) in order to turn data into products or services such as web platforms or mobile apps. For instance, the value proposition of one of these companies is offering consultancy services for simplification and implementation of farming operations through the use of maps and geodata services.

Another company is offering commodity price references that could help purchasing teams track efficiently commodity prices. For another company, the value proposition is to visualize big statistical data in a more friendly and concise way in order to make better use of it in their business activities. Companies that are related to the tourist sector are collecting data and producing smartphone apps that are helping tourists to find location-based content features and helping to city officials to analyze these data for economic and security reasons. Companies that are working with the health sector are offering to innovate solutions such as reducing time to process information about new therapies or deliver nutrient information to diabetics by providing a mobile solution, based on a scale and photo/nutrition app, in order to deliver measured weight to determine nutrients of meals and helping customers how

to manage their energy consumptions. Another company is describing its value proposition removing language barriers through the implementation of machine translation (MT) services across multiple languages. The MT engines help customers to translate content that is individually tailored for their style, language, and terminology. Furthermore, the engine renders secure translation for customers, without any risk of data leaks or sharing data with third parties.

An additional company that is also using open data as part of their production process and that its value proposition is offering enterprise-level solutions for managing spatial data; particularly, for support in land administration and agriculture processes. This company developed a solution called "Sentinel Hub" that brings Earth Observation (EO) data to customers' applications in the form they need, reducing the time and effort of downloading and processing the EO data.

In the real estate sector, there is a company that offers solutions based on the collection, generation, and combination of different datasets to make informed decisions concerning real estate e.g. including the search for a property, negotiations on the prices, how to handle the mortgage, moving strategies. Their value proposition is composed of performance, customization, design, price, cost reduction, risk reduction, accessibility, and usability of their platform.

There is another company providing innovative solutions also using open data in the legal sector. The value proposition that this company is offering to their customers is based on a web platform that improves the accessibility and user experience for better access to legal data sources. Moreover, the platform lets the community (citizens, businesses, legal experts, scholars and public bodies) collaborate and share their legal knowledge e.g. highlighting, tagging, commenting, sharing, open access publications.

Finally, we extracted the frequency of each word appearing in the 20 value propositions described by entrepreneurs and publicly available on the ODINE website. Figure 3.13 shows the terms that were mentioned more often (the bigger the word appears the higher the frequency).

FIGURE 3.13: Illustrates a word cloud representation the word frequency

Channels

The channel section describes how a company communicates with their customer segment and/or what is the most effective way to reaches them in order to deliver them the value proposition that they are offering. Exploring our data shows that entrepreneurs described different ways to attract them. For instance, some companies mentioned that a direct contact avoiding any kind of intermediaries that could increase the cost of their products with customers is the most effective way to create a stable and sustainable relationship. Others mentioned that they establish communication with their customers through their online platform, social media, newsletters, and chat services because it allows them to reach a broader audience.

Other companies are using their professional networks or creating videos and whitepapers. Others participating in seminars, expos, conferences and industry events as well as ads in specialized forums as a medium to keep or reach new customers. Another option described by some entrepreneurs is following the traditional channels through a marketing department

delivering customized services by customer relation manager system (CRM), phone, email, and free trials.

Customer Segment

The next block in the business model canvas is customer segments and it describes for whom companies are creating value and/or who are the most relevant customers for companies. Due to the diversity of companies and services that they are offering (even though some of them are in the same industrial sector) the customer segment varies according to their value proposition. For example, companies related to the "Professional, Scientific And Technical Activities" sector their main Customer Segments are Agricultural input and service providers, agrobusiness (seeds, chemicals), Large farms with innovation budget, drinking water companies, public authorities and financial institutions.

Companies related to the "Information And Communication" sector, their main customers' segments are professional services businesses (e.g. legal professional management consultants and industry-specific services professionals). Other companies in this sector are targeting customers in industrial companies, agrobusiness companies, Retailers, SMEs with high commodities impact, consulting & auditing firms. Other customer segment described by entrepreneurs covers research institutes, universities, hospital with research labs, drug discovery companies, and diabetics patients. Another company has its customer segment in marketing (tourist) departments of local governments in different cities in Europe. They also are targeting tourists and new residents in Europe. Another company is targeting customers such as political organisations, (local) governments, NGO's, and media. A further company offering solutions about energy consumption are targeting customers such as homeowners, contractors, energy agencies, property management agencies. A company offering translation services is pointing out customers not only in the private sector such as language service providers (LSP), multi-language vendors (MLV), localization departments at international corporations but also in the public sector (including EU governments) and to the European Commission. Companies related to the "human health and social work activities" and sectors, has specific customer segments such as patients with high medical need (e.g. oncology, rare diseases) and/or rapid development (e.g. inflammatory or neural diseases). Companies in the "real estate activities" sector also has specific target market such as users (searchers for a home) and service providers (banks, real estate brokers, architect, etc).

Cost Structure

The section of cost structure identifies what are the most important costs described in the business plan and/or which are the most expensive key resources or activities. According to the descriptions made by these entrepreneurs, we can categorize the cost structure in 3 main areas. The first one is concerning to the human resources, this investment is about hiring the qualified people that have the technical background to extract and transform data into

insights, user interfaces and platforms. Besides, hiring people that have the administrative skill to organize the business and sell their products.

The second category is related to the development of a platform. These costs are about the investment need it in building the digital product such as cloud computing, hardware (computers, internal switches, and routers, phones, etc), software (licenses in case of needed them), complementary data sets, and specific training for the technical team.

The third category is related to the fixed cost which involves rent, salaries, electricity, line phone, administrative services such as legal, accounting, etc.

Revenue Stream

The revenue stream section in the business model canvas framework refers to how entrepreneurs are describing how to monetize their business idea. This means, what kind of strategies are entrepreneurs implementing in order to incentivize customers to pay for their products or services. Examples of these revenues streams are monthly or annual subscription or a license fee, etc.

According to these entrepreneurs' descriptions, the main revenue stream is a mix of freemium and premium services in which customers start using the basic services through a freemium option and then they can switch to more advanced features in the premium service. A freemium schema is a strategy to advertise their service and attract customers, offering them a set number of features or a full availability options for a certain period of time of their products or services. The premium schema gives customer full access and supports to their services through a monthly or annual subscription, this relationship helps companies to consolidate their revenues strategy and customer loyalty.

Entrepreneurs are implementing this mixed model but targeting different customers. For instance, some companies are offering their services in the enterprise -business to business (B2B)- approach, others companies are looking to sell their products or services directly to customers -business to customers (B2C)-. An additional option described was targeting the public sector as their clients -business to government (B2G)-. Others companies have a more broad approach mixing all these segments and also offering a freemium and premium schema. These companies tend to label it as service as a software (SaaS) as their revenue strategy.

Another way how entrepreneurs monetize their idea is offering a license for specific purposes e.g. a number of calls to the API or merging different datasets. Consulting services is another type of revenue strategy in which the company offers specialized services to collect, clean, merge or extract explicit information. Figure 3.14 illustrates the number of word frequency for each revenue strategy described by these entrepreneurs.

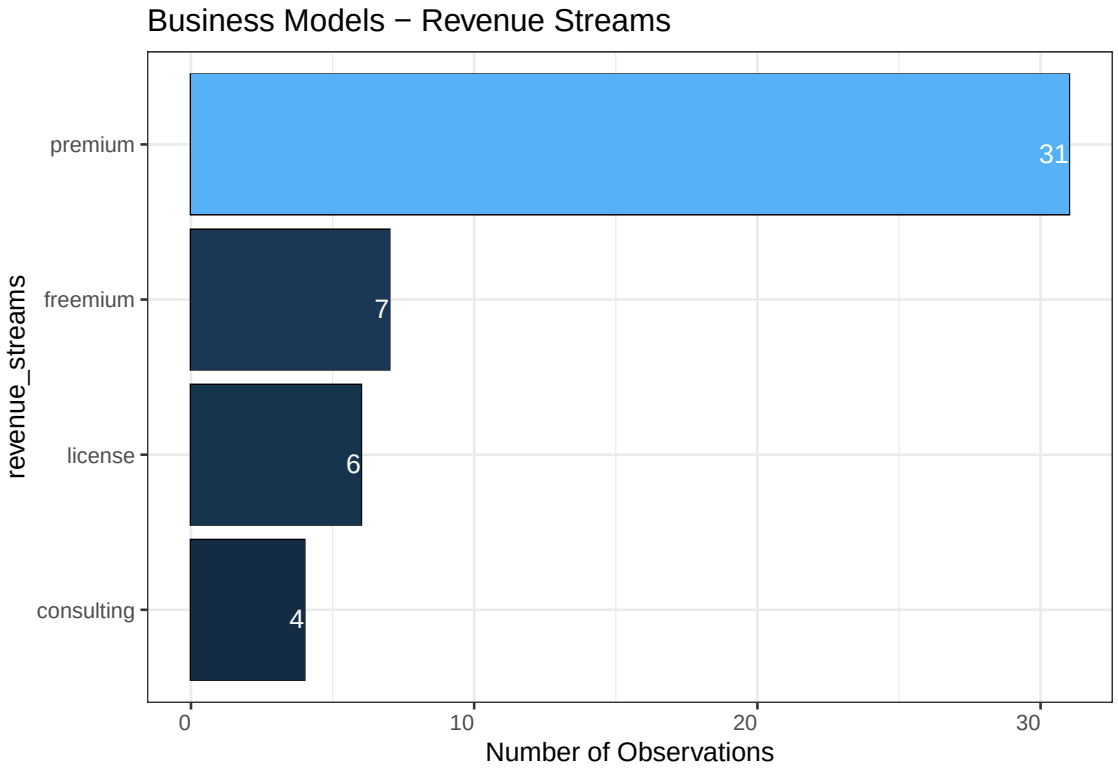


FIGURE 3.14: Represents the type of revenue streams

These are the descriptions made by the entrepreneurs of their business models of 20 of a total of 57 selected companies. The business model canvas of the others 37 companies is not publicly available due to the entrepreneurs decided to no publish it. Therefore, there is no chance to access this data and make the respective analysis. In the next section, we will analyze the data collected by ODINE regarding the companies that submitted their business proposals but that were not selected by the evaluators. We will extract the main idea, impact and team composition as a complementary analysis of our research.

3.6.5 ODINE: Business Proposals, Idea, Impact, and Team composition

In this section, we examined the rest of the companies which are more than 1,000 business proposals that applied to the ODINE programme but they were not selected to obtain the grant and the incubation process. However, these companies generated valuable business proposal to study. We analyze their applications using text mining techniques in order to extract insights from this data collected by ODINE. Mainly, we are focus understanding what is the main Idea that entrepreneurs are proposing, what is the type of Impact (economic, social or environmental) that they are looking for and how is the team composition (human capital) of these companies.

3.6.5.1 Idea: Describing the Core Idea

Entrepreneurs that are using open (government) data as an asset for their business were asked in the application process to briefly describe the core idea in their applications. Several entrepreneurs also included in this description what is the problem that they are offering to solve to their customer and how are they proposing to do it. All these descriptions together could be used as a proxy to estimate and analyze the business value proposition.

In order to perform our analysis, we first started reading the descriptions made by entrepreneurs to have a better context of their explanation. Our results show that there is a broad variety of ideas and implementations of open data. In particular, these sectors concentrate the majority of business proposals: "Information And Communication" (606 submissions), "Professional, Scientific And Technical Activities" (208 submissions), "Other Service Activities (63 submissions)", "Agriculture, Forestry, and Fishing (53 submissions)". Figure 3.15 shows the main word frequency when we apply text mining and we extract the bigrams (word pairs) to the description made by entrepreneurs about what is the core idea of their business proposals.

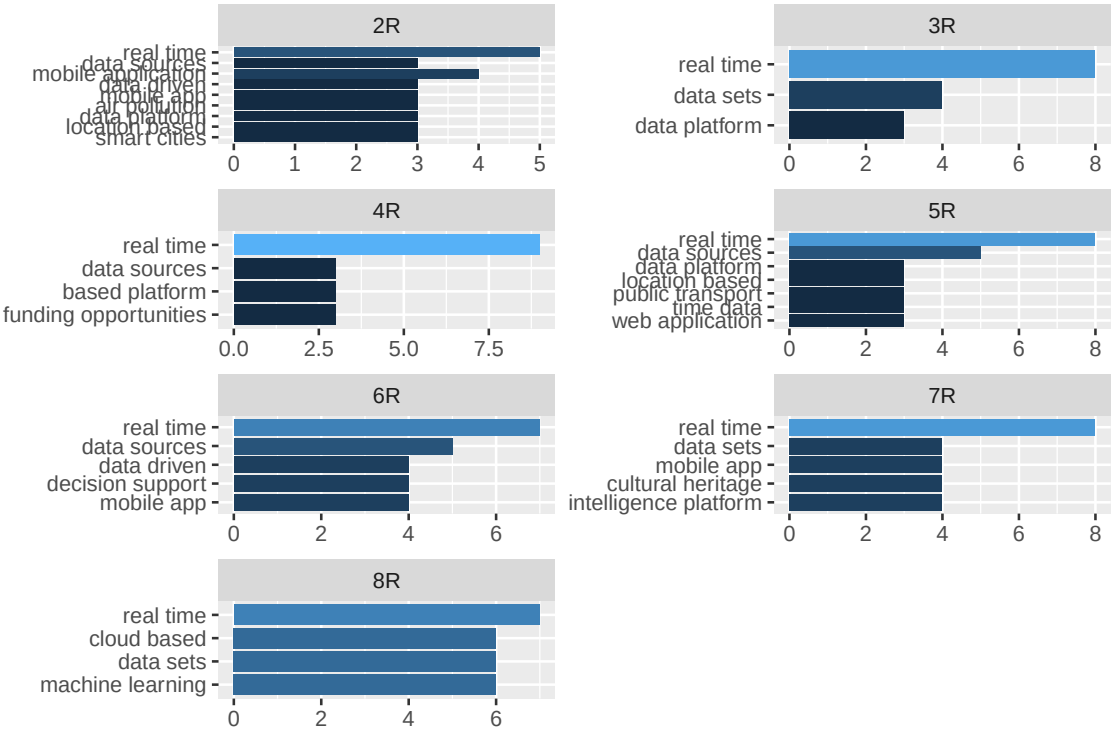


FIGURE 3.15: ODINE: Core idea (Bigrams).

Our analysis also shows a wide range of value propositions based on real-time data. For instance, an business idea based on open data from a Spain company is to provide real-time maps updates from specific regions in which traffic is a big trouble. Another company from Germany is proposing to solve the parking spaces problem recommending and updating data in real time. Other ideas are related to display in real time the level of pollution in order to

improve people's lives. Some companies from the agriculture sector also propose to update in real-time data in order to monitor and analyse the status of the crops and produce output estimations.

Another point that we found is that entrepreneurs are developing and offering solutions mainly on web and mobile platforms. For instance, a company based on Romania in the Agricultural sector is offering a solution creating a multifunctional website and mobile application for open urban planning data. Another company from Greece aimed at resolving existing key issues in european biomass supply chains by optimizing their management and decision support through mobile and web application. Another company from Germany is offering solutions through a mobile app featuring a cultural heritage museum guide based on open data about cultural artefacts, places, artists, and their connections that helps users identify, and interact around treasures of cultural heritage through knowledge graphs. Moreover, a company in Italy brings together the power of Wikipedia and mobile technologies to provide the largest free mobile audio guides collection in the world, giving tourists a new, convenient way to discover and experience cities and territories.

A company located in Poland is proposing a web application solution which aim is getting dispersed information possessed by two and a half thousand of smallest administrative public units in Poland (gminas) to make a product for business in need of information and analyses necessary for strategic and every-day decisions. In the UK a company is offering a web-app that helps businesses buy renewable energy projects in a few clicks, using satellite imagery, machine learning and a comprehensive marketplace. In addition, in Austria, a company developed a novel interactive tool for visual exploration of statistical indicators derived from open data and enterprise metrics, fully integrated into a web intelligence platform.

Other interesting insights of these data are related to the extraction, use and transformation of the open data adopting a data analytics approach and implementing machine learning techniques. For example, a company in the UK is using data analytics in agriculture to improve farm profitability, optimise resource usage and reduce environmental impact. Another company in Germany is offering solutions based on a data analytics engine that consumes available environmental open data to provide actionable insights about the urban climate. Furthermore, a company in Serbia is using data analytics on remote monitoring of patients based on real-time data collected by wearable sensors.

The implementation of machine learning algorithms is another approach used by entrepreneurs in order to transform open data into useful decision-making tools. As an example, a UK company is developing a machine learning (ML) service that allows energy providers to integrate truly intelligent features into their products, improving the efficiency of their service and promoting savings and comfort for their customers. Another company in the educational sector uses machine learning to personalise maths education at scale. This solution includes diagnostic assessment, automated feedback, progress tracking and live learning via whiteboard, messaging and audio, through which tutorial services are being delivered. Another

company in Greece is using machine learning to continuously adapt to conditions in real time and more efficiently control wind power production, with further application to biogas production.

3.6.5.2 Impact: Economic. Social or Environmental

As part of the application process, entrepreneurs were also asked to give a concrete example of the economic, environmental and/or social impact of their business proposal. The economic aspect refers to the potential benefits that entrepreneurs are offering in their business proposals some of these includes job creation, saving cost, improving processes and decision-making or increasing productivity. The social aspect of using open (government) data in their proposals is looking for impacts such as empowering society or specific less privileged groups (e.g. ethnic minorities, migrants, people with disabilities, isolated elderly people) through releasing, using or accessing information. The social impact is also focused on promoting governmental transparency, accountability and supporting culture. The environmental impact refers to applications solving problems such as reducing carbon emissions, encouraging reuse or water quality.

According to our results on extracting word pairs (bigrams) to the data collected by ODINE, we found that the economic impact was mainly mentioned by entrepreneurs, followed closely by social impact (in combination both are described in the literature such as share value in terms of open data impact) and in third place the environmental impact. Figure 3.16 shows the extraction of the main word pairs (bigrams) in the to the question stated in the ODINE application: what impact will your project have?

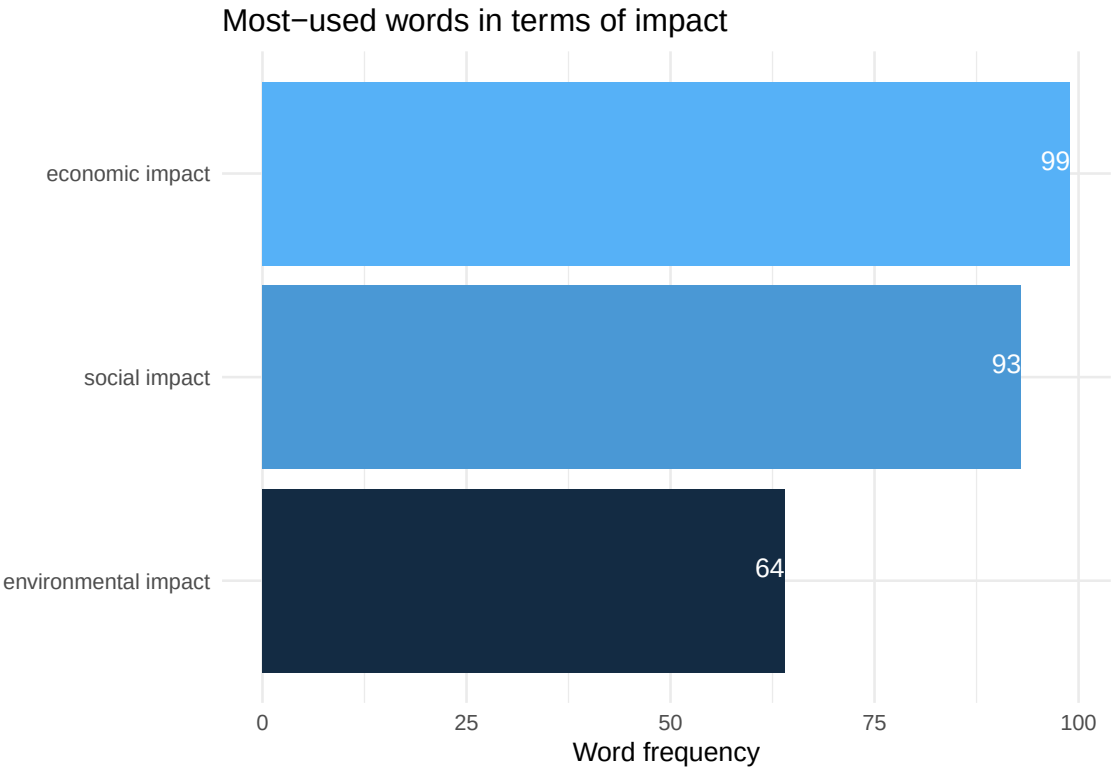


FIGURE 3.16: ODINE: Impact (Bigrams).

The impacts of open (government) data are still under research due to its recent adoption. However, open (government) data has the potential to offer an economic impact because the access to information promotes informed decisions, improves the allocation of resources and reduce the transaction costs of individuals and firms and governments. As an example of these benefits a company based on the UK argue that open (government) data can steer decision making surrounding home finding and property ownership aspirations using data analytics. They also argue that improves social mobility and personal economics by creating an open rental prices standard because of easy comparison. Others companies in Spain claims that the use of open data allows creating a bigger and more competitive ecosystem connecting bidders and claimants. Furthermore, democratizing the information access allows all type of companies (large and smalls) to access and have the same opportunities to public funds. Other company in Germany claims that using open data help them to reduce costs and promote the innovation developing open source projects, creating a synergy cycle between companies that pay for the product and support and it benefits the entire ecosystem.

The social impact using open (government) data relates to the benefits of empowering citizens and minority groups through information access, transparency and accountability. For instance, a company in the UK is using open (government) data to develop an app for automated booking of interpreters/professionals focused on deaf people. In addition, this business proposal is looking for identifies needs of service for deaf people and increase the efficiency of services for them. Another company uses data analysis techniques highlighting districts with poorer health outcomes or outlier prescribing patterns allow hidden issues to

be addressed in order to improve health services in these areas. Another company in Spain is using open (government) data in order to improve the efficiency of cultural management and better access by citizens to cultural events related to own preferences. They are using open (government) data in order to design event-policies according to citizen participation. Furthermore, they are developing better estimations of events attendance and the impact on the city. They are also focusing on cultural and linguistic minorities.

Open data can also be used in order to have an environmental impact. For instance, a company in Belgium is connecting organic food actors (farmers, suppliers, retailers etc.) food, aiming to improve organic food delivery processes and consumer satisfaction. This company developed an application powered by open (government) data to choose nearby products and delivery options. Through the process, users are informed about the carbon and ecological footprint, the nutritional value, and who gets paid what in the value chain. Another company in Spain dedicated to the agricultural sector are using open (government) data to facilitates the work of the farmers through data analytics, not only reducing costs and saving on pesticides and water but also decreasing carbon emissions to the natural environment. A company in the UK uses satellite data analysis and cleantech marketplace aiming to analyse the viability of renewable energy projects. According to this company, environmental projects have a dual impact (environmental and economic). For example, it could help Ireland increase its products and services from the marine environment sector. In addition, UK Carbon trust estimate wave & tidal alone will contribute £68 Bn to GDP. Ernst & Young[15] estimates new 1.2m direct jobs in ocean energy by 2050.

3.6.5.3 Team: Human capital

An important point for the ODINE consortium was to identify and analyze the human capital behind each submission in terms of academic background, technical and management skills, among other aspects. For this reason, the ODINE application contained a series of questions related to the team composition (see appendix A.8.3). However, due to the scope of our analysis and the anonymization of the data (this section contains the name of each entrepreneur and their members' name of the company), we only focus on this question of the application: what are the skills of each member?.

This question is relevant to analyze because these companies are related to various sectors offering diverse values propositions. Therefore, these companies are facing different technical challenges in their data pipeline process such as collecting, cleaning, transforming and extracting useful insights from data. Another big challenge is related to the processes to manage a company in terms of business models selection, scalability, market segment and so on. Hence it is important to understand what are the skills that these teams have in order to achieve these tasks. In addition, there is a gap in the literature on this topic.

[15] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52014SC0013

According to our results, we found that the team composition that applied to the ODINE program has a high-level education.Our analysis shows a high frequency of the words PhD (289 mentions) and Dr (151 mentions) as academic titles of members of the team (for practical purposes, we merged both terms as PhD). Followed by members with academic titles as Master degree (MSc) and Bachelor degree (BSc). Figure 3.17 shows the number of word frequency for each academic title.

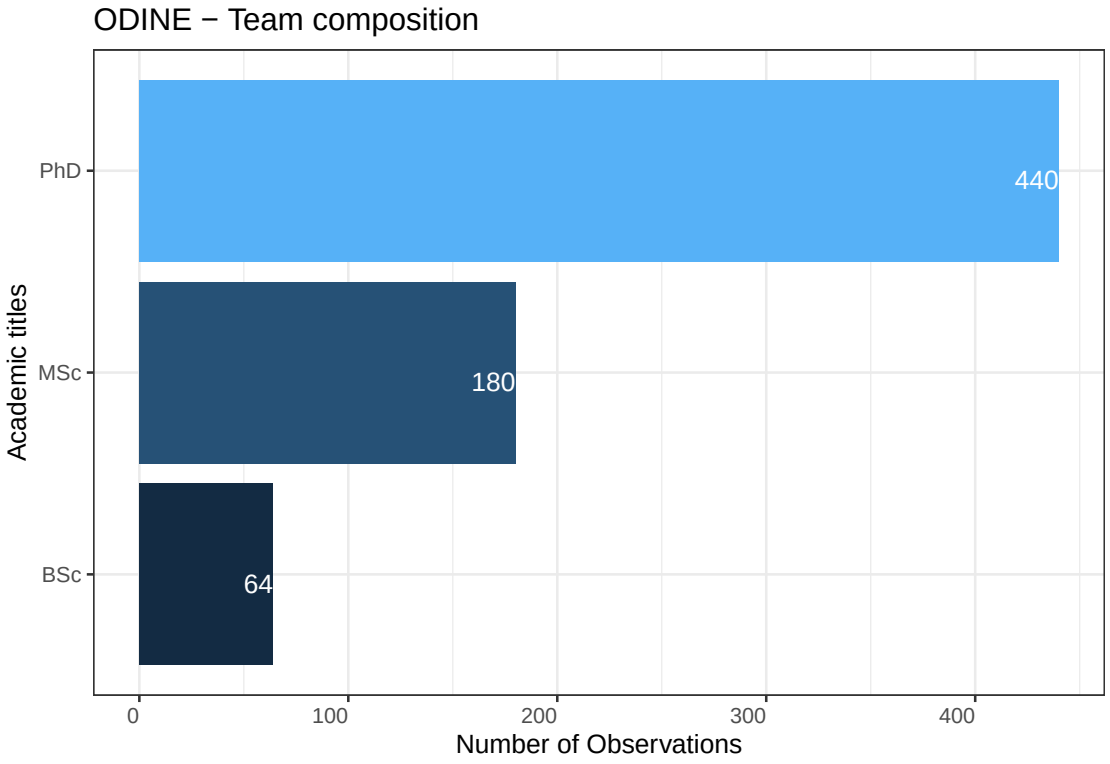


FIGURE 3.17: ODINE: Team composition

As we can see, the human capital of these companies is composed of a high academic degree. Extracting information about what are the areas in which these entrepreneurs finished their PhD, we found that computer science appears several times. The term artificial intelligence also appears as a recurrent field as well as areas such as natural language processing, web personalisation and innovation strategy. It is important to notice that no all entrepreneurs described their academic title; however, these numbers give us a good idea of their academic background. A list of these academic titles can be found in the appendix section A.9.13.1

Now, when we focus our analysis on their skills in order to collect, transform and extract insights from data and convert them into a business idea, we found that the most frequent bigrams are also related to their technical skills in areas such as computer science, project management, software development, data science and machine learning. An explanation of the frequency of these skills could be that the use of open data as an asset is a relatively new raw material to develop products and services and these are some of the skills required to

develop their business catalogs. Figure 3.18 illustrates the most frequent bigrams mentioned by entrepreneurs.

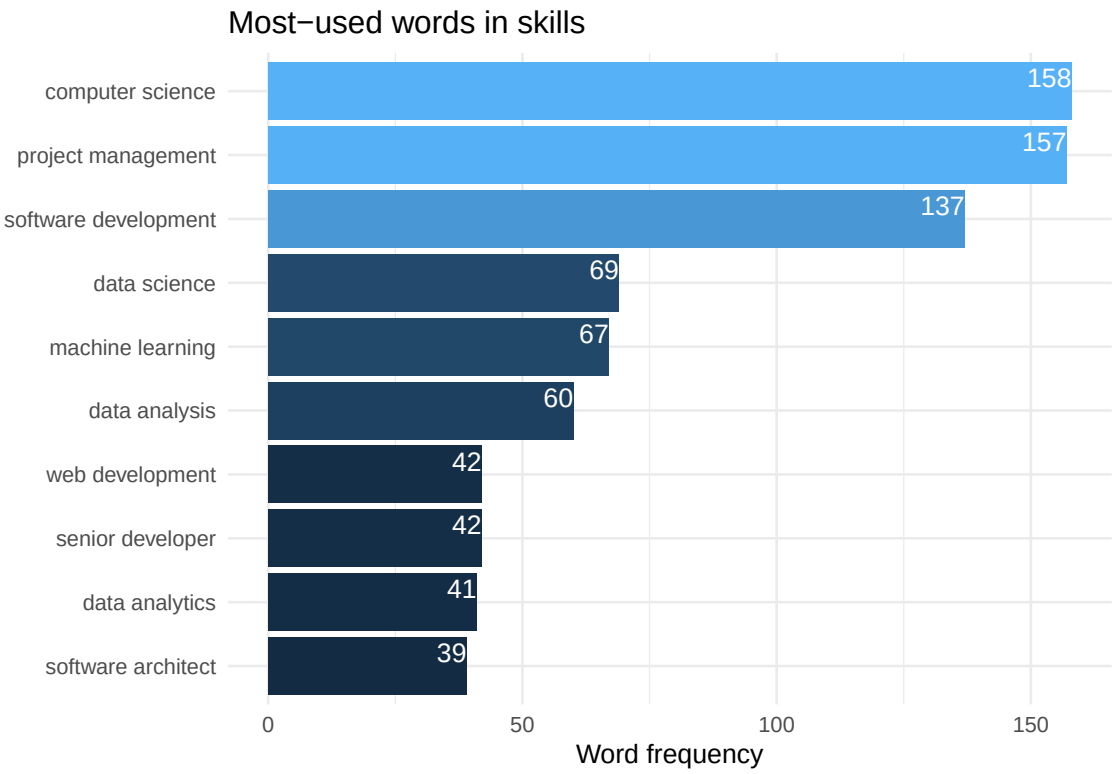


FIGURE 3.18: ODINE: Team composition skills (Bigrams).

Regarding the management aspects, the number of applications shows that The Chief Executive Officer (CEO) appears as the main role in charge of the business proposal. This illustrates that these entrepreneurs have an interdisciplinary background sharing technical and management skills. This is also consistent with the literature that argues that being an entrepreneur requires a combination of skills (technical, management, finance, networking). The second term that appears in order of frequency mentioned by entrepreneurs was the role of the Chief Technology Officer (CTO) who is the person in charge of scientific and technical issues within an organization. The reason of this is because as we described earlier, the strong influence of the technical background and the skills need it in order to transform Open (government) data as raw material into a business proposal. Furthermore, most of the applications received were companies in the information and communication and professional, scientific and technical Activities sectors that according to reference and management of nomenclatures (RAMON) in Europe these sectors are performing activities such as computer programming, consultancy and Information service activities. The next position that appears in the management combination is the Chief Operating Officer (COO) who is responsible for the everyday operation of a firm. The fourth position is occupied by the Chief Financial Officer (CFO) who is in charge of keeping the strategies and financial records of the company healthy and sustainable. The fifth position is for the Chief Marketing Officer (CMO) who is the position responsible for the design and implement the marketing, sales,

distribution and customer service strategy. Figure 3.19 shows the management composition according to the word frequency in the business application process.

