A LONGITUDINAL ANALYSIS OF CSR DISCLOSURE FOR BIST COMPANIES: A TEXT MINING APPROACH

by
ÖYKÜ AĞKOÇ AYRADİLLİ

Submitted to the Sabancı Graduate Business School
in partial fulfilment of
the requirements for the degree of Master of Science in Business Analytics

Sabancı University
July 2022

A LONGITUDINAL ANALYSIS OF CSR DISCLOSURE FOR BIST COMPANIES: A TEXT MINING APPROACH

Approved by:

Assoc. Prof. Dr. Ayşe Kocabıyıkoğlu . . . .
(Thesis Supervisor)

Asst. Prof. Dr. Burak Gökgür . . . .

Assoc. Prof. Dr. Cemal Deniz Yenigün . . . .

Date of Approval: July 6, 2022

ABSTRACT

A LONGITUDINAL ANALYSIS OF CSR DISCLOSURE FOR BIST COMPANIES: A TEXT MINING APPROACH

ÖYKÜ AĞKOÇ AYRADİLLİ

Business Analytics M.Sc. THESIS, JULY 2022

Thesis Advisor: Assoc. Prof. Dr. Ayşe Kocabıyıkoğlu
Thesis Co-Advisor: Asst. Prof. Dr. Celile Itır Göğüş

Keywords: corporate social responsibility, CSR disclosure, ESG reporting, corporate annual report, text-mining, dictionary-based text analysis, clustering

With environmental problems being more salient in daily life, corporations increasingly started to report activities they undertake that do not solely entail financial interest but are mostly related to their impact on the society, otherwise known as their Corporate Social Responsibility (CSR) activities. In the business and management literature, CSR reporting has particularly become a major research interest and a great source for understanding CSR behavior. Despite the wide interest in analyzing CSR reporting in the last decades, the range of methods for analysis remain narrow, mainly dominated by the widely used content analysis method. In this thesis, we followed a novel text mining approach to examine the annual reports of BIST companies from 2007 to 2020. For this purpose, we firstly prepared an ESG dictionary to extract keywords from the annual reports and assigned aggregate environment, social and governance scores to each report. Descriptive results for all data showed that governance related information has the highest salience among all ESG categories while environment salience has an upward trend. As a secondary task, we employed two different clustering algorithms, k-medoids and hierarchical (agglomerative) clustering, to group all reports based on their ESG salience. Our analysis revealed 3 distinct groups of reports and showed that the share of the group with high environment scores have increased significantly in 2020.

ÖZET

BİST ŞİRKETLERİ İÇİN KSS AÇIKLAMALARININ UZUNLAMASINA ANALİZİ: BİR METİN MADENCİLİĞİ YAKLAŞIMI

ÖYKÜ AĞKOÇ AYRADİLLİ

İş Analitiği YÜKSEK LİSANS TEZİ, Temmuz 2022

Tez Danışmanı: Prof. Dr. Ayşe Kocabıyıkoğlu
İkinci Tez Danışmanı: Asst. Prof. Dr. Celile Itır Göğüş

Anahtar Kelimeler: kurumsal sosyal sorumluluk, KSS açıklamaları, ESG raporlaması, kurumsal yıllık rapor, metin madenciliği, sözlük tabanlı metin analizi, kümeleme

Çevre sorunlarının günlük yaşamda daha belirgin hale gelmesiyle birlikte şirketler, finansal çıkarlarına ek olarak toplum üzerindeki etkileriyle de ilgili olan, diğer bir deyişle Kurumsal Sosyal Sorumluluk (KSS) faaliyetleri olarak bilinen faaliyetlerini giderek daha fazla raporlamaya başladılar. İşletme ve yönetim literatüründe, KSS raporlaması özellikle önemli bir araştırma konusu ve KSS davranışını anlamak için büyük bir kaynak haline geldi. Son on yılda, KSS raporlamasını analiz etmeye yönelik geniş ilgiye rağmen, pek çok araştırmanın içerik analizi yöntemiyle yapılması sebebiyle bu alandaki analiz yöntemleri sınırlı kalmaktadır. Bu tezde, 2007-2020 yılları arasında BİST şirketlerinin yıllık faaliyet raporlarını incelemek için yeni bir yöntem olan bir metin madenciliği yaklaşımını izledik. İlk olarak yıllık faaliyet raporlarından anahtar kelimeler çıkararak bir ESG sözlüğü hazırladık ve her rapor için toplam çevre, sosyal ve kurumsal yönetişim puanlarını hesapladık. Birincil sonuçlar, kurumsal yönetişim ile ilgili bilgilerin tüm ESG kategorileri arasında en yüksek görünürlüğe sahip olduğunu ve çevre görünürlüğünün yukarı yönlü bir artış eğiliminde olduğunu gösterdi. İkincil bir görev olarak, tüm raporları ESG görünürlüklerine göre gruplamak için k-medoids ve hiyerarşik kümeleme olmak üzere iki farklı kümeleme algoritması kullandık. Analizimiz, 3 farklı rapor grubu tespit etti ve 2020'de çevre puanları yüksek olan grubun payının önemli ölçüde arttığını ortaya koydu.

ACKNOWLEDGEMENTS

This thesis was completed with great support from a number of people without whom I would not have made it through my master's degree. I would firstly like to thank my thesis advisor Assoc. Prof. Dr. Ayşe Kocabıyıkoğlu for her academic and personal support during this project, and her sincere encouragement for my future academic endeavors. It was a pleasure and privilege to work with her. And special thanks to my co-advisor Asst. Prof. Dr. Celile Itır Goğuş for her valuable suggestions, thorough feedback and for providing expert guidance particularly on management and organizations related subjects. I am also grateful to my colleague and friend Bardia Alizadeh Moghtader, for teaching me a lot on data science throughout this research project. Thanks to his contribution, valuable time and honest feedback, the project has improved substantially. Finally, deepest thanks to my mother, father, and sister for all the support they have shown me so far and for being there for me all the time. And my dear husband who always made me feel confident in my abilities and without whom I wouldn't have dared to step into an academic journey.

I dedicate my dissertation work to all living creatures that do and will inevitably suffer from the greed of humanity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

With environmental problems being more salient in daily life and corporations and consumerism being major causes of such problems, a growing body of literature from various disciplines has focused on exploring the antecedents and consequences of corporate actions and seeking socially responsible alternatives. In line with this expanding interest, corporations increasingly started to report activities they undertake that do not solely entail financial interest but are mostly related to their impact on the society, otherwise known as their Corporate Social Responsibility (CSR) activities. With the establishment of widely recognized CSR reporting standards such as Global Reporting Initiative's (GRI) G3 Standards, AccountAbility's AA1000 Series, and the United Nations (UN) Global Compact's Communication on Progress (COP), CSR reporting practices have flourished, and corporations have been using various media channels to communicate their CSR practices to their stakeholders.

In the business and management literature, CSR reporting has particularly become a major research interest and a great source for understanding CSR behavior of companies. Initial attempts of CSR reporting research focused on the extent and nature of CSR disclosure within corporate annual reports and were followed by various research topics such as CSR's role in financial performance and corporate reputation, determinants of CSR reporting and different industries' CSR reporting practices (Khan et al., 2020). Today, researchers use not only corporate annual reports but also standalone CSR reports as well as websites and social media platforms of companies to analyze their CSR practices.

Despite the wide interest in analyzing CSR reporting in the last decades, the range of methods for analysis remain narrow, mainly dominated by the widely used content analysis method. Although both qualitative and quantitative content analysis methods are used for studying CSR reporting practices, and quantitative content analysis is becoming more and more common (Aureli, 2017), research employing the quantitative content analysis approach mostly analyzed limited samples and failed to cover a wide range of industries and companies as trained coders must read data in person which takes considerable amount of time. Besides, to the best of our knowledge, researchers mostly employed qualitative content analysis for their studies on CSR reporting in developing countries (Aggarwal & Singh, 2018; Ahmad et al., 2017; Gao, 2011; Kemp & Vinke, 2012). Using a text mining approach, the present study aims to demonstrate quantitative results portraying the CSR disclosure of Turkey representing an example for developing nations.

The structure of this thesis is as follows:

- Following the introduction, chapter 2 reviews the literature on CSR reporting. Section 2.1 briefly talks about the background of CSR acting as a starting point for CSR reporting. In Section 2.2, the origins of CSR reporting, its expansion and the historical development of international reporting standards are described. We review the literature on research analyzing CSR disclosures and outline limitations of existing studies in section 2.3, leading into the present study and the research questions which are introduced in section 2.4.

- Chapter 3 reports the empirical setup and describes the data employed for the study. In section 3.1, we introduce Borsa Istanbul (BIST) as the starting point of our data source and the regulations and practical implications of CSR reporting in Turkey. After setting the scene for analysis, in section 3.2, we describe the data used for the present study with the data collection, dictionary construction and data pre-processing steps. We close the chapter by defining the final variables to be used in the analysis.

- Chapter 4 lists the initial analysis conducted on the clean data and displays the descriptive results. It begins by presenting dictionary-based text analysis, the fundamental methodology adopted for the analysis of reports in this study in section 4.1. Next, section 4.2 illustrates the descriptive results gathered from the initial analysis of data.

- Chapter 5 introduces the methods used for grouping the initial results and reports the secondary results of the study. K-means, k-medoids and hierarchical clustering algorithms are summarized in section 5.1. Then, the outcomes of these applied methods are presented in section 5.2.

- Finally, in Chapter 6, the results of the empirical research are discussed addressing the research questions of the study. In addition, we explain the limitations of the present study and provide recommendations for future research.

2. Literature Review

2.1 Background of Corporate Social Responsibility

Corporate contributions, mostly in the form of philanthropy, were major practices leading to CSR as a concept, which began to take form in the 1950s (Carroll, 2008). Prior to the 1950s, criticisms about the emerging factory system in UK lead to the industrial welfare movement and individual philanthropy of businessman started to appear. However, socially responsible business behavior was still not the common practice in the increasingly corporate period (Carroll, 2008). Howard Bowen, who is considered as the father of CSR, used the term Corporate Social Responsibility for the first time in his book "Social Responsibilities of the Businessman" and defined the obligations of businessman towards the society (Bowen, 1953). From there on, CSR activities incrementally grew throughout the years by introducing the concept of social contract between businesses and society in 1970s (Carroll, 2008), companies starting to consider the social consequences of their actions in 1980s, CSR starting to become a widespread act by business schools introducing ethical education into their training for entrepreneurs in 1990s (Rodriguez-Gomez et al., 2020). By 2000s, CSR movement has become a global phenomenon (Carroll, 2008). Today, various companies have integrated CSR as an essential strategy for the sustainability of their businesses in line with the UN 2030 Agenda for Sustainable Development (UNDP, n.d.).

2.2 CSR Reporting and International Standards

The origins of research on non-financial corporate reporting dates back to 1970s when reporting on corporate contributions to society began to emerge (Fifka, 2013). Back then, particularly the multinational corporation (MNCs) were on spotlight for responsibility disclosure and were pressured for information disclosure both by supranational bodies like the UN and OECD, and the local governments and societies where MNCs operate (Gray et al., 1990). These disclosures initially started as part of the regular corporate annual reports, whereas businesses from western Europe soon began to publish stand-alone reports disclosing social responsibility information in the second half of 1970s (Fifka, 2013). While in 1970s, the content of reporting conveyed mostly social issues, the attention was shifted towards environmental issues in 1980s and businesses replaced their social reports with environmental ones (Fifka, 2013). In 1997, John Elkington published his book “Cannibals with Forks: The Triple Bottom Line of 21st Century Business” and defined the three bottom lines of sustainability for a corporation namely economic, social and environmental, an approach still widely used in evaluating corporate performance (Elkington & Rowlands, 1999). It was only in 2000s that businesses merged these two subjects, usually following the newly developed triple bottom line approach and issued reports called Sustainability Report, Corporate Social Responsibility Report, and Corporate Citizenship Report (Fifka, 2013). Current reports aiming to disclose corporate responsibility information cover a long list of topics ranging from social and environmental issues to health, sustainability, and philanthropy (Tschopp & Nastanski, 2014). Today, disclosure on social and environmental practices have flourished for companies from all sorts of regions, industries, and sizes as there is an immense interest of governments, non-governmental organizations, consumers, and all other kinds of stakeholders in footprints of corporations.

Addressing the growing interest in responsibility disclosure, various organizations worldwide started to design CSR reporting standards intending to establish a unified approach to CSR reporting. As an early attempt, The Coalition for Environmentally Responsible Economies (CERES) formed by social investment professionals and huge environmental groups of the time in the USA released a set of ten principles called “Valdez principles” named after the catastrophic Valdex spill, encouraging companies to address the impact of their products and production processes on the employees, the society, and the environment (Feder, 1989). Later, the same corporation, with involvement of the UN Environment Programme, launched the Global Reporting Initiative (GRI), to create an accountability mechanism that would ensure companies abide by responsible social, economic and governance principles. In 2000, the first version of GRI guidelines was launched by GRI. Since then, GRI have been publishing revised versions of GRI guidelines as well as new standards for different subject matters like Tax Standards, Waste Standards, Sustainability Reporting Standards etc. (GRI, n.d.). Another institution formed to encourage transparency in corporate reporting is the Institute of Social and Ethical Accountability. The institute published the AA1000 Assurance Standard in 1999 aiming to provide a basis for improving the sustainability performance of organizations (Jose, 2017). Soon after AA1000, the UN launched the UN Global Compact in 2000. It was designed as a guide for building socially responsible action and reporting (Tschopp & Nastanski, 2014). Consisting of ten principles covering issues relating to human rights, labor, the environment, and anti-corruption, the UN Compact has partnered with GRI and recommended reporting through GRI or any other method of reporting that addresses its principles. A more recent addition to the responsibility reporting standards is the Integrated Reporting (IR) Framework, developed by The International Integrated Reporting Council (IIRC) which was formed as a global coalition of regulators, investors, companies, standard setters, the accounting profession, academia, and NGOs. The initial IR framework was published back in 2013 to provide an alternative approach to corporate reporting by communicating all factors affecting the value creation process of an organization (Integrated Reporting, n.d.).

While these are some of the most widely recognized international standards, there are hundreds of other reporting guidelines, both on domestic and international level (Tschopp & Nastanski, 2014). A common feature of all these reporting standards is that they report the environment, social and governance (ESG) related matters for corporations. While there is no "best" framework agreed upon and used by all corporations, GRI is the most widely adopted standard for sustainability reporting (Bose, 2020).

2.3 CSR Reporting Analysis

As the reporting practices of companies and international reporting frameworks extended gradually, so did the research on CSR reporting. Scholars from all over the world began to investigate CSR's relationship to various subjects such as regulations, standards and certification, firm reputation, financial performance, customer satisfaction, stakeholder salience, firm environment, industry, alignment with firm mission and values, firm structure, firm size and so on employing CSR reports of various kinds (Aguinis & Glavas, 2012). Likewise, scholars conducted research analyzing different aspects of CSR reporting practices of companies in Turkey. However, most of these studies were either concentrated on a certain industry (Kiliç et al., 2015), limited in number of reports analyzed (Aksoy et al., 2020; Hoştut & Hof, 2014; Şahin et al., 2016) or focused on a narrow time interval (Ertuna & Tukel, 2010; Yüksel et al., 2008).

With the accumulation of available data sources, along with researchers from other research areas, scholars interested in CSR disclosure started to make use of advanced computational methods to analyze large amounts of data rapidly. Many researchers used text mining techniques to gather descriptive results from annual reports, CSR reports or sustainability reports. For instance, Goloshchapova et al. (2019) performed topic modelling (LDA) on CSR reports of publicly listed companies in Europe and the UK to discover the common CSR-related topics disclosed by these companies. The results of the study on more than 4.000 reports from 1999 to 2017 revealed that while CSR topics are sector dependent, "employees safety," "employees training support," "carbon emission," "human right," "efficient power," and "healthcare medicines" are the common topics reported by publicly listed companies in Europe and the UK. Similarly, Székely and Vom Brocke (2017) analyzed 9.514 sustainability reports retrieved from GRI website and published between 1999 and 2015 to identify common topics and practices related to sustainability. They found out 42 topics related to sustainability and reported overall observations for environmental, social, and economic sustainability categories. Ning et al. (2021) also gathered data from the GRI database and identified three themes from the topic modeling analysis: environment, social, and governance (ESG).

In addition to these industry-wide analysis, some studies focused on particular industries for CSR disclosure analysis. Li and Zhao (2021) described the key themes of detailed practices disclosed in CSR reports of global fashion companies using a dictionary-based text analysis method supported with LDA. Investigating the CSR reports of 24 top fashion companies in developed countries and 5 fashion companies in developing countries, the study revealed common environmental and social topics adopted by fashion companies. Results demonstrated that waste management and human rights are the top two most popular themes in fashion companies' CSR reports and that companies within the same product categories tend to follow similar sustainability practices. Cai et al. (2021) analyzed the Environmental information disclosure (EID) of China's heavy pollution industry from 2013 to 2017 and found that the overall quality of EID for these companies is low and that there are differences in EID quality between the 16 heavily polluting industries. Sustainability trends and practices in the 4 main sectors of process industries, namely oil/petrochemicals, bulk/specialty chemicals, pharmaceuticals, and consumer products sectors, were identified by Te Liew et al. (2014). Applying text mining on the sustainability reports of largest companies listed in the Forbes ranking of the global top 2000 companies in 2011, the study found out that all sectors have similar top focuses in terms of sustainability: health and safety, human rights, reducing GHG, conserving energy/energy efficiency, and community investment. Besides, the study showed that environment is the predominant sustainability topic in the process industries. Liao et al. (2017)'s systematic content analysis followed a similar method to text mining and analyzed the CSR communication of companies operating in the construction industry in 4 different regions: Asia, European Union (EU), US/Canada and China. The study collected the CSR-related reports of a sample of 310 international contractors from the Engineering News-Record list from 2009 to 2014 and revealed important differences between the levels of CSR communication of companies in different regions. According to the study, European contractors hold the highest levels of CSR communication, while Chinese contractors rank the lowest in CSR communication.

Descriptive studies are not the sole type of analysis using text mining to gather results on CSR disclosure. Some studies analyzed CSR's relationship to financial and operational performance. Myšková and Hájek (2019) investigated the financial and CSR-related information published in the annual reports of 1.380 listed US companies, aiming to uncover a relationship between their CSR activities and financial performance. The study revealed a correlation between financial results of the companies and the number of information companies presented concerning CSR and linked worse financial results to less information on CSR. Another study by Lee and Huang (2020) developed a model for operating performance forecasting by assessing the relationship between corporate operating performance and CSR. The study conducted LDA analysis to group CSR into numerous dimensions and measured each dimension's influence on a firm's financial performance so that managers can make efficient resource allocations.

Although these examples support increasing interest in advanced computational methods for CSR disclosure analysis, the number of studies conducted in such manner is far from optimal. In fact, studies analyzing or making use of CSR reports reported many different limitations. Khan et al. (2020)'s systematic literature review on CSR reporting research outlined the common limitations of CSR reporting research. To begin with, the study revealed that 22% of studies sampled an insufficient number of companies which is an obvious threat to the generalizability of the research results and that 16% of studies use a single source of data to examine. Furthermore, a short time period and similar type of companies were examined in the majority of the articles. In addition to these limitations on sample characteristics and sources of data, a frequently reported methodological limitation was the issue of subjectivity in applying the content analysis technique. Khan et al. (2020) found out that 80% of the studies examine CSR reporting activities using content analysis. Previously, Ali et al. (2017)'s and Fifka (2013)'s literature reviews also revealed that content analysis is the most commonly used method for analyzing CSR content in both developed and developing countries. The analysis has also detected the need to enhance research towards different geographies, particularly to developing nations (Khan et al., 2020).

