COMILLAS

UNIVERSIDAD PONTIFICIA

ICAI

MASTER'S DEGREE IN

INDUSTRIAL ENGINEERING

FINAL MASTER THESIS

Do ESG investments generate alpha?

Author: Javier Olavarría Múgica

Supervisor: Gabriele Casati

Co-Supervisor: Angel Sanz

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Do ESG Investments generate alpha?

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2020/2021 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: Javier Olavarría Múgica Fecha: 18/ 01/ 2021

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Gabriele Casati Fecha: 18/ 01/ 2021

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. Javier Olavarría Múgica

DECLARA ser el titular de los derechos de propiedad intelectual de la obra: Do ESG Investments generate alpha?, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.

b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.

c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.

d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.

e) Asignar por defecto a estos trabajos una licencia Creative Commons.

f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente)*.

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

a) Que la Universidad identifique claramente su nombre como autor de la misma

b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.

c) Solicitar la retirada de la obra del repositorio por causa justificada.

d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.

b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.

c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.
La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a …18….. de ……01…………………... de …2021.

ACEPTA

Fdo……………………………………………………

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

MASTER’S DEGREE IN
INDUSTRIAL ENGINEERING

FINAL MASTER THESIS

Do ESG investments generate alpha?

Author: Javier Olavarría Múgica

Supervisor: Gabriele Casati

Co-Supervisor: Angel Sanz

Madrid

DO ESG INVESTMENTS GENERATE ALPHA?

Author: Olavarría Múgica, Javier.
Supervisor: Casati, Gabriele.
Collaborator Entity: ICAI – Universidad Pontificia Comillas

Abstract

The aim of this project is to determine the relationship between ESG Ratings and the alpha of a company. To do this, several clustering and regression techniques were used on a dataset containing information on more than 8,000 companies. The final model shows that a correlation does exist.

Keywords: clustering, regression, acquisition, dataset, ESG, alpha.

1. Introduction

The concept of selective investment is not new. For decades, investors have been looking at different performance indicators to design their investment strategies and help them allocate their money. Financial indicators such as Return on Equity (ROE), Dividend Yield, P/E Ratio and P/CF Ratio, in conjunction with non-financial measures like the Customer Satisfaction Index, have been used to determine the performance of a company. However, investors have become increasingly more concerned about how their investments are impacting the world. This caused the birth of metrics that assess companies on an ethical level. One of these indicators is the ESG Rating.

ESG Ratings are valuations given to a company in regard to three areas: Environmental, Social and Governance. To obtain these ratings, a number of factors are assessed. Among many others, the Environmental Score takes into account the carbon emissions of a company, their energy efficiency and their climate change vulnerability. As for the Social Score, key metrics such as labor management, health or safety are considered. Finally, for the Governance Score, issues such as corruption, instability, business ethics or fraud are evaluated [1].

The Overall ESG Rating has become one of the metrics used for screening and strategic investment decisions. Wealthy individuals and large corporations are looking to create an impact with their money, resulting in funds that only invest in companies when their ESG score is higher than a certain threshold. As investors reward companies that have high ESG ratings, it is believed that these companies tend to outperform their peers.

Numerous studies have been conducted, most of them showing a positive correlation between ESG ratings and performance [2], some showing inconclusive results and a few others showing a negative relationship [3]. Therefore, the aim of this project is to shed some light on this issue and analyze the relationship between these factors and a company's alpha.

A company's alpha a proxy that will be used as the performance of a company. Alpha is the company's return above a particular index or benchmark. This metric is very relevant, since it eliminates the proportion of the return that might be due to the overall market or industry overperforming, leaving the added value of the company [4].

2. Project definition

As previously stated, the aim of this thesis is to explore the correlation between ESG Ratings and alpha. This relationship is studied using different regression techniques, all of which are performed with the aid of programming tools. Some of these methods include simple Linear Regression, Partial Least Squares, Random Forests and Gradient Boosting. Non-linear regression such as Polynomial Regression are also applied to have a variety of options to select the model that fits the data the best.

In addition to this, clustering is performed to find patterns and groups within the dataset. Companies are divided using clustering techniques and pre-established criteria (i.e., firms will also be classified based on country and industry, since the same benchmark cannot be used for all of them).

The model that best fits the data is used to determine the correlation between alpha and ESG Ratings. The most important factors, and whether these factors have a positive or negative relationship to alpha, are determined.

3. Model description

To examine this correlation, several steps were taken. These stages are: data acquisition, data storage, data cleaning and preprocessing and regression and clustering techniques.

Firstly, the data is acquired from three sources: MSCI ESG Direct, Bloomberg and Yahoo Finance. It is then stored in either CSV/Excel files or in SQLite. SQLite was used because of the large amount of data being handled, as well as the difficulty and time-consuming process to obtain it. SQLite allows ease of retrieval and administration.

Secondly, the treatment of missing values, outliers and qualitative variables was performed, constituting the data cleaning and preprocessing step.

Finally, the regression and classification techniques were applied. The base model to which the rest are compared is a simple linear regression with the main features selected based on intuition and common knowledge. Afterwards, the methods Partial Least Squares, Gradient Boosting, Random Forests and Polynomial Regression were tried, as well as an improved Linear Regression.

4. Results

Every model for each regression technique was evaluated using its Adjusted R-Squared, that is, the percentage of variance of the response explained by the model. In addition to this, whether or not a model has low p-values was also taken into account since p-values show the significant features. In the table below a summary of the models and the quality assessment can be viewed.

	Initial Linear Regression	PLS Regression	Random Forests	Gradient Boosting	Polynomial Regression	Final Linear Regression
Adjusted R-Squared	3.2%	9.1%	12.1%	4%	Negative	Avg 14%
P – value Coefficients	Yes	No	No	No	Yes	Yes
Overall quality	5th	3rd	2nd	4th	6th	1st

Table 1. Models Quality Assessment.

As such, the final model is a Linear Regression that uses the most important features obtained from Random Forests. As well, in this regression the companies were filtered based on geographical area or industry. This model shows an average of 14% of variance explained and most p-values below 5%. Therefore, this model was the most successful, improving over the rest.

5. Conclusions

Using the Linear Regression model, we can determine that a correlation does exist. However, this correlation is weak, explaining, at most, a 22% of the variance in some cases. It is extremely low when taking all companies in the dataset, and higher when filtering by region or industry.

Even though the correlation exists, a relationship with the overall ESG Rating was not found, but rather with some of its underlying categories. The most notable factors that are related to alpha are the Environmental Score and Social Score, which correlate positively overall. As well, low alphas tend to correspond to companies where the CEO's equity is not related to the company's performance. The Human Capital Development Score also has, generally, a negative relationship.

These relationships depend on the geographical region and industry. The regions with the highest correlations are Canada and South America, followed by the United States, Oceania, Japan, United Kingdom and China. At the bottom end we find Europe and Africa.

Regarding the classification in industries, the Energy, Telecommunication Services and Utilities sectors are the highest correlated. The lowest are Industrials, Financials, Consumer Discretionary and Consumer Staples. Falling in the middle, the Health Care and the Materials sectors are present.

The volatility of these companies was also contrasted against alpha, showing a high correlation. The regression model shows that companies that violate its debt covenants have had higher volatility in the last 360 days. The same happens when a CEOs pay is not related to its performance. Finally, companies with lower carbon emissions have had lower volatility. We can determine that, although none of the three pillar scores are correlated, the volatility is related strongly to some of their underlying factors.

6. Bibliography

[1] "Sustainable Investing: Investing for the Long-Term | Harvard Management Company." Harvard Management Company, 20 Nov. 2020, www.hmc.harvard.edu/sustainable-investing/.

[2] Gompers, P. A., Ishii, J. L., & Metrick, A. (2001). Corporate Governance and Equity Prices. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.278920

[3] Lopez de Silanes, F., McCahery, J. A., & Pudschedl, P. (2019). ESG Performance and Disclosure: A Cross-Country Analysis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3506084.

[4] Alpha Vs. Beta: What's the Difference? (2021). Investopedia. https://www.investopedia.com/ask/answers/102714/whats-difference-between-alpha-and-beta.asp

LAS INVERSIONES ESG, ¿GENERAN ALFA?

Autor: Olavarría Múgica, Javier.
Director: Casati, Gabriele.
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

Resumen del proyecto

El objetivo de este proyecto es determinar la relación entre los ratings ESG y el alfa de una compañía. Para ello, distintas técnicas de agrupación y regresión se han usado sobre un conjunto de datos, que contienen la información de más de 8.000 compañías. El modelo que explica mejor la respuesta muestra la existencia de una correlación, aunque ésta no es muy fuerte.

Palabras clave: clustering, regresión, adquisición, dataset, ESG, alfa.

1. Introducción

El concepto de inversión selectiva no es nuevo. Durante décadas, analistas financieros han examinado distintos indicadores del rendimiento de las compañías para diseñar sus estrategias de inversión y, de esta forma, decidir cómo distribuir su dinero o el de sus clientes. Indicadores financieros como el Return On Equity (ROE), el Dividend Yield, o los ratios P/E y P/CF, junto con métricas no financieras como el índice de satisfacción del cliente, se han usado para determinar el rendimiento o performance de una empresa. Sin embargo, estos inversores están cada vez más preocupados por el impacto que tienen sus inversiones en el mundo. Esto ha provocado el nacimiento de indicadores que evalúan las empresas a nivel ético. Uno de estos indicadores es el Rating ESG.

Las calificaciones ESG son valoraciones que se le dan a una empresa con respecto a tres áreas: ambiental, social y de gobernanza. Para obtener estas calificaciones, se evalúan más de doscientos factores. Por ejemplo, la Valoración Medioambiental tiene en cuenta las emisiones de carbono de una empresa, su eficiencia energética y su vulnerabilidad al cambio climático. En la Valoración Social se consideran métricas clave como la gestión laboral, la salud o la seguridad. Por último, para la Valoración de Gobernanza, se evalúan cuestiones como la corrupción, la inestabilidad, la ética empresarial o el fraude [1].

La calificación de ESG se ha convertido en una de las métricas utilizadas para la selección y las decisiones de inversión estratégica. Los inversores particulares y las grandes corporaciones buscan tener un impacto positivo con su dinero, lo que ha resultado en fondos que solo invierten en empresas cuyo nivel ESG es superior a un cierto umbral. Dado que los inversores recompensan a las empresas que tienen altas calificaciones ESG, gran parte de la comunidad financiera cree que estas empresas tienden a superar a sus pares.

Se han realizado numerosos estudios sobre esto, la mayoría de ellos mostrando una correlación positiva entre las calificaciones ESG y el rendimiento de una compañía [2]. Sin embargo, algunos muestran resultados no concluyentes y otros muestran una relación negativa [3]. Por lo tanto, el objetivo de este proyecto es arrojar algo de luz sobre este tema y analizar la relación entre estos factores y el alfa de una empresa.

El alfa de una compañía es el proxy que se utilizará como rendimiento o performance de ésta. Es el retorno de una empresa por encima de un índice o referencia. Debido a esto, la medida alfa elimina la proporción del retorno que se debe al mercado o la industria, dejando el valor añadido de la empresa [4].

2. Definición del proyecto

El objetivo de este proyecto es entender la correlación entre las calificaciones ESG y alfa. Esta relación se ha estudiado mediante diferentes técnicas de regresión, las cuales se han realizado con la ayuda de herramientas de programación. Algunos de estos métodos incluyen: Regresión Lineal, Partial Least Squares, Random Forests y Gradient Boosting. Técnicas de regresión no lineal, como la Regresión Polinómica, también se ha aplicado para obtener una variedad de opciones para seleccionar el modelo que mejor se ajusta a los datos.

Además de esto, se han realizado técnicas de clustering (agrupación) para encontrar patrones y grupos dentro del conjunto de datos. Las empresas se dividen utilizando técnicas de clustering y criterios preestablecidos: las empresas también se clasifican según su región geográfica o industria a la que pertenecen.

El modelo que mejor se ajusta a los datos se ha utilizado para determinar la correlación entre las calificaciones alfa y ESG. Se han determinado, a su vez, los factores más importantes, y su relación positiva o negativa con alfa.

3. Análisis de datos

Este proyecto se centra en la utilización de técnicas de regresión y agrupación para determinar la correlación entre alfa y los Rating ESG. Los pasos para lograr este objetivo son: adquisición, almacenamiento, limpieza y preprocesamiento de los datos; para después aplicar las técnicas de regresión y clustering.

En primer lugar, los datos se obtienen de tres fuentes: MSCI ESG Direct, Bloomberg y Yahoo Finance. Se almacenan en archivos de tipo csv/Excel o en SQLite. Se ha usado SQLite en determinadas situaciones a lo largo del proyecto ya que permite una fácil recuperación y administración de los datos. Esto resulta muy importante debido a la gran cantidad de observaciones que se manejan en la dataset, así como su dificultad de obtención.

En segundo lugar, se realizó el tratamiento de datos: se eliminaron observaciones repetidas, los outliers y observaciones con variables perdidas. También se gestionaron las variables cualitativas, finalizando así la depuración y preprocesamiento de datos.

Finalmente, se aplicaron las técnicas de regresión y clasificación. El modelo base con el que se compara el resto es una regresión lineal simple con las variables predictivas seleccionadas en base al conocimiento financiero previo y literatura que trata este tema. Posteriormente, se probaron los métodos Partial Least Squares, Gradient Boosting, Random Forest y Polinomial Regression, así como una Regresión Lineal mejorada.

4. Resultados

Cada modelo de regresión ha sido evaluado usando el R Cuadrado Ajustado, parámetro que indica la proporción de varianza explicada por el modelo. Además, también se ha tenido en cuenta los p-valor de los predictores, ya que valores bajos nos señalan las variables significativas del modelo. A continuación, puede verse una tabla resumen que contiene los parámetros R Cuadrado Ajustado de los modelos.

	Initial Linear Regression	PLS Regression	Random Forests	Gradient Boosting	Polynomial Regression	Final Linear Regression
Adjusted R-Squared	3.2%	9.1%	12.1%	4%	Negative	Avg 14%
P – value Coefficients	Yes	No	No	No	Yes	Yes
Overall quality	5th	3rd	2nd	4th	6th	1st

Tabla 1. Comparativa calidad de los modelos.

El modelo final es una Regresión Linear mejorada, que utiliza las variables más importantes obtenidas usando el método de Random Forests. No sólo esto, sino que las compañías han sido divididas por región geográfica e industria, aportando mejores resultados que en los modelos anteriores. Este modelo muestra una media de proporción de varianza explicada del 14%.

5. Conclusions

Usando este modelo de Regresión Lineal, se puede determinar que una correlación entre alfa y los Rating ESG existe. Sin embargo, esta correlación es débil, explicando, en el mayor de los casos, un 22% de la varianza en la respuesta. Esta correlación es extremadamente baja cuando se aplica el modelo a todo el universo de compañías, sin la división por región e industria.

Aunque esta correlación exista, no se ha encontrado una relación con el índice ESG total, sino con los factores subyacentes. Entre los factores más destacables se encuentran la Valoración Medioambiental y la Valoración Social, que se relacionan positivamente con alfa en la mayoría de los casos. También es importante recalcar la correlación con el capital humano, que es sorprendentemente negativa.

Esta relación depende de la región geográfica y la industria. Las regiones con correlaciones más altas son Canadá y América del Sur, seguidas de Estados Unidos, Oceanía, Japón, Reino Unido y China. Como últimas regiones se encuentran Europa y África.

En cuanto a la clasificación por industria, los sectores de Energía, Servicios de la Telecomunicación y Utilities son los más correlacionados. Entre los más bajos se encuentra el sector financiero, el industrial, además de los sectores de Consumer Discretionary y Consumer Staples. En el medio se encuentran los sectores Salud y Materiales.

La volatilidad de las compañías también fue contrastada con alfa, mostrando una alta correlación. El modelo de regresión muestra que las compañías que violan las condiciones contractuales de deuda presentan una mayor volatilidad en el periodo estudiado. Lo mismo ocurre cuando el pago al CEO de una compañía no es proporcional al rendimiento de la compañía.

6. Bibliografía

[1] "Sustainable Investing: Investing for the Long-Term | Harvard Management Company." Harvard Management Company, 20 Nov. 2020, www.hmc.harvard.edu/sustainable-investing/.

[2] Gompers, P. A., Ishii, J. L., & Metrick, A. (2001). Corporate Governance and Equity Prices. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.278920

[3] Lopez de Silanes, F., McCahery, J. A., & Pudschedl, P. (2019). ESG Performance and Disclosure: A Cross-Country Analysis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3506084.

[4] Alpha Vs. Beta: What's the Difference? (2021). Investopedia. https://www.investopedia.com/ask/answers/102714/whats-difference-between-alpha-and-beta.asp

Index

Figure Index

Table Index

Chapter 1: INTRODUCTION

The concept of selective investment is not new. For decades, investors have been looking at different performance indicators to design their investment strategies and help them allocate their money. Financial indicators such as Return on Equity (ROE), Dividend Yield, P/E Ratio and P/CF Ratio, in conjunction with non-financial measures like the Customer Satisfaction Index, have been used to determine the performance of a company. However, investors have become increasingly more concerned about how their investments are impacting the world. This caused the birth of metrics that assess companies on an ethical level. One of these indicators is the ESG Rating.

ESG stands for environmental, social and governance and they do not only define a company's ethics but rather their commitment to standards in the fields previously mentioned. Their adherence to these standards is evaluated and result in an ESG Rating. Some examples of the factors that are analyzed are: carbon footprint and energy efficiency for the environmental area, product safety and human development for social and business ethics and fraud for governance.

The ESG Rating has become one of the metrics used for screening and strategic investment decisions. Wealthy individuals and large corporations are looking to create an impact with their money, resulting in funds that only invest in companies when their ESG score is higher than a certain threshold. As investors reward companies that have high ESG ratings, it is believed that these companies tend to outperform their peers.

Numerous studies have been conducted, some of them showing a positive correlation between ESG ratings and performance, some showing inconclusive results and a few others showing a negative relationship. Therefore, the aim of this project is to shed some light on this issue and analyze the relationship between these factors and a company's alpha.

As it is usual in finance, there is no simple answer and finding an overall correlation might be impossible. Nevertheless, finding patterns and trends can motivate investors to further incorporate

ESG ratings to their screening process, as well as allowing funds or other financial vehicles to better defend their ethical position to their clients.

Chapter 2: STATE OF THE ART

One of the first ESG Indexes surfaced in 1990 with the MSCI KLD 400 Social Index, previously called Domini 400 Social Index. This index was manufactured by a committee that weighted size, ESG and sector. However, in recent years this committee has evolved and presents a transparent quantitative assessment of more than 1,000 companies.

This trend in handling ESG quantitatively is shared by many organizations, that have steered away from a qualitative evaluation that could be biased, to offer measurable metrics. This change is important, since it allows investors to compare different companies and benchmark them against others.

However, the task of assessing a high number of companies can be both expensive and time consuming and it has led to a lack of public information on ESG ratings. Some of the main players in this field are Bloomberg, MSCI, Sustainalytics and Refinitiv. Bloomberg has ESG data on 11,700 companies, MSCI on 8,500, and both Sustainalytics and Refinitiv on more than 10,000. Some of these institutions offer their ESG coverage for a fee, allowing few external consultations on their data for research purposes. Not only that, but the differences on the factors being evaluated, their weight on the overall rating and the general methodology, can make the cross examination of ESG ratings from multiple sources difficult.

There have been multiple studies conducted and articles written on the topic. They have shifted away the perception of the early seventies, which show a much more pessimistic view of social action in a corporation [1]. One of the most famous articles, written by Milton Friedman in 1970, describes the idea that the only responsibility of a corporate executive is to increase profits for the shareholders [2].

A different research study by Paul A. Gompers, professor of Business Administration at Harvard Business School, shows an overall positive relationship between some key governance factors and financial performance [3]. At the same time, Florencio Lopez-de-Silanes, finance professor at SKEMA Business School, found a negative relationship with the riskiness of companies in countries such as United Kingdom, Australia and France, although this relation was not statistically significant [4].

A Refinitiv study conducted in 2018 shows either positive or non-existing correlation between the ESG rating and the Excess Return of a company during the years 2016, 2017 and 2018, while showing a wide range of correlations in previous years [5][6].

The techniques used for these studies vary widely, some using linear regression while others examine the data with non-linear models. The metrics employed are also substantially different. Some researchers test the relationship of ESG ratings versus alpha, while others use Return on Equity, Return on Assets or Tobin's Q to assess a company's performance. In addition to this, some studies use as a proxy for ESG the ESG Index and others only take a look a particular pillar of the three: Environmental, Social and Governance. The studies where the underlying factors of each of these levels are used are scarce. Finally, some research papers conducted on the subject perform a simulated portfolio analysis with only the top and bottom performers [7].

Chapter 3: PROJECT DEFINITION

3.1 Motivation

Firstly, as previously mentioned, there is an interest in ESG analysis that has grown over the past decade. Investors understand the importance of this metric and have started to incorporate it in their process, but there is still a lot of misinformation due to contradictory studies on the subject and widespread assumptions. Being able to provide conclusive results can help investors and financial institutions make more knowledgeable decisions.

Secondly, studies that cover this subject have encouraged companies to share their ESG practices. While in some countries it is mandatory to report the ESG related information of your company, US firms are not yet required to do so. Nevertheless, each year, companies that were previously not covered have increased their disclosure of ESG reporting. Keeping the conversation on this topic alive, as well as providing new insights can motivate a company's stakeholders to demand the disclosure of this information, making companies more concerned with their practices and enhancing responsible business operations.

