NOVA
IMS
Information
Management
School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Break On Through (to the other side) of Political Connections: An exploratory Empirical Analysis of Portuguese Companies

<Data Science>

Gabriel Ravi de Sousa dos Santos

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in < Data Science and Advanced Analytics with major in Data Science>

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Break On Through (to the other side) of Political Connections: An exploratory Empirical Analysis of Portuguese Companies

I would like to dedicate this thesis to my parents that always believed in me, Vera and Wilton. To my friends that always supported me, Cave and Schuberry. To professor Josemberg Andrade that was the first person to convinced me to do a Masters Degree and to my supervisors Bruno and David.

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Professor Bruno Damásio, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I would also like to thank my tutor, David Neres, for his valuable guidance throughout my studies. You provided me with the tools that I needed to choose the right direction and successfully complete my dissertation.

I would like to acknowledge my colleagues from the Masters, Abdallah Zaher and Cristina Mousinho for their wonderful collaboration. They're the best dream team anyone can imagine.

And finally, I would like to thank my father, Wilton Santos. You have never doubt me for a single second and you are the most important person to me in the world.

"Bem, a vida é feita de mudanças. Temos que guardar a fase anterior com carinho e com boas lembranças, mas temos que encarar a nova fase com esperança de dias melhores. E só depende de nós fazer com que todas as novas fases da nossa vida sejam melhores do que todas as anteriores." (Wilton José Pereira dos Santos, my father)

Abstract

Transparency International's Global Corruption Barometer 2013 reveals that political parties, Parliament, the judiciary and the military are the most corrupt institutions in Portugal. Transparency International's 2017 Corruption Perception Index ranks the country 29th place out of 180 countries.

Although politicians have a great influence on company profits, there is no such research in Portugal that examines the impact of political connections on firms. This thesis tries to complete this task differently.

With the information publicized on the Portuguese parliament website, databases that contain information about the public employees, and one of the Orbis Van Dijk databases about the Portuguese companies, this paper tries to evaluate the performance of these companies compared to the ones that don't have connections.

This dissertation describes the difference between companies that has government connections and compare to companies that don't. The dissertation explains the origin of each database explains the keys to merge them, and then compares each one of them. In the end, the reader will be able to understand their behavior and how private companies with government connections have an advantage compared to the market.

This thesis in among the first ones to examine the influence of political power connections in Portugal.

Keywords: Big Data. Analytics. Hyphotesis Tests. Advanced Statistics.

Resumo

O Barómetro de Corrupção Global 2013 da Transparency International revela que os partidos políticos, o Parlamento, o sistema judiciário e os militares são as instituições mais corruptas em Portugal. O Índice de Percepção de Corrupção de 2017 da Transparency International classifica o país em 29º lugar entre 180 países.

Embora os políticos tenham uma grande influência nos lucros das empresas, não existe em Portugal qualquer investigação que examine o impacto das ligações políticas nas empresas. Esta dissertação tenta completar essa tarefa de maneira diferente.

Com a informações divulgadas no site do parlamento português com bases de dados públicas que contém informações sobre os funcionários públicos, e uma das bases de dados Orbis Van Dijk sobre as empresas portuguesas e seus respectivos desempenhos ao longo dos anos, esta tese procura avaliar o desempenho destas empresas face às que não o possuem conexões.

Esta dissertação descreve a diferença entre empresas que têm ligações com o governo e compara com empresas que não têm. A dissertação explica a origem de cada banco de dados, explica as chaves para mesclá-los e, a seguir, compara cada um deles. Ao final, o leitor poderá entender seu comportamento e como empresas privadas com vínculo governamental levam vantagem em relação ao mercado.

Palavras-chave: Big Data. Analytics. Teste de Hipótese. Estatística Avançada.

Contents

List of Figures

List of Tables

Glossary

CIG - The companies without connections inside the government 28

DBScan - Density-based spatial clustering of applications with noise 41

EIG - The companies with connections inside the government 30

Chapter 1

Introduction

The productivity of each government is highly dependent on the quality of the government parliament. The potency of each country is dependent on the private companies and the spirit of individuals employed in the public sector is a crucial determinant of government.

Existing literature suggests that dominant political leaders use their power to the advantage of favored firms, which in turn benefit from government-created rents. Empirical studies around the world support the view that benefits extracted by connected firms are significant. Political connections will offer the beneficiary lucrative government concessions, monopolies, licenses, government contracts, tax breaks, and easy access to bank loans, all of which protect the beneficiary from competition in their domestic market.

In Portugal, companies face an overall moderate risk of corruption when doing business. Corruption and abuse of power are most prevalent, in the areas of urban planning and public procurement. The Portuguese Criminal Code makes it illegal to give or accept a bribe, and the Law on Corruption in International Commerce and the Private Sector (in Portuguese) establishes the terms of liability for corruption offenses in international trade and private activities. Individuals and companies are criminally liable for corruption offenses, including bribery of foreign public officials in international commerce. Facilitation payments are prohibited, and gifts and hospitality may be considered illegal depending on their intent. While the country has made significant progress in the past decade, recurring corruption scandals involving high-level politicians, local administrators, and businesses abusing public funds have revealed that safeguards to counter corruption, and abuses of power have been somewhat inefficient in Portugal.

As in other developed countries, such a political connection is important in Portugal. Companies, no matter whether they are private enterprises, can only grow into giants and record decent earning growths if they can build certain political connections with dominant leaders for the long period of involvement of the government in the economy.

This dissertation combines almost 40 years of micro data from Portugal on around 154.682 companies since their foundations with public employees databases (deputies, senators, public employees, ...)

We will combine each of two databases with the public employees and one with every private companies that was active in Portugal. The parliament data set contains all the 230 elected deputies and their partners in 2018 and a public database that contains all the 7166 employees that have worked in the public sector since 1976. We will connect and merge these two previous databases with a database that has every employee and enterprise that has ever been in Portugal.

The first and most difficult challenge was to clean, transform and manage the Orbis Database (private companies database). We worked with more than 400 different variables with more than 10 million rows. The collected Orbis database has more than a thousand diverse files and formats.

After deep cleaning and processing in all the databases, we will apply five different merge techniques to detect the common names between them. The merging technique apply different sampling and statistical measures; after it, we will make a double check in each politician to make sure that he was matched on the right year. For example, a politician elected in 1990 won't be merged with a company that was founded in 2015.

The profile of each connection is important to be analyzed because we will be able to identify some patterns in this connection system.

In chapter 2 we will review the Literature Review. The chapter 3 is the methodology and it will explain the origin and development of each database.

The descriptive analysis, in chapter 4, will present and compare each gender, political party, business segment, and more. The reader will understand the behavior of these unusual connections. The chapter 5 will compare econometric variables between companies with connections with companies without connections. And Finally, the chapter 6 will give final conclusions and insights for future works.

This thesis in among the first ones to examine the influence of political power connections in Portugal.

1.1 Research Questions

To evaluate the impact of enterprises with public connections in a economy, the research questions that will be answered:

- **In Portugal, is it possible to infer that companies with public connections have, economically, a better performance than companies without connections?**
- **What are the main profile clusters of the privileged companies in Portugal? And what are their behavior compared to the companies with political connections?**
- **Talking about the companies with public connections, what is their importance on the economy? In other words, in the last decade, how many percentages of the national profits are exclusive to them**

1.2 Literature Review

You can check in Ramalho, 2003 the relation between the impeachment of president Fernando Collor de Melo in Brazil and the performance of the private companies that had connections with the president. He was president 1990 until 1992 and he was a right handed president known by his private connections. This work talk about two types of politically connected companies: Enterprises proven to be connected to him in a parliamentary investigation and companies owned by friends and relatives of the impeached President Fernando Collor. Using an event study procedure, it establishes that companies with private connections have on average significantly negative daily abnormal returns while the competitors companies have positive abnormal returns on these dates. And more, these companies do not experience a significant decline in their stock market valuation during the impeachment process while their competitors do not react positively to the episode. Even though the stock prices of firms connected to the president through family or friends fell initially, this decline was reversed entirely within a year of the impeachment. This paper is going to analyze not only one presidential mandate, but the whole parliament through more than 10 year. Here we will analyze the Portuguese parliament instead of the Brazilian and we will have to opportunity to measure the economic impact based on eight indicators.

Another important paper that will guide this thesis is the Faccio, 2006. It examine, for more than 42 countries, firms with controlling shareholders and top managers who are members of national parliaments or governments.The paper shows that connected companies enjoy easy access to debt financing, low taxation, and higher market share. These benefits are particularly pronounced when companies are connected through their owner, a seasoned politician, or a minister. Benefits are generally greater when connected firms operate in countries with higher degrees of corruption, resulting in a significant increase in value. A five-day event study around announcements of directors or dominant shareholders entering politics and of politicians joining boards documents that connections result in a significant increase in value when companies operate in highly corrupt countries, thus reflecting the great benefits they obtain. Connections do not add value to firms operating in countries with low levels of corruption; benefits are marginal. Even that we will analyse different variables (Profit Margin, Operating Revenue, and more), in another period of time (from 2010 until 2020), in this thesis we will design and compare the impact of public connections with exclusively Portuguese private companies and their performances. In other words, we will be able to understand and see if Portugal evolved positively since Mara Faccio article was published.

And finally, one of the papers that inspired this thesis was Shleifer and Vishny, 1994. "Politicians and Firms"presents a model of bargaining between politician and private company managers and the consequences about their commercialization, and privatization. Like Brazil, in many countries is easier to direct resources towards favored firms. Based on the assumptions that politicians cater to interest groups rather than the median voter, and that the relationship between politicians and managers is governed by incomplete contracts, they derive implications of bargaining between politicians and enterprise managers over what enterprises do. Here we are going to measure with the same parameters the impact of the biggest companies in the country economy performance. And to complement the analysis, we are going to describe how many percentage of this companies are represented in the total performance of the country.

Chapter 2

Theoretical Background

The basic statistics sections were based on Bussab and Morettin, 2010.

2.1 Machine Learning

Machine Learning is an Application of Artificial Intelligence (AI) it gives devices the ability to learn from their experiences and improve their self without doing any coding. For Example, when you shop from any website it's shows related search like:- People who bought also saw this.

Machine learning (ML) algorithms are broadly categorized as either supervised or unsupervised. Supervised learning algorithms have both input data and desired output data provided for them through labeling, while unsupervised algorithms work with data that is neither classified nor labeled. An unsupervised algorithm might, for example, group unsorted data according to similarities and differences.

However, many ML approaches, including transfer learning and active learning, involve what are more accurately described as semi-supervised algorithms. Transfer learning uses knowledge gained from completing one task to help solve a different but related problem, while active learning allows an algorithm to query the user or some other source for more information. Both systems are commonly used in situations where labeled data is scant.

Reinforcement learning, sometimes considered a fourth category, is based on rewarding desired behaviors and/or punishing undesired ones to direct unsupervised machine learning through rewards and penalties.