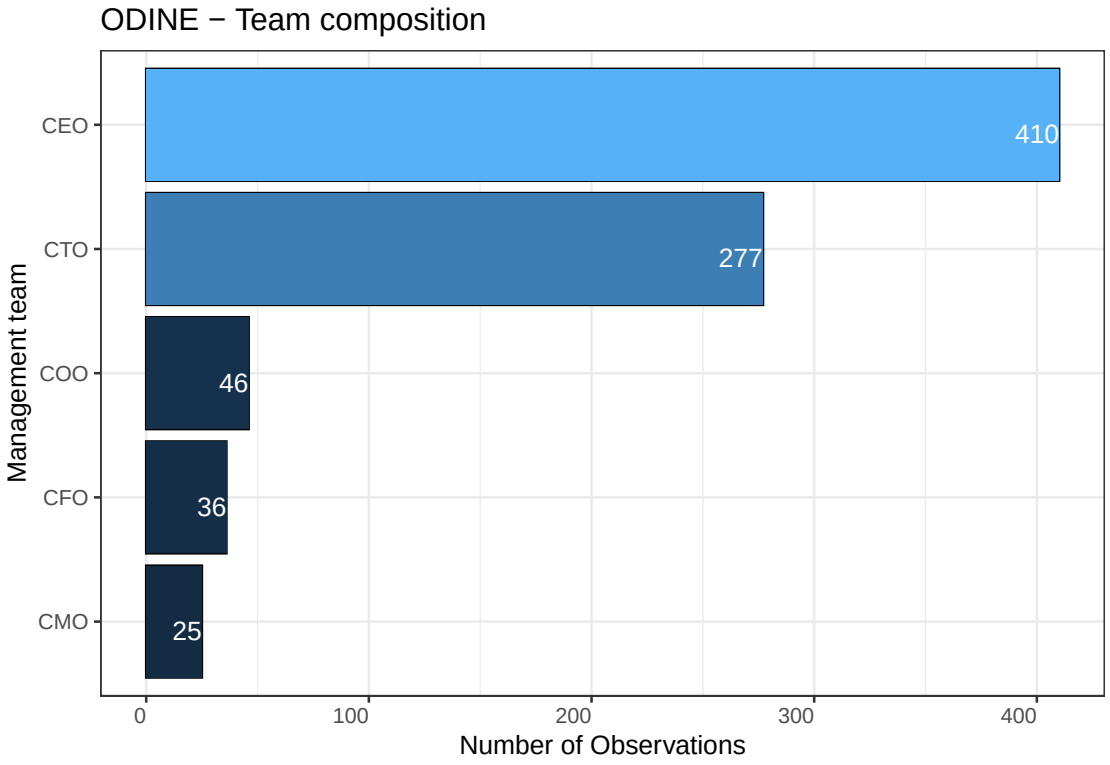


FIGURE 3.19: ODINE: Team management composition

After having analysed the core idea proposed by entrepreneurs, the type of impact (social-economic-environment) that they are describing and the team composition of these companies, we will move to the conclusion section in which we describe the main findings of this chapter.

3.7 Conclusions

The aim of this chapter is to analyze the perception and business opportunities generated by the release of open (government) data and understand how entrepreneurs are transforming open data into business proposals in Europe. The analysis of this research is using information collected by the Open Data Incubator for Europe (ODINE) which contains more than 1000 business proposals submitted by entrepreneurs using open data as part of their production process. The methodology implemented in this research is based on a descriptive text mining approach because these business proposals are composed of text descriptions about the company profile, structure of the business model adopted, the core idea about how they are using open (government) data as part of their business proposals, description of the economic, social or environmental impact implementing open (government) data and team compositions. All these textual descriptions are considered in the data science field as unstructured data.

The sequence of our analysis is structured as follows. First, we explore the companies profile to extract the country origin, economic sector, and the year in which the companies were created. Second, we analyze 20 of 57 companies that were selected as successful for the grant and incubation process by the ODINE consortium (we only analyse these 20 companies due to this is the only information publicly available). We split the analysis of these 20 companies by first selecting 5 uses cases in order to produce a complete context about their business models using the canvas framework describing their key partners and activities, the value propositions that are they offering, customer relationships, channels, segments, cost structure and revenue streams. Then, we analyze the other 15 companies through systematically implementing a text mining approach in order to find patterns in their business model canvas framework. Third, we study the business ideas generated by entrepreneurs that are using open data as part of their production process. Later, we analyze the economic, social or environmental impact that entrepreneurs are defining as their contribution to society. Finally, we analyse the team composition stated by the entrepreneurs that applied to ODINE.

These are the main findings of this research. The use of open (government) data is considered by these entrepreneurs as a business opportunity to enhance or produce new products or services and the chance to implement innovative business models. According to the ODINE submissions, this adoption of open (government) data as a digital asset by entrepreneurs is more perceptible in Southern and Northern regions (both areas generated 684 business proposals) than Western and Eastern of Europe (these areas generated 339 business proposals). Furthermore, the implementation of open (government) data by entrepreneurs is mainly in high-income countries (978 business proposals) rather than upper middle and lower income (46 and 1 respectively).

At the country level, the UK is the country with the higher number of business proposals submitted (241), followed by Spain (146), Germany (120), Italy (96) and Greece (53) which according to the World Bank these economies are considered as high-income. This result

agrees with our previous model measuring the relationship between entrepreneurship and open (government) data at the country level in which high-income economies show a positive and statistical significance relationship. Some of the determinants of this relationship could be explained because of the technological development, educational level, infrastructure and public policies adopted for these economies.

Another important finding is that most of the companies that submitted business proposals to the ODINE program are related to the technology sector. For instance, 656 entrepreneurs stated that they companies belong to the "Information and Communication" sector (companies that belong to this sector are related to the production, processing and distribution of data, communication, information technology and other information services activities) described by statistical classification of economic activities in the European Community using the Reference And Management Of Nomenclatures (RAMON). The second most mentioned sector by entrepreneurs is "Professional Scientific and Technical Activities" (220 business proposals) which is also related to the development of technological activities. The third sector is the "Other Services and Activities" (which in its description includes Repair services of computers and communication equipment) and Agriculture, Forestry and Fishing (70 and 52 business ideas submitted respectively). Other sectors mentioned in the business proposals are Education (26), Health (15), Finance (11), Real Estate (7), Entertainment (6) and Transportation (6). An important point of these companies associated with these sectors is that they are developing products and services base on data such as analytics, visualizations, and predictions and they are mainly related to the IT domain.

Regarding the companies profile, we find interesting points. One of them is that companies are relatively young (although not all companies provide this information because it was not explicitly requested in the ODINE's application form). Our results show that 154 entrepreneurs stated that they created their companies between 2011 and 2017. This could be associated with the fact of more business digitalization, technological evolution and that these are data-driven companies (IT domain) implementing solutions in diverse sectors. Another potential aspect is the release of government data has been growing because of its symbiotic relationship to the open government partnership movement (an initiative that started in 2010 and nowadays involves 75 economies) which promotes the development and implementation of policies that encourage the release of open (government) data in order to promote transparency, accountability and empower citizens looking for a shared value (social and economic). An economic and positive externality of these policies is the creation of new products, services and innovative business models based on the release of this public good.

Another result that we found applying text mining to the companies profile available through the ODINE application process is that these data-driven companies are mainly implementing technological methods based on data science, software development, cloud computing, linked data, semantic web, knowledge management, and information systems in order to extract and transform open data into products and services. This is an important distinction in terms of business approach and strategy because of the challenges of these technological process

involves and the set of knowledge and skills required to solve it. These challenges also could imply that the adoption of open (government) data for commercial purposes brings some barriers in terms of digital literacy and there is not extensive research about the skills and knowledge required to transform open (government) data into a product or services. Furthermore, open data by itself could be worthless if nobody is using this public good. One of the commercial values of open (government) data resides in its extraction, transformation, and combination with other datasets because these elements allow the creation of unique products.

As stated in the literature, the comparison and analysis of business models is a complex task because of the diversity of products and services that each company offer. For this reason, we adopt a uses cases approach describing 5 companies that are using open (government) data as an asset and part of their production process. We explore the business models proposed by these companies using the canvas framework in order to illustrate all their elements (key partners and activities, the value propositions, customer relationships, channels, segments, cost structure, and revenue streams). Furthermore, we expand our descriptive analysis but in this case using text mining due to the nature of the (unstructured) data to other 15 companies in order to find patterns in their answers.

Our results show that the key partners of these companies are mainly the private sector (composed of companies, consultancies and manufactures), followed by the public sector (conformed of governmental entities at the federal or regional level and public agencies across the European Union) and academic sector (research centers, universities or institutions that collaborate with entrepreneurs). The key activities of these companies are divided in the technical aspect such as the data pipeline processing that involves tasks such as collection, curation, transformation, integration and quality assurance. Besides, the development of their platforms are mainly based on web and/or mobile technologies. On the other hand, the key activities related to administrative tasks are the design of marketing, sales, customer care and management strategies. The value proposition refers to what kind of products or services the companies are offering in order to solve customer needs. This component is considered as the main element in terms of differentiation among companies and due to each company is offering a different solution, in different economic sectors. However, we found as a common factor of these companies is that they are supporting their value proposition implementing software engineering methodologies and cloud computing resources such as infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS). Moreover, these companies are also implementing data science techniques (collecting, cleaning, transforming, integrating and enhancing information using machine learning algorithms) in order to turn data into products and services.

There are several channels used by entrepreneurs to communicate with their customers. One of these is to have direct contact in terms of sales avoiding any kind of intermediation in order to offer better prices. Another is to use traditional tools such as customer relation manager systems (CRM), phone, email, and free trials. Other entrepreneurs prefer to disseminate their

products and services using online platforms, newsletters, social media and chat services. Another way to communicate with their customers is developing professional networks or creating videos and/or whitepapers. An alternative option described by other entrepreneurs is participating in seminars, expos, conferences, and industry events.

The customer segment refers for whom companies are creating value and/or who are the most relevant customers for companies. As stated before, due to the diversity of companies and services that they are offering (even though some of them are in the same industrial sector) the customer Segment varies according to the value proposition. For example, some companies related to the "Professional, Scientific And Technical Activities" sector their main customer segments are Agricultural input and service providers, agrobusiness (seeds, chemicals), large farms with innovation budget, drinking water companies, public authorities, and financial Institutions.

The cost structure is the section in which entrepreneurs describe the costs associated with their investment in order to produce their product and/or services. According to the descriptions made by the entrepreneurs, we can categorize the cost structure in 3 main areas. The first one is concerning to the human resources, this investment is about hiring the qualified people that have the technical background to extract and transform data into insights, user interfaces, and platforms. Besides, hiring people that have the administrative skill to organize the business and sell their products. The second category is related to the development of the platform. These costs are associated with the investment required to build the products or services such as hardware (computers, the internal switches, and routers, phones, etc.) software (cloud computing services, licenses in case of needed), complementary data sets (in case they need it), and specific training for the technical team. The third category is related to the fixed costs such as rent, salaries, electricity, line phone, administrative services such as legal, accounting, etc.

The revenue stream refers to how entrepreneurs are describing to monetize their business idea. This is an important distinction to mention because very often in the literature the term of the business model is only referred as the way a company generates revenues without taking into account the other components of the business model. According to the descriptions made by the entrepreneurs in the business model canvas, the main revenue stream is a mix of freemium and premium services in which customers start using the basic services through a freemium option and then they can switch to more advanced features in the premium service. A freemium schema is a strategy to advertise their service and attract customers, offering them a set of features or a full availability options for a certain period of time of their products or services. The premium plan gives to customer full access and supports to their services through a monthly or annual subscription, this schema helps companies to consolidate their revenues strategy and customer loyalty. Furthermore, some entrepreneurs are implementing this mixed model but targeting different customers. For instance, some companies are offering their services through an enterprise segment -business to business (B2B). Others are looking to sell their products or services directly to customers -business

to customers (B2C)-. An additional option described by some entrepreneurs were targeting the public sector as their clients -business to government (B2G)-. Others companies have a more broad approach mixing all these segments and also offering a freemium and premium schema. These companies tend to label it as Service as a Software (SaaS) as their revenue strategy. Another way how entrepreneurs monetize their idea is offering a license for specific purposes e.g. a number of calls to the API or merging different datasets. Consulting services is another type of revenue strategy in which the company offers specialized services to collect, clean, merge or extract explicit information.

Although only 57 companies were selected by the ODINE consortium to access their fund (for up to €100,000) and continue the incubation process, this project collected valuable information of more than 1,000 business proposals based on open (government) data in which entrepreneurs explained different aspects of their propositions. One of these aspects was to describe the core idea of their product or service in which entrepreneurs explained how they are using open (government) data as a digital asset in their business proposals. Our results show a wide range of business proposals mainly emanated from the "Information and Communication" and "Professional Scientific and Technical Activities" sectors (606 and 208 submissions respectively) in which entrepreneurs are offering technological solutions using open data into diverse domains such as agriculture, education, health, finance, real estate, transportation. Furthermore, entrepreneurs are describing in their proposals that they are developing their digital solutions mainly on web and mobile platforms. The implementation of machine learning algorithms is another common approach described by entrepreneurs in order to transform open (government) data into useful decision-making tools.

Concerning the economic, social or environmental impact described by entrepreneurs when they are using open (government) data. Our analysis shows that the economic impact was mainly commented by entrepreneurs, followed closely by social impact and in third place the environmental impact. The economic aspect refers to the benefits that entrepreneurs are offering in their business proposals such as job creation, save cost, improve decision-making, innovation, or business opportunities. This means, open (government) data allows access to information and it promotes informed decision, improves the allocation of resources and reduces the cost of households, firms, and governments. The social aspect of using open data in their proposals is related to empowering society, or specific less privileged groups of it (e.g. ethnic minorities, migrants, people with disabilities, isolated elderly people). The social impact is also focused on promoting governmental transparency, accountability and supporting culture. The environmental impact refers to applications solving problems such as reducing carbon emissions, encouraging reuse or water quality.

Finally, the last element that we analyzed was the human capital composition of these companies. Because of the scope of our analysis and the anonymization of the data, we focus our analysis extracting the entrepreneurs' descriptions of what are the skills of each member. We found that the team composition of these companies tends to have a high-level education profile. Our analysis shows a high frequency of the words PhD/Dr (440 mentions)

as academic titles of members of the team. Followed by other academic titles such as Master degrees (MSc, 180 mentions) and Bachelor degree (BSc, 64 mentions). As we can see, the human capital of these companies is composed of a high academic degree. It is important to notice that no all entrepreneurs described their academic title; however, these numbers give us a good idea of their academic background.

Extracting information about what are the specialization areas of these entrepreneurs, we found that computer science appears several times. The term "artificial intelligence" also appears as a recurrent field as well as areas such as natural language processing, web personalisation, and innovation strategy. Moreover, our analysis also reveals that the most frequent bigrams are also related to their technical skills in areas such as project management, software development, data science, and machine learning. Regarding their description of management titles and skills, the number of applications shows that the term Chief Executive Officer -CEO- appears (410 mentions) as the main role in charge of the business proposal. This suggests these entrepreneurs should have an interdisciplinary understanding due to they are implementing technical and management skills in their proposals. The second term that appears in order of frequency mentioned by entrepreneurs was the role of the Chief Technology Officer -CTO- (277 mentions). The reason for this is because as we described earlier, the strong influence of the technical background and the skills need it in order to transform open data as raw material into a business proposal. Other management titles descriptions include Chief Operating Officer -COO- (46 mentions) Chief Financial Officer -CFO- (36 mentions) and Chief Marketing Officer -CMO- (26 mentions).

Chapter 4

Measuring Risks and Challenges Using Open Data

4.1 Abstract

Government policies and progress in technology are allowing the increasing adoption of open data as part of the business processes in enterprises. The purpose of this research is to analyse what are the risks and challenges that companies are facing when they are using open (government) data. Previous studies have characterised these risks by conducting interviews with stakeholders that are mainly in the and academic public sectors. In this chapter, we consider the perspective of the entrepreneurs that are using open (government) data as part of a business proposition by analysing risks assessment information provided by 1173 applicants to Open Data Incubator for Europe (ODINE) program, applying machine learning algorithms to identify and cluster the risks, yielding a list of risks/challenges that are ranked according to its frequency and compared across countries in Europe.

Keywords: Open Data, Entrepreneurship, Risks and Challenges.

.

4.2 Introduction

Recently, governments are implementing open data policies that allow the use and reuse of datasets from different agencies by members of the public and companies, in order to promote transparency, accountability and create social and economic value. The economic impact of open (government) data has been estimated by the McKinsey Global Institute (Manyika et al., 2013) to be at least $3 trillion a year globally across several sectors such as finance, health, energy, retail, transportation and education. (Cappgiemini Consulting, 2013) estimated in €32 billion the impact of open (government) data on the European Union. This firm also reports that the private sector benefits from open data thanks to the creation of new business opportunities and better decision making based on the dissemination of governmental data. Finally, (Stott, 2014) lists economic growth and business and job creation as economic benefits related to the adoption of open (government) data.

A critical cornerstone to create economic value is to stimulate the creation of products and services based on open data. An example of this stimulus is the open data incubator for Europe (ODINE)[1], a program funded by the European Commission aiming at incubating companies with open (government) data centred business ideas. However, when centring a business idea on open (government) data, companies need to assess the risks concerning its provision and use that could affect their business processes such as data availability, accessibility, usability, or quality,. Policy makers and data publishers need to be aware of these risks to provide policies that facilitate the task of creating value for companies.

Previous research has been conducted about the barriers and limitations that governments are facing when they are adopting and implementing an open data policy. According to the literature, the spectrum of these barriers covers cultural (reluctant to share data, fear to expose data, appropriation of the information), legal (licensing, privacy, use of data), economic (some governments are not opening their datasets related to such as transportation, geolocation, tenders), technical (interoperability, different formats) and language (dataset just in one language. For instance, Spanish, English, Germany) impediments. The methodology most frequently reported in the literature to determine those risks is conducting workshops and interviews with academics and government authorities. However, there is scarce information about the impediments that companies deal with when they are using open (government) data as raw material to innovate or create products and services. These issues lead to our research questions:

1. Are the risks considered by companies using open (government) data the same as the ones described in the literature?

2. Which are the most critical risks and challenges for companies using open (government) data?

[1] `https://opendataincubator.eu/`

3. Are there any differences in the perceived risks depending on the country of origin of the companies?

Our contribution is the uncovering of an entrepreneurial perspective about what are the limitations that end-users face when using open (government) data. We developed a quantitative approach to identify and rank these barriers according to its frequency of occurrences mentioned by businessmen. Policy makers and data publishers need to be aware of these risks to provide policies that facilitate the task of creating value for companies and promote a virtuous economic cycle that involves innovation, job creation, and growth.

This chapter is organised as follows: Section 1 explains the aim, justification, and methodology used in this investigation. Section 2 reviews previous works on the barriers, limitations, risks and challenges described by data stakeholders when they are using open (government) data, and analyses their data collection methodology. In Section 3, we describe our methodology, data and model. The results and discussion of our analysis are presented in section 4. Finally, Section 5 draws conclusions, outlines key findings and discusses future work.

4.3 Background and Literature

This section provides a literature review of how different stakeholders such as academics, public servants, and entrepreneurs are perceiving risks and challenges in the adoption and use of open (government) data.

4.3.1 Risks and impediments from academic and public servants perspective

The release and implementation of open data present several risks, challenges, and barriers that have been studied by several authors. (Anneke Zuiderwijk et al., 2012a) propose the following list of socio-technological impediments availability, accessibility, usability, quality, metadata, interaction with the data provider, and opening and uploading. The methodology used to identify these impediments was through interviews with 6 key actors selected based on their academic background and open data experience, 4 workshops at international events with data stakeholders that are involved in government and academia sectors, and a literature analysis. They suggest 3 main groups of data impediments: access (related to creating, opening, finding and obtaining data), use (related to restrictions on data use) and deposition (associated with difficult to store, discuss and provide feedback).

(Conradie & Choenni, 2012) conduct their research at the municipal level in the Netherlands, finding several barriers related to the legal and technical frameworks, namely licensing, ownership of data, lack of policy and priority releasing data, privacy, use of data, data sources, data storage, and sustainability of data for release. They reached these results through interviews, workshops, questionnaires, and desk research. The stakeholders involved in this research were mainly in the public sector.

(Janssen, Charalabidis, & Zuiderwijk, 2012) describe and categorise several barriers related to the adoption of open data at the institutional level, these include the complexity of handling of data, use and participation, legal regulations, and data quality.

(Janssen et al., 2012) describe and categorize several barriers related to the adoption of open (government) data at the institutional level, these include the complexity of handling of data in terms of volume or formats, participation in the process to release these datasets due to the legal regulations, and data quality. Moreover, the authors argue that these institutional barriers have other negatives externalities that affect either other public entities, civil servants or end users that need to have access to this information. The data collection process of this research was made through interviews and group sessions which involved civil servants from different organizations at the federal and municipal level.

(Martin, Foulonneau, Turki, & Ihadjadene, 2013) developed a topology of challenges, risks, limitations, and barriers associated with the implementation of open data in the public sector. Authors suggest developing an open data initiative adopting a framework that considers

risks, contingency actions and expected opportunities due to the complexity that this task involves. For example, political will, stakeholders at different level in the public sector, technical infrastructure, economic issues, accessibility, data literacy and skills, licenses and legal regulations involve. The research was carried out analyzing 3 different open data initiatives and platforms developed on distinct governance levels (municipal and national) in France, Germany, and the UK.

(Barry & Bannister, 2014) develop a taxonomy of barriers related to the adoption of an open data initiative in Irish central and local government. The methodology used to the data collection in this research was through semi-structured interviews and study of internal documents. Authors describe in their investigation a series of barriers from different angles. For instance, economic barriers are related to the resources constraints in order to fund this initiative due to there are a cost and effort associated with the release of open (government) data. The legal barriers involve legislation and licensing the data. The cultural barriers are associated with the control and power of diverse stakeholders in terms of data sharing. The technical barriers include the use of legacy systems that constrains the interoperability and data publishing process. Furthermore, the lack of human capital to deal with these issues or other process such as data anonymization or manage dynamic data.

4.3.2 Risks and impediments from an entrepreneurs perspective

According to (Kitsios, Papachristos, & Kamariotou, 2017) governments are facing important challenges when they are collecting and releasing open data. One of these difficulties is generating interest in using open data in order to create a virtuous open data ecosystem, the challenge is based on the diversity of actors involved in it. In particular, developing an ecosystem for economic growth through actors that perceive open (government) data as a digital asset in order to use for the development of new products or services. This research was developed by implementing a qualitative methodology interviewing 6 actors understanding how this open data ecosystem is creating new business opportunities. The author agrees and confirms the main risks stated in the literature and provided by academics and public servants stated above. However, the author includes the technical skills required to transform data as an entrepreneurial idea as a barrier, the lack of readiness of some data sources, data quality, and the complexity to access regional data.

(Open Data Institute, 2015) conducted a study based on surveys, interviews and desk research that involves 270 companies that are using, producing or investing in open data. In their survey, participants are asked to rate from 1 to 5 (where 1 means little influence and 5 means great influence), eight data attributes that could have an impact on its adoption. They got 74 answers, reporting the percentage of companies that assigned 5 to each risk as follows: licensing of dataset 55%, provenance of data 42%, accuracy of data 39%, ease of access to datasets 35%, timeliness of data 31%, format of data 12%, applying documentation 5%, help and support from publishers 3%.

(Walker, Simperl, Capgemini, agent. European Data Portal:European Data Portal, & corporate-body. PUBL:Publications Office, 2020) raise several benefits when entrepreneurs are using open (government) data in order to create or improve business or products. However, the authors also claim that entrepreneurs are facing several challenges in adopting open (government) data as a digital asset. One of these challenges is the lack of openness. This means that entrepreneurs have identified valuable data that could be incorporated in their business process but this information is not publicly available or under a license of use. Another challenge is that the data is involved in a process or legal agreement between large companies and governmental entities that limit its consumption partially or totally. Authors also stated that entrepreneurs are facing the lack of networking or communication channels with data publishers and this situation is having an impact in terms of engagement. Besides, company owners declared issues related to open data reuse culture. Other challenges proposed by the authors and that are in syntony with previous research are lack of data standardization, accessibility, platforms, discoverability, use, quality, and licensing.

Open data advocates claim that open data encourages entrepreneurship and innovation; nevertheless, all the risks and challenges stated above are limiting its spread and adoption. There are several conclusions of this literature review: first, most scholars follow a qualitative methodology based on interviews that is strongly biased towards the public sector. Second, most scholars suggest that the risks associated with the provision of open data are related with a complex combination of economic, legal, cultural and technological factors that could limit the benefits of open data. In addition, the literature that seeks to analyse the impact of open data on entrepreneurship is still scarce, and many basic questions of the economic impact of open data, to the best of my knowledge, are unanswered.

4.4 Data

In order to solve our research questions, we collect and analyze data from two different sources. The former is the list of risk, challenges, barriers, and impediments described by stakeholders (academics and government authorities). This data is collected through workshops and interviews, that we reported in the literature review section. An extracted and detailed list of these risks are presented in the Appendix section 1.10.1-Risks, barriers and impediments described in the Literature Review.

The second source is data collected by the open data incubator for Europe (ODINE)[2], a programme funded by the European Commission (H2020)[3] aiming at fostering and supporting the next generation of digital businesses that are using or producing open data at the core of a business idea. After a dissemination process consisting of 8 open calls from May 2015 to August 2016, the ODINE consortium received 1173 business proposals. However, we

[2] https://opendataincubator.eu/about/
[3] https://ec.europa.eu/programmes/horizon2020/

proceeded to carry out an exclusion process due to incomplete answers (some entrepreneurs did not answer to the question provided in the application "What risks/challenges in using open data in the context of your product/service you envision?"). We also exclude duplicate submissions (some companies applied two or three times during the open calls). Our final dataset is composed of 989 applications submitted by entrepreneurs from different countries and industrial sectors across Europe.

4.5 Methodology

We apply a text mining approach to extract the perception of risks declared by participants in the survey described in the two data sources stated above. The pipeline process to collect and extract the risks declared by the academics and government authorities that participated in the workshops and interviews mentioned in the literature review section follows the next procedure. First, we perform a systematic search for all documents in indexed journals related to the risks, barriers, challenges, impediments, in using open data. For this purpose, we develop a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol. Our selection criteria were searching in the title words that match with "open data" OR "open government data". We also select matching the strings "open data" OR "open government data" in the topic section of an academic paper. Later, we search for the terms "risks" & "barriers" & "challenges" & "impediments" in the topic section. After that, we included the terms "quantitative" OR "qualitative" in the search in order to identify which type of methodology was used by the researchers. The time period is from 1970 to 2020 and the search indexes are SCI-EXPANDED, SSCI. Our results display 721 academic papers. However, we exclude academic papers that are not in the English language. Therefore, our final result shows 681 academic papers. Figure 4.1 illustrates the identification, eligibility, inclusion, and exclusion of elements that are part of our protocol.

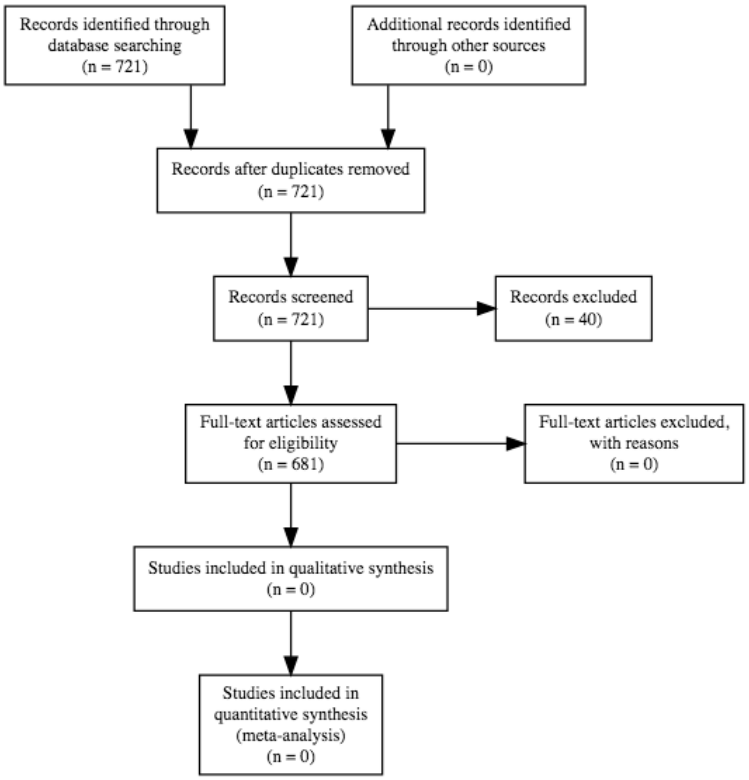


FIGURE 4.1: illustrates the PRISMA protocol.

The next step in our research is extracting and categorization a list of risks, barriers or impediments according to the description made by the papers' authors (this list is presented in appendix 1.8) implementing text mining libraries, using the R open source software (Feinerer et al., 2015; Hornik, 2017; Silge & Robinson, 2016). We clean, extract and group the most common words stated by the stakeholders involved in these academic papers.

Regarding the ODINE data, we merged all applications submitted to the incubation program into a small database and we analyzed the answers provided by the entrepreneurs (Appendix A.10.2) to the following question:

What risks/ challenges in using open data in the context of your product/service you envision?

Then, we identify common risks among the answers through implementing a text mining approach by transforming unstructured data into a bag of words and extracting insights from them (Zhang, Jin, & Zhou, 2010) and (Blei, 2012). By implementing this approach, it is possible to estimate how documents, as well as specific terms, are similar according to its inferred variables (Everitt, 1984). In what follows, we describe the feature extraction process to the two datasets (literature review and ODINE), implemented using the open source software R:

1. We collect data from academic databases such as Scopus, and Web of knowledge in order to extract the risks, challenges, and barriers stated by academic and government authorities in academic papers. The next step is to merge the data collected by the ODINE consortium in which entrepreneurs answered to the question stated in the application form *"what risks/challenges in using Open Data in the context of your product/service you envision?"*

2. The next step is to transform unstructured data (text) into a bag of words in order to perform the preprocessing (data cleaning). Preprocessing involves standardize to lower case, removing punctuation, special characters, numbers, stop words, connectors, articles. The bag of words approach allows us to filter terms with a specific number of occurrences (Griffiths & Steyvers, 2004) and to compute the well-known similarity measure such as tf-idf (term-frequency inverse document frequency) (Srivastava & Sahami, 2009).

3. With the similarity values, we cluster the answers by the occurrence of risks and categorized them by country. Finally, we use this cluster as a measure for ranking the more common words frequency in terms of risks and challenges.

4. To have a better insight of how the term is used within the answers (context and semantics), we read them and highlight common patterns. Figure 4.2 illustrates the whole pipeline process that we followed.

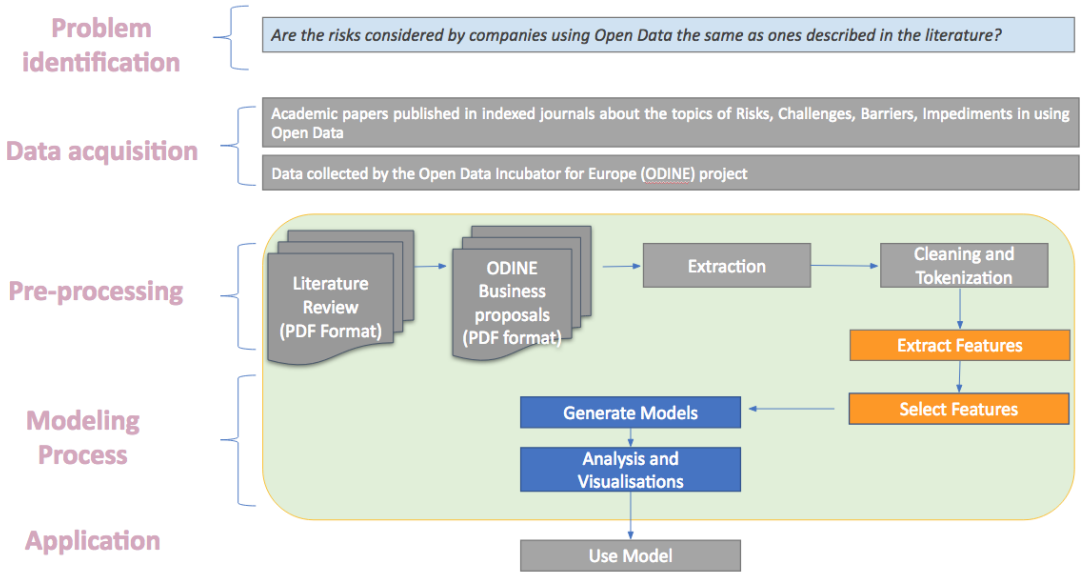


FIGURE 4.2: Risk analysis: Research workflow.