Finally, if the main hypothesis of this project is correct (i.e., a positive correlation exists between ESG ratings and financial performance), companies that are socially responsible will be rewarded in the long term, as even more investors will realize their potential to generate revenue. Companies with low ESG ratings will be penalized with less investments and they may be motivated to improve.

3.2 Objectives

Companies will be divided using clustering techniques and pre-established criteria: Clustering will be performed to find patterns and groups within the dataset of companies. Firms will also be classified based on country and industry, since the same benchmark cannot be used for all of them.

The relationship between ESG ratings and alpha will be explored using programming tools. Linear regression methods will be used to shed some light on this correlation. Some of these include simple Linear Regression, PLS, Random Forests and Gradient Boosting. Non-linear regression such as Polynomial Regression will also be performed to have a variety of options for the model selection.

The model that best fits the data will be used to determine the correlation between alpha and ESG Ratings. The most important factors, and whether these factors have a positive or negative relationship to alpha, will be determined.

3.3 Methodology

The methodology of this project will consist of 7 steps:

1. Research: the starting point of this thesis was to look into the different studies and literature that exist concerning the topic of ESG.

2. Data acquisition: The data is acquired from three sources: MSCI ESG Direct, Bloomberg and Yahoo Finance. The process requires web scraping from Yahoo Finance.

3. Data storage: Because of the large amount of data being handled, as well as the difficulty and time-consuming process to obtain it, it is necessary to store the acquired data in a way that allows ease of retrieval and administration. Excel and csv are used for small subsets of the data while SQL is utilized for larger pieces.

4. Data cleaning and preprocessing: Due to the nature of the project, data cleaning and processing is an integral part of the thesis. The treatment of missing values, outliers and qualitative variables, as well as the joint of data is presented in this section.

5. Regression and classification techniques: The main part of the thesis is the use of regression techniques to extract the relationship between ESG ratings and alpha.

6. Analysis and conclusions: After obtaining the main results from the data, a comprehensive analysis is done, completing the main objective of this thesis.

3.4 Alignment with the Sustainable Development Goals (SDG)

This project focuses on the topic of ESG and, therefore, it's alignment with the Sustainable Development Objectives is almost self-explanatory.

To start with, the objectives Good Health and Well-being (Goal 3), Clean Water and Sanitation (Goal 6), Affordable and Clean Energy (Goal 7), Climate Action (Goal 13), Life Below Water (Goal 14) and Life On Land (Goal 15) are directly related to some key issues assessed in the ESG Environmental score. Among other factors, Opportunities in Clean Tech, Opportunities in Renewable Energy, Toxic Emissions, Waste, Biodiversity and Land Use are evaluated on this subcategory.

Secondly, the goals of Decent Work and Economic Growth (Goal 8), Industry Innovation and Infrastructure (Goal 9), Responsible Consumption and Production (Goal 12) and Partnership for the Goals (Goal 17) are present in the following scores, among others: Financial System Instability, Business Ethics, Fraud, Product Safety, Human Capital Development and Access to Communication.

All of these factors are directly considered to develop the model and extract the correlation between ESG and financial performance. The effect that these features have on this correlation can be checked and conclusions on how they affect companies are analyzed.

On a larger picture, one of the objectives of this project is to spread awareness on how these factors can impact the performance of a company, therefore motivating them to adopt measures that can ensure a sustainable, responsible and environmentally friendly development. Not only that, but it will also encourage ESG investing by showing that the ESG rating is a driving factor in the performance of a company.

Chapter 4: Data Analysis

In this section the main focus of the project is discussed. Firstly, the dataset and variables is described, since this is the starting point of the regression, and understanding the data will make the explanations of future sections easier to understand. Secondly, the steps taken in acquiring this data are detailed. Thirdly, the process of cleaning and preprocessing the dataset, as well as a description of the final dataset, is shown. Finally, the different techniques and methods used to obtain the correlation are reported.

4.1 Dataset

The final data for the thesis consists of both explanatory variables and response variables. Both of them are explained in depth in the next subparts.

4.1.1 Explanatory variables

The explanatory variables are the ESG ratings of more than 8,000 companies and these will be the variables used to try to explain and predict the company's performance. They will also be called features, predictors and independent variables throughout this report.

These variables have three different importance levels:

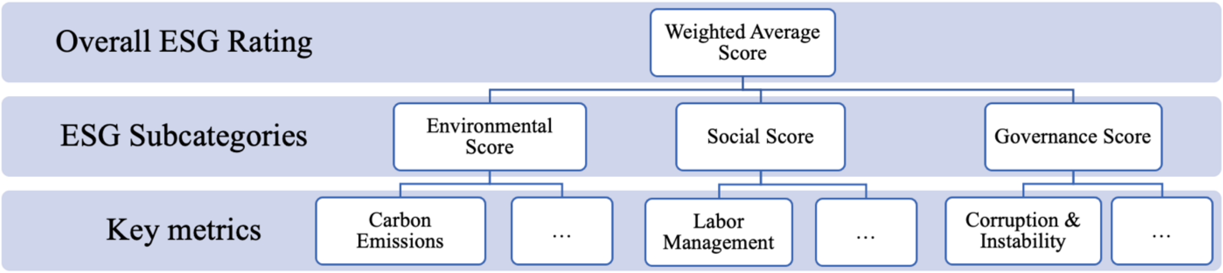


Figure 1. Explanatory variables schematic.

The overall ESG Rating is the most important one, since it is the aggregate of the other variables and represents the final score of the company. Two variables represent this score: IVA

Company Rating and Weighted Average Score. IVA Company Rating is the result of assigning a ranking from best (AAA) to worst (CCC) using the Weighted Average Score. As such, the IVA variable is categorical. The Weighted Average Score is a continuous variable and, as most of the other variables, operates on a 0-10 scale. Since both of these variables convey the same information, the Weighted Average Score is the one used for the regression.

The ESG subcategories are three and they are the Environmental, Social and Governance scores. They represent the score in each of these three fields. They are continuous, based on a 0-10 scale. Each of these three variables has a weight associated to them that is used to calculate the overall rating. As well, these weights can vary among different companies. The reason for applying different weights is that a metric may be much more relevant to a company than other. For example, the Carbon Emissions Score, is more important to a utility company than a software firm.

The final level in importance contains the key ESG metrics. These are the factors from which the other three ESG subcategories are calculated. They are more than a hundred and they are of three different types:

- Continuous on a 0-10 scale. Some examples of these variables are: Toxic Emissions and Waste Score, Responsible Investment Score and the Human Development Score.
- Continuous on a 0-100 scale. Some examples of these variables are: Board Percentile Rank or the Governance Percentile Rank.
- Binary (either a 0 or a 1 score). Some examples of these variables are: Debt Covenant Breach or Executive Misconduct.

In most cases, the higher the score the better the company does in these areas. The exception are the binary variables, that are used to flag occurrences and are marked as 1 when the situation has happened. For example, in the case of the variable Executive Misconduct, it would be flagged with a 1 if a senior executive of the company has faced prosecution for personal misconduct. Therefore, whether or not a 1 in these variables is beneficial or damaging depends on the particular factor that is being looked at.

As explained, these factors are aggregated and weighted depending on their relevance. For example, for companies in the Utilities industry, metrics such as Carbon Footprint may be more relevant than Access to Finance (metric that describes “the extent to which a company provides lending, financing, or products to underrepresented or underbanked communities”). Therefore, the first factor will have a bigger impact in the overall ESG Rating of a Utilities company than the second one. This methodology allows investors to compare different companies, since they are valuated based on the key issues that they face.

Once the regression models are run and the most relevant variables can be distinguished, a more through explanation of each of these variables will be delivered.

4.1.2 Response variables

The response variables, also called dependent variables, are the financial metrics. These are the variables for which we are trying to find a correlation with the ESG ratings.

The main response variable examined at is alpha. Alpha is “the excess return earned on an investment above the benchmark return” [8]. In a portfolio of securities, the return of the portfolio is compared to a specific benchmark, resulting in alpha. As such, this metric is often used to evaluate the value that the portfolio manager has added with its managing skills, decisions and strategy.

This concept is also applied to individual stocks, where the company’s alpha is the company’s return above a particular index or benchmark. This metric is very relevant, since it eliminates the proportion of the return that might be due to the overall market or industry doing well, leaving the added value of the company.

Alpha is one of the key parameters in the Capital Asset Pricing Model, a widely spread and accepted model to calculate the required rate of return of an investment. This model computes alpha as the formula shown below.

$$\alpha = R_s - [R_f + (R_m - R_f) \cdot \beta]$$

Where:

- R_s : Return of stock or portfolio
- R_f : Risk-free rate
- R_m : Market return

In the formula above, the parameter β is a measure of the volatility of a stock, and it is used to assess how the stock's return behaves in comparison to the market. For example, if the beta of Company A is 1.2, when the market goes up 100 basis points (an increase of 1%), Company A would theoretically move up 120 basis points (1.2% increase).

It is worth noting that, when referring to the market return, it is understood that a proxy for it is used. Normally this proxy is an index, such as the S&P 500 in the United States. In a similar fashion, a proxy such as the one-month Treasure Bill can be used for the Risk-Free Rate.

In this particular case, the alpha used in our models is the CAPM alpha computed over a time period of two years, for the years 2015-2016, 2016-2017, 2017-2018, 2018-2019. Finally, a one-year alpha is also studied, corresponding to the year 2020.

In addition to alpha, the volatility in the last 360 days is also contemplated as a dependent variable. ESG stocks are often perceived as more stable, so looking for a correlation between ESG ratings and volatility would be interesting as a side goal. By running regression models with the volatility, we can try to determine the truth behind this perception [9].

4.2 Data acquisition

The data will be obtained from three sources: ESG MSCI Direct, Yahoo Finance and Bloomberg. Each of these actions will be explained in depth in this part of the report but a general overview of the process can be seen below (*Figure 2*).

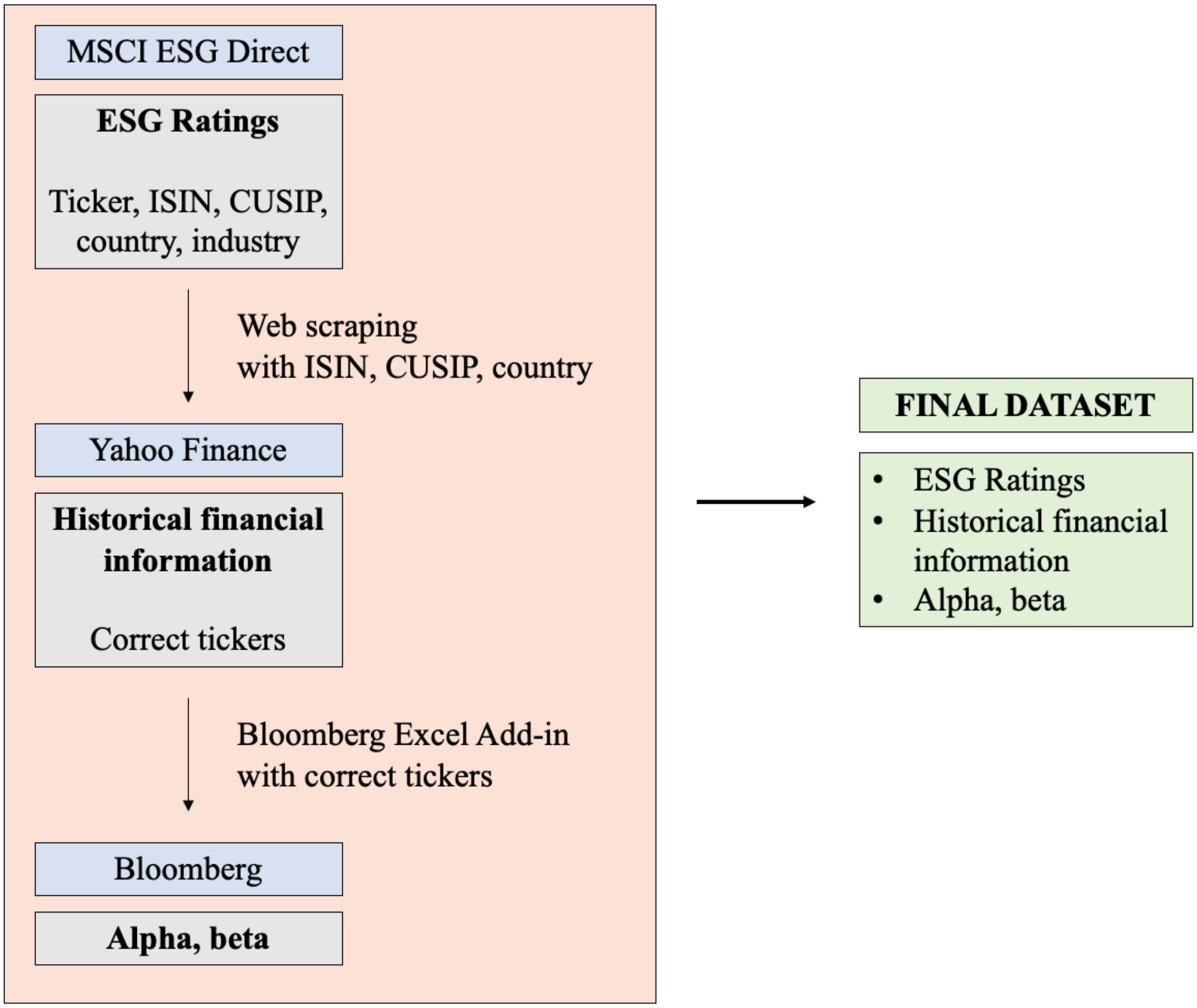
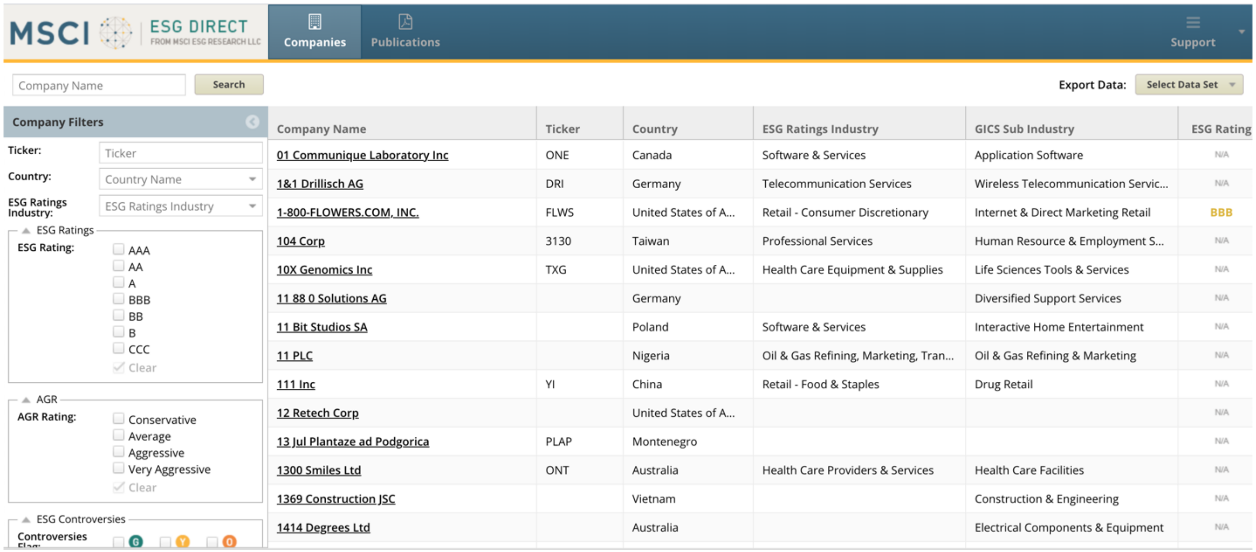

MSCI ESG Direct
ESG Ratings
Ticker, ISIN, CUSIP, country, industry
Web scraping with ISIN, CUSIP, country
Yahoo Finance
Historical financial information
Correct tickers
Bloomberg Excel Add-in with correct tickers
Bloomberg
Alpha, beta
FINAL DATASET
ESG Ratings
Historical financial information
Alpha, beta

Figure 2. Data acquisition overview.

4.2.1 ESG Ratings

The main component of the dataset will be the ESG ratings of more than 8,000 companies. Several organizations provide similar databases on ESG information, but the major ones are Bloomberg, Sustainalytics, MSCI and Refinitiv.

Since the evaluation of ESG factors is expensive and time consuming, these resources charge a fee to make use of their data. This project is purely done for research purposes, and as such, paying these high fees is not possible. Fortunately, Columbia University Library Resources grants access to the MSCI ESG database. Therefore, this will be the dataset that will be used for the entirety of this thesis. A snippet of this dataset is shown below (*Figure 3*).



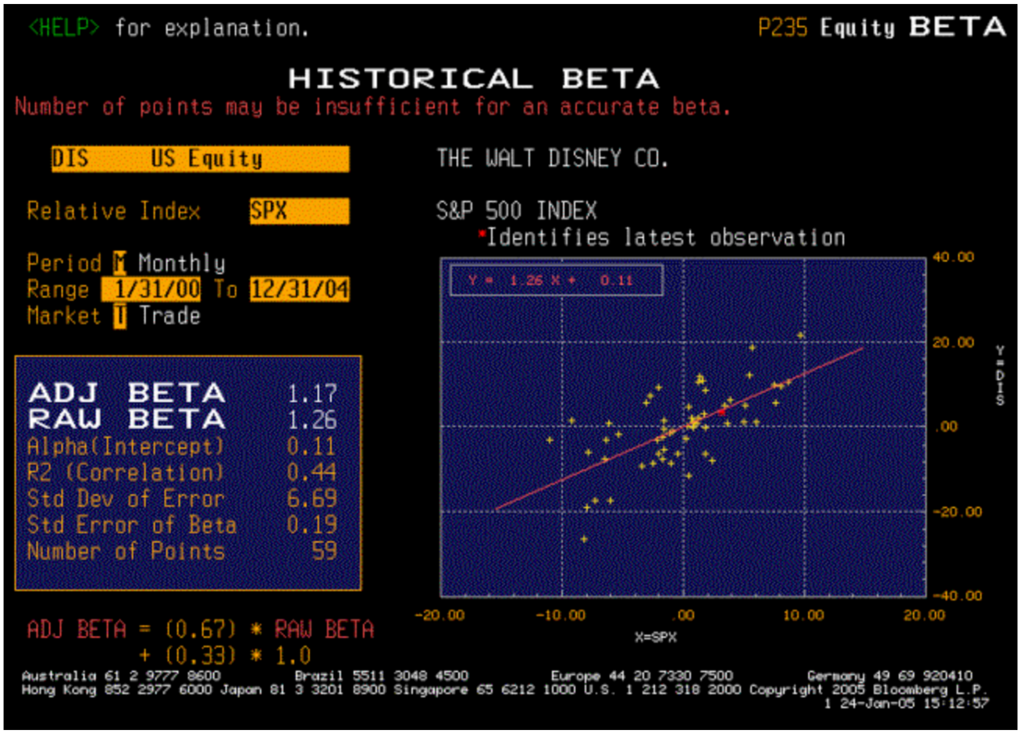
Company Name	Ticker	Country	ESG Ratings Industry	GICS Sub Industry	ESG Rating
01 Communique Laboratory Inc	ONE	Canada	Software & Services	Application Software	N/A
1&1 Drillisch AG	DRI	Germany	Telecommunication Services	Wireless Telecommunication Servic...	N/A
1-800-FLOWERS.COM, INC.	FLWS	United States of A...	Retail - Consumer Discretionary	Internet & Direct Marketing Retail	BBB
104 Corp	3130	Taiwan	Professional Services	Human Resource & Employment S...	N/A
10X Genomics Inc	TXG	United States of A...	Health Care Equipment & Supplies	Life Sciences Tools & Services	N/A
11 88 0 Solutions AG		Germany		Diversified Support Services	N/A
11 Bit Studios SA		Poland	Software & Services	Interactive Home Entertainment	N/A
11 PLC		Nigeria	Oil & Gas Refining, Marketing, Tran...	Oil & Gas Refining & Marketing	N/A
111 Inc	YI	China	Retail - Food & Staples	Drug Retail	N/A
12 Retech Corp		United States of A...			N/A
13 Jul Plantaze ad Podgorica	PLAP	Montenegro			N/A
1300 Smiles Ltd	ONT	Australia	Health Care Providers & Services	Health Care Facilities	N/A
1369 Construction JSC		Vietnam		Construction & Engineering	N/A
1414 Degrees Ltd		Australia		Electrical Components & Equipment	N/A

Figure 3. MSCI ESG Direct Dataset.

4.2.2 Alpha: Bloomberg

Once the ESG ratings have been obtained, some financial information is needed to assess the performance of companies. As described, alpha is the metric that has been chosen as the proxy for financial performance.

The alpha was obtained from Bloomberg using its Excel Add-in. Specifying the correct company ticker and the metric to be obtained, Bloomberg automatically fills the spreadsheet with the information requested. However, each request is counted and, if the data download limit is reached, the Bloomberg Terminal user will not be able to make any more requests until the next period, normally the next month. Therefore, with such a large number of companies, it is essential to be precise and request only for the data that is needed without mistakes. In this case, the downloaded data corresponds to the alphas, betas and volatility.