2.1.1 Supervised, Unsupervised and Reinforcement Learning

In supervised learning, the algorithms try to generalize patterns in the data based on a target variable (James et al., 2013). The learned patterns can then be used for predicting the value of the target variable for observations where the value of the target variable is unknown. In other words, for each observation of the predictor measurement, there is an associated response measurement. A machine learning model, which is fitted on data, relates the response to the predictors, intending to accurately predict the response for future observations or better understanding the relationship between the response and the predictors. Moreover, supervised learning can be separated into regression and classification. In a regression machine learning model, the algorithm learns and predicts the value of a numeric target variable. An example would be the regression of the age of a person, based on variables describing the person. In classification, the target value is categorical, e.g. the gender.

In contrast to supervised learning, in unsupervised learning, a target variable does not exist (James et al., 2013). Unsupervised learning algorithms seek to understand relations between variables or observations by observing the values of these variables so that they can be separated into groups. Grouping similar variables or grouping similar observations is possible. This can be used for grouping customers who have similar buying behaviors.

Reinforcement Learning is a reward- and penalty-based learning system (Mitchell, 1997). The learning algorithm tries to solve a specific task and receives positive and negative feedback for respective solutions. Regarding the feedback the algorithm receives, it can figure out ways to solve the desired task by itself, inside the controlled environment.

2.2 Hypothesis Test

In statistics, hypothesis testing is the method used to verify whether data are compatible with a hypothesis, which can often suggest the non-validity of a hypothesis. It is based on the analysis of a sample, through probability theory, used to evaluate certain parameters that are unknown in a population. Most of the time it consists of two hypotheses:

H_0 : Null Hypothesis to be tested

H_1 : Alternative Hypothesis

Significance Level is the probability with which it is liable to run the risk of a Type I Error (error of rejecting a given hypothesis that it is true). It is usually identified by the Greek letter α and is determined before sample extraction. For example, if we use a significance level equal to 0.05 it means that there is a probability of 5 in 100 that the hypothesis is rejected when it should be accepted, that is, a 95% chance of having made a correct decision.

P-value, in classical statistics, the p-value or descriptive level is a statistic used to synthesize the result of a hypothesis test. Formally, the p-value is defined as the probability of obtaining a test statistic equal to or more extreme than that observed in a sample, assuming true the null hypothesis H_0. If the p-value is less than the established significance level, the null hypothesis is rejected. Otherwise, H0 is not rejected, and the probability that it is true is acceptable.

2.2.1 T-test

Consider two samples $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ and these samples are must be paired, in other words, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Be $D_i = x_i - y_i$ for $i = 1, 2, ..., n$. The sample $D_1, D_2, ..., D_n$ is obtained from the differences between the values of each pair. The assumption is that the population of differences follows Normal distribution with mean μ_D and variance σ_D^2

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

where $\mu_D = \mu_1 - \mu_2$, be μ_1 - μ_2 the averages to the population X and Y, respectively. The formulation of hypotheses varies according to what is desired

The parameters sample mean μ_D and sample variance σ_D^2:

$$\bar{D} = \textstyle\sum_{i=1}^{n} D_i$$
$$S_D^2 = 1/(n-1) \textstyle\sum_{i=1}^{n} (D_i - \bar{D})^2$$

The statistic $\bar{D}$ has Gaussian distribution with mean μ_D and variance σ_D^2/n.

The statistics will have the Student's t-distribution with $(n-1)$ degrees of freedom with values being calculated with:

$$T = \frac{n^{1/2}(\bar{D} - \mu_D)}{S_D} \sim t_{n-1} \tag{2.1}$$

with p-value $= 2xP(|T| > t)$

If the p-value is lower than the significance level, the null hypothesis is rejected.

2.2.2 Shapiro Wilk test

The Shapiro Wilk Test is used to verify the adherence of any quantitative variable to the Normal Distribution model. This test is more suitable when the sample is smaller than 30. The assumption of normality is important to determine the test to be used. The beginning of the Shapiro Wilk Test procedure begins with the definition of hypotheses:

H_0 : The variable is normally distributed

H_1 : The variable is not normally distributed

Then the sample must be sorted in ascending order to obtain order statistics. The statistics is defined by:

$$W = \frac{1}{\partial}\left[\sum_{i=1}^{k} a_i \left(X^{(n-1+1)} - X^{(s)}\right)\right]^2 \tag{2.2}$$

where:

$K \sim n/2$; X^i is the i-order and $D = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ e $\bar{X} = \frac{1}{n}\sum_{i=1}^{x} X_i$

The decision can be taken through the critical value already tabulated or by the p-value, calculated by some statistical software. If the p-value is less than the significance level adopted, the null hypothesis is rejected.

2.2.3 Bartlet test

Bartlett's test (Snedecor and Cochran, 1983) is used to test if k samples have equal variances. Equal variances across samples is called homogeneity of variances. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Bartlett test can be used to verify that assumption.

The Bartlett test is defined as:

$$H_0 : \sigma_1 = \sigma_2 ... \sigma_N$$

$$H_1 : \sigma_i \neq \sigma_j \text{ for at least one pair (i,j).}$$

The Bartlett test statistic is designed to test for equality of variances across groups against the alternative that variances are unequal for at least two groups.

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i - 1)\ln s_i^2}{1 + (1/(3(k-1)))\left(\left(\sum_{i=1}^{k} 1/(N_i - 1)\right) - 1/(N-k)\right)} \tag{2.3}$$

In the above, s_i^2 is the variance of the i group, N is the total sample size, N_i is the sample size of the i group, k is the number of groups, and s_i^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$s_p^2 = \sum_{i=1}^{k} (N_i - 1)s_i^2/(N-k) \sim T > \chi^2_{1-\alpha,k-1} \tag{2.4}$$

with p-value = $2P(|\chi| > x)$

If the p-value is lower than the significance level, the null hypothesis is rejected.

2.2.4 Mann-Whitney test

The Mann Whitney Wilcoxon test is a non-parametric test that aims to verify whether two random samples X and Y have the same distribution function. As a result, this test can also be used to compare means between dependent samples.

Consider two samples $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ and these samples are must be paired, in other words, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Be $D_i = x_i - y_i$ for $i = 1, 2, ..., n$. The sample $D_1, D_2, ..., D_n$ is obtained from the differences between the values of each pair. The assumption is that the population of differences follows Normal distribution with mean μ_D and variance σ_D^2

$$H_0 : \mu_x - \mu_y = \mu_D = 0 H_1 : \mu_x \neq \mu_y \neq 0 \text{ or } H_1 : \mu_x > \mu_y \text{ ou } H_1 : \mu_x < \mu_y \tag{2.5}$$

where μ_x is the sample of X and μ_y is the sample of Y.

The test is made from the ordering of the variable D_i and ranks (or ranks) are assigned to each observation. Some observations can be assigned the same rank in the order. This phenomenon is called a tie. The test statistic is calculated as follows:

where $R(x_i, y_i)$ is the rank. So, if there are no matching values, the statistic to use is

$$T^+ = \sum (R_i \,\text{with}\, D_i > 0) \tag{2.6}$$

If there are ties, the statistic to be used has Normal distribution with zero mean and variance 1 and is given by:

$$V = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}} \tag{2.7}$$

If the p-value is lower than the significance level, the null hypothesis is rejected.

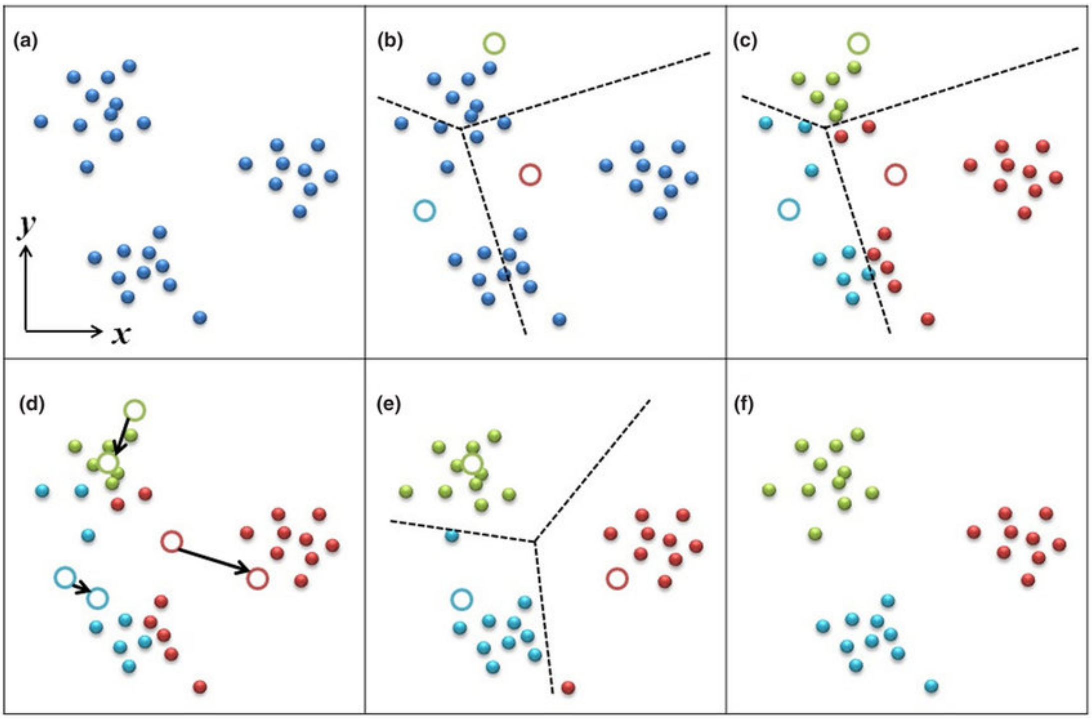
2.3 K-means Algorithm

Following Murtagh and Contreras, 2017, K-means clustering is one of the simplest and frequently used unsupervised learning algorithms, especially in data mining and statistics. Being a partitioning algorithm, its goal is to form groups of data points based on the number of clusters, represented by the variable k. K needs to be predefined before the execution. K-means uses an iterative refinement method to produce its final clustering based on the number of clusters defined by the user and the data set. Initially, k-means randomly chooses k as the mean values of k clusters, called centroids, and find the nearest data points of the chosen centroids to form k clusters. Then, it iteratively recalculates the new centroids for each cluster until the algorithm converges to one optimum value. K-means clustering would be suited with the numerical data with a low dimensionality because numerical data is used to compute the mean value. The type of data best suited for K-Means clustering would be numerical data with a relatively lower number of dimensions. The algorithm works as follow:

- K points are randomly initialized as centroids of clusters based on the predefined value of k.
- To form the k clusters, every data points of the data set are assigned to the nearest centroid by the distance. The Euclidean distance is used to in calculating the distance between each data points and the initialized centroids. Although there are many other metrics to find the closest distance, we apply Euclidean distance because several previous research about clustering analysis gained great outcomes using the Euclidean distance.
- The centroids are recalculated by averaging all of the data points assigned in each cluster so that the total intra-cluster variance can be reduced.
- Step 2 and 3 iterate until some criteria is met.

