ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

Máster de Ingeniería Industrial

Culture vs. financial performance: an assessment using web-scraped data

Autor: Jaime Domínguez
Directora: Sara Lumbreras

Table of Contents

Resumen ejecutivo

Durante las últimas décadas ha habido un gran número de teorías acerca del liderazgo y cómo éste puede impactar el desempeño económico de las empresas. Sin embargo, pocas de estas teorías demostraron el impacto real de los distintos tipos de liderazgo. Este proyecto aborda esta cuestión llevando a cabo diversas técnicas de análisis de datos.

Este reporte provee una detallada metodología y justifica las diferentes tareas que se llevaron a cabo para desempeñar el análisis. Este análisis es muy innovador ya que la fuente de información no es convencional, técnicas de "web scraping" fueron utilizadas para reunir los datos.

El proceso comienza por la extracción de los datos desde distintas fuentes. Se requirió un exhaustivo proceso de limpieza que se llevó a cabo explorando valor límites, datos ausentes, distribución de variables… Además, dado el alto número de variables y la correlación entre ellas, se llevó a cabo un análisis de componentes principales para lidiar con estos problemas.

Se llevaron a cabo diferentes técnicas estadísticas y de visualización para evaluar la correlación entre las variables. Los análisis más relevantes fueron anova, o análisis de varianza, usado para seleccionar los casos relevantes que se estudiarían en profundidad; y regresiones, para calcular la correlación real entre las variables.

El proyecto finaliza proveyendo conclusiones. Las correlaciones para los casos relevantes seleccionados fueron analizadas y explicados cuando fue posible. Algunos de estos casos fueron rechazados debido a inconsistencias en la regresión o al sinsentido de las conclusiones que estas arrojaban.

Executive summary

Over the last few decades, there have been many theories regarding leadership and how it could impact the financial performance of the companies. However, few of these theories proved the real impact of the different leadership styles. This project assesses this question by providing an analysis using different data science techniques.

This report provides a comprehensive methodology and justifies the different tasks to perform the analysis. This analysis utilizes web scraping, an innovative information source that is rarely used in reports such as this one.

The process starts by importing the data from different sources. A process of cleaning was required by exploring missing values, outliers, distribution of variables… Additionally, given the high number of variables and the correlation between them, a principal component analysis was carried out to deal with these problems.

Different statistical and visualization techniques were used to evaluate the correlation between the variables. The most relevant ones were the anova analysis, used to select the relevant cases for further analysis; and regression, to calculate the actual correlation between the variables.

The project finishes by providing conclusions. The correlation for the selected relevant cases are analyzed and explained when possible. Some promising correlations are rejected because of the inconsistency of the regression or the illogical nature of the conclusion.

1. Introduction

Over the years, the impact of multinational companies' decisions on the different spheres that shape our society has been a hot topic. Many experts across academia claim that companies must be a vehicle to improve social welfare while making a profit to be sustainable. Companies are making significant improvement in this regard through the implementation of Corporate Social Responsibility.

The analysis of the impact of companies' policies and decisions in the society and the environment regarding Corporate Social Responsibility is a common topic in reports and research theses. However, the impact of these decisions on the companies themselves is an issue that has been poorly explored in the last few years.

Even though it is a common belief that the impact of investing in social capital is positive for the different layers of the economy, the dimension of this impact is not clear. Another topic that is still unexplored is the aspects of the Corporate Social Responsibility that a company should give priority to base on their core business to produce a greater impact.

All these questions have high importance for the sustainable development of the society. The socially concerned companies could use these conclusions to understand how to make a bigger impact. The companies focused only on revenue would know how to spend their money on social issues to get a higher return for their investors.

A quantitative approach for the analysis of these topics is required to provide a trustful analysis and to understand the importance of developing a good leadership style. Different statistical and mathematical techniques are implemented throughout the project for a professional analysis of the data.

That said, this document explores quantitatively different aspects of the leadership of the companies and the impact on their financial performance. A comprehensive methodology is provided for a better understanding of the issue, including the definition of leadership strategies, the statistical techniques, the financial indicators, and the data management techniques.

2. State of art

The interest of Social Capital and its impact in the financial/economic sphere experienced a rise after the publication of *Making Democracy Work* in 1993 [1]. In this book, Putnam explored the reasons why some governments succeeded while others failed. His research was based on the situation in Italy in 1970 when different governments were created for different regions. Putnam found patterns of associationism, trust, and cooperation that promoted the economic growth.

In 1997, Knack et al [2] started quantitatively exploring the idea that social capital is important for the economic development of countries. In this report, they started by introducing the idea of Trust and Civic, as well as a methodology to measure these variables. Then, the correlation between these variables and economic indicators like the GDP was explored.

The influence of leadership and culture on the companies' organizational performance was analyzed by Ogbonna et al [3] in 2000. They used a survey system to measure leadership and culture and reported their methodology. After carrying out different statistical methods, they found empirical evidence to support their theory.

Jean B. McGuire et al [4] analyzed the impact of companies' reputation on their financial performance. They used Fortune magazine's ratings of corporate reputations and different financial indicators to find correlations between corporate social responsibility and financial performance. They discovered that market returns and risk are more correlated with the corporate social responsibility than previous reports had claimed.

The evolution of the study of this field has moved progressively from the analysis of governments to the analysis of companies. This project will contribute to the development of this field by exploring an idea that has not been widely discussed before: the impact of companies' leadership style on their financial performance.

3. Motivation

Different leadership styles, which are based on the core business of the company, create different economic outcomes. Therefore, the main motivation of this project is to identify the different leadership factors that can potentially make a company improve its financial performance.

The second objective of this project is to prove that investing in corporate social responsibility can have a positive financial impact. This discovery would translate into more investment from the company in the development of a sustainable society.

The two main objectives of the project are going to be addressed at the same time given that the indicators of social capital and leadership style are contained in the same variables. These variables will be explained in the corresponding section.

At the end of the project, the features of leadership that can help improve the financial performance of the companies, depending on the industry in which they are developing their activities, will be identified.

4. Methodology

The project will be carried out following a well-structured process, reporting and justifying the different tasks and providing conclusions when it is possible. Among the different identified tasks, the most important ones are:

- Web scraping: data mining from different websites.
- Leadership definition: to define what each type of leadership consists of and identify the keywords for the sentimental analysis.
- Sentimental analysis: to quantify the level of leadership that each company is expressing in their mission statement.
- Data cleaning: identify wrong data, missing values, and standardize categorical values.
- Standardization of variables: standardize the numerical variables given the industry the company is operating in.
- Principal component analysis: to reduce the number of variables and obtain non-correlated variables. In this step, it would be ideal to identify the meaning of each of the relevant components.
- Regressions and anova tests: to identify the correlation between the different predictors and the predicted variables and the importance and relevancy of these correlation with statistical indicators.
- Selection of relevant cases: find the most meaningful correlation cases to carry out a deeper investigation.
- Data visualization: for a better understanding of the obtained information.
- General conclusions: analysis of the results.

5. Web scraping

5.1. Definition of web scraping

Web scraping is defined as the technique or group of techniques that extract data from websites on the internet. It has gained great importance across many industries during the last few years given the increase of speed and ease of extracting the data, processes that took very long periods of time in the past. These techniques can be classified in two main groups depending on the way data is obtained from the internet:

- Conventional web scraping: the data is extracted directly from the source. The main inconvenience of the implementation of this technique is the time required to obtain the data and the messiness of the data. There are even specific packages to deal with this messiness like BeautifulSoup. On the bright side, it is possible to obtain all the information that is wanted from the internet as long as it is not protected by the owner. The types of information that can be extracted include text, tables, pictures, videos, animations, and hyperlinks.
- API web scraping: a connection with the database of the source is established. The acquisition of the data is done easily and standardized by requesting the data to the server in a formal process. On the other hand, proper training in the use of the API might be required and the learning curve will depend on the documentation of the API, which in some cases is very poor. In addition, the required data might not be available and the use of the API might be expensive.

The different options of web scraping must be assessed to choose the proper option to obtain the data. The main requirements for the chosen technique are that it must be free and that it is able to obtain all the data that the project requires. The election and its justification are provided in the following section.

5.2. The data on the internet

5.2.1. Financial data

The indicators to measure the financial performance of the companies must describe the entire picture of the situation of the company. This includes the profit that the companies are generating, not only in terms of money, but also compared to their own size and the rest of the companies in their industry/sector; the willingness of shareholders to invest in those companies which is strongly correlated with the price of the stocks; and the economic profit that the shareholders perceive with their investment which is reflected in the dividends. Given the financial features that will be described in this paper, the following indicators were identified:

- ROE (Return on equity): the amount of net income as a percentage of shareholders' equity. It is a measure of how efficient a company is at generating profits. This is the main financial performance indicator that is going to be used in this paper.
- EBITDA (Earnings before interest, taxes, depreciation and amortization): it is a useful financial indicator to understand the size of the company and the potential earnings for the company.
- Annual dividend: the amount of money that an investor that holds a share receives at the end of the year. Dividends can be issued as stock options, cash payments, or other properties.

A complete picture of the financial situation and performance of a company can be built using these indicators to answer the questions exposed in this section. To provide further analysis of the situation of the company in the external context, the market, stock index, sector, and industry are included in the database.

The companies are classified into the following categories based on the sector and industry in which they are operating:

- Consumer goods: sector in which companies provide a finished product to the final customer:
    - Automobile industry.
    - Food and beverages industry.
    - Personal products industry.
    - Pharmacy products and biotechnology
    - Retailing.
    - Technology products.
    - Textiles, clothing and shoes.

- Consumer services: sector in which companies provide a service to the final customer:
    - Communication and publicity.
    - Healthcare.
    - Internet.
    - Leisure, tourism and hotel industry.
    - Transport and distribution.

- Financial sector: companies that provide financial services:
    - Bank.
    - Insurance.
    - Other financials.

- Manufacturing: companies that are or participate in the process of manufacturing special equipment or construct buildings:
    - Construction.
    - Engineering and others.
    - Manufacturing and assembly of capital goods
    - Materials

- Petrol and power: companies involved in the production of energy:
    - Electricity and gas.

- Technology and telecommunications: companies specialized in developing new technologies:
    - Electronics and software.
    - Telecommunications.

Having successfully identified the required information for the project, the next step is to extract this data from the internet. Fortunately, all the required information is available in the database of Expansion, a reputable newspaper specializing in financial information. The information is provided in the following format in the webpage:

Ratios fundamentales

Ratio	2018 *	2017 *	2016	2015	2014
PER	**25,35**	**28,35**	**30,87**	**33,78**	**38,97**
PRECIO / CASH FLOW	18,90	20,94	23,01	24,86	28,53
PRECIO / VALOR CONTABLE	6,32	6,94	7,64	8,50	9,31
ROE	0,25	0,24	0,25	0,25	0,24
DIVIDENDO NETO POR ACCION	0,83	0,74	0,68	0,60	0,52
RENTABILIDAD POR DIVIDENDO		2,07	1,87	1,67	2,14
BENEFICIO NETO (MILLONES)	3.819,29	3.423,00	3.157,00	2.875,00	2.500,25
EBITDA (MILLONES)	6.097,00	5.504,56	5.083,00	4.699,00	4.103,00
BPA	1,23	1,10	1,01	0,92	0,80

Figure 1: example of financial data in Expansion

With the required data and source of information defined, it is essential to identify and implement a web scraping method. When exploring the format of the webpage, some obstacles for the extraction of the data were found.

The first barrier to extracting data from Expansion is that they do not have an API. This means that conventional ways of web scraping had to be used, making the process slower and more difficult to carry out. The second barrier was the inconsistency in the format of the links leading to the information. For this reason, a simple loop to extract the information for each company was not possible.

The solution was to use an existing tool for web scraping called Import.io. This tool follows the process for web scraping described below:

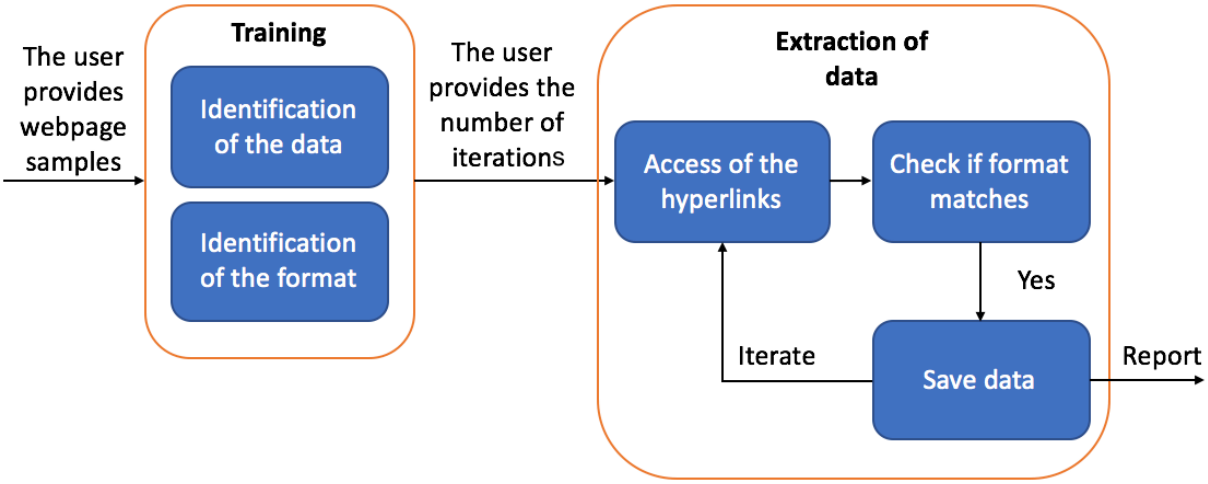

Figure 2: Import.io web scraping process

The process starts by requesting webpage samples to the user. In this part, the user shows 5 webpages to the tool with the format that the pages should have to be considered candidates for the scraping and identifies the data within those webpages. The tool receives this information and learns what the user wants to extract.

When the training process is completed, the user provides the number of iterations that the tool must perform. Then, the extraction of data starts. The tool accesses all the hyperlinks in the webpages that were provided initially and checks to see if they have relevant information. If there is relevant information, the data is stored.

When all the new webpages are checked, a new iteration is done by checking all the hyperlinks in the new webpages. This process is very time sensitive since it is $O(m^n)$, so the selection of the number of iterations must be done very carefully.

The tool was configured and the execution finished successfully, extracting all the required data in about 24 hours of computation.

5.2.2. Leadership data

The next layer of data in this project is the leadership data. The factors that drive the decisions in the management side of the company were the subject of analysis. The board of directors decide the ideology of the company and it is supposed to be a guide for the decision-making. In this paper, it is going to be assumed that their actions are consistent with their ideologies. Otherwise, the company would lose the trust of the investors.

It is difficult to extract the entire ideology or leadership system of a company since it is based on subjective concepts. To be able to capture as much information as possible about the companies, the following information was obtained for further analysis:

- Mission statement of the company: it defines the purpose of the company by identifying the product or service that they are offering, their main client, and the place in which they operate.
- Vision statement of the company: definition of the goal that the company wants to achieve in the medium/long term.
- Values of the company: brief description of the moral code of the company which will be taken into account in the decision-making process of the entire set of actions in all the layers of the company.

- Indeed reviews: obtained to quantify the good and bad aspects of working at a specific company. The following factors were obtained:
    - Overall
    - Culture and values
    - Work/life balance
    - Senior Management
    - Compensation and benefits
    - Career opportunities.

6. Sentimental Analysis

Sentimental analysis is defined as the process by which texts are categorized based on counting specific words related to different ideas. This process can consider good and bad words which would positively or negatively affect to the grade of a specific category.

In this analysis, the negative words are not considered when building the indicators since it is very unlikely that a company writes a negative concept in their mission, vision and values because this would drive away the investment.

6.1. Gallup leadership

In Assessing the Leadership Talent [5], Gallup Consulting establishes the different aspects of leadership that can predict the leader's performance:

- Direction: the ability to create goals, communicate the future of the company, develop the strategy and encourage people to work towards that vision of the company in an effective way.
- Drive: the ability to challenge people to improve their work and set high standards in the company.
- Influence: the ability to become a leader and a source of guidance for the workers in the company. This is obtained through charisma, persuasion, and emotion to build resilient trust.
- Relationship: the ability to take care of the employees, invest in them and want them to achieve their goals.
- Execution: the ability to complete the work effectively. This can be done by planning the timelines, assigning people to them and establish strategic plans for the different scenarios and situations.

With the different aspects of the Gallup leadership identified, a set of words must be defined to detect the different leadership styles with the sentimental analysis. It is important to point out that the sentimental analysis will not only consider the selected words but also the words that derivate from them:

- Direction: future, coming, imminent, expect, vision, strategy, plan, long-term, decision, direction, action, opportunity, chance, next, following, goal.
- Drive: require, overcome, demanding, improve, more, best, first, courage, advance, better, boost, increase, upgrade, extra, higher, further, new, leading, outstanding.
- Influence: motivation, hard-work, follow, encourage, incentive, influence, important, critical, work, team, group.
- Relationship: workers, learn, development, personal, achieve, team, group.
- Execution: Achieve, success, obtained, done, doing, dealing, operation, process, response, make, happen, solution.