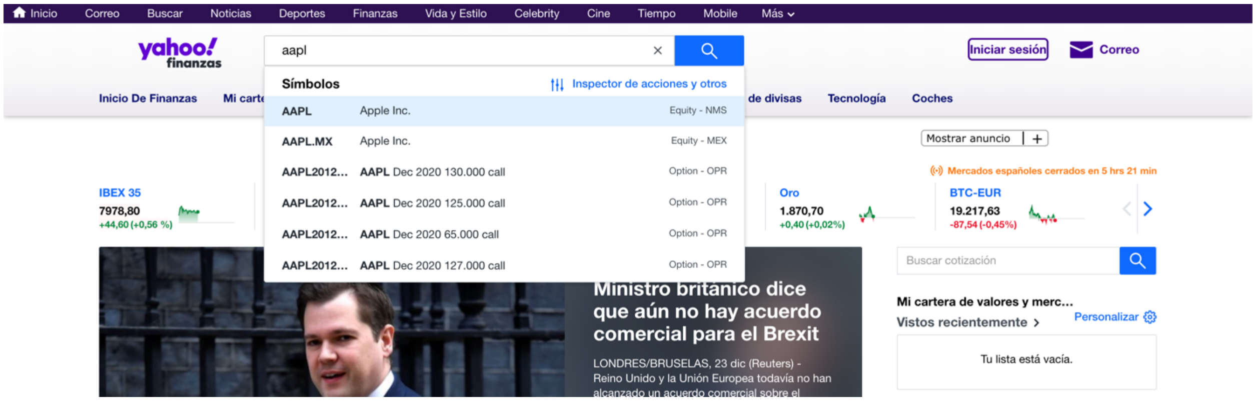
```
<HELP> for explanation.                              P235 Equity BETA
                      HISTORICAL  BETA
Number of points may be insufficient for an accurate beta.

  DIS      US Equity              THE WALT DISNEY CO.

 Relative Index    SPX            S&P 500 INDEX
                                    *Identifies latest observation
 Period M Monthly
 Range  1/31/00 To 12/31/04
 Market T Trade

 ADJ  BETA            1.17
 RAW  BETA            1.26
 Alpha(Intercept)     0.11
 R2 (Correlation)     0.44
 Std Dev of Error     6.69
 Std Error of Beta    0.19
 Number of Points     59

 ADJ BETA = (0.67) * RAW BETA
          + (0.33) * 1.0
Australia 61 2 9777 8600        Brazil 5511 3048 4500        Europe 44 20 7330 7500        Germany 49 69 920410
Hong Kong 852 2977 6000 Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2005 Bloomberg L.P.
                                                                                          1 24-Jan-05 15:12:57
```

Figure 4. Bloomberg Terminal: Beta and Alpha.

4.2.3 Financial information: Yahoo Finance

In addition to obtaining alpha, beta and the volatility, the historical prices of these companies were obtained to better understand them and obtain other metrics such as the return of the past two years. As previously mentioned, downloading large quantities of data from Bloomberg might not be the best option and, in this case, the data limit would have been reached very early on. As a consequence, Yahoo Finance was used.

Yahoo Finance is a Yahoo free service that contains financial information as well as commentary and analysis on companies. Despite being free to use, there is no free option on their website to download massive amounts of data, as the only way to check a company's information is to manually search for the company in the search bar. As a consequence, a Python program was written to automatically make requests on the website and obtain the information needed. This process is called web scraping.



aapl

Símbolos — Inspector de acciones y otros

AAPL	Apple Inc.	Equity - NMS
AAPL.MX	Apple Inc.	Equity - MEX
AAPL2012...	AAPL Dec 2020 130.000 call	Option - OPR
AAPL2012...	AAPL Dec 2020 125.000 call	Option - OPR
AAPL2012...	AAPL Dec 2020 65.000 call	Option - OPR
AAPL2012...	AAPL Dec 2020 127.000 call	Option - OPR

Mercados españoles cerrados en 5 hrs 21 min

IBEX 35	Oro	BTC-EUR
7978,80	1.870,70	19.217,63
+44,60 (+0,56 %)	+0,40 (+0,02%)	-87,54 (-0,45%)

Ministro británico dice que aún no hay acuerdo comercial para el Brexit

LONDRES/BRUSELAS, 23 dic (Reuters) - Reino Unido y la Unión Europea todavía no han alcanzado un acuerdo comercial sobre el

Mi cartera de valores y merc...

Vistos recientemente

Tu lista está vacía.

Figure 5. Yahoo Finance search bar.

There are some considerations to take into account when web scraping. Web scraping is not permitted in all websites, so it is important to read the Terms of Use of the website that contains the data that is being searched. In the case of Yahoo Finance, there is not specific reference to it. However, it says the following: "you agree not to […] interfere with or disrupt the Services or servers, systems or networks connected to the Services in any way". As well, it is stated: "[…] you may not access or reuse the Services, or any portion thereof, for any commercial purpose". As a consequence, the web scraping will be done responsibly, allowing time in between each request and with the only purpose of research.

The information obtained is in Json. Json is a text format to store and transport data similar to XML or BSON. Because of this, after obtaining the data, it had to be parsed (act of analyzing the syntax of a string of symbols). For this particular case, parsing can be viewed as a way to interpret information that is not on our programming language.

After parsing, the data was stored in SQLite. SQLite allows us to create and manage a database making it is easy to store and extract data when needed. SQLite offers some important advantages over CVS files.

To start with, the SQL language allows us to select and load a particular subset of data, as oppose to loading/reading a whole CVS file and then selecting the needed subset.

In addition, this program has built in functions that can be called through Python to check whether or not a piece of information is already in the dataset. This is very useful, since it takes time to gather data and we do not want to waste requests on information that we have already obtained.

Finally, the information is stored in the dataset "one by one". This means that the moment the historical information of one company is obtained, it is stored in the database without waiting for the rest of the companies' data. It takes an estimate of more than 6 hours to obtain all companies and, during that time, the internet connection can be interrupted, the computer can crash or even Python can run out of memory. As a consequence, the ability to be able to store the information one at a time is very important.

It is worth pointing out that the tickers (symbols that identify a company) from the MSCI ESG database do not correspond to either the Bloomberg or the Yahoo Finance tickers. Therefore, some transformations needed to be made to search for these companies on the two platforms. As a consequence, the companies where first searched on Yahoo Finance using their ISIN or CUSIP code from MSCI to obtain the correct ticker. However, not only some ISIN and CUSIP codes were not found in the dataset, but also these codes are not unique. Depending on the stock exchange where the company is traded, Bloomberg adds a suffix to the ticker, so this suffix was incorporated through Python.

To help visualize this process, a small sketch of the steps to store the companies “one by one” is shown below (*Figure 6*).

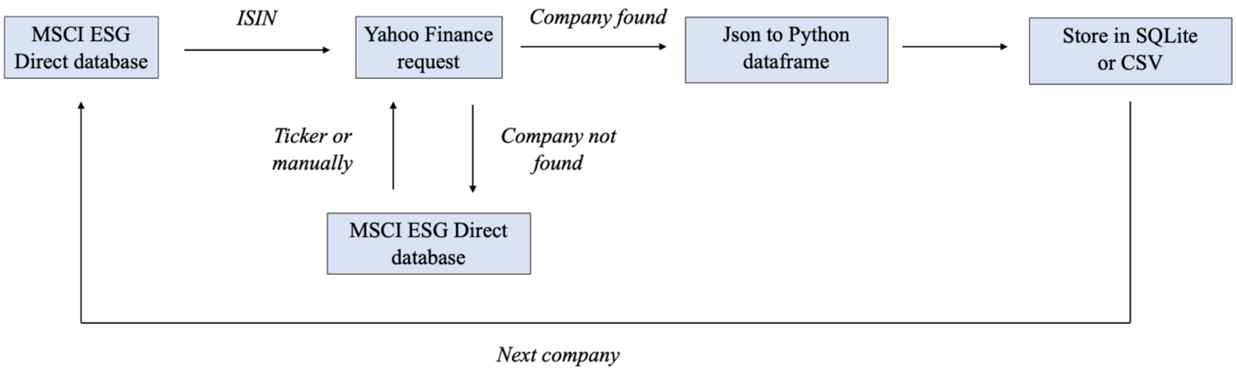


Figure 6. Yahoo Finance data acquisition and storage process.

4.3 Data cleaning and preprocessing

Due to the regression techniques that are going to be used on the data, the cleaning and preprocessing becomes a key element of this thesis.

Firstly, some variables from the original dataset need to be eliminated. These columns are primarily the weights of each ESG category and some information of the companies.

The overall ESG rating, as well as the three subcategories (Environmental, Social and Governance) are obtained by weighting the key metrics, such as environmental emissions or fraud cases. For this study, these weights are not relevant, since the overall ratings are already present in the dataset. As a consequence, these columns are eliminated.

In a similar fashion, the variables that describe a company such as their name, and their codes (ISIN, CUSIP, SEDOL, etc.) are also unneeded. Only the ticker of each company is spared as reference, but this variable will also be eliminated once we start applying regression methods to the dataset.

Secondly, the outliers need to be taken care of because they may have influence on the results or may decrease statistical power. Outliers can also represent errors in data collection, which we want to minimize as much as possible. A standard practice is to consider datapoints as outliers when they exceed the mean plus/minus three times the standard deviation.

Although it would be surprising to find outliers in the ESG ratings, since these categories are assessed and given a grade that fall within a 0-10 scale for example, it is imperative to verify it. The financial data of the companies do contain outliers (i.e., companies that have performed incredibly well or incredibly poorly) and so these companies are eliminated from the dataset. These companies only amount to a few dozen and given our +8,000 dataset, we can determine that it is not a substantial loss of data.

Thirdly, missing data is quite substantial on this dataset and, if not handled correctly, it can harm our statistical study. We need to distinguish between missing data on alpha or ESG variables, because they will be treated differently.

As previously mentioned, ESG information can be hard to gather, even with the company's collaboration. Because of the number of categories being evaluated (more than 100) it is only natural that a lot of these categories have missing data. A common practice for smaller datasets, in terms of independent variables, is to eliminate all rows that are missing one datapoint. However, if we decided to do this, we would end up with no companies at all. Another approach would be to eliminate data depending on the particular cluster and technique we are using. However, due to large number of methods that will be used, this idea is not optimal. The comprehensive solution that has been implemented in the end, is to eliminate the columns that have a missing number of datapoints over a certain threshold. Thanks to having the overall rating of each category, having a small number of missing key metrics will not result in a big problem from a statistical point of view.

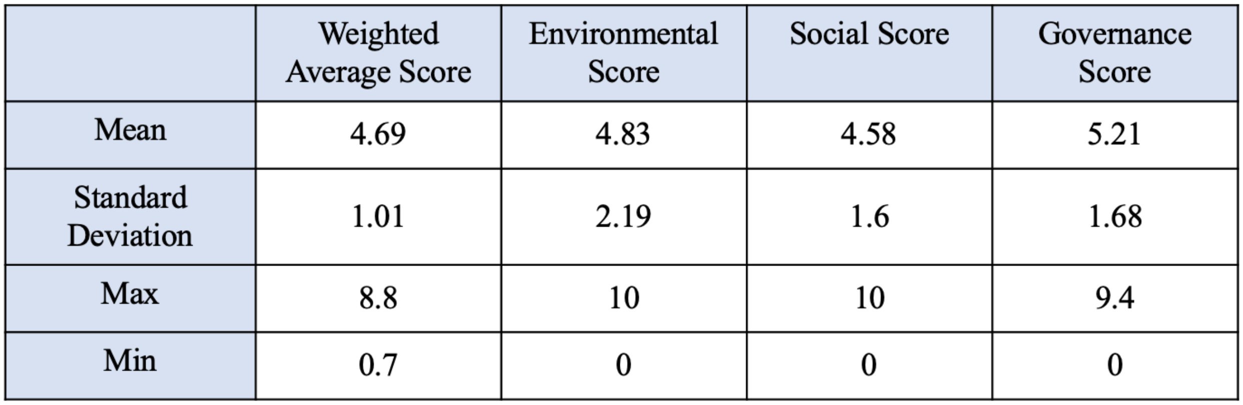
Regarding financial variables, not having missing data is much more important. We are using ESG independent variables to try to predict alpha. Therefore, if we have, for example, a missing alpha in a row, this particular company becomes irrelevant for the analysis. As a result, we eliminate all rows that miss this metric. It is worth noting that we will not eliminate them all at once before applying the regression methods, but rather eliminate them at the moment we perform them. This is due to the fact that it is fairly normal for companies to not have a financial metric for every year (company A may not have been public 3 years ago), but they may have data in more recent years, and therefore, we want to keep it when finding a correlation to "Alpha 2019", for example.

Finally, the dataset was checked for duplicates and invalid numerical values. In this regard, there was no loss of data due to this.

4.4 Final dataset

After cleaning and preprocessing the final dataset is obtained. This dataset has dimensions of 6,699 by 135. There is a row for each company (+6,500 of them) and a column for each variable (more than 120 between dependent and independent variables).

To understand the profile of the companies that are present in the dataset we compute the mean, standard deviation, the maximum value and the minimum value of the most important ESG variables and those of the dependent variables.



	Weighted Average Score	Environmental Score	Social Score	Governance Score
Mean	4.69	4.83	4.58	5.21
Standard Deviation	1.01	2.19	1.6	1.68
Max	8.8	10	10	9.4
Min	0.7	0	0	0

Table 1. Mean, standard deviation, max and min value of principal variables.

As it can be observed, the mean in all three categories, as well as the mean in the overall ESG rating, fall close to a score of 4.75, with standard deviations that indicate quite a spread in the data of the environmental, social and governance score, but not on the overall ESG rating. Although the maximum in all three categories is almost 10, the highest overall Weighted Average Score is 8.8.

	Alpha 2015-2016	Alpha 2016-2017	Alpha 2017-2018	Alpha 2018-2019	Alpha 2020	Beta 2019	Volatility last 360 days
Mean	0.17	0.18	0.07	0.041	0.1	0.97	49.47
Standard Deviation	0.49	1.24	0.63	0.47	0.72	0.3	17.82
Max	4.05	48.47	14.63	2.75	3.11	2.26	112.83
Min	-3.25	-53.3	-8.41	-2.53	-2.76	-0.33	0.01

Table 2. Mean, std, max and min of the dependent variables.

As for the financial variables, these measures are in line with our expectations. Overall, the means of the alphas are very similar, as well as their standard deviations. A special case, however, is Alpha 2016-2017, where we find not only a very high and very low maximum and minimum value respectively, but also a high standard deviation in comparison with other years.

As for the beta, the mean of 0.97 is appropriate, since that means the average company will have returns in line with the market. A negative beta is not impossible, but the opposite is much more common and so having a maximum of 2.26 and a minimum of -0.33 fits with our forecast. Lastly, the volatility can only be positive, hence the minimum value of almost zero.

Significant plots such as the ESG Score against alpha and the number of companies in each rating (AAA-CCC) are also key to understand the dataset.

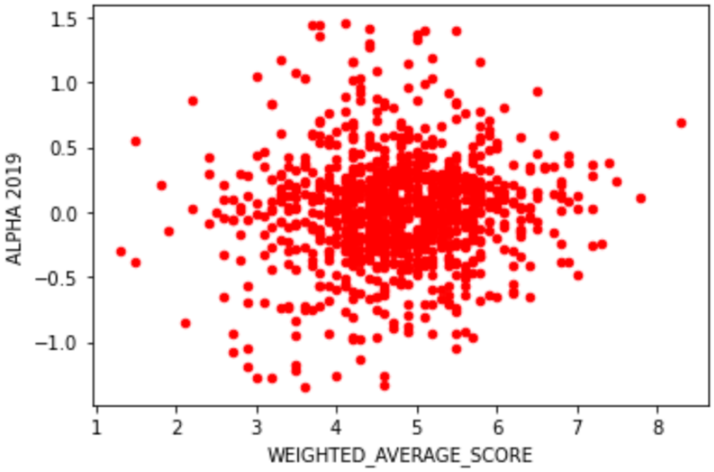


Figure 7. Plot of Weighted Average Score against Alpha 2018-2019.

The plot was done taking a random sample of 1,000 companies, since plotting all of them would result in a messier and more confusing plot. As it can be seen in the plot, there is no discernible pattern or relationship, which makes the use of different techniques and methods a must in order to find a correlation.

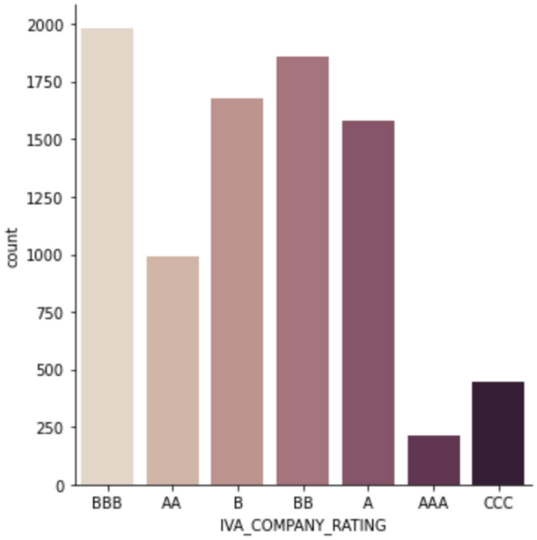


Figure 8. Count of companies in each rating (AAA-CCC).

It can be observed that, as expected, the categories in the extremes (AAA and CCC) contain the least number of companies. There are as well more companies on the lower half of the scale (CCC, B, BB) than on the upper half (AAA, AA, A).

4.5 Clustering

Clustering refers to the process of using statistical techniques and models to aggregate observations into meaningful subgroups, finding patterns. More precisely, clustering aims to find homogeneity and heterogeneity among the data, and it is a form of unsupervised learning, since the response is not considered in the models.

Two different clustering methods are performed: K-means Clustering and Hierarchical Clustering. The main difference between the two is that K-means Clustering divides the data into a pre-specified number of groups, also called clusters, while Hierarchical Clustering does not.

4.5.1 K-means clustering

The main idea behind K-means Clustering is to divide the data in groups and try to minimize the variability within the clusters. In order to do this, the algorithm performs a number of iterations, improving in each of them the clustering of the data.

As previously mentioned, in this method, a number of clusters K is specified, and the first iteration divides the data randomly in K clusters. This iteration serves as the initial groups for the data and as a first approach to finding the best possible grouping. The second iteration calculates the centroid of each cluster (the mean of all the observations within that group) and assigns each observation to the cluster with the closest centroid. The rest of iterations follow the same process as the second one, stopping when the clusters do not change anymore [10].

A visual representation of this process can be viewed below. In this graph, the final clusters are visualized easily, since the data only has two variables and, therefore, a 2D graph can be plotted. However, clustering with more variables is much messier and hard to visualize.

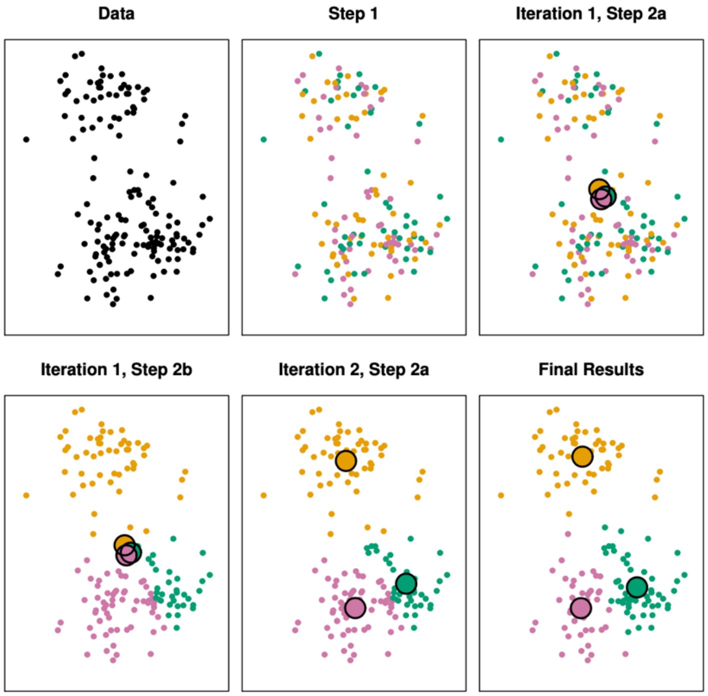


Figure 9. K-means Clustering iterations process example.

As previously mentioned, a number of clusters K needs to be specified. To correctly choose this parameter, an automated process of trial an error has been conducted. In each trial a different number of clusters is used, and the quality of the clusters is assessed. Thanks to this approach, a plot with the error and the number of clusters from each trial can be extracted. The error used to evaluate the quality is the SSE (Sum of Squared Estimate of Errors).

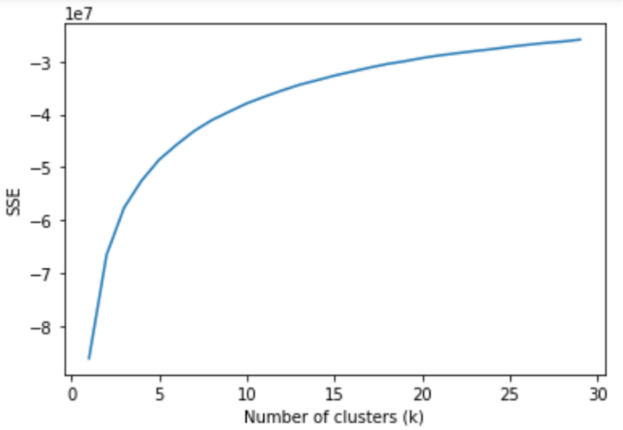


Figure 10. SSE vs Number of clusters plot in K-means Clustering.