Criteria are normally, when there are no changes in the centroids values, the sum of distances between the data points and the centroid of each cluster does not change anymore, the data points assigned to the clusters are the same as the previous assignment or the maximum iteration number has reached in case the algorithm is given a fixed iteration time.

Figure 2.1: A schematic illustration of the K-means algorithm for two-dimensional data clustering



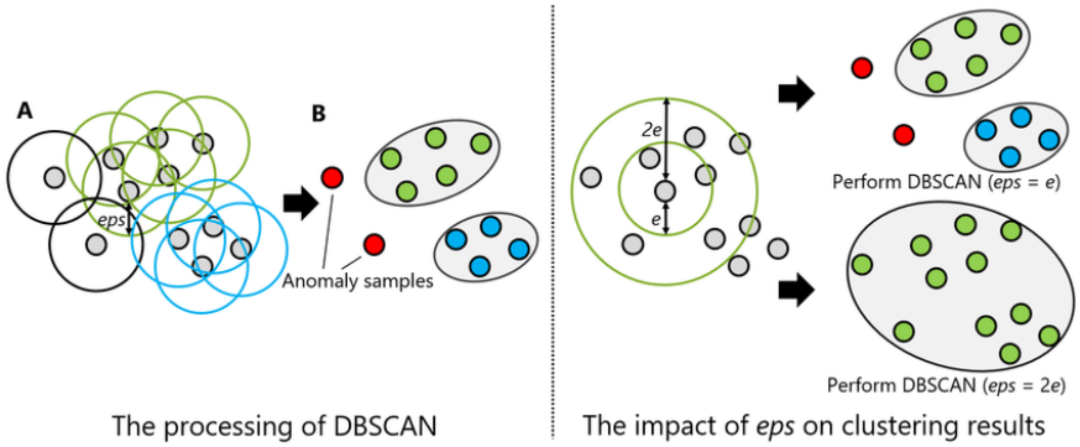
(a) The data points (solid blue circles) to be clustered in a 2D feature space. (b) For random locations of the cluster centers (aqua, green, and red hollow circles), each data point can be associated with the closest center. (c) The 2D space is divided into three regions through three decision boundaries (black dashed lines). (d) Each center moves to the centroid of the data points currently assigned to it (movements shown by the black arrows). (e) The updated cluster assignments of the data points are obtained according to the new center locations. The steps in (c) and (d) are repeated until convergence is achieved. (f) The final cluster assignments.

2.4 Density-based Spatial Clustering (DBSCAN)

One clustering technique that does not require the specification of the number of clusters is DBSCAN. However, DBSCAN basically requires two parameters: eps and min samples. Eps specifies the maximum distance between two data samples in which one of them is supposed to be a neighborhood of another which is a core point of a cluster. Min samples defines the minimum numbers of samples that have to be in the neighborhood together with the core sample. It assumes that a cluster is a dense region with data points that is greater than min samples within the range of eps of the core point and each cluster is separated from another by lower density.The steps in the DBSCAN algorithm includes:

- It starts with an arbitrary starting data point that has not been visited. The neighborhood of this point is extracted using the maximum distance eps.
- The points in the neighborhood of the core point are then used to search their respective neighborhood points since these points are still new unvisited points.
- This process of steps 2 and 3 is repeated until all points in the cluster have been visited and labeled forming the density-connected cluster.
- Then, the new un-visited point in the data set is retrieved and the algorithm repeats through step 1 to 4 until all points have been visited and become either noise or part of a cluster

Figure 2.2: A schematic illustration of the DBSCAN algorithm for two-dimensional data clustering



Chapter 3

Databases

The empirical strategy of this paper is geared toward investigating whether connections to politicians affect the business activities of firms.

Three types of data have been used to analyze the corruption episodes: Orbis - Bureau Van Dijk Data, Parliament Data and the Relational Data on Members of Portuguese Governments (1976–2014).

After we understand the behavior and profile of the three databases, we will create the final dataset. The Database is a combination of all three tables that we will explain above. For this combination, we applied five different merging techniques that are forms of multi-valued logic that deal with reasoning that is approximate rather than fixed and exact.

3.1 Primary Databases Description

3.1.1 Orbis - Bureau Van Dijk Data

Orbis is the world's most powerful comparable data resource on private companies. Orbis is Bureau van Dijk's flagship company database and resource for entity data. It has information on more than 400 million companies and entities across the globe – 40 million of these have detailed financial information. It contains information on companies across the world and focuses on private company information and also presenting companies in comparable formats.

The data that will be used in this project contains 2.673.436 records with 154.602 companies that got incorporated between 1950 and 2020 with more than 408.064 different people registered. There are 633 different columns with many several complexities of understanding. Firstly the simple variables about each company like tax number, number of managers/directors of the company, and more. Secondly, the variables that explain the profile of the partners, shareholders, owners and their profile variables like gender, birth date, current employee or not, and more.

Thirdly, some econometric variables explain percentages of shareholders, information dates, DUOs (Do Unto Others), and its percentages, GUOs, and the main section of the company. Some variables describe their fiscal years (since 2011), opening and closing dates (if it's the case), audit status, accounting practices, operating revenues, and more. And finally, there are many columns about the cash flow inside each company like profit margin, ROE using P/L, current ratios, ROCE, solvency ratios, number of employees, and more.

3.1.2 Parliament Data

The assembly consists of 230 members with their respective relatives. The members are elected by popular vote for legislative terms of four years from the country's twenty-two constituencies (eighteen in mainland Portugal corresponding to each district, one for each autonomous region, Azores and Madeira, one for Portuguese living in Europe, and the last one for those living in the rest of the world). All constituencies are represented within it, along with the plural political currents put to the vote that have obtained parliamentary representation and they represent all citizens, including non-electors, electors who did not vote, and those who did not give their electoral support to the Members elected.

Firstly, we combine profile information about the Portuguese Parliament like gender, age, political party, marital status, and more. Secondly, we scrapped the information about their marital status, campaigners, and their occupations. Finally, we subtracted the information about their last four professions, start and end dates, and the location that each one of them used to work. The database was extracted by Web Scrapping techniques inside the government official website https://www.parlamento.pt/

3.1.3 Relational Data on Members of Portuguese Governments (PtGov Database)

This data set contains relational data of all government members since the first constitutional government in Portugal after the Carnation Revolution (1976) until July 2013, comprising 19 governments. This information was collected through one-year research carried out by the authors using public records and official information (public and private institutions). Moniz and Campos, 2015

The first part of the data is described in this section portrays the basis of each connection. They specify the actor and the organization/institution to which it is connected, as well as the position held. The political party/group to which the actor is associated, as well as the party or group that in a specific relation the actor is representing, are also presented.

Firstly, the profile variables are name, political party, the year that the connection starts and when it ends, organization, and the position on the organization.

And second, as you can check in Moniz and Campos, 2015, the Metadata on the next table describes the connections between the government employees and private companies.

Table 3.1: Name, type of data and description of descriptive metadata variables.

Variable	Description
N_Years	Number of years between the begin and end year;
Econ_Sector	Connection to a given economic sector
ThreeMore_Econ_Sector	Indication that a given person has connections to three or more economic sectors.
Course_University	Course attended/completed.
Place_University	City of the university.
Gender	Gender of the person
MajorEconGroup	Connection to a given major economic group
PPP	Indication that a given connection involves a company involved in public-private partnerships.
Angola	Indication that a given connection involves a company with capital from Angola.
PSI20	Connection to a company indexed in the Portuguese Stock Market Index.
Government	Connection to a given government
Privatized_Company	Indication that a given connection involves a privatized company.

We will use this information later to analyse and compare the performance of the companies that has any connection with the government the information portrayed in the data set was collected during a one-year research conducted

3.2 Final Database description and methodology

Each data set above has a different structure. The biggest challenge about merging the Orbis Database with the public employees is that there is no Primary/Foreign Key between the bases. Firstly, it was used Full Name as the primary key and before each technique application, the data was cleaned and treated (lower-cased, remove punctuation, and more) and the error was minimized.

In a structured database, names are often treated the same as metadata for some other field like an email, phone number, or ID number. But what happens if you only have a name to look up a record?

When names are your only unifying data point, correctly matching similar names takes on greater importance, however, their variability and complexity make name matching a uniquely challenging task. Nicknames, translation errors, multiple spellings of the same name, and more all can result in missed matches. While search tools are abundant on the market, a name search is a different animal than document search and requires a fundamentally different approach.

Different name matching methods are best suited to solve different name matching challenges. There are many ways to match names, but no one universal solution. The best name matching software uses a hybrid of multiple methods to address the maximum number of name variations:

- Common key method
- List method
- Edit distance method
- Statistical similarity method
- Word embedding method

In this thesis we applied all five techniques above to confirm we selected the right combination of names. You can search more about each one of them in Gagliarducci and Manacorda, 2020

We we will emphasize a Edit distance method called Fuzzy String Matching. The Fuzzy String Matching, also known as Approximate String Matching, is the process of finding strings that approximately match a pattern.

Firstly, the Fuzzy logic is a form of multi-valued logic that deals with reasoning that is approximate rather than fixed and exact. Fuzzy logic values range between 1 and 0. For example, the value may range from completely true to completely false. In contrast, Boolean Logic is a two-valued logic: true or false usually denoted 1 and 0 respectively, that deals with reasoning that is exact. Fuzzy logic tends to reflect how people think and attempts to model our decision making hence it is now leading to new intelligent systems (expert systems).

Now, we have choosen which distance is better for our purpose. The first suggestion about the problem of string matching was treated in Navarro, 2001 when the author presents a number of experiments to compare the performance of the different algorithms and show which are the best choices. Each one them were applied in this paper, like Jaro Winkler distance, Hamming distance, Levenshtein distance and more.