2.4 The Present Study

To address some of the limitations of previous studies, the present study investigates the CSR reports of all companies listed on BIST from 2007 to 2020 using a text mining approach. The study reviews a large number of reports published by companies operating in different industries. Adopting a novel computational approach, the study aims to capture a snapshot of companies' CSR activities and to portrait a developing nation exemplar for CSR in a longitudinal and objective manner. The overarching research questions (RQ) of our study are:

(1) For BIST companies, how do the ESG categories compare?

(2) How are BIST companies grouped based on the ESG disclosures they make?

(3) What changes did these groupings undergo over time?

In the next chapter, we firstly introduce the empirical setting in which we addressed our research questions and describe the data employed for the study and its preprocessing steps.

3. Empirical Setting & Data

3.1 Borsa Istanbul and CSR Reporting in Turkey

3.1.1 Borsa Istanbul

Borsa Istanbul (BIST) was formed as an exchange entity in 2013 to combine Istanbul Stock Exchange (IMKB), Derivatives Exchange of Turkey (VOB) and Istanbul Gold Exchange (IAB) under one roof (Borsa Istanbul, 2022b). Currently its largest shareholder is Turkey Wealth Fund by 80.6%. The remaining shareholders are: 10.0% QH Oil Investments LLC, 1.3% Turkish Capital Markets Association, 2.32% Borsa Istanbul A.Ş. (acquired shares from shareholders) and 5.78% other corporations (intermediaries, banks, foreign exchange companies etc.) (Borsa Istanbul, 2022a). For our research, we collected data on companies that are traded on BIST as these companies are obliged to publicly disclose information regarding their financial and operational activities under Turkish Commercial Code (*Turkish Commercial Code, No.6102*, 2011). Information on these companies is provided by the Public Disclosure Platform (KAP) where we gathered the list of companies for our analysis (Public Disclosure Platform, 2022).

3.1.2 CSR Reporting in Turkey

Effective from 2005, the EU obligated its member countries to comply with International Financial Reporting Standards (IFRS), a set of rules that describe the way companies prepare their financial statements. This obligation resulted in many countries outside of the EU adopting these principles as well (Larson & Street, 2013). Turkey, in accordance with these developments, established Turkish Financial Reporting Standards (TFRS) as a translated version of IFRS (*Türkiye Finansal Raporlama Standartlarının ilk Uygulamasına ilişkin Türkiye Finansal Raporlama Standardı (TFRS 1) Hakkında Tebliğ*, 2006). In addition to these rules, Turkish Commercial Code No. 6102 was established on 2011 and required all enterprises, publicly traded or not, to prepare their financial statements in accordance with TFRS (PWC Turkey, 2011).

Despite the conformity of Turkish accounting standards with international ones, Turkish regulations on disclosure of environmental, social and governance (ESG) related matters are unparalleled to the international standards. There is no set-in stone international framework referenced by all companies traded on BIST, yet a lot of Turkish companies who disclose non-financial information follow the GRI guideline, a well-known international sustainability reporting standard, to report their ESG performance (Belverd et al., 2019).

More recently, IIRC's Integrated Reporting Framework have also gained importance as a new framework for publishing information about the relationship between companies' activities and significant financial and sustainability issues (Integrated Reporting Türkiye, 2022). On November 2017, IIRC and BIST signed a cooperation agreement to disseminate the publication of integrated reports in Turkey (Aras et al., 2019). Another attempt of BIST to encourage corporations on publishing non-financial information was forming the BIST Sustainability Index, which was launched on 2014 with the code XUSRD. The Index is aimed at reflecting companies' approach to various sustainability issues such as global warming, draining of natural resources, health, security, and employment. However, the number companies represented at the index are limited. Since 2014, only the companies traded on BIST 30 index have been evaluated for the BIST Sustainability Index. Starting from 2022, companies can voluntarily share data on their ESG practices with Refinitiv, the intermediary corporation assessing data for BIST Sustainability Index based on the international sustainability criteria.

In addition to BIST's attempts at encouraging firms to be transparent about their ESG practices, Capital Markets of Turkey (SPK) published a communique that requires all publicly traded companies to report compliance with corporate governance principles in 2014 (*Kurumsal Yönetim Tebliği (II-17.1)*, 2014). These principles, however, are predominantly governance related and binary.

3.2 Data

3.2.1 Data Collection

Numerous studies in the literature explore annual reports of companies to understand their CSR disclosure patterns (Ashcroft, 2012; Nobanee & Ellili, 2016; Qi et al., 2012; Sobhani et al., 2012). Using a similar approach, we collected corporate annual reports of companies traded on BIST from 2007 to 2020 for our study. We started the data collection process on 2021 and once all the reports on the company list previously gathered from KAP were collected, we reviewed the existing list of companies on BIST and based our final analysis on 551 companies traded on BIST on April 6, 2022.

The data were collected from the individual websites of companies in pdf format and were all written in Turkish. These reports convey information regarding the financial performance of the companies as well as the operational and social activities they engage in throughout the year. During the data collection process, we applied a preselection criterion to the data and eliminated those companies who published reports that does not contain any information about CSR activities by searching for keywords that are often used in CSR communication such as “social responsibility”, “donation” and “sustainability”. We also eliminated those pdf reports that are non-readable due to their pdf format as the text is not selectable and those that have less than 1000 words which is a low benchmark for a set of reports with 27.000 words length on average. Finally, we ended up with 3079 reports belonging to 338 unique companies.

While collecting the annual reports, we simultaneously filled a document-year matrix where rows represent company names and columns represent years. For each company, we marked the value cell as 1 if we collected the annual report corresponding to that particular year. For each company, in addition to the year variables, we also collected information on the industry in which the company is operating from KAP (Public Disclosure Platform, 2022).

3.2.2 Keyword Selection and Dictionary Construction

To apply dictionary-based text mining to our data, which will be explained further in the methods section, we constructed a dictionary with keywords in three different categories namely environmental, social and governance representing ESG categorization, an approach widely used in the literature to evaluate the social impact of a corporation on stakeholders beyond its financial performance. To the best of our knowledge, there is no pre-defined Turkish CSR dictionary. For this reason, to select the keywords for our ESG dictionary, we decided to take our own data as a basis. Firstly, we counted the number of words for each document in our data. Then, for each year, we selected two reports randomly from our data: each being within two standard deviations above and below the mean file length, respectively. Using this method, we were able to form a report sample that would represent both short and long reports for each year. We applied content analysis method to our sample and manually extracted keywords for the ESG categories. After forming the initial draft of the dictionary, we shortened the keywords by eliminating suffixes and removed repeating and unnecessary words. We then reviewed the CSR dictionary previously developed by Pencle and Mălăescu (2016) in English, publicly available at (https://provalisresearch.com/Download/CSR.zip) and made some additions to our dictionary of keywords. We covered climate change, carbon emissions, pollution, biodiversity, pollution, deforestation, energy efficiency etc. topics in the environmental category; gender, diversity, human rights, labor standards and rights, employee engagement, customer rights and satisfaction etc. topics in the social category; and board composition, executive compensation, corruption, and audit etc. topics in the governmental category. In total, 318 keywords were selected, 70 being in the environmental category, 153 being in the social category and 95 being in the governance category. We also categorized our keywords into N-grams, a sequence of n words, as 188 unigrams, 117 bigrams and 13 trigrams (See the Table A.1 in Appendix).

3.2.3 Data Pre-Processing

After we collected our data of corporate annual reports in pdf format, we applied the following preprocessing steps to our data prior to conducting initial analysis. Firstly, we used the Tika library in Python to parse our pdf documents into a text format ready for further analysis. We then employed the Natural Language Toolkit (NLTK) library in Python for most of the preprocessing operations. We began the work by tokenizing the corpus initially into sentences and then to words. We then transformed whole corpus of words into lowercase and applied stop word elimination to remove most frequently used words in Turkish language using NLTK's list of Turkish stop words. We also removed punctuation marks in the corpus and eliminated white spaces and words with single letter. Finally, we achieved lists of proper and clean words for each document.

Before conducting primary analysis on the list of words, we applied stemming, a method of reducing words into their word stems by chopping off the ends of words (Schütze et al., 2008), to each word in the list and produced a final list of stemmed words. By using this method, we ensured getting rid of affixes in each word and hence improving the performance of our retrieval task.

We decided to apply stemming instead of lemmatization, a similar technique of reducing words into a common base but instead of removing derivational affixes right away, taking the use of a vocabulary and morphological analysis of words into account (Schütze et al., 2008), due to the computational burden of lemmatization process given the large number of documents we aimed to analyze. Finally, we applied the same technique to our dictionary of keywords and produced a dictionary of stemmed keywords so that we can match words properly during word extraction. We reviewed the stemmed versions of keywords and removed keywords that were reduced to identical stems with another keyword. We also made additions to the keyword list by producing derivations of some keywords that we found useful for the analysis. The final dictionary of keywords selected along with their stemmed versions categorized in ESG categories and as unigrams, bigrams and trigrams can be found in Appendix Table A.1.

3.2.4 Final Variables

Once all the pre-processing steps were applied on the data, we assigned frequencies to each word by scanning each document and produced a document-term matrix. For repeating words, we calculated the frequency of unigrams by deducing the frequencies of bigrams and trigrams containing the repeating word. Once we had accurate frequencies for each word sequence in our dictionary, we constructed "environment", "social" and "governance" variables by summing all word frequencies in corresponding categories. Since the number of keywords searched in each category and the length of each report varied significantly, we scaled each variable by dividing the total frequency with the corresponding file length and the number of keywords searched in that category. Finally, we multiplied the outcome with 10.000 to achieve the number of keywords found in 10.000 words for each keyword searched. For instance, if the length of a file is 1000 words and the total frequency of keywords in the environment category is 10, we would calculate the environment score as (10/65/1.000) *10.000= 1,53. This would mean that for 1 environment keyword searched in the corpus, we would find 1,53 words in each 10.000 words.

In addition to these three variables, we added "file length" "industry" and "year" variables for each report in our data set to be further used in descriptive and clustering analysis. In the next chapter, we explain how we used these variables for our core method of dictionary-based text analysis. We then present our initial results and provide an answer to the RQ1.

4. Descriptive Methods & Results

4.1 Dictionary-Based Text Analysis

Exploring textual data has become much more automated and rapid with the development of computational linguistic analysis methods. Today, as vast amount of data is easily accessible, many researchers prefer automated text analysis over manually inspecting the data.

Essentially, text analysis approaches can be branched into two categories as closed-vocabulary and open-vocabulary approaches. Attempts of automated text analysis started with developing an algorithm following closed-vocabulary approaches at Harvard University. The goal was to discard the intuition aspect of the content analysis method and create a truly objective and systematic algorithm that would yield quantitative results (Stone & Hunt, 1963). This approach is also often called the dictionary-based approach as researchers build dictionaries of keywords to represent certain categories they aim to identify in the text. Researchers may prefer to employ existing dictionaries as building a new dictionary is time-consuming and demands extensive domain knowledge (Lowe, 2003). There is a wide range of dictionaries prepared in different languages, categories, and scope of coverage for different research areas. For instance, the Linguistic Inquiry and Word Count (LIWC) dictionary is the most commonly used tool for text analysis in the psychology literature. (Eichstaedt et al., 2021). Dictionaries of various other domains such as policy agendas, corporate philanthropy, characteristics of writing style, job description, aviation safety exist in the literature (Deng et al., 2019). Despite of the broad range of domains covered by existing dictionaries, the generic nature of these tools is often deficient in addressing the needs of a domain specific corpus and need adaptation to the changing meanings of words over time and space (Deng et al., 2019). Besides, some special-purpose dictionaries are simply not available in certain languages. Therefore, researchers often need to develop a dictionary well-suited for their research purpose and specific corpus. Deng et al. (2019) defined 3 distinct dictionary-building approaches: 1) manual; 2) semi-automatic; and 3) automatic and differentiated them by the level of automaticity employed to complete three core steps in the dictionary building process: 1) developing categories, 2) identifying entries and 3) categorizing entries. If the steps followed through the process are purely manual or automatic, the corresponding approach would also be viewed as purely manual or automatic. While manual approaches require a broad domain knowledge and are considered theory-driven, automatic approaches depend highly on programming knowledge and are by nature data-driven. When the researchers utilize their own judgements together with the assistance of text analysis software, it is considered a semi-automatic approach, which is the most-widely adopted dictionary building method. Each of these approaches has its pros and cons; therefore, researchers must consider the final aim of the research study before choosing the suitable one (Deng et al., 2019).

Once the dictionaries are built, the automated computer programs are directed to scan the preprocessed corpora to search for those keywords and assign frequencies to relevant categories (Eichstaedt et al., 2021). The results can be used by researchers in various ways in compliance with the research objective. For instance, the frequencies assigned to the categories can be employed as independent variables in a regression or classification problem.

Closed-vocabulary approaches, although very useful to focus on a specific domain, solely consider a limited number of words some of which may not even be present in the corpora and therefore result in a high dimensional space of vectors. Open-vocabulary approaches overcome this shortcoming by analyzing all words in the corpora and identifying semantically related clusters of words and thus lowering the dimensional space. However, they require more computational knowledge in their implementation, demand larger data sets, and are more complex to use than the closed-vocabulary programs (Eichstaedt et al., 2021).

Overall, automated text-analysis enables researchers to process vast amounts of data rapidly which is a distinctive advantage over the long-lasting manual content analysis. However, researchers must be aware of the disadvantages that comes with these methods and blend individual judgement with computational efficiency where needed.

Our study followed a closed-vocabulary approach. By their nature, annual reports contain information about financial and operational activities of companies that is irrelevant to our topic of interest. Therefore, to ensure that our results report ESG scores of companies, we preferred a dictionary-based approach over topic modelling. We built a custom ESG dictionary for our research purpose in a semi-automatic manner as explained in detail in the Empirical Setting and Data section, and the results yielded 3 variables namely environment, social and governance scores for each report in our data, to be further employed in clustering analysis.

4.2 Descriptive Analysis and Results

Prior to employing different clustering algorithms on our data, we applied descriptive analysis to get a primary understanding of the whole data set. In this section, we provide information on the initial results of our dictionary-based text analysis, for the whole data set as well as per industry and year.

Figure 4.1 and Figure 4.2 below illustrate how 3079 reports in our data set are distributed over the years and industries. Unsurprisingly, Figure 4.1 shows that the number of reports in our data set increases gradually each year. We applied a selection criterion concerning the ESG content of reports during the data collection process and as awareness and appreciation on ESG matters have an upward trend in the last few decades, this is not an unexpected result. Figure 4.2 exhibits the shares of industries represented in our data set. Nearly 70% of the reports belongs to production companies and financial institutions. Wholesale and retail trade companies, restaurants and hotels, holding companies and technology companies are next in line top industries in terms of the number of reports published.

In addition to the number of reports, we also reviewed the distribution of average file length over industries and years. Figure 4.3 illustrates the change in average file length in years per each industry. Looking at the figure, we can clearly see that most industries have published longer reports in the recent years and that holding companies and financial institutions produce longest reports among all. It is also interesting to see that companies operating in the technology industry and professional, scientific, and technical activity industries have started to publish significantly longer reports in the last few years.

To find out how ESG categories rank among each other and answer RQ1, for each category and report we calculated the average number of keywords found per 10.000 words for a single keyword searched in that category. The governance category scored top of the list by 3.5 keywords on average, and it was followed by the en-

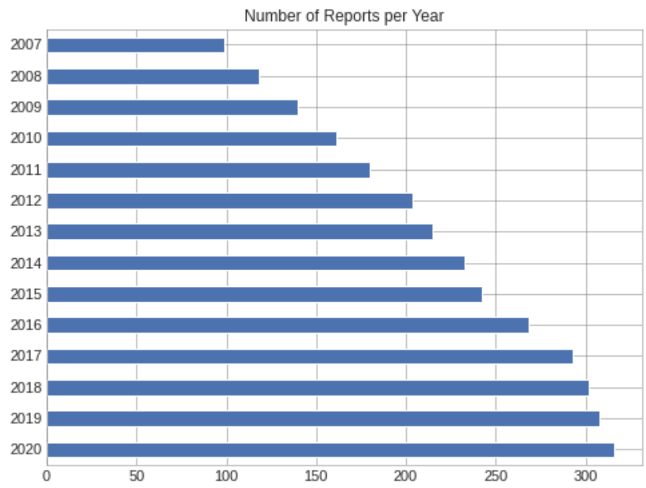


Figure 4.1 Number of Reports per Year

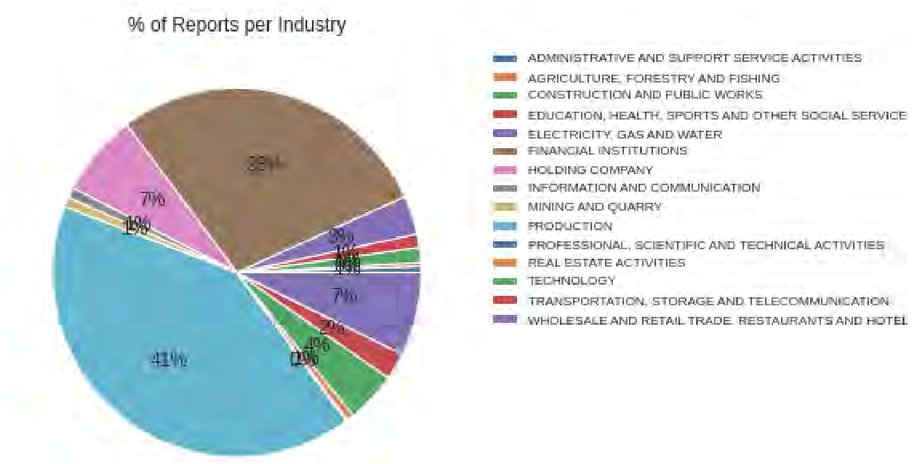


Figure 4.2 Percentages of industries represented in the data set

vironment category with 0.7 keywords. The social category ranked lowest with 0.5 keywords. Figure 4.4 and Figure 4.5 displays the distribution of average ESG scores over years and industries respectively and Table 4.1 and 4.2 reports the summary statistics for the same metrics. Overall, we can see that the salience of governance keywords increased substantially between 2008 and 2018 while keeping steady rates in the last two years. In contrast, environment keywords kept nearly steady in the first 11 years and almost doubled between 2018 and 2020. Again, this result looks like an anticipated impact of environmental issues being a major discussion subject and corporations being held more and more responsible for their environmental