The SSE in the graph (*Figure 10*) is inversed due to some programming constraints, so the lower in the vertical axis, the higher the error. Therefore, the higher the number of clusters the lower the error is. However, a point is reached in which performing the clustering with a high K is not worth it. This is because of three reasons: the improvement in the error is not as substantial, so many clusters can be very difficult to visualize and the higher the K, the more computationally exhausting is the process. Therefore, the elbow analysis is conducted, in which the number of clusters selected is a middle point (the elbow in the graph) where the error is not as high, but the number of clusters is also manageable. The number of clusters selected was 10.

An important note on this algorithm is that it uses categorical variables and, because of the large number of unique values of alpha, this variable had to be put in categories based on their values.

As the number of variables is high, not all the plots obtained from this process can be shown. However, the two most significant graphs can be found below, which plot the clusters with the Environmental Score against the Governance Score and alpha against the Carbon Emissions Score.

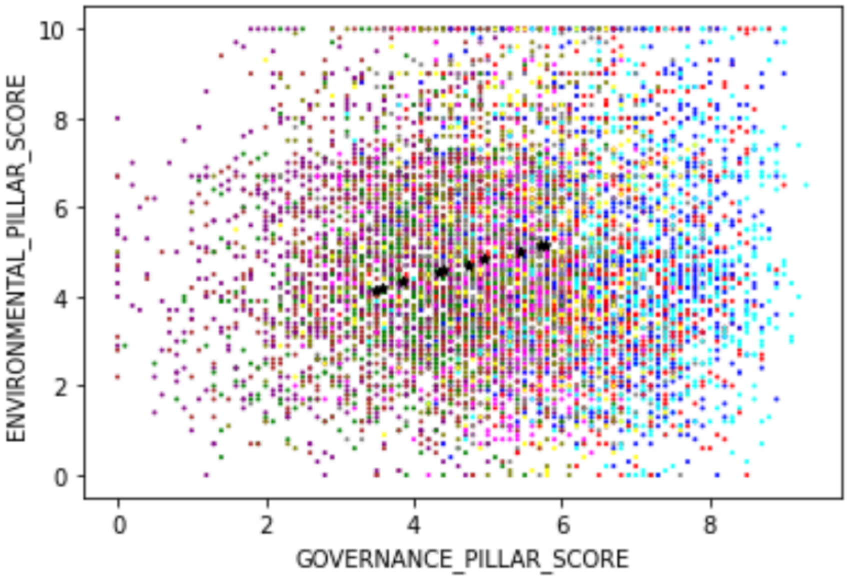


Figure 11. K-means clustering: Governance Score vs Environmental Score.

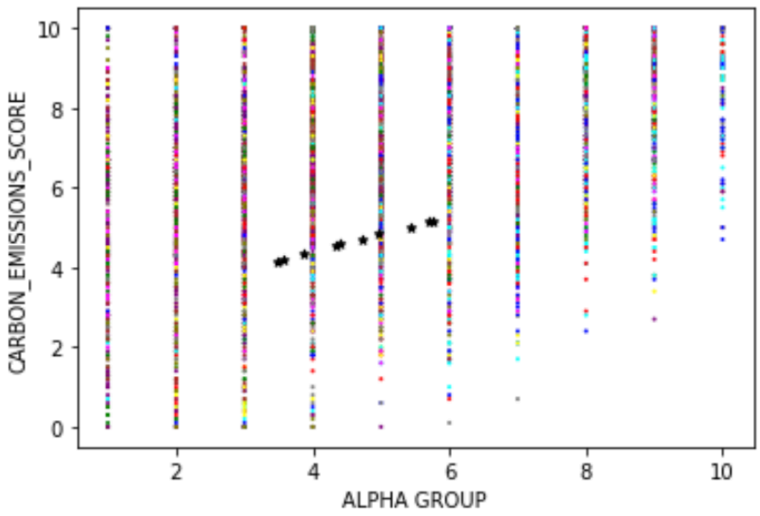


Figure 12. K-means clustering: Alpha vs Carbon Emissions Score.

With these plots we can see that the algorithm has grouped up clusters that represent quite well the Governance Score. Clusters in light blue can be seen having high scores in Governance and belong, at the same time, to the upper half of our alpha groups. We start seeing with this clustering that high ESG scores may lead to high alphas.

4.5.2 Hierarchical Clustering

In K-means, the first step was to assign observations into K clusters randomly. In Hierarchical Clustering the starting point is having each observation as its own cluster (bottom-up approach). The algorithm then computes the distance between these clusters and groups up the closest ones. This distance is often referred to as dissimilarities or linkages, and so, in each step, the method groups the least dissimilar clusters. Furthermore, in each step the groups become larger, forming a tree-like structure called dendrogram.

In the first step of the process the distance or linkage is just the Euclidean Distance between the points, but further down the algorithm, we need to specify what the distance between clusters is. There are four types: single, average, centroid and complete. Simple linkage is the distance between the closest points in the clusters, while the complete linkage is the opposite: distance between the furthest observations in the clusters. The centroid linkage is the distance between centroids and the average linkage is the average of the distances of each observation in the first cluster with each observation in the second cluster.

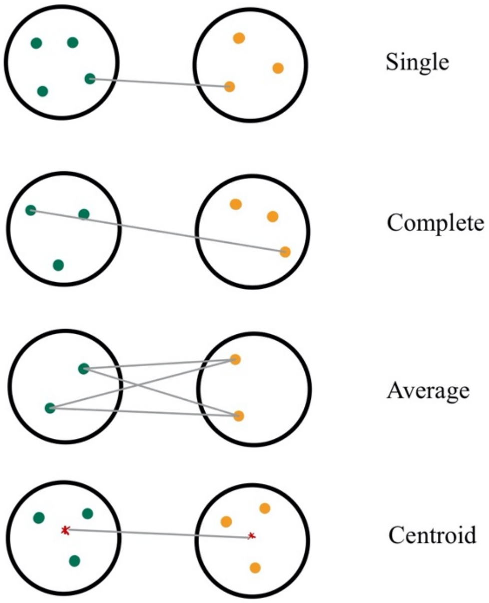


Figure 13. Types of linkages in Hierarchical Clustering.

Due to this linkage concept, it becomes necessary to scale the data. Since the distances in all the variables that define observations are being computed, it is imperative to give the same weight to all variables and, therefore, scale. For example, if there were two variables in a dataset: price, in the order of $1-100, and volume, in the thousands; and no scaling, the clusters would almost exclusively be decided based on the variable volume because of its order of magnitude. We make adjustments so that every feature has mean 0 and standard deviation 1.

These four linkage options were tried to try to obtain the best possible clustering. However, none of them contributed in a meaningful way, as the clusters did not show any specific pattern that was yet unknown. Therefore, another approach, called CART (Classification and Regression Trees) was tried. This algorithm is based on the concept of the Gini Impurity. Gini impurity is a criterion to determine the quality of a split (top-bottom structure). Roughly speaking, gini measures the probability of randomly chosen observations being wrongly classified.

After running the CART algorithm, the following dendrogram or decision tree was obtained. Because of its large size, only a part of the graph is shown below (*Figure 14*).

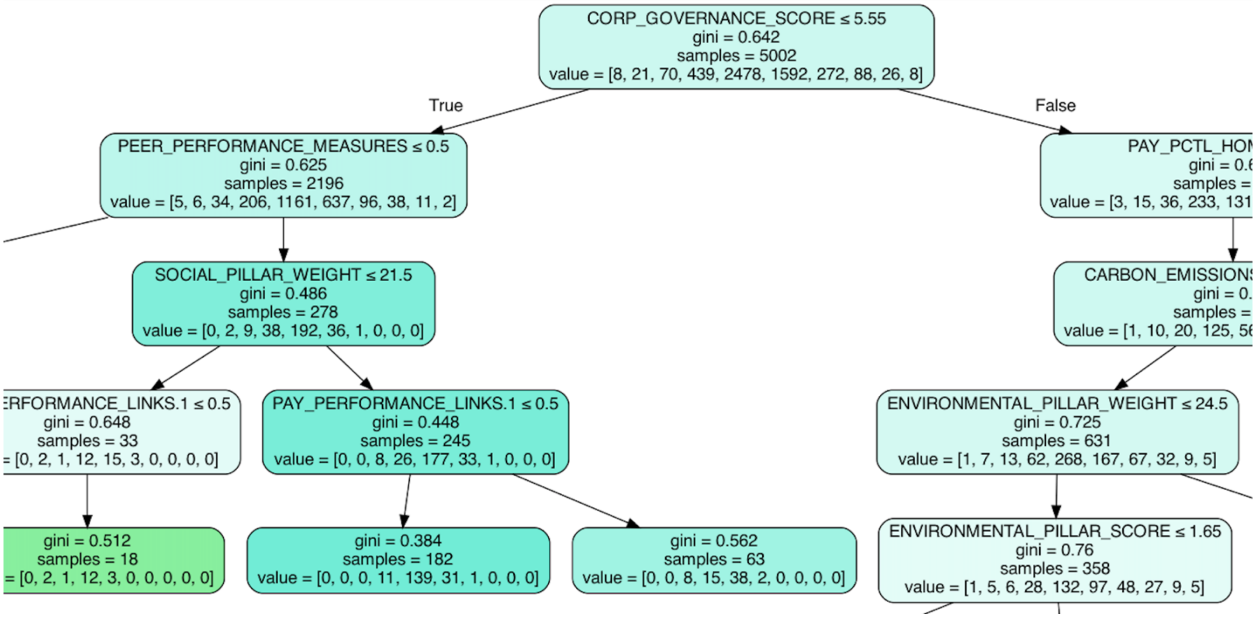

CORP_GOVERNANCE_SCORE ≤ 5.55
gini = 0.642
samples = 5002
value = [8, 21, 70, 439, 2478, 1592, 272, 88, 26, 8]
True
False
PEER_PERFORMANCE_MEASURES ≤ 0.5
gini = 0.625
samples = 2196
value = [5, 6, 34, 206, 1161, 637, 96, 38, 11, 2]
PAY_PCTL_HON
gini = 0.6
samples =
value = [3, 15, 36, 233, 131
SOCIAL_PILLAR_WEIGHT ≤ 21.5
gini = 0.486
samples = 278
value = [0, 2, 9, 38, 192, 36, 1, 0, 0, 0]
CARBON_EMISSIONS
gini = 0.
samples =
value = [1, 10, 20, 125, 56
ERFORMANCE_LINKS.1 ≤ 0.5
gini = 0.648
samples = 33
= [0, 2, 1, 12, 15, 3, 0, 0, 0, 0]
PAY_PERFORMANCE_LINKS.1 ≤ 0.5
gini = 0.448
samples = 245
value = [0, 0, 8, 26, 177, 33, 1, 0, 0, 0]
ENVIRONMENTAL_PILLAR_WEIGHT ≤ 24.5
gini = 0.725
samples = 631
value = [1, 7, 13, 62, 268, 167, 67, 32, 9, 5]
gini = 0.512
samples = 18
= [0, 2, 1, 12, 3, 0, 0, 0, 0, 0]
gini = 0.384
samples = 182
value = [0, 0, 0, 11, 139, 31, 1, 0, 0, 0]
gini = 0.562
samples = 63
value = [0, 0, 8, 15, 38, 2, 0, 0, 0, 0]
ENVIRONMENTAL_PILLAR_SCORE ≤ 1.65
gini = 0.76
samples = 358
value = [1, 5, 6, 28, 132, 97, 48, 27, 9, 5]

Figure 14. Dendogram cut using CART algorithm.

To visualize the clusters, they were plotted in pairs of two variables, such as the plot below, where the Environmental Score was plotted against the Corporate Governance Score and each of them against alpha. In this particular plot, Alpha 2019 was used, although the results were similar for the rest of alphas.

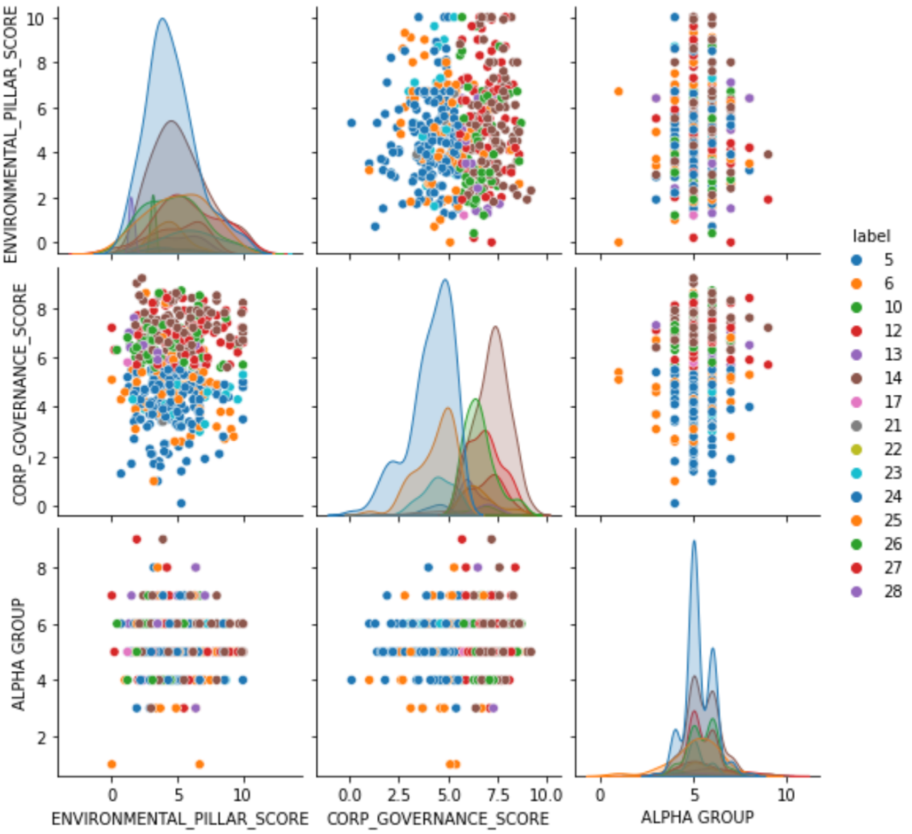


Figure 15. Hierarchical clustering plots of Alpha against Governance Score and Environmental Score.

These clusters, unfortunately, do not provide any insight on the relationship between ESG and alpha. The clusters are easily identifiable, in comparison with K-means, but they do not offer the same conclusion.

4.6 Regression

In this section, the main regression techniques will be described. The aim of this project is not to explain in depth the mathematics behind these methods. However, some understanding of how they operate may be necessary to understand the results and so, some brief explanation will be given.

These regression techniques will try to model the dependent variable as a response of the explanatory variables. The alphas are our dependent variables and the ESG Ratings our independent ones. The model will then try to make predictions of the response with these ESG Ratings.

Because of this, a core concept of regression analysis becomes apparent: dividing the dataset in training and test subsets. The idea emerges from the need to train the model and test it with different data. If we both train and test the regression on exactly the same companies, we are at risk of overfitting the data.

Overfitting is a regression error that occurs when the model is tailored to “fit the quirks and random noise of your specific sample, rather than reflecting the overall population” []. An example of this can be seen below [11].

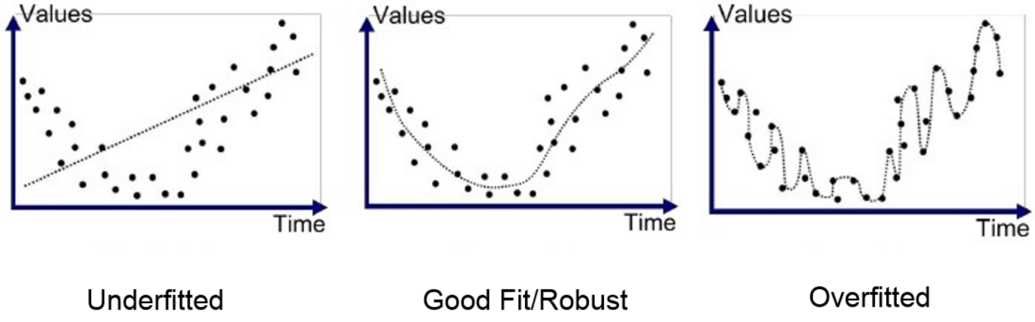


Figure 16. Underfitted vs robust vs overfitted.

Therefore, the dataset is divided in a subset containing 75% of the data (training) and in another, with the 25% remaning (test). We will then try out and train different models on the training dataset and test it on the test dataset. The percentages of 75-25 can be different but are commonly used, since usually we want our training subset to be larger than the test subset.

One of the parameters that this thesis focuses on is the R-Squared and Adjusted R-Squared. The R-Squared is a statistical measure that shows the proportion of the variance of the response explained by the independent variables. In other words, it represents how much of our variable alpha is explained by looking at the ESG ratings. If we do not overfit the data, the higher this percentage, the more satisfied we can be with our model.

The adjusted R-Squared is similar to R-Squared, but it also includes a penalization for the number of variables used in our model. Looking at this metric is standard practice in regression, since otherwise you would be able to add hundreds of variables of high order polynomials to force fit the data.

It is important to note that on the financial field, a high R-Squared is very rarely achieved. Alpha, stock price, 1-year return and numerous other metrics depend on a lot of factors, some of which are impossible to predict. Global events like a financial crisis or local occurrences such as a competitor acquiring a company can have great influence. Human psychology and market sentiment can also play a big role. If we were able to explain exactly and pinpoint all the reasons as to why the companies have their particular alpha value, we would be able to predict very precisely the stock market, which is not possible to an extent. Not only that, but in this particular thesis we are not looking at financial indicators like the ROE ratio, that could help us explain much more variance. As a result, our R-Squared and Adjusted R-Squared will be low, but as explained, it is expected.

4.6.1 Linear regression

As the name suggests, linear regression will try to model the dependent variable as a linear response of the explanatory variables. An example of linear regression can be seen below. The blue dots represent actual observations of the data and the red line shows the linear regression model.

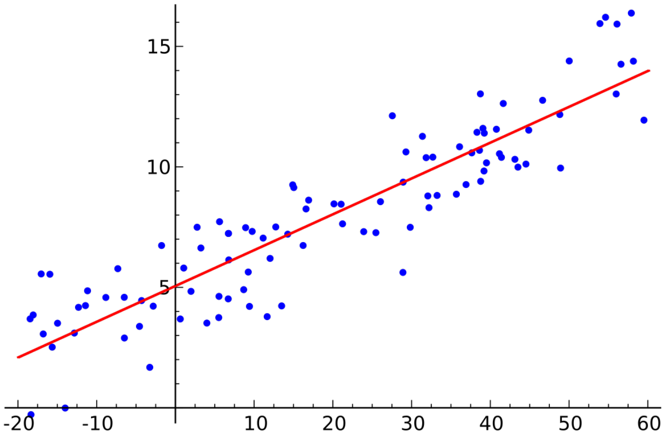


Figure 17. Example of linear regression [12].

The main structure for a linear regression can be viewed in the following equation, where Y is the response variable, X_i are the explanatory variables and β_i are the coefficients of each of these variables, with the exception of β_0, the intercept or constant. The parameter ϵ represents the noise in the data.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_0 X_3 + \cdots + \epsilon$$

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_0 X_3 + \cdots$$

As a starting point for our regression analysis, a simple linear regression is conducted. This model is not intended to be the final model, but rather a benchmark to which to compare the rest. Having this initial model allows us to see the improvement made further down the line.

A recurrent problem in linear regression is choosing the number of predictors to use in the model and which ones. Intuitively, it might be thought that the higher the number of predictors, the better results are obtained. This is not true, and one of the main reasons is that it can lead to a very overfitted model. As well, it takes more computational power, and the interpretation of the results is much worse. As a result, an automated process has been done, with 15 predictors to chose from. The operation starts with 1 predictor and tries fitting a model, each time with a different variable. It then selects the best one. In the next step two predictors are used, and so all the combinations of pairs of the 15 variables are tried. The process continues until arriving to the final best model. This process can be seen in the figure below.