6.2. Ignatian leadership

Saint Ignatius of Loyola was a Spanish priest that founded the Society of Jesus. He was not only a spiritual leader for the catholic community, but also a well-rounded person who wrote texts about many different topics.

He established the principles for the good development of leadership in the members of the Society of Jesus. These principles can be accepted as indicators of leadership in the management of a company since they develop emotional intelligence, one of the kinds of intelligence that the members of a board must have.

These principles are compiled by Chris Lowney in A Guide to Jesuit Leadership and Ignatian Spirituality [6] and are defined in the following categories:

- Spiritual: every leader must encourage people to behave morally and must show that the company is at the service of the people and not the opposite.
- Self-knowledge: the leadership starts with the knowledge of oneself. It is the path to achieve self-motivation.
- Trust: the relationships in a company must be based on trust. Developing trust between employees improves the performance of the company.
- Discernment: the action of being aware of the situation of the company and making good decisions based on it.
- Heroism: being confident about one's own capacity and feeling capable of overcoming the obstacles.
- Love: being committed to the human potential and work so people develop their own potential.
- Ingenuity: the capability of accepting change.

The chosen words for this sentimental analysis are identified as follows:

- Spiritual leadership: responsibility, morals, value, care, commit, environment, social, duty, noble, principle, ethic, climate, habitat, collective, people, need, right, clean.
- Self-knowledge: self-awareness, passion, drive, strength, weakness, talent.
- Trust: communication, collaboration, talent, synergies.
- Discernment: reflection, decision making, awareness.
- Heroism: big goals, ambition, aspirations.
- Love: service, community.
- Ingenuity: creativity, wit, innovation, deal with uncertainty, deal with change, flexibility.

7. Data cleaning

7.1. Missing values

There are different techniques to deal with missing values. A comprehensive description of the different options is provided in this section to justify the election of a technique. The techniques can be classified into the following categories:

- Leave the missing values: missing data can provide important information in certain analysis like the break of a sensor. In these cases, the missing data is not processed and left as a missing value.
- Fill in the missing values: there are different techniques to fill in the missing values using different hypothesis:
    - Fill in using a statistical indicator: the most common one is using the mean of the rest of the samples for that variable.
    - Fill in using the closest value of the variable in the future: if the variable is a time series, the value can be filled in with the value of the next sample.
    - Fill in using the closest value of the variable in the past: if the variable is a time series, the value can be filled in with the value of the previous sample.
- Remove the missing values: delete the samples with missing values.

Given that the data can be very time sensitive and the missing values do not provide any kind of information, it was decided that the samples with missing values would be removed from the data set.

7.2. Outliers

The outliers of the variables must be analyzed to conclude if they are actual outliers of the variables or wrong data. A table with different statistical functions is generated to help identify the outliers:

ROE	Dividend	Profit	Price	EBITDA
Min. :-0.5100	Min. :0.000	Min. :-0.70943	Min. : 1.07	Min. :-116.15
1st Qu.: 0.0900	1st Qu.:0.035	1st Qu.:-0.09981	1st Qu.: 18.39	1st Qu.: 788.7
Median : 0.1500	Median :0.675	Median : 0.03828	Median : 53.31	Median : 2433.0
Mean : 0.1843	Mean :1.062	Mean : 0.04218	Mean : 73.35	Mean : 6587.3
3rd Qu.: 0.2175	3rd Qu.:1.720	3rd Qu.: 0.17925	3rd Qu.: 93.32	3rd Qu.: 9056.3
Max. : 1.1500	Max. :7.300	Max. : 1.17783	Max. :758.88	Max. :82487.0

PER	Overall	workLife_balance	Compensation	Career_options
Min. : 0.00	Min. :2.800	Min. :2.000	Min. :2.600	Min. :2.400
1st Qu.: 13.00	1st Qu.:3.900	1st Qu.:3.600	1st Qu.:3.600	1st Qu.:3.225
Median : 19.13	Median :4.000	Median :3.800	Median :3.800	Median :3.400
Mean : 27.58	Mean :3.987	Mean :3.743	Mean :3.766	Mean :3.390
3rd Qu.: 25.68	3rd Qu.:4.175	3rd Qu.:3.900	3rd Qu.:4.000	3rd Qu.:3.500
Max. :624.30	Max. :4.600	Max. :4.500	Max. :4.500	Max. :4.300

Senior_management	Culture	DirectionPoints	DrivePoints	InfluencePoints
Min. :0.100	Min. :0.600	Min. :0.000000	Min. :0.00000	Min. :0.00000
1st Qu.:3.325	1st Qu.:3.600	1st Qu.:0.006055	1st Qu.:0.00778	1st Qu.:0.00000
Median :3.600	Median :3.800	Median :0.014600	Median :0.01482	Median :0.01046
Mean :3.499	Mean :3.751	Mean :0.018207	Mean :0.01947	Mean :0.01189
3rd Qu.:3.700	3rd Qu.:4.000	3rd Qu.:0.026013	3rd Qu.:0.02388	3rd Qu.:0.01808
Max. :4.300	Max. :4.600	Max. :0.100000	Max. :0.16667	Max. :0.06667

RelationshipPoints	ExecutionPoints	SpiritualPoints	SelfKnowledgePoints	TrustPoints
Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.000000	Min. :0.000000
1st Qu.:0.02057	1st Qu.:0.00000	1st Qu.:0.01410	1st Qu.:0.000000	1st Qu.:0.000000
Median :0.03648	Median :0.01087	Median :0.02800	Median :0.000000	Median :0.000000
Mean :0.04229	Mean :0.01294	Mean :0.03247	Mean :0.002952	Mean :0.002022
3rd Qu.:0.06045	3rd Qu.:0.01968	3rd Qu.:0.04559	3rd Qu.:0.005078	3rd Qu.:0.000000
Max. :0.23913	Max. :0.07143	Max. :0.16038	Max. :0.029850	Max. :0.050000

DiscernmentPoints	HeroismPoints	LovePoints	IngenuityPoints
Min. :0.0000000	Min. :0.0000000	Min. :0.000000	Min. :0.000000
1st Qu.:0.0000000	1st Qu.:0.0000000	1st Qu.:0.000000	1st Qu.:0.000000
Median :0.0000000	Median :0.0000000	Median :0.006450	Median :0.000000
Mean :0.0001143	Mean :0.0006816	Mean :0.009648	Mean :0.005811
3rd Qu.:0.0000000	3rd Qu.:0.0000000	3rd Qu.:0.014892	3rd Qu.:0.006783
Max. :0.0051300	Max. :0.0270300	Max. :0.066670	Max. :0.078950

Table 1: Basic statistics

The previous table shows that the maximum EBITDA is very big compared to the mean. A histogram is plotted for a better visualization of the distribution of the data:

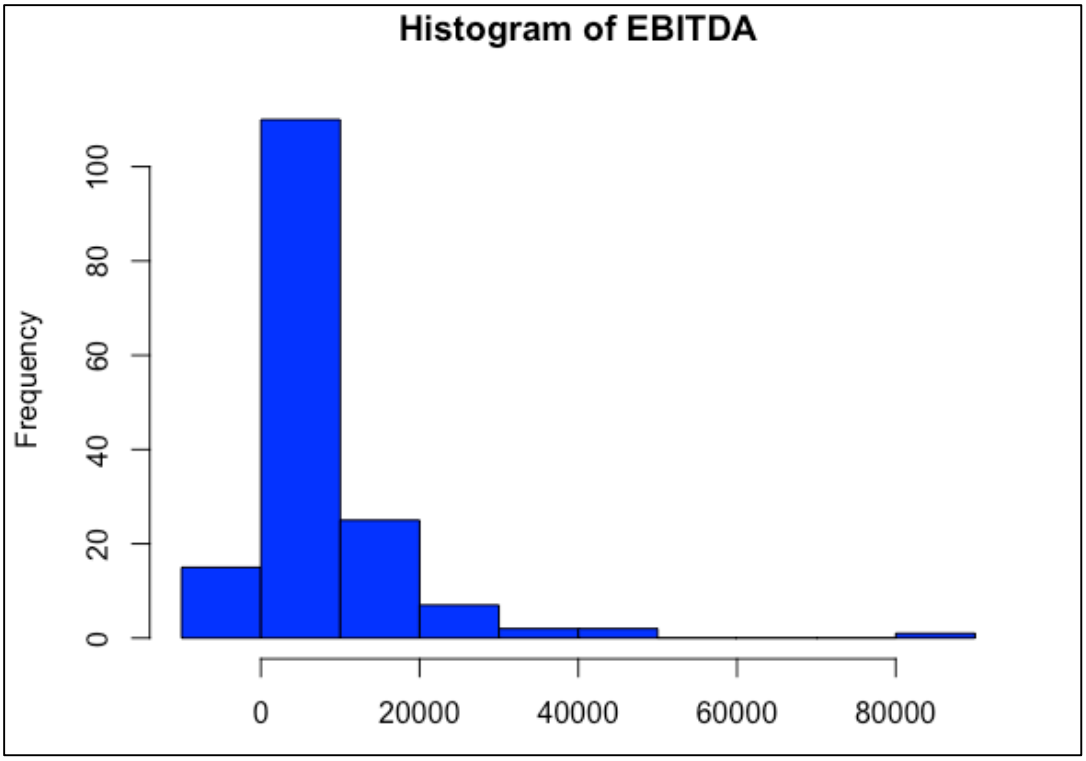

Figure 3: Histogram of EBITDA

There is clearly an outlier for the EBITDA variable. The first step to identify if the sample is an outlier consists of identifying the company with the outlier. The company with this high EBITDA is Apple. It can be concluded that this is not wrong data given the high levels of income of the company.

7.3. Variable selection

The variables in a correlation analysis must have sufficient variability so that they are meaningful. Therefore, an analysis of the variability of the variables is required to potentially narrow down the list by removing variables. Table 1 provides information on what variables might not have enough variability to be considered relevant.

Companies with discernment points in their statements are not very frequent and the level of importance is very low. A histogram is a good plot to visualize this information.

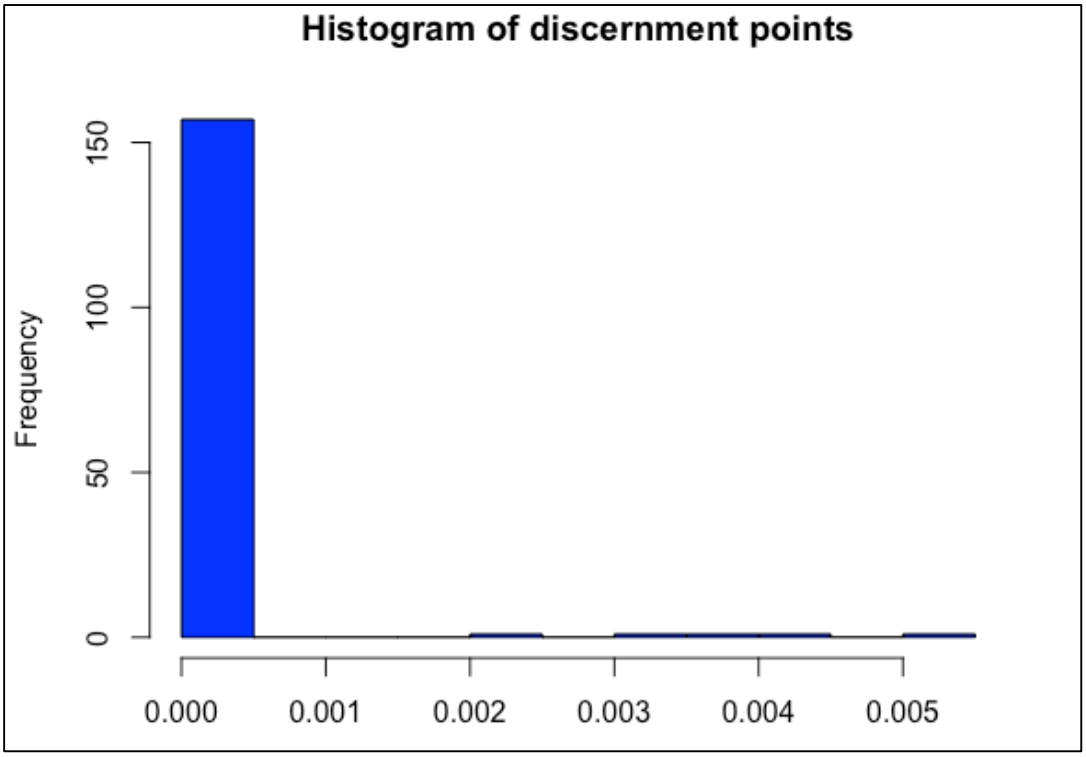

Figure 4: Histogram of discernment points

The figure shows that only five companies have discernment points with a maximum value of 0.5%. Therefore, it was decided to exclude this variable from the analysis. Another variable with very low levels of importance is the heroism. A histogram is provided for heroism points as follows:

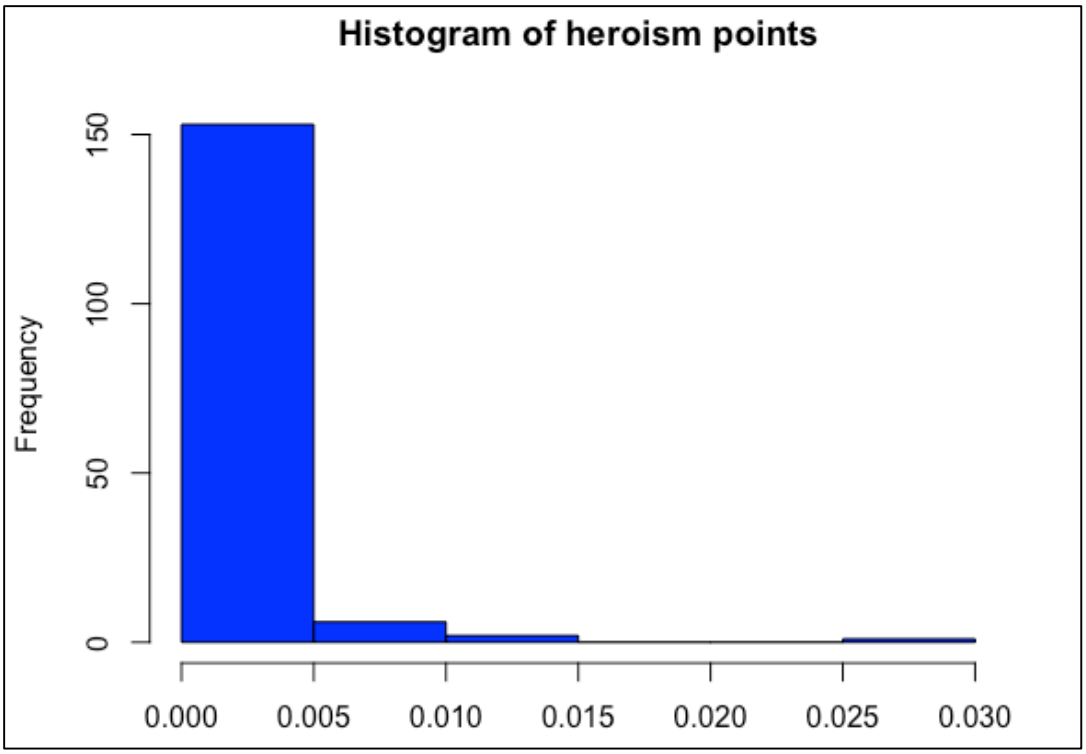

Figure 5: Histogram of heroism points

Most of the values have 0% heroism points and only one sample have a relevant ratio of 3%. With these levels of importance, this variable can not be considered relevant for the analysis and is therefore removed from the data set.

8. Data exploration

As a first step to analyzing the correlation between the financial performance and the culture of the companies, exploring the data is required to understand the story that it is being told. Different statistics and visualizations are going to be provided to justify the future steps of the project.