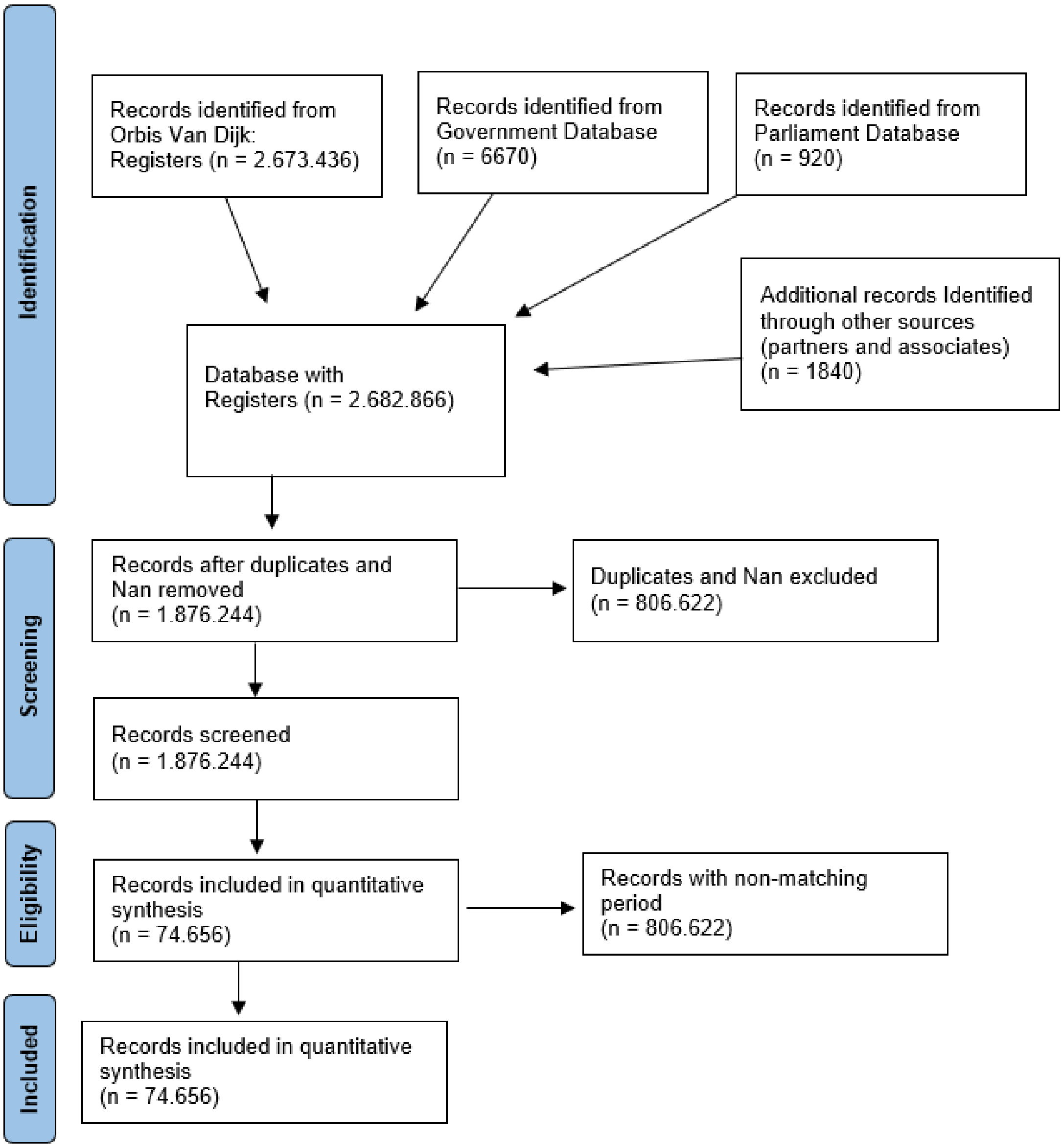
The paper written by D., 2019 provides a comparison of various algorithms for approximate string matching. Like in this thesis, the author provides an alternative approach for approximate string matching which are better suited for text retrieval.

Based on this, we have selected the Levenshtein Distance for the final merge. Based on Haldar and Mukhopadhyay, 2011, The Levenshtein distance is a number that tells you how different two strings are. The higher the rate, the more different the two strings are. It can be calculated by:

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j)+1 \\ \operatorname{lev}_{a,b}(i,j-1)+1 \\ \operatorname{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \end{cases}$$

According to MIT in "Longstanding problem put to rest, howpublished = https://news.mit.edu/2015/algorithm-genome-best-possible-0610, note = Accessed: 2021-10-06", n.d., it may very well be that Levenshtein's algorithm is the best that we'll ever get in terms of efficiency.

And finally, check the Data Flow bellow to understand the exhausting process to derive the final database.


Identification
Records identified from Orbis Van Dijk: Registers (n = 2.673.436)
Records identified from Government Database (n = 6670)
Records identified from Parliament Database (n = 920)
Additional records Identified through other sources (partners and associates) (n = 1840)
Database with Registers (n = 2.682.866)
Screening
Records after duplicates and Nan removed (n = 1.876.244)
Duplicates and Nan excluded (n = 806.622)
Records screened (n = 1.876.244)
Eligibility
Records included in quantitative synthesis (n = 74.656)
Records with non-matching period (n = 806.622)
Included
Records included in quantitative synthesis (n = 74.656)

As you can see in this data flow, the first part was focused on the extraction and merging part. Like we mention before, we had to make sure that we are selecting the right people and their respective enterprises. Secondly, we treated the database removing the duplicates and records that is missing any variable. And thirdly, we checked each company and person on the database to make sure that the period of analysis was correct. From now, we will make the analysis and clustering of this dataset.

3.2.1 Final Database

The final database contains 640 different columns to be analyzed and has all econometric and profile variables to make the analysis.

For further analysis, we selected for this database 74706 different people (and their respective companies). But only 76 of them have real connections with the government.

There are two merges inside this database. The first one is the parliament database with the Orbis Database that there are 17 deputies/partners that owe or are a shareholder of an outliers business. We collected their last four occupations for the previous 15 years.

And the second database is the merge between the PtGov database with the Orbis database. There are 59 connections between these public employees that are a shareholder/CEO of a rising company. These records start in 2002 and their chronology will be respected for the analysis and future comparisons

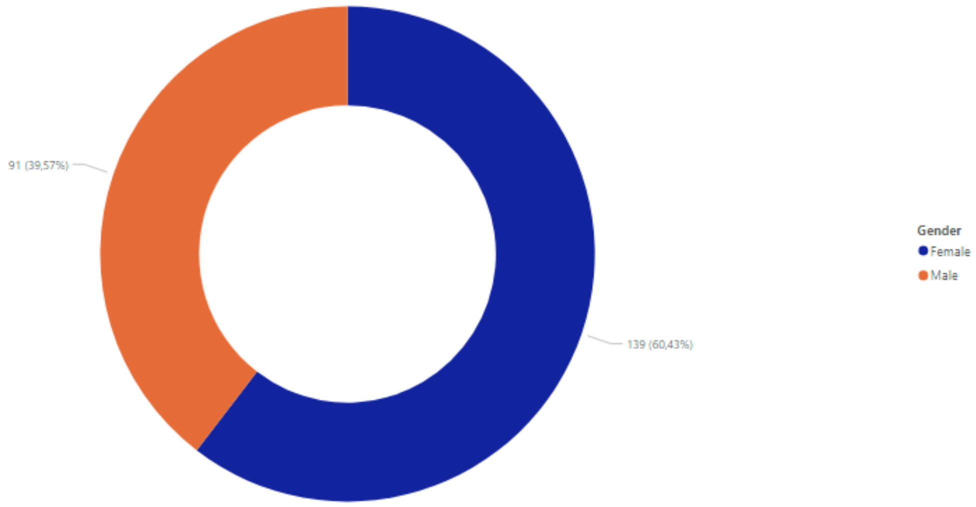
3.3 Analysis of the primary tables

3.3.1 Profile of the Portuguese Parliament

We will dispose the descriptive analysis of the Portuguese parliament to understand and explain that the Portuguese parliament is more diverse when you compare it to the worldwide Parliament.

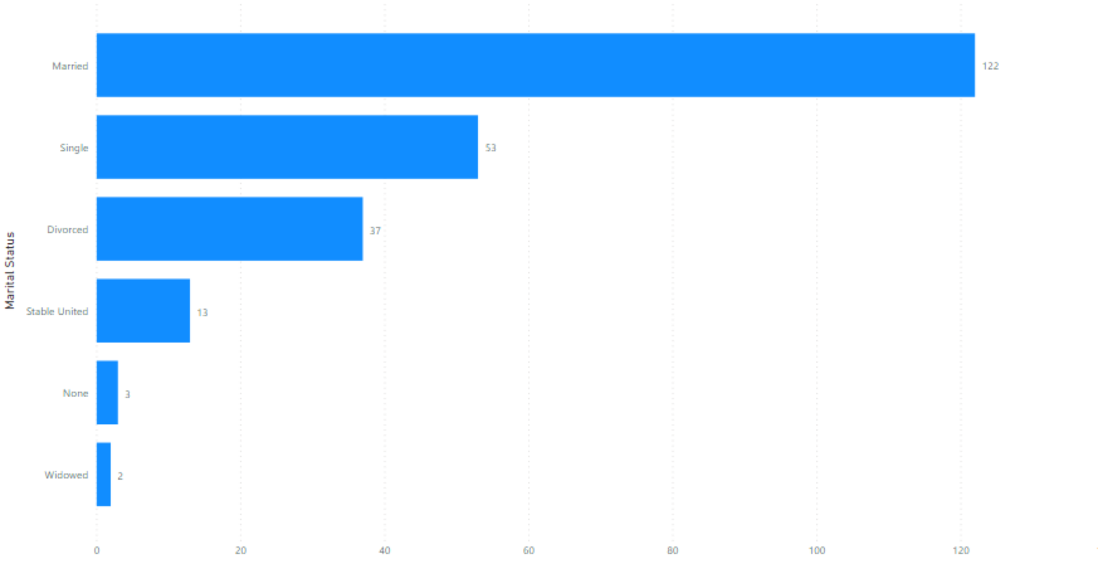
Firstly, the global average of women deputies in parliaments has now reached 25.5 percent but in Portugal is 40 percent. This percentage have been growing for the last five elections.

Figure 3.1: Gender distribution of the Portuguese Parliament



Male deputies are a majority in every region in 60 percent of the country.

Figure 3.2: Marital Status distribution of the Portuguese Parliament



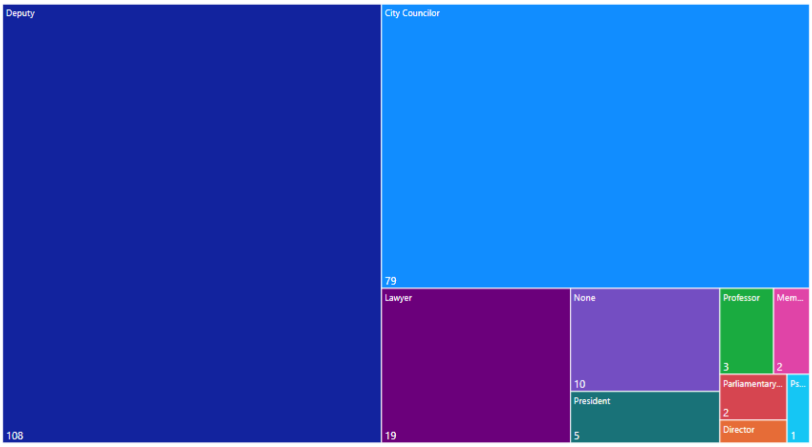
When we check the marital status proportions, we can read that more than 50 percent of the parliament is married and the global average is above 75 percent.