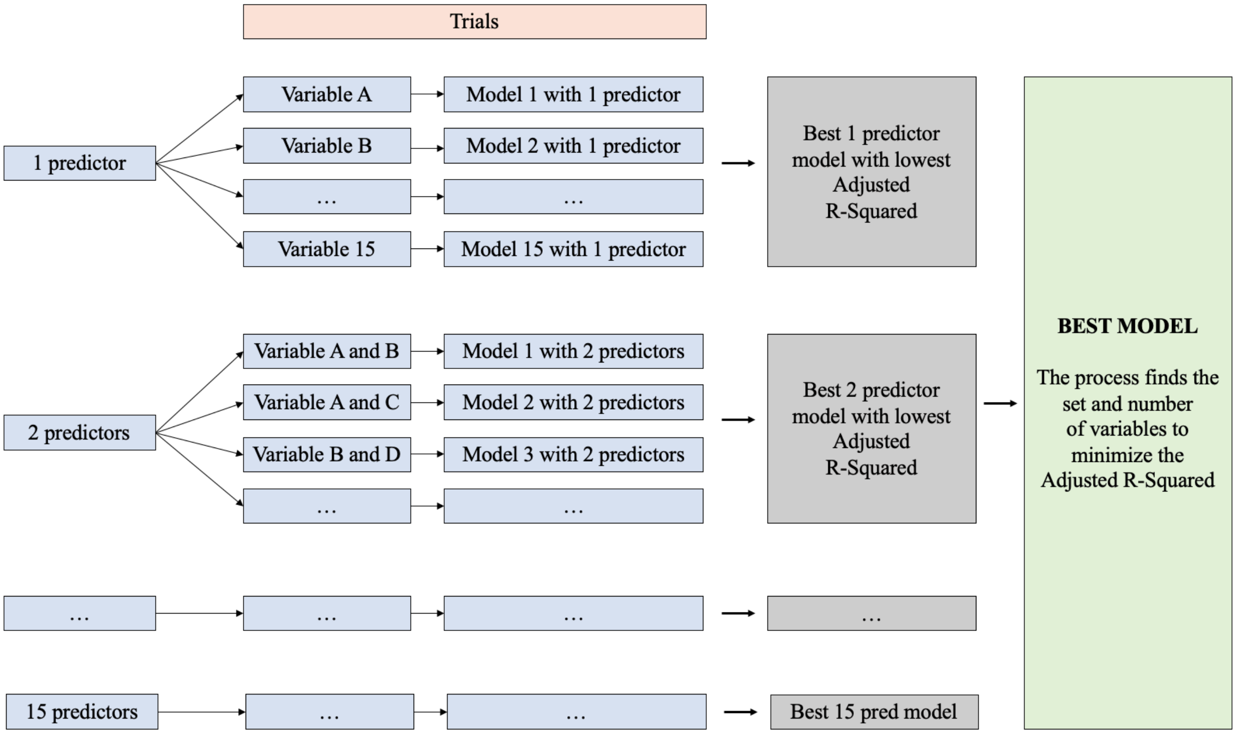
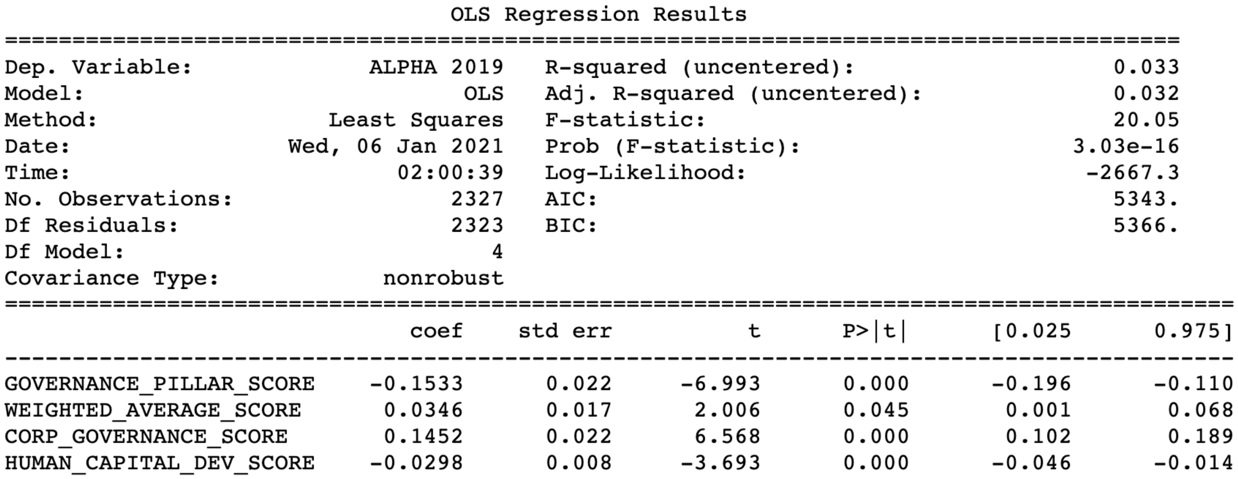

Trials
1 predictor
Variable A
Model 1 with 1 predictor
Variable B
Model 2 with 1 predictor
…
…
Variable 15
Model 15 with 1 predictor
Best 1 predictor model with lowest Adjusted R-Squared
2 predictors
Variable A and B
Model 1 with 2 predictors
Variable A and C
Model 2 with 2 predictors
Variable B and D
Model 3 with 2 predictors
…
…
Best 2 predictor model with lowest Adjusted R-Squared
…
…
…
…
15 predictors
…
…
Best 15 pred model
BEST MODEL
The process finds the set and number of variables to minimize the Adjusted R-Squared

Figure 18. Linear Regression choosing best set and best number of predictors.

The 15 predictors chosen to iterate through are the overall ESG Rating, followed by the three ESG subcategories and 11 key metrics. The best model results in one that uses only 4 predictors: Governance Pillar Score, Weighted Average Score, Corporate Governance Score and Human Development Score.

To clarify some of the undefined variables: Human Development Score is a key metric that evaluates "companies' ability to attract, retain and develop human capital based on their provision of benefits, training and development programs, and employee engagement". The Corporate Governance Score is "the extent to which companies' corporate governance practices in specific governance areas – audit, board, compensation/remuneration, shareholder rights -- pose financial risks to shareholders”.

The model obtained is the following:



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:             ALPHA 2019   R-squared (uncentered):                   0.033
Model:                            OLS   Adj. R-squared (uncentered):              0.032
Method:                 Least Squares   F-statistic:                              20.05
Date:                Wed, 06 Jan 2021   Prob (F-statistic):                    3.03e-16
Time:                        02:00:39   Log-Likelihood:                         -2667.3
No. Observations:                2327   AIC:                                      5343.
Df Residuals:                    2323   BIC:                                      5366.
Df Model:                           4
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025       0.975]
--------------------------------------------------------------------------------------------
GOVERNANCE_PILLAR_SCORE    -0.1533      0.022     -6.993      0.000      -0.196       -0.110
WEIGHTED_AVERAGE_SCORE      0.0346      0.017      2.006      0.045       0.001        0.068
CORP_GOVERNANCE_SCORE       0.1452      0.022      6.568      0.000       0.102        0.189
HUMAN_CAPITAL_DEV_SCORE    -0.0298      0.008     -3.693      0.000      -0.046       -0.014
```

Figure 19. Initial Linear Regression Model.

As we can observe, all four variables are significant, since their p-value is under 5%. However, the Adjusted R-Squared is 3.2%, which is very low but improved in later models. An analysis of the coefficients of each variable will be done with later linear models, since this one is just a benchmark for the rest.

4.6.2 PLS (Partial Least Squares)

PLS, or Partial Least Squares is a statistical method similar to PCA (Principal Component Analysis), in that it builds a smaller and new dataset in a lower dimensional space, reducing the number of features. Each of these new features is a linear combination of the original predictors.

PCA looks for features that are as informative as possible, that is, new features that capture as much of the variability of the original predictors as possible. However, PCA disregards the response variable in the reduction of dimensions. This is, in fact, the main difference between PCA and PLS and the reason why we use PLS rather than PCA in this thesis. Partial Least Squares makes use of the response in order to identify key features, which makes it more suited for our goal.

Partial Least Squares is a powerful tool in statistics, that can be used for both clustering and regression. In this particular case, we will be looking at PLS regression, since we are looking for some improvemen over our linear regression. PLS can bring to the analysis the following advantages over other techniques:

- PLS can help visualize in a graph the correlations among the predictors when reducing the number of variables to two.
- PLS allows us to keep track of less data without losing much analytic power.
- As we have seen in the previous model, it can be difficult and time consuming to conduct linear regression with so many explanatory variables. Because the technique helps us understand the correlations among predictors, it derives that it can be used to search for the variables that we could keep for our linear regression.

One of the biggest disadvantages of the method is the loss of interpretability. In our linear regression model, we could identify the influence that each predictor had on the response. This can be very useful to, for example, suggest companies to improve their environmental footprint, if a strong positive correlation is found. However, when PLS is performed, the features for which the relationship is found are Principal Component 1, Principal Component 2, etc. We might know there is a strong correlation to the response, but it is difficult to pinpoint which of the original factors have positive/negative correlations or which of them have the biggest impact on the dependent variable.

In a similar fashion to linear regression, the number of principal components needs be selected. To do this, the error (MSE, Mean Squared Error) will be plotted against the number of components.

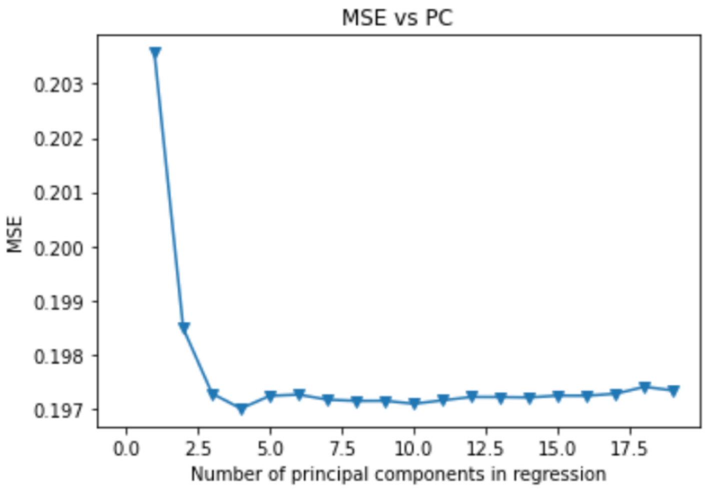


Figure 20. PLS: Number of principal components against the Mean Squared Error.

The lowest error is obtained with 4 principal components. The model is then fitted, obtaining an Adjusted R-Squared of 11.5%, much higher than the initial 3.2% of linear regression.

As previously explained, the principal components are composed from the original features of the dataset. However, not all of these features have the same impact on the principal components, since they are constructed to maximize the variability in the response and in the predictors. Therefore, the original predictors that have the highest impact can be found. This can be interesting for our linear regression model, due to the fact that the 15 variables used in the first linear model were selected by intuition. Plotting the loading vectors with two principal components is a method to visually find these features.

Since the plot contains all +120 variables, the size of the plot must be very large and, as a result, only a part of the plot is shown below (*Figure 21*).

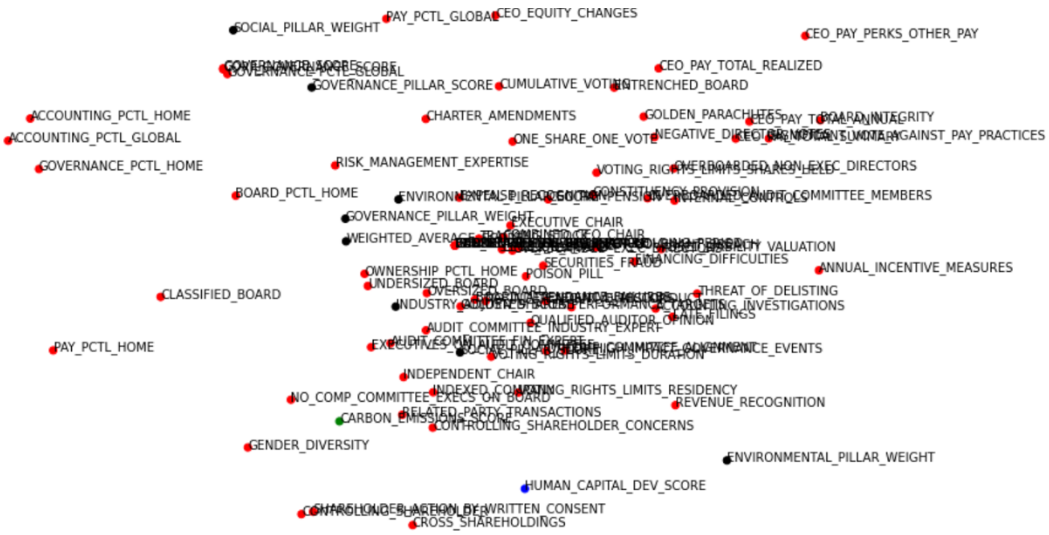


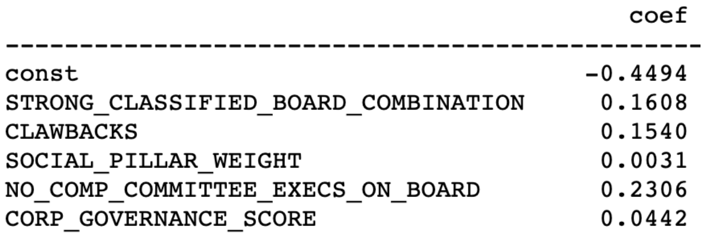
Figure 21. PLS Loading Vectors visualization.

The features furthest away from the center are the ones that contribute the most in PLS, and therefore, are the ones that were initially used in the improved linear model. Features such as Human Capital Development, Social Pillar Score, Environmental Pillar Score do not come as a surprise, since two of them were already used in the first model and the Social Pillar Score is one of the main ESG subcategories. Some other of these features are:

- Cross Shareholdings: “Is the company involved in a series of cross-shareholdings with other (related or unrelated) companies? Flagged if yes.”
- Accounting PCTL Global: “Company's Global Percentile Rank for Accounting”
- Strong Classified Board Combination: “Does the company have a classified or staggered board in combination with other limitations on shareholder rights that further limit shareholder ability to impact the make-up of the board? Flagged if yes”.
- Clawbacks: “Has the company adopted a clawback policy that would recoup incentive compensation based on accounts that were restated at a later date? Flagged if no”.
- No competency Committee Executives on Board: “Does the company lack a standing compensation committee, and also have current company executives serving on its board? Flagged if yes.”

It is worth noting that these features do not necessarily give the best possible linear regression model, since they have been chosen not only to explain variability in the response but also variability in the features.

The result from using these predictors is a similar Adjusted R-Squared to the initial model, with 5 variables being extremely significant (p-values lower than 1%).



```
                                              coef
--------------------------------------------------
const                                      -0.4494
STRONG_CLASSIFIED_BOARD_COMBINATION         0.1608
CLAWBACKS                                   0.1540
SOCIAL_PILLAR_WEIGHT                        0.0031
NO_COMP_COMMITTEE_EXECS_ON_BOARD            0.2306
CORP_GOVERNANCE_SCORE                       0.0442
```

Figure 22. Coefficients of Linear Regression Model after PLS.

It can be observed a positive correlation between alpha and the Social Score, as well as between alpha and the Corporate Governance Score.

4.6.3 Random forests

Random Forests is a method similar in essence to Decision Trees, described in the clustering section. The main difference is that Random Forests is a collection of trees and, therefore, it tends to outperform decision trees. This technique is not too sensitive to the dataset and does not fall as easily as Decision Trees in overfitting. It must be noted that this method is used in this thesis for regression, although it could have also been used for clustering.

Random Forests is a good method to reliably estimate the importance of features, although it is much slower than other techniques. Much like PLS, some interpretability is also lost.

The Adjusted R-Squared is 15%, which represents an improvement from all other previous models. The importance of the features can be seen below.

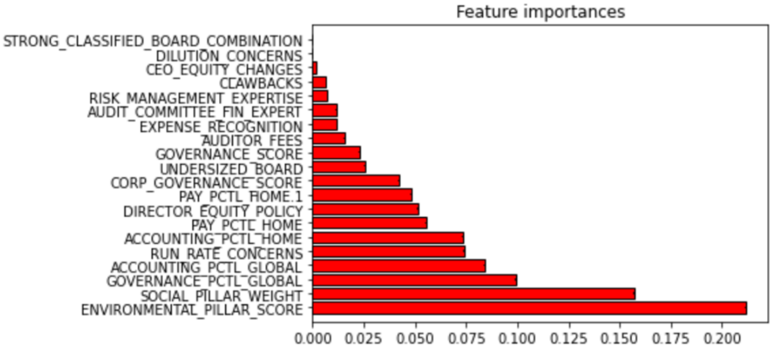


Figure 23. Feature importance using Random Forests.

Knowing which features are the most important ones for alpha is very useful for our linear model, which will be described at the end of the section.

4.6.4 Gradient Boosting

Gradient Boosting is a method that has become very popular in recent years. Kaggle, a platform focused on data science that organizes competitions with free to use datasets, is one of the main drivers in this popularization [13]. In the past three years, most of the competitions organized in the Kaggle community have been won by a team using gradient boosting as the primary method.

This technique uses simple models to create a composite one, that usually outperforms the rest. These simpler models are known as weak predictors and, in the case of gradient boosting, they are decision trees.[14]

The method starts in its first iteration with a decision tree, fitting the data to it. It then calculates the error, the difference between the predictions made by the model and the actual observations. In its second iteration, gradient boosting adds another decision tree that tries to minimize these errors, also known as residuals. As more iterations are performed, more weak learners that minimize the previous residuals are added to the method, becoming a composite model of the rest.

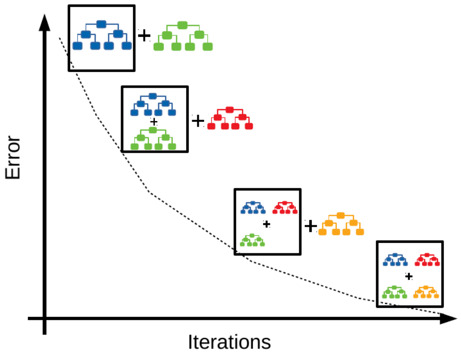


Figure 24. Gradient Boosting method.

After fitting the data to a gradient boosting model, a lower Adjusted R-Squared can be observed, in comparison to random forests or PLS.

4.6.5 Polynomial regression

As the name suggests, polynomial regression is a technique that fits the data with a polynomial equation. The explanatory variables could be raised to the power of two, three or more, depending on which power makes the model perform better. Not only that, but one variable could have multiple terms, each of them raised to a different power. The basic equation for polynomial regression can be seen below.

$$Y \sim \beta_0 + \beta_{11}X_1 + \beta_{12}X_1^2 + \cdots + \beta_{pd}X_p^d$$

Linear Regression is in essence a polynomial regression of first degree, where all variables present in the model are raised to the power of one. Therefore, polynomial regression could potentially add more prediction capabilities. A comparison of some polynomial models of different degrees can be found below.

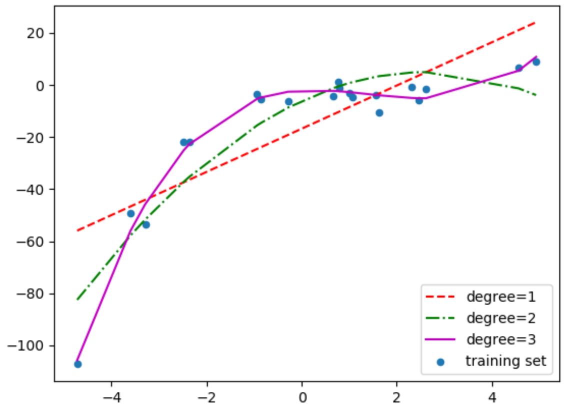


Figure 25. Polynomial regression example.

An important issue with polynomial regression needs to be addressed. In the image above, we can see that the third-degree polynomial is the one that fits the data the best. However, this model may be overfitting the data to accommodate outliers or the noise present in the observations. Because of this, it is imperative to divide the dataset in training and test subsets, as described in the beginning of this chapter. The risk of overfitting is then reduced considerably.

After fitting a polynomial model to the data, we obtain a negative Adjusted R-Squared. As explained at the beginning of this section, this parameter represents the percentage of variance explained hence it is impossible, conceptually, to obtain a negative value. However, the reality is that the Adjusted R-Squared can have in fact negative values, although it may be counter intuitive. Negative values just prove that the model is not suited to fit this data. Furthermore, it means that it is a worse model than a horizontal line.

Since the polynomial regression is not a good approach for this dataset, no further improvements are tried with this method.

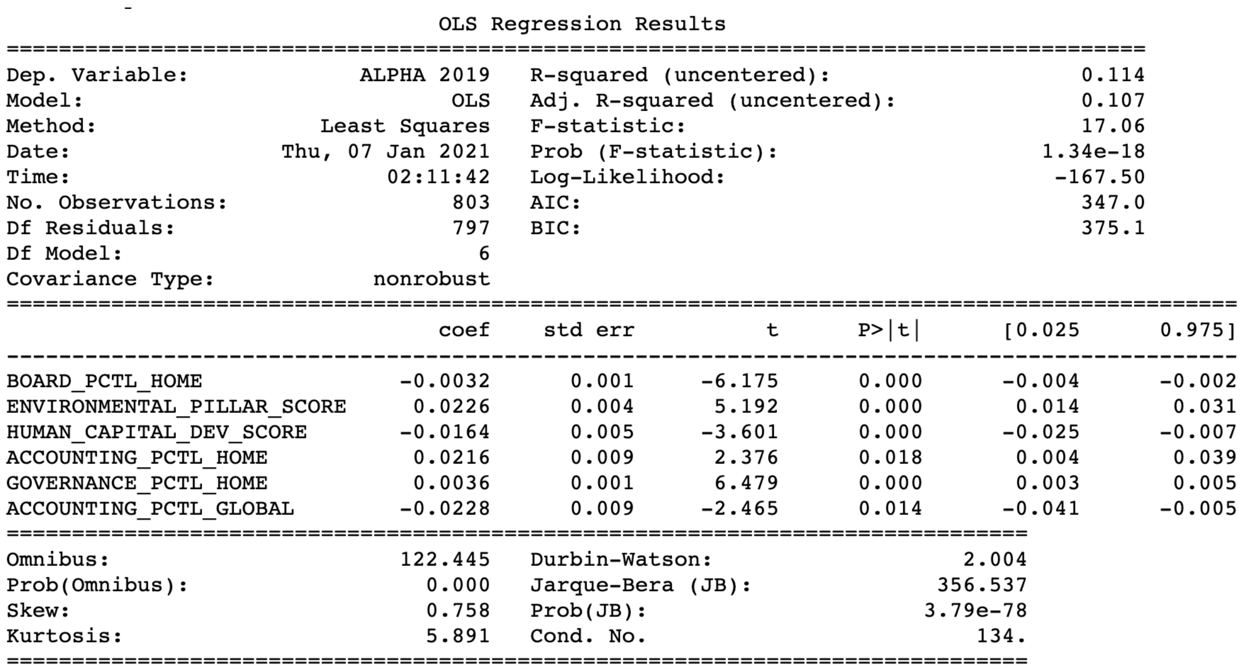
4.6.6 Linear regression with Random Forests features' importance and filtering

It would be reasonable to think that ESG impacts differently depending on the geographical area where the company operates. As well, there might be trends in a particular industry that do not apply to the rest. Because of this, companies have been filtered based on country, industry or both, using the MSCI sector list [15]. Not only that but having found the most important features through Random Forests, they can be used in a linear regression model. This approach gives the best results achieved so far, raising the Adjusted R-Squared to 10% in several countries and to 20% in some industries.

It is worth mentioning that when the companies are filtered by both country and industry the remaining number of companies is very low in most cases. Trying to fit a model would lead to a very strong overfitting, since the number of predictors is close to the number of observations. As a result, the companies have been filtered separately by industry or by country, not both at the same time.

As well, some countries do not have enough companies in the dataset to perform some meaningful regression, like Spain for example. Therefore, these countries have been grouped up by continent. As such, the geographical areas for the filtering of companies are not countries or continents, but a mix. For example, Spain has been put in the Europe group, but countries like United Kingdom can be studied as well by themselves since its universe of companies is bigger.

An example of one of these models, where only Japanese companies are taken, is shown below:



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              ALPHA 2019   R-squared (uncentered):                   0.114
Model:                             OLS   Adj. R-squared (uncentered):              0.107
Method:                  Least Squares   F-statistic:                              17.06
Date:                 Thu, 07 Jan 2021   Prob (F-statistic):                    1.34e-18
Time:                         02:11:42   Log-Likelihood:                         -167.50
No. Observations:                  803   AIC:                                      347.0
Df Residuals:                      797   BIC:                                      375.1
Df Model:                            6                                                  
Covariance Type:             nonrobust                                                  
===============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
BOARD_PCTL_HOME               -0.0032      0.001     -6.175      0.000      -0.004      -0.002
ENVIRONMENTAL_PILLAR_SCORE     0.0226      0.004      5.192      0.000       0.014       0.031
HUMAN_CAPITAL_DEV_SCORE       -0.0164      0.005     -3.601      0.000      -0.025      -0.007
ACCOUNTING_PCTL_HOME           0.0216      0.009      2.376      0.018       0.004       0.039
GOVERNANCE_PCTL_HOME           0.0036      0.001      6.479      0.000       0.003       0.005
ACCOUNTING_PCTL_GLOBAL        -0.0228      0.009     -2.465      0.014      -0.041      -0.005
==============================================================================
Omnibus:                      122.445   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              356.537
Skew:                           0.758   Prob(JB):                     3.79e-78
Kurtosis:                       5.891   Cond. No.                         134.
==============================================================================
```

Figure 26. Japanese companies Linear Regression Model after Random Forests.