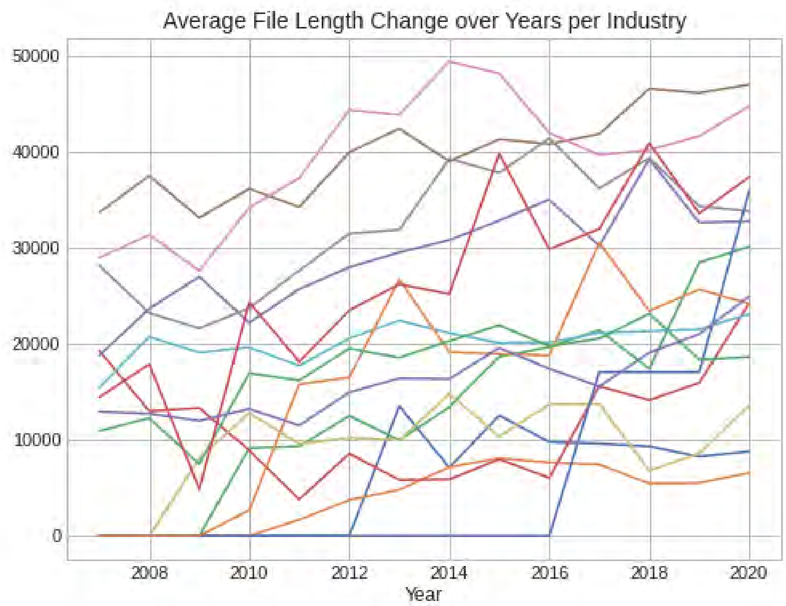


Figure 4.3 Change in average file length in years per each industry

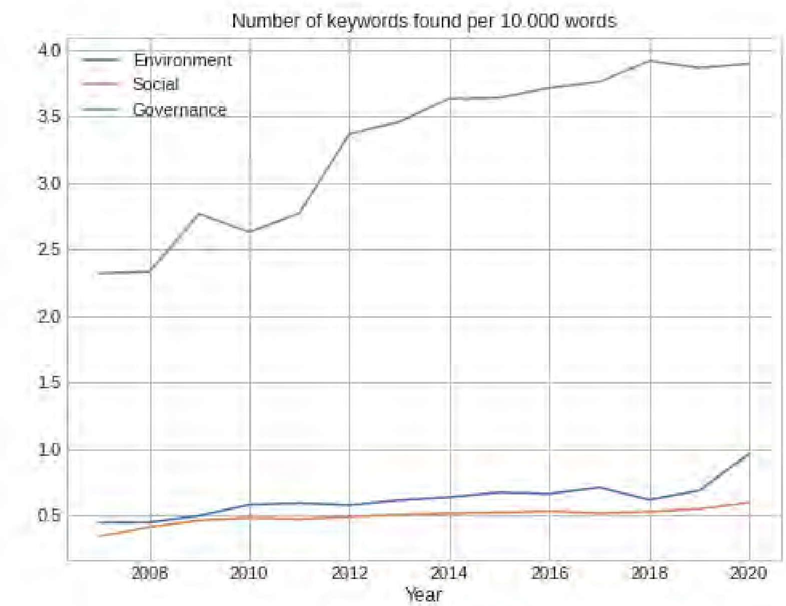


Figure 4.4 ESG scores per industry

footprints in the last couple of years. Looking from an industry angle, governance category takes the lead in all industries but one: Professional, Scientific and Technical Activities. Although environment score for this industry exceeds governance scores, we must say that looking at the figure, one should not reach to overarching conclusions for the whole industry since there is only a single engineering company represented for this industry in our data set. The next industry with a high envi-

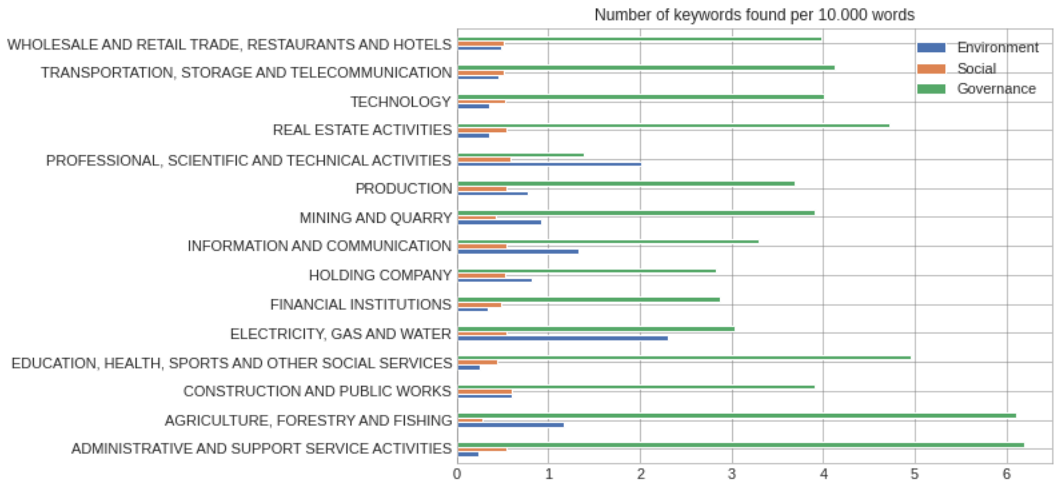


Figure 4.5 ESG scores per year

ronment score, is Electricity, Gas and Water. Based on the nature of the businesses in this category, this result shouldn't come as a surprise as companies in these industries tend to report more on environment related matter as part of daily business activities. Social scores, unlike governance and environment scores, are distributed somewhat uniformly among different industries and range between 0 and 2. Overall, although the average scores for environment and social categories are somewhat close, we can see that social scores lie on a narrower range and environment scores have much more variance.

		2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Environment	min	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	max	4,66	4,87	5,04	4,84	5,13	5,03	4,90	5,37	5,13	4,91	25,15	4,18	6,29	5,52
	mean	0,45	0,45	0,50	0,58	0,59	0,58	0,62	0,64	0,67	0,66	0,71	0,62	0,69	0,97
	median	0,28	0,28	0,32	0,38	0,39	0,40	0,41	0,47	0,47	0,46	0,44	0,45	0,46	0,64
	std	0,59	0,52	0,57	0,65	0,65	0,61	0,63	0,60	0,61	0,65	1,55	0,58	0,74	0,94
Social	min	0,04	0,09	0,00	0,13	0,00	0,00	0,08	0,00	0,00	0,11	0,00	0,05	0,17	0,18
	max	1,08	1,57	1,86	1,74	1,56	1,43	1,57	1,57	1,48	1,53	1,61	1,53	2,05	2,02
	mean	0,35	0,41	0,47	0,48	0,47	0,49	0,51	0,52	0,52	0,53	0,52	0,53	0,55	0,60
	median	0,30	0,38	0,43	0,46	0,45	0,47	0,48	0,48	0,49	0,49	0,47	0,49	0,50	0,54
	std	0,17	0,22	0,23	0,19	0,22	0,19	0,20	0,21	0,21	0,22	0,22	0,21	0,23	0,26
Governance	min	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,66	0,00	0,01
	max	7,24	7,61	8,54	8,40	8,94	8,73	9,28	8,26	7,76	8,83	9,03	9,03	9,11	9,23
	mean	2,32	2,33	2,77	2,63	2,77	3,37	3,46	3,63	3,64	3,71	3,76	3,91	3,86	3,89
	median	2,06	1,89	2,29	2,08	2,32	2,79	2,78	2,99	3,06	3,14	3,12	3,21	3,38	3,44
	std	1,28	1,49	1,75	1,61	1,63	1,93	1,86	1,89	1,87	1,94	1,96	2,02	1,98	1,87

Table 4.1 Summary statistics of ESG scores per year

		Administrative And Support Service Act.	Agriculture, Forestry, Fishing	Construction and Public Works	Education, Health, Sports	Electricity, Gas, Water	Financial Inst.	Holding	Information and Commun.
Environment	min	0,05	0,74	0,11	0,07	0,44	0,00	0,00	0,65
	max	0,64	2,47	1,23	0,57	25,15	4,00	3,98	1,73
	mean	0,23	1,17	0,61	0,25	2,31	0,34	0,82	1,32
	median	0,22	1,08	0,57	0,19	1,37	0,27	0,60	1,32
	std	0,17	0,51	0,34	0,13	2,73	0,31	0,76	0,31
Social	min	0,33	0,13	0,28	0,08	0,00	0,00	0,12	0,30
	max	0,68	0,35	1,29	1,25	1,57	2,02	1,35	0,80
	mean	0,54	0,27	0,60	0,44	0,53	0,48	0,53	0,54
	median	0,58	0,29	0,56	0,34	0,48	0,42	0,49	0,50
	std	0,11	0,07	0,21	0,31	0,24	0,25	0,20	0,16
Governance	min	3,73	3,11	0,60	1,88	0,21	0,00	0,00	2,08
	max	9,03	7,12	8,41	9,11	6,94	8,94	7,53	6,37
	mean	6,20	6,11	3,90	4,95	3,04	2,88	2,83	3,30
	median	6,05	6,43	3,66	5,34	2,64	2,45	2,28	3,20
	std	1,44	1,13	2,45	1,75	1,52	1,69	1,71	1,06

		Mining and Quarry	Production	Professional, Scientific and Technical Act.	Real Estate	Technology	Transportation, Storage, Telecomm.	Wholesale, Retail, Restaurants and Hotels
Environment	min	0,42	0,00	1,41	0,14	0,00	0,00	0,00
	max	2,19	5,52	2,22	1,09	2,10	1,74	3,44
	mean	0,92	0,78	2,02	0,36	0,35	0,45	0,48
	median	0,73	0,61	2,22	0,32	0,28	0,42	0,33
	std	0,49	0,61	0,41	0,20	0,26	0,31	0,51
Social	min	0,26	0,00	0,44	0,36	0,00	0,00	0,00
	max	0,53	2,05	0,62	1,09	1,20	1,01	1,13
	mean	0,42	0,54	0,58	0,54	0,53	0,51	0,51
	median	0,43	0,50	0,62	0,47	0,54	0,55	0,47
	std	0,06	0,22	0,09	0,18	0,17	0,21	0,18
Governance	min	1,06	0,00	0,94	2,43	0,00	0,84	0,00
	max	5,79	9,28	2,74	8,13	7,22	7,78	8,16
	mean	3,91	3,70	1,39	4,72	4,01	4,13	3,98
	median	3,40	3,01	0,94	3,67	4,58	4,63	3,14
	std	1,30	1,95	0,90	2,05	1,88	2,12	2,00

Table 4.2 Summary statistics of ESG scores per industry

Finally, we extracted the most frequent keywords searched in the whole data set to have an initial inference of top CSR related topics disclosed in the reports and to notice major changes throughout the years. Table 4.3 and Table 4.4 below portrait top 5 frequent keywords searched among all documents per year and industry respectively. In line with our previous takeaways, most of these keywords belong to the governance category. However, we can see that the word “energy” from the environment category ranked 5th on the list in 2019 and 2020 which shows that the category’s salience is improving. A glance at Table 4.4 too shows that for all industries, most of the keywords belong to the governance category. Nonetheless, keywords such as “energy”, “nature” and “environment” are included in the top 5 list for industries such as Construction and Public Works, Electricity Gas and Water and Information and Communication.

	1	2	3	4	5
2007	Yönetim kurulu	Denetim	Genel kurul	Kanun	Pay sahibi
2008	Yönetim kurulu	Denetim	Genel kurul	Kanun	Pay sahibi
2009	Yönetim kurulu	Denetim	Genel kurul	Kanun	Pay sahibi
2010	Yönetim kurulu	Denetim	Genel kurul	Kanun	Pay sahibi
2011	Yönetim kurulu	Denetim	Genel kurul	Kanun	Kurumsal yönetim
2012	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Kanun
2013	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Kanun
2014	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Pay sahibi
2015	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Pay sahibi
2016	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Pay sahibi
2017	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Pay sahibi
2018	Yönetim kurulu	Denetim	Kurumsal yönetim	Genel kurul	Kanun
2019	Yönetim kurulu	Denetim	Kurumsal yönetim	Genel kurul	Enerji
2020	Yönetim kurulu	Denetim	Kurumsal yönetim	Genel kurul	Enerji

Table 4.3 Most frequent keywords per year

	1	2	3	4	5
Administrative And Support Service Act.	Yönetim kurulu	Genel kurul	Komite	Denetim	Pay sahibi
Agriculture, Forestry, Fishing	Yönetim kurulu	Kurumsal yönetim	Genel kurul	Denetim	Pay sahibi
Construction and Public Works	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Enerji
Education, Health, Sports	Yönetim kurulu	Genel kurul	Denetim	Kurumsal yönetim	Hastane
Electricity, Gas, Water	Enerji	Yönetim kurulu	Denetim	Kurumsal yönetim	Genel kurul
Financial Inst.	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Kanun
Holding	Yönetim kurulu	Denetim	Doğanın	Genel kurul	Kurumsal yönetim
Information and Commun.	Yönetim kurulu	Doğanın	Denetim	Genel kurul	Kurumsal yönetim
Mining and Quarry	Yönetim kurulu	Genel kurul	Denetim	Kurumsal yönetim	Yetki
Production	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Pay sahibi
Professional, Technical and Scientific Act.	Enerji	Yönetim kurulu	Denetim	Yurt	Çevre
Real Estate	Yönetim kurulu	Denetim	Komite	Genel kurul	Kurumsal yönetim
Technology	Yönetim kurulu	Denetim	Kurumsal yönetim	Genel kurul	Pay sahibi
Transportation, Storage, Telecomm.	Yönetim kurulu	Denetim	Kurumsal yönetim	Genel kurul	Kanun
Wholesale, Retail, Restaurants and Hotels	Yönetim kurulu	Denetim	Genel kurul	Kurumsal yönetim	Pay sahibi

Table 4.4 Most frequent keywords per industry

In the next chapter, we present the final operations we applied to our data set prior to the clustering analysis as well as a number of decisions we made during the analysis. We then explain the clustering methods used to group the reports based on ESG disclosure and present overall and yearly results answering RQ2 and RQ3.

5. Clustering Methods & Results

5.1 Clustering Analysis

Clustering analysis is an unsupervised learning algorithm that is performed as part of an exploratory data analysis instead of a predictive task. Originally built at University of California, the algorithm was used to group blocks of cultures (Driver & Kroeber, 1932). It consists of a broad set of computational techniques to find subgroups, in other words clusters, in the data set. The goal of clustering algorithms is to group data in a way that the data points are similar within each group while data points in different groups are different from each other. In other words, it aims to build distinct homogeneous subgroups from the data set (Gareth et al., 2013). There are various clustering algorithms that follow different assumptions. For our study, we employed the two common clustering algorithms namely k-medoids clustering and hierarchical clustering to answer the rest of our research questions, RQ2 and RQ3. It is also worth mentioning that before applying these algorithms to our data, we normalized our variables using min-max scaling to range between 0 and 1 so that they lie on a common range and are comparable. We preferred normalization over standardization since some of our variables have outliers and do not follow a Gaussian distribution.

5.1.1 K-Means and K-Medoids Clustering Algorithms

K-means clustering algorithm is a partitioning method that divides the data into K distinct subgroups. It requires the researcher to define the hyperparameter K before clustering each data point into one of these K non-overlapping clusters. The objective of the algorithm is to minimize the within-cluster variation (Gareth et al., 2013). The mathematical representation of the objective function is as follows:

$$\min_{C_1,\ldots C_k} \sum_{k=1}^{K} W(C_k) \tag{5.1}$$

where C_k denotes cluster K and $W(C_k)$ denotes the amount by which the observations within a cluster differ from each other. In short, the objective is to create K clusters in the data set such that the total within-cluster variation, summed over all K clusters, is minimized. The objective function is subject to two constraints: 1) The clusters must ensure that each observation belongs to at least one of the K clusters. 2)No observation belongs to more than one cluster (Gareth et al., 2013).

There are different approaches to calculate the within cluster variation $W(C_k)$. The most common one is squared Euclidean distance. Using this distance metric, within cluster variation is calculated as the sum of all the pairwise squared Euclidean distances between data points within a cluster, divided by the total number of data point in the cluster. Rewriting the within cluster variation, the objective function becomes as follows:

$$\min_{C_1,\ldots C_k} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - xi'j)^2 \tag{5.2}$$

where p denotes number of dimensions (features). The algorithm works as follows to ensure that the within cluster variation decreases at each step:

1. Assign a random number, from 1 to K, to each of the data points as initial cluster assignments.

2. For each cluster, compute the cluster centroid as the vector of p feature means for all data points in the cluster.

3. Assign each data point to the closest cluster centroid.

4. Repeat Step 2 and Step 3 until cluster assignments no longer change.

There are several limitations to k-means clustering algorithm. Firstly, the initiation method of the initial clusters (Step 1) plays a vital role in the final results of the algorithm. Choosing arbitrary cluster centers might end up in a bad outcome in which clusters found by the algorithm do not represent the natural clusters in the data. One solution to this problem is to run multiple iterations with random initial centroids and compare the results. An alternative way was offered by Arthur and Vassilvitskii (2006). They proposed a novel approach of initiating cluster centers called k-means++ algorithm. The algorithm works as follows:

1a. Take one center C_1 , chosen uniformly at random from X.

1b. Take a new center C_i, choosing $x \in X$ with probability $\frac{D(x)^2}{\sum\limits_{x \in X} D(x)^2}$

1c. Repeat Step 1b until all K centers are chosen.

2-4. Apply steps in the standard k-means algorithm.

where $D(x)$ denotes the shortest distance from a data point to the closest cluster center. K-means++ algorithm tends to improve the quality of the clusters and lowers overall computational runtime.

Another challenge with the k-means algorithm is that the number of clusters (K) is an hyperparameter that must be set by the researcher. A widely used method to select the number of clusters in clustering is the elbow method (Thorndike, 1953). It is a heuristic method where within cluster variation is plotted as a function of number of clusters. The method takes its name from the elbow shape that occurs when the function is plotted. As the number of clusters increases, the within cluster variation will start to decrease until there are as many clusters as data points. The elbow point corresponds to the optimal number of clusters since the decrease in within cluster variation will cutback after this point. Another method that is widely used to determine the best value of K is Silhouette analysis, developed by Rousseeuw (1987). Silhouette analysis is an evaluation of cluster validity and for each data point in the data set, it assesses the average distance of the data point to the points in its cluster as well as the minimum average distance of the data point to the points in another cluster. It lies on a range of (0,1) and the closer it is to 1, the better clustered the data point. So, to determine the best value of K, a Silhouette plot displaying cluster validity for different numbers of K can be build.

A final drawback of the k-means algorithm is its' sensitivity to outliers. Since the cluster centers are computed by averaging features of data points in the corresponding cluster, the algorithm is highly susceptible to outlier data points. To overcome this issue, Kaufman and Rousseeuw (1990) proposed a new algorithm called k-medoids. This method is almost identical to k-means but instead of taking the average of data points, it assigns actual data points (medoids) as cluster centers. The algorithm of k-medoids works as follows:

1. Randomly assign K cluster medoids.

2. Assign each data point to the closest cluster medoid and compute the total distance of all data points in the cluster to the cluster medoid.

3. Randomly select another data point in each cluster and compute the total distance of all data points in the cluster to the new data point. If the sum of distances to the new data point is smaller than the sum of distances to the medoid, keep the new data point as medoid.

4. Repeat 2 and 3 until the medoids no longer change.

K-medoids algorithm, in contrast to k-means, is more robust to outlier points in the data set. As our yearly data consists of outliers, we applied k-medoids algorithm with k-medoids++ initiation and plotted an elbow graph and a silhouette plot to determine the number of clusters in each year. In the next section, we talk about hierarchical clustering algorithm, the second clustering method we applied to our data, and its results.

5.1.2 Hierarchical Clustering

Hierarchical clustering algorithm is an alternative to the k-means clustering algorithm, eliminating the need for a specified K. While applying hierarchical clustering to a data set, the researcher does not have to specify the number of clusters K in advance. There are two opposite forms of this clustering method: 1) a bottom-up (agglomerative) approach in which all data points are initially considered separate clusters and they gradually merge into larger clusters of similar data points until one unified cluster is formed 2) a top-down (divisive) approach in which all data points are initially considered one large cluster and recursively data points are split until all data points represent separate clusters. Agglomerative clustering is the most common type of hierarchical clustering (Gareth et al., 2013).