8.1. Indeed data

As explained before, the data from Indeed was collected to have a picture of what the employees think about the company they work for. The distribution and the correlation between these variables are explored in the following plot:

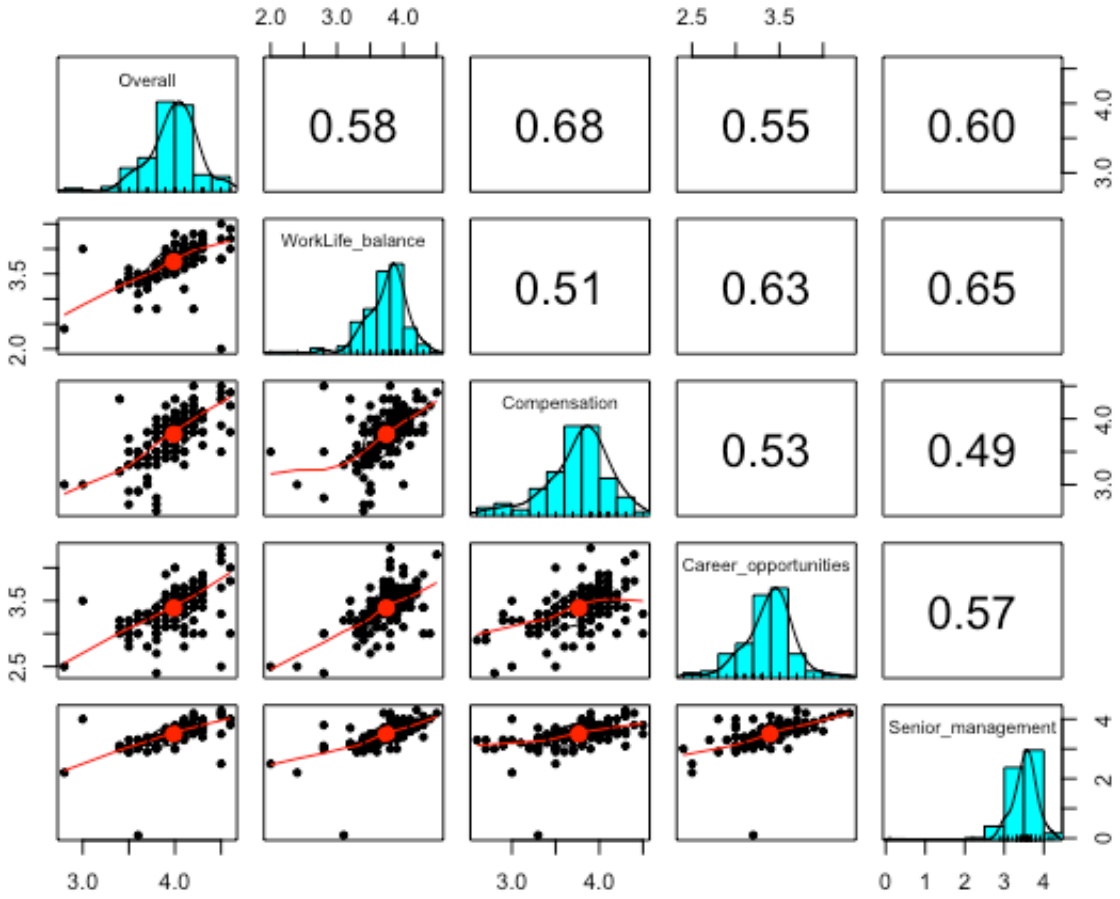

Figure 6: EDA of Indeed data

There is a lot of information contained in this plot. The diagonal shows distribution of the variables. The squares on the lower-left part show a scatter plot of the variables with a regression line that shows roughly how these variables are correlated. The squares on the top-right part show the correlation coefficient of the variables. The following conclusion can be obtained with the observation of this information:

- The variables are normally distributed, which is a good feature to build prediction models to understand the data, which is the goal of this project.
- There is a high correlation between the variables. The scatter plot and the correlation coefficients prove it. The fact that there is a strong correlation between the variables implies that:
    - The future correlation analysis will have big errors given the interaction between the variables.
    - A dimensionality reduction analysis will be required in order to get variables that represent the variance of the data.
- The variance of the variables needs to be explored further to conclude if the variables are meaningful for the analysis.

The following table shows quantitative indicators of the distribution of the data:

Overall	WorkLife balance	Compensation	Career_opportunities	Senior_management
Min: 2.800	Min: 2.000	Min: 2.600	Min: 2.400	Min: 0.100
1st Qu: 3.900	1st Qu: 3.600	1st Qu: 3.600	1st Qu: 3.225	1st Qu: 3.325
Median: 4.000	Median: 3.800	Median: 3.800	Median: 3.400	Median: 3.600
Mean: 3.987	Mean: 3.743	Mean: 3.766	Mean: 3.390	Mean: 3.499
3rd Qu: 4.175	3rd Qu: 3.900	3rd Qu: 4.000	3rd Qu: 3.500	3rd Qu: 3.700
Max: 4.600	Max: 4.500	Max: 4.500	Max: 4.300	Max: 4.300

Table 2: statistics of Indeed data

Given that these variables are on a scale of 0 to 5, the difference between the first and third quartile shows that the variance of the variables is enough for these variables to be representative for the analysis.

8.2. Gallup leadership indicators

Following the same methodology as with the Indeed data, the following plot shows the distribution and correlation of the Gallup leadership variables:

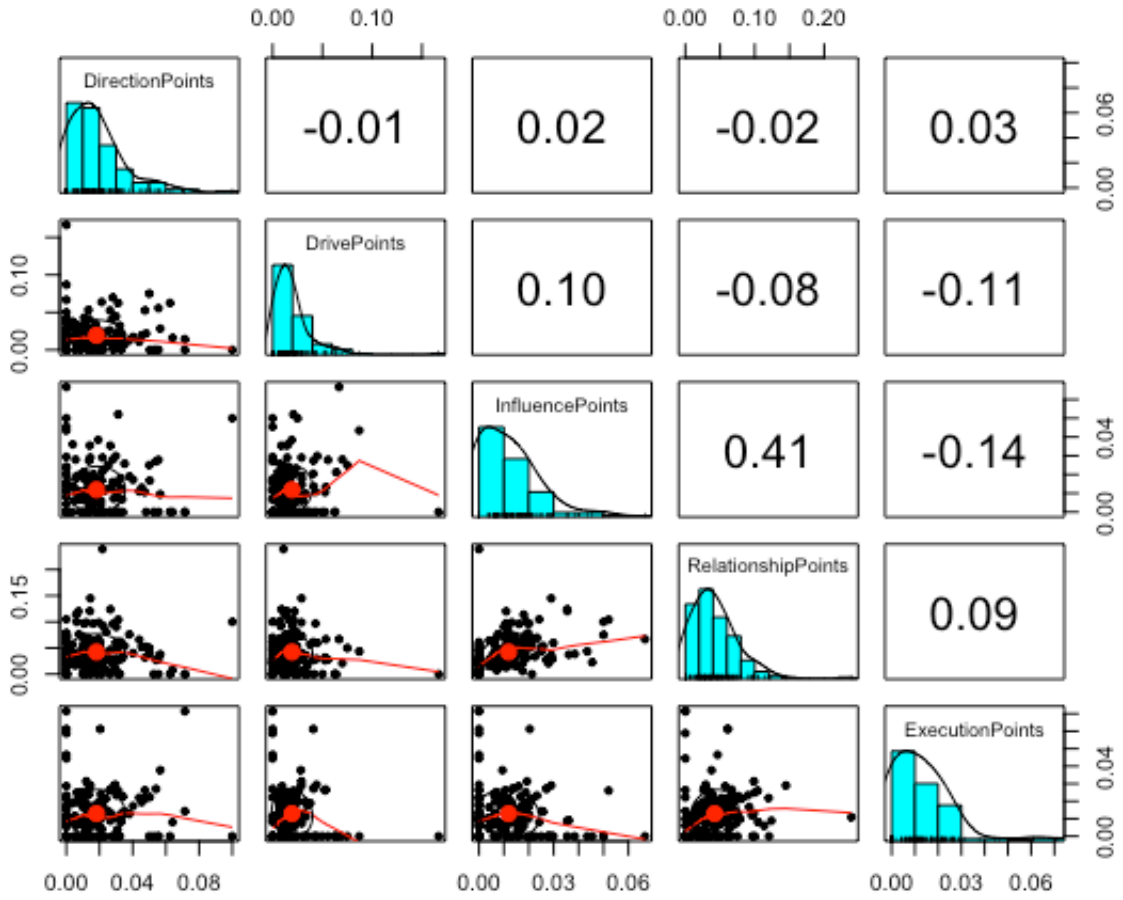

Figure 7: EDA of Gallup data

The following conclusions can be obtained with the observation of the plot:

- There is not a relevant correlation between the variables. The only correlation that is important and should be taken into account for further analysis is the relationship

between the influence and the relationship points. This idea will be explored during the PCA analysis.

- There is a significant portion of the samples with a value equal to zero for all the variables. However, as seen on the scatter plot, there are also many samples with values other than zero that will contribute to the analysis.

8.3. Ignatian leadership indicators

The following plot shows the distribution and correlation of the indicators of leadership defined by San Ignacio de Loyola:

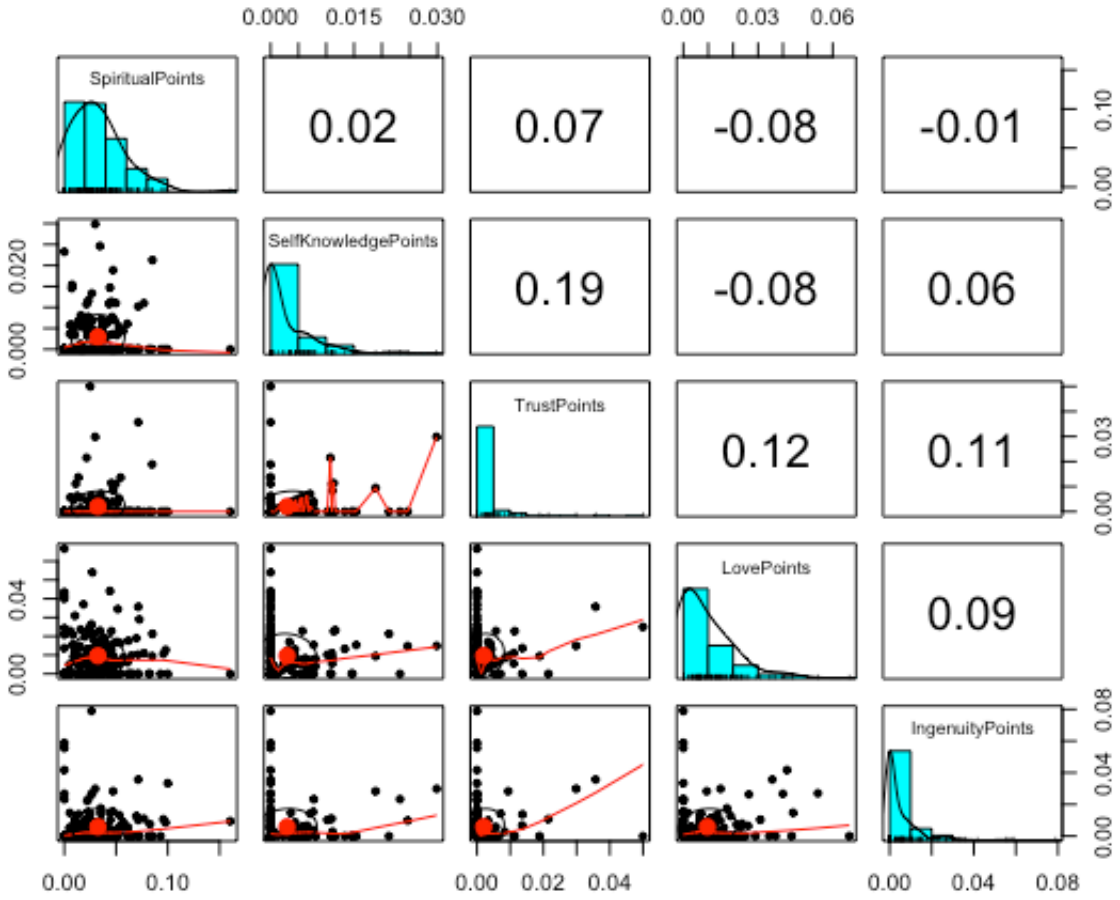

Figure 8: EDA of Ignatian leadership

The following conclusions can be obtained with the observation of the plot:

- There is not a strong correlation between any of the variables, which makes the analysis easier.
- The degree of appearance of these concepts in the culture of the companies is not very high. Therefore, it might be a good idea to change the degree of appearance for a boolean (True or False) to represent whether or not they exist.

9. Standardization

The process of standardization consists of calculating how many standard deviations fall between each sample and the mean. To calculate the standardized samples the following equations must be used:

$$x_s = \frac{x - \mu}{\sigma}$$

This is a very common analysis that is used with the following problems for data analysis:

- When the variables are in different scales and the purpose of the analysis is to study correlations or to make predictive/classification models.
- When scale sensitive analyses are carried out and the different units would make the model fail.
- When a variable is wanted to be analyzed taking into account the label of a different variable.

The project encounters all the previous problems. The culture indicators are measured with different units. Some of them are measured with a grade between 0 and 5, whereas others are measured with a ratio between the number of appearances in a text and the total number of words.

As it will be explained later in the project, an analysis called principal component analysis is going to be carried out to reduce the number of predictors to analyze the correlations. This analysis is very sensitive to the scale of the data. The standardization of the data is mandatory to deal with this problem.

As the different industries and sectors have different characteristics, the financial performance of the companies has limits depending on the industry they in which they operate. For this reason, the process of standardization is going to be done with the means of the variables grouped by industry.

10. Principal component analysis

Principal component analysis is a statistical method that is based on orthogonal projections to convert a set of n variables into a non-correlated set of variables with equal or lower number of dimensions. This method is widely used in the following cases:

- To reduce the number of variables to carry out further analysis by selecting the principal components that capture most of the variance of the data and therefore, most of the information is captured.
- To deal with correlation between variables that are going to be used to model correlation with another variable. The set of variables that are returned from the principal component analysis are not correlated.

The principal component analysis is based on the directions of maximum variance in the data. For this reason, this method is very scale-sensitive. This problem has been addressed previously by standardizing the variables. This effect can be seen in the following example plot in which the principal component analysis is carried out for the same data with different scales.

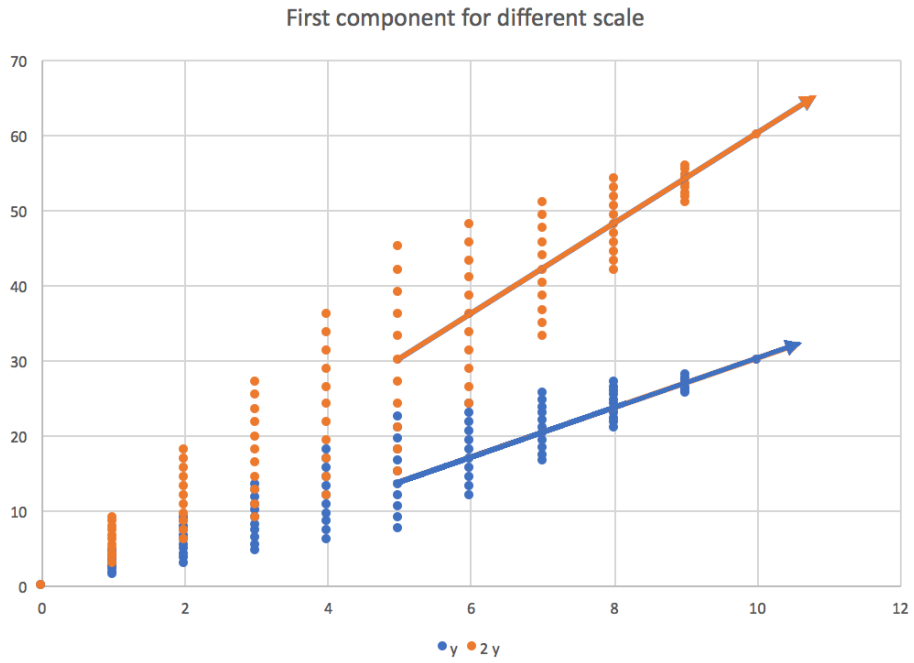

Figure 9: First component with different scale

For this analysis, it was decided to carry out independent principal component analyses for the three set of variables in the data set: indeed indicators, Gallup indicators, and Ignatian indicators. The motivation for this decision was to find, if the principal component analysis is reasonable, one variable for each of the set of indicators.

As one of the motivation to perform principal component analysis is to reduce the number of variables for further analysis, choosing the number of components to consider (hopefully one for each set of variables) is a critical decision. A common technique to make this decision is to plot the additional variance explained by each component, identify the elbow of the function and select the number of components the explain most of the variance.

After performing the principal component analysis with the data. The following plots are created with the explained variance. The components are ordered in the x-axis based on the additional variance that the component explains. This makes it easier to identify the elbow of the function and therefore, the number of components.