Chapter 5: RESULTS

In this section, the main results of the models are discussed. Two clustering and five regression techniques have been used, each with their advantages and limitations, that will be summarized.

5.1 Clustering

Both Hierarchical and K-means Clustering show a decent clustering of the data. However, Hierarchical does not provide any insight, while K-means does.

K-means show the existence of a relationship between the Governance Score and alpha. Observations belonging to clusters with low alphas, were also scoring low in the Governance Subcategory. Nevertheless, this correlation was very mild.

5.2 Regression

The five regression methods that were tried are: Linear Regression, Partial Least Squares, Random Forests, Gradient Boosting and Polynomial Regression. In addition to these techniques, the companies were filtered based on country or industry. The results of these approaches are the following.

1. The initial linear regression serves as a benchmark to which to compare other models. Fifteen predictors, chosen by studying other articles that focus on this subject, were used in an iterative manner to obtain the best possible linear model. The result is a very low Adjusted R-Squared, as expected. This model does not provide any particular insight on the data, since the correlation found is extremely low.

2. Partial Least Squares does not only provide a better model, but it also shows the features that capture the variability of the predictors and response the most. The disadvantage of this model

is the loss of interpretability, since the influence of each predictor cannot be identified. The result was a considerable improvement in comparison with the initial linear model.

3. The technique Random Forests involves the use of a collection of trees to make the regression. This method does not only provide a better model, in comparison to the initial linear regression, but also a list of the most important features to explain alpha. Random Forests, much like PLS, suffers as well from a loss of interpretability. The result of this regression was an improvement from both the initial model and PLS.

4. The technique Gradient Boosting performs an iterative process, creating a strong learner from weak learners (trees). This model, as well, does not provide the means to identify the contribution of each predictor. The results of the model are similar to the initial linear regression model, worse than PLS and Random Forests.

5. Polynomial Regression could provide some insight in the influence of each variable on the response. However, the Adjusted R-Squared was negative, which shows that this type of regression is not suited for the dataset and therefore, there are no meaningful conclusions that can be extracted from this method.

6. The final model is a Linear Regression using the most important features obtained from Random Forests. As well, the companies were filtered based on geographical area or industry. The reason why the companies were not filtered on both geographical area and industry at the same time is that the resulting number of companies would be too low to run reasonable regression techniques. This model is the most successful, improving over the rest.

A summary of the metrics of each model can be seen in the table below:

	Initial Linear Regression	PLS Regression	Random Forests	Gradient Boosting	Polynomial Regression	Final Linear Regression
Adjusted R-Squared	3.2%	9.1%	12.1%	4%	Negative	Avg 14%
P – value Coefficients	Yes	No	No	No	Yes	Yes
Overall quality	5th	3rd	2nd	4th	6th	1st

Table 3. Summary of regression models.

From Random Forests, the most important variables that affect alpha have been found. Nevertheless, depending on the geographical region or industry these metrics do not have the same level of significance. The details will be discussed further down this section, but these are, overall, the most relevant variables:

- Environmental Pillar Score: One of the main three ESG subcategories that, among other evaluations, assesses the carbon footprint, energy efficiency and climate change vulnerability. It summarizes the company’s approach regarding the environment.

- Social Pillar Score: Another main ESG subcategory that, among other factors, judges the human capital development, labor standards and product safety. It summarizes the company’s approach regarding social issues.

- Board Percentile Home: Company’s percentile ranking in their home market regarding their board. The Accounting and Governance Percentile in their home market are also important. These three factors are significant, but they do not have a big impact on alpha, in comparison with other variables.

- No Compensation Committee Executives on Board: As the name suggests, this variable is marked as a 1 if the company lacks a standing compensation committee and current company executives on the board.

- Human Capital Development Score: Score based on a company's ability to attract, retain and develop human capital.

- CEO Pay Total Summary: Flagged with a 1 if the total pay of the CEO falls into an extreme in comparison to the company's peers.

- Overboarded Executive Directors: Marked with a 1 if the company's executive board members serve on the board of more than one public companies.

- Corporate Governance Score: Although it may seem similar to the Governance Pillar Score, this item evaluates whether governance issues pose a risk to the company's shareholders.

- Business Combination Provision: Marked as a 1 if the company has a business provision in place or is subject to business combination protection.

- Run Rate Concerns: Evaluates whether the company's run rate is contrary to shareholders interests.

- Internal Pay Equity: Whether or not the CEO's pay exceeds three times the median pay of the rest of board members.

- Pay Performance Links: If the CEO's equity pay reflects the company's share price movement over the past 5 years, this variable is marked with a zero. Otherwise, it is flagged with a one.

It is important to notice the absence of the Governance Pillar Score and the overall ESG Rating, which implies that they don't have a strong correlation with alpha. However, a lot of the factors described are metrics used in computing the overall Governance Score and the overall ESG Rating, and therefore, those factors are also present in a way.

Having explained the meaning of these variables, we can look at the results from the improved linear regression, since it is the model that fits the data the best. Firstly, we will examine the companies divided by geographical area and afterwards, divided by industry. The tables below summarize with which regions the correlation was stronger or weaker and which variables have an impact in alpha. Whether or not these factors have a positive or negative relationship can also be viewed.

<table>
<tr><td rowspan="2">Correlation
(variability explained)</td><td colspan="2">🟥 Weak</td><td colspan="2">🟨 Medium</td><td colspan="2">🟩 Strong</td></tr>
<tr><td colspan="2">0-5%</td><td colspan="2">5.1-12.9%</td><td colspan="2">13%-20%</td></tr>
<tr><td rowspan="2">Variables</td><td colspan="3">🟥 Negative relation</td><td colspan="3">🟩 Positive relation</td></tr>
<tr><td colspan="6">The 3 variables with the highest impact are marked.</td></tr>
</table>

Table 4. Legend for tables of results.

It is important to note that even though a classification as “strong correlation” is given to a 13-22% of variance explained, this is still quite low if compared to other data science studies done on different fields. The reason it is classified as such is because in the finance world, the performance of companies is difficult to analyze and a company’s stock price depend on hardly predictable factors, such as worldwide events or human psychology. As a result, having an explained variability of 20% signifies an important relationship to consider.

	Africa	Europe	United Kingdom	United States	Canada	China	Japan	South America	Oceania
Overall Correlation	🟥 Weak	🟥 Weak	🟨	🟨	🟩 Strong	🟨	🟨	🟩 Strong	🟨
Environmental Score		🟩		🟩			🟩 1st		🟥
Social Score		🟩	🟩 3rd	🟩		🟩 2nd		🟥 2nd	
Board Percentile Home							🟥		
Accounting Percentile Home							🟩 2nd	🟩	
Governance Percentile Score				🟥			🟩		
Accounting Percentile Global							🟥 3rd		
No Compensation Committee						🟩 1st		🟩 1st	🟩 1st
Human Capital Development Score		🟥 3rd	🟥	🟥			🟥		🟥 3rd
CEO Pay Total		🟥 2nd							
Overboarded Executive Directors					🟩 3rd				
Corporate Governance Score				🟩 3rd				🟩 3rd	
Business Combination Provision					🟩 2nd				🟩 2nd
Run Rate Concerns			🟥 2nd	🟩 2nd					
Accounting Percentile Global			🟩	🟩					
Pay Performance Links		🟥 1st	🟥 1st	🟥 1st	🟥 1st				

Table 5. Results from Linear Regression divided by region.

The results show several insights:

- Africa and Europe show an extremely low variance explained, almost none. Not only that, but Africa does not have any ESG variable which is statistically significant and so we can determine that there is no correlation in that region.

- The strongest correlated regions are Canada and South America with a 17% and 21% of variance explained, respectively.

- Both the Environmental Pillar Score and Social Pillar Score affect most regions and, overall, the relationship is positive.

- The variable Internal Pay Equity (whether or not the CEO's pay exceeds three times the median pay of board members) was eliminated from the table, since it was not statistically significant in any case.

- Whether or not the CEO's equity pay reflects the company's share price movement over the past 5 years (Pay Performance Links) is one of the highest correlated variables and in all the cases where the variable was significant the relationship was negative. This means that having a high alpha is correlated to the CEO's equity pay being related to the company's performance. There are multiple articles exploring this idea and the overall consensus is that rewarding the CEO when the company overperforms or punishing the equity pay if it underperforms is good for a company [16]. This is in line with the result obtained. However, it is surprising that a company not having a compensation committee (No Compensation Committee factor) correlated positively with alpha in the regions of China, South America and Oceania [17].

- One surprising result is the relationship with the Human Capital Development factor, present in a lot of regions and always negative. Companies that actively invest in training programs, give benefit packages and have good employee engagement score high on this metric. Companies that rely heavily on high-skilled employees but do not show evidence of such employee engagement score poorly. Therefore, it is surprising that companies that invest in employee engagement are not rewarded for it.

As for the companies divided by industry, the following results are obtained:

	Industrials	**Financials**	**Consumer Discretion.**	**Consumer Staples**	**Health Care**	**Materials**	**Energy**	**Information Technology**	**Telecomm. Services**	**Utilities**
Overall Correlation	Weak	Weak	Weak	Weak	🟨	🟨	Strong	Weak	Strong	Strong
Environmental Score		3rd				🟥				2nd
Social Score	2nd								1st	
Accounting Percentile Global	3rd			🟩	🟩					
No Compensation Committee		3rd						1st		
Human Capital Development Score					🟥				2nd	
CEO Pay Total						1st				
Overboarded Executive Directors				3rd	2nd					1st
Corporate Governance Score						🟩		2nd		
Business Combination Provision				🟩						
Run Rate Concerns		2nd		2nd	3rd					
Internal Pay Equity						3rd	2nd			
Pay Performance Links	1st	1st	1st	1st	1st	1st	1st		1st	

Table 6. Table of results of Linear Regression with companies divided by industry.

The insights obtained from these results are the following:

- The ESG factors Board Percentile and Accounting Percentile in home market, as well as Governance Percentile globally were eliminated from the table of results, since neither of them correlated with any industry.

- The Social and Environmental Score are not present in most of the industries. However, for the Utilities sector there is a positive relationship, and it is quite impactful.

- The CEO's Pay Performance Links is again highly correlated and negative in all cases. This time, Internal Pay Equity (CEO earning 3 times more than the median of the rest of executives) is also significant in the Materials and Energy sectors, with a negative relationship.

- The Energy, Telecommunication Services and Utilities sectors are the highest correlated sectors, with Industrials, Financials, Consumer Discretionary and Consumer Staples being the lowest. It is worth noting that in the case of Consumer Discretionary, only one variable is relevant: Pay Performance Links.

Chapter 6: CONCLUSIONS

The objective of the project is to find a correlation between alpha and the ESG Ratings of companies, if it exists. In order to do so, several regression and clustering techniques have been performed, from which Linear Regression has been shown to be the best approach, since it fits the data the best.

We can determine that a correlation does exist. However, this correlation is weak, explaining, at most, a 22% of the variance in some cases. It is extremely low when taking all companies in the dataset, and higher when filtering by region or industry.

As a result, ESG Ratings should not be used in isolation to make investment decisions. As it is usual in the industry, several quantitative and qualitative assessments should be made before deciding to invest in a particular stock. ESG Ratings should be used as one of these metrics from which a final judgement can be made [17][18].

A topic to discuss with respect to this correlation is causation. A common misconception when studying data is that correlation implies causation. This is not true and the fact that a correlation between ESG Ratings and alpha exists, does not imply that a company's alpha is due to its ESG, but rather that these occurrences, to a degree, happen together. However, the point is still valid: ESG correlates to alpha.

Even though the correlation exists, a relationship with the overall ESG Rating was not found, but rather with some of its underlying categories. The most notable factors that are related to alpha are the Environmental Score and Social Score, which correlate positively overall. As well, low alphas tend to correspond to companies where the CEO's equity is not related to the company's performance. The Human Capital Development Score also has, generally, a negative relationship.

These relationships depend on the geographical region and industry. The regions with the highest correlations are Canada and South America, followed by the United States, Oceania, Japan, United Kingdom and China. At the bottom end we find Europe and Africa.

Regarding the classification in industries, the Energy, Telecommunication Services and Utilities sectors are the highest correlated. The lowest are Industrials, Financials, Consumer Discretionary and Consumer Staples. Falling in the middle, the Health Care and the Materials sectors are present.

The volatility of these companies in the year 2020 was also contrasted against alpha, showing a high correlation. The regression model shows that companies that violate its debt covenants have higher volatility in the last 360 days. The same happens when a CEOs pay is not related to its performance. Finally, companies with lower the carbon emissions have had lower volatility. We can determine that, although none of the three pillar scores are correlated, the volatility is related strongly to some of their underlying factors.

Chapter 7: BIBLIOGRAPHY

[1] "Sustainable Investing: Investing for the Long-Term | Harvard Management Company." Harvard Management Company, 20 Nov. 2020, www.hmc.harvard.edu/sustainable-investing/.

[2] "A Friedman Doctrine-- The Social Responsibility Of Business Is to Increase Its Profits (Published 1970)." The New York Times, 2021, www.nytimes.com/1970/09/13/archives/a-friedman-doctrine-the-social-responsibility-of-business-is-to.html.

[3] Gompers, P. A., Ishii, J. L., & Metrick, A. (2001). Corporate Governance and Equity Prices. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.278920

[4] Lopez de Silanes, F., McCahery, J. A., & Pudschedl, P. (2019). ESG Performance and Disclosure: A Cross-Country Analysis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3506084.

[5] How do ESG scores relate to financial returns? | Refinitiv Perspectives. (2020, August 26). Refinitiv Perspectives. https://www.refinitiv.com/perspectives/market-insights/how-do-esg-scores-relate-to-financial-returns/.

[6] "ESG Score VS Financial Performance." @Refinitiv, 2020, www.refinitiv.com/en/resources/white-paper/esg-score-vs-financial-performance.

[7] Lester, Anna, and Chen He. Harnessing ESG as an Alpha Source in Active Quantitative Equities.

[8] Alpha Vs. Beta: What's the Difference? (2021). Investopedia. https://www.investopedia.com/ask/answers/102714/whats-difference-between-alpha-and-beta.asp

[9] Ashwin Kumar, N. C., et al. "ESG Factors and Risk-Adjusted Performance: A New Quantitative Model." Journal of Sustainable Finance & Investment, vol. 6, no. 4, Oct. 2016, pp. 292–300, 10.1080/20430795.2016.1234909.

[10] Abhinav Choudhary. "K Means Clustering with Python." DataScience+, 25 Sept. 2019, datascienceplus.com/k-means-clustering/.

[11] Anup Bhande. "What Is Underfitting and Overfitting in Machine Learning and How to Deal with It." Medium, GreyAtom, 11 Mar. 2018, medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76.

[12] Wikipedia Contributors. “Linear Regression.” Wikipedia, Wikimedia Foundation, 17 Dec. 2020, en.wikipedia.org/wiki/Linear_regression.

[13] kashnitsky. “Topic 10. Gradient Boosting.” Kaggle.com, Kaggle, July 2020, www.kaggle.com/kashnitsky/topic-10-gradient-boosting.

[14] “Block-Distributed Gradient Boosted Trees.” Tvas.Me, 2019, tvas.me/articles/ 2019/08/26/Block-Distributed-Gradient-Boosted-Trees.html.

[15] “MSCI.” Msci.com, 2021, www.msci.com/gics.

[16] “CEO Incentives—It’s Not How Much You Pay, But How.” Harvard Business Review, May 1990, hbr.org/1990/05/ceo-incentives-its-not-how-much-you-pay-but-how.

[17] “The Role of the Compensation Committee | Diligent.” Diligent Insights, 22 May 2018, insights.diligent.com/compensation-committee/the-role-of-the-compensation-committee.

[18] Morningstar, Inc. “How Does Investing in ESG Companies Affect Returns?” Morningstar, Inc., 19 Feb. 2020, www.morningstar.com/insights/2020/02/19/esg-companies.

[19] Sullivan, Kristen, et al. “Advancing ESG Investing: A Holistic Approach for Investment Management Firms.” The Harvard Law School Forum on Corporate Governance, The Harvard Law School Forum on Corporate Governance, 11 Mar. 2020, corpgov.law.harvard.edu/ 2020/03/11/advancing-esg-investing-a-holistic-approach-for-investment-management-firms/.