The agglomerative clustering algorithm follows several simple steps:

1. Decide on a distance metric and let each data point be a cluster.

2. Merge the most similar clusters based on the distance metric.

3. Update the pairwise dissimilarities among remaining clusters.

4. Repeat 2 and 3 until there is only a single cluster.

Most often, Euclidean distance is used as a distance metric for the agglomerative clustering algorithm. Another key decision made by the researcher is concerning the update of pairwise dissimilarities among remaining clusters. To measure the dissimilarity between two clusters containing multiple data points, a method of linkage must be selected by the researcher. Common methods of linkage are complete linkage, group average linkage and single linkage. Complete linkage considers the largest dissimilarity between two clusters while single linkage does the opposite and considers the smallest dissimilarity between two clusters to update the pairwise dissimilarities. Alternatively, group average takes the mean inter cluster dissimilarities. Most often, researchers prefer complete and group average linkage methods over single linkage as these methods yield denser and more balanced clusters (Gareth et al., 2013).

The results of the hierarchical clustering algorithms are displayed in a tree-based representation of data points called a dendrogram. When applying agglomerative clustering, the dendrogram is built in a bottom-up manner, starting from the data points, and merging them into larger clusters. The data points that merge at the bottom of the tree are more similar to each other than those that merge later. The height of the branches therefore represents how similar the data points and hence the clusters are to each other. To determine the clusters, researchers must draw a horizontal line across the dendrogram and the grouped set of datapoints below the line will be considered as distinct clusters. To draw the line, researchers can compare the height of the merge to the average merge heights below and if its substantially higher, draw a cut on the merge as it is joining distinct clusters.

We applied agglomerative clustering algorithm to our data set, using Euclidean distance as our distance metric and complete linkage as the linkage method. Results of our clustering analysis will be further explained in the next section and clustering results for each year can be found in Appendix A,Yearly Results.

5.2 Clustering Results

Prior to applying the clustering algorithms to our data set, we partitioned our data into yearly subgroups. Since this operation significantly decreased the number of data points available for clustering analysis, we tried to avoid removing data points from our data set as much as possible. However, once we constructed the box-plots of variables for each year, we noticed the outliers close to the extreme borderline which particularly affected agglomerative clustering results and produced single data point clusters in 2017, 2011, 2009 and 2008 data sets. Figure 5.1 displays the box-plot for 2017 as an example. The environment category score of a particular data point is displayed on the top with a score of 25. In cases similar to this example, we removed these outliers before conducting our clustering analysis.

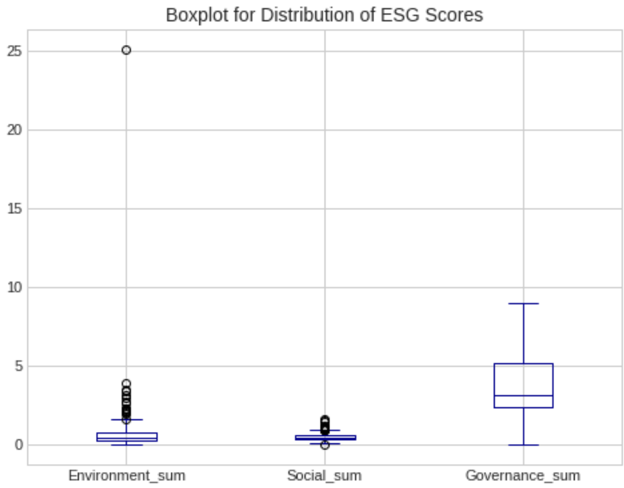


Figure 5.1 Distribution of ESG scores

Our secondary check before applying the clustering algorithm was the correlation between the variables. For each year, we constructed a correlation matrix displaying pairwise correlations of the variables. Figure 5.2 shows the correlation matrix for 2017. Similar to the result in this year, we found a moderate negative correlation between governance and file length variables in all years. Fortunately, this was not an obstacle for our analysis since we did not include file length as a clustering variable.

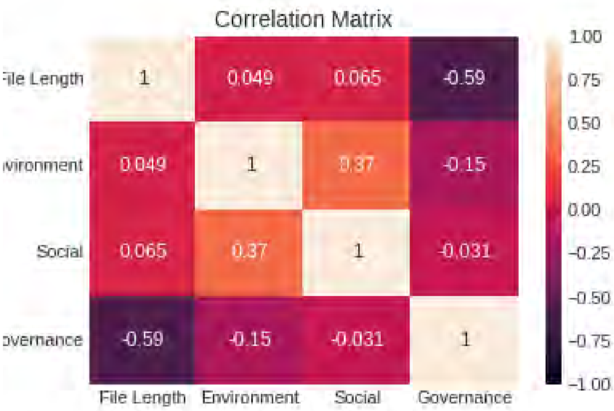


Figure 5.2 Correlation of Final Variables

A final decision we had to make during the clustering analysis was concerning the number of clusters. When applying K-medoids algorithm, we selected the number of clusters prior to applying the algorithm, by displaying the elbow and silhouette plots and choosing the elbow point with the highest silhouette coefficient. For agglomerative clustering, we followed a more heuristic approach in which we applied the clustering method, plotted the dendrogram and compared the height of the merge of top branches to the average merge heights below. We then compared the silhouette coefficients for different number of clusters and made a heuristic decision in which we took both factors into account. In some cases, we preferred a cluster quantity with a lower silhouette coefficient over a higher one since the distribution of data points to clusters were too unbalanced. If the silhouette coefficients were very close to each other, we preferred a larger cluster quantity over a smaller one to display the clusters in detail. In most cases, we selected 2 clusters with k-medoids and 3 or 4 clusters with agglomerative clustering algorithms. The number of clusters for each clustering method and year can be found in Appendix Table A.2.

After selecting the number of clusters and applying the clustering algorithms to each year's data, we mapped out the original (pre-scaled) data points on 3-d plots with each dimension representing a different category of the ESG score. Figure 5.3 and Figure 5.4 shows the 3-d plots for 2017 data clustered with k-medoids and agglomerative clustering methods respectively. It is clearly visible that in both cases, majority of the data points are clustered similarly. The most evident difference is that agglomerative clustering, in contrast to k-medoids, tends to break large clusters and therefore clustered a bunch of data points with a relatively higher environment score separately. This difference was present in nearly all years.

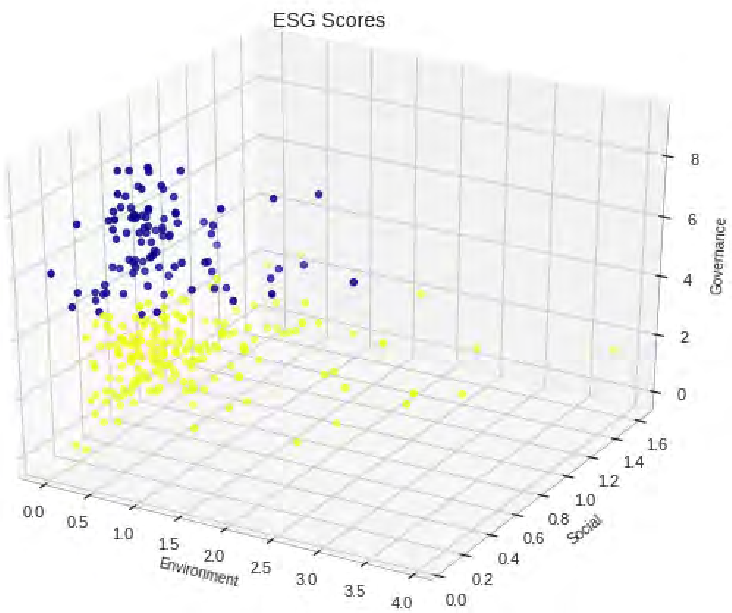


Figure 5.3 3-d plot for K-medoids clustering (2017)

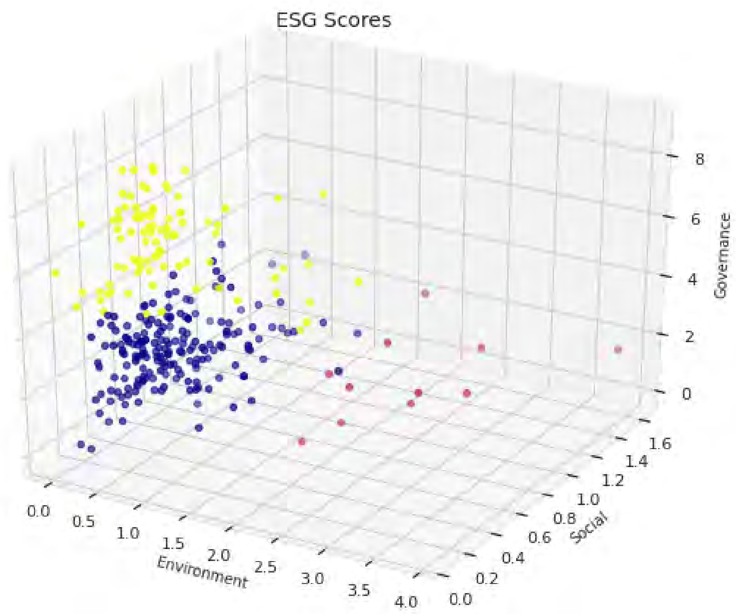


Figure 5.4 3-d plot for Hierarchical clustering (2017)

Once we had an approximate understanding of the clusters by looking at the 3-d plots, we computed the average statistics for different variables and compared the clusters. For each clustering method, we plotted the total number of reports, ESG scores and average document length per cluster. Figures 5.5, 5.6 and 5.7 illustrate these metrics for 2017 data clustered with the agglomerative clustering algorithm. Answering RQ2, the plots show that there are three major groups of reports. The first group, cluster 2, containing reports with less than 10.000 words per report on average, has high governance scores but low environment and social scores. On the contrary, the largest cluster is cluster 0 with reports of more than 35.000 words per report on average that has moderate governance scores and low environment and social scores that are similar to the previous group's scores. Finally, cluster 1 has the smallest number of reports, with an approximate length of 32.000 words per report on average, and contains reports with low social scores, moderate environment scores and moderate governance scores, slightly lower than environment scores. These three distinct groups appear in the results of agglomerative clustering for all years with different nominal ESG scores and slight changes between the rank of environment and governance scores in cluster 1. When the number of clusters is 4 instead of 3, we observe a new group resembling cluster 0 that has relatively higher scores in environment and social dimensions with a similar file length.

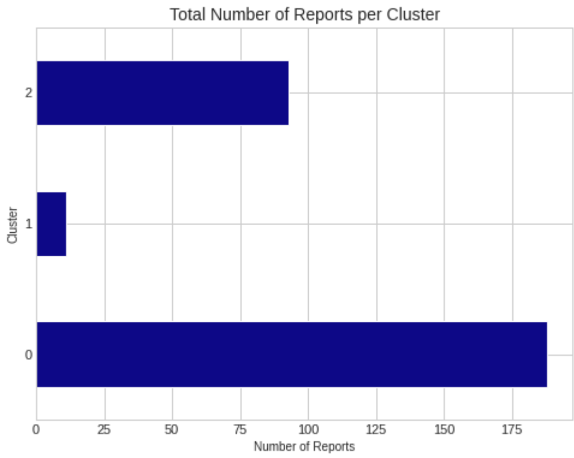


Figure 5.5 Total number of reports per cluster (2017)

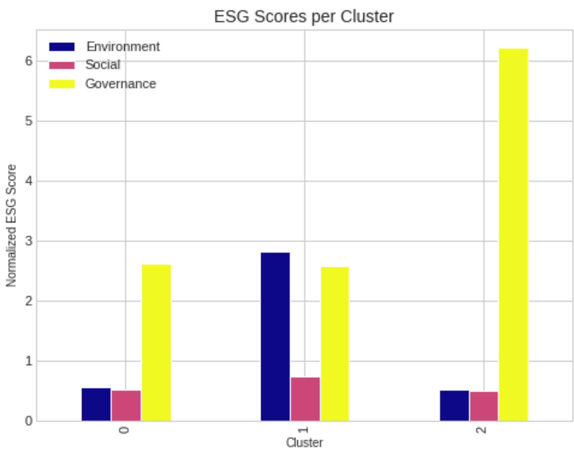


Figure 5.6 ESG scores per cluster (2017)

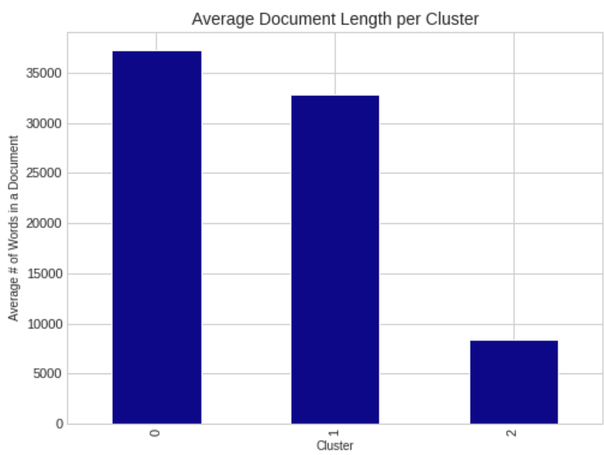


Figure 5.7 Average document length per cluster (2017)

The final plot we prepared for each clustering algorithm is a grid of pie charts illustrating the distribution of clusters for each industry. Figure 5.8 shows data on this metric for the agglomerative clustering results of 2017 data. A common result we observe in the pie charts of all years is that the group of reports with a relatively high environment score (cluster 1 for 2017) mainly come from the following industries: Electricity, Gas Water, Financial Institutions, Holding Companies, Wholesale, Retail, Horeca and Production. This is an expected outcome for companies in the Electricity, Gas Water industry by nature as well as for the rest of these industries dominated by large B2C corporations. Another shared result we observe is that the reports prepared by companies in administrative and support services industry and agriculture, forestry and fishing industries mostly belong to cluster 2 which has high governance scores. Given further supporting research, this result might be attributed to the fact that these companies conduct business highly regulated by the government entities.

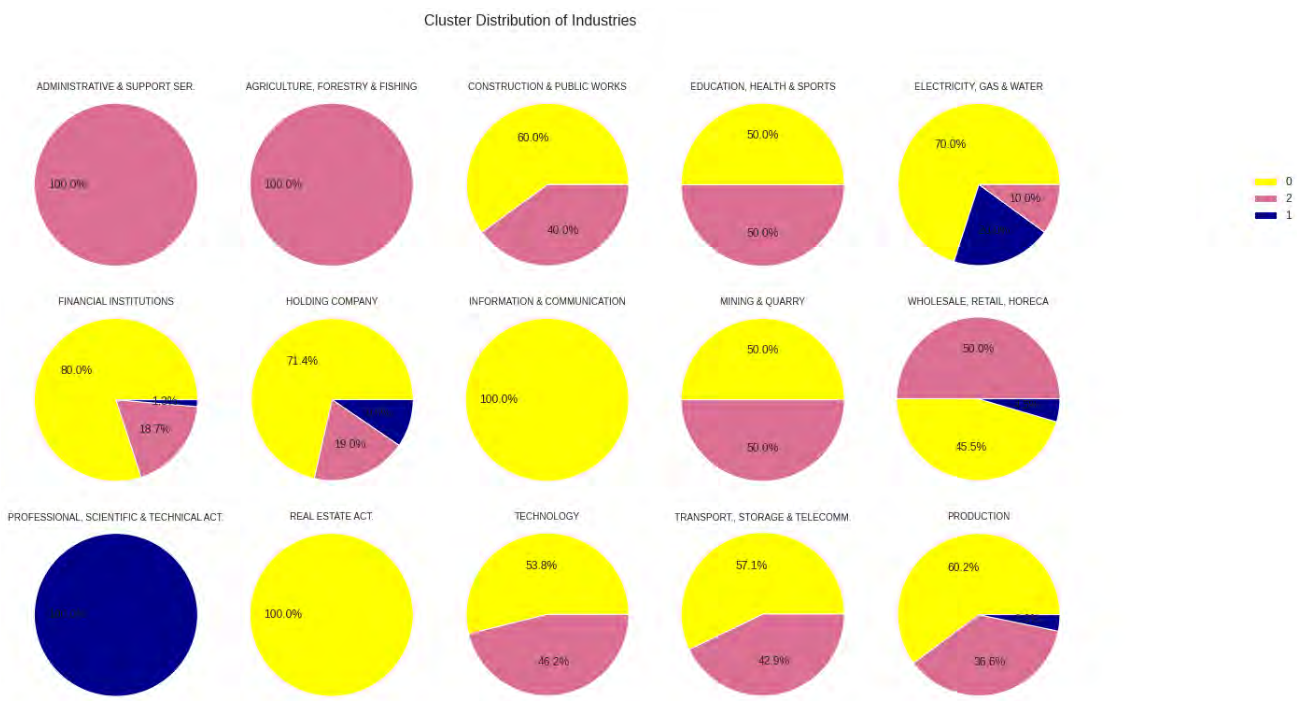


Figure 5.8 Distribution of clusters for each industry (2017)

The most evident difference we see between the years, particularly for the agglomerative clustering results, is that the proportion of companies grouped in the cluster with high environment scores increased significantly in 2020. This result clearly addresses our final research question, RQ3. Table 5.1 illustrates the corresponding percentages for each year. We can see that the highest numbers are from 2008 and 2020. However, 2008 data is not a reliable source for this comparison since the environment score of the cluster is relatively low (slightly above 1) compared to other years. We can see that there is a huge jump in 2020 with a 8.23% cluster distribution. This jump can be attributed to a numerous variables and further research is needed to uncover the antecedents and consequences of such a change.

Year	%
2007	2,02%
2008	8,55%
2009	2,16%
2010	1,86%
2011	3,35%
2012	0,98%
2013	4,65%
2014	1,29%
2015	1,65%
2016	1,49%
2017	3,77%
2018	1,32%
2019	1,30%
2020	8,23%

Table 5.1 Proportion of companies grouped in the cluster with high environment scores per year

In our final chapter, we discuss the results of our study, address the limitations, and suggest potential research questions for future work.

6. Conclusion & Discussion

This research studied the CSR disclosures of BIST companies from 2007 to 2020. Our study followed a text mining approach and examined the annual reports of BIST companies that are publicly available on company websites. We prepared an ESG dictionary to extract related keywords from the annual reports and assigned aggregate environment, social and governance scores to each report. Descriptive results for all data showed that governance related information had the highest salience among all ESG categories while environment salience has an upward trend. We then employed two different clustering algorithms, k-medoids and hierarchical (agglomerative) clustering, to group all reports based on their ESG salience. Our analysis revealed 3 distinct groups of reports and showed that the share of the group with high environment scores has increased significantly in 2020.

A previous study on SP 500 companies by Tamimi and Sebastianelli (2017) revealed that companies are most transparent regarding governance disclosures and that environmental and social practices are disclosed with significant deficiencies. In our study, in line with Tamimi and Sebastianelli (2017)'s results, governance category had the highest rank among all categories for each year. This is an expected result since governance is the only category regulated by regulating parties. It is, however, disappointing to see that there is a huge difference between the scores of governance and other categories for more than 90% of the companies each year. These findings indicate that companies do not have strong motivations to be transparent about their social and environmental practices. Unfortunately, our findings show that BIST companies do not comply with Freeman (2010)'s highly recognized Stakeholder Theory which emphasizes companies' responsibility towards all stakeholders including shareholders, employees, customers, suppliers, the government, and society. A recent study by Jackson et al. (2020) compared OECD countries in terms of the effects of non-financial disclosure (NFD) on CSR and found that companies operating in countries where non-financial disclosure is required adopt significantly more CSR activities than others. Complementing the results of our study, Jackson et al. (2020)'s findings motivates us to suggest that mandatory NFD for social and environmental disclosure could be used as a governmental policy to urge corporations become more transparent.