Table 3.2: Parties of the Portuguese Parliament

	N	%
PS	108	47%
PSD	79	37%
BE	19	8%
PCP	10	4%
CDS-PP	5	2%
PAN	4	1%
PEV	2	1%
Independent	2	1%
CH	1	~0%
IL	1	~0%

We can check above that 47 percent of the portuguese parliament is composed by PS deputies. And 37% of the deputies are connected to PSD.

Figure 3.3: Last Occupation distribution of the Portuguese Parliament



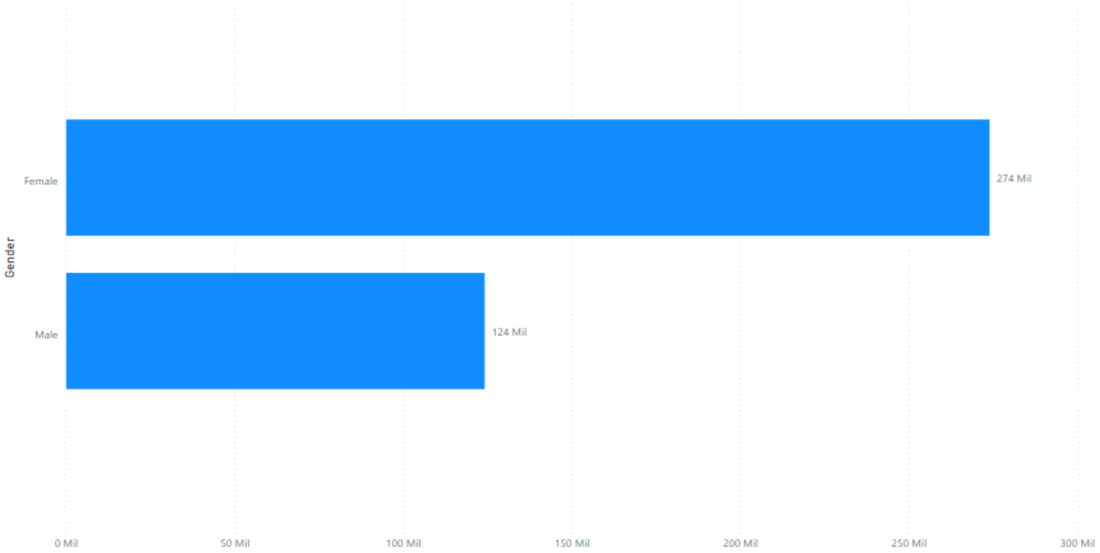
More than 50 percent of the actual parliament was connected to politics someway, for example, 34 percent of the deputies were deputies last election and 8 percent were city councilors. This fact is an opportunity to analyze even closer their previous occupations and if this is a real factor that may cause some type of advantage.

3.3.2 Profile of the Portuguese Companies

For this descriptive analysis, we applied time sampling, in other words, we applied a sampling technique only to extract the people that are or were working at the same period that the public employees that have connections inside the government. This sample database contains 398,245 companies with their respective variables.

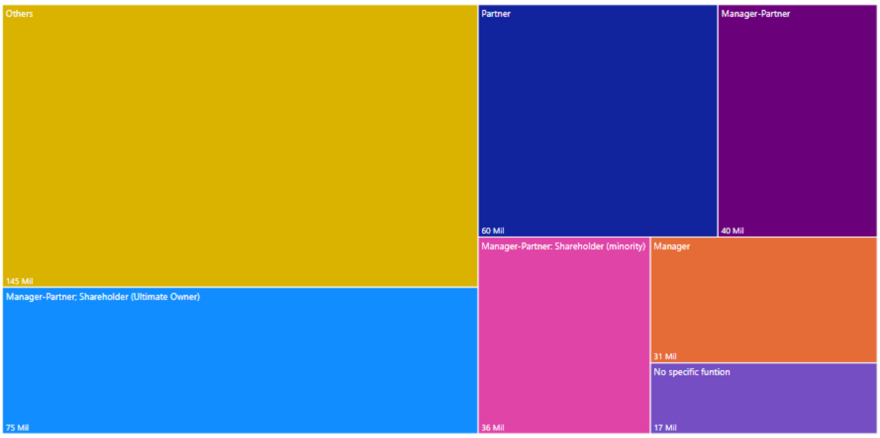
The current global labor force participation rate for women is close to 49 percent. For men, it's 75 percent. That's a difference of 26 percentage points, with some regions facing a gap of more than 50 percentage points. In our case, about 69 percent of the people that worked in a Portuguese company are men

Figure 3.4: Gender Distribution of the Orbis Database



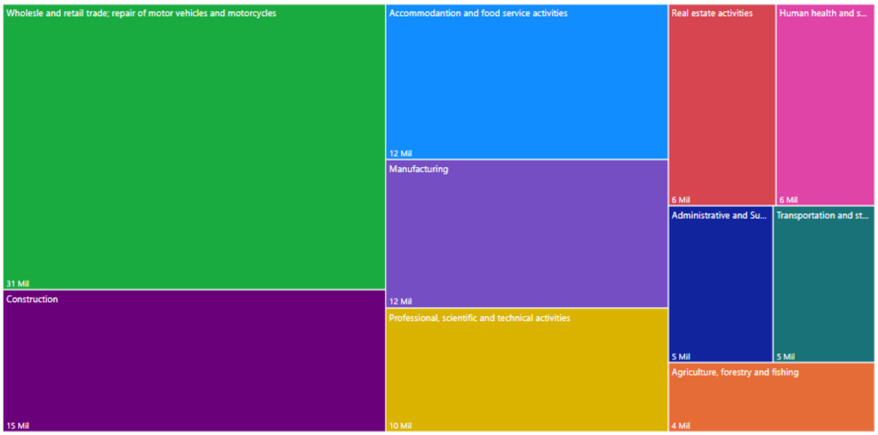
We can notice that only than 7% of the employees/employers is a manager of the company.

Figure 3.5: Shareholder Title distribution of the Orbis Database



And finally, the table below shows the distribution of the segment of the companies, here we have available 93,696 records with the specification about the segment.

Figure 3.6: Sector distribution of the Orbis Database



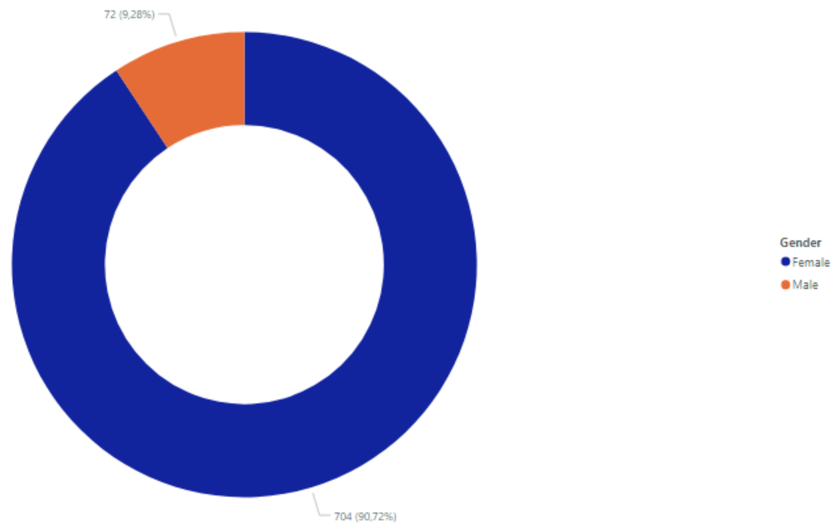
The biggest part of the companies are in the Wholesale and retail trade; repair of motor vehicles and motorcycles market. The other segment is the sum of the sectors Information and communication, Financial and insurance activities, Arts, entertainment and recreation, Education, Mining and quarrying, Water supply; sewerage, waste management, and remediation activities, Electricity, gas, steam, and air conditioning supply, Public administration and defense; compulsory social security, Activities of extraterritorial organizations and bodies

3.3.3 Profile of the Relational Data on Members of Portuguese Governments (PtGov Database)

The Relational Data on Members of Portuguese Governments (1976–2014) by Moniz and Campos, 2015 is a data set containing information on the explicit connections concerning all members of Portuguese governments from 1976 until July 2013 and it contains more than 20 different columns with useful information about the member of the Portuguese Governments. Here we will only describe the main variables that will be useful for the posterior analysis.

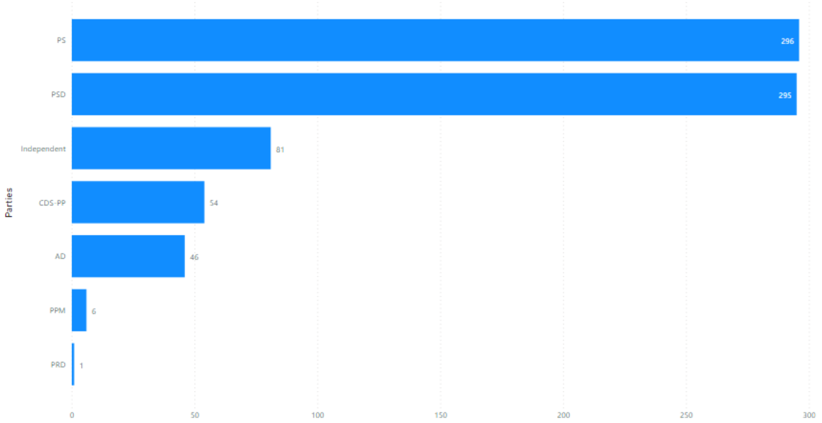
This database contains 776 public employees and their gender distribution is:

Figure 3.7: Gender Distribution of the PtGov Database



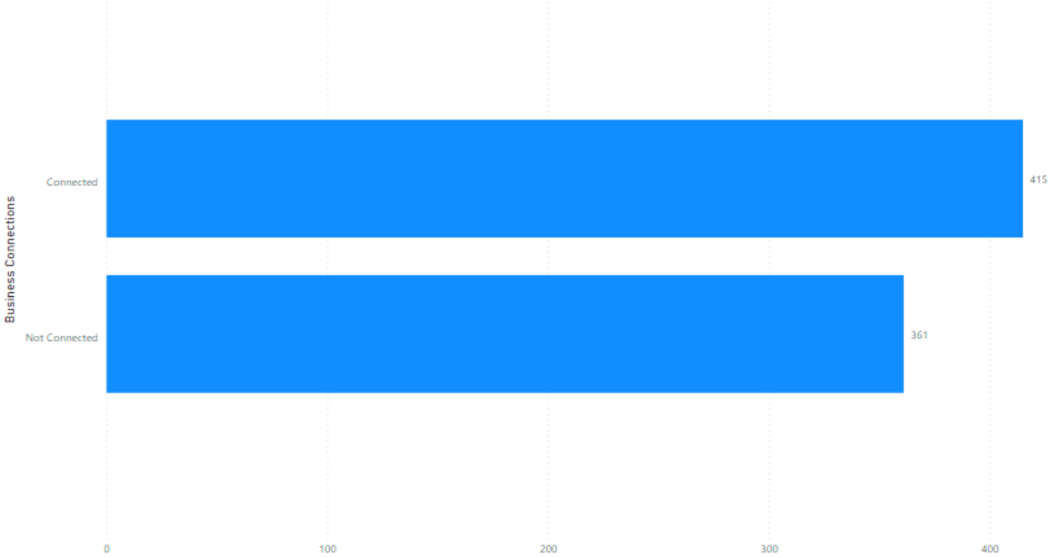
As you can see, more than 90% of the public employees for the last 20 years are men.