4.6 Results and Discussion

Regarding our two research questions: *"Are the risks considered by companies using open (government) data the same as the ones described in the literature?"* and *"Which are the most critical risks and challenges for companies using open (government) data?"* we find an association between the barriers and limitations proposed by the literature and the risks and challenges mentioned by ODINE's applicants across Europe. Figure 4.3 compares results in both methodologies. On the left side, the list of risks in the literature, and on the right side, the list of risk-related terms found in our corpus, ranked by occurrence frequency. In what follows, we describe the risks that we find, and based on our reading of the answers that mention each risk, we provide their context and their link to the risks in the literature.

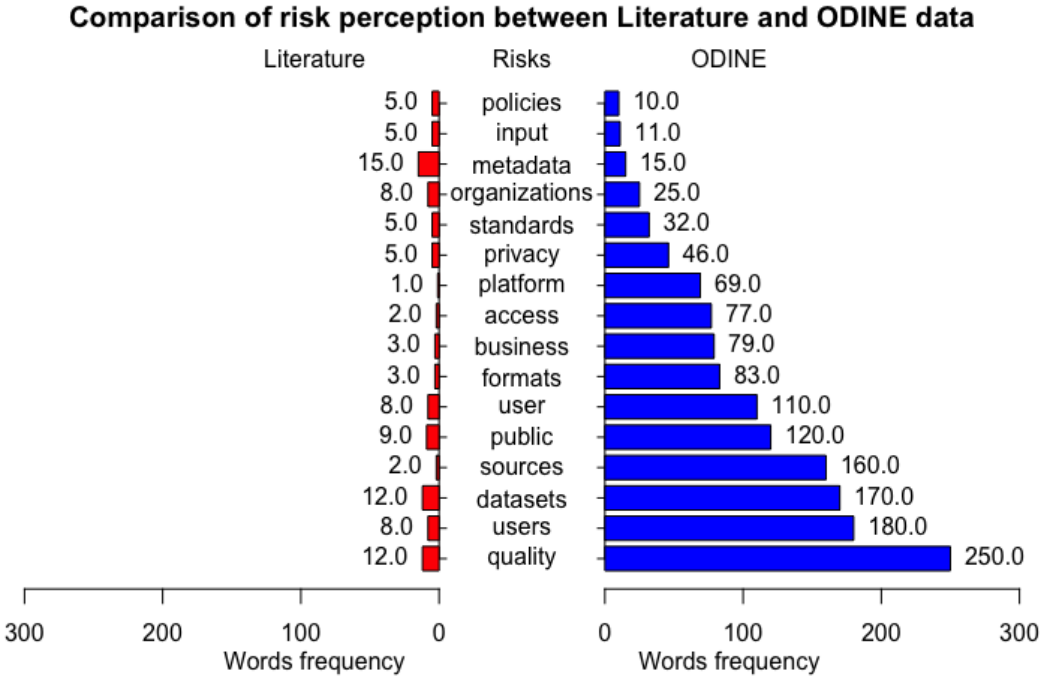


FIGURE 4.3: Shows the risk comparison between literature and entrepreneurs.

4.6.1 Risks in the Literature Review

We implement a text mining approach and we extract the frequency of words stated by the stakeholders (academics and government authorities) in the literature about open data risks. Our results show that for these stakeholders metadata is considered as a main risks. Metadata (which is referred data about the data) is an important element for open data discovery, classification, and contextual information. Furthermore, metadata plays a crucial role in the use and implementation of Linked Open Data (LOD) (Anneke Zuiderwijk et al., 2012b). These are some textual descriptions made by academics and government authorities regarding

metadata "Metadata cannot be found, is not provided, is incomplete or insufficient", "there is no commonly agreed metadata", "contextual metadata is lacking" or "lack of a single standard to describe datasets". Figure 4.4 illustrates the top 10 risks described by academics and government authorities through interviews.

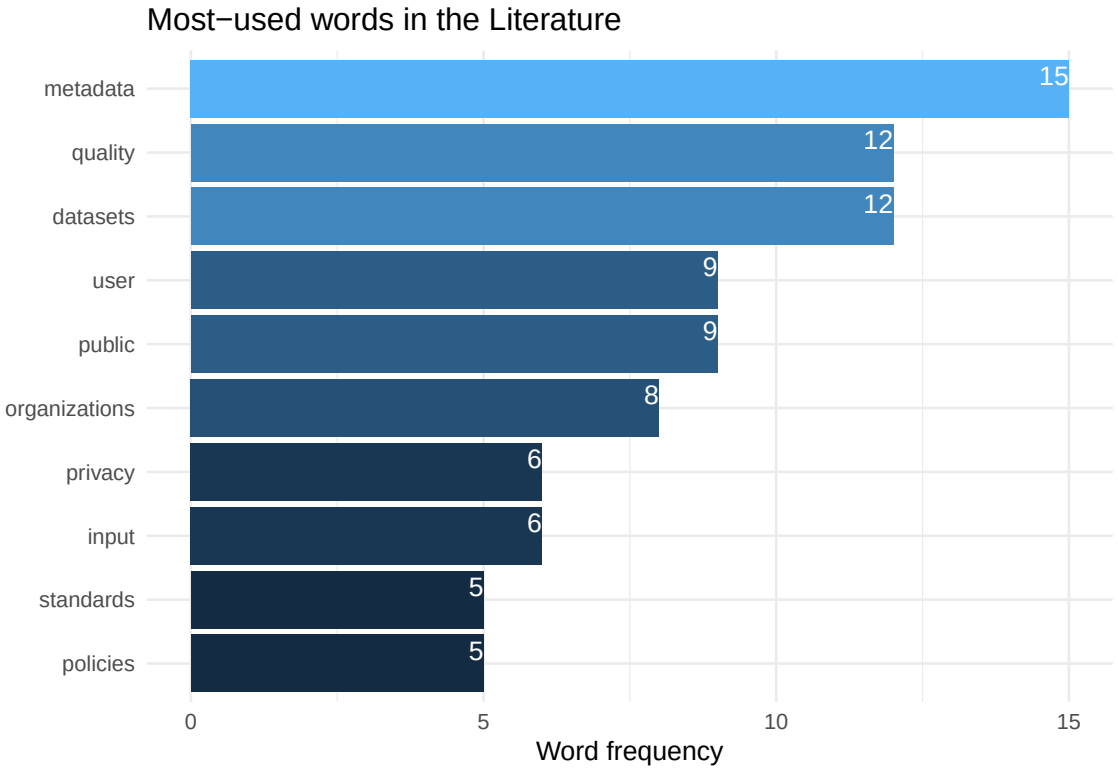


FIGURE 4.4: Shows the top 10 risks described in the literature revised

The second risk mentioned by academics and government authorities is data quality. The context of this word is related to issues such as "Data are not reliable", "lack of accuracy", "Debatable quality for user input" "incomplete information or essential information is missing" "obsolete and non-valid data" "the format of the data is arbitrary or not easily accessible" "Absence of standards".

Datasets are in the third place mentioned as a risk. The context of this word involves different aspects such as licenses and legal framework "lack of heterogeneous licenses across datasets". Metadata "Lack of standards to describe datasets" and quality "Datasets are not complete", "No version management"

The fourth element in the list is User which is related to the interaction between people and data, the frequency of this word involves descriptions such as "Data are only available to a certain group of users (e.g. commercial users, researchers, or governmental organizations)", "Users are forced to employ various arbitrary data transformations to make data usable and comparable", "No incentives or no added value for users to make use of open data", "No discussion between the data provider and the data user possible".

Public is the fifth element that stakeholders described as a risk, the repetition of this word is connected to descriptions such as Public sector e.g. "The use of open data might require a considerable transformation of public sector organizations", "Public organizations do not react to user input", "The open data process is not viewed as an interaction process between the government and the public". This word also includes the topic of the data by itself. For instance, "Making public only non-value-adding data" or "Data are not understandable for the general public".

Organizations are another element that academics and government authorities often mentioned as risks. It refers to governmental institutions and these are some descriptions generated by the stakeholders interviewed "unclear which organization collects which data", "data are not published, as organizations keep these data for themselves", "governmental organizations sometimes use restrictions that are prohibited according to the law" "difference between organizations, for example, differences in terminology. This makes it very difficult to link and combine datasets".

Privacy is also considered as a risk by the academics and government authorities interviewees, some of the concerns described were "threat of privacy violation by publishing data" "unclear trade-off between public values (transparency vs privacy values)", "privacy violation", "privacy and policies for data management and regulation", privacy, opaque ownership or judicial issues" .

The eight word more often mentioned by stakeholders was Input, which in this context is related to the data composition and user feedback about the data. These are some descriptions expressed by the interviewees "debatable quality of user input", "public organizations do not react on user input", "no process for dealing with user input".

Standard is the next word stated by academics and government authorities, some of the descriptions include "Absence of standards", "lack of meta-standards", "no standard software for processing open data", "lack of single standard to describe datasets", "different data standards are available and used".

The tenth most frequently element described by the interviewees is Policies and it is related to the adoption/implementation of open data strategies. For instance, these are some description about this risk "different types of open data policies", "lack of research on differences between open data policies", "inconsistency of public policies", "not all countries worldwide have adopted national open data policies".

4.6.2 Risks in the ODINE application

Concerning the ODINE data, our results show that the main risks and challenges mentioned by entrepreneurs are: quality, users, datasets, sources, formats, business, access, platform,

availability, accuracy. Figure 4.5 shows the top 10 risks described by entrepreneurs in their business proposals submitted to ODINE.

In order to have a better understanding of these results, this research contextualizes the meaning of each word. In order to have a better understanding of these results, this research contextualizes the meaning of each word. For instance, the most critical risk mentioned by entrepreneurs is the quality of the data published (data quality is an abstract concept, that often serves to englobe many other risks such as completeness, availability, standards, or readability). Therefore, according to entrepreneurs' descriptions, governments at the federal and municipal levels (which are the main sources of the open government data) are releasing data without following a methodology that ensures its quality. The lack of standards involved in the production process of data is an additional factor that affects data quality. Entrepreneurs also mentioned that datasets usually come without any quality check and they need to invest resources (financial, technical, human, and time) before using the data. Data quality is also important in terms of demand, this means that if the quality is not improved over time, stakeholders stop using it. Data quality also plays an important role in the accuracy of the metrics or products developed. Some of the datasets mentioned by entrepreneurs that contain imprecisions or low quality are weather, energy, air, georeference, education, company registry, transportation, and health.

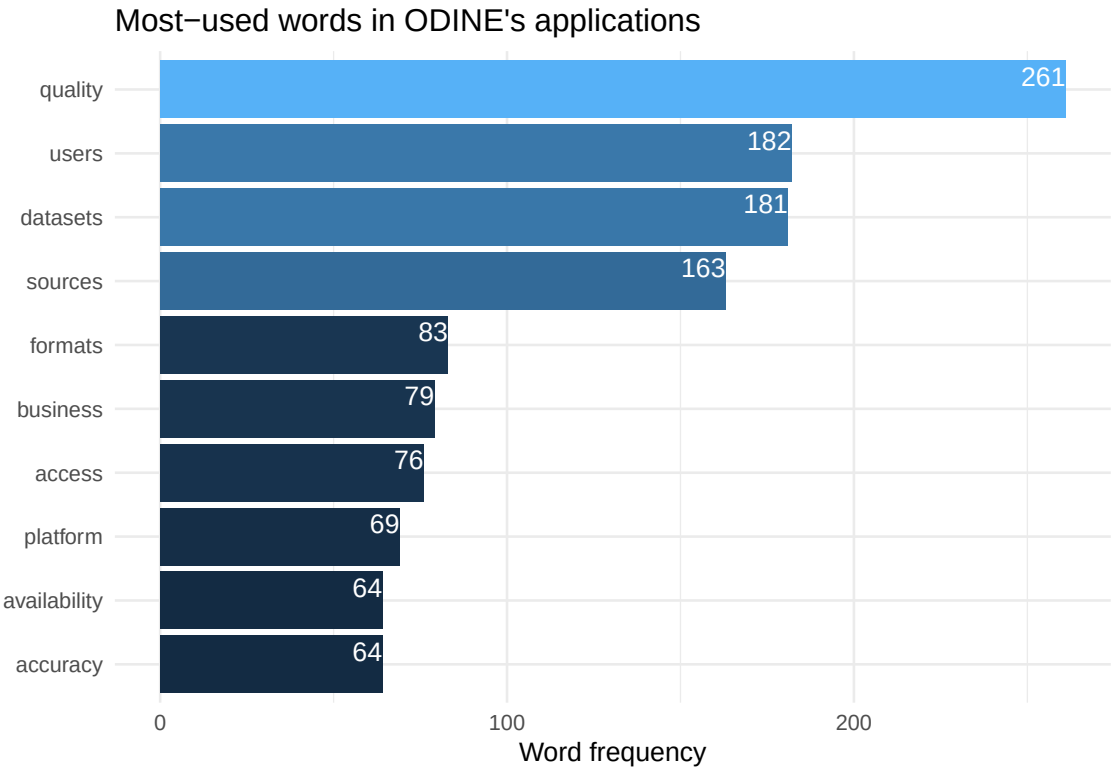


FIGURE 4.5: Displays the top 10 risks described by entrepreneurs

The second position in our results is for the term "users" which in this context refers either to the creation of a customer base and its engagement or to the requirement of some sort of

user participation to enrich their business proposal (e.g., "users create content through the use of the application", "our main challenge is to generate sufficient users", "how to attract and engage users", "the greatest challenge is to incentivise adoption by users."). Privacy and data protection is also a concern related to users due to the potential regulations that could inhibit the use of some type of data. Moreover, entrepreneurs anticipated that anonymization techniques should be implemented in order to preserve a sustainable business model. These are some descriptions about this concern "the risks involved are regarding the violation of users privacy.", "we will only make public anonymous or aggregated data to avoid privacy issues", "levels of open data and laws around anonymizing data sets will vary from country to country".

The following risk is "datasets". This refers to incompleteness in terms of missing values, metadata and description of its attributes. Other difficulties include the format in which the data is released and the type of accessibility. The problem arises when entrepreneurs try to combine and integrate different data sets and they find that the information is not completely reliable. This issue has a negative effect on the credibility of the data released and the data provider, that in the long term affects the open data movement. Some descriptions about this term involves "A risk is one or more of these datasets becoming unavailable", "the quality and the time to update of the datasets can be risky", "The datasets are noisy and not uniform, which represents a significant challenge", "The main challenge is the data sets are exceptionally large, often cumbersome, and were not designed to be integrated."

The next ranked risk is "sources". It refers to the challenges that businessmen face when they need to handle, process and combine data from different data producers and countries. One of these challenges expressed was to find reliable and complete sources of datasets. Other barriers associated with the source of the data is that it could contain personal information, as well as limitations in their commercial use. Diversity in the data format (pdf, xlsx, doc) and languages (english, spanish, italian, greek, german) is also cited. Finally, businessmen also consider availability as a part of this risk, as it is in the case of quality. These are some description stated by entrepreneurs "Data consistency and mapping varies among data sources", "Integration for open publication of datasets from different sources and of heterogeneous nature, and basic cross-linking, are the main challenges", "The key challenge is accommodating diverse data sources and quality".

Businessmen mentioned that the release of open data in diverse "formats" represents a challenge at the time to integrate different datasets. Furthermore, the lack of standardization across Europe affect data interoperability and semantics. Lastly, the lift of cultural barriers and the implementation of specific policies related to the adoption open standards are needed in order to enhance the open data movement. Some descriptions of this risk are "Data format and frequency will vary for each source", "data is provided in heterogeneous formats by different governmental authorities", "Fragmentation across sites also means a diversity of non-standard data formats"

The next risk is "business". When entrepreneurs refer to the risks related to business, they mainly describe challenges associated with the size of the market and user engagement. Concerns are mainly related to the impact of the scalability of the business idea, the legal framework of the data and its integration case. Furthermore, entrepreneurs expressed their concern about the Open Data policies sustainability due to political wills and its relation to their business model. These are some concerns stated by entrepreneurs "The most important risk/challenge is related to the scalability of the business model over national audiences with different political cultures, regulations, and market", "government will withdraw the data, or change the dataset to the extent that it interrupts the continuity of the business", "access to this data be taken away, then the viability of the business could be at risk". We highlight the fact that this is the only risk without a direct match with the ones described in the literature which could be explained due to the context of this data.

"Access" is an additional risk mentioned by entrepreneurs, it refers to the number of datasets that are currently open, the frequency in which they are updated, the right that businessmen have to use and reuse the data, the cultural barriers that governments have to publish data and the certainty and sustainability of the policies adopted to publish and release open data. Some description of this risk includes "will focus on the countries where these data are available and accessible.", "Without access to these open data, the AI is not able to provide customized and suitable financial advice to the user", "access to data might be challenging", "challenges involve: inconvenient open data API access".

"Platforms" in this context are related to the risks and challenges to build a Website or infrastructure that support their idea and engage users and the pertinence of their business model. Some mentions of this risk include "To minimize these risks we store all data in our platform", "The major challenge is related to the promotion/visibility of the platform", "The main challenge will be to win users who will integrate data into the platform"

"Availability" is an additional risk and entrepreneurs refer to it as the commitment of governments to publish and keep the data open, updated and with a high level of quality. These challenges are also related to legislation that policy makers need to do in order to guarantee the data opening. Businessmen also mention that there are still government datasets that remain closed in some sectors (eg. health, transportation, and maps) and there is uncertain if they will be open. Some description involves concerns such as "Our open data strategy presents a number of risks such as data availability due to the content, format or frequency of open data provision can change", "There are risks involved when it comes to data availability", "The major risks associated to open data are warranty and availability of data for medium/long terms"

"Accuracy" was mentioned as a risk since it is an important component of data quality, in this context, accuracy is associated in terms of usability, consistency, cross-lingual, validity, time of update frequency and completeness of its attributes. Entrepreneurs consider that whether data are not updated, it can lose its value. They also consider that the lack of

consistency and completeness makes data useless. These are some points of view described by entrepreneurs "the major risk has to do with the accuracy and reliability of the data", "major challenges are that of the latency and the accuracy of data", "the quality and accuracy of the data from the selected datasets could be a risk".

In order to provide a more detailed content analysis, in the appendix section 1.9 there is a text mining analysis showing the level of association among these 10 risks and giving further insights through a network graph of correlated terms.

4.6.3 Perception according to country of region

Regarding our third research question "Are there any differences in the perceived risks depending on the country of origin of the company?", The approach that we implemented to solve it, was selecting the top 5 countries (United Kingdom, Spain, Germany, Italy, and Greece) that applied to the ODINE program because these economies submitted more than 620 applications of a total of 1173 Figure 4.6 illustrates the number of applications submitted per country.

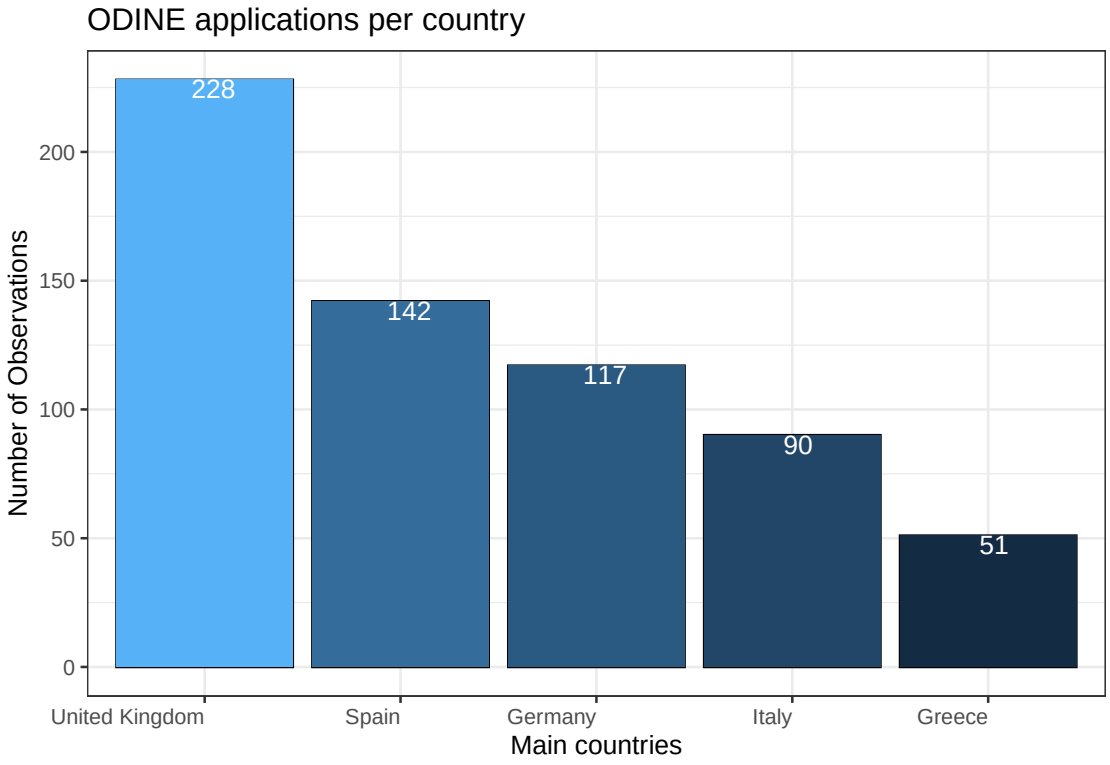


FIGURE 4.6: Illustrates the number of applications per country

Analyzing these 5 countries, our results show that "quality", "datasets" and "users" are the 3 words most frequently mentioned by entrepreneurs as risks or challenges dealing with open data. The quality of the data is described as one of the main concerns in these economies. Recalling that the concept of data quality could be abstract by definition; however, as our

previous analysis suggests, data quality is associated to topics such as readability, accuracy, completeness, format, and update of the data. Finally, there is no significant variance among the other risks mentioned by entrepreneurs when we analyzed risks/challenges per country. Figure 4.7 illustrates the top 10 risks described by entrepreneurs in our selected countries.

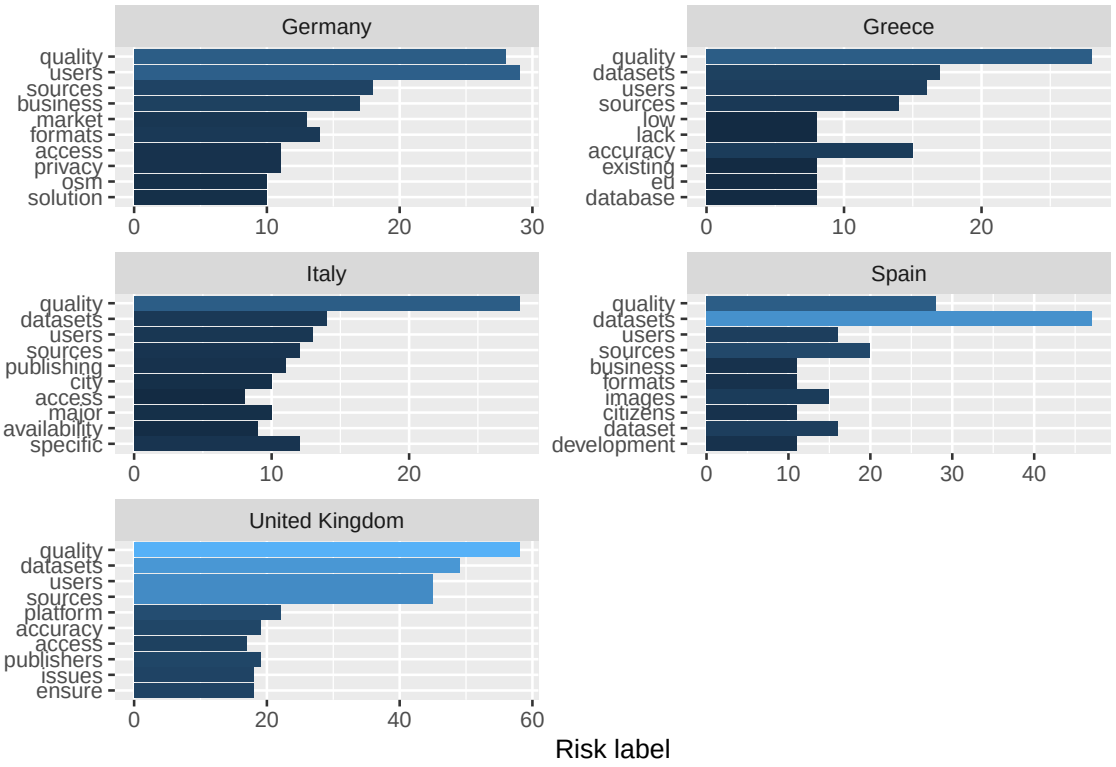


FIGURE 4.7: Shows the risks described by entrepreneurs in the top 5 countries

4.7 Conclusions

Open Data is a socio-technological movement that has been growing globally. Governments at the federal and municipal level, in concordance with other government initiatives that encourage openness, transparency, accountability and civil empowerment are one of the main promoters of this movement. Literature about open data have documented high expectations about its social and economic benefits; however, there are also several concerns about how to realize this potential.

In this chapter, we develop a comparative analysis of risks and challenges of using open (government) data. On the one hand, the risks and challenges described in the literature by academics and government authorities. On the other hand, the risks described by entrepreneurs that are using open data as part of their production process and that applied to the ODINE program. Furthermore, we analyze what are the most critical risks mentioned by entrepreneurs that are facing when they are collecting and transforming open data in order to foster innovation, adopt new business models and create jobs. Then, we compare these risks with respect to the country of origin of the companies, aiming to find similarities or differences among risks. We also conduct our analysis using 989 applications to the Open Data Incubator for Europe (ODINE) program, that answered the question: What risks/ challenges in using open data in the context of the product/service you envision?

Our results show that there is an association between the barriers and limitations proposed by the literature and the risks and challenges mentioned by ODINE's applicants across Europe. Data quality appears in the Literature and ODINE applications as a recurrent concern. The main risks stated in the literature according to their frequency of appearance are 1) metadata, 2) quality, 3) datasets, 4) user, 5) public, 6) organizations, 7) privacy, 8) user input, 9) standard, 10) policies.

Regarding the risks and challenges described by entrepreneurs when they are using open (government) data as raw material are, in order of relevance, 1) quality, 2) users, 3) datasets, 4) sources, 5) formats, 6) business, 7) access, 8) platform, 9) availability, 10) accuracy. Further analysis of the answers revealed that data quality is considered as the main concern since it has an impact in the products developed by entrepreneurs. In addition, low quality implies a cost in the extraction and transformation process that impacts the win share of the company. According to our entrepreneurs' sample, the quality of data is an issue that affects several governmental datasets across Europe, in sources such as weather, energy, air, georeference, education, company registry, transportation, and health. In addition, we conduct a more detailed study revealing that the attributes of datasets mostly associated with data quality are 'updated' (i.e., risks related to how often the data is updated or finding that the data is old), 'available' (issues related to finding datasets, when the data is not available or closed), 'inaccurate' (when the data is lack of consistency and completeness) and 'formats' (lack of standardization and heterogeneous formats).

Regarding country variability, our findings suggest that the variability in risks perceived across european countries is minimal. Data quality is strongly associated with the assessment of risks of companies of almost all european countries. A more in-depth analysis on the 5 countries with most ODINE applications reveals that quality is also perceived as the most important risk.

We conclude that data quality plays a crucial role in terms of open (government) data demand, this means that if the quality is not improved over time stakeholders could stop using this digital asset. Data quality has a direct effect on the open (government) data adoption, and the credibility and influence of its movement. Open data policymakers need to be aware of these concerns since low-quality data is a result of a low-quality production process. Furthermore, additional stakeholders such as private sector should be incorporated in the development of specific policies and strategies that promote more widely the use of government data since this commercial sector could supply and fill the expectations of economic growth and job creation. Future work will be focused on developing measures to assess the risks, barriers, and limitations of using open (government) data on a per-country and per-dataset basis, and developing methods for estimating the impact of these risks on the commercial and social demand of open (government) data.

Chapter 5

Conclusion

The benefits, challenges, impact, and release of open data from public, private, and academic institutions is a topic under exploration and research from different fields and perspectives. This work focuses on adopting an interdisciplinary approach based on economics theory and data science techniques analyzing the concept, relation, use, risks, and challenges of open (government) data from an entrepreneurship perspective. In particular, analyzing what is the association and effect of open (government) data in business opportunity recognition and the choice of individuals to become entrepreneurs. Then, exploring how entrepreneurs are using open (government) data in order to transform it and adopt a business model to monetize it. Later, investigating what are the main risks and challenges that entrepreneurs are facing when they are using open (government) data in their business process. The relevance of this chapter is summarizing the methodologies applied and the main findings of each section in order to provide an overview of the whole contribution of this research. Furthermore, this chapter includes a section of future work giving insights about the following research ideas based on this work and implementing an interdisciplinary approach formulated on economics and data science.

The main contribution of chapter 1 is to provide a theoretical frame explaining the interdisciplinary approach implemented in this research based on economic theory and data science techniques. Moreover, defining the concepts of open data and explaining the role that the public sector is playing in the open data ecosystem. Then, we conceptualize the entrepreneurship term and we focus on examining the interaction of these fields with others such as economic, business, and government policy in order to provide a framework that we can use as domain knowledge in our research.

Chapter 2, studies the economic effect of open (government) data on entrepreneurship. We use optimization techniques to develop a formal theoretical model that allows us to study the relationship between open (government) data and the choice of individuals to become entrepreneurs or employees. These alternatives have benefits and costs and the alternative of becoming an employee is obtaining a salary which represents the opportunity cost of

becoming an entrepreneur. Our model shows that open (government) data might increase the rate of return of becoming an entrepreneur by providing critical information to economic agents that might make the decision of becoming an entrepreneur more profitable. For this reason, more open (government) data might induce more individuals to choose to become an entrepreneur.

To test this hypothesis, in this chapter, we develop an empirical model using econometric and data science techniques to examine the empirical connection between entrepreneurship and open (government) data. To develop this regression analysis we create a sample with data of 135 countries for years 2013 to 2016. Our hypothesis of interest is the relationship between open data and entrepreneurship. For our empirical analysis, we use data from the Open Data Barometer (ODB) generated by The World Wide Web Foundation. To be more specific, we use the open data score which measures, among other things, the availability and quality of open data. We also use the global and entrepreneurship and development index, which is an aggregate measure of entrepreneurship activities. To properly test the relationship between open data and entrepreneurship we use a set of control variables to explain the degree of global heterogeneity of entrepreneurship such as the degree of business freedom in each country, the tax burden and infrastructure in each country, the efficiency of the legal rules, transparency of government policy making, the degree of market efficiency, intellectual property rights protections, etc.

The main findings of this chapter are the following. First, the model shows that changes in open (government) data are correlated with positive changes in the global entrepreneurship and development index (GEDI). Our estimates suggest that a 1% increase in the index of open (government) data leads to increases of 0.11% in the GEDI index. Therefore, open (government) data could provide the information needed for the identification of new business opportunities, strategic planning and the evaluation of investment projects. Moreover, open (government) data could give more access to information and this helps entrepreneurs to reach rational decisions when they have access to information about the needs of supply and demand in markets. Without the access to this crucial information, entrepreneurs might end up choosing dominated alternatives which, in turn, affects the efficiency in the allocation of resources and the rate of return of entrepreneurs.

The main contribution of this chapter to the literature is that, to the best of our knowledge, this is the first empirical analysis that seeks to quantify the economic effect of open (government) data on entrepreneurship. Our analysis not only provides a formal theoretical model that explains why open data affects the economic decisions of entrepreneurs but also provides empirical evidence, using recognized data science techniques (such as regression analysis) on a sample with data constituted by 135 countries for years 2013 to 2016. We believe, that this regression analysis provides strong support for the hypothesis that open government data have significant economic effects on entrepreneurship.

In chapter 3, our main contribution is to provide evidence of the relationship between open (government) data on entrepreneurship. For this reason, we focus our research on examining how entrepreneurs are transforming open (government) data into business ideas on Europe. For this purpose, we use the information collected by the Open Data Incubator for Europe (ODINE) which is a project funded from 2015 to 2017 by the European Commission through the program H2020.