Another major finding of our study is that the proportion of companies grouped in the cluster with relatively higher environment scores in their reports has significantly increased from 1.30% in 2019 to 8,23% in 2020. The cluster of companies with high environment score mostly belong to Electricity, Gas Water, Financial Institutions, Wholesale, Retail, Horeca and Production industries or are holding companies. The industrial differences on the extent of ESG disclosure is in line with previous research which reports that "more sensitive" consumer and energy supplying industries provide more CSR information in their reports (Gamerschlag et al., 2011; Reverte, 2009). On the other hand, the substantial increase in the proportion of these companies in 2020 is unanticipated but promising. Potential drivers of this increase could be addressed in further studies given that research for subsequent years portrait a steady or increasing rate for environment and/or social scores. For instance, future research can explore the change in stakeholder expectations in terms of environmental issues and how it changed with the pandemic.

While we find the novel text mining method a suitable approach for our research purpose, as with all other empirical studies, there are several limitations to the present study that needs to be acknowledged and addressed. Firstly, the data collected for this study solely consisted of corporate annual reports. Although annual reports do convey important CSR-related information, companies increasingly publish stand-alone CSR/sustainability reports, reserve a CSR page on their websites and communicate CSR related information on their social media accounts. Future studies can unite different sources of media to grasp a broader view on companies' CSR disclosures. Secondly, the industry classification adopted for our study is open to criticism since the distribution to different industries is unbalanced. We acquired the industry information from the Public Disclosure Platform (2022) together with the list of BIST companies and merged some of the categories together. While BIST companies, by their very nature, have an unbalanced industry distribution, some industries are represented by only three or fewer companies which makes them statistically less reliable. Finally, there are some potential limitations related to the text mining approach we adopted for this study. For instance, the dictionaries we constructed might not capture all of the relevant CSR aspects as they're build accustomed to the data used for the present study. Furthermore, we used the stemming technique to reach word stems instead of lemmatization, a more elaborate and complex method that takes morphological analysis of words into account. That is, in our analysis, words were detached from their textual background. For future studies, lemmatization can be used as an alternative technique and results of different techniques can be compared. Despite all these limitations, we believe that our results provide interesting insights into the CSR disclosure of companies, particularly in a developing nation context, and present fruitful findings for future research. Further investigations of our findings and acknowledgements of the limitations of our study can contribute highly to the CSR disclosure literature.

BIBLIOGRAPHY

Aggarwal, P., & Singh, A. K. (2018). Csr and sustainability reporting practices in india: An in-depth content analysis of top-listed companies. *Social Responsibility Journal*.

Aguinis, H., & Glavas, A. (2012). What we know and don't know about csr: A review and research agenda. *Journal of Management*, *38*(4), 932–968.

Ahmad, N. B. J., Rashid, A., & Gow, J. (2017). Board independence and corporate social responsibility (csr) reporting in malaysia. *Australasian Accounting, Business and Finance Journal*, *11*(2), 61–85.

Aksoy, M., Yilmaz, M. K., Tatoglu, E., & Basar, M. (2020). Antecedents of corporate sustainability performance in turkey: The effects of ownership structure and board attributes on non-financial companies. *Journal of Cleaner Production*, *276*, 124284.

Ali, W., Frynas, J. G., & Mahmood, Z. (2017). Determinants of corporate social responsibility (csr) disclosure in developed and developing countries: A literature review. *Corporate Social Responsibility and Environmental Management*, *24*(4), 273–294.

Aras, G., Özsözgün Çalışkan, A., Esen, E., & Kutlu Furtuna, Ö. (2019). Türkiye'de entegre raporlama: Mevcut durum, paydaşların algı ve beklentileri. *ACCA, CFGS, KPMG, İstanbul, Ocak*.

Arthur, D., & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding* (Tech. Rep.). Stanford.

Ashcroft, P. A. (2012). Extent of environmental disclosure of us and canadian firms by annual report location. *Advances in Accounting*, *28*(2), 279–292.

Aureli, S. (2017). A comparison of content analysis usage and text mining in csr corporate disclosure. *International Journal of Digital Accounting Research*, *17*.

Belverd, E., SARI, E. S., GÜNGÖR, N., TÜREL, A., & Mustafa, C. (2019). Sustainability reporting and integrated reporting: evidence from turkish high performance companies. *Hitit Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *12*(2), 257–270.

Borsa Istanbul. (2022a). *Kurumsal yönetim.* Retrieved from `https://www.borsaistanbul.com/tr/sayfa/218/kurumsal-yonetim`

Borsa Istanbul. (2022b). *Tarihsel gelişmeler.* Retrieved from `https://www.borsaistanbul.com/tr/sayfa/27/tarihsel-gelismeler`

Bose, S. (2020). Evolution of esg reporting frameworks. In *Values at work* (pp. 13–33). Springer.

Bowen, H. R. (1953). Social responsibility of the businessman. new york: Harper and row.

Cai, R., Lv, T., & Deng, X. (2021). Evaluation of environmental information disclosure of listed companies in china's heavy pollution industries: A text mining-based methodology. *Sustainability*, *13*(10), 5415.

Carroll, A. B. (2008). A history of corporate social responsibility: Concepts and practices. *The Oxford handbook of corporate social responsibility*, *1*.

Deng, Q., Hine, M. J., Ji, S., & Sur, S. (2019). Inside the black box of dictionary building for text analytics: a design science approach. *Journal of international technology and information management*, *27*(3), 119–159.

Driver, H. E., & Kroeber, A. L. (1932). *Quantitative expression of cultural relationships* (Vol. 31) (No. 4). Berkeley: University of California Press.

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H., Giorgi, S., Park, G., ... others (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398.

Elkington, J., & Rowlands, I. H. (1999). Cannibals with forks: The triple bottom line of 21st century business. *Alternatives Journal*, *25*(4), 42.

Ertuna, B., & Tukel, A. (2010). Traditional versus international influences: Csr disclosures in turkey. *European Journal of International Management*, *4*(3), 273–289.

Feder, B. (1989). Who will subscribe to the valdez principles? *The New York Times*, *10*, F6.

Fifka, M. S. (2013). Corporate responsibility reporting and its determinants in comparative perspective–a review of the empirical literature and a meta-analysis. *Business strategy and the environment*, *22*(1), 1–35.

Freeman, R. E. (2010). *Strategic management: A stakeholder approach.* Cambridge university press.

Gamerschlag, R., Möller, K., & Verbeeten, F. (2011). Determinants of voluntary csr disclosure: empirical evidence from germany. *Review of Managerial Science*, *5*(2), 233–262.

Gao, Y. (2011). Csr in an emerging country: a content analysis of csr reports of listed companies. *Baltic Journal of management*.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in r.* Spinger.

Goloshchapova, I., Poon, S.-H., Pritchard, M., & Reed, P. (2019). Corporate social responsibility reports: topic analysis and big data approach. *The European Journal of Finance*, *25*(17), 1637–1654.

Gray, S. J., Radebaugh, L. H., & Roberts, C. B. (1990). International perceptions of cost constraints on voluntary information disclosures: A comparative study of uk and us multinationals. *Journal of international business studies*, *21*(4), 597–622.

GRI. (n.d.). *Our mission and history.* Retrieved from `https://www.globalreporting.org/about-gri/mission-history/`

Hoştut, S., & Hof, S. D. V. H. (2014). Corporate social responsibility practices of transnational corporations: Examples from turkey and italy1. *European Journal of Research on Education*, *2*(2), 202–213.

Integrated Reporting. (n.d.). *About us.* Retrieved from `https://www.integratedreporting.org/the-iirc-2/`

Integrated Reporting Türkiye. (2022). *What is integrated reporting?* Retrieved from `http://www.entegreraporlamatr.org/en/integrated-reporting/what-is-integrated-reporting.aspx`

Jackson, G., Bartosch, J., Avetisyan, E., Kinderman, D., & Knudsen, J. S. (2020). Mandatory non-financial disclosure and its influence on csr: An international comparison. *Journal of Business Ethics*, *162*(2), 323–342.

Jose, T. (2017). Need for harmonisation of sustainability reporting standards.

Journal of Finance and Economics, *5*(6), 253–258.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, *344*, 68–125.

Kemp, L. J., & Vinke, J. (2012). Csr reporting: a review of the pakistani aviation industry. *South Asian Journal of Global Business Research*.

Khan, M., Hassan, A., Harrison, C., & Tarbert, H. (2020). Csr reporting: A review of research and agenda for future research. *Management Research Review*.

Kiliç, M., Kuzey, C., & Uyar, A. (2015). The impact of ownership and board structure on corporate social responsibility (csr) reporting in the turkish banking industry. *Corporate Governance*.

Kurumsal Yönetim Tebliği (II-17.1). (2014). T.C. Resmi Gazete (28871, 03.01.2014).

Larson, R. K., & Street, D. L. (2013). Ifrs teaching resources: Available and rapidly growing. In *Teaching ifrs* (pp. 25–46). Routledge.

Lee, M., & Huang, Y.-L. (2020). Corporate social responsibility and corporate performance: A hybrid text mining algorithm. *Sustainability*, *12*(8), 3075.

Li, M., & Zhao, L. (2021). Exploring global fashion sustainability practices through dictionary-based text mining. *Clothing and Textiles Research Journal*, 0887302X21998268.

Liao, P.-C., Xia, N.-N., Wu, C.-L., Zhang, X.-L., & Yeh, J.-L. (2017). Communicating the corporate social responsibility (csr) of international contractors: Content analysis of csr reporting. *Journal of Cleaner Production*, *156*, 327–336.

Lowe, W. (2003). The statistics of text: New methods for content analysis. In *Midwest conference* (pp. 1–17).

Myšková, R., & Hájek, P. (2019). Relationship between corporate social responsibility in corporate annual reports and financial performance of the us companies. *Journal of International Studies*, *12*(1).

Ning, X., Yim, D., & Khuntia, J. (2021). Online sustainability reporting and firm performance: Lessons learned from text mining. *Sustainability*, *13*(3), 1069.

Nobanee, H., & Ellili, N. (2016). Corporate sustainability disclosure in annual reports: Evidence from uae banks: Islamic versus conventional. *Renewable and Sustainable Energy Reviews*, *55*, 1336–1341.

Pencle, N., & Mălăescu, I. (2016). What's in the words? development and validation of a multidimensional dictionary for csr and application using prospectuses. *Journal of Emerging Technologies in Accounting*, *13*(2), 109–127.

Public Disclosure Platform. (2022). *Bist companies.* Retrieved from `https://www.kap.org.tr/en/bist-sirketler`

PWC Turkey. (2011). *New turkish commercial code, a blueprint for the future.* Retrieved from `https://www.pwc.com.tr/en/publications/ttk-assets/pages/ttk-a_blueprint_for_the_future.pdf`

Qi, Y., Taplin, R., & Brown, A. M. (2012). Natural environment disclosures in the annual reports of chinese listed entities. *Journal of Asian and African Studies*, *47*(6), 587–604.

Reverte, C. (2009). Determinants of corporate social responsibility disclosure ratings by spanish listed firms. *Journal of business ethics*, *88*(2), 351–366.

Rodriguez-Gomez, S., Arco-Castro, M. L., Lopez-Perez, M. V., & Rodríguez-Ariza,

L. (2020). Where does csr come from and where does it go? a review of the state of the art. *Administrative Sciences*, *10*(3), 60.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Şahin, Z., Çankaya, F., & Yılmaz, Z. (2016). Content analysis of sustainability reports: a practice in turkey. In *7th european business research conference* (Vol. 15, p. 16).

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press Cambridge.

Sobhani, F. A., Amran, A., & Zainuddin, Y. (2012). Sustainability disclosure in annual reports and websites: a study of the banking industry in bangladesh. *Journal of Cleaner Production*, *23*(1), 75–85.

Stone, P. J., & Hunt, E. B. (1963). A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the may 21-23, 1963, spring joint computer conference* (pp. 241–256).

Székely, N., & Vom Brocke, J. (2017). What can we learn from corporate sustainability reporting? deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PloS one*, *12*(4), e0174807.

Tamimi, N., & Sebastianelli, R. (2017). Transparency among s&p 500 companies: An analysis of esg disclosure scores. *Management Decision*, *55*(8), 1660–1680.

Te Liew, W., Adhitya, A., & Srinivasan, R. (2014). Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry*, *65*(3), 393–400.

Thorndike, R. L. (1953). Who belongs in the family. In *Psychometrika.*

Tschopp, D., & Nastanski, M. (2014). The harmonization and convergence of corporate social responsibility reporting standards. *Journal of business ethics*, *125*(1), 147–162.

Turkish Commercial Code, No.6102. (2011). Retrieved from https://www.mevzuat.gov.tr/mevzuatmetin/1.5.6102.pdf

Türkiye finansal raporlama standartlarının ilk uygulamasına ilişkin Türkiye finansal raporlama standardı (TFRS 1) hakkında tebliğ. (2006). T.C. Resmi Gazete (26125, 31.03.2006).

UNDP. (n.d.). *Sustainable development goals: United nations development programme.* Retrieved from https://www.undp.org/sustainable-development-goals

Yüksel, C., Bulut, D., & Tüzüner, V. (2008). Corporate social responsibility programs in turkey: A content analysis. *Corporate Responsibility & Sustainability Research, CRRC*.

APPENDIX A

Keyword	Stemmed keyword	Category-ESG	Category-N-gram
Ağaç	ağaç	Environment	Unigram
ağaçlandırma	ağaçlandırma	Environment	Unigram
Akaryakıt	akaryakıt	Environment	Unigram
Aleyhte bildirim	("aleyh", "bildir")	Environment	Bigram
Ambalaj	ambalaj	Environment	Unigram
arıtma	arıtma	Environment	Unigram
Atık	atık	Environment	Unigram
Bitki	bitki	Environment	Unigram
Biyoçeşitlilik	biyoçeşitlilik	Environment	Unigram
biyoçözünür	biyoçözünür	Environment	Unigram
biyoenerji	biyoenerji	Environment	Unigram
Biyolojik çeşitlilik	("biyolojik", "çeşitlilik")	Environment	Bigram
Carbon Disclosure Project	("carbo", "disclosur", "project")	Environment	Trigram
ÇEVKO	çevko	Environment	Unigram
Çevre	çevre	Environment	Unigram
Çevresel	çevresel	Environment	Unigram
Çevreci	çevreç	Environment	Unigram
deniz kirliliği	("de", "kirlilik")	Environment	Bigram
Doğa	dok	Environment	Unigram
Doğanın	doğa	Environment	Unigram
Doğal	doğal	Environment	Unigram
EFSIS	efsıs	Environment	Unigram
Egzoz	egzoz	Environment	Unigram
ekosistem	ekosiste	Environment	Unigram
Ekotasarım	ekotasar	Environment	Unigram
Emisyon	emisyo	Environment	Unigram
Enerji	enerji	Environment	Unigram
Farkındalık	farkındalık	Environment	Unigram
fidan	fida	Environment	Unigram
Geri dönüşüm	("ger", "dönüş")	Environment	Bigram
Geri kazanım	("ger", "kaza")	Environment	Bigram
Gıda israfı	("gı", "israf")	Environment	Bigram
gürültü ölçümü	("gürül", "ölç")	Environment	Bigram
Hayvan	hayva	Environment	Unigram
hektar	hektar	Environment	Unigram
İklim	iklim	Environment	Unigram
İyi tarım	("iyi", "tar")	Environment	Bigram
Karbon	karbo	Environment	Unigram
Karbondioksit	karbondioksit	Environment	Unigram
kimyasalların yönetimi	("kimyasal", "yönet")	Environment	Bigram
kirletmek	kirletmek	Environment	Unigram
Kloroflorokarbon	kloroflorokarbo	Environment	Unigram
Organik tarım	("organik", "tar")	Environment	Bigram
orman	orma	Environment	Unigram
ormanlık	ormanlık	Environment	Unigram
ozon tabakası	("ozo", "tabakas")	Environment	Bigram
plastik	plastik	Environment	Unigram
Rüzgar	rüzgar	Environment	Unigram
saf	saf	Environment	Unigram
Sera gazı	("sera", "gaz")	Environment	Bigram
Sorumlu tüketim	("sorumlu", "tüket")	Environment	Bigram
Sorumlu üretici	("sorumlu", "üretiç")	Environment	Bigram
Su havzaları	("su", "havza")	Environment	Bigram
Su kaynakları	("su", "kaynak")	Environment	Bigram
Su riskleri	("su", "risk")	Environment	Bigram
Su tüketimi	("su", "tüket")	Environment	Bigram
Su verimliliği	("su", "verimlilik")	Environment	Bigram

Su yönetimi	("su", "yönet")	Environment	Bigram
Sürdürülebilirlik	sürdürülebilirlik	Environment	Unigram
Tasarruf	tasarruf	Environment	Unigram
temiz	te	Environment	Unigram
temizlik	temizlik	Environment	Unigram
toprak	toprak	Environment	Unigram
ürün kalitesi	("ür", "kalites")	Environment	Bigram
Ürün yaşam döngüsü	("ür", "yaşa", "döngüs")	Environment	Trigram
Üstün üretim belgesi	("üs", "üret", "belges")	Environment	Trigram
Verimlilik	verimlilik	Environment	Unigram
Yakıt tüketimi	("yakıt", "tüket")	Environment	Bigram
Yeniden tüketilebilir	("yeni", "tüketilebilir")	Environment	Bigram
yeşil tedarik zinciri	("yeşil", "tedarik", "zincir")	Environment	Trigram
AÇEV	açev	Social	Unigram
Adalet	adalet	Social	Unigram
Adil ücret	("adil", "ücret")	Social	Bigram
Aile	ail	Social	Unigram
Akademik	akademik	Social	Unigram
araştırma	araştırma	Social	Unigram
Aşı kampanyası	("aş", "kampanyas")	Social	Bigram
ayrımcılık	ayrımcılık	Social	Unigram
ayrıştırıcı	ayrıştırıç	Social	Unigram
azınlık	azınlık	Social	Unigram
bağış	bağış	Social	Unigram
Bağlılık	bağlılık	Social	Unigram
Bakım	bak	Social	Unigram
Barış elçiliği	("barış", "elçilik")	Social	Bigram
Başarı kriterleri	("başar", "kriter")	Social	Bigram
Bayram harçlığı	("bayra", "harçlık")	Social	Bigram
Beceri	becer	Social	Unigram
beden eğitimi	("be", "eğit")	Social	Bigram
Bilgilendirme	bilgilendirme	Social	Unigram
bilim	bil	Social	Unigram
bilimsel	bilimsel	Social	Unigram
bilinçlendirme	bilinçlendirme	Social	Unigram
bordro	bordro	Social	Unigram
Burs	burs	Social	Unigram
bursiyer	bursiyer	Social	Unigram
cesaret	cesaret	Social	Unigram
Cinsel tercih	("cinsel", "tercih")	Social	Bigram
Cinsiyet	cinsiyet	Social	Unigram
çağdaş	çağdaş	Social	Unigram
çalışan bağlılığı	("çalışa", "bağlılık")	Social	Bigram
Çalışan hakları	("çalışa", "hak")	Social	Bigram
Çalışan odaklılık	("çalışa", "odaklılık")	Social	Bigram
çalışma standartları	("çalışma", "standart")	Social	Bigram
çalışma şartları	("çalışma", "şart")	Social	Bigram
çeşitlilik	çeşitlilik	Social	Unigram
Çocuk	çocuk	Social	Unigram
ÇYDD	çydd	Social	Unigram
demokrasi	demokras	Social	Unigram
Destekçi	destekçi	Social	Unigram
Dil	dil	Social	Unigram
Din	din	Social	Unigram
disiplin	disipl	Social	Unigram
doğal afet	("doğal", "afet")	Social	Bigram
duyarlılık	duyarlılık	Social	Unigram
Eğitim	eğit	Social	Unigram
el sanatları	("el", "sanat")	Social	Bigram