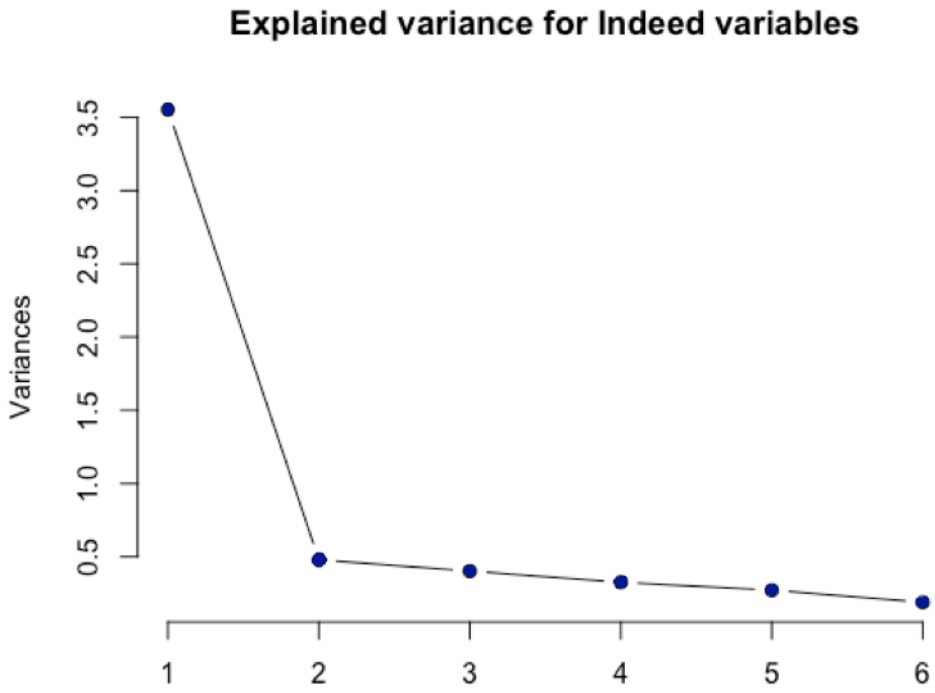

Figure 10: explained variance 1

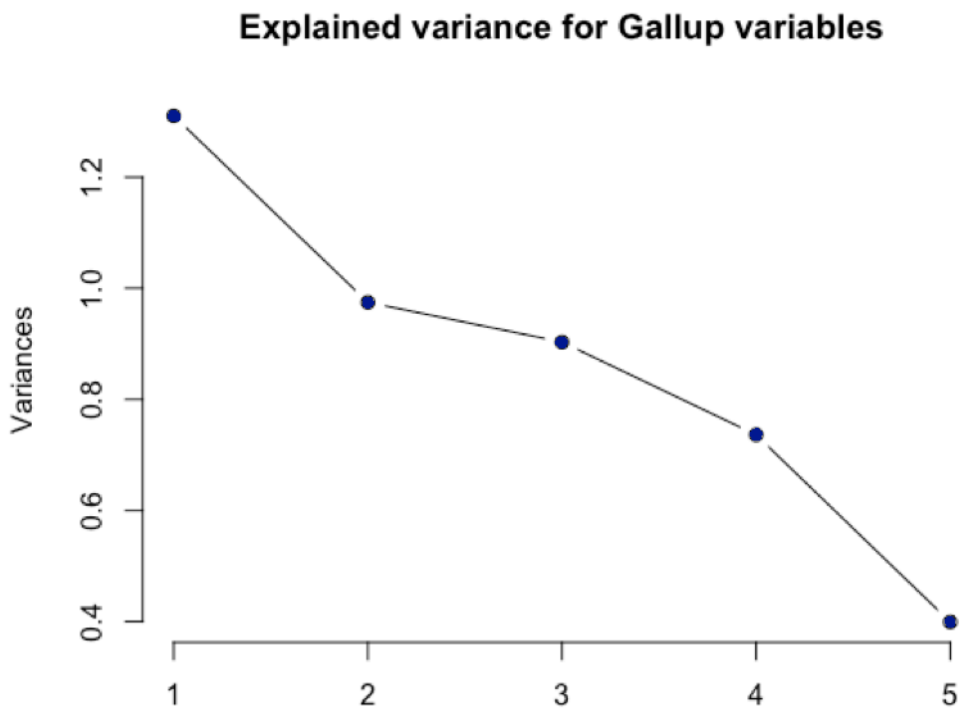

Figure 11: explained variance 2

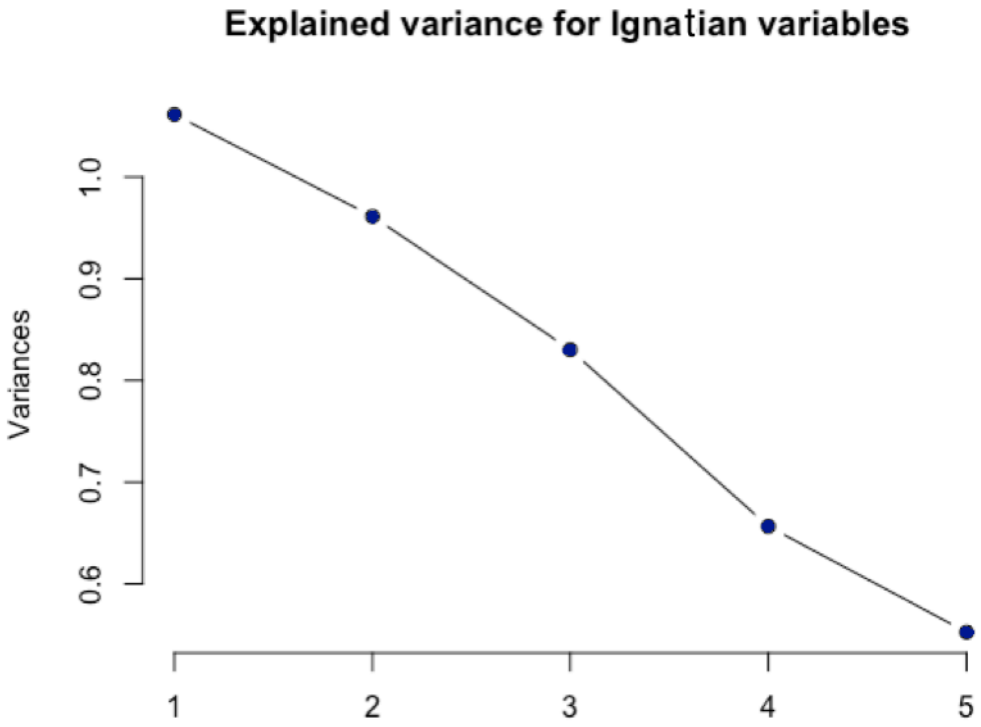

Figure 12: explained variance 3

The previous plots showed that the principal component analysis is only relevant for the variables included in the Indeed set. This is explained by the existence of a clear elbow and a lack of incremental explained variance in the principal components. The explained variance in the other set of variables follows a straight line, which means that all of them would be relevant for the analysis.

Given the results obtained with the principal component analysis, it was concluded that only the first component from the Indeed set of variables will be taken into account. In the case of the remaining variables, the principal component analysis has been proved useless and therefore the original variables are kept for further analysis.

11. Correlation analysis

After performing the previous steps described in this paper, a data analysis can be performed. A brief reminder of the purpose of the project is provided as follows. The ROE is an indicator of the financial performance of the company. The hypothesis of this paper is that the ROE can be affected by the leadership strategy of the company. In this section, this hypothesis is going to be explored by running different statistical and visualization techniques.

It was decided to analyze the different ROEs depending on the industry in which the company is operating. To justify this election, it must be proven that there are significant differences between industries and within each industry. The following box and whiskers plot is provided to explore this information.

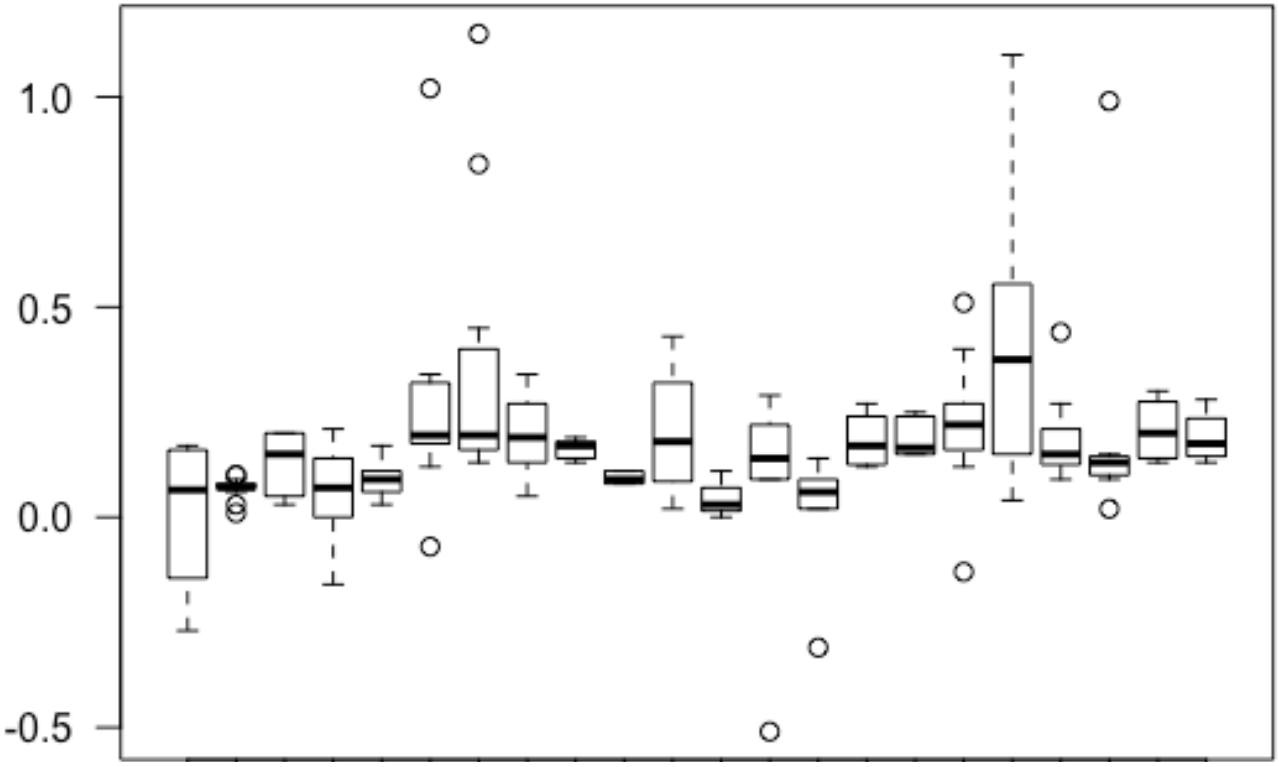

Figure 13: Boxplot of ROE by industry

By looking at the previous plot, several conclusions can be determined. The first one is that the ROE varies significantly depending on the industry and therefore it makes sense to perform independent studies. The second conclusion is that the ROE varies significantly in some of the industries and a study of the correlation might bring important insights.

Different techniques are performed to evaluate if there is actually a correlation between the leadership styles and the ROE.

11.1. Correlation exploration

Although in many cases the correlation between variables a variable depends not only on a set of variables, but also in the interaction between the variables, in other cases, the correlation is direct. For this reason, the ROE is going to be plotted against the different variables within the different industries.

Many plots showed promising patterns that will be further explored with other methods. There plots that unexpectedly suggest that the ROE is reduced with certain leadership styles. Different statistical methods must be carried out to understand what is happening in these cases.

Some of the plots with interesting information and in which the correlation is clear are shown as follows:

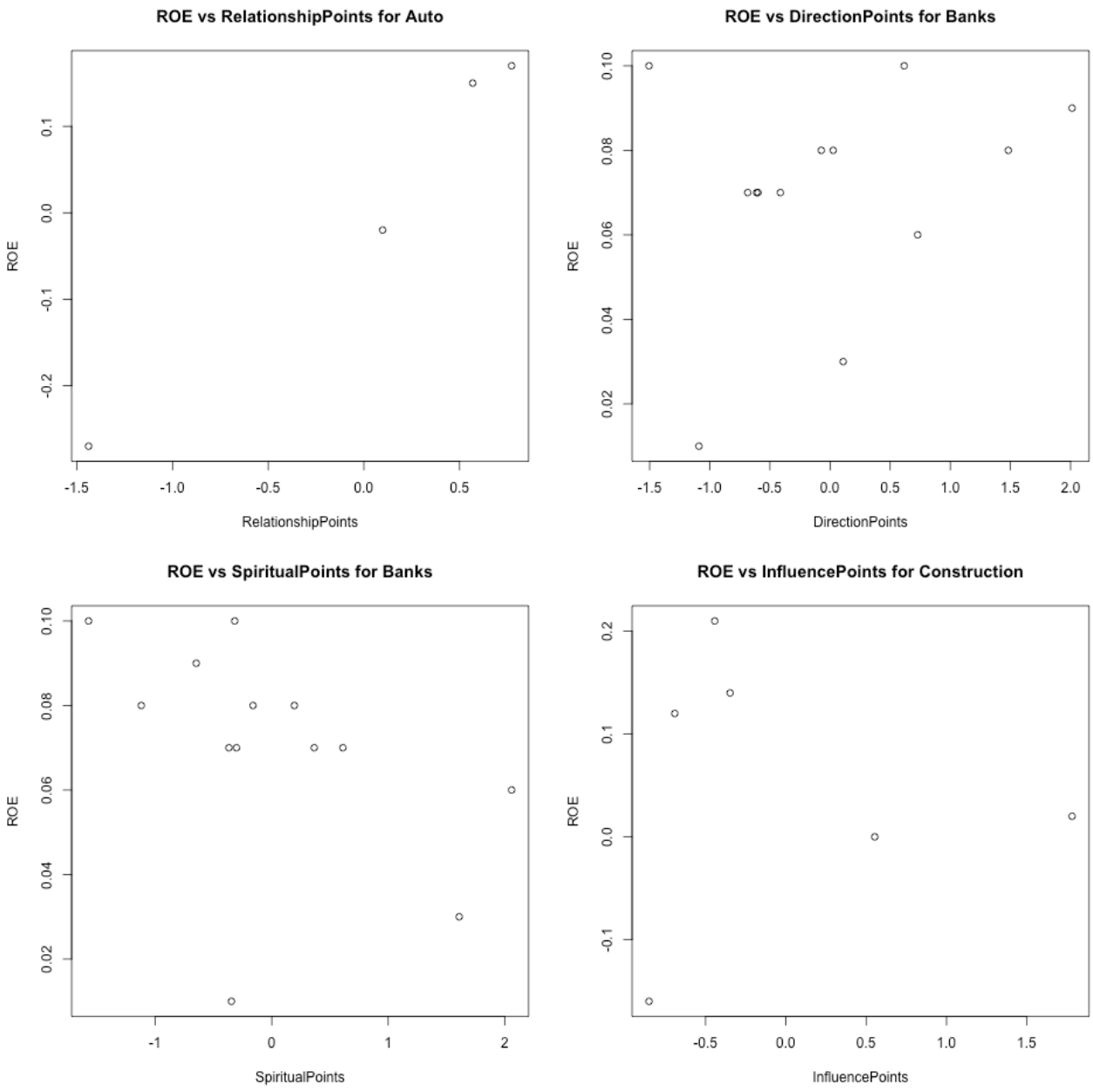

Figure 14: Examples of relationship 1

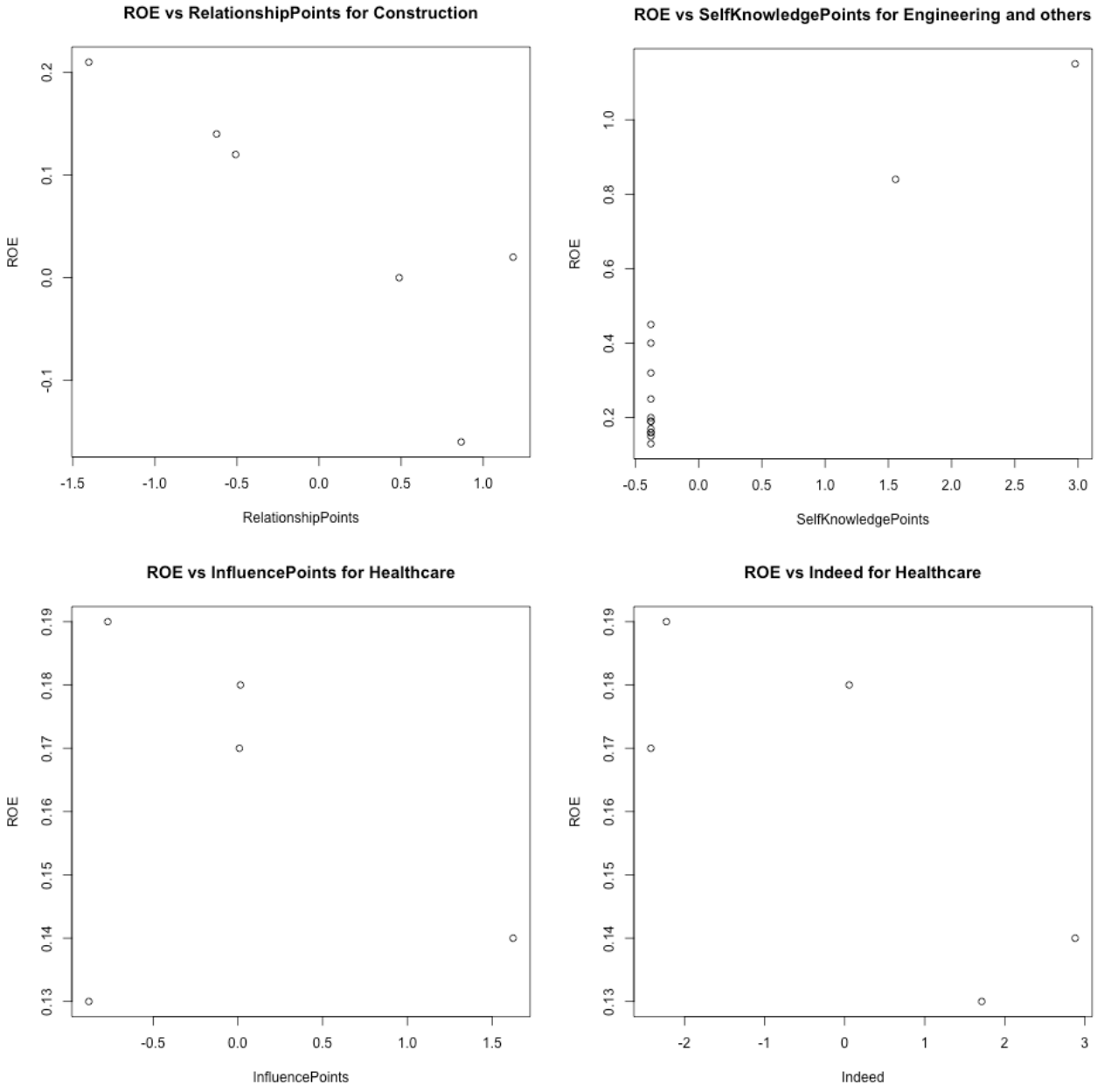

Figure 15: Examples of relationship 2

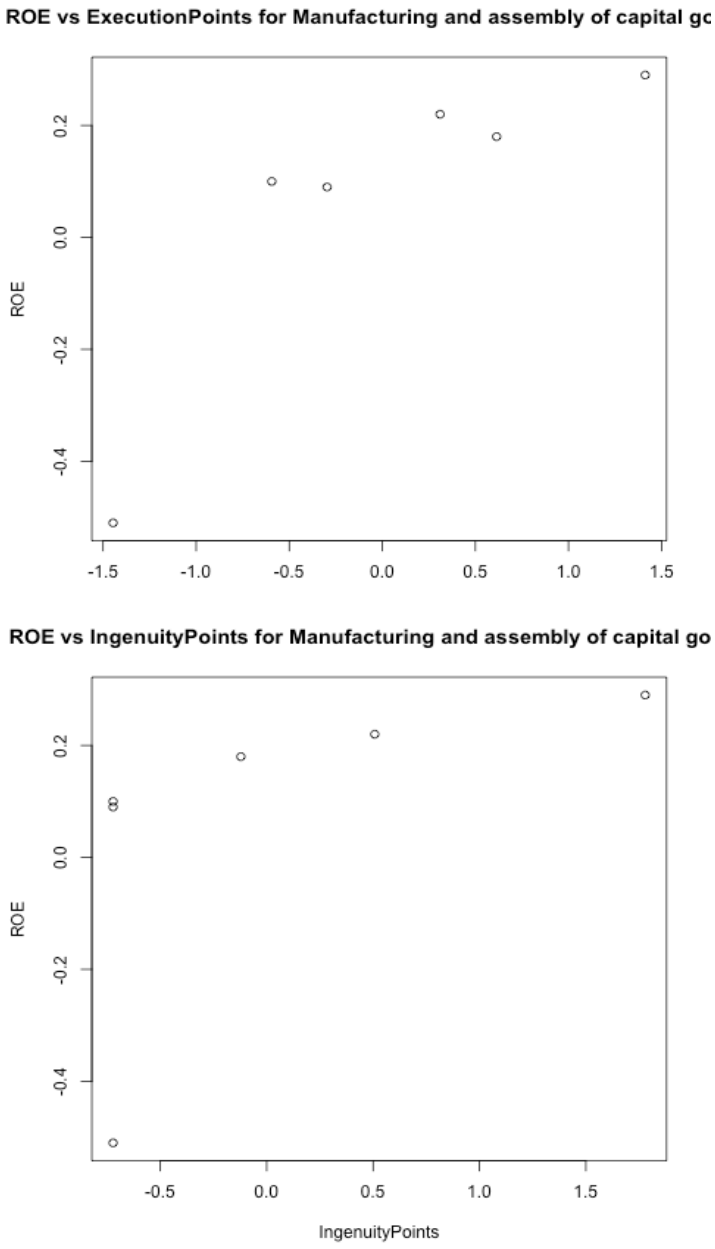

Figure 16: Examples of relationship 3

11.2. Anova analysis

The existence of relationships between leadership factors and ROE could be seen through different plots in the previous sections. However, a more detailed and quantitative analysis is essential to fully understand the correlation of the variables.

It was decided to run anova analyses to narrow down the interactions that must be explained and to quantify these interactions. In the annex 1, all the anova results can be checked. Below, the table shows the results of the analysis that are considered relevant. It was decided that the analysis is relevant if the p-value is smaller than 5%.

Industry	Variable	p-value
Engineering and others	SelfKnowledgePoints	3.43E-07
Electricity and gas	SelfKnowledgePoints	0.04801629
Materials	TrustPoints	0.002545835
Construction	RelationshipPoints	0.031662883
Technology products	RelationshipPoints	0.016719009
Technology products	TrustPoints	0.034588822
Textiles, clothing and shoes	SpiritualPoints	0.046045274
Retailing	ExecutionPoints	0.030581539
Retailing	SpiritualPoints	0.005118812
Food and beverage	TrustPoints	0.010786312
Personal products	InfluencePoints	0.035203032
Auto	RelationshipPoints	0.012992248
Manufacturing and assembly of capital goods	ExecutionPoints	0.029897927

Table 3: Anova for relevant cases

With these results, two goals have been accomplished. The first one is to prove quantitatively the existence of the correlations and their relevancy. The second is to identify the relevant cases to analyze in detailed their relationships. These cases are going to be used in the following section to calculate the function that these correlations follow.

11.3. Regression analysis

The final step of the project consists on figuring out whether if there is a positive or negative correlation between leadership styles and ROE for the relevant cases described in the previous section. After careful visualization of the data, it was decided that the regressions could only follow increasing or decreasing relationships with linear or exponential distributions. The r squared, which is a measure of how well a regression is fitting the data, was calculated for a linear and an exponential regression for the relevant cases. The results are shown in the following table:

Case	Linear	Exponential
Engineering and others - SelfKnowledgePoints	0.89362766	0.88012317
Electricity and gas - SelfKnowledgePoints	0.310206456	0.309474669
Materials - TrustPoints	0.918750614	0.907012167
Construction - RelationshipPoints	0.724137827	0.729989068
Technology products - RelationshipPoints	0.713780776	0.727133412
Technology products - TrustPoints	0.623898838	0.641518625
Textiles, clothing and shoes - SpiritualPoints	0.910029618	0.911314304
Retailing - ExecutionPoints	0.38769313	0.385435054
Retailing - SpiritualPoints	0.559966121	0.573547507
Food and beverage - TrustPoints	0.757864568	0.751084535
Personal products - InfluencePoints	0.710010921	0.712773474
Auto - RelationshipPoints	0.974184303	0.984923813
Manufacturing and assembly of capital goods - ExecutionPoints	0.731514038	0.647710218

Table 4: R-squared for relevant cases

Some conclusions can be extracted from this table. The first conclusion is that the R^2 value is very similar for the exponential and linear model for all of the cases. Since the relevant information for this project is the trend, positive or negative, it is going to be assumed that the data follows linear distributions for all of the cases.

Additionally, the R^2 values for most of the models are very high which proves that the regression is very accurate. However, there are several values with a low R^2. The reason for this low R^2 is that there are many samples with the predictor equals to zero that increases the sum of residuals.

An example of this behavior is displayed on Figure 19. This could be a problem to make a predictive model, but that is not the purpose of the project. Since the correlation was proved with the anova analysis, the relevant information from the regression is the sign of the slope. The R^2 error says that these models don't fit well the data, which is true, but the sign of the slope of the function is correctly calculated. These cases need to be analyzed carefully.

The following step is therefore to calculate whether if there is a positive or a negative correlation between the variables in the relevant cases. The following table shows these results:

Case	Slope	Sign
Engineering and others - SelfKnowledgePoints	0.283450173	Positive
Electricity and gas - SelfKnowledgePoints	0.020012109	Negative
Materials - TrustPoints	0.155030694	Negative
Construction - RelationshipPoints	0.111570288	Negative
Technology products - RelationshipPoints	0.103784834	Positive
Technology products - TrustPoints	0.097030565	Positive
Textiles, clothing and shoes - SpiritualPoints	0.077254628	Negative
Retailing - ExecutionPoints	0.194331011	Positive
Retailing - SpiritualPoints	0.233549495	Positive
Food and beverage - TrustPoints	0.095193986	Positive
Personal products - InfluencePoints	0.038675009	Negative
Auto - RelationshipPoints	0.201048577	Positive
Manufacturing and assembly of capital goods - ExecutionPoints	0.247969013	Positive

Table 5: Slope for relevant cases

It is important to point that when reading the table, the magnitude of the slope is not relevant because the analysis has been done with standardize values and it might have been skewed due to a high number of samples equal to zero in the analysis. The complete results of the linear regression analysis can be checked on the annex 2.

11.4. Visualization of the results

The relevant cases are plotted with their respective linear regressions for a better understanding:

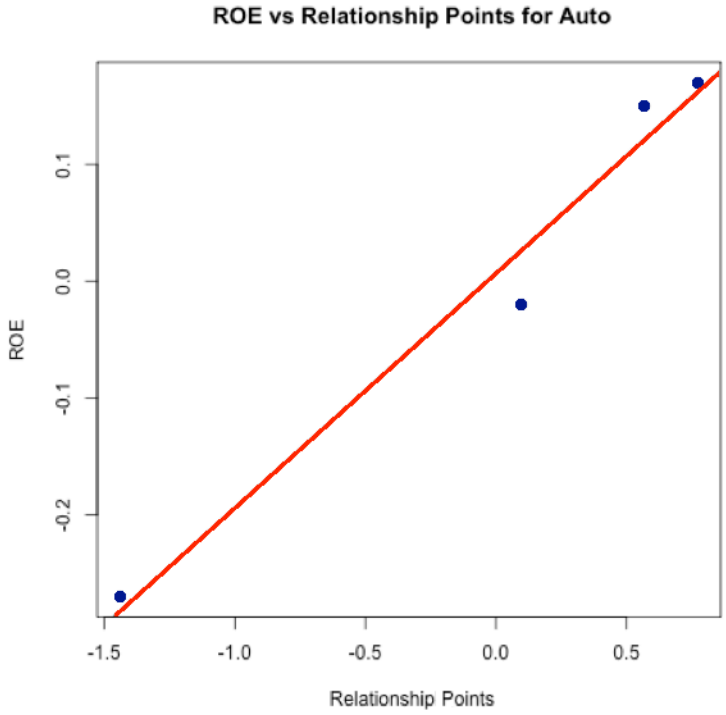

Figure 17: Relevant case 1

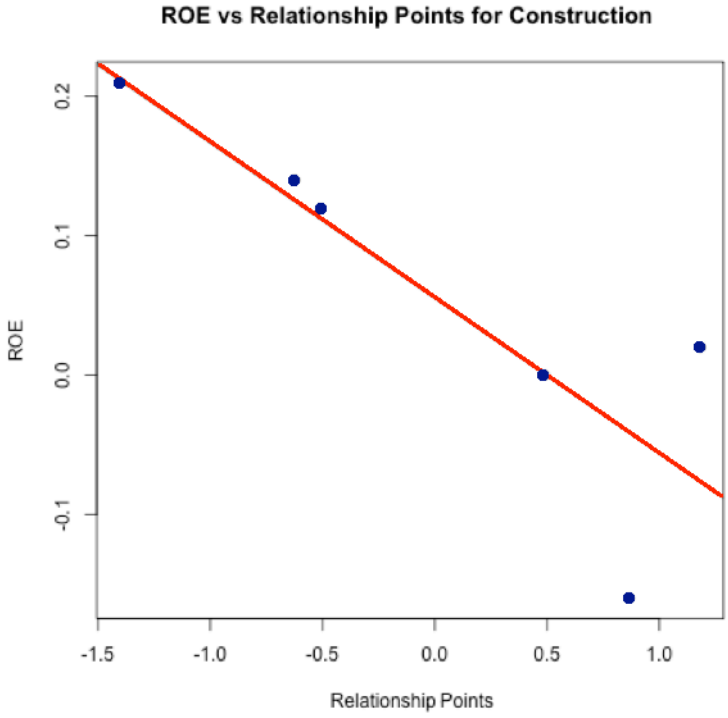

Figure 18: Relevant case 2

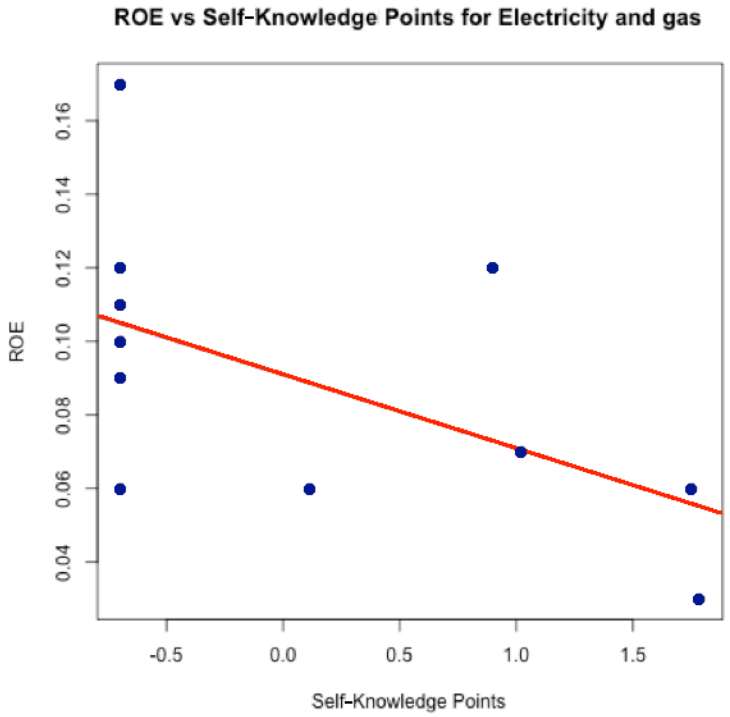

Figure 19: Relevant case 3

Figure 20: Relevant case 4

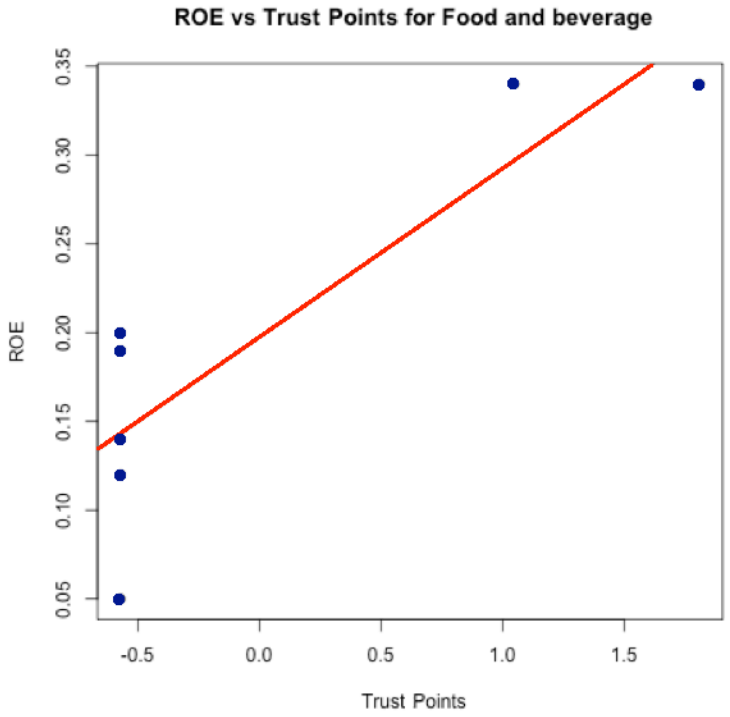

Figure 21: Relevant case 5

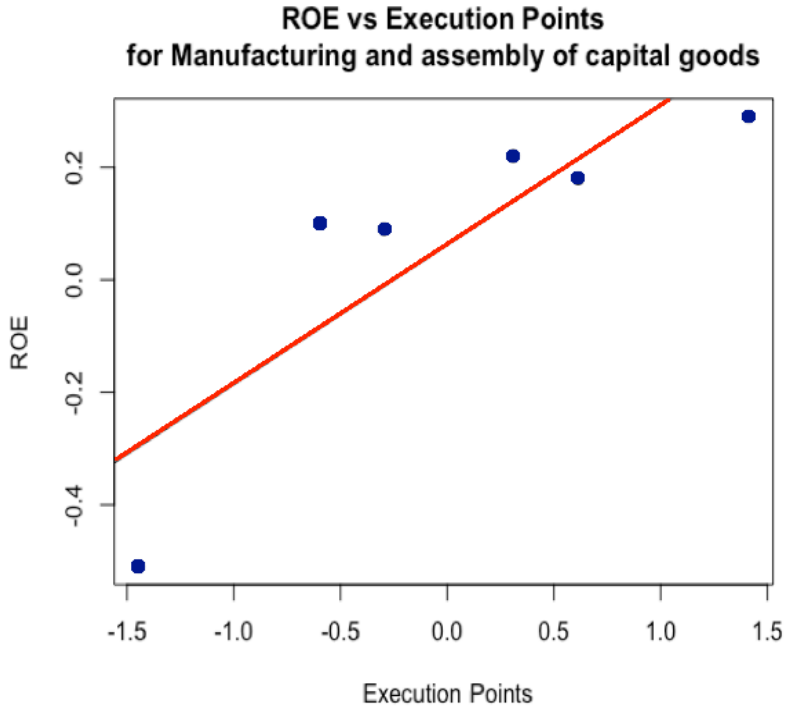

Figure 22: Relevant case 6

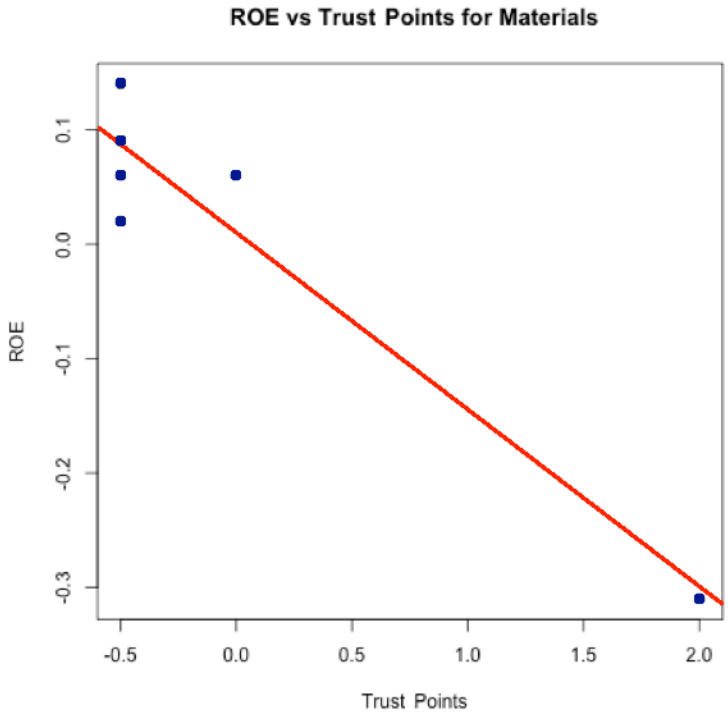

Figure 23: Relevant case 7

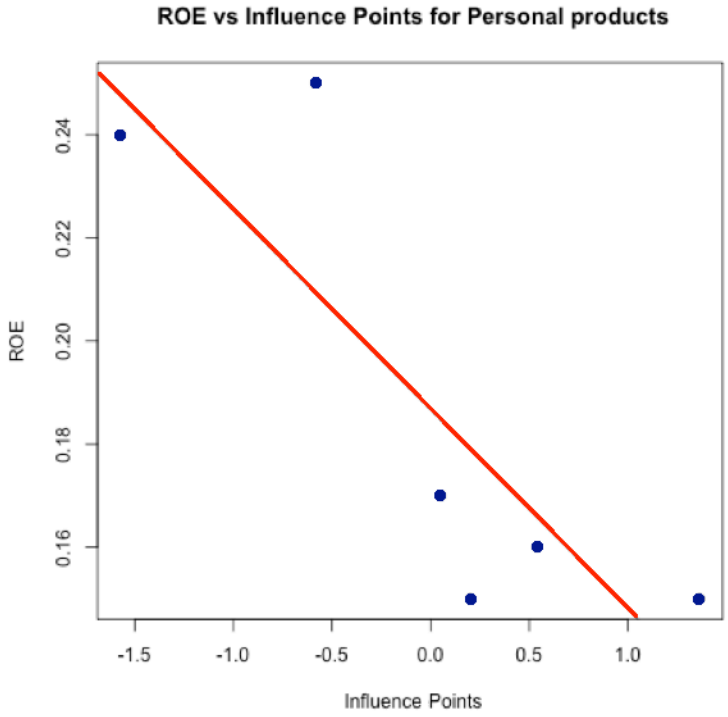

Figure 24: Relevant case 8

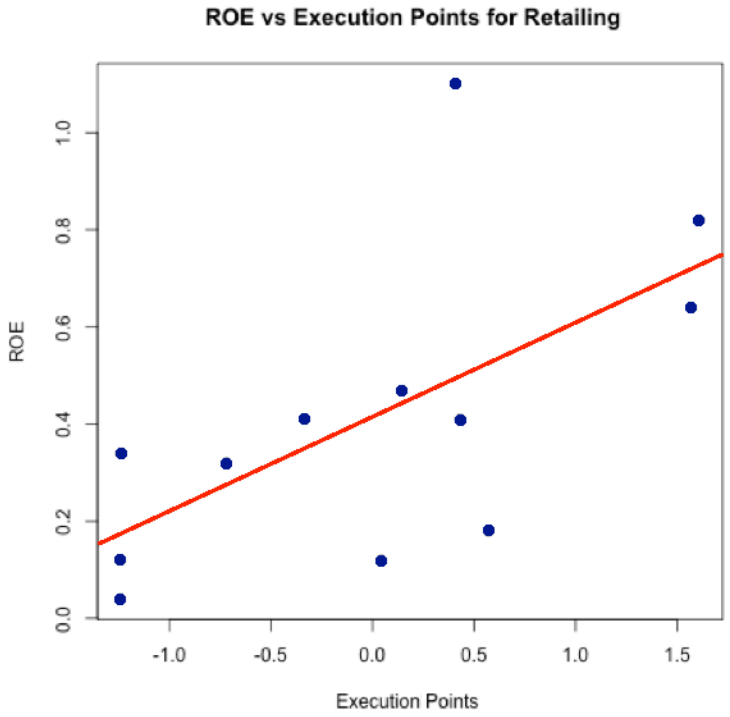

Figure 25: Relevant case 9

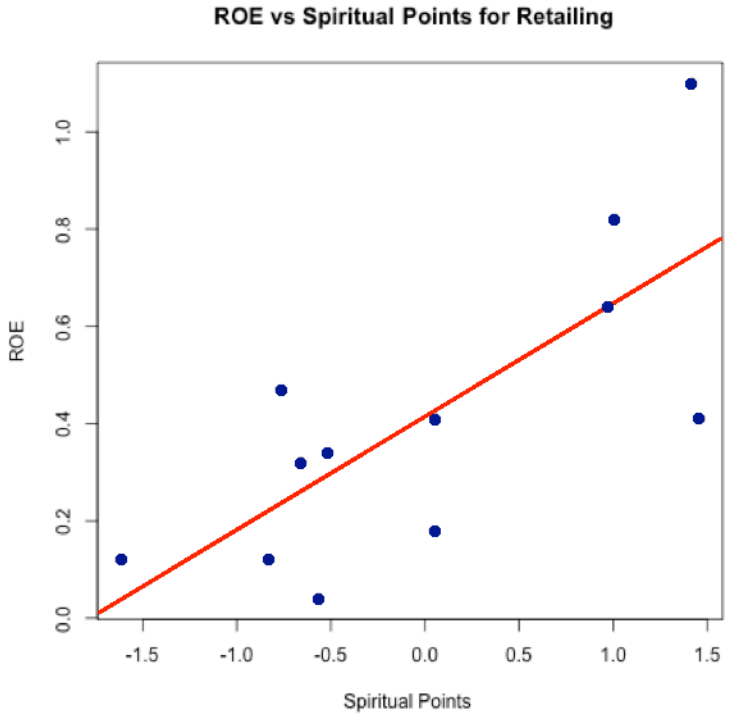

Figure 26: Relevant case 10

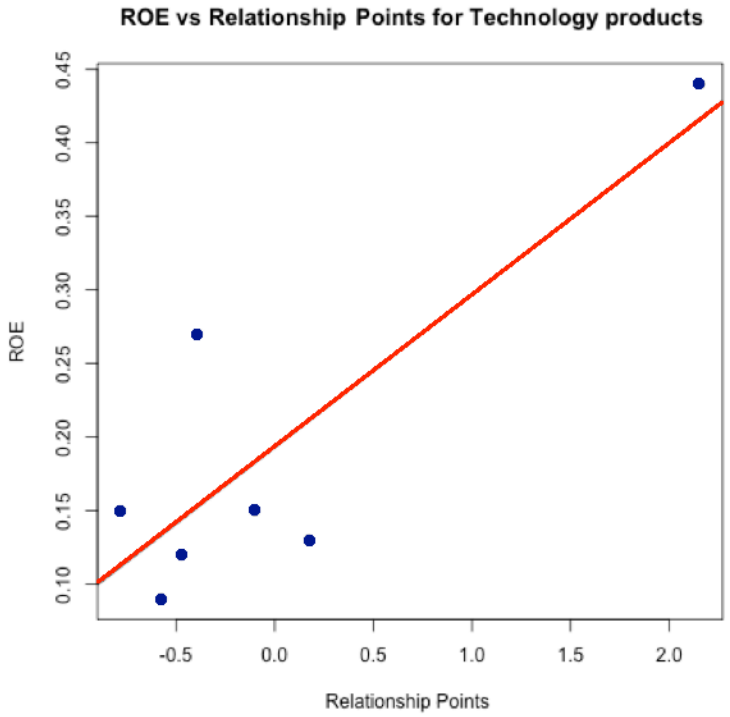

Figure 27: Relevant case 11

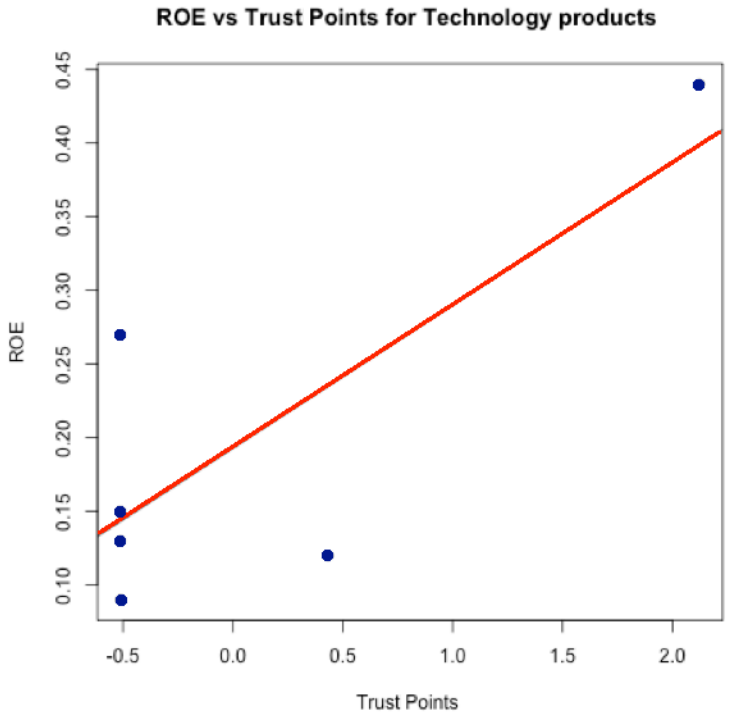

Figure 28: Relevant case 12

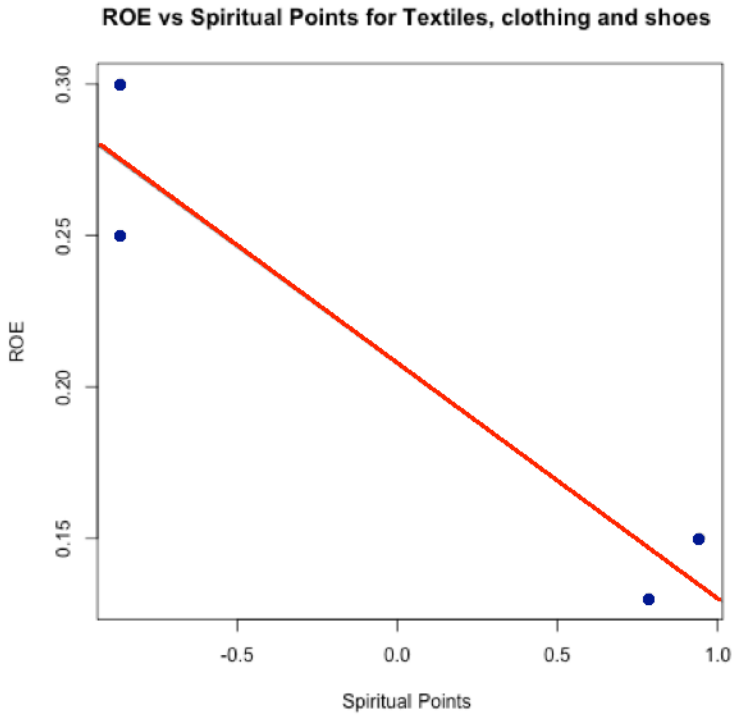

Figure 29: Relevant case 13

The previous figures showed certain cases like the 7, 11 and 12 that have a correlation due to an outlier. These cases are not considered in the conclusions section since the correlations are not significant.

12. Conclusions

The main goal of the project is to conduct an analysis to determine the relationship between the mission of the company and the financial performance achieved through the expressed means. Having said this, it is important to point out that the use of the conclusions of this project would put a company on a good track to achieve a good financial performance, but it does not ensure the positive performance of the company because their success depends on numerous factors including the successful implementation of the desired leadership style. After delivering a comprehensive analysis by explaining and justifying the different steps of the project, and careful interpretation of the data both quantitatively and visually, the project concludes with the following statements divided into the ideas explored in the introduction:

Indeed reviews

- The data from Indeed was gathered to be able to compare the leadership expressed by the companies in their mission statements with the actual leadership that is implemented in the company.
- The PCA showed a strong correlation between the variables obtained from Indeed. This shows that the grades provided by the workers of the company consider the overall quality of the company, assigning similar grades to the different fields.
- Even though the analysis of the opinions of Indeed and the financial performance of the company was promising, the results showed that there is not a single correlation between their factors and the ROE for any of the studied industries.

Gallup leadership

In order to refresh the ideas explained in the respective section and to make conclusions, the definition of the different fields is provided again as follows:

- Direction: the ability to create goals, communicate the future of the company, develop the strategy and encourage people to work towards that vision of the company in an effective way. Words: **future, coming, imminent, expect, vision, strategy, plan, long-term, decision, direction, action, opportunity, chance, next, following, goal.**
- Drive: the ability to challenge people to improve their work and set high standards in the company. Words: **require, overcome, demanding, improve, more, best, first, courage, advance, better, boost, increase, upgrade, extra, higher, further, new, leading, outstanding.**
- Influence: to ability to become a leader and a source of guidance for the workers in the company. This is obtained through charisma, persuasion, and emotion to build resilient trust. Words: **motivation, hard-work, follow, encourage, incentive, influence, important, critical, work, team, group.**
- Relationship: the ability to take care of the people, invest in them and want them to achieve their goals. Words: **workers, learn, development, personal, achieve, team, group.**
- Execution: the ability to complete the work effectively. This can be done by planning the timelines, assigning people to them and stablish strategic plans for the different scenarios and situations. Words: **achieve, success, obtained, done, doing, dealing, operation, process, response, make, happen, solution**.

After delivering a quantitative analysis and observing the plots, the following statements can be concluded:

- Relevant case #1 showed that there is a positive correlation between the ROE and the relationship points in the automotive industry. Thus, companies in the automotive industry seem to perform better if they state that they care about the workers of the company. Assuming that these companies implement this leadership style, the increase in financial performance is caused by the fact that this is an efficiency-based industry in which improving the production system places the company among the leaders of the market. This improvement can only be achieved through the formation and care of the employees.
- Relevant case #2 showed that there is a negative correlation between the ROE and the relationship points in the construction industry. The most relevant capabilities to succeed in the construction industry are the capital, the credibility, and experience in the field. Therefore, the conclusion is that the companies that prioritize other factors, like the employees in this case, are making a wrong strategy and seem to have a worse financial performance.
- Relevant case #6 showed that there is a positive correlation between the ROE and the execution points in the manufacturing industry. This shows that expressing the importance of effective planning and strategy seems to have a very positive impact on the manufacturing industry. Assuming that these companies implement this leadership style, due to the fact that the manufacturing industry has a high reliance on resource allocation both geographically and technologically, their financial performance would rely on good long-term planning.
- Relevant case #8 showed that there is a negative correlation between the ROE and the influence points in the personal products industry. This means that companies that are not a source of guidance for the workers seem to have a better performance. This effect is unexpected and would be a good subject for future study.
- Relevant case #9 showed that there is a positive correlation between the ROE and the execution points in the retail industry. Companies expressing their priority towards effective planning seem to have a better financial performance. Assuming that this

leadership is implemented in these companies, similar to the manufacturing case, the financial performance would correlate with the good long-term planning.