Figure 3.8: Parties distribution of the PtGov Database



And here we can check that there is a balance between the two more representative parties in Portugal (PS and PSD with 38% each), and there is a preponderance part of the sample that is Independent (11%). For further analysis, here is the distribution of people that are connected to the business.

Figure 3.9: Bussines conections of the PtGov Database



As you can read, there are 415 public employees with business connections. These people were analyzed to be added to the final dataset (next chapter).

3.4 Analysis of the final tables

And here we come to the final database, the Connections Inside the Government, in other words, this is the database where we compile every database with public employees, parliament for their last four occupations, public employees partners with the Orbis Van Dijk Database.

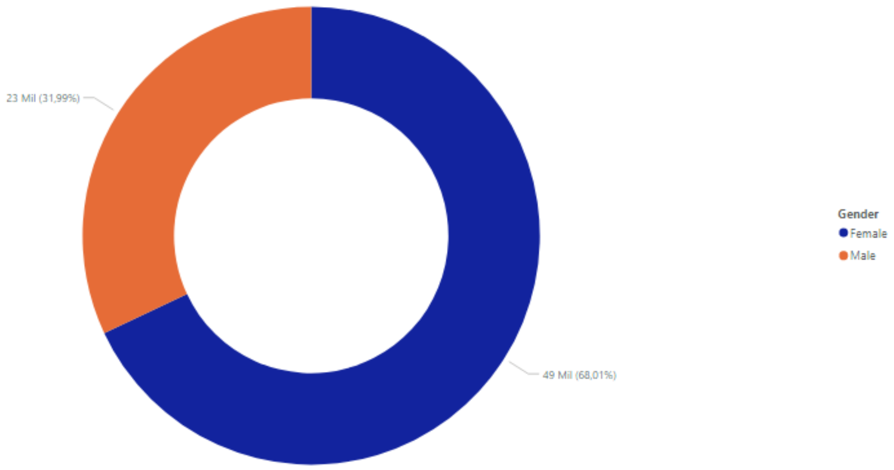
The final database contains 640 different columns to be analyzed and has all econometric and profile variables to make the analysis. Here we will analyze the profile of the enterprises that contain connections inside the government.

3.4.1 Enterprises without connections inside the government (CIG Database)

This section will describe the CIG - The companies without connections inside the government set in other words, this analysis will narrate the profile of the companies that have connections inside the Portuguese government, later we will use this to infer some different behavior

Firstly, this sample contains 74706 companies with their respective companies.

Figure 3.10: Gender Distribution of the CIG Database



Here we will describe the Connections distribution by principal GUO/CEO of the CIG database.

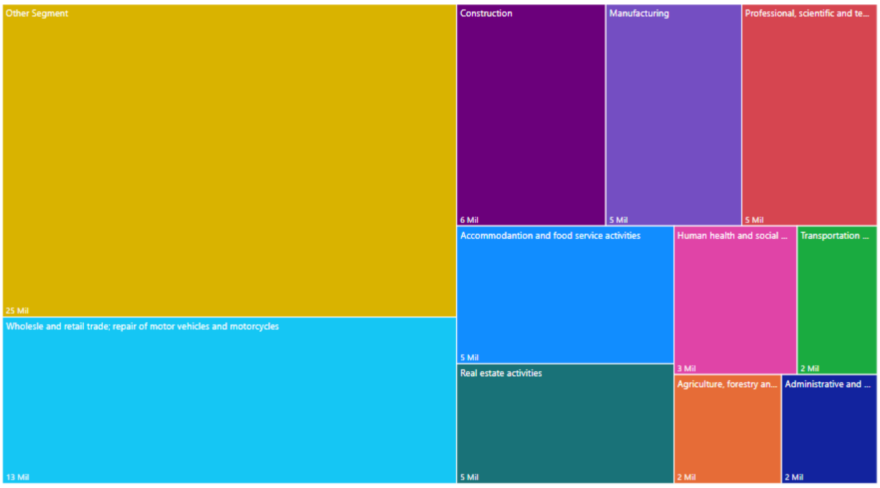
Table 3.3: Connections to privatized companies distribution

	N	%
Connected	76	1%
Not Connected	74580	99%

As you can notice, there are 76 companies that have connections with the government in their existence time.

The table below shows the distribution of the segment of the companies,

Figure 3.11: Sector distribution of the CIG Database



The biggest part of the companies are in the Wholesale and retail trade; repair of motor vehicles and motorcycles market. The other segment is the sum of the sectors Information and communication, Financial and insurance activities, Arts, entertainment and recreation, Education, Mining and quarrying, Water supply; sewerage, waste management and remediation activities, Electricity, gas, steam and air conditioning supply, Public administration and defence; compulsory social security, Activities of extraterritorial organisations and bodies.

3.4.2 Enterprises with connections inside the government (EIG Database)

Here we are going to specify some important variables inside the EIG - The companies with connections inside the government (1% of companies that has connections)

Table 3.4: Gender Distribution of the connections data sample

	N	%
M	65	85%
F	11	15%

Like some previous databases, the proportion of men is equal to 85%

Table 3.5: Political Party Distribution of the connections data sample

Connection per Party	N	%
Independent	31	42%
PSD	20	27%
PS	15	20%
CDS	5	6%
AD	5	6%

The most representative party is the Independent party employees, followed by PSD and PS, with 27% and 20% respectively.

Table 3.6: Main Section of the connections data sample

Main Section	N	%
G - Wholesale and retail trade; repair of motor vehicles and motorcycles	18	24%
F - Construction	11	16%
I - Accommodation and food service activities	11	16%
C - Manufacturing	9	11%
M - Professional, scientific and technical activities	6	8%
L - Real estate activities	5	6%
H - Transportation and storage	3	3%
Q - Human health and social work activities	3	3%
K - Financial and insurance activities	2	2%
N - Administrative and support service activities	2	2%
J - Information and communication	2	2%
S - Other service activities	2	2%
A - Agriculture, forestry and fishing	1	1%
P - Education	1	1%

The biggest part of the companies are in the Wholesale and retail trade; construction and Accommodation and food service activities.

Table 3.7: Shareholder Type of the connections data sample

Shareholder Type	N	%
One or more named individuals or families	50	68%
Corporate	5	8%
Financial company	1	1%
Unnamed private shareholders, aggregated	1	1%
Not informed	19	23%

The majority of the Shareholder type to the EIG Database is companies with One or more named individuals or families, 68%.

Table 3.8: Number of employee of the connections data sample

Number of employees (Last avail. yr)	N	%
One	12	17%
[2, 10]	29	39%
[11, 20]	2	2%
20 or more	10	14%
Not specified	23	32%

And finally, the majority of the companies contain 2 to 10 employees and there are 11 companies with more than 20 people. About 17% of the companies have only one employee, sometimes is habitual for a person to create a company to protect their own interests

Chapter 4

Results

The principal purpose of this section is to compare and analyze the some significant econometric Ratios between the companies that contain connections inside the government (EIG Database) with the companies that don't (CIG Database). After this chapter, you will understand that companies with connections usually have better financial performance.

We will compare the performance of the companies based on their Operating Revenue (Turnover) EURO, Cash Flow EURO, Total assets EURO, Profit Margin %, ROE using P/L be-fore tax, ROCE using P/L before tax, Solvency ratio (Asset-based) %, and shareholders funds EURO.

We specified that our list was distributed among time, but for this analysis, we only have variables from 2011 until 2020. The information that we will describe below is only for these recent cases.

After the comparative analysis, some hypothesis test will be applied to test if the ratios of the EIG database are more significant than the CIG database. But before, a Kolmogorov-Smirnov will be applied to validate if the data has the Gaussian Distribution and a Bartlett Test will be applied to test the homocedasticity. If the data is Gaussian and homocedastic we will apply paired sample t-test, if they're not, we will apply Mann-Whitney non-parametric test.

4.1 Comparative Analysis

Is it possible to infer that companies with public connections have, economically, a better performance than companies without connections?

4.1.1 Operating revenue

Operating revenue is the revenue that a company generates from its primary business activities.

For example, a retailer produces its operating revenue through merchandise sales; a physician derives their operating revenue from the medical services that they provide. What constitutes operating revenue varies based on the business or the industry.

Table 4.1: Operating Revenue (thousand EURO) means between CIG and EIG databases from 2016 to 2020

Year	CIG Database (€)	EIG Database (€)	Delta (€)
2020	385.0	Nan	Na
2019	139.0	167.0	+43%
2018	135.0	172.0	+27%
2017	137.0	158.0	+17%
2016	116.0	132.5	+14%

4.1.2 Cash Flow

Cash flow is the net amount of cash and cash equivalents being transferred into and out of a business. Cash received represents inflows, while money spent represents outflows.

At a fundamental level, a company's ability to create value for shareholders is determined by its ability to generate positive cash flows or, more specifically, maximize long-term free cash flow (FCF). FCF is the cash that a company generates from its normal business operations after subtracting any money spent on capital expenditures (CapE)

Here is the Cash flow between the two databases:

Table 4.2: Cash Flow (thousand EURO) means between CIG and EIG databases from 2011 to 2020

Year	CIG Database (€)	EIG Database (€)	Delta (%)
2020	-39.0	Nan	Nan
2019	20.41	36.35	+78%
2018	26.33	35.45	+34%
2017	28.43	31.93	+12%
2016	26.02	27.30	+4%
2015	15.74	34.92	+121%
2014	7.14	34.39	+381%
2013	9.71	20.46	+110%
2012	-4.06	6.75	+67%
2011	5.947	3.321	-55%

4.1.3 Profit Margin

Profit margin is one of the commonly used profitability ratios to gauge the degree to which a company or a business activity makes money. It represents what percentage of sales has turned into profits. Simply put, the percentage figure indicates how many cents of profit the business has generated for each dollar of sale.