APPENDIX

LIBRARIES USED:

```
from sys import stdout #System specific parameters and functions
import numpy as np # Arrays, linear algebra
import pandas as pd # Dataframe usage and capabilities

#Plotting
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
from scipy.signal import savgol_filter

import seaborn as sns #k-means visualization

import graphviz
from sklearn.tree import export_graphviz
import pydot

# Statistical library for regression and clustering techniques
from scipy import stats
from sklearn.cross_decomposition import PLSRegression
from sklearn.model_selection import cross_val_predict
from sklearn import metrics
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import scale #For scaling the data
from sklearn import model_selection
from sklearn.ensemble import RandomForestRegressor #For Random Forests
Regression
from sklearn.ensemble import GradientBoostingRegressor #For Gradient Boosting
from sklearn.model_selection import train_test_split #For dividing data in
training and test
from sklearn.preprocessing import StandardScaler #For scaling the data
from sklearn.preprocessing import PolynomialFeatures #For polynomial
regression
from sklearn.linear_model import LinearRegression #For polynomial regression

from sklearn.tree import DecisionTreeClassifier #For hierarchical clustering
from sklearn.metrics import classification_report, confusion_matrix #For
hierarchical clustering
from sklearn import tree #For hierarchical clustering

import statsmodels.api as sm #To perform regression models
from statsmodels.api import add_constant

import xgboost as xgb #Gradient Boosting

from sklearn.cluster import KMeans #K-means clustering

# Time display and other capabilities
import time
import itertools
```