Emeklilik	emeklilik	Social	Unigram
Emniyet	emniyet	Social	Unigram
Engelli	engelli	Social	Unigram
Ergenlik	ergenlik	Social	Unigram
esnek çalışma	("esnek", "çalışma")	Social	Bigram
Eşit fırsat	("eşit", "fırsat")	Social	Bigram
Eşit hak	("eşit", "hak")	Social	Bigram
Eşit kariyer fırsatları	("eşit", "kariyer", "fırsat")	Social	Trigram
Eşit olanaklar	("eşit", "olanak")	Social	Bigram
Eşit ücret	("eşit", "ücret")	Social	Bigram
eşitliği	eşitlik	Social	Unigram
eşit	eşit	Social	Unigram
eşitsizliği	eşitsizlik	Social	Unigram
etnik köken	("etnik", "köke")	Social	Bigram
Farklılıklar	farklılık	Social	Unigram
Fayda	fay	Social	Unigram
festival	festival	Social	Unigram
galeri	ga	Social	Unigram
gelenek	gelenek	Social	Unigram
Gelişim	geliş	Social	Unigram
gençlik	gençlik	Social	Unigram
geri bildirim	("ger", "bildir")	Social	Bigram
Gönüllü	gönüllü	Social	Unigram
Güvenilir	güvenilir	Social	Unigram
hamilelik	hamilelik	Social	Unigram
hassasiyet	hassasiyet	Social	Unigram
hasta	has	Social	Unigram
Hastane	hastane	Social	Unigram
hayır işleri	("hayır", "iş")	Social	Bigram
hibe	hip	Social	Unigram
Hobi	hobi	Social	Unigram
Irk	ırk	Social	Unigram
İhtiyaç	ihtiyaç	Social	Unigram
İkramiye	ikrami	Social	Unigram
İKSV	iksv	Social	Unigram
İlkokul	ilkokul	Social	Unigram
ilköğretim	ilköğret	Social	Unigram
İnanış	inanış	Social	Unigram
insan hakları	("in", "hak")	Social	Bigram
insan kaynağı	("in", "kaynak")	Social	Bigram
İnsan onuru	("insa", "onur")	Social	Bigram
İsraf	israf	Social	Unigram
İstihdam	istihda	Social	Unigram
İş elbisesi	("iş", "elbises")	Social	Bigram
İş emniyeti	("iş", "emniyet")	Social	Bigram
İş güvencesi	("iş", "güvences")	Social	Bigram
iş güvenliği	("iş", "güvenlik")	Social	Bigram
İş kazası	("iş", "kazas")	Social	Bigram
iş sağlığı	("iş", "sağlık")	Social	Bigram
İşçi	işçi	Social	Unigram
İşe alım	("işe", "al")	Social	Bigram
İyileştirme	iyileştirme	Social	Unigram
Kadın	kadı	Social	Unigram
Kamu sağlığı	("kamu", "sağlık")	Social	Bigram
Kamu yararı	("kamu", "yarar")	Social	Bigram
kanser	kanser	Social	Unigram
Kapsayıcı	kapsayıç	Social	Unigram
Kar amacı gütmeyen	("kar", "amaç", "gütmeye")	Social	Trigram
kariyer fırsatı	("kariyer", "fırsat")	Social	Bigram

Kariyer gelişimi	("kariyer", "geliş")	Social	Bigram
Kariyer planlama	("kariyer", "planla")	Social	Bigram
kariyer yönetimi	("kariyer", "yönet")	Social	Bigram
katılım	katıl	Social	Unigram
Katkı sağlama	("katkı", "sağla")	Social	Bigram
Kaza	kaz	Social	Unigram
Kıdem tazminatı	("kıde", "tazminat")	Social	Bigram
Kişilik onuru	("kişilik", "onur")	Social	Bigram
kişisel beceriler	("kişisel", "beceri")	Social	Bigram
Kişisel gelişim	("kişisel", "geliş")	Social	Bigram
kitap	kitap	Social	Unigram
Koçluk	koçluk	Social	Unigram
Kötü muamele	("köt", "muamel")	Social	Bigram
lider	lider	Social	Unigram
lise	lis	Social	Unigram
MEB	meb	Social	Unigram
mentorluk	mentorluk	Social	Unigram
mentor	mentor	Social	Unigram
meslek hastalığı	("meslek", "hastalık")	Social	Bigram
mezhep	mezhep	Social	Unigram
mikrokredi	mikrokredi	Social	Unigram
milliyet	milliyet	Social	Unigram
motivasyon	motivasyo	Social	Unigram
müdahele programları	("müdahel", "program")	Social	Bigram
müze	müz	Social	Unigram
özgür	özgür	Social	Unigram
özgürlük	özgürlük	Social	Unigram
sanat	sanat	Social	Unigram
sendika	sendika	Social	Unigram
sigarayı bırakma	("sigara", "bırakma")	Social	Bigram
sivil toplum kuruluşu	("sivil", "topl", "kuruluş")	Social	Trigram
sosyal	sosyal	Social	Unigram
sosyal sorumluluk	("sosyal", "sorumluluk")	Social	Bigram
sponsor	sponsor	Social	Unigram
şiddet	şiddet	Social	Unigram
taciz	taciz	Social	Unigram
takım çalışması	("tak", "çalışmas")	Social	Bigram
tiyatro	tiyatro	Social	Unigram
toplum	topl	Social	Unigram
vakıf	vakıf	Social	Unigram
yan haklar	("yan", "hak")	Social	Bigram
yardım	yar	Social	Unigram
yardımlaş	yardımlaş	Social	Unigram
yetenek	yetenek	Social	Unigram
yetkinlik	yetkinlik	Social	Unigram
yıllık izin	("yıllık", "iz")	Social	Bigram
yoksul	yoksul	Social	Unigram
yurt	yurt	Social	Unigram
açıklık	açıklık	Governance	Unigram
Adet	adet	Governance	Unigram
Adil	adil	Governance	Unigram
Adli yaptırım	("adli", "yaptır")	Governance	Bigram
Ahlak	ahlak	Governance	Unigram
azınlık hakları	("azınlık", "hak")	Governance	Bigram
Azlık hakkı	("azlık", "hakkı")	Governance	Bigram
Bağımsızlık	bağımsızlık	Governance	Unigram
Bilgi edinme	("bilgi", "edinme")	Governance	Bigram
bilgi gizliliği	("bilgi", "gizlilik")	Governance	Bigram
bilgilendirme politikası	("bilgilendirme", "politikas")	Governance	Bigram

Birleşmiş Milletler Kalkınma	("birleş", "millet", "kalkınma")	Governance	Trigram
Dava	dav	Governance	Unigram
Değer	değer	Governance	Unigram
denetim	denet	Governance	Unigram
Denetleme	denetle	Governance	Unigram
Divan heyeti	("diva", "heyet")	Governance	Bigram
Doğruluk	doğruluk	Governance	Unigram
Dürüst	dürüst	Governance	Unigram
Etik	etik	Governance	Unigram
Genel kurul	("genel", "kurul")	Governance	Bigram
Gizlilik	gizlilik	Governance	Unigram
Global Reporting Initiative	("global", "reportingi", "initiativ")	Governance	Trigram
GRI	grı	Governance	Unigram
Güvenli	güvenli	Governance	Unigram
haksız rekabet	("haksız", "rekabet")	Governance	Bigram
Hesap verebilirlik	("hesap", "verebilirlik")	Governance	Bigram
hesapverebilirlik	hesapverebilirlik	Governance	Unigram
huzur hakkı	("huzur", "hakkı")	Governance	Bigram
İç kontrol	("iç", "kontrol")	Governance	Bigram
İçeriden bilgi	("içeri", "bilgi")	Governance	Bigram
İçsel bilgi	("içsel", "bilgi")	Governance	Bigram
ifşa	ifşa	Governance	Unigram
İmtiyaz	imtiyaz	Governance	Unigram
İş disiplini	("iş", "disipl")	Governance	Bigram
kalite	kali	Governance	Unigram
kamuyu aydınlatma	("kamu", "aydınlatma")	Governance	Bigram
Kanun	kan	Governance	Unigram
Kar dağıtım	("kar", "dağıt")	Governance	Bigram
Kar payı	("kar", "pa")	Governance	Bigram
karapara aklanması	("karapar", "aklanmas")	Governance	Bigram
Komite	komi	Governance	Unigram
Kurumsal değerler	("kurumsal", "değer")	Governance	Bigram
Kurumsal kültür	("kurumsal", "kül")	Governance	Bigram
Kurumsal risk	("kurumsal", "risk")	Governance	Bigram
Kurumsal vatandaş	("kurumsal", "vatandaş")	Governance	Bigram
kurumsal web sitesi	("kurumsal", "web", "sites")	Governance	Trigram
Kurumsal yönetim	("kurumsal", "yönet")	Governance	Bigram
Küresel ilkeler sözleşmesi	("küresel", "ilke", "sözleşmes")	Governance	Trigram
Mali haklar	("mali", "hak")	Governance	Bigram
Menfaat	menfaat	Governance	Unigram
misyon	misyo	Governance	Unigram
organizasyonel iklim	("organizasyonel", "ikl")	Governance	Bigram
Organizasyonel yönetim	("organizasyonel", "yönet")	Governance	Bigram
otonomi	otonomi	Governance	Unigram
oy hakkı	("oy", "hakkı")	Governance	Bigram
oylama	oyla	Governance	Unigram
Örf	örf	Governance	Unigram
özerklik	özerklik	Governance	Unigram
pay devri	("pay", "devri")	Governance	Bigram
Pay sahibi	("pay", "sahip")	Governance	Bigram
Paydaş	paydaş	Governance	Unigram
Prim	prim	Governance	Unigram
Rekabet	rekabet	Governance	Unigram
risk faktörleri	("risk", "faktör")	Governance	Bigram
Risk yönetim	("risk", "yönet")	Governance	Bigram
Sağlık sigortası	("sağlık", "sigortas")	Governance	Bigram
samimiyet	samimiyet	Governance	Unigram
Saydamlık	saydamlık	Governance	Unigram
Saygınlık	saygınlık	Governance	Unigram

Sır	sır	Governance	Unigram
sorumluluk	sorumluluk	Governance	Unigram
stratejik hedef	("stratejik", "hedef")	Governance	Bigram
Sürdürülebilir Kalkınma	("sürdürülebilir", "kalkınma")	Governance	Bigram
Şeffaflık	şeffaflık	Governance	Unigram
Şirket blgilendirme politikası	("şirket", "blgilendirme", "politikas")	Governance	Trigram
Tarafsızlık	tarafsızlık	Governance	Unigram
teftiş	teftiş	Governance	Unigram
Temettü	temettü	Governance	Unigram
ticari sır	("ticar", "sır")	Governance	Bigram
Tutarlılık	tutarlılık	Governance	Unigram
Ulusal değerler	("ulusal", "değer")	Governance	Bigram
UN Global Compact	("un", "global", "compact")	Governance	Trigram
UNDP	undp	Governance	Unigram
uyum	u	Governance	Unigram
Üst yönetim	("üst", "yönet")	Governance	Bigram
veto hakkı	("veto", "hakkı")	Governance	Bigram
vizyon	vizyo	Governance	Unigram
yasal mevzuat	("yasal", "mevzuat")	Governance	Bigram
Yasalar	yasa	Governance	Unigram
yatırımcı toplantıları	("yatırımcı", "toplantı")	Governance	Bigram
yetki	yetki	Governance	Unigram
Yolsuzlukla mücadele	("yolsuzluk", "mücadel")	Governance	Bigram
Yönetim kurulu	("yönet", "kurul")	Governance	Bigram
liyakat	liyakat	Governance	Unigram

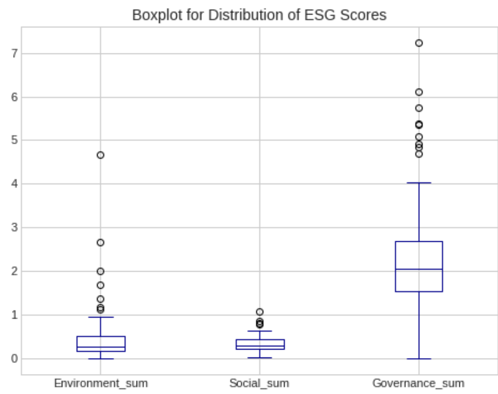
Table A.1 Table of keywords selected (stemmed versions categorized in ESG categories and N-grams)

Year	K-Medoids Clustering	Hierarchical Clustering
2007	2	3
2008	3	3
2009	3	4
2010	2	4
2011	2	4
2012	2	4
2013	2	4
2014	3	4
2015	2	4
2016	2	4
2017	2	3
2018	2	3
2019	2	4
2020	2	3

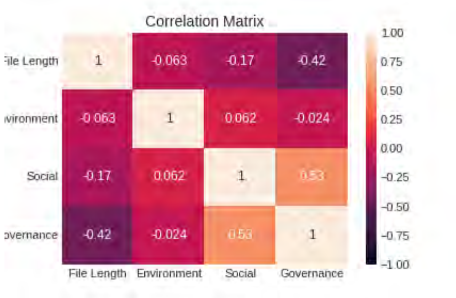
Table A.2 Number of clusters per year

Yearly Results

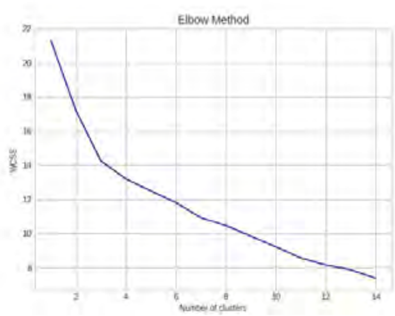
2007



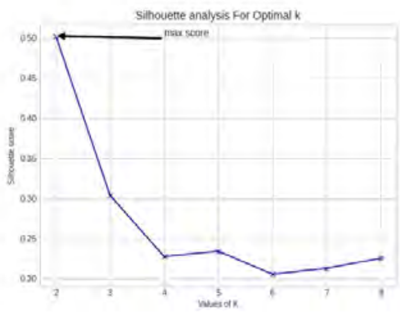
BoxplotforDistributionofESGScores2007



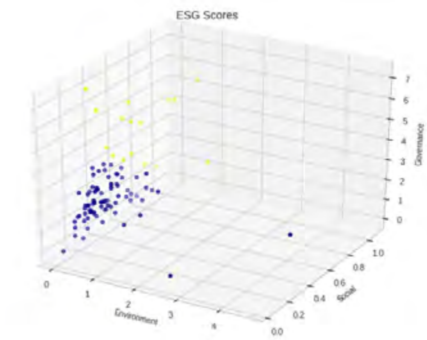
CorrelationMatrix2007



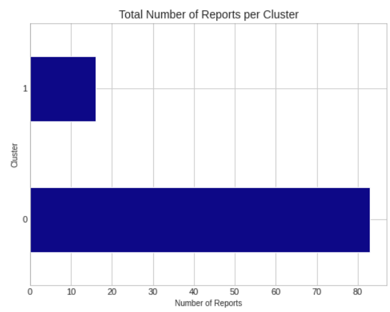
ElbowMethod(K-Medoids,2007)



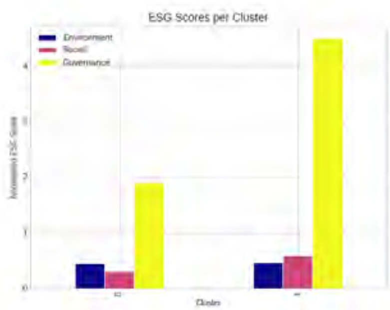
SilhouettePlot((K-Medoids,2007)



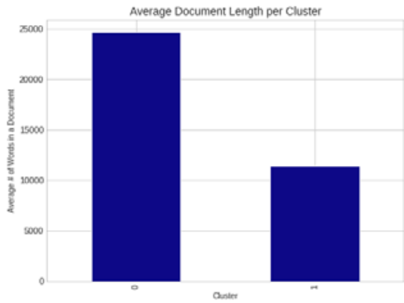
3-dplotofdatapoints(K-Medoids,2007)



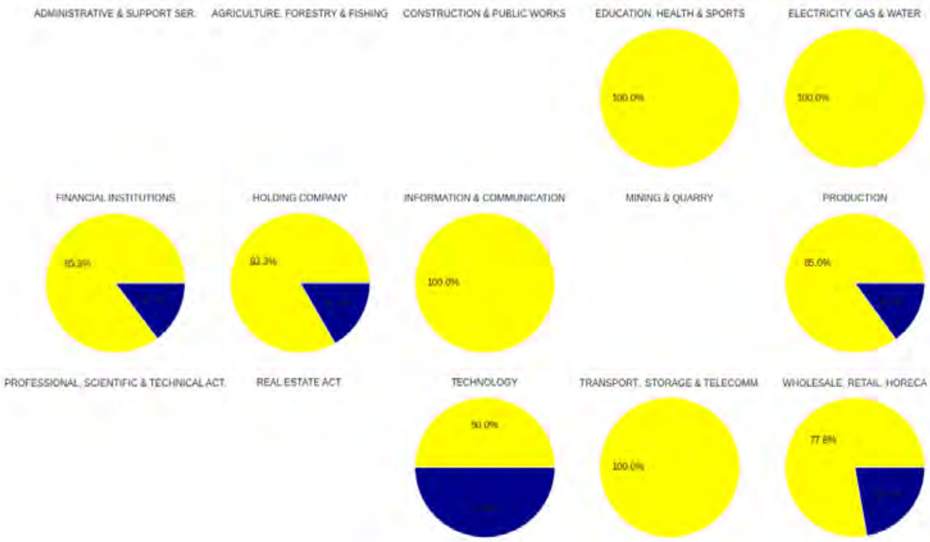
TotalNumberofReportsperCluster(K-Medoids,2007)



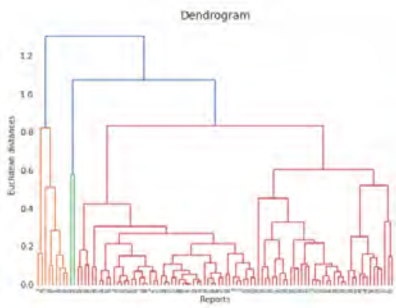
ESGScoresPerCluster(K-Medoids,2007)



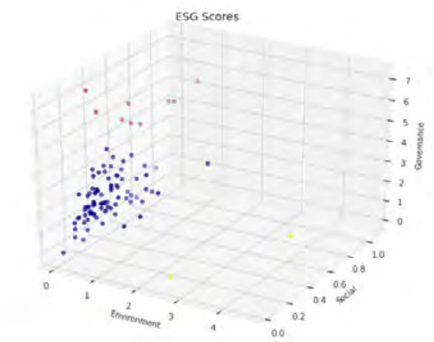
AverageDocumentLengthperCluster(K-Medoids,2007)



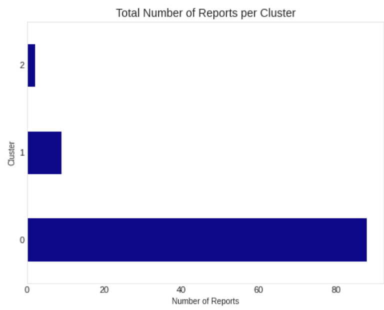
ClusterDistributionofIndustries(K-Medoids,2007)