We adopted a text mining approach in order to extract the descriptions made by entrepreneurs that applied to this incubation program. In particular, our interest is analyzing the companies profile, presenting information about the country of origin, the economic sector that they belong, and the year in which the companies were created. Furthermore, we are interested in examining what are the business models adopted by these entrepreneurs across Europe. Exploring the business models structure is important because it describes the key partners, key activities, value propositions, customer relationships, channels, segments, cost structure, and revenue streams that entrepreneurs are adopting as a strategy to commercialize and differentiate their business ideas.

Our results in chapter 3 based on ODINE data is that companies using data as an asset for their production process are relatively young due to some of them were created between 2010 and 2017. This could be associated with the fact of more data availability by governments and other private institutions, business digitalization, and technological evolution. Another finding is that the adoption of open (government) data as a digital asset by entrepreneurs is more perceptible in Southern and Northern than Western and Eastern regions of Europe. Furthermore, the implementation of open (government) data by entrepreneurs is mainly in high-income countries rather than upper middle and lower income. At the country level, the United Kingdom is the country with the highest number of business proposals generated followed by Spain, Germany, Italy and Greece. This result is consistent with the analysis developed in chapter 2 in which the relationship between entrepreneurship and open (government) data tends to have better effects in high-income countries. Some of the determinants of this relationship could be explained because of the technological development, educational level, infrastructure and public policies adopted for these economies.

Another important finding of chapter 3 is that most of the companies that are transforming open (government) data into business proposals are related to the technology sector. For instance, 60% of the entrepreneurs that created a business proposal based on open (government) data belong to the information and communication sector (companies that belong to this sector are related to the production, processing and distribution of data, communication, information technology and other information services activities). In addition, a 20% of the business proposal were created by entrepreneurs registered in the professional scientific and technical activities sector, this area is also related to the development of technological activities. The other 20% of entrepreneurs that created a business proposal are related to different sectors such as agriculture, finance, health, education, real estate activities among others. An important point to note is that of these data-driven companies is that they are mainly

implementing technological methods based on data science, software development, cloud computing, linked data, semantic web, knowledge management, and information systems in order to extract and transform open (government) data into products and services.

Regarding the business model analysis, our results show that key partners of these companies are mainly the private sector, followed by the public and academic sector. This means that entrepreneurs are combining data from private/public and/or academic sector in order to enhance their value proposition and create a differentiator in their products or services. The key activities of these companies are divided into the technical aspects (data gathering, cleaning, processing and algorithms development) administrative tasks (marketing, sales, customer care and management strategies). The value proposition that these data-driven companies varies depending on the sector and solution that they are offering; however, we found that entrepreneurs are supporting their value proposition implementing software engineering methodologies and cloud computing resources such as infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS). Moreover, these companies are also implementing data science techniques (collecting, cleaning, transforming, integrating and enhancing information using machine learning algorithms) in order to turn data into products and services. The channels used by these companies covers the traditional options such as direct contact with customer, CRM, phone, email, and free trials, online platforms, newsletters, and social media. The customer segment is associated to the value proposition. For example, a company related to the professional, scientific and technical activities sector their main customer segments are agricultural input and service providers, agrobusiness (seeds, chemicals), large farms with innovation budget, drinking water companies, public authorities, and financial Institutions. The main cost structure reported by these entrepreneurs are payroll (hiring the right people in terms of technical background due to the challenges that collect and transform data implies), infrastructure (computers and cloud services), administrative and fixed costs. The main revenue stream found is a mix of freemium and premium services in which customers start using the basic services through a freemium option and then they can switch to more advanced features in the premium service

According to our results, we can identify a wide range of companies offering value proposition based on open (government) data as a core idea, some of these companies are not only looking for an economic impact but also they are focusing on the social and/or environmental field. However, a crucial point for these data-driven companies is the human capital, we found that the team composition of these companies is highly skilled holding Ph.D. and MSc degrees in very specific areas of technology such as artificial intelligence, semantic web technologies, natural language processing among others.

In chapter 4, the contribution of this research lies contrasting the limitations, risks, barriers, obstacles that were described by academics and civil servants in the open data literature with the risks and challenges described by entrepreneurs when they are using open (government) data in their business. For this purpose, we collect data through a systematic review process in the literature of open (government) data collecting a list of risk, challenges, barriers, and

impediments described by academics and government authorities through workshops and interviews. Then, we collect the descriptions made by entrepreneurs when they answered the question of "what risks/challenges in using open data in the context of your product/service you envision?" that the ODINE team made in their application process to be selected to their incubation process.

Later, using text mining techniques, we extract and compare both risks assessment .Our results show that there is an association between the barriers and limitations proposed in both groups. On one hand, the list of risk mentioned in the literature review by academic and civil servants are 1) metadata 2) quality 3) datasets 4) users 5) public 6) organizations 7) privacy 8) input 9) standards 10) policies. On the other hand, the categorization of these risks described by entrepreneurs when they are using open (government) data as raw material are, in order of relevance, 1) quality, 2) users, 3) datasets, 4) sources, 5) formats, 6) business, 7) access, 8) platform, 9) availability, 10) accuracy.

An important finding of this research is that entrepreneurs put a special emphasis on the quality of the data that governments are releasing. They argue that data quality has a direct effect on the credibility and influence of the open (government) data movement due to low data quality implies a cost in the extraction and transformation process that impacts the win share of the company. Regarding country variability, our findings suggest that the variability in risks perceived across European countries is minimal, entrepreneurs describe that the quality of data is an issue that affects several governmental datasets across Europe, this involves sources such as weather, energy, air, georeference, education, company registry, transportation, and health. Therefore, entrepreneurs describe that open data policymakers need to be aware of these concerns since low-quality data is a result of a low-quality production process. Our results in this chapter conclude that data quality plays a crucial role in terms of open (government) data demand, this means that if the quality is not improved over time stakeholders could stop using this digital asset.

5.1 Future work

This session covers the next steps and the proposed research ideas based on the experience gained in the development of this thesis. An interesting research topic and which there is scarcity information in the open data community is analyzing the supply of open (government) data. In particular, we are interested in studying what are the determinants and cross-country differences in terms of supplying open data by governments. The aim of this research is to investigate factors such as political participation, sociodemographic characteristics, demographic, and global income distribution that help us to explain the country's supply of open (government) data. Another research idea that gives continuity to my previous work is analyzing the demand for open (government) data by entrepreneurs at the country level.

Another research idea that gives continuity to my previous work is analyzing the demand for open (government) data by entrepreneurs at the country level clustering by economic differences (high, medium, low income). The goal of this research is to study the adoption and implementation of open (government) data by entrepreneurs, focusing on comparing geographical regions due to the political, economic, institutional, financial, and cultural differences. In particular, we are interested in analyzing what are the main datasets that entrepreneurs are using in their business ideas and then compared the availability of this information in developing countries.

An additional research idea is using economic theory and data science in order to analyze the public policies adopted by each government that belongs to the Open Government Partnership (OGP) in which each member of this global initiative proposes a specific action plan to develop a national open data policy. The goal is to explore these action plans, then, clustering similar economies based on the World Bank classification. Later, classifying and comparing these open data policies based on the economic factors of each country.

Finally, the rationale for these research questions is based on the premise that there is a gap or null research work in these areas. Moreover, these questions are relevant and pretend to contribute to open data literature.

Appendix A

Appendix

A.1 Bibliometric analysis

```
# This chunk ensures that the thesisdown package is
#rm(list=ls())

library(thesisdown)
# Set how wide the R output will go
options(width = 70)

library(remotes)
library(thesisdown)
library(bookdown)
library(tidyverse)
library(vroom)
library(rmarkdown)
library(tinytex)
library(pagedown)
library(citr)
library(bibliometrix)

#treemaps
library(treemapify)
library(treemap)

#tables
library(knitr)
library(kableExtra)
```

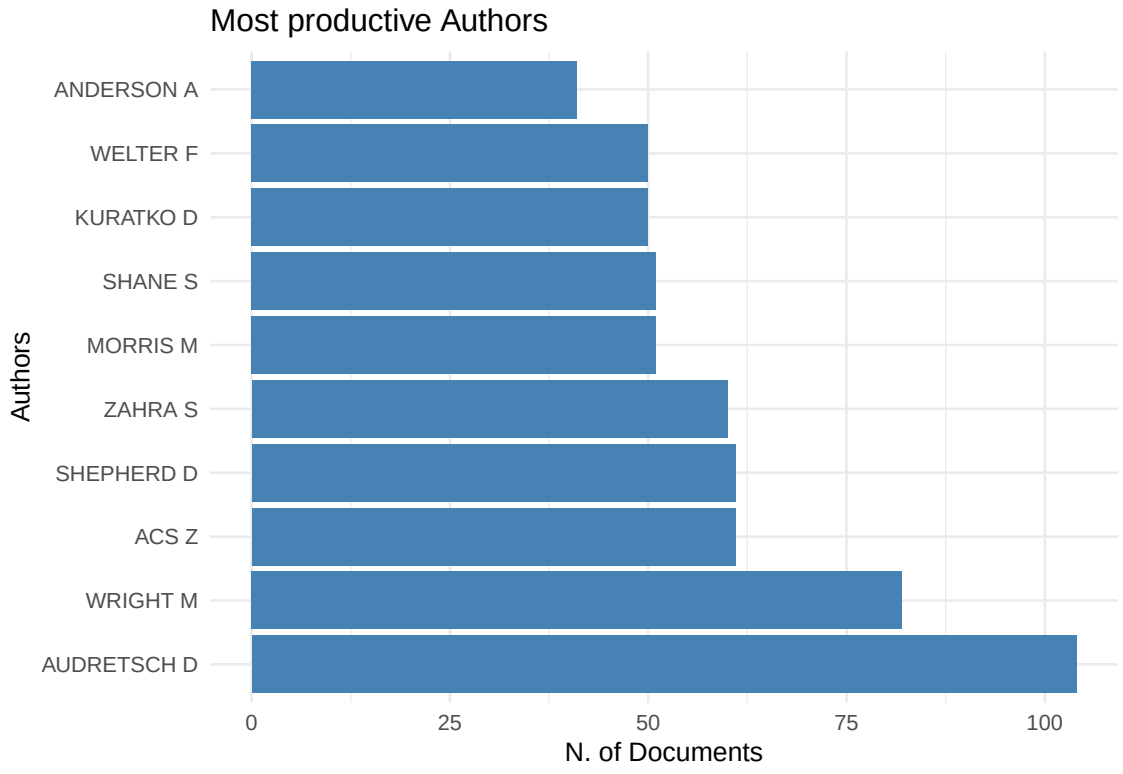
```
library(data.table)
library(flextable)
library(xtable)
library(stargazer)
library(sjPlot)
library(gtsummary)

#NAs
library(naniar)

#ML:PCA
library(rJava)
library(FSelector)
library(missMDA)
library(FactoMineR)
library(factoextra)
library(caret)
library(olsrr)
library(broom)

#Text mining
library(stringi)
library(stringr)
library(tidytext)
library(tidyr)
library(wordcloud)
library(scales)
library(tokenizers)
library(widyr)
library(quanteda)
library(tm)
library(quanteda)
library(stm)
library(SnowballC)
library(igraph)
library(ggraph)
```

```
bibliometric_data_df <-vroom("data/0_data_bibliometric.csv")
```

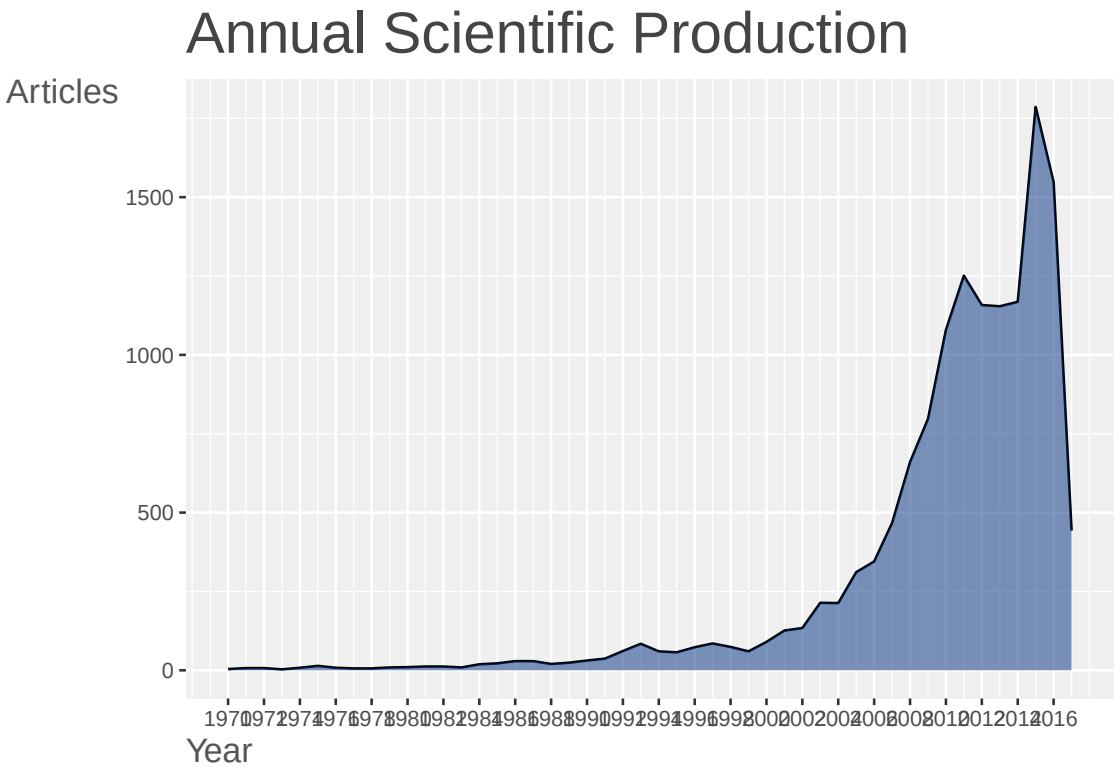
```
Rows: 13,824
Columns: 17
```

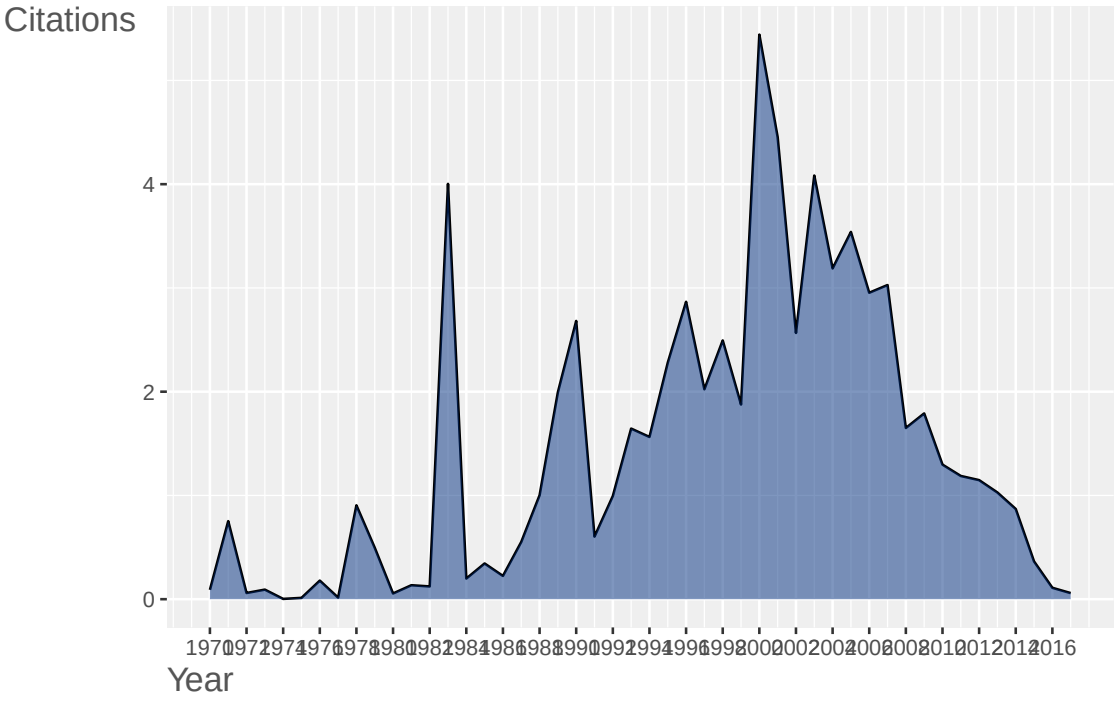
```
Delimiter: ","
chr [14]: AU, TI, SO, JI, DT, DE, ID, AB, C1, CR, SC, UT, RP, DB
dbl [ 3]: number, TC, PY

Use ‘spec()‘ to retrieve the guessed column specification
Pass a specification to the ‘col_types‘ argument to quiet this message
```

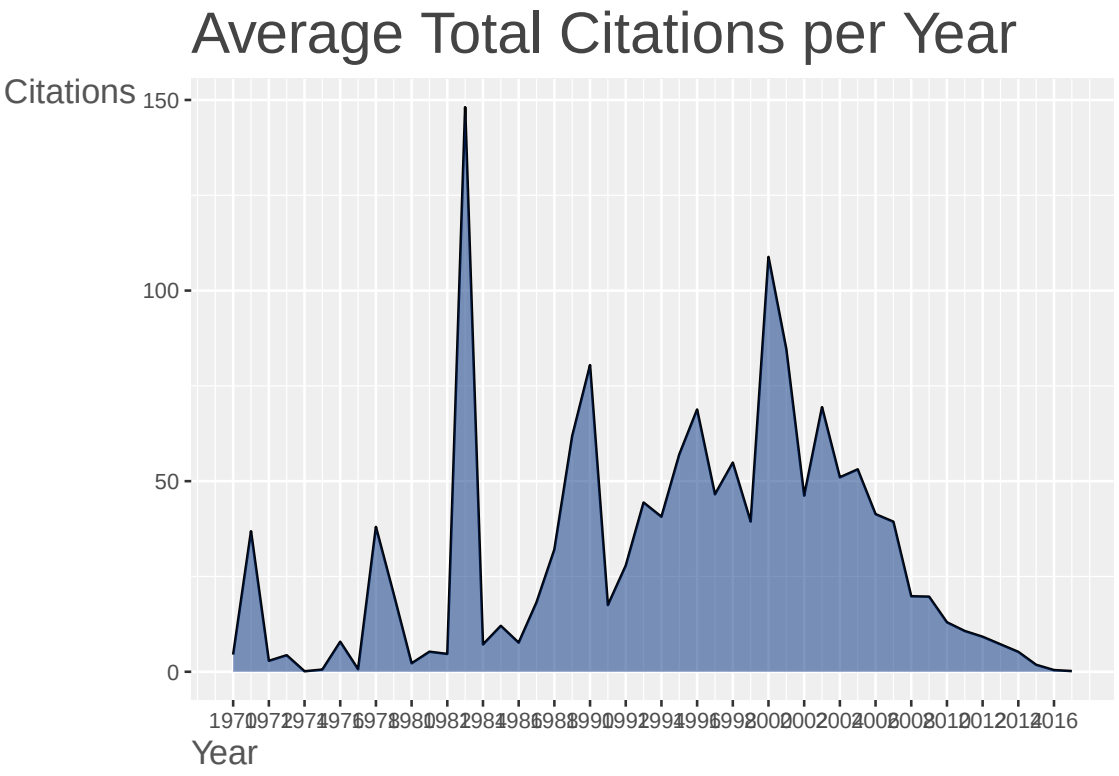
```r
results <- biblioAnalysis(bibliometric_data_df, sep = ";")

bibliometric_summary <- summary(object = results, k = 10, pause = FALSE)

plot(x = results, k = 10, pause = FALSE)
```



UNITED KINGDOM" "PORTUGAL" "UNITED KINGDOM" "BULGARIA" "FRANCE" NA "SWEDEN" "AUSTRIA" "SWEI


Annual Scientific Production
Articles
1500
1000
500
0
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016
Year


Average Article Citations per Year
Citations
4
2
0
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016
Year




```
#Loading data
data_bibliometric <-vroom("data/1_PhD_Graphs_data.csv")
```

```
New names:
* `` -> ...9

Rows: 48
Columns: 28
Delimiter: ","
chr [12]: Articles_Information, Articles_numbers, Authors_Information, Articles_Authors
dbl [16]: Authors_numbers, Articles_Authors_number, Year, Articles_Year, ...9, Number_A

Use `spec()` to retrieve the guessed column specification
Pass a specification to the `col_types` argument to quiet this message
```

```
#Selecting articles per year
articles_year <- data_bibliometric %>%
  select(Articles_Year, Year) %>%
  na.omit()
```

```r
#Selecting articles per country
country_articles <- data_bibliometric %>%
  select(CountryArticles, ArticlesperCountry) %>%
  na.omit()
```

```r
#Selecting articles per category
subject_category <- data_bibliometric2 %>%
  select(Subject_Category) %>%
  na.omit() %>%
  group_by_all() %>%
  count() %>%
  arrange(desc(n)) %>%
  filter(n > 160)
```

```r
#Selecting articles per source
relevant_sources <- data_bibliometric %>%
  select(RelevantSources, RelevantSourcesArticles) %>%
  na.omit()
```

```r
#Creating a dataframe after a text mining extraction process
methodology <- data.frame(
                  type = factor(c("Qualitative","Quantitative",
                                  "Qualitative & Quantitative"),
                          levels=c("Qualitative","Quantitative",
                                   "Qualitative & Quantitative")),
                  total_type = c(1105, 395,  694)
                  )
```

In Chapter 2:

A.2 Data collection: Dependent and independent variables

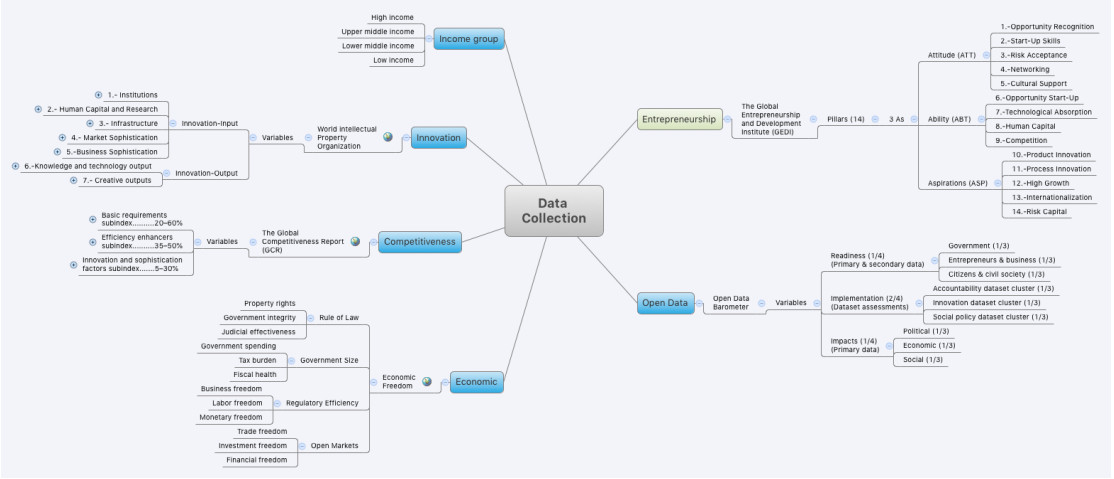
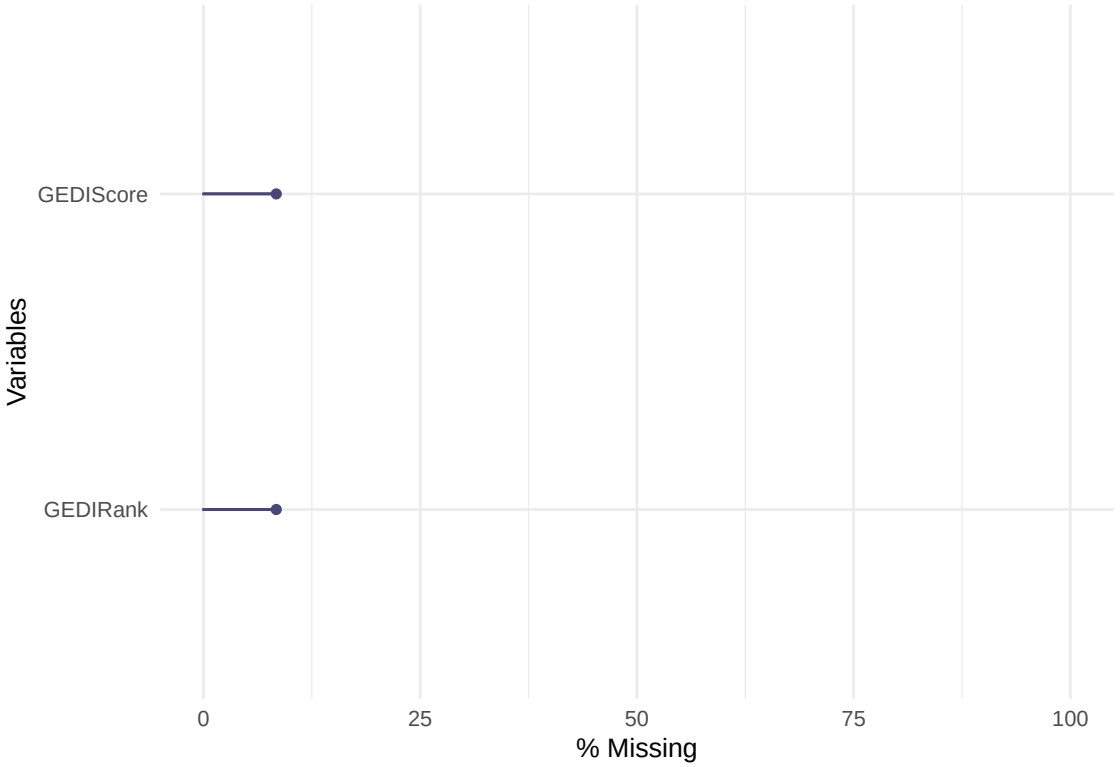


FIGURE A.1: Dependent and independent variables.

A.3 Descriptive summary, % of NAs in each index

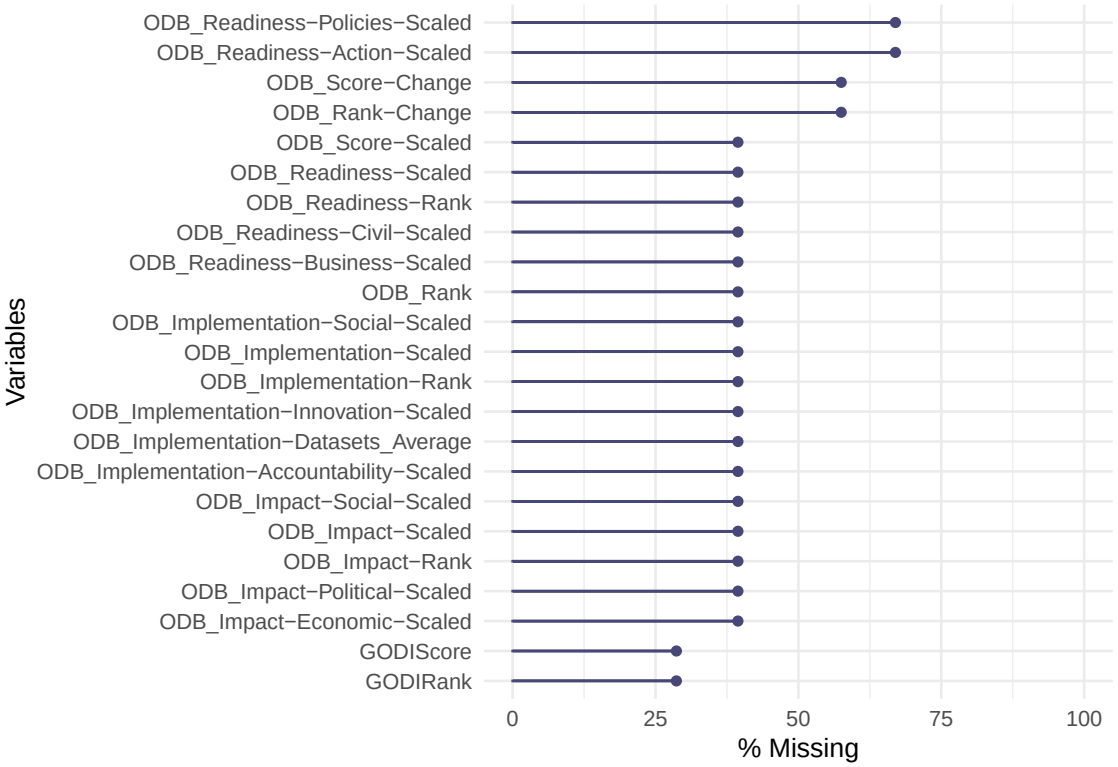
A.3.1 The Global Entrepreneurship and Development Index (GEDI)

```
gedi_na <- all_variables[1:548, 13:14]

#Percentage of missing values.
gg_miss_var(gedi_na, show_pct = TRUE) + ylim(0, 100)
```



A.3.2 The Open Data Barometer (ODB)

```
odb_na <- all_variables[1:548, 16:38]

#Percentage of missing values.
gg_miss_var(odb_na, show_pct = TRUE) + ylim(0, 100)
```



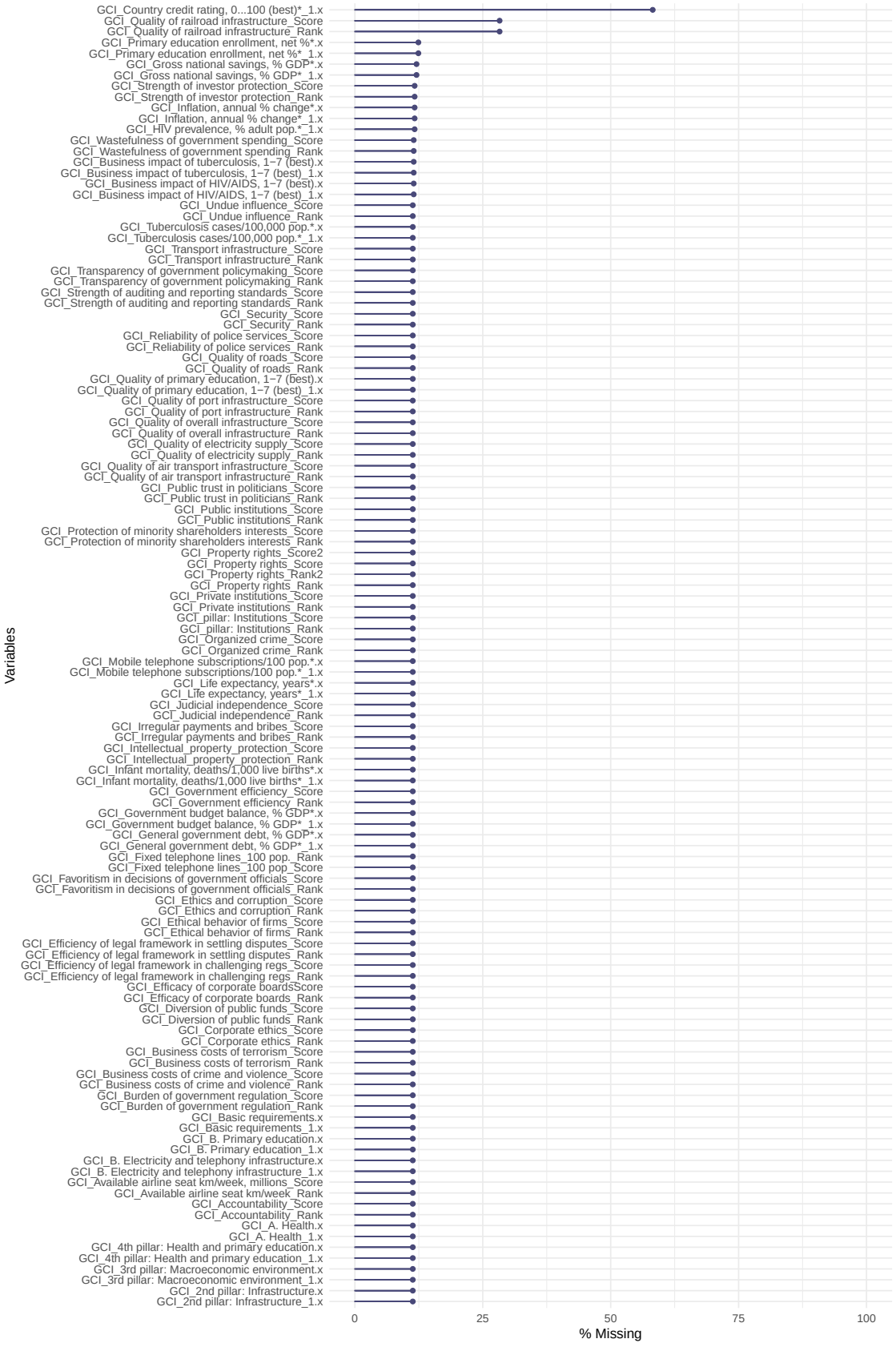
A.3.3 The Economic Freedom Index (EF)

```
ef_na <- all_variables[1:548, 39:63]

#Percentage of missing values.
gg_miss_var(ef_na, show_pct = TRUE) + ylim(0, 100)
```