# Dataset:

```
file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
```

```
df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

df.describe(); #Information on the variables: mean, std, min, max, 25%, ...


# REGRESSION:

# PLS Regression:

        # Functions:

#Removes unneeded independent variables

def removeColumns(df, lower_bound, word):

    for column in df.columns[20:209]:
        if (df[column].isna().sum()>lower_bound) or (word in column):
            del df[column]

    return df

#Removes unneeded dependent variables

def RegressionON(df, dep_var):

    lst = ['INDEX', 'ALPHA 2016', 'ALPHA 2017', 'ALPHA 2018', 'ALPHA 2019',
'BETA 2019', 'ALPHA 2020', 'VOLATILITY_360D_CALC']

    for column in lst:
        if column != dep_var:
            del df[column]

    return df

#Removes columns that contain a particular word. Used to remove "weigth"
columns.

def cleanData(df, column, word):
    df[column] = df[~df[column].isin([word])]
    return df

#Removes columns that contain a particular word

def cleanData2(df, column, word):
    df = df.where(df[column] != word)
    df = df[df[column].notna()]
    return df

# Division in training and test subsets

def divideTrainTest(df, perc, dep_variable):

    train = df.sample(frac = perc)
    test = df.drop(train.index)

    ytrain = train[dep_variable].values
    ytest = test[dep_variable].values
    xtrain = train.drop([dep_variable], axis=1).values
    xtest = test.drop([dep_variable], axis=1).values

    return train, test, xtrain, ytrain, xtest, ytest

# Run 10k cross-validation with PLS to find number of components that
minimizes MSE while balancing the number of PC
    # plot: Default value: yes
    # maxPC: maximum number of principal components

def MSEvsPrincipalComponents(datax, datay, maxPC, plot = 'yes'):

    n = len(datax)

    #Run 10k fold cross-validation with shuffle
    kf_10 = model_selection.KFold(n_splits=10, shuffle=True, random_state=1)

    mse = [] #Empty list to store MSE values

    for i in np.arange(1, maxPC):
        pls = PLSRegression(n_components=i, scale = True)
        score = model_selection.cross_val_score(pls, scale(datax), datay,
cv=kf_10, scoring='neg_mean_squared_error').mean()
        mse.append(-score)

    print('Lowest MSE:', min(mse), 'for number of components:',
mse.index(min(mse))+1)

    if plot == 'yes':
        # Plot results
        plt.plot(np.arange(1, maxPC), np.array(mse), '-v')
        plt.xlabel('Number of principal components in regression')
        plt.ylabel('MSE')
        plt.title('MSE vs PC')
        plt.xlim(xmin=-1)

#PLS Regression and plot

def runPLSRegression(xtrain, ytrain, xtest, ytest, n_components = 2,
plot_predicted = 'yes'):

    pls = PLSRegression(n_components)
    pls.fit(scale(xtrain), ytrain)

    MSE = mean_squared_error(ytest, pls.predict(scale(xtest)))
    Rsquare = r2_score(ytest, pls.predict(scale(xtest)))
    AdjRsquare = 1-((1-Rsquare)*((len(xtrain)-1)/(len(xtrain)-7-1)))

    print('MSE:', MSE)
    print('R Squared:', Rsquare)
    print('Adj R Squared:', AdjRsquare)

    if plot_predicted == 'yes':
        predicted = pls.predict(scale(xtest))
        actual = ytest
        fig1, ax1 = plt.subplots(figsize=(10, 10))
        ax1.scatter(actual, predicted)
        ax1.set_xlabel('Actual')
        ax1.set_ylabel('Predicted')
        plt.show()

    return pls

#Obtain PLS Coefficients

def PLSCoefficients(pls, variables, plot = 'all'):

    PLS_Coef = pd.DataFrame(pls.coef_, columns = ['Coefficients'])
    Variables = pd.DataFrame(variables[:-1], columns = ['Variables'])
    PLS_Coef = pd.merge(PLS_Coef, Variables, left_index = True, right_index =
True)
    PLS_Coef_L20 = PLS_Coef.nlargest(20, 'Coefficients')
    PLS_Coef_L15 = PLS_Coef.nlargest(15, 'Coefficients')
    PLS_Coef_L10 = PLS_Coef.nlargest(10, 'Coefficients')

    if plot == 'all':
        figsize = (15, 15)
        axx = PLS_Coef.Variables
        axy = PLS_Coef.Coefficients
    else:
        figsize = (10, 10)
        if plot == '20':
            axx = PLS_Coef_L20.Variables
            axy = PLS_Coef_L20.Coefficients
        elif plot == '15':
            axx = PLS_Coef_L15.Variables
            axy = PLS_Coef_L15.Coefficients
        elif plot == '10':
            axx = PLS_Coef_L10.Variables
            axy = PLS_Coef_L10.Coefficients

    if plot == 'all' or plot == '20' or plot == '15' or plot == '10':
        fig, ax = plt.subplots(figsize = figsize)
        ax.barh(axx, axy, color = 'white', edgecolor = 'black')
        plt.show()

    return PLS_Coef

#PLS Parameters

def PLS_Parameters(pls, xtest, ytest):
    return pls.fit_transform(xtest,ytest), pls.x_loadings_, pls.coef_

#Classigy ESG variables into Environmental, Social and Governance categories.

def classifyVariablesPLSLoad(df, PLS_loads, dep_var):

    ESG_Dictionary = pd.read_excel('ESGRatings_Data_Dictionary.xlsx',
index_col=0) #Read from file

    lst = list()
    for column_name in df.columns:
        try:
            dic = dict()
            cat = ESG_Dictionary['Category'].where(ESG_Dictionary['Column 
Name']==column_name).dropna().item()
        except:
            if '1' in column_name: continue
            elif column_name == 'AUDITOR_FEES': continue
            elif column_name == dep_var: continue
            else:
                cat = ESG_Dictionary['Category'].where(ESG_Dictionary['Column 
Name']==column_name).dropna().reset_index()['Category'][0]
        dic['Variables'] = column_name
        dic['Category'] = cat
        lst.append(dic)

    categories = pd.DataFrame(lst, columns = ['Variables', 'Category'])
    categories['Category'] = categories['Category'].replace(['ESG Ratings - 
Governance - KeyMetrics'],'ESG Ratings - Governance')

    PLS_Loadings = pd.DataFrame(PLS_loads, columns = ['PC1', 'PC2'])
    columns = pd.DataFrame(df.columns, columns = ['Variables'])
    PLS_Loadings = pd.merge(PLS_Loadings, columns, left_index = True, 
right_index = True)
    PLS_Loadings = pd.merge(PLS_Loadings, categories, on = 'Variables')

    return PLS_Loadings

#Plot PLS loading vectores

def plotPLSLoadings(PLS_Loadings, names = 'yes'):

    colors = {'ESG Ratings - Governance':'red', 'ESG Ratings - 
Summary':'black', 'ESG Ratings - Social':'blue', 'ESG Ratings - 
Environment':'green'}

    if names == 'yes':
        figsize=(20, 20)

    else: figsize=(10, 10)

    fig, ax = plt.subplots(figsize = figsize)
    ax.scatter(PLS_Loadings.PC1, PLS_Loadings.PC2, 
c=PLS_Loadings['Category'].map(colors))

    if names == 'yes':
        for i, txt in enumerate(PLS_Loadings.Variables):
            ax.annotate(txt, (PLS_Loadings.PC1[i], PLS_Loadings.PC2[i]))
    plt.savefig('PLS.png')
    plt.show()

# Runs whole PLS Regression

def PLS(dep_var):

    #Get our dataset:
    file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG 
Ratings
```

```
    file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
    df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

    #Eliminate columns that we do not need:
    df = removeColumns(df, 1500, 'WEIGHT') #Remove ind variables with more
than 1500 missing values and the weights

    for column in df.columns[0:12]: #Eliminate columns such as name, ISIN,
CUSIP, industry...
        del df[column]

    del df['GM_HOME_MARKET']
    del df['ASSESSMENT_CHANGE_DATE.1']
    del df['ASSESSMENT_CHANGE_DATE']

    df = RegressionON(df, dep_var) #Eliminate dependent variables that are
not going to be looked at in this regression

    #Eliminate missing values:
    df = cleanData2(df, dep_var, '#N/A N/A')
    df = cleanData2(df, dep_var, '#N/A Invalid Security')
    df = cleanData2(df, dep_var, 'NaN')
    df = cleanData2(df, dep_var, '#N/A Field Not Applicable')

    df[dep_var] = df[dep_var].dropna()

    df = df.astype(float) #Change type of values to float
    df = df[~df.isin([np.nan, np.inf, -np.inf]).any(1)] #Eliminate values
close to infinity

    #Eliminate outliers:
    df['z_score']=stats.zscore(df[dep_var])
    df = df.loc[df['z_score'].abs()<=3]
    del df['z_score']

    #Divide data in training and test:
    train, test, xtrain, ytrain, xtest, ytest = divideTrainTest(df, 0.75,
dep_var)

    #PLS:
    MSEvsPrincipalComponents(xtrain, ytrain, maxPC = 20, plot = 'yes')

    variables = train.columns
    pls = runPLSRegression(xtrain, ytrain, xtest, ytest, n_components = 4,
plot_predicted = 'yes')
    pls = runPLSRegression(xtrain, ytrain, xtest, ytest, n_components = 2,
plot_predicted = 'no')
    PLS_Coef = PLSCoefficients(pls, variables, plot = 'yes')
    PLS_score, PLS_loads, PLS_coef = PLS_Parameters(pls, xtest, ytest)
    PLS_Loadings = classifyVariablesPLSLoad(df, PLS_loads, dep_var = dep_var)
    PLS_Scores = pd.DataFrame(PLS_score[0], columns = ['PC1', 'PC2'])
    plotPLSLoadings(PLS_Loadings, names = 'yes')

    return df, PLS_score, PLS_loads, PLS_Coef

        # Main:
```

```
plsdf, PLS_score, PLS_loads, PLS_Coef  = PLS(dep_var = 'ALPHA 2019') #PLS
execution

# Linear Regression

        # Functions:

# Linear Regression preparation: cleaning, adding constant,...

def prepareOLSRegression(df, dep_var, filter_type, filter_coun, filter_ind,
n_features):

    df = removeColumns(df, 1500, 'WEIGHT') #Remove ind variables with more
than 1500 missing values and the weights

    ind_var = PLS_Coef.nlargest(n_features,
'Coefficients').Variables.unique()

    for column in df.columns:
        if (column != dep_var) and (column not in ind_var) and
(column!='GM_HOME_MARKET') and (column!='IVA_INDUSTRY'):
            del df[column]

    df = cleanData2(df, dep_var, '#N/A N/A') #Eliminate missing values
    df = cleanData2(df, dep_var, '#N/A Invalid Security') #Eliminate missing
values
    df = cleanData2(df, dep_var, 'NaN') #Eliminate missing values

    df[dep_var] = df[dep_var].dropna() #Eliminate missing values

    if filter_type == 'industry':
        del df['GM_HOME_MARKET']
        df = df.where(df['IVA_INDUSTRY']==filter_ind)
        df = df[df['IVA_INDUSTRY'].notna()]
        del df['IVA_INDUSTRY']

    if filter_type == 'country':
        del df['IVA_INDUSTRY']
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        del df['GM_HOME_MARKET']

    if filter_type == 'both':
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        df = df.where(df['IVA_INDUSTRY']==filter_ind)
        df = df[df['IVA_INDUSTRY'].notna()]
        del df['GM_HOME_MARKET']
        del df['IVA_INDUSTRY']

    if filter_type == 'no':
        del df['GM_HOME_MARKET']
        del df['IVA_INDUSTRY']

    df = df.astype(float) #Change type of values to float
    df = df[~df.isin([np.nan, np.inf, -np.inf]).any(1)] #Eliminate values
close to infinity

    y = df[dep_var]
    x = df[ind_var]
    x = sm.add_constant(x)

    return x, y, ind_var

#Performs Linear Regression

def performOLSRegression(X, Y, until_n, show_all = 'no', show_best = 'yes'):

    models_best = pd.DataFrame(columns=["RSS", "model"])

    tic = time.time()
    for i in range(1,until_n):
        models_best.loc[i] = getBest(i)

    toc = time.time()
    print("Total elapsed time:", (toc-tic), "seconds.")

    if show_all == 'yes':
        print(models_best.apply(lambda row: row[1].rsquared_adj, axis=1))
    if show_best == 'yes':
        best_feat = models_best.apply(lambda row: row[1].rsquared_adj,
axis=1).nlargest(1).index[0]

    print(models_best.loc[best_feat, "model"].summary())
def processSubset(feature_set):
    # Fit model on feature_set and calculate RSS
    model = sm.OLS(Y,X[list(feature_set)])
    regr = model.fit()
    RSS = ((regr.predict(X[list(feature_set)]) - Y) ** 2).sum()
    return {"model":regr, "RSS":RSS}
def getBest(k):

    tic = time.time()

    results = []

    for combo in itertools.combinations(X.columns, k):
        results.append(processSubset(combo))

    models = pd.DataFrame(results)

    # Choose best model
    best_model = models.loc[models['RSS'].argmin()]

    toc = time.time()
    print("Processed", models.shape[0], "models on", k, "predictors in",
(toc-tic), "seconds.")

    return best_model

        # Main:
```

```
file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

X, Y, ind_var = prepareOLSRegression(df, dep_var = 'ALPHA 2019', filter_type
= 'no',
                            filter_coun = 'Japan', filter_ind = 'Steel',
n_features = 20)

print('Independent variables:', ind_var)

performOLSRegression(X, Y, 3, show_all = 'no', show_best = 'yes');

# Random forests:

        # Functions:

#Prepare data for Random Forests

def prepareRFRegression(df, dep_var, filter_type, filter_coun, filter_ind,
n_features):

    deletecolumns = ['ISSUER_NAME','issuerId',
    'ISSUER_TICKER','ISSUER_CUSIP', 'ISSUER_SEDOL',
    'ISSUER_ISIN', 'CIK_NUM',
    'ISSUER_CNTRY_DOMICILE', 'GICS_SUB_IND',
    'IVA_RATING_DATE', 'IVA_COMPANY_RATING', 'INDUSTRY_ADJUSTED_SCORE']

    lower_bound = 1500

    for column in deletecolumns:
        del df[column]

    for column in df.columns:
        if (df[column].isna().sum()>lower_bound):
            del df[column]

    for column in df.columns:
        if 'WEIGHT' in column:
            del df[column]

    df = RegressionON(df, dep_var)

    ind_var = df.columns.tolist()
    ind_var.remove(dep_var)
    ind_var.remove('GM_HOME_MARKET')
    ind_var.remove('IVA_INDUSTRY')

    for column in df.columns:
        if (column != dep_var) and (column not in ind_var) and
(column!='GM_HOME_MARKET') and (column!='IVA_INDUSTRY'):
            del df[column]

    df = cleanData2(df, dep_var, '#N/A N/A') #Eliminate missing values
    df = cleanData2(df, dep_var, '#N/A Invalid Security') #Eliminate missing
values
    df = cleanData2(df, dep_var, 'NaN') #Eliminate missing values
    df = cleanData2(df, dep_var, '#N/A Field Not Applicable')

    df[dep_var] = df[dep_var].dropna() #Eliminate missing values

    if filter_type == 'industry':
        del df['GM_HOME_MARKET']
        df = df.where(df['IVA_INDUSTRY']==filter_ind)
        df = df[df['IVA_INDUSTRY'].notna()]
        del df['IVA_INDUSTRY']

    if filter_type == 'country':
        del df['IVA_INDUSTRY']
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        del df['GM_HOME_MARKET']

    if filter_type == 'both':
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        df = df.where(df['IVA_INDUSTRY']==filter_ind)
        df = df[df['IVA_INDUSTRY'].notna()]
        del df['GM_HOME_MARKET']
        del df['IVA_INDUSTRY']

    if filter_type == 'no':
        del df['GM_HOME_MARKET']
        del df['IVA_INDUSTRY']

    df = df.astype(float) #Change type of values to float
    df = df[~df.isin([np.nan, np.inf, -np.inf]).any(1)] #Eliminate values
close to infinity

    y = df[dep_var]
    x = df[ind_var]
    x = sm.add_constant(x)

    return x, y, ind_var

#Perform Random Forests

def performRF(xtrain, xtest, ytrain, ytest, print_stats = 'yes', print_png =
'yes', plot = 'yes'):

    sc = StandardScaler()
    xtrain = sc.fit_transform(xtrain)
    xtest = sc.transform(xtest)

    adj_R_square = 0

    best_max_depth = 7

    for i in range(5, 50):
        rf = RandomForestRegressor(n_estimators = 25, max_depth = 7,
random_state = 42)
```

```
        # Train the model on training data
        rf_model = rf.fit(xtrain, ytrain)
        ypred = rf.predict(xtest)
        if adj_R_square < metrics.r2_score(ytest, ypred):
            adj_R_square = metrics.r2_score(ytest, ypred)
            best_max_depth = i

    # Fit model
    rf = RandomForestRegressor(n_estimators = 25, max_depth = best_max_depth,
random_state = 42)
    # Train the model on training data
    rf_model = rf.fit(xtrain, ytrain)

    # Use the forest's predict method on the test data
    ypred = rf.predict(xtest)
    # Calculate the absolute errors
    errors = abs(ypred - ytest)

    #Print metrics
    if print_stats == 'yes':
        # Print out the mean absolute error (mae)
        print('MSE:', round(metrics.mean_squared_error(ytest, ypred), 2))
        print('Adj R-Square:', round(metrics.r2_score(ytest, ypred), 2))

    # Create and save graph:
    if print_png == 'yes':
        tree = rf.estimators_[8]
        export_graphviz(tree, out_file = 'tree.dot', feature_names = ind_var,
rounded = True, precision = 1)
        (graph, ) = pydot.graph_from_dot_file('tree.dot')
        graph.write_png('tree.png')

    if plot == 'yes':
        importances = rf_model.feature_importances_
        std = np.std([tree.feature_importances_ for tree in
rf_model.estimators_], axis=0)
        indices = np.argsort(importances)[::-1]

        dic = dict()
        for i in range (0, len(ind_var)):
            e = indices[i]
            dic[e] = ind_var[e]

        # Print the feature ranking
        print("Feature ranking:")

        for f in range(X.shape[1]):
            pos = indices[f]
            a = f+1
            print(a, 'feature:', dic[pos], '(%f)' %
(importances[indices[f]]))

        # Plot
        plt.figure()
        plt.title("Feature importances")
        plt.barh(range(X.shape[1]), importances[indices], color='red',
edgecolor = 'black', yerr=std[indices], align="center")
```

```
        plt.yticks(range(X.shape[1]), dic.values())
        plt.ylim([-1, X.shape[1]])
        plt.show()

    return dic.values()

        # Main:

file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

X, Y, ind_var = prepareRFRegression(df, dep_var = 'VOLATILITY_360D_CALC',
filter_type = 'no',
                            filter_coun = 'Japan', filter_ind = 'Steel',
n_features = 20)

del X['const']

y = np.array(Y)
x = np.array(X)

#Split data
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = 0.25,
random_state = 42)

imp_features = performRF(xtrain, xtest, ytrain, ytest, print_stats = 'yes',
print_png = 'yes', plot = 'yes')
ind_var = list(imp_features)[:15];

# Linear Regression after RF:

        # Functions:

def prepareOLSafterRF(df, dep_var, ind_var, filter_type, filter_coun,
filter_ind, n_features):

    deletecolumns = ['ISSUER_NAME','issuerId',
    'ISSUER_TICKER','ISSUER_CUSIP', 'ISSUER_SEDOL',
    'ISSUER_ISIN', 'CIK_NUM',
    'ISSUER_CNTRY_DOMICILE', 'GICS_SUB_IND',
    'IVA_RATING_DATE', 'IVA_COMPANY_RATING', 'INDUSTRY_ADJUSTED_SCORE']

    lower_bound = 1500

    for column in deletecolumns:
        del df[column]

    for column in df.columns:
        if (df[column].isna().sum()>lower_bound):
            del df[column]

    for column in df.columns:
        if 'WEIGHT' in column:
            del df[column]
    df = RegressionON(df, dep_var)

    for column in df.columns:
        if (column != dep_var) and (column not in ind_var) and
(column!='GM_HOME_MARKET') and (column!='IVA_INDUSTRY'):
            del df[column]

    df = cleanData2(df, dep_var, '#N/A N/A') #Eliminate missing values
    df = cleanData2(df, dep_var, '#N/A Invalid Security') #Eliminate missing
values
    df = cleanData2(df, dep_var, 'NaN') #Eliminate missing values
    df = cleanData2(df, dep_var, '#N/A Field Not Applicable')

    df[dep_var] = df[dep_var].dropna() #Eliminate missing values

    if filter_type == 'industry':
        del df['GM_HOME_MARKET']
        df = df.where(df['IVA_INDUSTRY']==filter_ind)
        df = df[df['IVA_INDUSTRY'].notna()]
        del df['IVA_INDUSTRY']

    if filter_type == 'country':
        del df['IVA_INDUSTRY']
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        del df['GM_HOME_MARKET']

    if filter_type == 'both':
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        df = df.where(df['IVA_INDUSTRY']==filter_ind)
        df = df[df['IVA_INDUSTRY'].notna()]
        del df['GM_HOME_MARKET']
        del df['IVA_INDUSTRY']

    if filter_type == 'no':
        del df['GM_HOME_MARKET']
        del df['IVA_INDUSTRY']

    df = df.astype(float) #Change type of values to float
    df = df[~df.isin([np.nan, np.inf, -np.inf]).any(1)] #Eliminate values
close to infinity

    y = df[dep_var]
    x = df[ind_var]
    x = sm.add_constant(x)

    return x, y, ind_var

        # Main:

file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings
```

```
X, Y, ind_var = prepareOLSafterRF(df, dep_var = 'VOLATILITY_360D_CALC',
ind_var = ind_var, filter_type = 'no',
                            filter_coun = 'Japan', filter_ind = 'Industrial
Conglomerates', n_features = 20)

print('Independent variables:', ind_var)
print('')

performOLSRegression(X, Y, 10, show_all = 'no', show_best = 'yes')
With regions, not countries
def countCountries(df):

    #COUNT HOW MANY OF EACH COUNTRY/REGION:
    countriesCount = dict()

    #Set count to zero
    for item in df['GM_HOME_MARKET'].unique().tolist():
        countriesCount[item] = 0

    #Count
    for country in df['GM_HOME_MARKET']:
        if country in countriesCount.keys():
            countriesCount[country] += 1

    #WE WILL GROUP COUNTRIES BY REGION if the companies of a country are less
than 400:

    countriesToGroup = list()
    countriesNotToGroup = list()

    for country, count in countriesCount.items():
        if count < 400 :
            countriesToGroup.append(country)
        else:
            countriesNotToGroup.append(country)

    return countriesToGroup
def substituteCountries(df, file, countriesToGroup):

    for country in countriesToGroup:
        cont = file.loc[file['COUNTRY'] == country, 'CONTINENT']
        cont2 = ''.join(cont.values)
        df['GM_HOME_MARKET'] = df['GM_HOME_MARKET'].replace(country, cont2)

    return df
file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

countriesToGroup = countCountries(df)

file3 = pd.read_csv('Countries.csv') #Read file containing ESG Ratings

df = substituteCountries(df, file3, countriesToGroup)
```

```
X, Y, ind_var = prepareOLSafterRF(df, dep_var = 'ALPHA 2019', ind_var =
ind_var, filter_type = 'country',
                            filter_coun = 'Oceania', filter_ind = 'Industrial
Conglomerates', n_features = 20)

print('Independent variables:', ind_var)
print('')

performOLSRegression(X, Y, 8, show_all = 'no', show_best = 'yes')

# With industries

def cleanwithInd(df, dep_var, ind_var):

    deletecolumns = ['ISSUER_NAME','issuerId',
    'ISSUER_TICKER','ISSUER_CUSIP', 'ISSUER_SEDOL',
    'ISSUER_ISIN', 'CIK_NUM',
    'ISSUER_CNTRY_DOMICILE', 'GICS_SUB_IND',
    'IVA_RATING_DATE', 'IVA_COMPANY_RATING', 'INDUSTRY_ADJUSTED_SCORE']

    lower_bound = 1500

    for column in deletecolumns:
        del df[column]

    for column in df.columns:
        if (df[column].isna().sum()>lower_bound):
            del df[column]

    for column in df.columns:
        if 'WEIGHT' in column:
            del df[column]

    df = RegressionON(df, dep_var)

    for column in df.columns:
        if (column != dep_var) and (column not in ind_var) and
(column!='GM_HOME_MARKET') and (column!='GICS_INDUSTRY'):
            del df[column]

    df = cleanData2(df, dep_var, '#N/A N/A') #Eliminate missing values
    df = cleanData2(df, dep_var, '#N/A Invalid Security') #Eliminate missing
values
    df = cleanData2(df, dep_var, 'NaN') #Eliminate missing values

    df[dep_var] = df[dep_var].dropna() #Eliminate missing values

    return df

def filterIndustries(df_industries, dep_var, ind_var, filter_ind,
n_features):

    df_industries =
df_industries.where(df_industries['GICS_INDUSTRY']==filter_ind)
    df_industries = df_industries[df_industries['GICS_INDUSTRY'].notna()]
    del df_industries['GICS_INDUSTRY']
    df_industries = df_industries.astype(float) #Change type of values to
float
    df_industries = df_industries[~df_industries.isin([np.nan, np.inf, -
np.inf]).any(1)] #Eliminate values close to infinity

    y = df_industries[dep_var]
    x = df_industries[ind_var]
    x = sm.add_constant(x)

    return x, y, ind_var

def prepareOLSwithIndustries(df, dep_var, ind_var, filter_type, filter_coun,
filter_ind, n_features):

    deletecolumns = ['ISSUER_NAME','issuerId',
    'ISSUER_TICKER','ISSUER_CUSIP', 'ISSUER_SEDOL',
    'ISSUER_ISIN', 'CIK_NUM',
    'ISSUER_CNTRY_DOMICILE', 'GICS_SUB_IND',
    'IVA_RATING_DATE', 'IVA_COMPANY_RATING', 'INDUSTRY_ADJUSTED_SCORE']

    lower_bound = 1500

    for column in deletecolumns:
        del df[column]

    for column in df.columns:
        if (df[column].isna().sum()>lower_bound):
            del df[column]

    for column in df.columns:
        if 'WEIGHT' in column:
            del df[column]

    df = RegressionON(df, dep_var)

    for column in df.columns:
        if (column != dep_var) and (column not in ind_var) and
(column!='GM_HOME_MARKET') and (column!='GICS_INDUSTRY'):
            del df[column]

    df = cleanData2(df, dep_var, '#N/A N/A') #Eliminate missing values
    df = cleanData2(df, dep_var, '#N/A Invalid Security') #Eliminate missing
values
    df = cleanData2(df, dep_var, 'NaN') #Eliminate missing values

    df[dep_var] = df[dep_var].dropna() #Eliminate missing values

    if filter_type == 'industry':
        del df['GM_HOME_MARKET']
        df = df.where(df['GICS_INDUSTRY']==filter_ind)
        df = df[df['GICS_INDUSTRY'].notna()]
        del df['GICS_INDUSTRY']

    if filter_type == 'country':
        del df['GICS_INDUSTRY']
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        del df['GM_HOME_MARKET']

    if filter_type == 'both':
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        df = df.where(df['GICS_INDUSTRY']==filter_ind)
        df = df[df['GICS_INDUSTRY'].notna()]
        del df['GM_HOME_MARKET']
        del df['GICS_INDUSTRY']

    if filter_type == 'no':
        del df['GM_HOME_MARKET']
        del df['GICS_INDUSTRY']

    df = df.astype(float) #Change type of values to float
    df = df[~df.isin([np.nan, np.inf, -np.inf]).any(1)] #Eliminate values
close to infinity

    y = df[dep_var]
    x = df[ind_var]
    x = sm.add_constant(x)

    return x, y, ind_var

def performOLSRegressionwithIndustries(X, Y, until_n, show_all = 'no',
show_best = 'yes'):

    models_best = pd.DataFrame(columns=["RSS", "model"])

    tic = time.time()
    for i in range(1,until_n):
        models_best.loc[i] = getBest(i)

    toc = time.time()
    print("Total elapsed time:", (toc-tic), "seconds.")

    if show_all == 'yes':
        print(models_best.apply(lambda row: row[1].rsquared_adj, axis=1))
    if show_best == 'yes':
        best_feat = models_best.apply(lambda row: row[1].rsquared_adj,
axis=1).nlargest(1).index[0]

    print(models_best.loc[best_feat, "model"].summary())

    return models_best.loc[best_feat, "model"]

        # Main:

file1 = pd.read_csv('esg_ratings_factors3.csv') #Read file containing ESG
Ratings
file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

industries = list()
coeff = list()
```

```
pval = list()
adjrsq = list()

df = cleanwithInd(df, dep_var = 'ALPHA 2019', ind_var = ind_var)

del df['GM_HOME_MARKET']

for industry in df['GICS_INDUSTRY'].unique():

    print('Industry:', industry)

    X, Y, ind_var = filterIndustries(df, dep_var = 'ALPHA 2019', ind_var =
ind_var, filter_ind = industry, n_features = 20)

    model = performOLSRegressionwithIndustries(X, Y, 6, show_all = 'no',
show_best = 'yes')

    industries.append(industry)
    coeff.append(model.params)
    pval.append(model.pvalues)
    adjrsq.append(model.rsquared_adj)

# Gradient boosting:

        # Functions:

def performGBoosting(xtrain, ytrain, xtest, ytest, plot = 'yes'):

    best_n_estimators = 1

    for i in range(1, 200, 4):

        n_estimators = i
        GBoosting = GradientBoostingRegressor(n_estimators = n_estimators,
max_depth = 100, learning_rate = 0.01, criterion = 'mse')
        GBoosting_model = GBoosting.fit(xtrain, ytrain)
        ypred = GBoosting_model.predict(xtest)
        r2 = metrics.r2_score(ytest, ypred)
        if i == 1:
            r2_before = r2
        if r2 > r2_before:
            best_n_estimators = i
            r2_before = r2
        mse = mean_squared_error(ytest, ypred)

    n_estimators = best_n_estimators
    GBoosting = GradientBoostingRegressor(n_estimators = n_estimators,
max_depth = 100, learning_rate = 0.01, criterion = 'mse')
    GBoosting_model = GBoosting.fit(xtrain, ytrain)
    ypred = GBoosting_model.predict(xtest)
    r2 = metrics.r2_score(ytest, ypred)

    print('R-squared:', r2)
    print('MSE:', mse)

    test_score = np.zeros((n_estimators,), dtype=np.float64)
    for i, y_pred in enumerate(GBoosting.staged_predict(xtest)):
        test_score[i] = GBoosting.loss_(ytest, ypred)

    if plot == 'yes':
        fig = plt.figure(figsize=(6, 6))
        plt.subplot(1, 1, 1)
        plt.title('Deviance')
        plt.plot(np.arange(n_estimators) + 1, GBoosting.train_score_, 'b-',
                 label='Training Set Deviance')
        plt.plot(np.arange(n_estimators) + 1, test_score, 'r-',
                 label='Test Set Deviance')
        plt.legend(loc='upper right')
        plt.xlabel('Boosting Iterations')
        plt.ylabel('Deviance')
        fig.tight_layout()
        plt.show()

        # Main:

file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

X, Y, ind_var = prepareOLSRegression(df, dep_var = 'ALPHA 2019', filter_type
= 'country',
                            filter_coun = 'United States', filter_ind =
'Steel', n_features = 20)

del X['const']

# Labels are the values we want to predict
y = np.array(Y, dtype = float)
# Convert to numpy array
x = np.array(X, dtype = float)
# Split the data into training and testing sets
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = 0.25,
random_state = 2)

sc = StandardScaler()
xtrain = sc.fit_transform(xtrain)
xtest = sc.transform(xtest)

performGBoosting(xtrain, ytrain, xtest, ytest, plot = 'no')


# Polynomial regression:

        # Functions:

def preparePolyRegression(df, dep_var, filter_type, filter_coun, filter_ind,
n_features):

    df = removeColumns(df, 1500, 'WEIGHT') #Remove ind variables with more
than 1500 missing values and the weights
```

```
    ind_var = PLS_Coef.nlargest(n_features,
'Coefficients').Variables.unique()

    for column in df.columns:
        if (column != dep_var) and (column not in ind_var) and
(column!='GM_HOME_MARKET') and (column!='IVA_INDUSTRY'):
            del df[column]

    df = cleanData2(df, dep_var, '#N/A N/A') #Eliminate missing values
    df = cleanData2(df, dep_var, '#N/A Invalid Security') #Eliminate missing
values
    df = cleanData2(df, dep_var, 'NaN') #Eliminate missing values

    df[dep_var] = df[dep_var].dropna() #Eliminate missing values

    if filter_type == 'industry':
        del df['GM_HOME_MARKET']
        df = df.where(df['IVA_INDUSTRY']==filter_ind)
        df = df[df['IVA_INDUSTRY'].notna()]
        del df['IVA_INDUSTRY']

    if filter_type == 'country':
        del df['IVA_INDUSTRY']
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        del df['GM_HOME_MARKET']

    if filter_type == 'both':
        df = df.where(df['GM_HOME_MARKET']==filter_coun)
        df = df[df['GM_HOME_MARKET'].notna()]
        df = df.where(df['IVA_INDUSTRY']==filter_ind)
        df = df[df['IVA_INDUSTRY'].notna()]
        del df['GM_HOME_MARKET']
        del df['IVA_INDUSTRY']

    df = df.astype(float) #Change type of values to float
    df = df[~df.isin([np.nan, np.inf, -np.inf]).any(1)] #Eliminate values
close to infinity

    df['z_score']=stats.zscore(df[dep_var])
    df = df.loc[df['z_score'].abs()<=3]
    del df['z_score']

    y = df[dep_var]
    x = df[ind_var]
    x = sm.add_constant(x)

    return x, y, ind_var

def performPolyRegression(x, y):

    poly_reg = PolynomialFeatures(degree=4)
    x_pol = poly_reg.fit_transform(x)

    #Divide data in training and test:
    xtrain_pol, xtest_pol, ytrain_pol, ytest_pol = train_test_split(x_pol, y,
test_size = 0.25, random_state = 4)
```

```
    pol_reg = LinearRegression()
    pol_model = pol_reg.fit(xtrain_pol, ytrain_pol)

    MSE = mean_squared_error(ytest_pol, pol_reg.predict(xtest_pol))
    Rsquare = r2_score(ytest_pol, pol_reg.predict(xtest_pol))
    AdjRsquare = 1-((1-Rsquare)*((len(xtest_pol)-1)/(len(xtest_pol)-7-1)))

    print('MSE:', MSE)
    print('AdjRsquare:', AdjRsquare)


        # Main:

file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

x, y, ind_var = preparePolyRegression(df, dep_var = 'ALPHA 2019', filter_type
= 'country',
                            filter_coun = 'Japan', filter_ind = 'Steel',
n_features = 20)

performPolyRegression(x, y)


# Clustering:

# Decision trees:

        # Functions:

def preHierarchical(dep_var):

    #Get our dataset:
    file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
    file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
    df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

    #Eliminate columns that we do not need:
    df = removeColumns(df, 1500, 'WEIGHT') #Remove ind variables with more
than 1500 missing values and the weights

    for column in df.columns[0:12]: #Eliminate columns such as name, ISIN,
CUSIP, industry...
        del df[column]

    del df['GM_HOME_MARKET']
    del df['ASSESSMENT_CHANGE_DATE.1']
    del df['ASSESSMENT_CHANGE_DATE']

    df = RegressionON(df, dep_var) #Eliminate dependent variables that are
not going to be looked at in this regression
```

```
    #Eliminate missing values:
    df = cleanData2(df, dep_var, '#N/A N/A')
    df = cleanData2(df, dep_var, '#N/A Invalid Security')
    df = cleanData2(df, dep_var, 'NaN')

    df[dep_var] = df[dep_var].dropna()

    df = df.astype(float) #Change type of values to float
    df = df[~df.isin([np.nan, np.inf, -np.inf]).any(1)] #Eliminate values
close to infinity

    #Eliminate outliers:
    df['z_score']=stats.zscore(df[dep_var])
    df = df.loc[df['z_score'].abs()<=3]
    del df['z_score']

    df[dep_var] = df[dep_var]
    df['ALPHA GROUP'] = pd.cut(df[dep_var], bins=10, labels=False) + 1
    df['ALPHA GROUP'] = df['ALPHA GROUP'].astype(int)
    del df[dep_var]

    return df

def performHierarchical(df, dep_var):

    y = df[dep_var].values
    x = df.drop([dep_var], axis=1).values
    ind_var = df.columns[:-1]
    ind_var = ind_var.tolist()
    train, test, xtrain, ytrain, xtest, ytest = divideTrainTest(df, 0.75,
dep_var)
    classifier = DecisionTreeClassifier(criterion = 'gini', max_features =
15, max_leaf_nodes = 15)
    classifier.fit(xtrain, ytrain)
    y_pred = classifier.predict(xtest)
    print(confusion_matrix(ytest, y_pred))
    print(classification_report(ytest, y_pred))

    dot_tree = tree.export_graphviz(classifier,
out_file='tree_classifier.dot', feature_names = ind_var, filled=True,
rounded=True, special_characters=True)
    (photo, ) = pydot.graph_from_dot_file('tree_classifier.dot')
    photo.write_png('tree_classifier.png')

    return classifier, x, y

def Treepairplot(classifier, df, dep_var, sample_size):

    node_features = classifier.tree_.feature.tolist()
    node_features = list(filter(lambda a: a != -2, node_features))
    lst = list()
    for feature_number in node_features:
        if df.columns[feature_number] not in lst:
            lst.append(df.columns[feature_number])
    lst.append('label')
    lst.append(dep_var)
    labels = classifier.apply(x)
    df['label'] = labels
    clustered = df[lst].sample(sample_size)
    sns.pairplot(data = clustered, hue = 'label', palette = 'tab10', kind =
'scatter')

    return lst


lst.append('ENVIRONMENTAL_PILLAR_SCORE')
lst.append('CORP_GOVERNANCE_SCORE')

def Treepairplotvar(classifier, df, dep_var, sample_size, var1, var2):

    lst = list()
    lst.append('label')
    lst.append(var1)
    lst.append(var2)
    lst.append(dep_var)
    labels = classifier.apply(x)
    df['label'] = labels
    clustered = df[lst].sample(sample_size)
    sns.pairplot(data = clustered, hue = 'label', palette = 'tab10', kind =
'scatter')

        # Main:

df = preHierarchical(dep_var = 'ALPHA 2019')
classifier, x, y = performHierarchical(df, 'ALPHA GROUP')
cluster_imp_features = Treepairplot(classifier, df, dep_var = 'ALPHA GROUP',
sample_size = 500)

Treepairplotvar(classifier, df, dep_var = 'ALPHA GROUP', sample_size = 500,
var1 = 'ENVIRONMENTAL_PILLAR_SCORE', var2 = 'CORP_GOVERNANCE_SCORE')

# K-means:

        # Functions:


def preK_means(dep_var):

    #Get our dataset:
    file1 = pd.read_csv('esg_ratings_factors.csv') #Read file containing ESG
Ratings
    file2 = pd.read_csv('Bloomberg.csv') #Read file containing Bloomberg data
    df = pd.merge(file1, file2, on = 'ISSUER_NAME') #Combine bloomberg
information with ESG Ratings

    #Eliminate columns that we do not need:
    df = removeColumns(df, 1500, 'WEIGHT') #Remove ind variables with more
than 1500 missing values and the weights

    for column in df.columns[0:12]: #Eliminate columns such as name, ISIN,
CUSIP, industry...
        del df[column]
    del df['GM_HOME_MARKET']
    del df['ASSESSMENT_CHANGE_DATE.1']
    del df['ASSESSMENT_CHANGE_DATE']

    df = RegressionON(df, dep_var) #Eliminate dependent variables that are
not going to be looked at in this regression

    #Eliminate missing values:
    df = cleanData(df, dep_var, '#N/A N/A')
    df = cleanData(df, dep_var, '#N/A Invalid Security')
    df = cleanData(df, dep_var, 'NaN')

    df[dep_var] = df[dep_var].dropna()

    df = df.astype(float) #Change type of values to float
    df = df[~df.isin([np.nan, np.inf, -np.inf]).any(1)] #Eliminate values
close to infinity

    #Eliminate outliers:
    df['z_score']=stats.zscore(df[dep_var])
    df = df.loc[df['z_score'].abs()<=3]
    del df['z_score']

    df[dep_var] = df[dep_var]
    df['ALPHA GROUP'] = pd.cut(df[dep_var], bins=10, labels=False) + 1
    df['ALPHA GROUP'] = df['ALPHA GROUP'].astype(int)
    del df[dep_var]

    sc = StandardScaler()
    x = sc.fit_transform(df)

    return df

def plotSSEvsClusters(df):

    # Plot to check best number of clusters.
    nc = range(1, 30) # Number of iterations to perform.
    kmeans = [KMeans(n_clusters=i) for i in nc]
    score = [kmeans[i].fit(df).score(df) for i in range(len(kmeans))]
    plt.xlabel('Number of clusters (k)')
    plt.ylabel('SSE')
    plt.plot(nc,score)
    plt.show()

def performKmeans(df, n_clusters):

    kmeans = KMeans(n_clusters = n_clusters).fit(df)
    centroids = kmeans.cluster_centers_

    return kmeans, centroids

def plotKmeans(df, kmeans, dep_var, centroids, var1, var2, var3, var4):

    labels = kmeans.predict(df)
    df['label'] = labels
    #cluster_imp_features.remove('ALPHA GROUP')
    #cluster_imp_features.append(dep_var)
    clustered = df[cluster_imp_features].sample(500)

    #%% Plot k-means clustering.
    colors=['red','green','blue','yellow','fuchsia', 'brown', 'gray',
'olive', 'cyan', 'purple']
    assign=[]
    for row in labels:
         assign.append(colors[row])

    plt.scatter(df[var1], df[var2], c=assign, s=1)
    plt.scatter(centroids[:, 0], centroids[:, 1], marker='*', c='black',
s=20) # Marco centroides.
    plt.xlabel(var1)
    plt.ylabel(var2)
    plt.show()

    plt.scatter(df[var3], df[var4], c=assign, s=1)
    plt.scatter(centroids[:, 0], centroids[:, 1], marker='*', c='black',
s=20) # Marco centroides.
    plt.xlabel(var3)
    plt.ylabel(var4)
    plt.show()

    return clustered

        # Main:

dep_var = 'ALPHA 2019'
df = preK_means(dep_var)
plotSSEvsClusters(df)
kmeans, centroids = performKmeans(df, n_clusters = 10)
clustered = plotKmeans(df, kmeans, dep_var, centroids, var1 = 'ALPHA GROUP',
var2 = 'CARBON_EMISSIONS_SCORE',
                       var3 = 'GOVERNANCE_PILLAR_SCORE', var4 =
'ENVIRONMENTAL_PILLAR_SCORE')

sns.pairplot(clustered, hue = 'label', palette = 'tab10')
dep_var = 'ALPHA 2019'
df = preK_means(dep_var)
plotSSEvsClusters(df)
kmeans, centroids = performKmeans(df, n_clusters = 10)
clustered = plotKmeans(df, kmeans, dep_var, centroids, var1 = 'ALPHA GROUP',
var2 = 'CARBON_EMISSIONS_SCORE',
                       var3 = 'GOVERNANCE_PILLAR_SCORE', var4 =
'ENVIRONMENTAL_PILLAR_SCORE')

sns.pairplot(clustered, hue = 'label', palette = 'tab10')
```