Ignatian leadership

In order to refresh the ideas explained in the respective section and to make conclusions, the definition of the different fields is provided again as follows:

- Spiritual: every leader must encourage people to behave morally and must show that the company is at the service of the people and not the opposite. Words: **responsibility, morals, value, care, commit, environment, social, duty, noble, principle, ethic, climate, habitat, collective, people, need, right, clean.**
- Self-knowledge: the leadership starts with the knowledge of yourself. It is the path to achieve the self-motivation. Words: **self-awareness, passion, drive, strength, weakness, talent.**
- Trust: the relationships in a company must be based on the trust. Developing the trust between employees improves the performance of the company. Words: **communication, collaboration, talent, synergies.**
- Discernment: the action of being aware of the situation of the company and make good decisions based on it. Words: **reflection, decision making, awareness.**
- Heroism: being confident about your own capacity and feeling capable of overcoming the obstacles. Words: **big goals, ambition, aspirations.**
- Love: being committed to the human potential and work so people develop their own potential. Words: **service, community.**
- Ingenuity: the capability of accepting the change and the excellence. Words: **creativity, wit, innovation, deal with uncertainty, deal with change, flexibility.**

After delivering a quantitative analysis and observing the plots, the following statements can be concluded:

- Relevant case #3 showed that there is a negative correlation between the ROE and the self-knowledge points in the electricity and gas industry. This means that a company seems to perform better if their mission statement does not reflect on their perceived strengths and weaknesses. This effect is unexpected and would be a good subject for future study.
- Relevant case #4 showed that there is a positive correlation between the ROE and the self-knowledge points in the engineering market. The companies seem to perform better if they express what their purpose is. Assuming that this leadership style is implemented, given that engineering companies are based on technical products or services, companies expressing their purpose might attract more investors since they would understand what they are investing their money in.
- Relevant case #5 showed that there is a positive correlation between the ROE and the trust points in the beverage and food industry. This means that expressing the willingness to develop a relationship of trust between employees within a beverage and food company would increase the financial performance. Assuming that this strategy is implemented, companies that retain talented employees through trust would have a better performance.
- Relevant cases #10 showed a positive correlation between the ROE and the spiritual points for the retail industry. This means that expressing the will to care about customers seems to have a positive impact on the financial performance. Assuming that this leadership is implemented successfully, retailing companies would perform better because they understand and fulfill their customers' needs.
- Relevant case #13 showed that there is a negative correlation between the ROE and the spiritual points in the textile industry. This means that companies in the textile industry seem to perform worse if they express that they care about their customers. This effect is unexpected and would be a good subject for future study.

13. Future work

This project was a good first approach on how the different kinds of leadership can affect the financial performance of the company. In the conclusions section, it was determined that there are negative correlations between good leadership styles and the financial performance in certain industries. A further research on the reasons behind that, using other quantitative or non-quantitative techniques, could round up this analysis and throw very interesting conclusions.

The most important future work is to repeat this analysis over time. The data analysis in this project considers the data from 2015 and makes conclusions about the financial performance given the leadership styles. However, delivering a good leadership can create a sustainable long-term financial performance. For this reason, this analysis should be delivered to evaluate the financial performance evolution over time.

References

[1] Putnam, R., Leonardi, R., & Nanetti, R. (1993). Making democracy work. Princeton University Press.

[2] Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. The Quarterly Journal of Economics, 112(4), 1251-1288.

[3] Ogbonna, E., & Harris, L. (2000). Leadership style, organizational culture and performance: Empirical evidence from UK companies. The International Journal of Human Resource Management, 11(4), 766-788. doi:10.1080/09585190050075114

[4] McGuire, J., Sundgren, A., & Schneeweis, T. (1988). Corporate social responsibility and firm financial performance. The Academy of Management Journal, 31(4), 854-872.

[5] Assessing the leadership talent . (2014). ().Gallup Consulting.

[6] Lowney, C. (2003). A guide to jesuit leadership and ignatian spirituality . ().

Annex 1: Results of anova analysis

industry	variable	pvalue
Engineering and others	DirectionPoints	0.82002283
Engineering and others	DrivePoints	0.921543744
Engineering and others	InfluencePoints	0.846858552
Engineering and others	RelationshipPoints	0.109053894
Engineering and others	ExecutionPoints	0.529846515
Engineering and others	SpiritualPoints	0.778875398
Engineering and others	SelfKnowledgePoints	3.43E-07
Engineering and others	TrustPoints	0.051872992
Engineering and others	LovePoints	0.355654444
Engineering and others	IngenuityPoints	0.93767857
Engineering and others	Indeed	0.272103378
Transport and distribution	DirectionPoints	0.655602796
Transport and distribution	DrivePoints	0.254656506
Transport and distribution	InfluencePoints	0.461960326
Transport and distribution	RelationshipPoints	0.597794798
Transport and distribution	ExecutionPoints	0.539260504
Transport and distribution	SpiritualPoints	0.411585639
Transport and distribution	SelfKnowledgePoints	0.3827866
Transport and distribution	TrustPoints	0.462517433
Transport and distribution	LovePoints	0.768386032
Transport and distribution	IngenuityPoints	0.6913933
Transport and distribution	Indeed	0.730969576
Electricity and gas	DirectionPoints	0.834103496
Electricity and gas	DrivePoints	0.943912577
Electricity and gas	InfluencePoints	0.328037
Electricity and gas	RelationshipPoints	0.209365307
Electricity and gas	ExecutionPoints	0.22544612
Electricity and gas	SpiritualPoints	0.21511236
Electricity and gas	SelfKnowledgePoints	0.04801629
Electricity and gas	TrustPoints	0.093884577
Electricity and gas	LovePoints	0.910953938
Electricity and gas	IngenuityPoints	0.685557483
Electricity and gas	Indeed	0.369601764
Leisure, tourism and hotel industry	DirectionPoints	0.801514102
Leisure, tourism and hotel industry	DrivePoints	0.663889804

Leisure, tourism and hotel industry	InfluencePoints	0.590327415
Leisure, tourism and hotel industry	RelationshipPoints	0.760339889
Leisure, tourism and hotel industry	ExecutionPoints	0.121472126
Leisure, tourism and hotel industry	SpiritualPoints	0.194680771
Leisure, tourism and hotel industry	SelfKnowledgePoints	NA
Leisure, tourism and hotel industry	TrustPoints	NA
Leisure, tourism and hotel industry	LovePoints	0.857994828
Leisure, tourism and hotel industry	IngenuityPoints	0.496718397
Leisure, tourism and hotel industry	Indeed	0.525201389
Materials	DirectionPoints	0.85781492
Materials	DrivePoints	0.581250299
Materials	InfluencePoints	0.337793397
Materials	RelationshipPoints	0.752033961
Materials	ExecutionPoints	0.65233947
Materials	SpiritualPoints	0.545184382
Materials	SelfKnowledgePoints	0.880600996
Materials	TrustPoints	0.002545835
Materials	LovePoints	0.560153021
Materials	IngenuityPoints	NA
Materials	Indeed	0.473148408
Construction	DirectionPoints	0.831892625
Construction	DrivePoints	0.461055564
Construction	InfluencePoints	0.879620879
Construction	RelationshipPoints	0.031662883
Construction	ExecutionPoints	0.739226424
Construction	SpiritualPoints	0.46990725
Construction	SelfKnowledgePoints	0.948802569
Construction	TrustPoints	0.935073126
Construction	LovePoints	0.287108794
Construction	IngenuityPoints	0.8319893
Construction	Indeed	0.410486123
Technology products	DirectionPoints	0.465818301
Technology products	DrivePoints	0.349259557
Technology products	InfluencePoints	0.202367153
Technology products	RelationshipPoints	0.016719009
Technology products	ExecutionPoints	0.756208915
Technology products	SpiritualPoints	0.917496444
Technology products	SelfKnowledgePoints	0.508830055
Technology products	TrustPoints	0.034588822

Technology products	LovePoints	0.6923628
Technology products	IngenuityPoints	0.360704734
Technology products	Indeed	0.152382126
Textiles, clothing and shoes	DirectionPoints	0.56824035
Textiles, clothing and shoes	DrivePoints	0.377798067
Textiles, clothing and shoes	InfluencePoints	0.425301764
Textiles, clothing and shoes	RelationshipPoints	0.597818497
Textiles, clothing and shoes	ExecutionPoints	0.347517002
Textiles, clothing and shoes	SpiritualPoints	0.046045274
Textiles, clothing and shoes	SelfKnowledgePoints	0.362010252
Textiles, clothing and shoes	TrustPoints	NA
Textiles, clothing and shoes	LovePoints	0.526652767
Textiles, clothing and shoes	IngenuityPoints	0.444871787
Textiles, clothing and shoes	Indeed	0.392314311
Pharmacy products and biotechnology	DirectionPoints	0.258267833
Pharmacy products and biotechnology	DrivePoints	0.941145578
Pharmacy products and biotechnology	InfluencePoints	0.908412772
Pharmacy products and biotechnology	RelationshipPoints	0.240814666
Pharmacy products and biotechnology	ExecutionPoints	0.393116765
Pharmacy products and biotechnology	SpiritualPoints	0.94487993
Pharmacy products and biotechnology	SelfKnowledgePoints	0.993394587
Pharmacy products and biotechnology	TrustPoints	0.938792669
Pharmacy products and biotechnology	LovePoints	0.980860525
Pharmacy products and biotechnology	IngenuityPoints	0.75522823
Pharmacy products and biotechnology	Indeed	0.892800359
Insurance	DirectionPoints	0.794830951
Insurance	DrivePoints	0.239005182
Insurance	InfluencePoints	0.600998004
Insurance	RelationshipPoints	0.76650392
Insurance	ExecutionPoints	0.590377814
Insurance	SpiritualPoints	0.248310279
Insurance	SelfKnowledgePoints	0.081716401
Insurance	TrustPoints	NA
Insurance	LovePoints	0.138792341
Insurance	IngenuityPoints	0.093656727
Insurance	Indeed	0.38327408
Electronics and software	DirectionPoints	0.677756254
Electronics and software	DrivePoints	0.822117303
Electronics and software	InfluencePoints	0.451792493

Electronics and software	RelationshipPoints	0.91464378
Electronics and software	ExecutionPoints	0.864472249
Electronics and software	SpiritualPoints	0.30217725
Electronics and software	SelfKnowledgePoints	0.943125112
Electronics and software	TrustPoints	0.88382798
Electronics and software	LovePoints	0.387367768
Electronics and software	IngenuityPoints	0.773632125
Electronics and software	Indeed	0.635728503
Retailing	DirectionPoints	0.915414861
Retailing	DrivePoints	0.90193575
Retailing	InfluencePoints	0.800632019
Retailing	RelationshipPoints	0.587132366
Retailing	ExecutionPoints	0.030581539
Retailing	SpiritualPoints	0.005118812
Retailing	SelfKnowledgePoints	0.184291891
Retailing	TrustPoints	0.340001152
Retailing	LovePoints	0.580235936
Retailing	IngenuityPoints	NA
Retailing	Indeed	0.980605693
Other financials	DirectionPoints	0.643695285
Other financials	DrivePoints	0.601927782
Other financials	InfluencePoints	0.094626703
Other financials	RelationshipPoints	0.09739661
Other financials	ExecutionPoints	0.27301201
Other financials	SpiritualPoints	0.450627728
Other financials	SelfKnowledgePoints	0.628116282
Other financials	TrustPoints	0.506258069
Other financials	LovePoints	0.766095583
Other financials	IngenuityPoints	0.741373274
Other financials	Indeed	0.290125801
Food and beverage	DirectionPoints	0.619249944
Food and beverage	DrivePoints	0.537534714
Food and beverage	InfluencePoints	0.842162117
Food and beverage	RelationshipPoints	0.56114048
Food and beverage	ExecutionPoints	0.086237324
Food and beverage	SpiritualPoints	0.90192546
Food and beverage	SelfKnowledgePoints	0.379883157
Food and beverage	TrustPoints	0.010786312
Food and beverage	LovePoints	0.763877868

Food and beverage	IngenuityPoints	0.175859624
Food and beverage	Indeed	0.984973718
Communication and publicity	DirectionPoints	0.788963599
Communication and publicity	DrivePoints	0.869427871
Communication and publicity	InfluencePoints	0.740611018
Communication and publicity	RelationshipPoints	0.737352331
Communication and publicity	ExecutionPoints	NA
Communication and publicity	SpiritualPoints	0.288043567
Communication and publicity	SelfKnowledgePoints	0.367171395
Communication and publicity	TrustPoints	0.863288792
Communication and publicity	LovePoints	0.925754671
Communication and publicity	IngenuityPoints	0.367171395
Communication and publicity	Indeed	0.278978719
Internet	DirectionPoints	0.970785533
Internet	DrivePoints	0.162258636
Internet	InfluencePoints	0.289114644
Internet	RelationshipPoints	0.289114644
Internet	ExecutionPoints	0.829291206
Internet	SpiritualPoints	0.475344777
Internet	SelfKnowledgePoints	0.289114644
Internet	TrustPoints	NA
Internet	LovePoints	0.289114644
Internet	IngenuityPoints	NA
Internet	Indeed	0.546602753
Banks	DirectionPoints	0.377827429
Banks	DrivePoints	0.122201354
Banks	InfluencePoints	0.502444854
Banks	RelationshipPoints	0.837126839
Banks	ExecutionPoints	0.336035238
Banks	SpiritualPoints	0.102414006
Banks	SelfKnowledgePoints	0.733429529
Banks	TrustPoints	0.912880998
Banks	LovePoints	0.582625296
Banks	IngenuityPoints	0.506110741
Banks	Indeed	0.110494922
Personal products	DirectionPoints	0.559496575
Personal products	DrivePoints	0.064789855
Personal products	InfluencePoints	0.035203032
Personal products	RelationshipPoints	0.933956099

Personal products	ExecutionPoints	0.229846669
Personal products	SpiritualPoints	0.823436525
Personal products	SelfKnowledgePoints	0.453042535
Personal products	TrustPoints	NA
Personal products	LovePoints	0.580486651
Personal products	IngenuityPoints	0.358183614
Personal products	Indeed	0.918813137
Auto	DirectionPoints	0.374382745
Auto	DrivePoints	0.787096377
Auto	InfluencePoints	0.295219383
Auto	RelationshipPoints	0.012992248
Auto	ExecutionPoints	0.834955773
Auto	SpiritualPoints	0.941450479
Auto	SelfKnowledgePoints	0.468159181
Auto	TrustPoints	NA
Auto	LovePoints	0.455356649
Auto	IngenuityPoints	0.468159181
Auto	Indeed	0.150612698
Manufacturing and assembly of capital goods	DirectionPoints	0.431134105
Manufacturing and assembly of capital goods	DrivePoints	0.418384127
Manufacturing and assembly of capital goods	InfluencePoints	0.284407815
Manufacturing and assembly of capital goods	RelationshipPoints	0.990559611
Manufacturing and assembly of capital goods	ExecutionPoints	0.029897927
Manufacturing and assembly of capital goods	SpiritualPoints	0.442974422
Manufacturing and assembly of capital goods	SelfKnowledgePoints	0.871809686
Manufacturing and assembly of capital goods	TrustPoints	NA
Manufacturing and assembly of capital goods	LovePoints	0.716855255
Manufacturing and assembly of capital goods	IngenuityPoints	0.230352769
Manufacturing and assembly of capital goods	Indeed	0.118709359
Telecommunications	DirectionPoints	0.824477888

Telecommunications	DrivePoints	0.126257755
Telecommunications	InfluencePoints	0.45862284
Telecommunications	RelationshipPoints	0.261482519
Telecommunications	ExecutionPoints	0.646975752
Telecommunications	SpiritualPoints	0.841814101
Telecommunications	SelfKnowledgePoints	0.542600717
Telecommunications	TrustPoints	0.722365152
Telecommunications	LovePoints	0.130345281
Telecommunications	IngenuityPoints	0.682267125
Telecommunications	Indeed	0.774308566
Healthcare	DirectionPoints	0.117287819
Healthcare	DrivePoints	0.451389596
Healthcare	InfluencePoints	0.651261253
Healthcare	RelationshipPoints	0.582157078
Healthcare	ExecutionPoints	0.720838832
Healthcare	SpiritualPoints	0.537400379
Healthcare	SelfKnowledgePoints	0.693886886
Healthcare	TrustPoints	0.693886886
Healthcare	LovePoints	0.477043767
Healthcare	IngenuityPoints	0.720533502
Healthcare	Indeed	0.090585035