A.3.4 The Global Competitiveness Index (GCI)

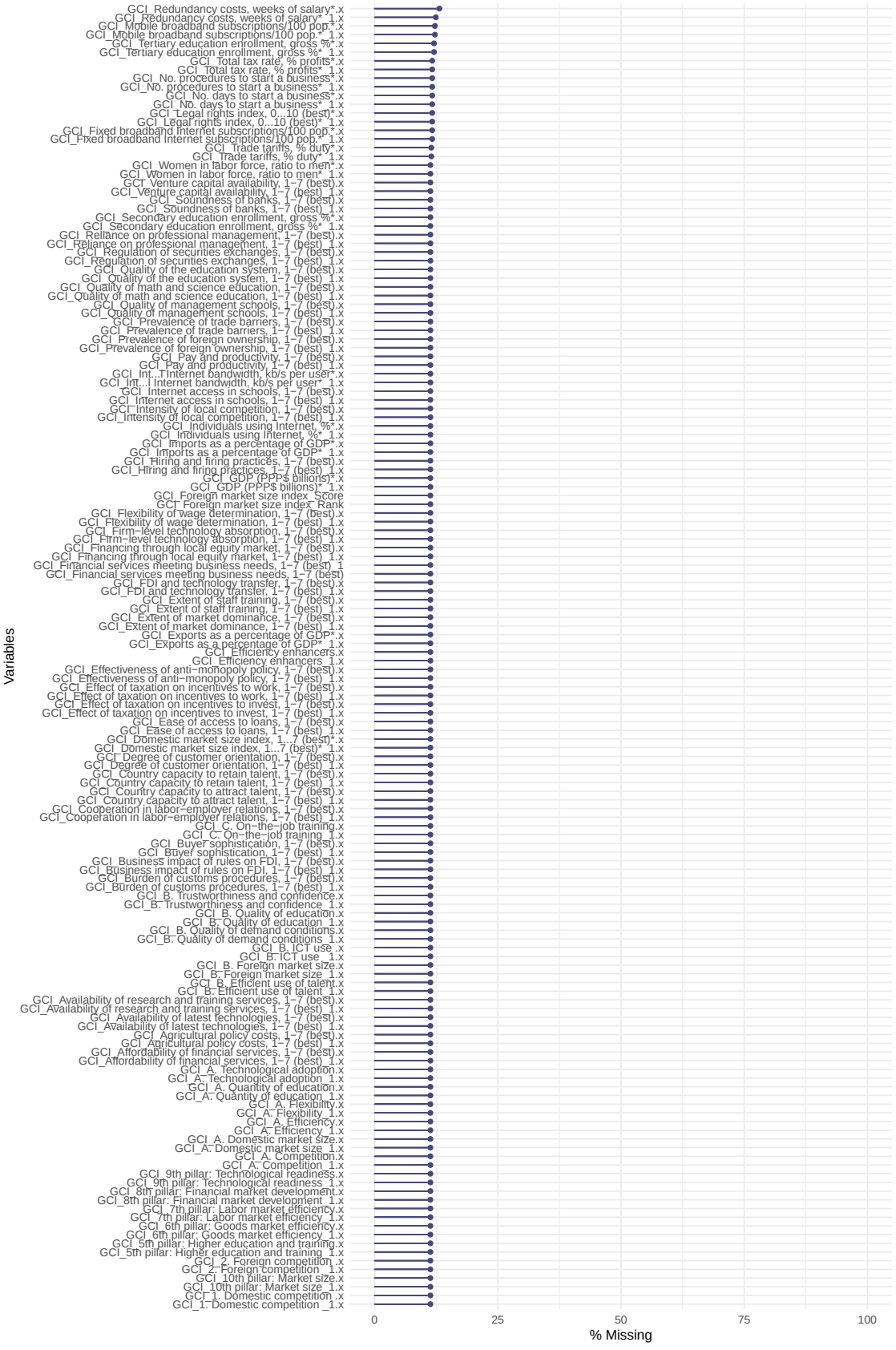
A.3.4.1 Basic Requirement Factor Driven Sub-Pillar

```
### --------Subsetting this indicador  --------

#### Basic Requirement Factor Driver
gci_basic_factor_na <- all_variables[1:548, 64:183]

#Percentage of missing values.
gg_miss_var(gci_basic_factor_na, show_pct = TRUE) + ylim(0, 100)
```


GCI_Country credit rating, 0...100 (best)*_1.x
GCI_Quality of railroad infrastructure_Score
GCI_Quality of railroad infrastructure_Rank
GCI_Primary education enrollment, net %*.x
GCI_Primary education enrollment, net %*_1.x
GCI_Gross national savings, % GDP*.x
GCI_Gross national savings, % GDP*_1.x
GCI_Strength of investor protection_Score
GCI_Strength of investor protection_Rank
GCI_Inflation, annual % change*.x
GCI_Inflation, annual % change*_1.x
GCI_HIV prevalence, % adult pop.*_1.x
GCI_Wastefulness of government spending_Score
GCI_Wastefulness of government spending_Rank
GCI_Business impact of tuberculosis, 1−7 (best).x
GCI_Business impact of tuberculosis, 1−7 (best)_1.x
GCI_Business impact of HIV/AIDS, 1−7 (best).x
GCI_Business impact of HIV/AIDS, 1−7 (best)_1.x
GCI_Undue influence_Score
GCI_Undue influence_Rank
GCI_Tuberculosis cases/100,000 pop.*.x
GCI_Tuberculosis cases/100,000 pop.*_1.x
GCI_Transport infrastructure_Score
GCI_Transport infrastructure_Rank
GCI_Transparency of government policymaking_Score
GCI_Transparency of government policymaking_Rank
GCI_Strength of auditing and reporting standards_Score
GCI_Strength of auditing and reporting standards_Rank
GCI_Security_Score
GCI_Security_Rank
GCI_Reliability of police services_Score
GCI_Reliability of police services_Rank
GCI_Quality of roads_Score
GCI_Quality of roads_Rank
GCI_Quality of primary education, 1−7 (best).x
GCI_Quality of primary education, 1−7 (best)_1.x
GCI_Quality of port infrastructure_Score
GCI_Quality of port infrastructure_Rank
GCI_Quality of overall infrastructure_Score
GCI_Quality of overall infrastructure_Rank
GCI_Quality of electricity supply_Score
GCI_Quality of electricity supply_Rank
GCI_Quality of air transport infrastructure_Score
GCI_Quality of air transport infrastructure_Rank
GCI_Public trust in politicians_Score
GCI_Public trust in politicians_Rank
GCI_Public institutions_Score
GCI_Public institutions_Rank
GCI_Protection of minority shareholders interests_Score
GCI_Protection of minority shareholders interests_Rank
GCI_Property rights_Score2
GCI_Property rights_Score
GCI_Property rights_Rank2
GCI_Property rights_Rank
GCI_Private institutions_Score
GCI_Private institutions_Rank
GCI_pillar: Institutions_Score
GCI_pillar: Institutions_Rank
GCI_Organized crime_Score
GCI_Organized crime_Rank
GCI_Mobile telephone subscriptions/100 pop.*.x
GCI_Mobile telephone subscriptions/100 pop.*_1.x
GCI_Life expectancy, years*.x
GCI_Life expectancy, years*_1.x
GCI_Judicial independence_Score
GCI_Judicial independence_Rank
GCI_Irregular payments and bribes_Score
GCI_Irregular payments and bribes_Rank
GCI_Intellectual_property_protection_Score
GCI_Intellectual_property_protection_Rank
GCI_Infant mortality, deaths/1,000 live births*.x
GCI_Infant mortality, deaths/1,000 live births*_1.x
GCI_Government efficiency_Score
GCI_Government efficiency_Rank
GCI_Government budget balance, % GDP*.x
GCI_Government budget balance, % GDP*_1.x
GCI_General government debt, % GDP*.x
GCI_General government debt, % GDP*_1.x
GCI_Fixed telephone lines_100 pop._Rank
GCI_Fixed telephone lines_100 pop_Score
GCI_Favoritism in decisions of government officials_Score
GCI_Favoritism in decisions of government officials_Rank
GCI_Ethics and corruption_Score
GCI_Ethics and corruption_Rank
GCI_Ethical behavior of firms_Score
GCI_Ethical behavior of firms_Rank
GCI_Efficiency of legal framework in settling disputes_Score
GCI_Efficiency of legal framework in settling disputes_Rank
GCI_Efficiency of legal framework in challenging regs_Score
GCI_Efficiency of legal framework in challenging regs_Rank
GCI_Efficacy of corporate boardsScore
GCI_Efficacy of corporate boards_Rank
GCI_Diversion of public funds_Score
GCI_Diversion of public funds_Rank
GCI_Corporate ethics_Score
GCI_Corporate ethics_Rank
GCI_Business costs of terrorism_Score
GCI_Business costs of terrorism_Rank
GCI_Business costs of crime and violence_Score
GCI_Business costs of crime and violence_Rank
GCI_Burden of government regulation_Score
GCI_Burden of government regulation_Rank
GCI_Basic requirements.x
GCI_Basic requirements_1.x
GCI_B. Primary education.x
GCI_B. Primary education_1.x
GCI_B. Electricity and telephony infrastructure.x
GCI_B. Electricity and telephony infrastructure_1.x
GCI_Available airline seat km/week, millions_Score
GCI_Available airline seat km/week_Rank
GCI_Accountability_Score
GCI_Accountability_Rank
GCI_A. Health.x
GCI_A. Health_1.x
GCI_4th pillar: Health and primary education.x
GCI_4th pillar: Health and primary education_1.x
GCI_3rd pillar: Macroeconomic environment.x
GCI_3rd pillar: Macroeconomic environment_1.x
GCI_2nd pillar: Infrastructure.x
GCI_2nd pillar: Infrastructure_1.x
Variables
0
25
50
75
100
% Missing

A.3.4.2 Efficiency Enhancers Sub-Pillar

```
### Efficiency Enhancers
gci_efficiency_na <- all_variables[1:548, 184:333]

#Percentage of missing values.
gg_miss_var(gci_efficiency_na, show_pct = TRUE) + ylim(0, 100)
```


GCI_Redundancy costs, weeks of salary*.x
GCI_Redundancy costs, weeks of salary*_1.x
GCI_Mobile broadband subscriptions/100 pop.*.x
GCI_Mobile broadband subscriptions/100 pop.*_1.x
GCI_Tertiary education enrollment, gross %*.x
GCI_Tertiary education enrollment, gross %*_1.x
GCI_Total tax rate, % profits*.x
GCI_Total tax rate, % profits*_1.x
GCI_No. procedures to start a business*.x
GCI_No. procedures to start a business*_1.x
GCI_No. days to start a business*.x
GCI_No. days to start a business*_1.x
GCI_Legal rights index, 0...10 (best)*.x
GCI_Legal rights index, 0...10 (best)*_1.x
GCI_Fixed broadband Internet subscriptions/100 pop.*.x
GCI_Fixed broadband Internet subscriptions/100 pop.*_1.x
GCI_Trade tariffs, % duty*.x
GCI_Trade tariffs, % duty*_1.x
GCI_Women in labor force, ratio to men*.x
GCI_Women in labor force, ratio to men*_1.x
GCI_Venture capital availability, 1–7 (best).x
GCI_Venture capital availability, 1–7 (best)_1.x
GCI_Soundness of banks, 1–7 (best).x
GCI_Soundness of banks, 1–7 (best)_1.x
GCI_Secondary education enrollment, gross %*.x
GCI_Secondary education enrollment, gross %*_1.x
GCI_Reliance on professional management, 1–7 (best).x
GCI_Reliance on professional management, 1–7 (best)_1.x
GCI_Regulation of securities exchanges, 1–7 (best).x
GCI_Regulation of securities exchanges, 1–7 (best)_1.x
GCI_Quality of the education system, 1–7 (best).x
GCI_Quality of the education system, 1–7 (best)_1.x
GCI_Quality of math and science education, 1–7 (best).x
GCI_Quality of math and science education, 1–7 (best)_1.x
GCI_Quality of management schools, 1–7 (best).x
GCI_Quality of management schools, 1–7 (best)_1.x
GCI_Prevalence of trade barriers, 1–7 (best).x
GCI_Prevalence of trade barriers, 1–7 (best)_1.x
GCI_Prevalence of foreign ownership, 1–7 (best).x
GCI_Prevalence of foreign ownership, 1–7 (best)_1.x
GCI_Pay and productivity, 1–7 (best).x
GCI_Pay and productivity, 1–7 (best)_1.x
GCI_Int...TInternet bandwidth, kb/s per user*.x
GCI_Int...I Internet bandwidth, kb/s per user*_1.x
GCI_Internet access in schools, 1–7 (best).x
GCI_Internet access in schools, 1–7 (best)_1.x
GCI_Intensity of local competition, 1–7 (best).x
GCI_Intensity of local competition, 1–7 (best)_1.x
GCI_Individuals using Internet, %*.x
GCI_Individuals using Internet, %*_1.x
GCI_Imports as a percentage of GDP*.x
GCI_Imports as a percentage of GDP*_1.x
GCI_Hiring and firing practices, 1–7 (best).x
GCI_Hiring and firing practices, 1–7 (best)_1.x
GCI_GDP (PPP$ billions)*.x
GCI_GDP (PPP$ billions)*_1.x
GCI_Foreign market size index_Score
GCI_Foreign market size index_Rank
GCI_Flexibility of wage determination, 1–7 (best).x
GCI_Flexibility of wage determination, 1–7 (best)_1.x
GCI_Firm–level technology absorption, 1–7 (best).x
GCI_Firm–level technology absorption, 1–7 (best)_1.x
GCI_Financing through local equity market, 1–7 (best).x
GCI_Financing through local equity market, 1–7 (best)_1.x
GCI_Financial services meeting business needs, 1–7 (best)_1
GCI_Financial services meeting business needs, 1–7 (best)
GCI_FDI and technology transfer, 1–7 (best).x
GCI_FDI and technology transfer, 1–7 (best)_1.x
GCI_Extent of staff training, 1–7 (best).x
GCI_Extent of staff training, 1–7 (best)_1.x
GCI_Extent of market dominance, 1–7 (best).x
GCI_Extent of market dominance, 1–7 (best)_1.x
GCI_Exports as a percentage of GDP*.x
GCI_Exports as a percentage of GDP*_1.x
GCI_Efficiency enhancers.x
GCI_Efficiency enhancers_1.x
GCI_Effectiveness of anti–monopoly policy, 1–7 (best).x
GCI_Effectiveness of anti–monopoly policy, 1–7 (best)_1.x
GCI_Effect of taxation on incentives to work, 1–7 (best).x
GCI_Effect of taxation on incentives to work, 1–7 (best)_1.x
GCI_Effect of taxation on incentives to invest, 1–7 (best).x
GCI_Effect of taxation on incentives to invest, 1–7 (best)_1.x
GCI_Ease of access to loans, 1–7 (best).x
GCI_Ease of access to loans, 1–7 (best)_1.x
GCI_Domestic market size index, 1...7 (best)*.x
GCI_Domestic market size index, 1...7 (best)*_1.x
GCI_Degree of customer orientation, 1–7 (best).x
GCI_Degree of customer orientation, 1–7 (best)_1.x
GCI_Country capacity to retain talent, 1–7 (best).x
GCI_Country capacity to retain talent, 1–7 (best)_1.x
GCI_Country capacity to attract talent, 1–7 (best).x
GCI_Country capacity to attract talent, 1–7 (best)_1.x
GCI_Cooperation in labor–employer relations, 1–7 (best).x
GCI_Cooperation in labor–employer relations, 1–7 (best)_1.x
GCI_C. On–the–job training.x
GCI_C. On–the–job training_1.x
GCI_Buyer sophistication, 1–7 (best).x
GCI_Buyer sophistication, 1–7 (best)_1.x
GCI_Business impact of rules on FDI, 1–7 (best).x
GCI_Business impact of rules on FDI, 1–7 (best)_1.x
GCI_Burden of customs procedures, 1–7 (best).x
GCI_Burden of customs procedures, 1–7 (best)_1.x
GCI_B. Trustworthiness and confidence.x
GCI_B. Trustworthiness and confidence_1.x
GCI_B. Quality of education.x
GCI_B. Quality of education_1.x
GCI_B. Quality of demand conditions.x
GCI_B. Quality of demand conditions_1.x
GCI_B. ICT use_.x
GCI_B. ICT use_1.x
GCI_B. Foreign market size.x
GCI_B. Foreign market size_1.x
GCI_B. Efficient use of talent.x
GCI_B. Efficient use of talent_1.x
GCI_Availability of research and training services, 1–7 (best).x
GCI_Availability of research and training services, 1–7 (best)_1.x
GCI_Availability of latest technologies, 1–7 (best).x
GCI_Availability of latest technologies, 1–7 (best)_1.x
GCI_Agricultural policy costs, 1–7 (best).x
GCI_Agricultural policy costs, 1–7 (best)_1.x
GCI_Affordability of financial services, 1–7 (best).x
GCI_Affordability of financial services, 1–7 (best)_1.x
GCI_A. Technological adoption.x
GCI_A. Technological adoption_1.x
GCI_A. Quantity of education.x
GCI_A. Quantity of education_1.x
GCI_A. Flexibility.x
GCI_A. Flexibility_1.x
GCI_A. Efficiency.x
GCI_A. Efficiency_1.x
GCI_A. Domestic market size.x
GCI_A. Domestic market size_1.x
GCI_A. Competition.x
GCI_A. Competition_1.x
GCI_9th pillar: Technological readiness.x
GCI_9th pillar: Technological readiness_1.x
GCI_8th pillar: Financial market development.x
GCI_8th pillar: Financial market development_1.x
GCI_7th pillar: Labor market efficiency.x
GCI_7th pillar: Labor market efficiency_1.x
GCI_6th pillar: Goods market efficiency.x
GCI_6th pillar: Goods market efficiency_1.x
GCI_5th pillar: Higher education and training.x
GCI_5th pillar: Higher education and training_1.x
GCI_2. Foreign competition_.x
GCI_2. Foreign competition_1.x
GCI_10th pillar: Market size.x
GCI_10th pillar: Market size_1.x
GCI_1. Domestic competition_.x
GCI_1. Domestic competition_1.x
Variables
0
25
50
75
100
% Missing

A.3.4.3 Innovation and Sophistication Factors Sub-Pillar

```
### Innovation and Sophistication
gci_innovation_na <- all_variables[1:548, 334:373]

#Percentage of missing values.
gg_miss_var(gci_innovation_na, show_pct = TRUE) + ylim(0, 100)
```

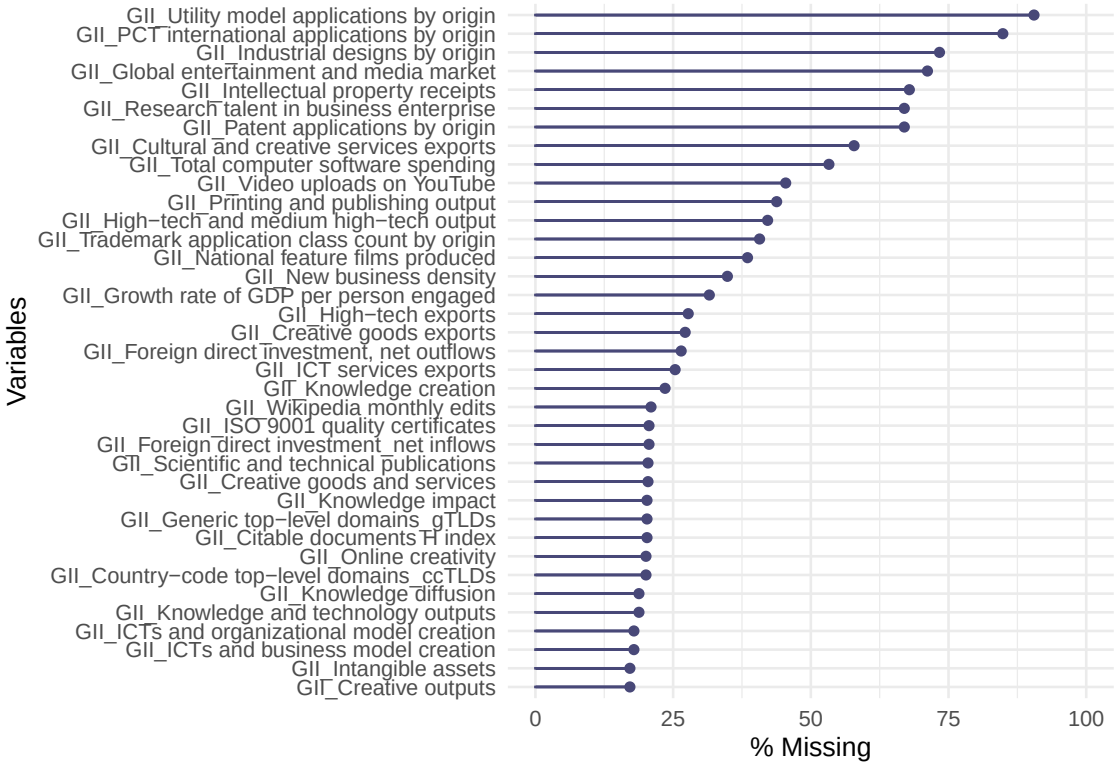
A.3.5 The Global Innovation Index (GII)

A.3.5.1 Innovation -Input Sub-Pillar

```
### --------Subsetting this indicador  --------

### Innovation-Input
rgci_innovation_input_na <- all_variables[1:548, 374:450]

#Percentage of missing values.
gg_miss_var(rgci_innovation_input_na, show_pct = TRUE) + ylim(0, 100)
```

A.3.5.2 Innovation Output Sub-Pillar

```
### Innovation-Output
gci_innovation_output_na <- all_variables[1:548, 451:487]

#Percentage of missing values.
gg_miss_var(gci_innovation_output_na, show_pct = TRUE) + ylim(0, 100)
```

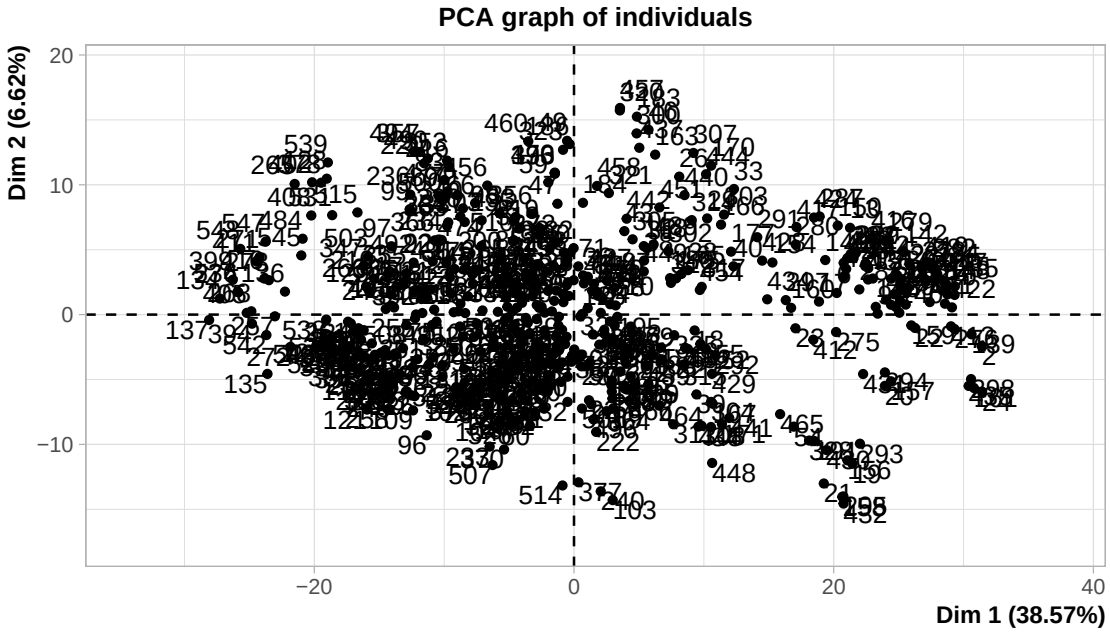


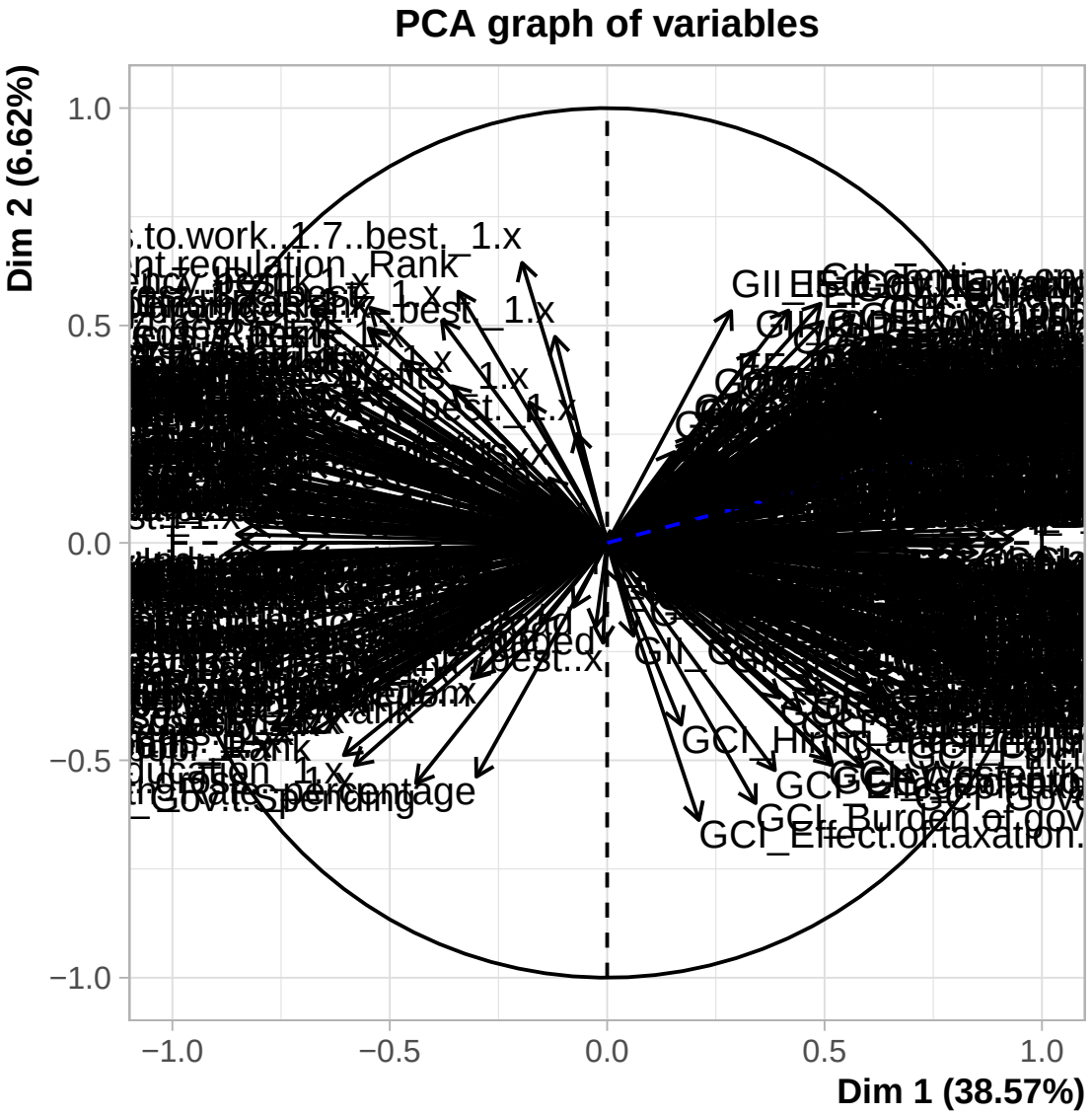
A.4 Principal Components Analysis (PCA)

A.4.1 Eigenvalues

```
GEDIV3 <- read.csv("~/Google Drive/PhD/Software/R/Code/Bookdown/PhD_Thesis/index/data/5_GEDI

#Selecting numeric variables
gedi_all_data <- GEDIV3[1:548, c(14:487)]

# Perform PCA
gedi_pillars_pca <- PCA(gedi_all_data, quanti.sup = 1:1)
```

```
Warning in PCA(gedi_all_data, quanti.sup = 1:1): Missing values are imputed by the mean of the
the imputePCA function of the missMDA package
```





A.4.2 PCA Dimension 1 (Dim1) composition using all indicators

```
#### Dimension description (it describes the Dim1)
gedi_pillars_desc <- dimdesc(gedi_pillars_pca,
                             axes = c(1,2), proba = 0.05)

# Description of dimension 1
gedi_pillars_desc$Dim.1
```

A.4.3 Subset of variables

```
###################### Selecting variables
gedi_model_df <- GEDIV3 %>%
  select(GEDIScore,
    ################### Open Data Index
              ODB_Score.Scaled,
                  ODB_Readiness.Scaled,
                  ODB_Implementation.Scaled,
                  ODB_Impact.Scaled,
                      ODB_Impact.Economic.Scaled,

################### Open Government Partnership
                        OGP,

################### Economic Freedom
                  EFScore,
                     EF_Property.Rights,
                     EF_Freedom.from.Corruption,
                     EF_Fiscal.Freedom,
                     EF_Labor.Freedom,
                     EF_Corporate.Tax.Rate_percentage,
                     EF_Tax.Burden_percentage_of_GDP,
                     EF_Unemployment_percentage,
                     EF_Business.Freedom,
                     EF_GDP_per_Capita_PPP,

################### GCR Competitiveness
  ################ 1.- Institutions
       GCI_Public.institutions_Score,
             GCI_Property.rights_Score,
                      GCI_Intellectual_property_protection_Score,
              GCI_Ethics.and.corruption_Score,
                      GCI_Diversion.of.public.funds_Score,
                      GCI_Public.trust.in.politicians_Score,
                     GCI_Irregular.payments.and.bribes_Score,
              GCI_Undue.influence_Score,
                     GCI_Judicial.independence_Score,
                     GCI_Efficiency.of.legal.framework.in.challenging.regs_Score,
              GCI_Government.efficiency_Score,
                     GCI_Wastefulness.of.government.spending_Score,
                     GCI_Burden.of.government.regulation_Score,
```

```
                    GCI_Efficiency.of.legal.framework.in.settling.disputes_Score,
                # NO   GCI_Efficiency.of.legal.framework.in.challenging.regs_Score,
                    GCI_Transparency.of.government.policymaking_Score,
      GCI_Private.institutions_Score,

  ################## 2.- Infrastructure
              GCI_2nd.pillar..Infrastructure.x,
                  GCI_B..Electricity.and.telephony.infrastructure.x,

  ################ 3.- MacroEconomic Environment
      GCI_3rd.pillar..Macroeconomic.environment.x,

  ################ 5.- Higher Education and Training
     GCI_5th.pillar..Higher.education.and.training.x,
              GCI_A..Quantity.of.education_1.x,
                    GCI_Secondary.education.enrollment..gross....x,
                    GCI_Tertiary.education.enrollment..gross....x,
              GCI_B..Quality.of.education.x,
                    GCI_Quality.of.the.education.system..1.7..best..x,
                    GCI_Quality.of.math.and.science.education..1.7..best..x,
                    GCI_Quality.of.management.schools..1.7..best..x,
                    GCI_Internet.access.in.schools..1.7..best..x,
              GCI_C..On.the.job.training.x,
                    GCI_Availability.of.research.and.training.services..1.7..best..
                    GCI_Extent.of.staff.training..1.7..best..x,

          ####### 6.- Goods Market  Efficienty
                GCI_6th.pillar..Goods.market.efficiency.x,

          ####### 7.-Labor Market Efficienty
                GCI_7th.pillar..Labor.market.efficiency.x,

          ####### 8.- Labor Market Efficienty
                GCI_8th.pillar..Financial.market.development.x,

          ####### 9.- Technological readiness
                GCI_9th.pillar..Technological.readiness.x,

          ####### 10.- Market Size
                GCI_10th.pillar..Market.size.x,
```

```
          ####### 11.- Business sophistication
                GCI_11th.pillar..Business.sophistication..x,

          ####### 12.- R&D Innovation
                GCI_12th.pillar..Innovation.x,


############################ GII Innovation
         ######### 1.- Institutions
      GII_Institutions,
                GII_Political.environment,
                      GII_Political.stability.and.absence.of.violence.terrorism,
                      GII_Government.effectiveness,
               GII_Regulatory.environment,
                      GII_Regulatory.quality,
                      GII_Rule.of.law,
               GII_Business.environment,
                      GII_Ease.of.starting.a.business,
                      GII_Ease.of.getting.credit,
                      GII_Ease.of.paying.taxes,

        ######### 2.- Human Capital and Research
      GII_Human.capital.and.research,
              GII_Education,
                      GII_Expenditure.on.education,
                      GII_Government.expenditure.on.education.per.pupil..secondary,
                      GII_School.life.expectancy,
                      GII_Assessment.in.reading..mathematics..and.science,
                      GII_Pupil.teacher.ratio..secondary,
              GII_Tertiary.education,
                    GII_Tertiary.enrolment,
                    GII_Graduates.in.science.and.engineering,
                    GII_Tertiary.inbound.mobility,
              GII_Research.and.development._R_D,
                    GII_Researchers,
                    GII_Gross.expenditure.on.R_D_GERD,
                    GII_Global.R_D.companies_average.expenditure.top.3,
                    GII_QS.university.ranking.average.score.top.3.universities,

        ######### 3.-Infrastructure
        GII_Infrastructure,
```