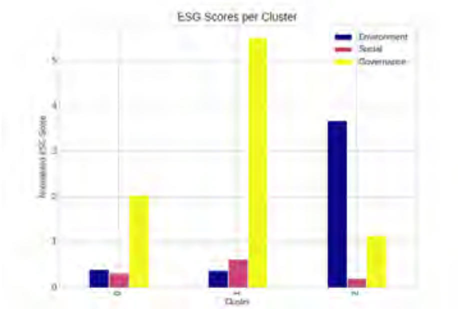
Dendrogram(Hierarchical,2007)



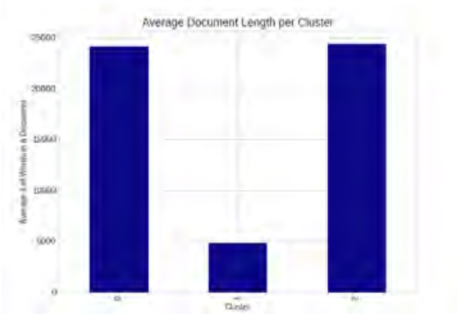
3-dplotofdatapoints(Hierarchical,2007)



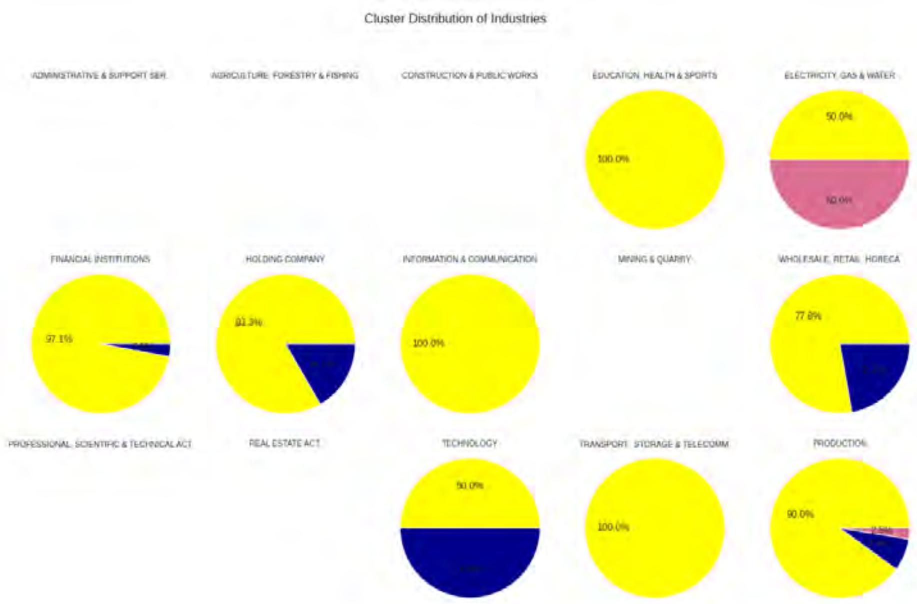
TotalNumberofReportsperCluster(Hierarchical,2007)



ESGScoresPerCluster(Hierarchical,2007)

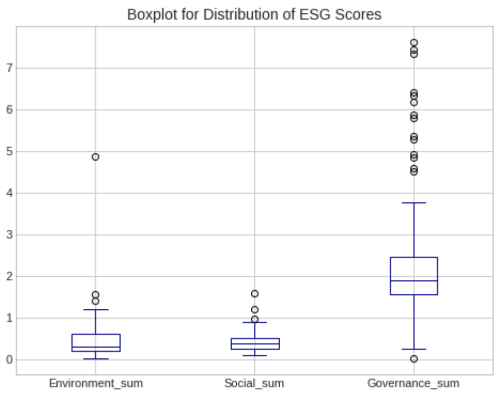


AverageDocumentLengthperCluster(Hierarchical,2007)

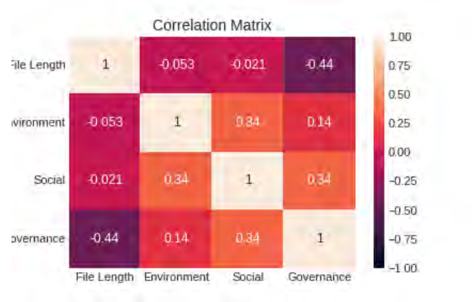


ClusterDistributionofIndustries(Hierarchical,2007)

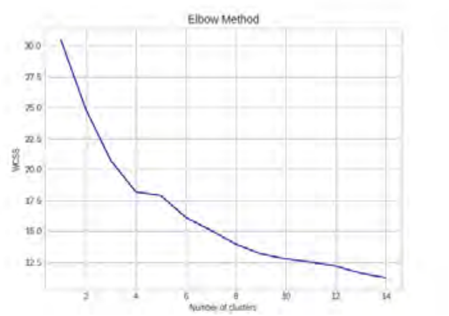
2008



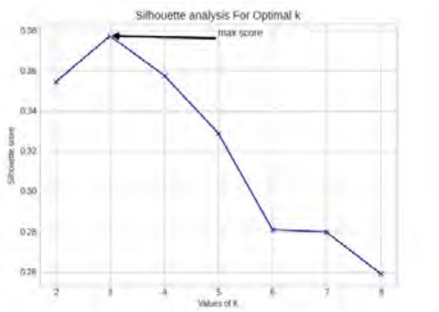
BoxplotforDistributionofESGScores2008



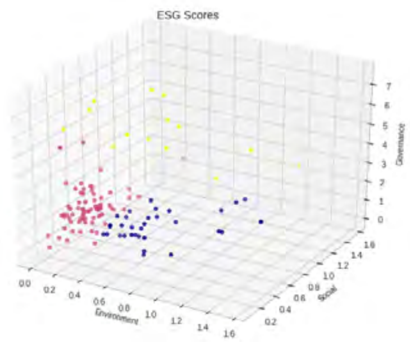
CorrelationMatrix2008



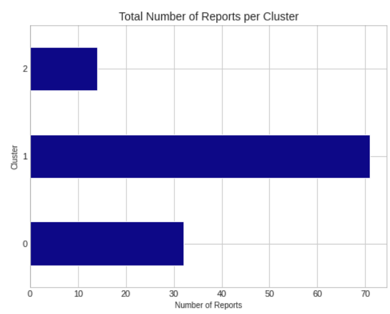
ElbowMethod(K-Medoids,2008)



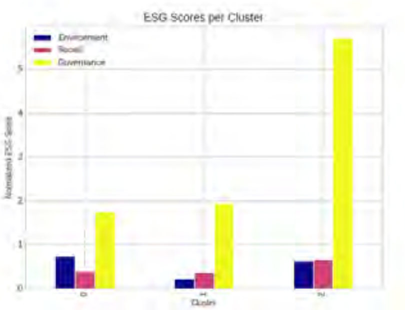
SilhouettePlot((K-Medoids,2008)



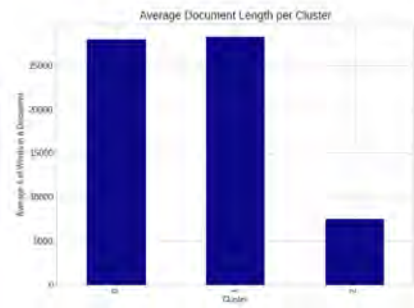
3-dplotofdatapoints(K-Medoids,2008)



TotalNumberofReportsperCluster(K-Medoids,2008)

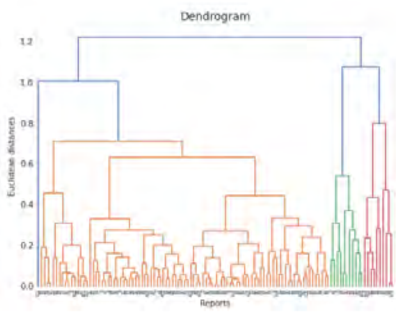


ESGScoresPerCluster(K-Medoids,2008)

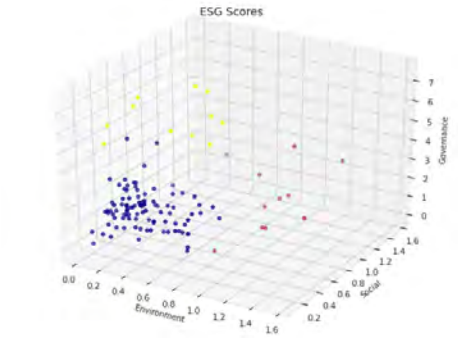


AverageDocumentLengthperCluster(K-Medoids,2008)

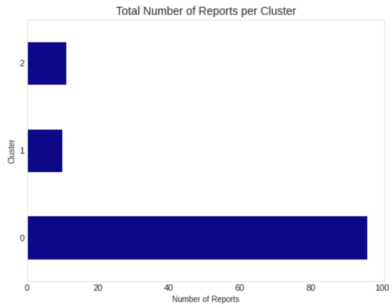
ClusterDistributionofIndustries(K-Medoids,2008)



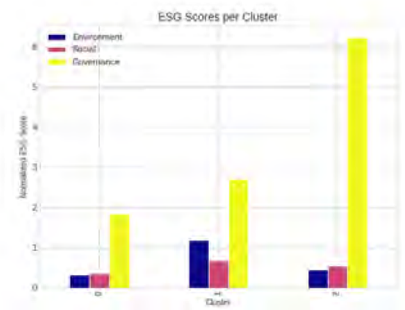
Dendrogram(Hierarchical,2008)



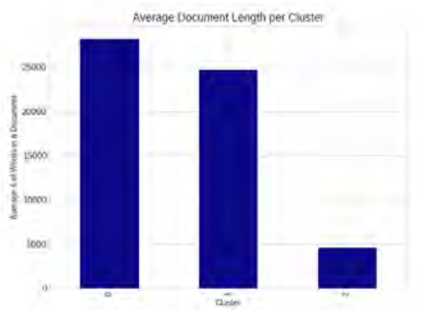
3-dplotofdatapoints(Hierarchical,2008)



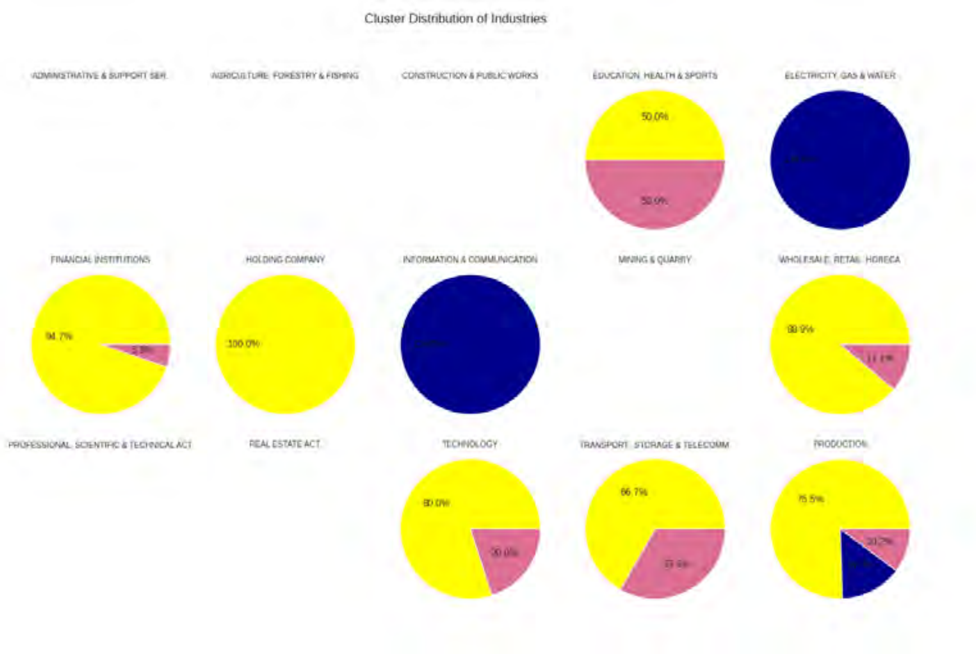
TotalNumberofReportsperCluster(Hierarchical,2008)



ESGScoresPerCluster(Hierarchical,2008)

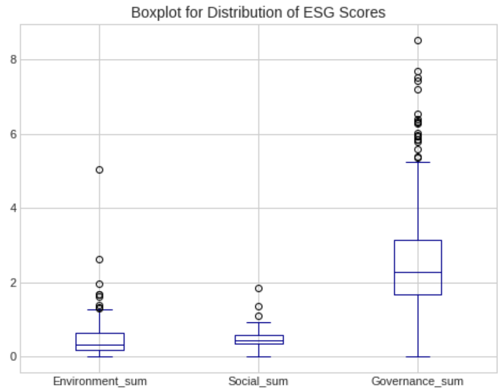


AverageDocumentLengthperCluster(Hierarchical,2008)

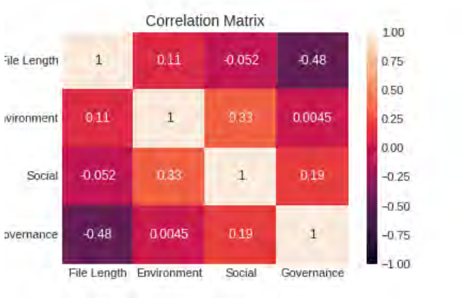


ClusterDistributionofIndustries(Hierarchical,2008)

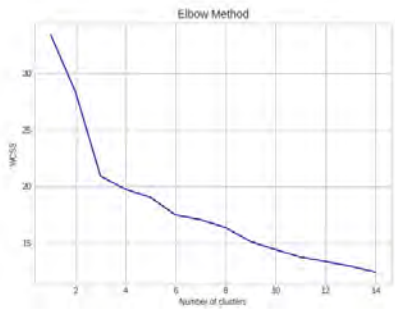
2009



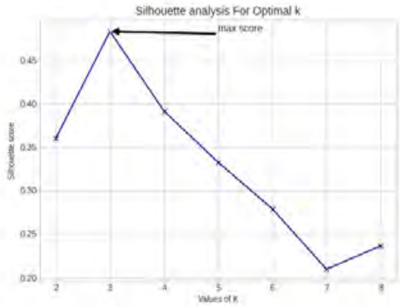
BoxplotforDistributionofESGScores2009



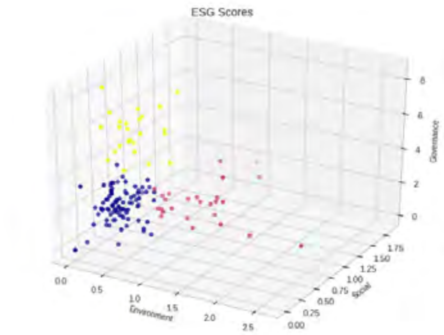
CorrelationMatrix2009



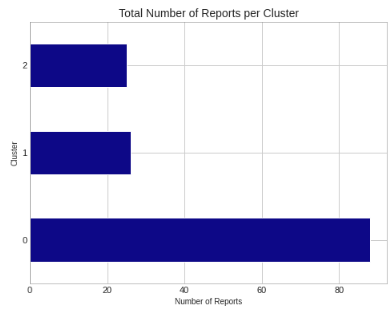
ElbowMethod(K-Medoids,2009)



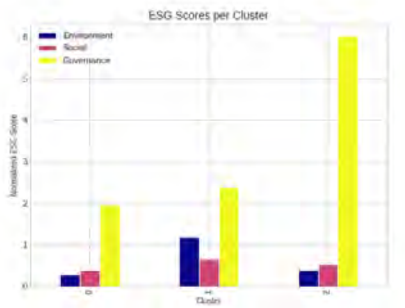
SilhouettePlot((K-Medoids,2009)



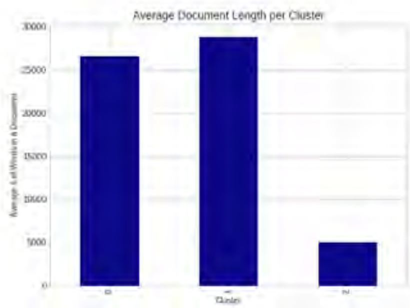
3-dplotofdatapoints(K-Medoids,2009)



TotalNumberofReportsperCluster(K-Medoids,2009)

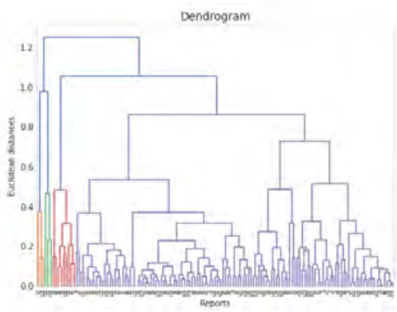


ESGScoresPerCluster(K-Medoids,2009)

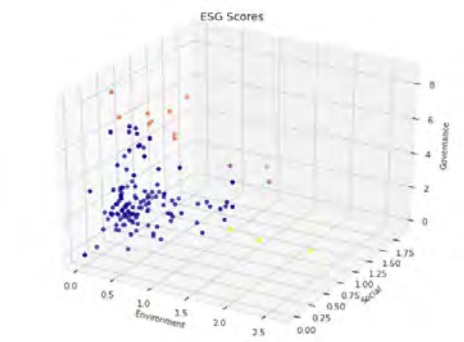


AverageDocumentLengthperCluster(K-Medoids,2009)

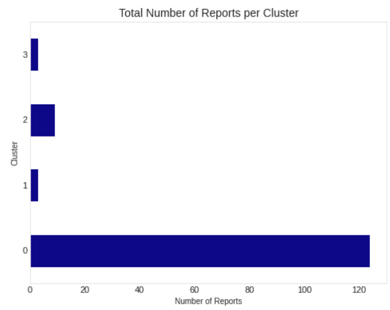
ClusterDistributionofIndustries(K-Medoids,2009)



Dendrogram(Hierarchical,2009)

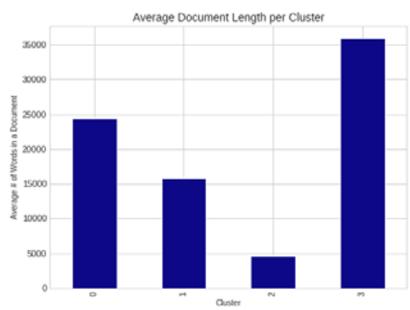


3-dplotofdatapoints(Hierarchical,2009)



TotalNumberofReportsperCluster(Hierarchical,2009)

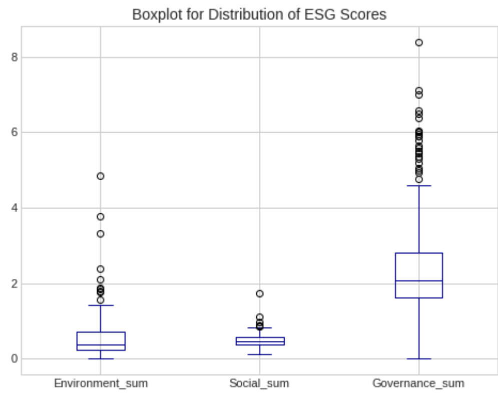
ESGScoresPerCluster(Hierarchical,2009)



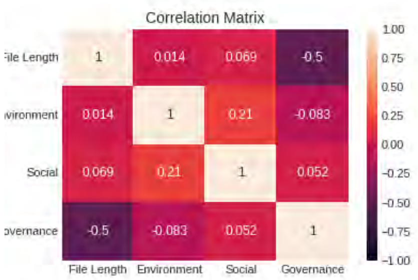
AverageDocumentLengthperCluster(Hierarchical,2009)

ClusterDistributionofIndustries(Hierarchical,2009)