Annex 2: linear regression results for relevant cases

case	intercept	pvalue_intercept	slope	pvalue_slope	rsquared
Engineering and others - SelfKnowledgePoints	0.34	3.07E-08	0.283450173	3.43E-07	0.89362766
Electricity and gas - SelfKnowledgePoints	0.090769231	4.53E-07	-0.020012109	0.04801629	0.310206456
Materials - TrustPoints	0.01	0.659418658	-0.155030694	0.002545835	0.918750614
Construction - RelationshipPoints	0.055	0.155044487	-0.111570288	0.031662883	0.724137827
Technology products - RelationshipPoints	0.192857143	0.000865759	0.103784834	0.016719009	0.713780776
Technology products - TrustPoints	0.192857143	0.001613369	0.097030565	0.034588822	0.623898838
Textiles, clothing and shoes - SpiritualPoints	0.2075	0.005099833	-0.077254628	0.046045274	0.910029618
Retailing - ExecutionPoints	0.414166667	0.000227286	0.194331011	0.030581539	0.38769313
Retailing - SpiritualPoints	0.414166667	6.02E-05	0.233549495	0.005118812	0.559966121
Food and beverage - TrustPoints	0.197142857	0.00030631	0.095193986	0.010786312	0.757864568
Personal products - InfluencePoints	0.186666667	7.81E-05	-0.038675009	0.035203032	0.710010921
Auto - RelationshipPoints	0.0075	0.744192957	0.201048577	0.012992248	0.974184303
Manufacturing and assembly of capital goods - ExecutionPoints	0.061666667	0.419317484	0.247969013	0.029897927	0.731514038

Annex 3: Python code for sentimental analysis

```
import re
import xlrd
import xlwt

file_location = "/Users/jaimedominguezdp/Desktop/Proyecto/Performance.xlsx"
workbook=xlrd.open_workbook(file_location)
sheet=workbook.sheet_by_index(0)
book=xlwt.Workbook(encoding="utf-8")
sheet1=book.add_sheet("Sheet 1")

for i in range(1,sheet.nrows):
    phrase=str(sheet.cell_value(i,31))
    company=str(sheet.cell_value(i,0))
    direction_tag =
re.compile(r'\bfuture\b|\bcoming\b|\bimminent\b|\bexpect|\bvision|\bstrategy?|\bplan|\bl
ong-
term\b|\bdecision|\bdirection|\baction|\bopportunity?|\bchance|\bwill\b|\bnext\b|\bfollo
wing\b|\bgoal', re.MULTILINE | re.IGNORECASE)
    drive_tag =
re.compile(r'\brequire?|\bovercome?|\bdemanding\b|\bimprove|\bmore\b|\bbest\b|\bfirst\
b|\bcourage|\badvance\b|\bbetter\b|\bboost|\bincrease\b|\bupgrade|\bextra\b|\bhigher\
b|\bfurther\b|\bnew\b|\bleading\b|\boutstanding\b', re.MULTILINE | re.IGNORECASE)
    influence_tag = re.compile(r'\bmotivation|\bhard-
work|\bfollow|\bencourage|\bincentive|\binfluence?|\bimportant\b|\bcritical\b|\bwork\b|\
bteam|\bgroup', re.MULTILINE | re.IGNORECASE)
    relationship_tag =
re.compile(r'\bworker|\blearn\b|\bdevelopment\b|\bpersonal\b|\bachieve?|\bwe\b|\bteam
|\bgroup', re.MULTILINE | re.IGNORECASE)
    execution_tag =
re.compile(r'\bhave\b|\bachieve?|\bsuccess\b|\bobtained\b|\bdoing\b|\bdealing\b|\bopera
tion|\bprocess|\bresponse|\bmake\b|\bhappen\b|\bsolution\b', re.MULTILINE |
re.IGNORECASE)
    spiritual_tag =
re.compile(r'\bresponsibility?|\bmoral|\bvalue|\bcare\b|\bcommit|\benvironment\b|\bsocia
l\b|\bduty?|\bnoble\b|\bprinciple\b|\bethic|\bclimate\b|\bhabitat\b|\bcollective|\bpeople\
b|\bneed\b|\bright\b|\bclean', re.MULTILINE | re.IGNORECASE)
    selfknow_tag = re.compile(r'\bself-
awareness\b|\bpassion|\bdrive|\bstrength|\bweakness|\btalent', re.MULTILINE |
re.IGNORECASE)
```

```
   trust_tag =
re.compile(r'\bcommunication|\bcollaboration|\bcollaborative\b|\btalent|\bsynergy?',
re.MULTILINE | re.IGNORECASE)
   discernment_tag = re.compile(r'\breflection|\bdecision making\b|\bawareness',
re.MULTILINE | re.IGNORECASE)
   heroism_tag = re.compile(r'\bbig goal|\bambition|\baspiration', re.MULTILINE |
re.IGNORECASE)
   love_tag = re.compile(r'\bservice|\bcommunity?', re.MULTILINE | re.IGNORECASE)
   ingenuity_tag = re.compile(r'\bcreativity\b|\bwit\b|\binnovation|\bdeal with change|\bdeal
with uncertainty?|\bflexibility', re.MULTILINE | re.IGNORECASE)

   match_pattern = re.findall(r'\b[a-z]{1,15}\b', phrase)

   direction_temp=re.findall(direction_tag, phrase)
   direction=len(direction_temp)
   drive_temp=re.findall(drive_tag, phrase)
   drive=len(drive_temp)
   influence_temp=re.findall(influence_tag, phrase)
   influence=len(influence_temp)
   relationship_temp=re.findall(relationship_tag, phrase)
   relationship=len(relationship_temp)
   execution_temp=re.findall(execution_tag, phrase)
   execution=len(execution_temp)
   spiritual_temp=re.findall(spiritual_tag, phrase)
   spiritual=len(spiritual_temp)
   selfknow_temp=re.findall(selfknow_tag, phrase)
   selfknow=len(selfknow_temp)
   trust_temp=re.findall(trust_tag, phrase)
   trust=len(trust_temp)
   discernment_temp=re.findall(discernment_tag, phrase)
   discernment=len(discernment_temp)
   heroism_temp=re.findall(heroism_tag, phrase)
   heroism=len(heroism_temp)
   love_temp=re.findall(love_tag, phrase)
   love=len(love_temp)
   ingenuity_temp=re.findall(ingenuity_tag, phrase)
   ingenuity=len(ingenuity_temp)
   count=0

   for word in match_pattern:
      count = count+1

   sheet1.write(i,0,company)
   sheet1.write(i,1,direction)
```

```
    sheet1.write(i,2,drive)
    sheet1.write(i,3,influence)
    sheet1.write(i,4,relationship)
    sheet1.write(i,5,execution)
    sheet1.write(i,6,spiritual)
    sheet1.write(i,7,selfknow)
    sheet1.write(i,8,trust)
    sheet1.write(i,9,discernment)
    sheet1.write(i,10,heroism)
    sheet1.write(i,11,love)
    sheet1.write(i,12,ingenuity)
    sheet1.write(i,13,count)


sheet1.write(0,0,"Company name")
sheet1.write(0,1,"Direction points")
sheet1.write(0,2,"Drive points")
sheet1.write(0,3,"Influence points")
sheet1.write(0,4,"Relationship points")
sheet1.write(0,5,"Execution points")
sheet1.write(0,6,"Spiritual points")
sheet1.write(0,7,"Self-knowledge points")
sheet1.write(0,8,"Trust points")
sheet1.write(0,9,"Discernment points")
sheet1.write(0,10,"Heroism points")
sheet1.write(0,11,"Love points")
sheet1.write(0,12,"Ingenuity points")
sheet1.write(0,13,"Total words")

book.save("Results.xls")
```

Annex 4: R code for data analysis

```
library(psych)

# Import the data

setwd("/Users/jaimedominguezdp/Desktop/Proyecto")
df<-read.csv("Performance_good.csv")

# Selecting relevant data

df<-
df[1:162,c("Company_name","Market","Index","Sector","Industry","ROE_2015","Dividend_201
5",
           "Profit_2015","Price_2015","EBITDA_2015","PER_2015","Overall","WorkLife_balance",

"Compensation","Career_opportunities","Senior_management","Culture","DirectionPoints",

"DrivePoints","InfluencePoints","RelationshipPoints","ExecutionPoints","SpiritualPoints",

"SelfKnowledgePoints","TrustPoints","DiscernmentPoints","HeroismPoints","LovePoints",
           "IngenuityPoints")]

# Exploring the data

summary<-summary(df)
# write.csv(summary, "/Users/jaimedominguezdp/Desktop/Proyecto/basic_stats.csv")

# Fixing standardization problems

levels(df$Market)
levels(df$Index)
levels(df$Sector)
levels(df$Industry)

# Fixing data types

df$Market<-factor(df$Market)
df$Index<-factor(df$Index)
df$Sector<-factor(df$Sector)
df$Industry<-factor(df$Industry)

# Exploring suspicioues values
```

```
summary(df)

## There is a sample with a very high EBITDA

hist(df$EBITDA_2015, main='Histogram of EBITDA',xlab = 'EBITDA')
df[df$EBITDA_2015==max(df$EBITDA_2015),"Company_name"] # It happens to be Apple so it
makes sense

## There is also a sample with very high PER

hist(df$PER_2015)
df[df$PER_2015==max(df$PER_2015),"Company_name"] # It happens to be Amazon so it
makes sense

# The values seem to be correct
# Now I am going to get rid of the variables without relevant dispersion

for (i in colnames(df)[seq(12,29,1)]){
  hist(df[[i]], main=i, bins=20)
}

# There are not many companies with discernment points, and the ratios are very low
# I am going to check the case to see if it is worth to include it in the analysis

head(df[order(df$DiscernmentPoints,decreasing=T),c("Company_name","DiscernmentPoints")]
)

# There are only 4 companies with a ratio different from 0, and they are not industry disruptors
# I decide to get rid of the variable

df["DiscernmentPoints"]<-NULL

# There are not many companies with heroism points, and the ratios are very low
# I am going to check the case to see if it is worth to include it in the analysis

head(df[order(df$HeroismPoints,decreasing=T),c("Company_name","HeroismPoints")])

# There are companies with relevant values
# I decide to not keep the variable

df["HeroismPoints"]<-NULL

# I want to start analyzing if there are correlations between the predictos
```

```
cor(df[,c(12:27)])

# The correlation between the predictors can't be considered very high, however, the number of variables
# is a problem. It would be good to apply dimensionality reduction techniques

# Standardize
# At this point the level of standardization must be considered

stand <- data.frame(scope= character(), scope_value= character(), variable= character(), avg= numeric(),
                    std = numeric())

scope<-c("Sector","Industry")

for(i in scope){
  for(j in levels(factor(df[[i]]))){
    for(k in colnames(df)[12:27]){
      stand<-rbind(stand,data.frame(scope=i,scope_value=j,variable=k,
                                    avg=mean(df[df[i]==j,k]),
                                    std=sd(df[df[i]==j,k])))
    }
  }
}

# Since I do not know yet the level of standardization that I have to apply, I will make the code to
# change it as I want

level_of_factorization<-"Industry"

for(i in 1:nrow(df)){
  for(j in colnames(df)[12:27]){
    if(stand[stand$scope==level_of_factorization &
             stand$scope_value==as.character(df[[i,level_of_factorization]]) &
             stand$variable==j,"std"]!=0){
      df[i,j]<-(df[i,j]-stand[stand$scope==level_of_factorization &
                              stand$scope_value==as.character(df[[i,level_of_factorization]]) &
                              stand$variable==j,"avg"])/
               stand[stand$scope==level_of_factorization &
                     stand$scope_value==as.character(df[[i,level_of_factorization]]) &
                     stand$variable==j,"std"]
    }else{
```

```
    df[i,j]<-0
   }
  }
}

# Now tha t there has been a very unique process of standardization, it is required to check the
# correlation again

cor(df[,c(12:27)])

# The good news are that the variables are now highly uncorrelated
# The bad thing, is that there are to many of them now
# I am going to apply PCA to reduce the dimension of the variables

### Apply PCA

pca_indeed = df[,c(12:17)]
pca_gallup = df[,c(18:22)]
pca_Ignatian = df[,c(23:27)]

df_pca_indeed <- prcomp(~.,data=pca_indeed, center = FALSE, scale. = FALSE)
df_pca_gallup <- prcomp(~.,data=pca_gallup, center = FALSE, scale. = FALSE)
df_pca_Ignatian <- prcomp(~.,data=pca_Ignatian, center = FALSE, scale. = FALSE)

plot(df_pca_indeed, type = "l", main='Explained variance for Indeed variables')
plot(df_pca_gallup, type = "l", main='Explained variance for Gallup variables')
plot(df_pca_Ignatian, type = "l", main='Explained variance for Ignatian variables')

pca_first<-data.frame(predict(df_pca_indeed, newdata=pca_indeed))
df<-cbind(df,pca_first[,1])
colnames(df)[28] <- "Indeed"

# Select the final dataframe

df<-
df[,c("Company_name","Industry","ROE_2015","DirectionPoints","DrivePoints","InfluencePoint
s",

"RelationshipPoints","ExecutionPoints","SpiritualPoints","SelfKnowledgePoints","TrustPoints",
        "LovePoints","IngenuityPoints","Indeed")]

# ROE of the different industries

boxplot(ROE_2015~Industry,data=df)
```

```
# Plot ROE against variables

for(i in unique(df$Industry)){
  for(j in colnames(df)[4:14]){

    png(filename=paste("/Users/jaimedominguezdp/Desktop/Proyecto/Plots/",i,j,".png",sep=""))
    plot(df[df$Industry==i,"ROE_2015"]~df[df$Industry==i,j],xlab=j,ylab='ROE',
         main=paste('ROE vs',j,'for',i))
    dev.off()

  }}

# Create anova matrix

anova_results<-data.frame(industry=character(),variable=character(),pvalue=double())

for(i in unique(df$Industry)){
  for(j in colnames(df)[4:14]){

    df_temp<-data.frame(industry=i,variable=j,
                        pvalue=summary(aov(df[df$Industry==i,"ROE_2015"]~
                                             df[df$Industry==i,j]))[[1]][["Pr(>F)"]][1])
    anova_results<-rbind(anova_results,df_temp)
  }}

write.csv(anova_results,'anova.csv')

# Generate table of pearson

pearson_results <- data.frame(case=character(),linear=double(),exp=double())
slope_results <- data.frame(case=character(),slope=double(),sign=character())
regression_results <- data.frame(case=character(),intercept=double(),
                                 pvalue_intercept=double(),slope=double(),pvalue_slope=double(),
                                 rsquared=double())

cases <- matrix(c('Engineering and others','Electricity and gas','Materials','Construction',
                  'Technology products','Technology products','Textiles, clothing and shoes',
                  'Retailing','Retailing','Food and beverage','Personal products','Auto',
                  'Manufacturing and assembly of capital goods','SelfKnowledgePoints',
                  'SelfKnowledgePoints','TrustPoints','RelationshipPoints','RelationshipPoints',
                  'TrustPoints','SpiritualPoints','ExecutionPoints','SpiritualPoints',
                  'TrustPoints','InfluencePoints','RelationshipPoints','ExecutionPoints'),
                  nrow=13,ncol=2
)
```

```
for(i in 1:13){

 linear_reg<-lm(df[df$Industry==cases[i,1],'ROE_2015']~df[df$Industry==cases[i,1],cases[i,2]])
 exp_reg<-nls(paste('ROE_2015~I(a+b^',cases[i,2],')',sep=''),data=df[df$Industry==cases[i,1],],
          start=list(a=1,b=1))

 df_temp<-data.frame(case=paste(cases[i,1],'-',cases[i,2]),
                linear=summary(linear_reg)$r.squared,
                exp=1-(sum(residuals(exp_reg)^2) /sum((df[df$Industry==cases[i,1],'ROE_2015']-
                              mean(df[df$Industry==cases[i,1],'ROE_2015']))^2)) )

 pearson_results<-rbind(pearson_results,df_temp)

 slope_temp<-linear_reg$coefficients[[2]]

 if(slope_temp>0){
  sign_temp='Positive'
 }else{
  sign_temp='Negative'
 }

 df_temp2 <- data.frame(case=paste(cases[i,1],'-',cases[i,2]),
                  slope=slope_temp,
                  sign=sign_temp)

 slope_results<-rbind(slope_results,df_temp2)

 df_temp3 <- data.frame(case=paste(cases[i,1],'-',cases[i,2]),
                  intercept=linear_reg$coefficients[[1]],
                  pvalue_intercept=summary(linear_reg)$coefficients[7],
                  slope=slope_temp,
                  pvalue_slope=summary(linear_reg)$coefficients[8],
                  rsquared=summary(linear_reg)$r.squared)
 regression_results<-rbind(regression_results,df_temp3)

 png(filename=paste("/Users/jaimedominguezdp/Desktop/Proyecto/Regressions/",
             paste(cases[i,1],'-',cases[i,2]),".png",sep=""))
 plot(df[df$Industry==cases[i,1],'ROE_2015']~df[df$Industry==cases[i,1],
    cases[i,2]],xlab=cases[i,2],ylab='ROE',
    main=paste('ROE vs',cases[i,2],'for',cases[i,1]))
 abline(linear_reg)
 dev.off()
}
```