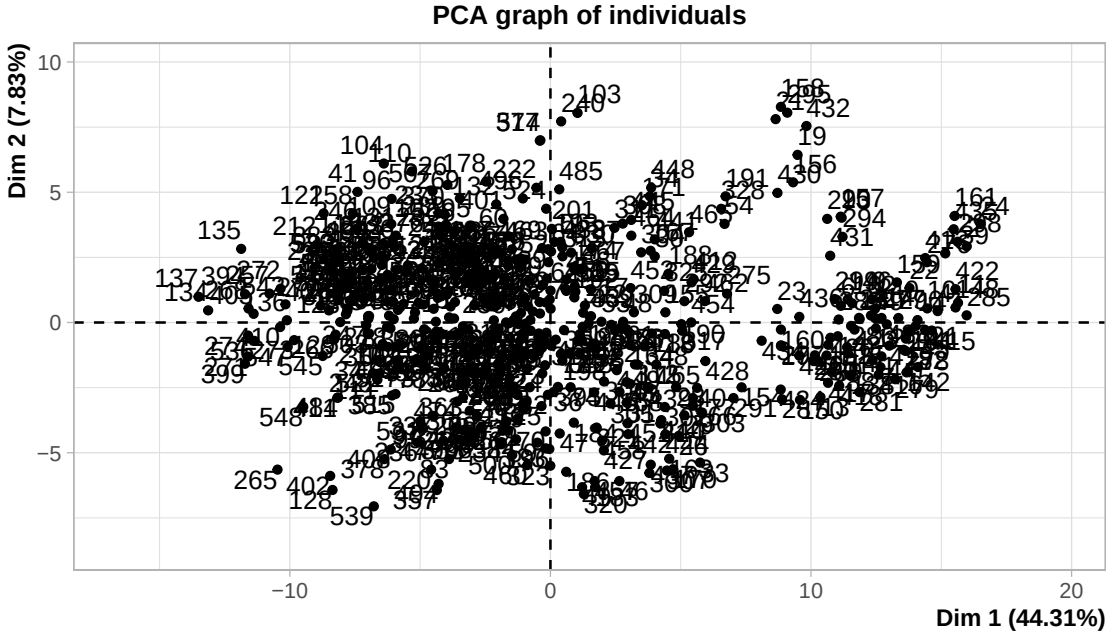
```
        ############ 4.- Market sophistication
        GII_Market.sophistication,

        ############ 5.- Business Sophistication
        GII_Business.sophistication,

   ############ Innovation Output
        ############ 6.- Knowledge and technology output
        GII_Knowledge.and.technology.outputs,

        ############ 7.- Creative Ouputs
      GII_Creative.outputs,
          GII_Intangible.assets,
                    GII_Trademark.application.class.count.by.origin,
                    GII_Industrial.designs.by.origin,
                    GII_ICTs.and.business.model.creation,
                    GII_ICTs.and.organizational.model.creation,
          GII_Creative.goods.and.services,


############################ Income Group
                    IG_High_income,
                    IG_Upper_middle_income,
                    IG_Lower_middle_income,
            #Economy,
            #Income.group

                    IG_Low_income


  ) #End select

gedi_model_df
```

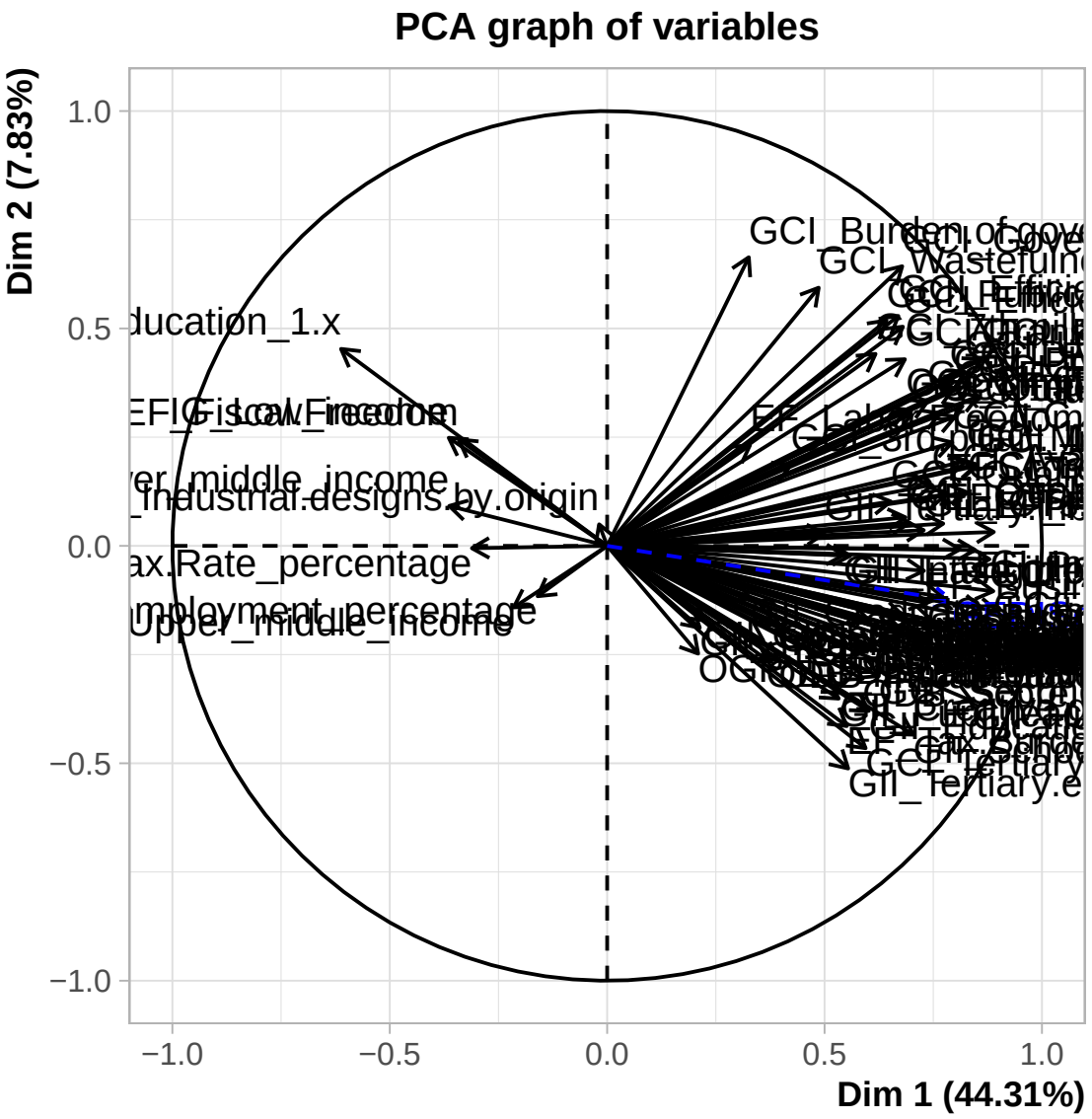
A.5 PCA on Subset of variables

```
#Selecting numeric variables
gedi_pillars_active <- gedi_model_df[1:548, 1:97]
```

```
gedi_6_pillars_pca <- PCA(gedi_pillars_active, quanti.sup = 1:1) # Perform PCA
```

```
Warning in PCA(gedi_pillars_active, quanti.sup = 1:1): Missing values are imputed by the mean
should use the imputePCA function of the missMDA package
```





A.5.1 PCA Dimension 1 (Dim1) composition using a subset

```
#### Dimension description (it describes the Dim1)
gedi_pillars_desc <- dimdesc(gedi_pillars_pca,
                             axes = c(1,2), proba = 0.05)

# Description of dimension 1
gedi_pillars_desc$Dim.1
```

A.6 Selec Features

A.6.1 Stepwise Forward Regression

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	GCI_9th.pillar..Technological.readiness.x	0.5668	0.5660	405.7134	4241.2193	11.5553
2	IG_High_income	0.6296	0.6283	270.0646	4157.4133	10.6949
3	GII_Research.and.development._R_D	0.6696	0.6677	184.4803	4096.8726	10.1110
4	GCI_5th.pillar..Higher.education.and.training.x	0.6843	0.6820	154.1415	4073.8487	9.8918
5	ODB_Score.Scaled	0.6944	0.6916	133.9707	4058.0077	9.7411
6	GCI_Efficiency.of.legal.framework.in.challenging.regs_Score	0.7038	0.7005	115.3789	4042.9005	9.5991
7	GII_Infrastructure	0.7086	0.7049	106.7972	4035.8972	9.5294
8	GII_Creative.goods.and.services	0.7133	0.7091	98.5070	4028.9977	9.4611
9	GCI_2nd.pillar..Infrastructure.x	0.7186	0.7139	88.9446	4020.8225	9.3824
10	EF_Business.Freedom	0.7225	0.7173	82.4091	4015.1879	9.3259
11	GCI_Transparency.of.government.policymaking_Score	0.7253	0.7197	78.1936	4011.5608	9.2868
12	GCI_6th.pillar..Goods.market.efficiency.x	0.7284	0.7223	73.5370	4007.4695	9.2439
13	GII_Ease.of.starting.a.business	0.7324	0.7258	66.7998	4001.3701	9.1845
14	GCI_Intellectual_property_protection_Score	0.7362	0.7292	60.4450	3995.5117	9.1274
15	GCI_Efficiency.of.legal.framework.in.settling.disputes_Score	0.7391	0.7317	56.1258	3991.4922	9.0859
16	GII_Trademark.application.class.count.by.origin	0.7415	0.7337	52.7937	3988.3610	9.0520
17	GCI_Public.institutions_Score	0.7442	0.7360	48.9096	3984.6423	9.0134
18	GCI_Public.trust.in.politicians_Score	0.7462	0.7376	46.4640	3982.2816	8.9861
19	GII_Institutions	0.7481	0.7391	44.2221	3980.0882	8.9602
20	GII_Education	0.7508	0.7414	40.2686	3976.1482	8.9203
21	GII_ICTs.and.business.model.creation	0.7541	0.7443	35.0751	3970.8840	8.8698
22	GII_Tertiary.enrolment	0.7563	0.7461	32.3640	3968.0739	8.8393
23	GII_Political.stability.and.absence.of.violence.terrorism	0.7586	0.7480	29.3602	3964.9182	8.8062
24	GII_Expenditure.on.education	0.7605	0.7496	27.0144	3962.4007	8.7784
25	GCI_10th.pillar..Market.size.x	0.7624	0.7510	25.0427	3960.2395	8.7535
26	GII_Intangible.assets	0.7642	0.7525	22.9054	3957.8708	8.7270
27	GII_Researchers	0.7658	0.7536	21.5765	3956.3301	8.7072
28	GII_Market.sophistication	0.7676	0.7550	19.5705	3954.0388	8.6816
29	GII_School.life.expectancy	0.7686	0.7556	19.4337	3953.7361	8.6717
30	GII_QS.university.ranking.average.score.top.3.universities	0.7699	0.7565	18.6042	3952.6718	8.6559
31	GII_Creative.outputs	0.7708	0.7570	18.5613	3952.4487	8.6467
32	GII_ICTs.and.organizational.model.creation	0.7724	0.7583	16.9266	3950.4710	8.6237
33	GCI_Ethics.and.corruption_Score	0.7734	0.7588	16.8902	3950.2298	8.6145
34	GCI_Government.efficiency_Score	0.7744	0.7595	16.6200	3949.7203	8.6031
35	GII_Government.effectiveness	0.7751	0.7597	17.0651	3949.9949	8.5980
36	GII_Knowledge.and.technology.outputs	0.7757	0.7599	17.7051	3950.4812	8.5945
37	GII_Political.environment	0.7764	0.7602	18.2315	3950.8365	8.5900
38	EF_Property.Rights	0.7771	0.7604	18.8136	3951.2492	8.5860
39	EF_Corporate.Tax.Rate_percentage	0.7776	0.7606	19.5526	3951.8337	8.5834
40	EF_Fiscal.Freedom	0.7792	0.7618	18.0292	3949.8591	8.5607
41	GCI_Tertiary.education.enrollment..gross....x	0.7798	0.7620	18.7857	3950.4494	8.5582
42	GCI_Quality.of.management.schools..1.7..best..x	0.7807	0.7625	18.8145	3950.2074	8.5491
43	IG_Low_income	0.7812	0.7625	19.7517	3950.9947	8.5482

FIGURE A.2: Stepwise forward regression.

A.6.2 Stepwise Backward Regression

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	GCI_9th.pillar..Technological.readiness.x	0.5668	0.5660	405.7134	4241.2193	11.5553
2	IG_High_income	0.6296	0.6283	270.0646	4157.4133	10.6949
3	GII_Research.and.development._R_D	0.6696	0.6677	184.4803	4096.8726	10.1110
4	GCI_5th.pillar..Higher.education.and.training.x	0.6843	0.6820	154.1415	4073.8487	9.8918
5	ODB_Score.Scaled	0.6944	0.6916	133.9707	4058.0077	9.7411
6	GCI_Efficiency.of.legal.framework.in.challenging.regs_Score	0.7038	0.7005	115.3789	4042.9005	9.5991
7	GII_Infrastructure	0.7086	0.7049	106.7972	4035.8972	9.5294
8	GII_Creative.goods.and.services	0.7133	0.7091	98.5070	4028.9977	9.4611
9	GCI_2nd.pillar..Infrastructure.x	0.7186	0.7139	88.9446	4020.8225	9.3824
10	EF_Business.Freedom	0.7225	0.7173	82.4091	4015.1879	9.3259
11	GCI_Transparency.of.government.policymaking_Score	0.7253	0.7197	78.1936	4011.5608	9.2868
12	GCI_6th.pillar..Goods.market.efficiency.x	0.7284	0.7223	73.5370	4007.4695	9.2439
13	GII_Ease.of.starting.a.business	0.7324	0.7258	66.7998	4001.3701	9.1845
14	GCI_Intellectual_property_protection_Score	0.7362	0.7292	60.4450	3995.5117	9.1274
15	GCI_Efficiency.of.legal.framework.in.settling.disputes_Score	0.7391	0.7317	56.1258	3991.4922	9.0859
16	GII_Trademark.application.class.count.by.origin	0.7415	0.7337	52.7937	3988.3610	9.0520
17	GCI_Public.institutions_Score	0.7442	0.7360	48.9096	3984.6423	9.0134
18	GCI_Public.trust.in.politicians_Score	0.7462	0.7376	46.4640	3982.2816	8.9861
19	GII_Institutions	0.7481	0.7391	44.2221	3980.0882	8.9602
20	GII_Education	0.7508	0.7414	40.2686	3976.1482	8.9203
21	GII_ICTs.and.business.model.creation	0.7541	0.7443	35.0751	3970.8840	8.8698
22	GII_Tertiary.enrolment	0.7563	0.7461	32.3640	3968.0739	8.8393
23	GII_Political.stability.and.absence.of.violence.terrorism	0.7586	0.7480	29.3602	3964.9182	8.8062
24	GII_Expenditure.on.education	0.7605	0.7496	27.0144	3962.4007	8.7784
25	GCI_10th.pillar..Market.size.x	0.7624	0.7510	25.0427	3960.2395	8.7535
26	GII_Intangible.assets	0.7642	0.7525	22.9054	3957.8708	8.7270
27	GII_Researchers	0.7658	0.7536	21.5765	3956.3301	8.7072
28	GII_Market.sophistication	0.7676	0.7550	19.5705	3954.0388	8.6816
29	GII_School.life.expectancy	0.7686	0.7556	19.4337	3953.7361	8.6717
30	GII_QS.university.ranking.average.score.top.3.universities	0.7699	0.7565	18.6042	3952.6718	8.6559
31	GII_Creative.outputs	0.7708	0.7570	18.5613	3952.4487	8.6467
32	GII_ICTs.and.organizational.model.creation	0.7724	0.7583	16.9266	3950.4710	8.6237
33	GCI_Ethics.and.corruption_Score	0.7734	0.7588	16.8902	3950.2298	8.6145
34	GCI_Government.efficiency_Score	0.7744	0.7595	16.6200	3949.7203	8.6031
35	GII_Government.effectiveness	0.7751	0.7597	17.0651	3949.9949	8.5980
36	GII_Knowledge.and.technology.outputs	0.7757	0.7599	17.7051	3950.4812	8.5945
37	GII_Political.environment	0.7764	0.7602	18.2315	3950.8365	8.5900
38	EF_Property.Rights	0.7771	0.7604	18.8136	3951.2492	8.5860
39	EF_Corporate.Tax.Rate_percentage	0.7776	0.7606	19.5526	3951.8337	8.5834
40	EF_Fiscal.Freedom	0.7792	0.7618	18.0292	3949.8591	8.5607
41	GCI_Tertiary.education.enrollment..gross....x	0.7798	0.7620	18.7857	3950.4494	8.5582
42	GCI_Quality.of.management.schools..1.7..best..x	0.7807	0.7625	18.8145	3950.2074	8.5491
43	IG_Low_income	0.7812	0.7625	19.7517	3950.9947	8.5482

FIGURE A.3: Stepwise backward regression.

A.6.3 Stepwise Regression

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	GCI_9th.pillar..Technological.readiness.x	addition	0.567	0.566	405.7130	4241.2193	11.5553
2	IG_High_income	addition	0.630	0.628	270.0650	4157.4133	10.6949
3	GII_Research.and.development._R_D	addition	0.670	0.668	184.4800	4096.8726	10.1110
4	GCI_5th.pillar..Higher.education.and.training.x	addition	0.684	0.682	154.1420	4073.8487	9.8918
5	ODB_Score.Scaled	addition	0.694	0.692	133.9710	4058.0077	9.7411
6	GCI_Efficiency.of.legal.framework.in.challenging.regs_Score	addition	0.704	0.701	115.3790	4042.9005	9.5991
7	GII_Infrastructure	addition	0.709	0.705	106.7970	4035.8972	9.5294
8	GCI_9th.pillar..Technological.readiness.x	removal	0.708	0.705	105.2720	4034.3043	9.5241
9	GII_Creative.goods.and.services	addition	0.713	0.709	96.9690	4027.4007	9.4558
10	GCI_2nd.pillar..Infrastructure.x	addition	0.717	0.712	91.3030	4022.6802	9.4067
11	EF_Business.Freedom	addition	0.721	0.716	84.0540	4016.4612	9.3451
12	GCI_Transparency.of.government.policymaking_Score	addition	0.724	0.719	79.0040	4012.1127	9.2998
13	GCI_6th.pillar..Goods.market.efficiency.x	addition	0.727	0.721	74.3690	4008.0690	9.2572
14	GII_Ease.of.starting.a.business	addition	0.731	0.725	68.2210	4002.5555	9.2026
15	GCI_Intellectual_property_protection_Score	addition	0.734	0.727	63.8230	3998.5829	9.1611
16	GII_Institutions	addition	0.736	0.730	59.8890	3994.9850	9.1230
17	GII_ICTs.and.business.model.creation	addition	0.739	0.732	55.6230	3991.0106	9.0819
18	GII_Intangible.assets	addition	0.743	0.735	49.0560	3984.7354	9.0221
19	GII_Education	addition	0.746	0.738	43.9730	3979.7978	8.9736
20	GII_Human.capital.and.research	addition	0.749	0.741	39.6780	3975.5578	8.9311
21	GCI_10th.pillar..Market.size.x	addition	0.751	0.742	37.7170	3973.5950	8.9073
22	GII_Tertiary.enrolment	addition	0.753	0.744	35.5490	3971.3929	8.8816
23	GII_Expenditure.on.education	addition	0.755	0.745	33.3810	3969.1592	8.8558
24	GCI_Efficiency.of.legal.framework.in.settling.disputes_Score	addition	0.757	0.746	31.5130	3967.2005	8.8323
25	GII_Political.stability.and.absence.of.violence.terrorism	addition	0.758	0.748	29.7850	3965.3575	8.8098
26	GII_Researchers	addition	0.760	0.749	28.6510	3964.1064	8.7921
27	GII_Research.and.development._R_D	removal	0.759	0.749	27.4400	3962.9263	8.7902
28	GCI_Public.institutions_Score	addition	0.761	0.750	26.3510	3961.7084	8.7728
29	GCI_Public.trust.in.politicians_Score	addition	0.762	0.751	25.1080	3960.3084	8.7540
30	GII_Market.sophistication	addition	0.764	0.752	23.9050	3958.9291	8.7355
31	GII_Human.capital.and.research	removal	0.764	0.752	22.2570	3957.3012	8.7301
32	GII_School.life.expectancy	addition	0.765	0.753	21.2970	3956.1630	8.7134

FIGURE A.4: Stepwise regression.

```
######## Stepwise graphs
gedi_sfr_graph <- data.frame(
                  type = factor(c("Open Data",
                                  "Economic Freedom", "GCR Competitiveness",
                                  "GII Innovation", "Income Group"),
                                levels=c("Open Data","Open Government Data",
                                  "Economic Freedom", "GCR Competitiveness",
                                  "GII Innovation", "Income Group")),
                  total_type = c(1,  4, 15, 20, 2)
                  )
```

A.7 GEDI Model

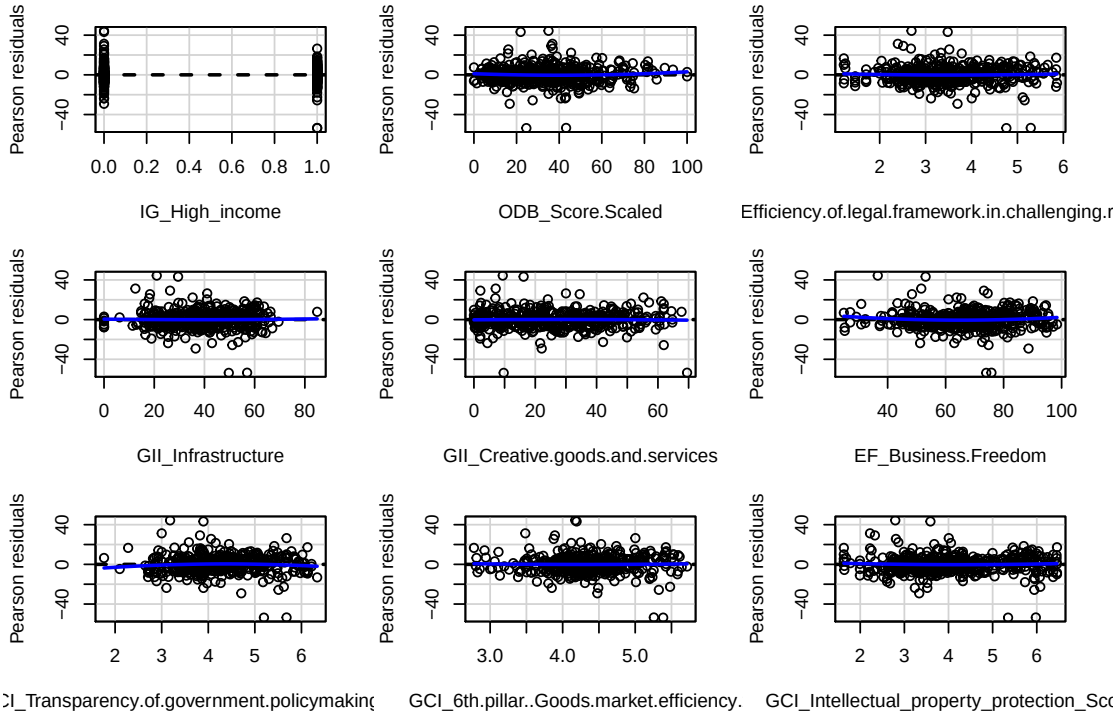
A.7.1 Model results

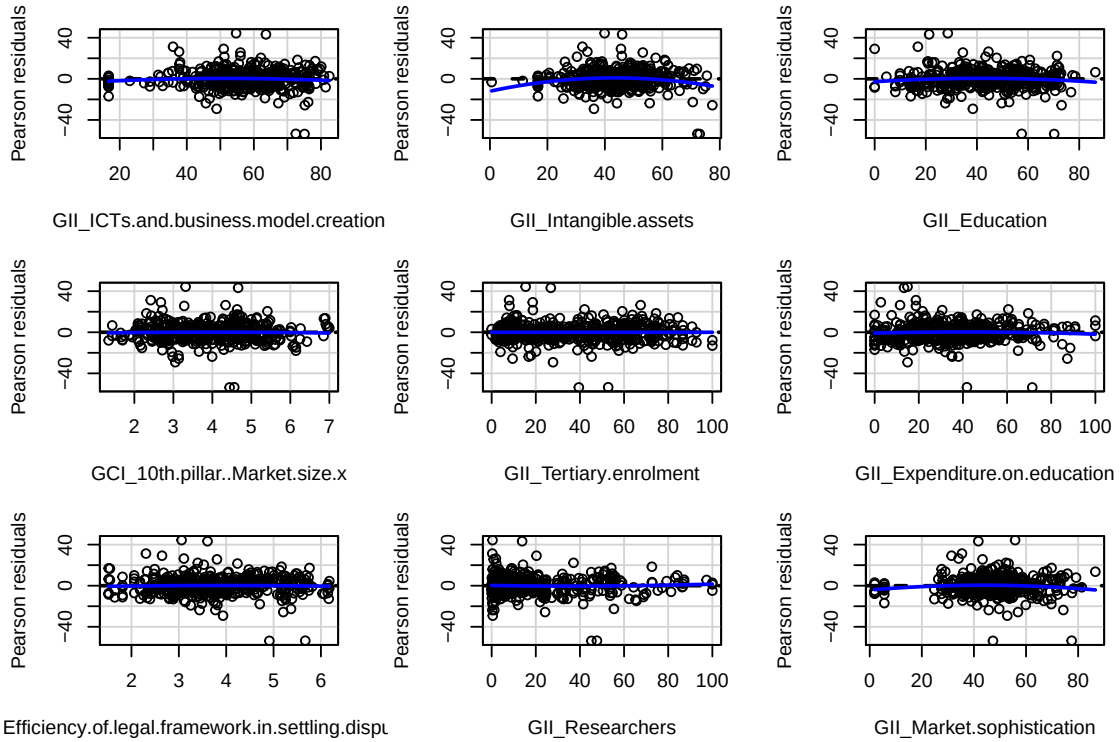
```
###################### Split and Test
# Train and test  using  "gedi_model_df"
set.seed(100)
train_rows <- sample(1:nrow(gedi_model_df),
                           size=0.85*nrow(gedi_model_df))


gedi_training <- gedi_model_df[train_rows, ]
gedi_test <- gedi_model_df[-train_rows, ]


###################### Building a Linear Regressor
gedi_model_training <- lm(GEDIScore ~
                  IG_High_income +

                  ODB_Score.Scaled +
                  GCI_Efficiency.of.legal.framework.in.challenging.regs_Score +
                  GII_Infrastructure +

                  GII_Creative.goods.and.services +

                  EF_Business.Freedom +
                  GCI_Transparency.of.government.policymaking_Score +
                  GCI_6th.pillar..Goods.market.efficiency.x +

                  GCI_Intellectual_property_protection_Score +

                  GII_ICTs.and.business.model.creation +
                  GII_Intangible.assets +
                  GII_Education +

                  GCI_10th.pillar..Market.size.x +
                  GII_Tertiary.enrolment +
                  GII_Expenditure.on.education +
                  GCI_Efficiency.of.legal.framework.in.settling.disputes_Score +
```

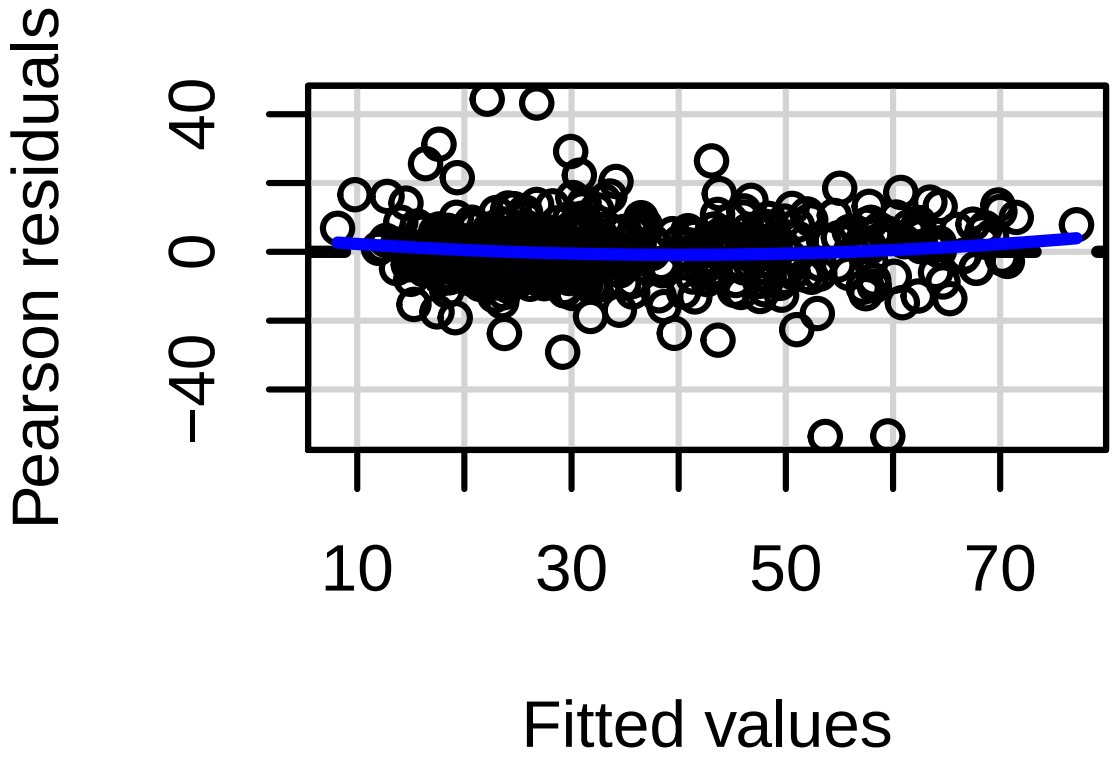
```
                    GII_Researchers +

                    GII_Market.sophistication
                ,
                    data = gedi_training)

summary(gedi_model_training)
```

A.7.2 Analisis of variance

```
# Which features are contributing to the curvature
car::residualPlots(gedi_model_training)
```




Pearson residuals
GII_ICTs.and.business.model.creation
GII_Intangible.assets
GII_Education
GCI_10th.pillar..Market.size.x
GII_Tertiary.enrolment
GII_Expenditure.on.education
Efficiency.of.legal.framework.in.settling.dispu
GII_Researchers
GII_Market.sophistication


Pearson residuals
Fitted values

A.7.3 Relationships between GEDI and independent variables

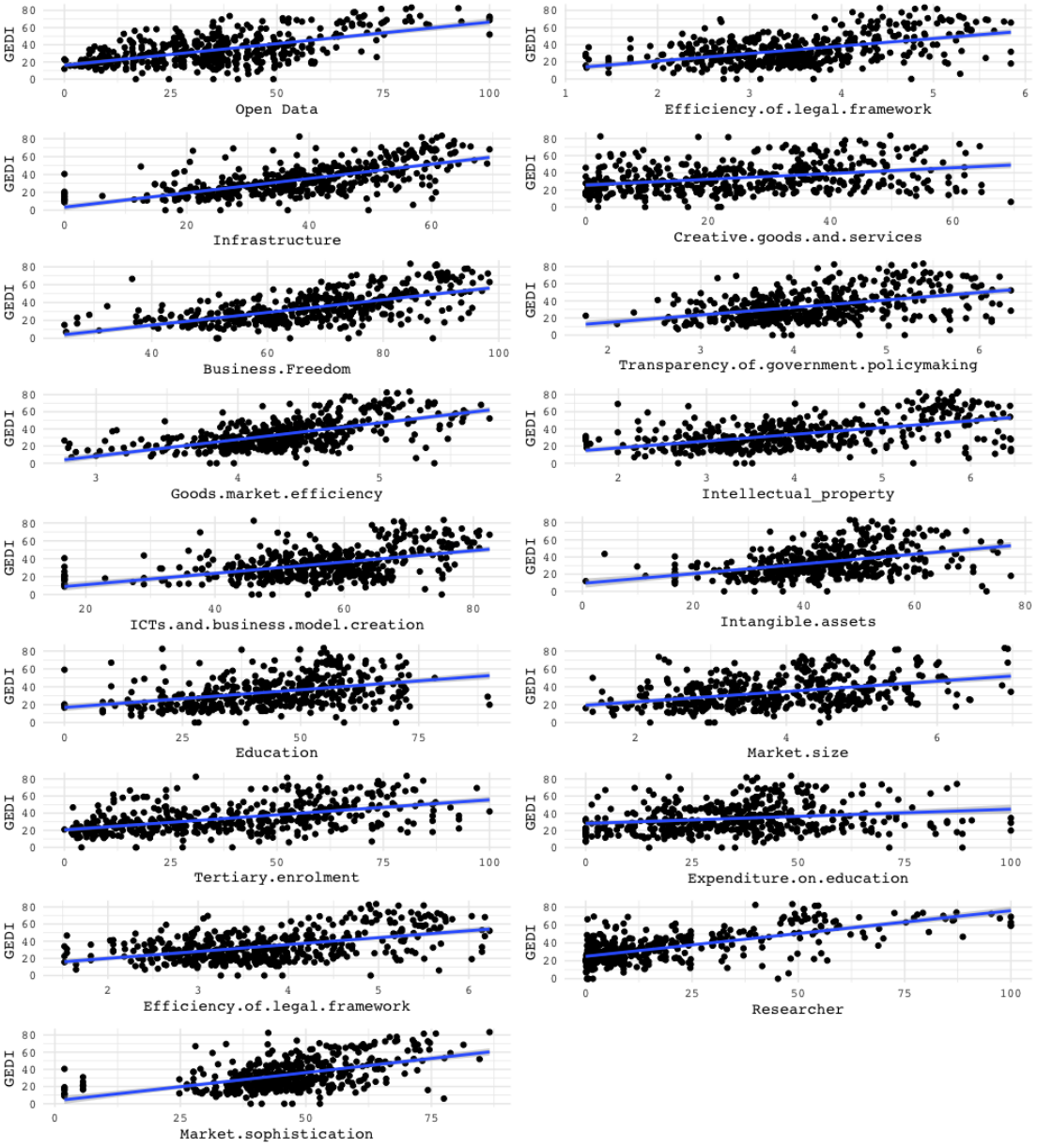


FIGURE A.5: Dependent and independent variables.