Table 4.3: Profit Margin (%) means between CIG and EIG databases from 2011 to 2020

Year	CIG Database (%)	EIG Database (%)	Delta (%)
2020	-9.02	Nan	Nan
2019	5.82	5.81	-.01%
2018	5.57	3.54	-2.03%
2017	5.41	6.16	+.75%
2016	4.18	4.11	-.07%
2015	3.32	6.18	+2.86%
2014	1.71	5.41	+3.7%
2013	0.20	2.08	+1.88%
2012	-1.78	3.52	+5.3%
2011	-0.26	2.14	+2.4%

4.1.4 Total Assets

An asset is a resource with economic value that an individual, corporation, or country owns or controls with the expectation that it will provide a future benefit. Assets are reported on a company's balance sheet and are bought or created to increase a firm's value or benefit the firm's operations. An asset can be thought of as something that, in the future, can generate cash flow, reduce expenses, or improve sales, regardless of whether it's manufacturing equipment or a patent.

Table 4.4: Total Assets (thousands €) means between CIG and EIG databases from 2012 to 2020

Year	CIG Database (%)	EIG Database (%)	Delta (%)
2020	45.00	Nan	Nan
2019	567.98	863.48	+52%
2018	707.33	822.39	+16%
2017	727.57	809.12	+11%
2016	695.09	763.75	+9%
2015	702.69	807.25	+14%
2014	751.01	901.09	+19%
2013	801.03	1129.04	+40%
2012	797.54	1120.38	+42%

4.1.5 Shareholders Equity

For corporations, shareholder equity (SE), also referred to as stockholders' equity, is the corporation's owners' residual claim on assets after debts have been paid. Shareholder equity is equal to a firm's total assets minus its total liabilities.

Table 4.5: Shareholders Funds (thousands €) means between CIG and EIG databases from 2011 to 2020

Year	CIG Database (%)	EIG Database (%)	Delta (%)
2020	-2628.00	Nan	Na
2019	254.64	463.17	+82%
2018	281.23	399.72	+42%
2017	268.25	368.94	+138%
2016	239.63	340.84	+42%
2015	235.56	466.12	+98%
2014	274.36	504.77	+83%
2013	269.46	540.26	+100%
2012	286.45	557.22	+94%
2011	294.40	558.79	+89%

4.1.6 Return on Equity

Return on equity (ROE) is a measure of financial performance calculated by dividing net income by shareholders' equity. Because shareholders' equity is equal to a company's assets minus its debt, ROE is considered the return on net assets. ROE is considered a measure of a corporation's profitability in relation to stockholders' equity.

Table 4.6: Return of Equity (%) between CIG and EIG databases from 2011 to 2020

Year	CIG Database (%)	EIG Database (%)	Delta (%)
2020	-0.95	Nan	Na
2019	12.74	1.95	+1000%
2018	12.84	19.15	+49%
2017	17.36	1.67	+1000%
2016	14.03	15.21	+8%
2015	2.92	7.94	+271%
2014	11.49	8.67	-25%
2013	6.39	-2.84	-444%
2012	-0.72	-0.85	-18%
2011	2.22	-4.44	-200%

4.1.7 Return on Capital Employed

Return on capital employed (ROCE) is a financial ratio that can be used in assessing a company's profitability and capital efficiency. In other words, this ratio can help to understand how well a company is generating profits from its capital as it is put to use.

Table 4.7: Return of Equity (%) between CIG and EIG databases from 2011 to 2020

Year	CIG Database (%)	EIG Database (%)	Delta (%)
2020	5.85	Nan	Na
2019	8.99	5.85	-34%
2018	11.91	8.99	-24%
2017	7.38	11.91	+61%
2016	3.92	7.38	+188%
2015	10.79	3.92	+275%
2014	-5.00	10.79	+215%
2013	-7.10	-5.00	+29%
2012	-1.88	-7.10	-333%
2011	1.29	1.88	+14%

4.1.8 Solvency Ratio

A solvency ratio is a key metric used to measure an enterprise's ability to meet its long-term debt obligations and is used often by prospective business lenders. A solvency ratio indicates whether a company's cash flow is sufficient to meet its long-term liabilities and thus is a measure of its financial health. An unfavorable ratio can indicate some likelihood that a company will default on its debt obligations.

Table 4.8: Solvency asset based (%) between CIG and EIG databases from 2011 to 2020

Year	CIG Database (%)	EIG Database (%)	Delta (%)
2020	Nan	Nan	Na
2019	41.02	39.58	-3%
2018	40.15	40.31	+0%
2017	37.77	39.00	+3%
2016	35.83	38.28	+7%
2015	35.65	35.29	-1%
2014	35.17	37.89	+7%
2013	34.33	32.96	-9%
2012	34.41	28.85	-16%
2011	34.62	35.59	+1%

4.1.9 Hypothesis Tests

This section will focus on the hypothesis tests to compare the means between the same variables between the two different databases.

Before any hyphotesis test, we have to test two assumptions:

- If the populations are gaussian with a Shapiro-Wilk test, and
- If the populations are Homocedastic with a Bartlett Test.

Firstly, the Shapiro-Wilk test for normality is available when using the distribution platform to examine a continuous variable in some population.

The null hypothesis for this test is that the data are normally distributed. The P-value < limit value listed in the output is the p-value. If the chosen alpha level is 0.05 and the p-value is less than 0.05, then the null hypothesis that the data are normally distributed is rejected. If the p-value is greater than 0.05, then the null hypothesis is not rejected.

Check the p-values for each database and variable.

Table 4.9: Shapiro-Wilk test to check normality in each variable.

Variable	p-value CIG	p-value EIG	Hypothesis Test
Operating Revenue	0.99	0.56	T-test
Cash Flow	0.85	0.52	T-test
Profit Margin	0.89	0.22	Mann-Whitney
Total Assets	0.01	0.03	T-test
Shareholders Equity	0.01	0.01	T-test
ROE	0.01	0.01	Mann-Whitney
ROCE	0.01	0.09	Mann-Whitney
Solvency Ratio	0.97	0.13	T-test

Based on the hyphotesis that we explained, we will apply the Mann-Whitney non parametric tests on the variables: Profit Margin, ROE and ROCE.

Secondly, we will test the homoscedasticity with the Barlett test. Bartlett's test of Homogeneity of Variances is a test to identify whether there are equal variances of a continuous or interval-level dependent variable across two or more groups of a categorical, independent variable. It tests the null hypothesis of no difference in variances between the groups.

Table 4.10: Bartlett test to check homoscedasticity between variables

Variable	p-value Barlett
Operating Revenue	0.88
Cash Flow	0.13
Profit Margin	0.22
Total Assets	0.14
Shareholders Equity	0.98
ROE	0.77
ROCE	0.41
Solvency Ratio	0.37

Checking each one the p-value we can check that there are no difference between the variances.

And finally, we will be able to check if the averages of each varaible are equal. All the tests bellow have the same hypothesis We will carry out tests of the null hypothesis that the means of the populations from which the two samples were taken are equal and the alternative hypothesis is that the CIG mean is greater than the EIG mean. Check the results below:

Table 4.11: T test to check the difference between non parametric data

Variable	p-value t-test
Operating Revenue	0.001
Cash Flow	0.08
Total Assets	0.001
Shareholders Equity	0.03
Solvency Ratio	0.51

As you can check, for the parametric hypothesis tests, we can infer with 5% of significance that the Operating Revenue, Total Assets and Shareholders Equity are greater in companies that have connections inside the government than companies that don't.

Table 4.12: Mann-Whitney to check the difference between parametric data

Variable	P-value t-test
Profit Margin	0.01
ROE	0.71
ROCE	0.53

With 5% of significance, we can assume the profit margin of a company with connections inside the government is greater than the profit margin in a company that doesn't.

4.2 Cluster Analysis

What are the main profile clusters of the privileged companies in Portugal? And what are their behavior compared to the companies with political connections?

4.2.1 Cluster Selection

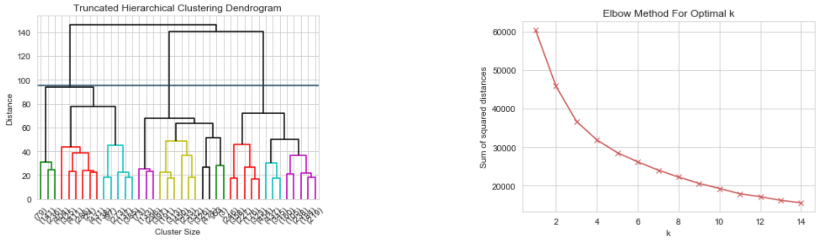
We will create clusters for all companies and then we will compare the performance of each clusters with the EIG Database.

One of the most important techniques in pattern recognition is Clustering. Clustering of high-performance companies is very important not only for investors, but also for the creditors, financial creditors, stockholders, etc Momeni et al., 2015. To this end, we have the economic variables Operating Revenue, Cash Flow, Total Assets, ROE and ROCE.

The K-means is one of the most popular clustering algorithms (Hajizadeh et al., 2010 and Kanungo et al., 2002). This clustering algorithm was first described by Macqueen (1967). This algorithm is a partitioning clustering or non hierarchical method that splits dataset (objects) into k clusters Jain, 2010.

The stages of k-means are described in Celebi, 2011.

At the end of this part, we are expecting to have different clusters for all companies with public connections. The techniques that will be used are: K-means, K-medoids, Self Organizing Maps, DBScan and Mean Shift. After we apply them, each approach will be complemented with a silhouette graph and a dendrogram. Both graphs will be supported by the K-means technique.



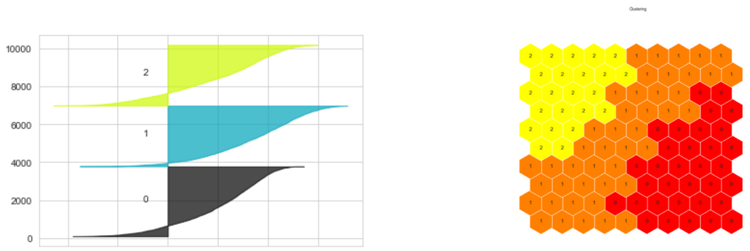
(a) Dendogram Graph

(b) Elbow Graph

Figure 4.1: Support Graphs to decide the number of clusters (Dendogram and Elbow Graphs)

When we analyse the dendogram and the elbow graph above, we can suppose that the best number of clusters for our companies are three.

Now we will apply the other techniques to make sure about this decision. The techniques bellow are Self Organizing Maps, Mean Shift, DBScan and the Sillouette graph (Michaud, 1997, Comaniciu and Meer, 1999 and Kohonen, 1990)

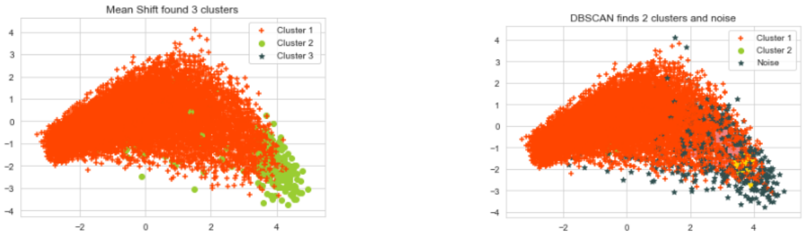


(a) Sillhouette Graph

(b) Self Organizing Map graph

Figure 4.2: Support Graphs to decide the number of clusters (SOM and Sillhouette)

As you can check, both techniques suggest three balanced clusters. Now we will check the performance of DB Scan and Mean Shift clustering techniques.