A.8 Open Data Incubator for Europe (ODINE): Application template

A.8.1 Idea

Proposal title

1. Idea

1.1 Strength and novelty of the idea

*Describe the core idea of your application in **one sentence**.*	
How are you different from your competitors? *Why are you using and/or producing open data?*	

1.2 Dataset description and use

What data sets (open and proprietary) will you use and how? *Give an example of how open data will be used.*	
What risks/challenges in using open data in the context of your product/service you envision?	

1.3 Open by default

Give an example of how you are contributing to the open data ecosystem.	
Do you rely on personal data?, if so, how do you deal with it?	

FIGURE A.6: Application proposal: Core idea.

A.8.2 Impact

2. Impact

2.1 Value proposition and potential scale

What is the problem you solve? Who are your users? How do you solve it?	
How will you make money? What are your revenue model and monetisation strategy?	
What is the market segment and size you are addressing?	

2.2 Market opportunity and timing

Why is now a good time? Give an example.	
How many users or customers do you already have?	

2.3 What impact will your project have

What impact will your solution have?	
Give a concrete example (where appropriate) of the economic, environmental and/or social impact of your idea.	

FIGURE A.7: Application proposal: Impact.

A.8.3 Team and Budget

3. Team and budget

3.1 Knowledge and skills of the team

List the core members of your team What are their skills?	
How many members are working full/part time on the project?	
Why should we back your team?	

3.2 Capacity to realise the idea

How much short-term funding do you need?	
What is your current monthly cash burn rate?	
What is your time-to-market? What is the customer acquisition cost (actual or predicted)?	
Indicate other sources of funding and how likely you are to secure them.	

Revenue forecasts	Year 0	Year 1	Year 2	Year 3
Revenues (€)				
Headcount (#)				

Year 0 = Last Year | Revenue, profits and headcount can be zero.

Please provide a brief justification (1 paragraph) for your revenue forecast (e.g. customers, pricing, and market size).

FIGURE A.8: Application proposal: Team and budget.

A.8.3.1 Budget

3.3 Budget for the incubation period (6 months)

Give a breakdown of how you will use ODINE's funding for personnel, subcontracting, travel, equipment, and other goods and services. Respect the following rules. Your application might be declared non-eligible if you fail to do so:

1. *Describe costs only for ODINE's incubation period: 6 months and for a maximum of €100 000.*
2. *Remember that a flat overhead rate of 25% is applied to costs (except subcontracting)*
3. *Remember that due to European regulation, only 15% of purchased equipment can be reimbursed. Consult the Guide for Applicants for more details on eligible and reimbursed costs.*
4. *You may remove this instruction notice.*

	Cost over 6 months	**Overhead (25%)**	**Total in Euro**
Personnel			
Travel			
Equipment			
Other goods and services			
Subcontracting		n/a	
		Grand total in Euro	

Please provide a brief explanation of in what you are going to spend the funds (e.g. CEO Salary, subcontract legal advice, travel to XYZ conference, etc). This can be provided inside the cells or as a separate paragraph. You may delete this notice.

FIGURE A.9: Application proposal: Budget.

A.9 ONDINE's Overview

```
#Loading data
odine_data <-read_csv("~/Google Drive/PhD/Software/R/Code/Bookdown/PhD_Thesis/index/da
```

```
Parsed with column specification:
cols(
  counter = col_double(),
  id_round = col_double(),
  round = col_character(),
  submission = col_double(),
  Economy = col_character(),
  Region = col_character(),
  company_name = col_character(),
  company_profile = col_character(),
  extraction = col_character(),
  sector = col_character(),
  activity = col_character(),
  customers = col_character(),
  target_markets = col_character(),
  role = col_character(),
  number = col_double(),
  Code = col_character(),
  Region2 = col_character(),
  `Income group` = col_character()
)
```

A.9.1 Applications per region

```
#Selecting odine application per region
odine_region <- odine_data %>%
  group_by(Region) %>%
  count()

#odine_region
```

A.9.2 Applications per income group

```
#Selecting odine application per region
odine_income <- odine_data %>%
  group_by(`Income group`) %>%
  count()

#odine_income
```

A.9.3 Applications per country

```
#Selecting odine application per country
odine_economy <- odine_data %>%
  group_by(Economy) %>%
  count() %>%
  filter( n> 1)

#odine_economy
```

A.9.4 Applications per sector

```
#Selecting odine application per country
odine_sector_graph <- odine_data %>%
  group_by(sector) %>%
  count() %>%
  arrange(desc(n)) %>%
  filter( n > 5)

odine_sector_graph
```

A.9.5 Company creation

```
#Selecting odine company profile description
odine_company_year <- odine_data %>%
  select(company_profile)
```

```
odine_company_year

#Extracting the year
company_profile_year_extraction  <- stri_extract_all(odine_company_year, regex = "\\d
```

```
Warning in stri_extract_all_regex(str, regex, ...): argument is not an atomic vector; c
```

```
company_profile_year_extraction

#Transforming list to character
company_profile_year_unlist <- unlist(company_profile_year_extraction)
#company_profile_year_unlist

#Transforming character to data frame
company_profile_df <- tibble(line = 1:608, text = company_profile_year_unlist)

#company_profile_df

#Transforming character to numeric
company_profile_df$text <- as.numeric(as.character(company_profile_df$text))

#company_profile_df

company_profile_year <-company_profile_df %>%
  group_by(text) %>%
  filter( text > 1995 & text < 2018) %>%
  count() %>%
  arrange(desc(n))

#company_profile_year
```

A.9.6 Company profile

```
#Selecting odine company profile description
odine_company_profile <- odine_data %>%
  select(company_profile, round)

odine_company_profile
```

```
### This is the same procees of tokenization, just assigting the value a "tidy_books""
# "text" is the column name in the dataframe
tidy_company_profile <- odine_company_profile %>%
  unnest_tokens(word, company_profile) %>%
  anti_join(stop_words)
```

```
Joining, by = "word"
```

```
#tidy_company_profile

## Let's remove some of these less meaningful words to make a better, more meaningful plot
cp_mystopwords <- data_frame(word = c("data", "company", "based", "project", "services",
                                      "public", "here", "me", "pre", "mrx", "f", "step"))

tidy_company_profile <- anti_join(tidy_company_profile,
                                  cp_mystopwords, by = "word")
```

```
################################ BIGRAMS ##########################
## we are examining pairs of two consecutive words, often called \bigrams"
company_profile_bigrams <- odine_company_profile %>%
  unnest_tokens(bigram, company_profile, token = "ngrams", n = 2)

company_profile_bigrams
#View(company_profile_bigrams)

############################## Bigrams: Most common words ###################################

## the most common bigrams
company_profile_bigrams %>%
  count(bigram, sort = TRUE)

## Let's remove some of these less meaningful words to make a better, more meaningful plot
cp_bigram_mystopwords <- data_frame(bigram = c("is a", "in the", "of the",  "and the",
                                               "fgm amor", "dil e.v", "fp5 ist",
                                               "leonardo da", "vinci project",
                                               "da vinci", "we are", "with the",
                                               "for the",
                                               "technologie gmbh", "dil technologie",
                                               "ist 2001", "mondragon corporation"))
```

```
company_profile_bigrams <- anti_join(company_profile_bigrams,
                                     cp_bigram_mystopwords, by = "bigram")
```

A.9.7 Business Models.- Key partners

```
#Loading data
odine_business_model <-read_csv("~/Google Drive/PhD/Software/R/Code/Bookdown/PhD_Thes
```

```
Parsed with column specification:
cols(
  count = col_double(),
  round = col_character(),
  submisssion = col_double(),
  company_name = col_character(),
  sector = col_character(),
  `Key Partners` = col_character(),
  `Key Activities` = col_character(),
  `Key Resources` = col_character(),
  `Value Propositions` = col_character(),
  `Customer Relationships` = col_character(),
  Channels = col_character(),
  `Customer Segments` = col_character(),
  `Cost Structure` = col_character(),
  `Revenue Streams` = col_character(),
  `Year Created` = col_double(),
  Website = col_character(),
  `Technology development` = col_character()
)
```

```
### This is the same procees of tokenization, just assigting the value a "tidy_books"
# "text" is the column name in the dataframe
tidy_key_partners <- odine_business_model %>%
  unnest_tokens(word, `Key Partners`) %>%
  anti_join(stop_words)
```

```
Joining, by = "word"
```

```
#tidy_key_partners

key_partners_data <- data.frame(
                      Key_partners = factor(c("private sector", "public sector", "academic i
                                     levels=c("private sector", "public sector", "academic in
                      observations = c(17, 14, 11)
                    )
key_partners_data
```

A.9.8 Business Models.- Key activities

```
### This is the same procees of tokenization, just assigting the value a "tidy_books""
# "text" is the column name in the dataframe
tidy_key_activities <- odine_business_model %>%
  unnest_tokens(word, `Key Activities`) %>%
  anti_join(stop_words)
```

```
Joining, by = "word"
```

```
tidy_key_activities
#View(tidy_key_activities)


## Let's remove some of these less meaningful words to make a better, more meaningful plot
key_activities_mystopwords <- data_frame(word = c("mt", "language", "custom"))

tidy_key_activities <- anti_join(tidy_key_activities,
                                          key_activities_mystopwords, by = "word")
```

A.9.9 Business Models.- Value proposition

```
### This is the same procees of tokenization, just assigting the value a "tidy_books""
# "text" is the column name in the dataframe
tidy_value_proposition <- odine_business_model %>%
  unnest_tokens(word, `Value Propositions`) %>%
  anti_join(stop_words)
```

```
Joining, by = "word"
```

```
#tidy_value_proposition


# Word Count
tidy_value_proposition_word_count <- tidy_value_proposition %>%
                                     anti_join(stop_words) %>%
                                     count(word, sort = TRUE)
```

```
Joining, by = "word"
```

```
#tidy_value_proposition_word_count

## Let's remove some of these less meaningful words to make a better, more meaningful
tidy_value_proposition_mystopwords <- data_frame(word = c("customer", "bike"))

tidy_value_proposition <- anti_join(tidy_value_proposition,
                                    tidy_value_proposition_mystopwords, by =
```

A.9.10 Business Models.- Revenue streams

```
#Selecting the revenue stream column
odine_revenue_streams <- odine_business_model %>%
  select(`Revenue Streams`)

#odine_revenue_streams


pattern <- "(free|freemium|premium|Premium|license|licensing|Trials|trial|Subscriptio

revenue_streams_extraction <- str_extract_all(odine_revenue_streams,
                                              pattern)
```

```
Warning in stri_extract_all_regex(string, pattern, simplify = simplify, : argument is n
```

```
revenue_streams_extraction <- unlist(revenue_streams_extraction)

#table(revenue_streams_extraction)

revenue_streams_data <- data.frame(
```

```
                      revenue_streams = factor(c("freemium", "premium", "license",
                                                 "consulting"),
                                               levels=c("freemium", "premium", "license",
                                                 "consulting")),
                      observations = c(7, 31, 6,  4)
                    )
#revenue_streams_data
```

A.9.11 Core idea

```
#Loading data
odine_impact <-read_csv("~/Google Drive/PhD/Software/R/Code/Bookdown/PhD_Thesis/index/data/S
```

```
Parsed with column specification:
cols(
  count = col_double(),
  count_round = col_double(),
  round = col_character(),
  submission = col_double(),
  submission_core = col_double(),
  question = col_character(),
  impact_answer = col_character(),
  sector = col_character(),
  company_name = col_character(),
  company_profile = col_character(),
  economy = col_character(),
  region = col_character(),
  country_code = col_character(),
  global_region = col_character(),
  income_group = col_character()
)
```

```
### This is the same procees of tokenization, just assigting the value a "tidy_books""
# "text" is the column name in the dataframe
tidy_core_idea <- odine_core_ideal %>%
  unnest_tokens(word, core_idea_answer) %>%
  anti_join(stop_words)
```

```
Joining, by = "word"
```

```
#tidy_core_idea

################################ BIGRAMS ########################
## we are examining pairs of two consecutive words, often called \bigrams"
core_idea_bigrams <- odine_core_ideal %>%
  unnest_tokens(bigram, core_idea_answer, token = "ngrams", n = 2)

#core_idea_bigrams
#View(core_idea_bigrams)


############################## Bigrams: Most common words ##############################

## the most common bigrams
core_idea_bigrams %>%
  count(bigram, sort = TRUE)

## Let's remove some of these less meaningful words to make a better, more meaningful
cp_bigram_mystopwords <- data_frame(bigram = c("is a", "in the", "of the",  "and the"
                                               "fgm amor", "dil e.v", "fp5 ist",
                                               "leonardo da", "vinci project",
                                               "da vinci", "we are", "with the",
                                               "for the",
                                               "technologie gmbh", "dil technologie",
                                               "ist 2001", "mondragon corporation"))

core_idea_bigrams <- anti_join(core_idea_bigrams,
                                              cp_bigram_mystopwords, by = "bigram")
```

A.9.12 Impact

```
#Loading data
odine_core_ideal <-read_csv("~/Google Drive/PhD/Software/R/Code/Bookdown/PhD_Thesis/i
```

```
Parsed with column specification:
cols(
  count = col_double(),
  count_round = col_double(),
  round = col_character(),
```

```
  submission = col_double(),
  question = col_character(),
  core_idea_answer = col_character(),
  economy = col_character(),
  region = col_character(),
  company_name = col_character(),
  company_profile = col_character(),
  sector = col_character(),
  country_code = col_character(),
  global_region = col_character(),
  income_group = col_character()
)
```

```
### This is the same procees of tokenization, just assigting the value a "tidy_books""
# "text" is the column name in the dataframe
tidy_impact <- odine_impact %>%
  unnest_tokens(word, `impact_answer`) %>%
  anti_join(stop_words)
```

```
Joining, by = "word"
```

```
#tidy_impact

################################ BIGRAMS ##########################
## we are examining pairs of two consecutive words, often called \bigrams"
impact_bigrams <- odine_impact %>%
  unnest_tokens(bigram, `impact_answer`, token = "ngrams", n = 2)

impact_bigrams



############################## Bigrams: Most common words ###################################

## the most common bigrams
impact_bigrams %>%
  count(bigram, sort = TRUE)

## Let's remove some of these less meaningful words to make a better, more meaningful plot
impact_bigram_mystopwords <- data_frame(bigram = c("is a", "in the", "of the",  "and the",
                                                   "fgm amor", "dil e.v", "fp5 ist",
                                                   "leonardo da", "vinci project",
```

```
                                                "da vinci", "we are", "with the",
                                                "for the",
                                                "technologie gmbh", "dil technologie",
                                                "ist 2001", "mondragon corporation"))

impact_bigrams <- anti_join(impact_bigrams,
                                              impact_bigram_mystopwords, by = "bigram")
```

A.9.13 Team

```
#Loading data
odine_team <-read_csv("~/Google Drive/PhD/Software/R/Code/Bookdown/PhD_Thesis/index/da
```

```
Parsed with column specification:
cols(
  count = col_double(),
  count_round = col_logical(),
  round = col_character(),
  submission = col_double(),
  submission_core_idea = col_double(),
  sector = col_character(),
  company_name = col_character(),
  company_profile = col_character(),
  economy = col_character(),
  region = col_character(),
  country_code = col_character(),
  global_region = col_character(),
  income_group = col_character(),
  question = col_character(),
  team_answer = col_character()
)
```

```
#Selecting the revenue stream column
odine_team_academic <- odine_team %>%
  select(team_answer)

pattern <- "(Phd|PhD|Dr|MSc|MBL|MBA|BSc|CEO|CFO|CMO|CTO|Director|data scientist|develo

team_extraction <- str_extract_all(odine_team_academic,
                                           pattern)
```

```
Warning in stri_extract_all_regex(string, pattern, simplify = simplify, : argument is not an a
```

```
team_extraction <- unlist(team_extraction)

#table(team_extraction)

team_data <- data.frame(
                      team = factor(c("PhD", "MSc", "BSc"),
                                    levels=c("PhD", "MSc", "BSc")),
                      observations = c(440, 180, 64)
                    )
team_data
```

```
################################ BIGRAMS ########################
## we are examining pairs of two consecutive words, often called \bigrams"
## we are examining pairs of two consecutive words, often called \bigrams"
team_bigrams <- odine_team %>%
  unnest_tokens(bigram, team_answer, token = "ngrams", n = 2)

team_bigrams


############################## Bigrams: Most common words ###################################

## Let's remove some of these less meaningful words to make a better, more meaningful plot
team_bigram_mystopwords <- data_frame(bigram = c("business development", "project manager",
                                                 "software engineer", "software developer",
                                                 "yrs experience", "serial entrepreneur",
                                                 "software engineering", "data scientist",
                                                 "lead developer"))

team_bigrams <- anti_join(team_bigrams,
                                            team_bigram_mystopwords, by = "bigram")
```

```
#Selecting the revenue stream column
odine_team_academic <- odine_team %>%
  select(team_answer)

pattern <- "(Phd|PhD|Dr|MSc|MBL|MBA|BSc|CEO|CFO|CMO|CTO|Director|data scientist|developer|jo

team_extraction <- str_extract_all(odine_team_academic,
                                         pattern)
```

```
Warning in stri_extract_all_regex(string, pattern, simplify = simplify, : argument is n
```

```
team_extraction <- unlist(team_extraction)

#table(team_extraction)

team_administrative_data <- data.frame(
                      team_administrative = factor(c("CEO", "CMO", "CTO", "CFO", "COO
                                    levels=c("CEO", "CMO", "CTO", "CFO", "COO")),
                      observations = c(410, 25, 277,36, 46)
                    )
team_administrative_data
```

A.9.13.1 Academic titles

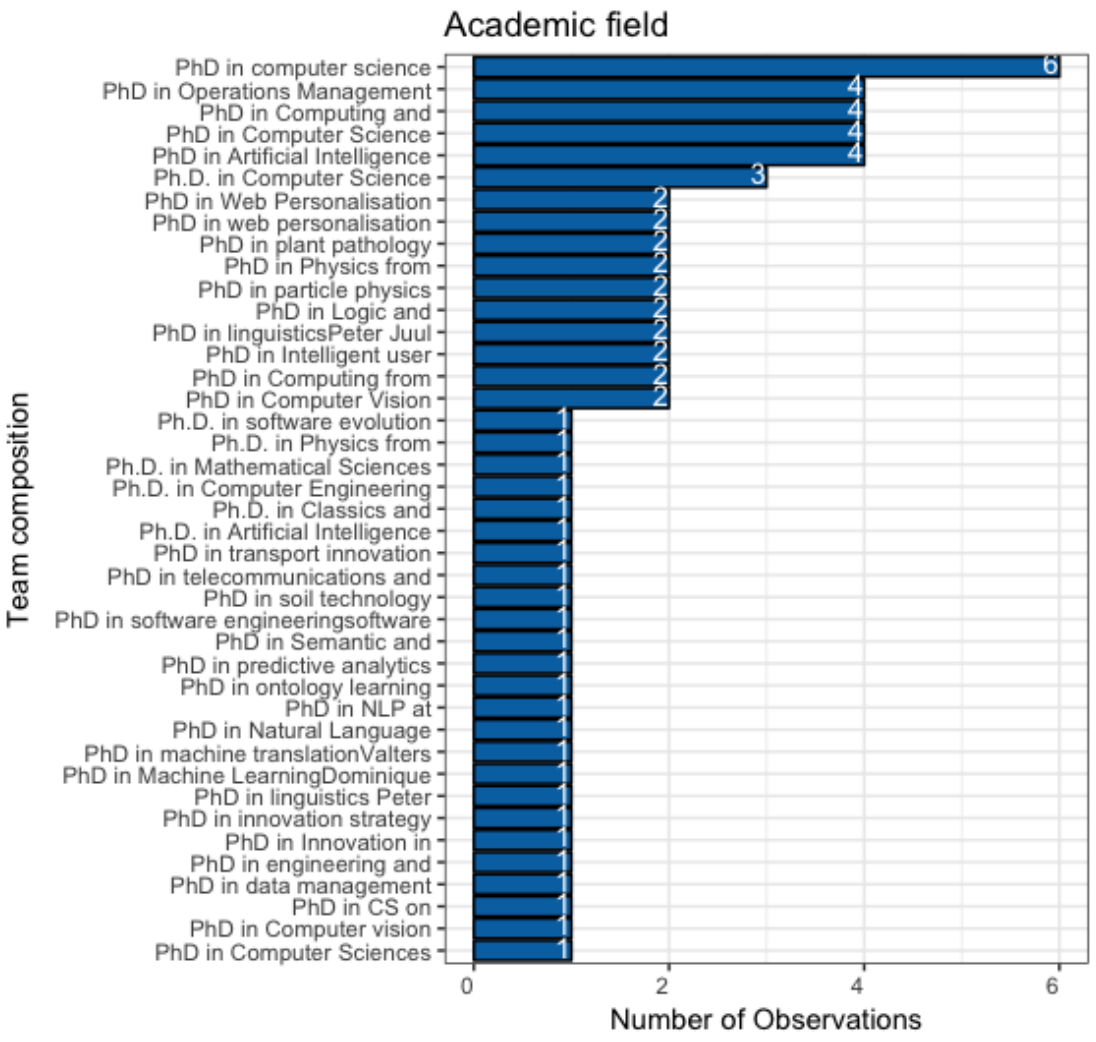


FIGURE A.10: Academic titles: Budget.

A.10 Risks and challenges

A.10.1 Literature risks

```
#Loading data
literature_risks <-read_csv("~/Google Drive/PhD/Software/R/Code/Bookdown/PhD_Thesis/index/da
```

```
Parsed with column specification:
cols(
  Year = col_double(),
  Year2 = col_double(),
  Paper = col_character(),
  Paper2 = col_character(),
  Category = col_character(),
  Label = col_character(),
  Description = col_character(),
  text = col_character()
)
```

A.10.2 ODINE risks

```
#Loading data
odine_risks <-read_csv("~/Google Drive/PhD/Software/R/Code/Bookdown/PhD_Thesis/index/data/12
```

```
Parsed with column specification:
cols(
  count = col_double(),
  count_round = col_double(),
  round = col_character(),
  submission = col_double(),
  submission_core = col_double(),
  sector = col_character(),
  company_name = col_character(),
  company_profile = col_character(),
  economy = col_character(),
  region = col_character(),
  country_code = col_character(),
  global_region = col_character(),
```

```
  income_group = col_character(),
  question = col_character(),
  risks_answer = col_character(),
  text = col_character()
)
```

```
  #Selecting odine application per region
  # "text" is the column name in the dataframe
  tidy_risks_question <- odine_risks %>%
    unnest_tokens(word, `risks_answer`) %>%
    anti_join(stop_words)
```

```
Joining, by = "word"
```

```
  ## Let's remove some of these less meaningful words to make a better, more meaningful
  tidy_risks_mystopwords <- data_frame(word = c("data", "0", "1", "2", "6", "risk", "ch
                                                "information", "main", "risks", "challe
                                                "sets", "provide", "project", "service"
                                                "related", "news", "tail", "bil", "coun
                                                "local", "public", "source", "user"))

  tidy_risks_question <- anti_join(tidy_risks_question,
                                   tidy_risks_mystopwords, by = "word")
```

```
  risks_question_data_df <- read_csv("data/12_what_risks.csv") %>% glimpse()
```

```
Parsed with column specification:
cols(
  count = col_double(),
  count_round = col_double(),
  round = col_character(),
  submission = col_double(),
  submission_core = col_double(),
  sector = col_character(),
  company_name = col_character(),
  company_profile = col_character(),
  economy = col_character(),
  region = col_character(),
  country_code = col_character(),
  global_region = col_character(),
  income_group = col_character(),
```

```
  question = col_character(),
  risks_answer = col_character(),
  text = col_character()
)
```

```
Rows: 989
Columns: 16
$ count           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
$ count_round     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
$ round           <chr> "2R", "2R", "2R", "2R", "2R", "2R", "2R", "2R", "2R", "2R", "2R", "2R
$ submission      <dbl> 4, 5, 7, 9, 10, 11, 12, 13, 14, 15, 19, 20, 21, 22, 24, 25, 26, 27, 29
$ submission_core <dbl> 4, 5, NA, 9, 10, 11, 12, 13, 14, 15, 19, 20, 21, 22, 24, 25, 26, 27, 2
$ sector          <chr> "Other Service Activities", "Professional, Scientific And Technical Ac
$ company_name    <chr> "Snap Election Ltd", "Rental Hotspots Ltd", "Entimoss Ltd", "Proanimal
$ company_profile <chr> "Snap Election is an opinion-sharing social platform giving users in a
$ economy         <chr> "United Kingdom", "United Kingdom", "United Kingdom", "Spain", "United
$ region          <chr> "Northern Europe", "Northern Europe", "Northern Europe", "Southern Eur
$ country_code    <chr> "GBR", "GBR", "GBR", "ESP", "GBR", "ROU", "ITA", "CHE", "GBR", "ESP",
$ global_region   <chr> "Europe & Central Asia", "Europe & Central Asia", "Europe & Central As
$ income_group    <chr> "High income", "High income", "High income", "High income", "High inco
$ question        <chr> "What risks/challenges in using open data in the context of your produ
$ risks_answer    <chr> "Our main challenge is to generate sufficient users; we are solving th
$ text            <chr> "Our main challenge is to generate sufficient users; we are solving th
```

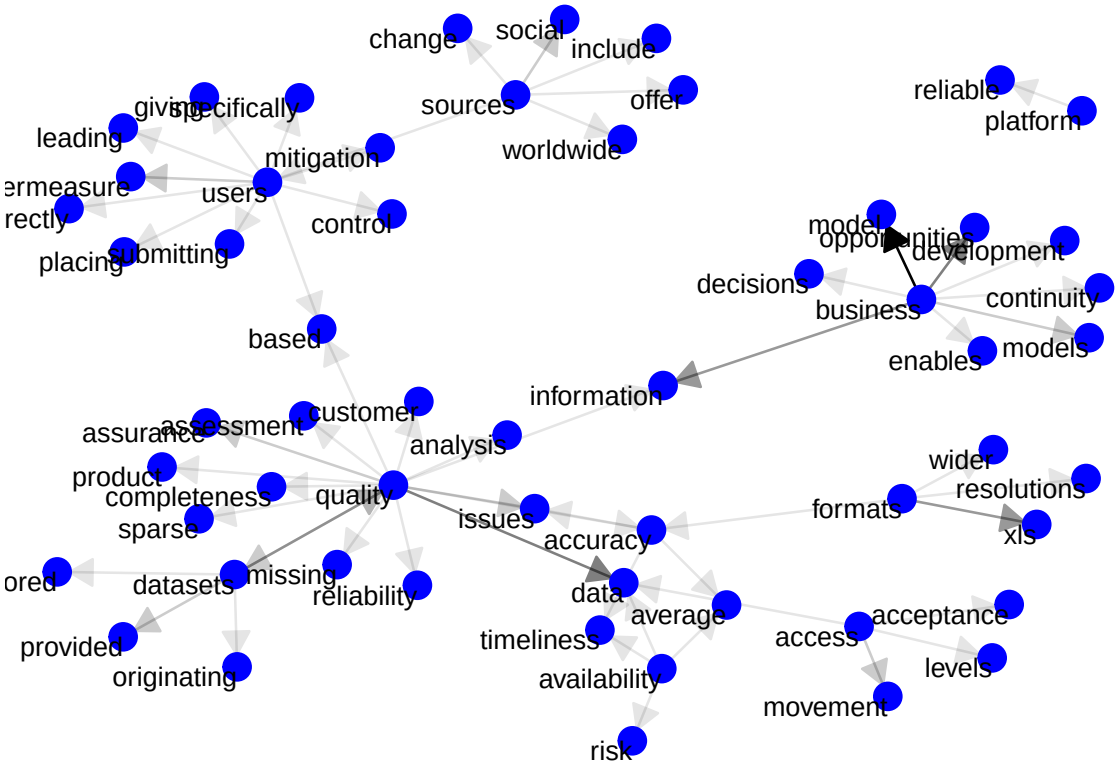
```
tidy_risks_question <- risks_question_data_df %>%
  unnest_tokens(word, `risks_answer`) %>%
  anti_join(stop_words)
```

```
Joining, by = "word"
```

```
#tidy_risks_question %>% glimpse()

rq_bigrams <- risks_question_data_df %>%
  unnest_tokens(bigram, `risks_answer`, token = "ngrams", n = 2)

rq_bigrams_separated <- rq_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

rq_bigrams_filtered <- rq_bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)
```

```
rq_bigrams_counts <- rq_bigrams_filtered %>%
  count(word1, word2, sort = TRUE)

#rq_bigrams_counts %>% glimpse()


rq_bigrams_counts_selection <- rq_bigrams_counts %>%
  filter(word1 %in% c("quality", "users", "datasets", "sources",
                      "formats", "business", "access", "platform",
                      "availability", "accuracy"))


#rq_bigrams_counts_selection
#View(rq_bigrams_counts_selection)

## Let's remove some of these less meaningful words to make a better, more meaningful
risks_mystopwords <- data_frame(word2 = c("i.e", "due", "3", "w.r.t", "e.g",
                                          "don't", "ii", "eu", "e.g", "we're",
                                          "willbe", "google", "output", "we've"))

rq_bigrams_counts_selection <- anti_join(rq_bigrams_counts_selection,
                                         risks_mystopwords, by = "word2")

#rq_bigrams_counts_selection %>% glimpse()
#View(rq_bigrams_counts_selection)

# filter for only relatively common combinations
rq_bigrams_graph <- rq_bigrams_counts_selection %>%
  filter(n >= 2) %>%
  graph_from_data_frame()

#rq_bigrams_graph

#library(ggraph)
#set.seed(2017)

#library(tidygraph)

set.seed(2020)
```

```
a <- grid::arrow(type = "closed", length = unit(.15, "inches"))

ggraph(rq_bigrams_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "blue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()
```



References

Acs, Z. J., Szerb, L., & Lloyd, A. (2017). Enhancing entrepreneurial ecosystems: A GEI approach to entrepreneurship policy. In Z. J. Acs, L. Szerb, & A. Lloyd (Eds.), *Global entrepreneurship and development index 2017* (pp. 81–91). Cham: Springer International Publishing.

Afuah, A. (2002). *Internet business models and strategies: Text and cases* (2nd ed.). New York, NY, USA: McGraw-Hill, Inc.

Aghion, P., & Howitt, P. (1990). A model of growth through creative destruction.

Aldrich, H. (2000). Learning together: National differences in entrepreneurship research. *The Blackwell Handbook of Entrepreneurship*.

Amit, R., & Zott, C. (2001). Value creation in e-business. *Strategic Manage. J.*, *22*(6-7), 493–520.

Applegate, L. M. (2001). E-business models: Making sense of the internet business landscape. *Information Technology and the Future Enterprise: New Models for Managers*, 49–94.

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An r-tool for comprehensive science mapping analysis. *J. Informetr.*, *11*(4), 959–975.

Audretsch, D. B., Thurik, R., Verheul, I., & Wennekers, S. (2006). *Entrepreneurship: Determinants and policy in a European-US comparison*. Springer Science & Business Media.

Barne, D. (2014). What can open data entrepreneurs do for development? `http://blogs.worldbank.org/voices/what-can-open-data-entrepreneurs-do-development`.

Barry, E., & Bannister, F. (2014). Barriers to open data release: A view from the top. *Information Polity*, *19*(1,2), 129–152.

Becker, G. S., Murphy, K. M., & Tamura, R. (1990). Human capital, fertility, and economic growth. *J. Polit. Econ.*, *98*(5, Part 2), S12–S37.

Bedini, I., Farazi, F., Leoni, D., Pane, J., Tankoyeu, I., & Leucci, S. (2014). Open government data: Fostering innovation. *JeDEM - eJournal of eDemocracy and Open Government*, *6*(1), 69–79.

Bellman, R., Clark, C. E., Malcolm, D. G., Craft, C. J., & Ricciardi, F. M. (1957). On the construction of a Multi-Stage, Multi-Person business game. *Oper. Res.*, *5*(4), 469–503.

Benhabib, J., & Spiegel, M. M. (1994). The role of human capital in economic development evidence from aggregate cross-country data. *Journal of Monetary Economics*.

Bernus, P. (2001). Some thoughts on enterprise modelling. *Prod. Plan. Control*, *12*(2), 110–118.

Bettin, G., Lucchetti, R., & Zazzaro, A. (2012). Endogeneity and sample selection in a model for remittances. *J. Dev. Econ.*, *99*(2), 370–384.

Blalock, H. M. (1963). Correlated independent variables: The problem of multicollinearity. *Soc. Forces*, *42*(2), 233–237.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, *55*(4), 77–84.

Bonina, C. M. (2013). New business models and the value of open data: Definitions, challenges and opportunities. *RCUK Digital Economy Theme, Www. Nemode. Ac. Uk/Wpcontent/Uploads/2013/11/Bonina-Opendata-Report-FINAL. Pdf*.

Briel, F. von, Davidsson, P., & Recker, J. (2018). Digital technologies as external enablers of new venture creation in the IT hardware sector. *Entrepreneurship Theory and Practice*, *42*(1), 47–69.

Brynjolfsson, E. (2011). ICT, innovation and the e-economy. *EIB Papers*, *16*(2), 60–76.

Bull, I., & Willard, G. E. (1993). Towards a theory of entrepreneurship. *J. Bus. Venturing*, *8*(3), 183–195.

Buuren, S. van, Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations inR. *J. Stat. Softw.*, *45*(3).

Cabinet Office, U. K. (2012). *Cabinet office open data strategy - data.gov.uk*. Knowledge; Information Management Unit.

Cappgiemini Consulting, U. K. (2013). *The open data economy unlocking economic value by opening government and public data*. Cappgiemini Consulting.

Carlo, J. L., Lyytinen, K., & Boland, R. J. (2012). Dialectics of collective minding: Contradictory appropriations of information technology in a High-Risk project. *Miss. Q.*, *36*(4), 1081–1108.

Carree, M. A., & Thurik, A. R. (2003). The impact of entrepreneurship on economic growth.

In Z. J. Acs & D. B. Audretsch (Eds.), *Handbook of entrepreneurship research* (pp. 437–471). Springer US.

Casson, M. (2005). Entrepreneurship and the theory of the firm. *J. Econ. Behav. Organ.*, *58*(2), 327–348.

Chambers, J. M., & Hastie, T. (1992). *Statistical models in S*. Wadsworth & Brooks/Cole Advanced Books & Software.

Chan, C. M. L. (2013). From open data to open innovation strategies: Creating E-Services using open government data. In (pp. 1890–1899). IEEE.

Chattapadhyay, S. (2013). Towards an expanded and integrated open government data agenda for india. In *ACM international conference proceeding series* (pp. 202–205). Seoul: Association for Computing Machinery.

Chatterjee, S., & Hadi, A. S. (2015). Regression analysis by example.

Chesbrough, H., & Rosenbloom, R. S. (2002). The role of the business model in capturing value from innovation: Evidence from xerox corporation's technology spin-off companies. *Ind Corp Change*, *11*(3), 529–555.

Chesbrough, H., Vanhaverbeke, W., & West, J. (2008). *Open innovation: Researching a new paradigm*. OUP Oxford.

Cole, J. H. (2003). Contribution of economic freedom to world economic growth, 1980-99. *Cato J.*, *23*, 189.

Conradie, P., & Choenni, S. (2012). Exploring process barriers to release public sector information in local government. In *Proceedings of the 6th international conference on theory and practice of electronic governance* (pp. 5–13). New York, NY, USA: ACM.

Crumpton, M. A. (2012). Innovation and entrepreneurship. *The Bottom Line*, *25*(3), 98–101.

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harv. Bus. Rev.*, *90*(10), 70–6, 128.

Davies, T. (2013). Open data barometer: 2013 global report. *World Wide Web Foundation and Open Data Institute*.

Dawes, S. S. (2012). A realistic look at open data. *Center for Technology in Government, University at Albany/SUNY Available at Http://Www. W3. Org/2012/06/Pmod/Pmod2012_submission Pdf*.

De Loo, I., & Soete, L. (1999). The impact of technology on economic growth: Some new ideas and empirical considerations.

Dıaz-Casero, J. C., Dıaz-Aunion, D. A. M., Sanchez-Escobedo, M. C., Coduras, A., &

Hernandez-Mogollon, R. (2012). Economic freedom and entrepreneurial activity. *Management Decision*, *50*(9), 1686–1711.

Dong, G., & Liu, H. (2018). *Feature engineering for machine learning and data analytics*. CRC Press.

Dos Santos Brito, K., Da Silva Costa, M. A., Garcia, V. C., & De Lemos Meira, S. R. (2014). Brazilian government open data: Implementation, challenges, and potential opportunities. In *ACM international conference proceeding series* (pp. 11–16). Aguascalientes: Association for Computing Machinery.

European Commission. (2003). European legislation on reuse of public sector information. `https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information`.

Everitt, B. S. (1984). *An introduction to latent variable models*. Dordrecht: Springer Netherlands.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.*, *49*(1), 92–107.

Feinerer, I., Hornik, K., Software, A., & GPL Ghostscript), I. (. ps T. F. (2015, July). Tm: Text mining package.

Ferreira, J. J., Fayolle, A., Fernandes, C., & Raposo, M. (2017). Effects of schumpeterian and kirznerian entrepreneurship on economic growth: Panel data evidence. *Entrep. Reg. Dev.*, *29*(1-2), 27–50.

Ferro, E., & Osella, M. (2013). Eight business model archetypes for PSI re-use. In *Open data on the web workshop, google campus, shoreditch, london*.

Filippi, P. de, & Maurel, L. (2014). The paradoxes of open data and how to get rid of it? Analysing the interplay between open data and sui-generis rights on databases. *International Journal of Law and Information Technology*, *22*(4), 1–22.

Foundation, R. (2013). R: The R project for statistical computing. `https://www.r-project.org/`.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*. SAGE Publications.

Galper, J. (2001). Three business models for the stock exchange industry. *The Journal of Investing*, *10*(1), 70–78.

Garfield, E., Malin, M. V., & Small, H. (1983). Citation data as science indicators.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. INSTITUTE FOR SCIENTIFIC INFORMATION INC PHILADELPHIA PA; apps.dtic.mil.

Garg, A., & Tai, K. (2013). Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *Int. J. Model. Ident. Control*, *18*(4), 295–312.

Gartner, W. B. (1988). Who is the entrepreneur? Is the wrong question. In *American journal of small business*.

Gartner, W. B. (2001). Is there an elephant in entrepreneurship? Blind assumptions in theory development*. In A. Cuervo, D. Ribeiro, & S. Roig (Eds.), *Entrepreneurship: Concepts, theory and perspective* (pp. 229–242). Berlin, Heidelberg: Springer Berlin Heidelberg.

Gebauer, J., & Ginsburg, M. (2003). The US wine industry and the internet: An analysis of success factors for online business models. *Electronic Markets*, *13*(1), 59–66.

Gelman, A., & Hill, J. (2006, December). Data analysis using regression and multilevel-hierarchical models | statistical theory and methods. `http://www.cambridge.org/mx/academic/subjects/statistics-probability/statistical-theory-and-methods/data-analysis-using-regression-and-multilevelhierarchical-models?format=HB&isbn=9780521867061#vXMTXfss7xmxC1ZR.97`; Cambridge University Press.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *AMJ*, *59*(5), 1493–1507.

Godfrey, L. G. (1978). Testing for multiplicative heteroskedasticity. *J. Econom.*, *8*(2), 227–236.

Goldfarb, A., Greenstein, S. M., & Tucker, C. E. (2015). Introduction to" economic analysis of the digital economy". In *Economic analysis of the digital economy* (pp. 1–17). University of Chicago Press.

Gonzalez-Pernia, J. L., Pe na-Legazkue, I., & Vendrell-Herrero, F. (2012). Innovation, entrepreneurial activity and competitiveness at a sub-national level. *Small Bus. Econ.*, *39*(3), 561–574.

Gordijn, J., & Akkermans, J. M. (2003). Value-based requirements engineering: Exploring innovative e-commerce ideas. *Requirements Engineering*, *8*(2), 114–134.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. U. S. A.*, *101*(suppl 1), 5228–5235.

Grossman, G., & Helpman, E. (1991). Innovation and growth in the global economy MIT press. *Cambridge, MA*.

Gurin, J. (2014). *Open data now: The secret to hot startups, smart investing, savvy marketing, and fast innovation*. McGraw Hill Professional.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*(Mar), 1157–1182.

Gwartney, J. D., Lawson, R. A., & Holcombe, R. G. (1999). Economic freedom and the environment for economic growth. *J. Inst. Theor. Econ.*, *155*(4), 643–663.

Gwartney, J., Lawson, R., Park, W., & Skipton, C. (2005). Economic freedom of the world: 2005 annual report, vancouver: The fraser institute. *Data Retrieved from Www. Freetheworld. Com*.

Gylfason, T., & Zoega, G. (2006). Natural resources and economic growth: The role of investment. *The World Economy*.

Hall, M. A., & Smith, L. A. (1999). Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *FLAIRS conference* (Vol. 1999, pp. 235–239). aaai.org.

Hamel, G., & Ruben, P. (2000). *Leading the revolution* (Vol. 286). Harvard Business School Press Boston, MA.

Harrison, T. M., & Sayogo, D. S. (2014). Transparency, participation, and accountability practices in open government: A comparative study. *Gov. Inf. Q.*, *31*(4), 513–525.

Hassan, M., Hossny, M., Nahavandi, S., & Creighton, D. (2012). Heteroskedasticity variance index. In *2012 UKSim 14th international conference on computer modelling and simulation* (pp. 135–141).

Hebbali, A. (2017). Olsrr: Tools for teaching and learning OLS regression.

Hebert, R. F., & Link, A. N. (1989). In search of the meaning of entrepreneurship. *Small Bus. Econ.*, *1*(1), 39–49.

Hornik, K. (2017, August). Natural language processing infrastructure [r package NLP version 0.1-11]. Comprehensive R Archive Network (CRAN).

Huang, J., Henfridsson, O., Liu, M. J., & Newell, S. (2017). Growing on steroids: Rapidly scaling the user base of digital ventures through digital innovaton. *Miss. Q.*, *41*(1).

Huber, F., Rentocchini, F., & Wainwright, T. (2016). Open innovation: Revealing and engagement in open data organisations franz. *SSRN Electronic Journal*.

Immonen, A., Palviainen, M., & Ovaska, E. (2014). Towards open data based business: Survey on usage of open data in digital services. *International Journal of Research in Business and Technology*, *4*(1).

Ivanova, E., & Gibcus, P. (2003). The decision-making entrepreneur. *Recuperado Junio*, *23*, 2006.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013a). An introduction to statistical learning.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013b). *An introduction to statistical learning: With applications in R*. Springer, New York, NY.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258–268.

John Lu, Z. Q. (2010). The elements of statistical learning: Data mining, inference, and prediction. *J. R. Stat. Soc. Ser. A Stat. Soc.*, *173*(3), 693–694.

Johnson, D. (2001). What is innovation and entrepreneurship? Lessons for larger organisations. *Industrial and Commercial Training*, *33*(4), 135–140.

Jon-Arild Johannessen, Bjørn Olsen, & G.T. Lumpkin. (2001). Innovation as newness: What is new, how new, and new to whom? *Euro Jrnl of Inn Mnagmnt*, *4*(1), 20–31.

Kaasenbrood, M., Zuiderwijk, A., Janssen, M., De Jong, M., & Bharosa, N. (2015). *Exploring the factors influencing the adoption of open government data by private organisations*. IGI Global.

Kapetanios, G., Marcellino, M. G., & Papailias, F. (2014, June). *Variable selection for large unbalanced datasets using Non-Standard optimisation of information criteria and variable reduction methods*.

Karimi, J., & Walter, Z. (2015). The role of dynamic capabilities in responding to digital disruption: A Factor-Based study of the newspaper industry. *Journal of Management Information Systems*, *32*(1), 39–81.

Katrutsa, A., & Strijov, V. (2017). Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Syst. Appl.*, *76*, 1–11.

Katz, R., Koutroumpis, P., & Fernando, M. C. (2014). Using a digitization index to measure the economic and social impact of digital agendas. *Info*, *16*(1), 32–44.

Ken Farr, W., Lord, R. A., & Wolfenbarger, J. L. (1998). Economic freedom, political freedom, and economic Well-Being: A causality analysis. *Cato J.*, *18*, 247.

Kenney, M., & Zysman, J. (2016). The rise of the platform economy. *Issues Sci. Technol.*, *32*(3), 61.

Kirzner, I. M. (1973). *Competition and entrepreneurship*. University of Chicago Press.

Kitsios, F., Papachristos, N., & Kamariotou, M. (2017). Business models for open data ecosystem: Challenges and motivations for entrepreneurship and innovation. In *2017 IEEE 19th conference on business informatics (CBI)* (Vol. 1, pp. 398–407). ieeexplore.ieee.org.

Klaus Schwab, W. E. F. (2017). *The global competitiveness report 2017–2018 - www3.weforum.org* . (No. NA). World Economic Forum.

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R*. Springer Science & Business Media.

Korez-Vide, R., & Tominc, P. (2016). Competitiveness, entrepreneurship and economic growth. In *Competitiveness of CEE economies and businesses* (pp. 25–44). Springer, Cham.

Kotabe, M., & Scott Swan, K. (1995). The role of strategic alliances in high-technology new product development. *Strat. Mgmt. J.*, *16*(8), 621–636.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, *1*(2), 111–117.

Kraemer, K. L., Dedrick, J., & Yamashiro, S. (2000). Refining and extending the business model with information technology: Dell computer corporation. *The Information Society*, *16*(1), 5–21.

Kreft, S. F., & Sobel, R. S. (2005). Public policy, entrepreneurship, and economic freedom. *Cato J.*, *25*, 595.

Krueger, C., Swatman, P., & Beek, K. van der. (2004). New and emerging business models for online news: A survey of 10 european countries. *BLED 2004 Proceedings*, 28.

Kuffner, T. A., & Walker, S. G. (2017). Why are p-values controversial? *Am. Stat.*, 0–0.

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

Lab, N. G. (2014). Open data 500. `http://www.opendata500.com/`.

Lakomaa, E., & Kallberg, J. (2013). Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs. *IEEE Access*, *1*, 558–563.

Laudien, S. M., Bouncken, R., & Pesch, R. (2018). Understanding the acceptance of digitalization-based business models: A qualitative-empirical analysis. *Global Proc*, *Surrey*(2018), 104.

Lee, M., Almirall, E., & Wareham, J. (2015). Open data and civic apps: First-generation failures, second-generation improvements. *Commun. ACM*, *59*(1), 82–89.

LeHong, H. (2019). *Digital business overview: Major frameworks in one report* (No. G00405058). Gartner Research.

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access*, *5*, 7776–7797.

Linder, J. C., & Cantrell, S. (2000). Changing business models: Surveying the landscape | sci-napse | academic search engine for paper. `https://scinapse.io/papers/1584872405`.

Liu, H., & Motoda, H. (1998). *Feature extraction, construction and selection: A data mining perspective*. Norwell, MA, USA: Kluwer Academic Publishers.

Lumpkin, G. T., & Dess, G. G. (1996). Clarifying the entrepreneurial orientation construct and linking it to performance. *Acad. Manage. Rev.*, *21*(1), 135–172.

Lüdecke, D., Makowski, D., & Waggoner, P. (2019). Performance: Assessment of regression models performance. *R Package Version 0. 4*, *2*.

Lyytinen, K., Sørensen, C., & Tilson, D. (2017). Generativity in digital infrastructures. In R. D. Galliers & M.-K. Stein (Eds.), *The routledge companion to management information systems* (1st ed., pp. 253–275). Abingdon, Oxon ; New York, NY : Routledge, 2017.: Routledge.

MacInnes, I., Moneta, J., Caraballo, J., & Sarni, D. (2002). Business models for mobile content: The case of M-Games. *Electronic Markets*, *12*(4), 218–227.

Magalhaes, G., Roseira, C., & Manley, L. (2014). Business models for open government data. In *Proceedings of the 8th international conference on theory and practice of electronic governance* (pp. 365–370). New York, NY, USA: ACM.

Magretta, J. (2002). Why business models matter. *Harv. Bus. Rev.*, *80*(5), 86–92, 133.

Mahadevan, B. (2000). Business models for Internet-Based E-Commerce: An anatomy. *Calif. Manage. Rev.*, *42*(4), 55–69.

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). Open data: Unlocking innovation and performance with liquid information. *McKinsey Global Institute*, 21.

Martin, K. (2018). The penalty for privacy violations: How privacy violations impact trust online. *J. Bus. Res.*, *82*, 103–116.

Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. (2013). Open data: Barriers, risks and opportunities. In Castelnovo, W and Ferrari, E (Ed.), *PROCEEDINGS OF THE 13TH EUROPEAN CONFERENCE ON EGOVERNMENT* (pp. 301–309). CURTIS FARM, KIDMORE END, NR READING, RG4 9AY, ENGLAND: ACAD CONFERENCES LTD.

Mergel, I., Kattel, R., Lember, V., & McBride, K. (2018). Citizen-oriented digital transformation in the public sector. In *Proceedings of the 19th annual international conference on digital government research: Governance in the data age* (pp. 1–3). New York, NY, USA: Association for Computing Machinery.

Miller, T., & Kim, A. B. (2013). Defining economic freedom. *Miller AT, Holmes KR, Feulner EJ (Eds)*, 87–94.

Morando, F. (2013). Legal interoperability: Making open government data compatible with businesses and communities. *JLIS. It*, *4*(1), 441.

Mueller, P. (2007). Exploiting entrepreneurial opportunities: The impact of entrepreneurship on growth. *Small Bus. Econ.*, *28*(4), 355–362.

Nakamura, A., & Nakamura, M. (1998). Model specification and endogeneity. *J. Econom.*, *83*(1), 213–237.

OECD. (1998). *Fostering entrepreneurship*. OECD Publishing.

OECD. (2016). *OECD digital government studies open government data review of mexico data reuse for public sector impact and innovation: Data reuse for public sector impact and innovation*. OECD Publishing.

Ooms, J. (2017). Pdftools: Text extraction, rendering and converting of PDF documents.

Open Data Institute. (2015). Research: Open data means business. `http://theodi.org/open-data-means-business`.

Open Data Institute, Lateral Economics. (2016). Research: The economic value of open versus paid data | open data institute. `https://theodi.org/research-economic-value-open-paid-data`.

Osareh, F. (1996). Bibliometrics, citation analysis and Co-Citation analysis: A review of literature I. *Libri*, *46*(3), 102.

Osterwalder, A., & Pigneur, Y. (2004). An ontology for e-business models. *Value Creation from E-Business Models*, *1*, 65–97.

Osterwalder, A., Pigneur, Y., & Tucci, C. L. (2005). Clarifying business models: Origins, present, and future of the concept. *Communications of the Association for Information Systems*, *16*(1).

Ovans, A. (2015). What is a business model? *Harvard Business Review*.

Pateli, A. (2003). A framework for understanding and analysing ebusiness models. *BLED 2003 Proceedings*, 4.

Pato, M. L., & Teixeira, A. A. C. (2016). Twenty years of rural entrepreneurship: A bibliometric survey. *Sociol. Ruralis*, *56*(1), 3–28.

Peter B. Seddon, T. U. of M., Geoffrey P. Lewis, M. B. S., Phil Freeman, S. U. of T., & Graeme Shanks, U. of M. (2004). The case for viewing business models as abstractions of strategy. *Communications of the Association for Information Systems*, *13*(1), 25.

Petrovic, O., Kittl, C., & Teksten, R. D. (2001, October). *Developing business models for ebusiness*.

Phethean, C., Simperl, E., Tiropanis, T., Tinati, R., & Hall, W. (2016). The role of data science in web science. *IEEE Intell. Syst.*, *31*(3), 102–107.

Pooran Wynarczyk, Panagiotis Piperopoulos, & Maura McAdam. (2013). Open innovation in small and medium-sized enterprises: An overview. *Int. Small Bus. J.*, *31*(3), 240–255.

Porter, M. E. (1986). *Competition in global industries*. Harvard Business Press.

Porter, M. E. (1998). The competitive advantage of nations. *Pak. Dev. Rev.*, *37*(1), 90–94.

Powell, B. (2002). Economic freedom and growth: The case of the celtic tiger. *Cato J.*, *22*, 431.

Rao, C. R., & Toutenburg, H. (1995). Linear models. In *Springer series in statistics* (pp. 3–18).

Reynolds, P., Bosma, N., Autio, E., Hunt, S., De Bono, N., Servais, I., ... Chin, N. (2005). Global entrepreneurship monitor: Data collection design and implementation 1998–2003. *Small Bus. Econ.*, *24*(3), 205–231.

Rocha, H. O. (2004). Entrepreneurship and development: The role of clusters. *Small Bus. Econ.*, *23*(5), 363–400.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*.

Sandoval-Almazan, R., Gil-Garcia, J. R., Luna-Reyes, L. F., Luna, D. E., & Rojas-Romero, Y. (2012). Open government 2.0: Citizen empowerment through open data, web and mobile apps. In *Proceedings of the 6th international conference on theory and practice of electronic governance* (pp. 30–33). New York, NY, USA: ACM.

Schumpeter, J. A. (1934). *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle*. Transaction Publishers.

Schumpeter, J. A. (1939). *Business cycles* (Vol. 1). McGraw-Hill New York.

Schumpeter, J. A. (1942). *Socialism, capitalism and democracy*. Harper; Brothers.

Selander, L., & Jarvenpaa, S. L. (2016). Digital action repertoires and transforming a social movement organization. *Miss. Q.*, *40*(2), 331–352.

Sepulveda, F., & Mendez, F. (2005). Optimal government regulations and red tape in an economy with corruption. *SSRN Electronic Journal*.

Shane, S. (2000). The promise of entrepreneurship as a field of research. *Acad. Manage. Rev.*, *25*(1), 217–226.

Shubar, A., & Lechner, U. (2004). The public WLAN market and its business models-an empirical study. In *Proceedings of the 17th bled eCommerce conference*. domino.fov.uni-mb.si.

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R. *The Journal of Open Source Software*, *1*(3), 37.

Slappendel, Carol. (1996). Perspectives on innovation in organizations. *Organization Studies*, *17*(1), 107–129.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269.

Smith, G., Ofe, H. A., & Sandberg, J. (2016). Digital service innovation from open data: Exploring the value proposition of an open data marketplace. In *2016 49th hawaii international conference on system sciences (HICSS)* (pp. 1277–1286). ieeexplore.ieee.org.

Solow, R. M. (1956). A contribution to the theory of economic growth. *Q. J. Econ.*, *70*(1), 65–94.

Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics*.

Srinivasan, A., & Venkatraman, N. (2018). Entrepreneurship in digital platforms: A network-centric view. *Strategic Entrepreneurship Journal*, *12*(1), 54–71.

Srivastava, A. N., & Sahami, M. (2009). *Text mining: Classification, clustering, and applications*. Chapman; Hall/CRC.

St"ahler, P. (2002). Business models as an unit of analysis for strategizing. In *International workshop on business models, lausanne, switzerland* (Vol. 45, pp. 2990–2995).

Stevenson, H. H., & Jarillo, J. C. (1990). A paradigm of entrepreneurship: Entrepreneurial management. *Strategic Manage. J.*, *11*, 17–27.

Stott, A. (2014). *Open data for economic growth (transport & ICT global practice). Washington, DC: The world bank* (No. 89606). The World Bank.

Swan, T. W. (1956). ECONOMIC GROWTH and CAPITAL ACCUMULATION. *Economic Record*.

Szerb, L., Aidis, R., & Acs, Z. J. (2013). The comparison of the global entrepreneurship monitor and the global entrepreneurship and development index methodologies. *Foundations and Trends in Entrepreneurship*, *9*(1), 1–142.

Takagi, S. (2014). Research note: An introduction to the economic analysis of open data. *Rev Socionetwork Strat*, *8*(2), 119–128.

Tan, C.-M., Wang, Y.-F., & Lee, C.-D. (2002). The use of bigrams to enhance text categorization. *Inf. Process. Manag.*, *38*(4), 529–546.

The World Bank Group. (2013). Open data essentials | data. `http://opendatatoolkit.worldbank.org/en/essentials.html`.

Thorhildur Jetzek, M. A. A. N. B.-A. (2013). Generating value from open government data. In *ICIS, at http://aisel.aisnet.org/icis2013/proceedings/GeneralISTopics/5/, volume: 2013* (Vol. 2).

Timmers, P. (1998). Business models for electronic markets. *Electronic Markets*, *8*(2), 3–8.

Tinati, R., Carr, L., Halford, S., & Pope, C. (2012). Exploring the impact of adopting open data in the UK government. In (p. 3).

Ubaldi, B. (2013). *Open government data towards empirical analysis of open government data initiatives*.

Van Praag, C. M. (1999). Some classic views on entrepreneurship. *Economist*, *147*(3), 311–335.

Varian, H. (2018, July). *Artificial intelligence, economics, and industrial organization*. National Bureau of Economic Research.

Verbeek, M. (2017). *A guide to modern econometrics 5th edition*. (Wiley, Ed.). Wiley.

Verheul, I., Wennekers, S., Audretsch, D., & Thurik, R. (2002). An eclectic theory of entrepreneurship: Policies, institutions and culture. In U. S. Springer (Ed.), *Economics of science, technology and innovation* (pp. 11–81).

Vesper, K. H. (1988). Entrepreneurial academics—how can we tell when the field is getting somewhere? *J. Bus. Venturing*, *3*(1), 1–10.

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, *28*(2), 118–144.

Vijayarani, S., & Janani, R. (2016). Text mining: Open source tokenization tools – an analysis. *ACII*, *3*(1), 37–47.

Vrande, V. van de, Jong, J. P. J. de, Vanhaverbeke, W., & Rochemont, M. de. (2009). Open innovation in SMEs: Trends, motives and management challenges. *Technovation*, *29*(6-7), 423–437.

Wainwright, T., Huber, F., & Rentocchini, F. (2014). Open wide? Business opportunities and risks in using open data. In. eprints.soton.ac.uk.

Walker, J., Simperl, E., Capgemini, A., agent. European Data Portal:European Data Portal, & corporate-body. PUBL:Publications Office. (2020). *Open data and entrepreneurship* (No. 10). European Data Portal; Publications Office of the European Union.

Weill, P., & Vitale, M. (2002). What IT infrastructure capabilities are needed to implement e-business models. *MIS Quarterly Executive*, *1*(1), 17–34.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *J. R. Stat. Soc. Ser. C Appl. Stat.*, *22*(3), 392–399.

Williams, S. (1967). Business process modeling improves administrative control. *Automation, December*, *44*, 50.

Wipo, I., Cornell University. (2017). *The global innovation index 2017* (No. 10). Cornell University,

Wortmann, J. C., Hegge, H. M. H., & Goossenaerts, J. B. M. (2001). Understanding enterprise modelling from product modelling. *Prod. Plan. Control*, *12*(3), 234–244.

Yu, C.-C. (2016). A value-centric business model framework for managing open data applications. *JOURNAL OF ORGANIZATIONAL COMPUTING AND ELECTRONIC COMMERCE*, *26*(1-2, SI), 80–115.

Zarembka, P. (1990). Transformation of variables in econometrics. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Econometrics* (pp. 261–264). London: Palgrave Macmillan UK.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators, (10), 17.

Zeleti, F. A., & Ojo, A. (2017). The 6-values open data business model framework. In A. Ojo & J. Millard (Eds.), *Government 3.0 – next generation government technology infrastructure and services: Roadmaps, enabling technologies & challenges* (pp. 219–239). Cham: Springer International Publishing.

Zeleti, F. A., Ojo, A., & Curry, E. (2016). Exploring the economic value of open government data. *Gov. Inf. Q.*, *33*(3), 535–551.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1), 43–52.

Zhao, F. (2005). Exploring the synergy between entrepreneurship and innovation. *International Journal of Entrepreneurial Behavior & Research*, *11*(1), 25–41.

Zilibotti, F., Aghion, P., Howitt, P., & Garcia-Penalosa, C. (1999). Endogenous growth theory. *The Canadian Journal of Economics / Revue canadienne d'Economique*.

Zimmermann, H.-D., & Pucihar, A. (2015). *Open innovation, open data and new business models* (No. ID 2660692). Rochester, NY: Social Science Research Network.

Zuiderwijk, A., Helbig, N., Gil-Garcia, J. R., & Janssen, M. (2014). Special issue on innovation through open data - a review of the state-of-the-art and an emerging research agenda: Guest editors' introduction. *Journal of Theoretical and Applied Electronic Commerce Research*, *9*(2), I–XIII.

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Gov. Inf. Q.*, *31*(1), 17–29.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Sheikh_Alibaks, R. (2012a). Socio-Technical impediments of open data. *ResearchGate*, *10*(2), 156–172.

Zuiderwijk, A., Jeffery, K., & Janssen, M. (2012b). The potential of metadata for linked open data and its value for users and publishers. *JeDEM - eJournal of eDemocracy and Open Government*, *4*(2), 222–244.