(a) Mean Shift Graph

(b) DBSCan graph

Figure 4.3: Support Graphs to decide the number of clusters (Mean Shift and DB Scan Graphs)

As we can check in Yang et al., 2014, DBScan - Density-based spatial clustering of applications with noise and Mean Shift struggled because the clusters here have a similar density. In our case, we don't have that many dimensions as necessary.

But we already selected our three clusters that we want to analyse. The next section will be about it.

4.2.2 Cluster Analysis and comparisson

In this section, we aim to develop the clusters and, in the end, to analyse them to build our strategy. Firstly, the cluster size is described below:

Table 4.13: Clusters size

	Cluster 1	Cluster 2	Cluster 3
Size	38578	24101	10901

The third cluster is unbalanced compared to the other but it describes a different behavior. When we compare the behavior of these clusters with the database with a government connection, we will check if this cluster is interesting or not.

Let's move on to analyse each cluster, and to check the conditional means for each group of variables.

Here's the table with the conditional mean of each one of the variables for the companies clusters and the EIG Database:

Table 4.14: Performance comparrison between Clusters and EIG Database

	Operating Revenue	Cash Flow	Total Assets	Shareholders Equity	Solvency Ratio
Cluster 1	52.25	-1.44	28.8	12.78	9.21
Cluster 2	117.00	14.01	95	291.1	19.23
Cluster 3	329.75	9.24	299.54	201.46	33.33
EIG Database	157.37	28.85	902.06	524.72	40.52

The EIG Database has the highest values in four out of five variables but we can notice that the clusters are well distributed and we can generalize some group of companies based on them.

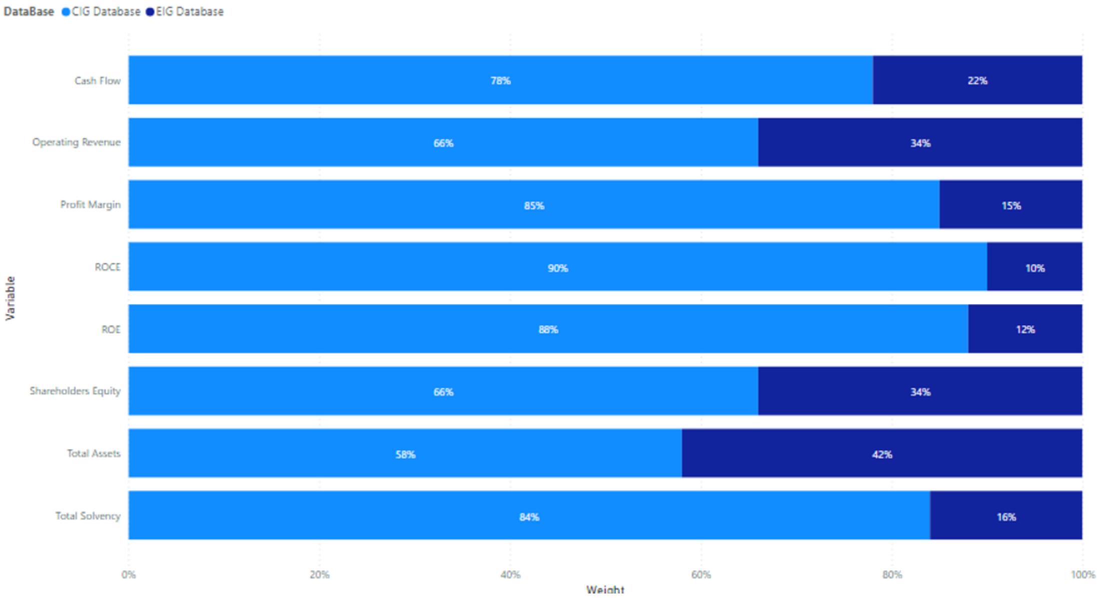
4.3 Performance Analysis

Talking about the companies with public connections, what are their importance on the economy? In other words, in the last decade, how many percentages of the national profits are exclusive to them

We have already described the behavior and profile of these 1% of companies in the section 3.4.2. But we are going to compare their economic behavior compared to the rest of the companies. In Bilan et al., 2020 you can read that income distribution can cause large-scale transformations in human resources structure, essential changes of economic outputs via its impact on life satisfaction and motivation of work. In this section, we will compare the performance between the companies with government connections (EIG Database) and the companies without them (CIG Database).

The following graph will demonstrate if the 1% of the companies are significant over the total performance from 2011 until 2020.

Figure 4.4: Weight of each econometric variable



As you can check, 42% of the total assets belong to the EIG Database. 22% of the total cash flow belongs to the companies with government connections but another important piece of information is that these companies only have 15% of the total profit margin in Portugal. Like Güleç and Bektaş, 2019 explains, Cash Flow and Profit Margin are not necessarily positively correlated. Many reasons can explain the performance of these ratios.

Chapter 5

Conclusion

Even politicians have a huge influence on company profits and performance, there is no strong evidence in Portugal to examine if the effects of the political connections on firms. it's noticeable that companies with connections inside the government have a significant positive performance speaking of analysis of econometric variables.

Before this thesis, only Moniz and Campos, 2015 described the relation between private companies and the government and their profile description.

This thesis solves these problems by taking specific angles. With the information provided on government websites via Web Scrapping, we collect a name list of politicians during the past years. But, the problem is that they were not combined in a single database.

Some useful insights were obtained about the profile of the Portuguese government and all companies that were founded in Portugal, like, the majority is composed of men, and more than 50% of the database worked with politics for more than 8 years. Wholesale and retail trade (24%) and Construction (16%) are the main components of the companies inside the total distribution and consequently, the ones with more connections inside the government.

This thesis fits into a new and growing literature that examines the impact of connections on business activities. It suggests that companies with benefits extracted are significant. These benefits will provide in the future monopolies, licenses, contracts, tax breaks, easy access to bank loans, and concessions to the beneficiaries ones. Consequently, these companies will have an advantage over the competition in their respective markets.

This thesis is among the first ones to examine the influence of political power connections in Portugal. In general, proportionally, a company with connections has a better performance than a company that does not.

5.1 Answering Research Questions

In Portugal, It is it possible to infer that companies with public connections have, economically, a better performance than companies without connections?

Based on the analysis of the variables Operating Revenue, Cash Flow, Total Assets and Profit Margin and their respective hypothesis tests, we can infer that these companies have some advantages compared to others.

What are the main profile clusters of the privileged companies in Portugal? And what are their behavior compared to the companies with political connections?

We selected three different clusters for these privileged companies. Cluster one is companies with lower operating revenue, negative cash flow, and lower shareholder equities. The second cluster is mainly composed of companies the higher equities and higher cash flow over the years. And thirdly, cluster three are companies mainly composed of Retail, Real State and Other service companies that contain the highest values of Operating Revenue, Total Assets, and Solvency. In the previous chapter, we had compared the absolute results of each cluster with the performance of the EIG Database and we inferred that these companies have a better performance than the two clusters.

Talking about the companies with public connections, what is their importance on the economy? In other words, in the last decade, how many percentages of the national profits are exclusive to them

The companies without government connections still have the best performance in every single variable. But speaking of Operating Revenue, Total Assets, and Shareholders Equity, the companies with government connections still have a significant part of the country's performance.

5.2 Recommendations for Future Works

The recommendations for potential future works are strongly related to the limitations of this thesis. More performance variables can be analyzed and compared to infer if these political connections are indeed significant or not.

The CEOs, GUO's and directors of the companies can be investigated as well to find if they have other companies with benefits as well. And with the right data, their respective connections can be investigated in contemplation of finding another source of benefits.

The Cluster analysis can be developed with more details, in this thesis we had created only three clusters to compare with the EIG database but it was reasonable to build more clusters and new AI approaches to specify more the companies and find more information about them

And finally, this thesis opens an opportunity to analyze and compare the performance of these companies in the future, conducive to analyzing and compare the results among the future years.

Bibliography

Bilan, Y., Mishchuk, H., Samoliuk, N., & Yurchyk, H. (2020). Impact of income distribution on social and economic well-being of the state. *Sustainability*, *12*(1), 429.

Bussab, W. d. O., & Morettin, P. A. (2010). Estatıstica básica. *Estatıstica básica* (pp. xvi–540).

Celebi, M. E. (2011). Improving the performance of k-means for color quantization. *Image and Vision Computing*, *29*(4), 260–271.

Comaniciu, D., & Meer, P. (1999). Mean shift analysis and applications. *Proceedings of the seventh IEEE international conference on computer vision*, *2*, 1197–1203.

D., A. (2019). A fuzzy approach to approximate string matching for text retrieval in nlp. *Journal of Computational Information Systems*, *15*, 26–32.

Faccio, M. (2006). Politically connected firms. *American economic review*, *96*(1), 369–386.

Gagliarducci, S., & Manacorda, M. (2020). Politics in the family: Nepotism and the hiring decisions of italian firms. *American Economic Journal: Applied Economics*, *12*(2), 67–95.

Güleç, Ö. F., & Bektaş, T. (2019). Cash flow ratio analysis: The case of turkey.

Hajizadeh, E., Ardakani, H. D., & Shahrabi, J. (2010). Application of data mining techniques in stock markets: A survey. *Journal of Economics and International Finance*, *2*(7), 109–118.

Haldar, R., & Mukhopadhyay, D. (2011). Levenshtein distance technique in dictionary lookup methods: An improved approach. *arXiv preprint arXiv:1101.1232*.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, *31*(8), 651–666.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 881–892.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Longstanding problem put to rest, howpublished = https://news.mit.edu/2015/algorithm-genome-best-possible-0610, note = Accessed: 2021-10-06. (n.d.).

Michaud, P. (1997). Clustering techniques. *Future Generation Computer Systems*, *13*(2-3), 135–147.

Momeni, M., Mohseni, M., & Soofi, M. (2015). Clustering stock market companies via k-means algorithm. *Kuwait Chapter of the Arabian Journal of Business and Management Review*, *4*(5), 1.

Moniz, N. M., & Campos, A. (2015). Relational data on members of portuguese governments (1976–2014).

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: An overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6), e1219.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, *33*(1), 31–88.

Ramalho, R. (2003). The effects of an anti-corruption campaign: Evidence from the 1992 presidential impeachment in brazil. *Unpublished Paper*.

Shleifer, A., & Vishny, R. W. (1994). Politicians and firms. *The quarterly journal of economics*, *109*(4), 995–1025.

Yang, Y., Lian, B., Li, L., Chen, C., & Li, P. (2014). Dbscan clustering algorithm applied to identify suspicious financial transactions. *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 60–65.