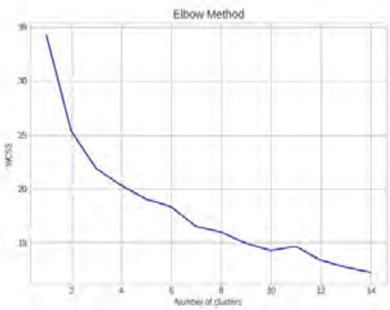
2010



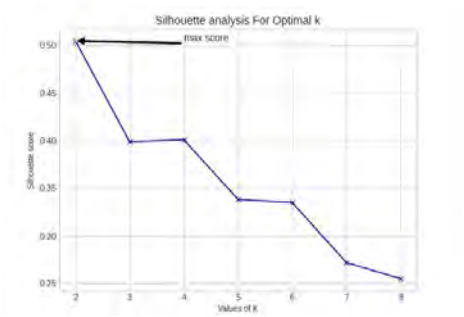
BoxplotforDistributionofESGScores2010



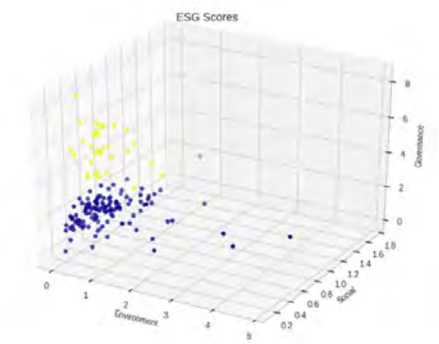
CorrelationMatrix2010



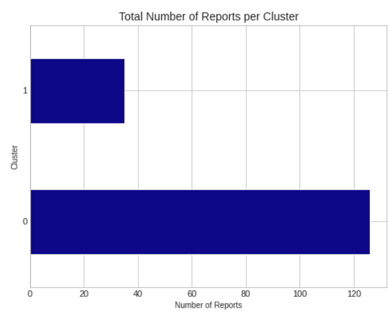
ElbowMethod(K-Medoids,2010)



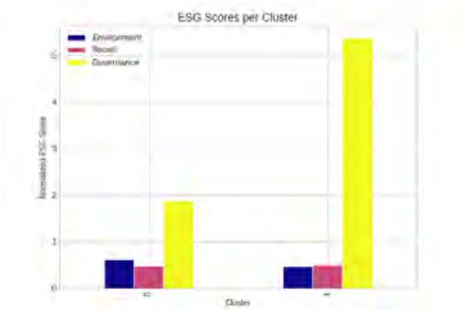
SilhouettePlot((K-Medoids,2010)



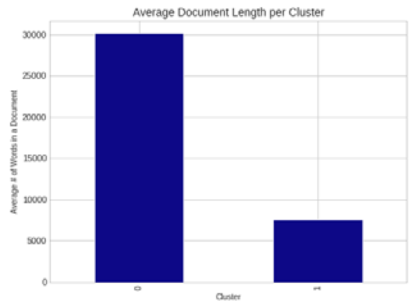
3-dplotofdatapoints(K-Medoids,2010)



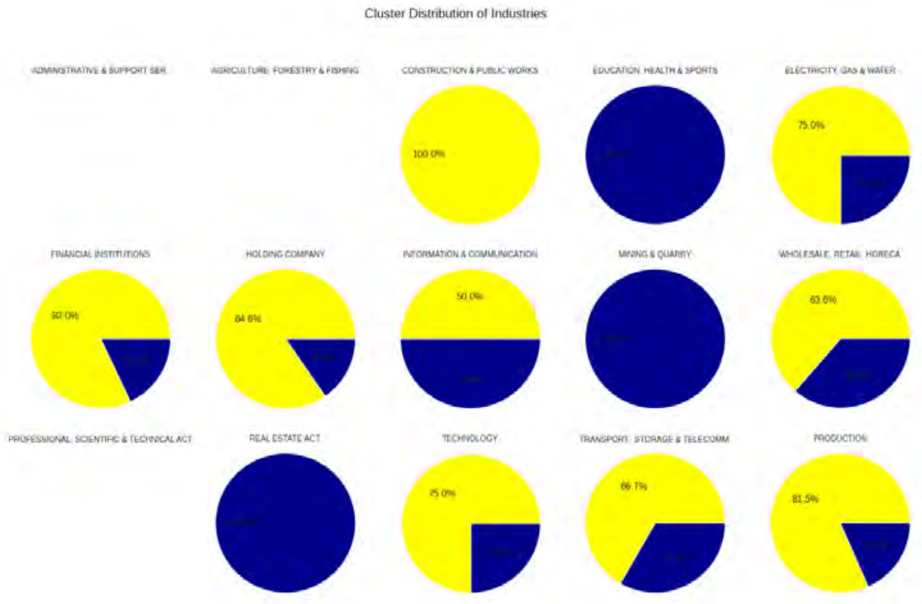
TotalNumberofReportsperCluster(K-Medoids,2010)



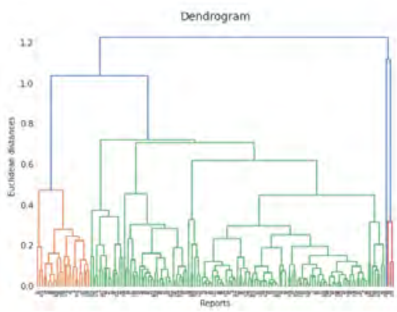
ESGScoresPerCluster(K-Medoids,2010)



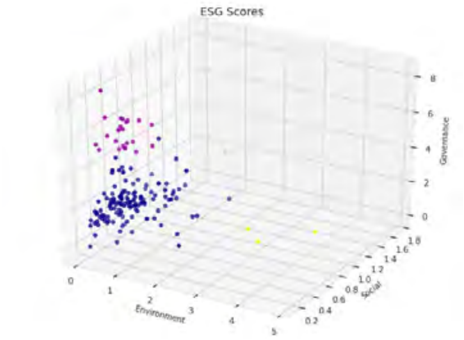
AverageDocumentLengthperCluster(K-Medoids,2010)



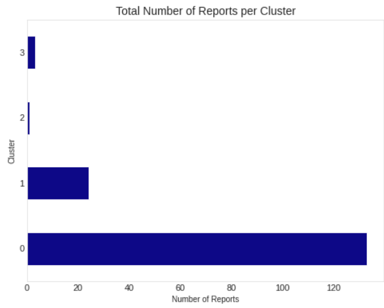
ClusterDistributionofIndustries(K-Medoids,2010)



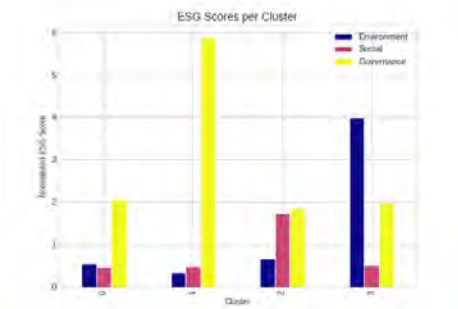
Dendrogram(Hierarchical,2010)



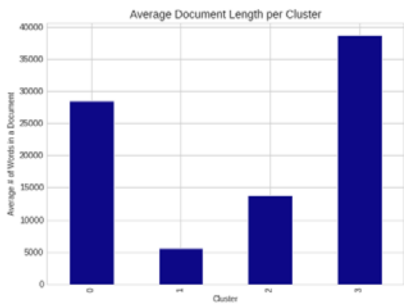
3-dplotofdatapoints(Hierarchical,2010)



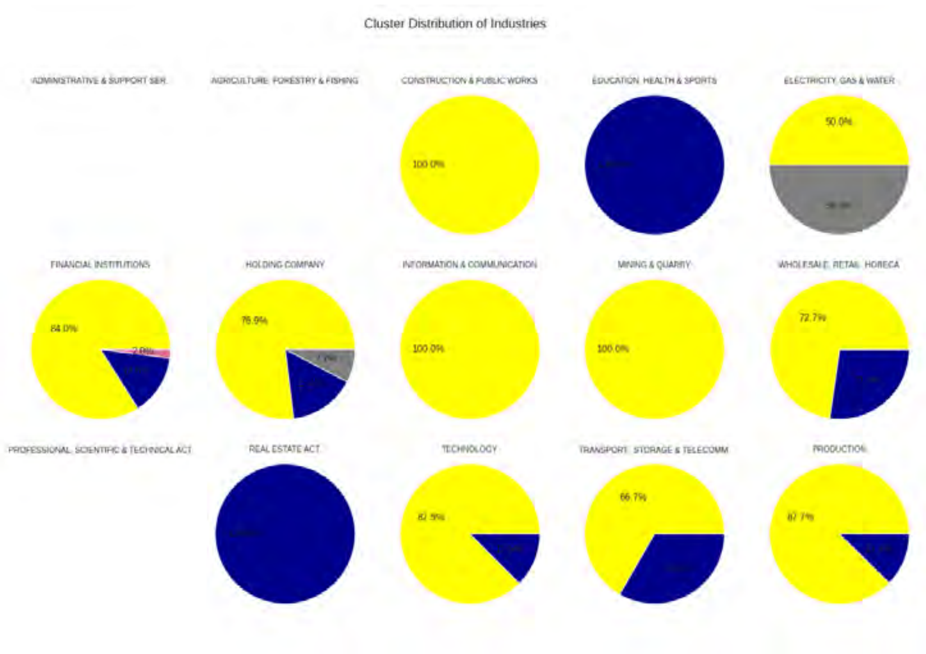
TotalNumberofReportsperCluster(Hierarchical,2010)



ESGScoresPerCluster(Hierarchical,2010)

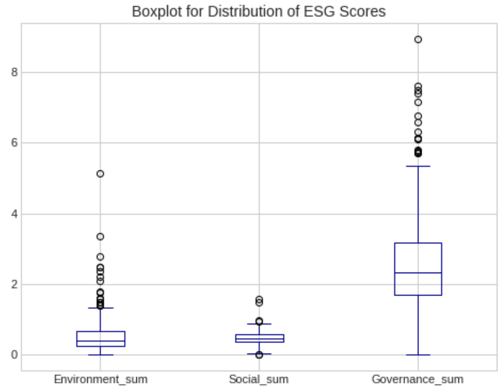


AverageDocumentLengthperCluster(Hierarchical,2010)

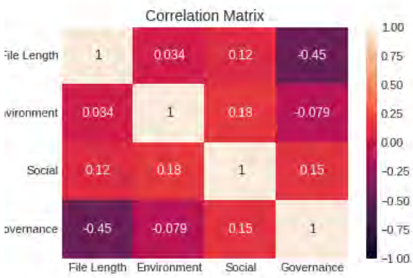


ClusterDistributionofIndustries(Hierarchical,2010)

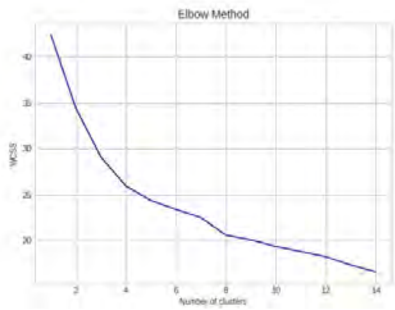
2011



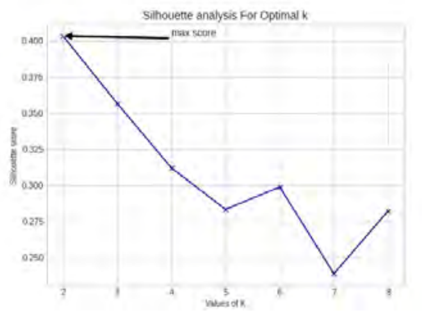
BoxplotforDistributionofESGScores2011



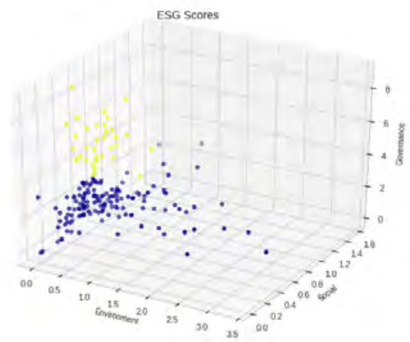
CorrelationMatrix2011



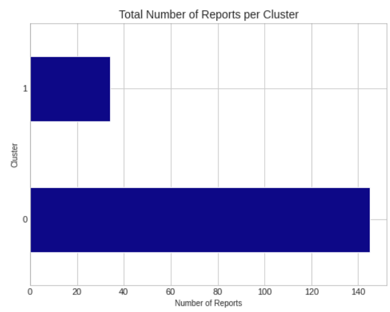
ElbowMethod(K-Medoids,2011)



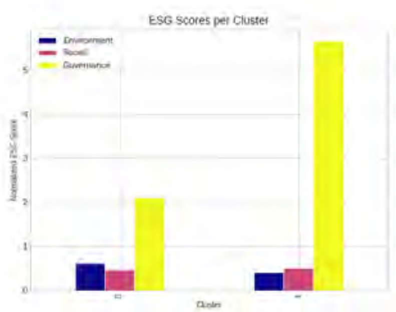
SilhouettePlot((K-Medoids,2011)



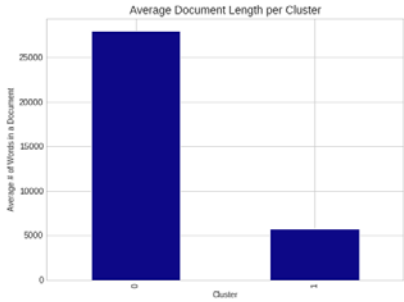
3-dplotofdatapoints(K-Medoids,2011)



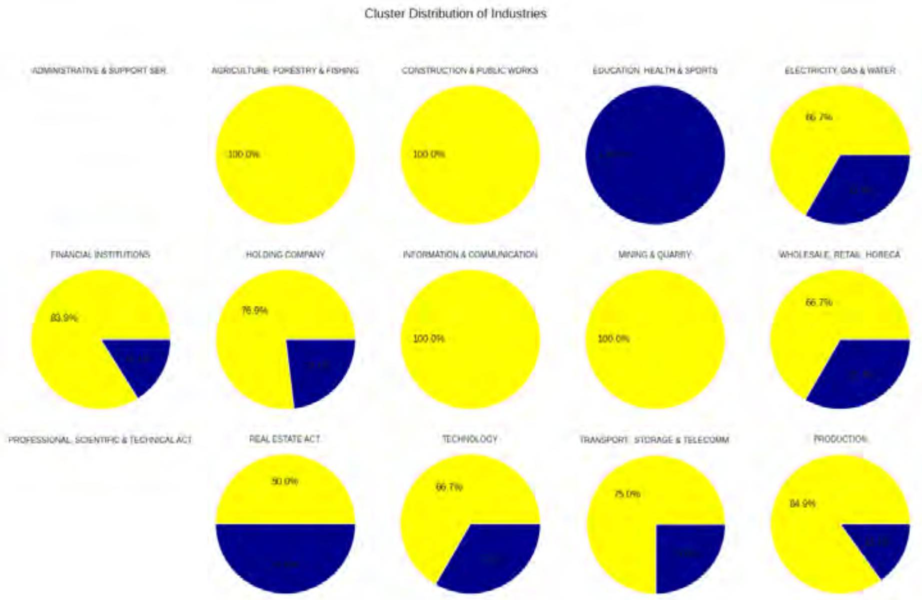
TotalNumberofReportsperCluster(K-Medoids,2011)



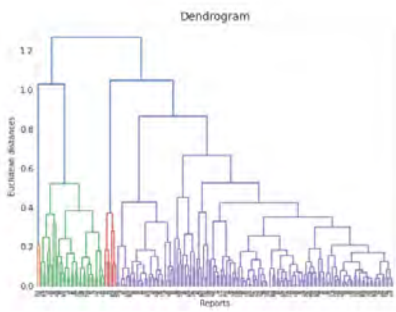
ESGScoresPerCluster(K-Medoids,2011)



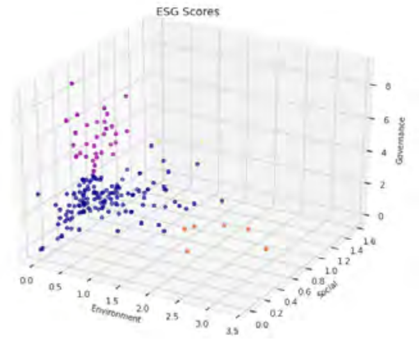
AverageDocumentLengthperCluster(K-Medoids,2011)



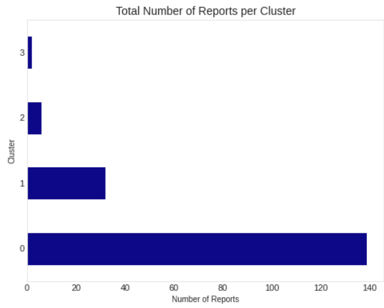
ClusterDistributionofIndustries(K-Medoids,2011)



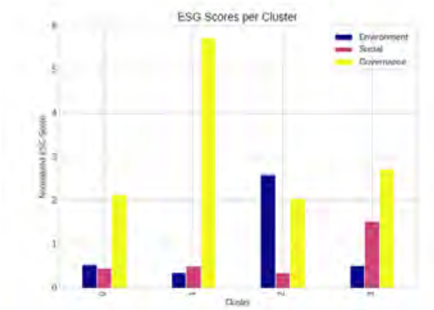
Dendrogram(Hierarchical,2011)



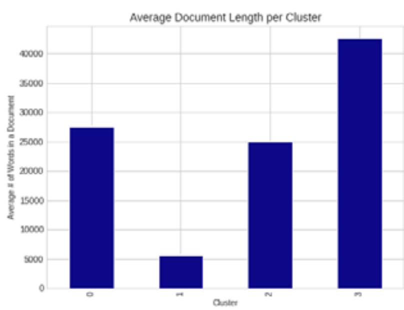
3-dplotofdatapoints(Hierarchical,2011)



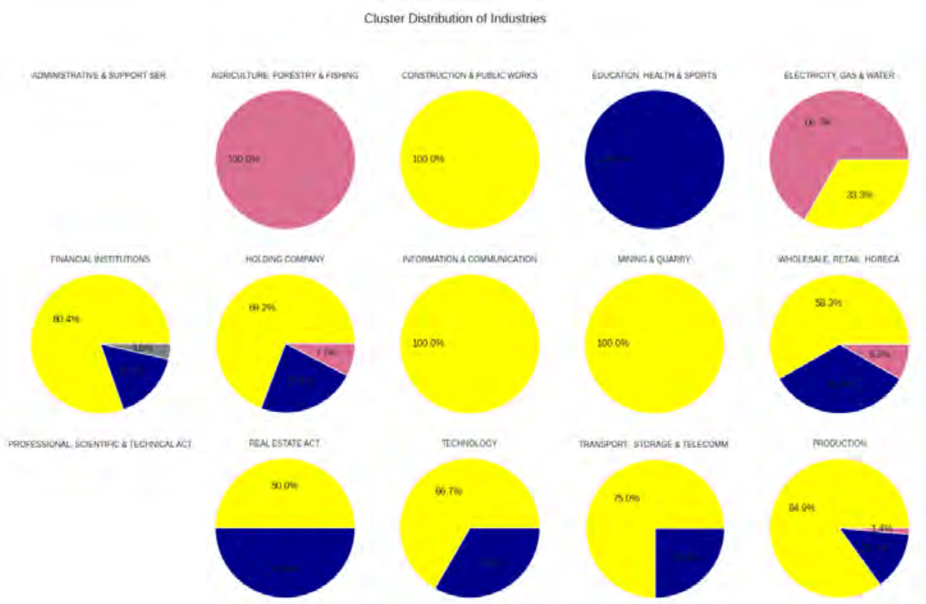
TotalNumberofReportsperCluster(Hierarchical,2011)



ESGScoresPerCluster(Hierarchical,2011)

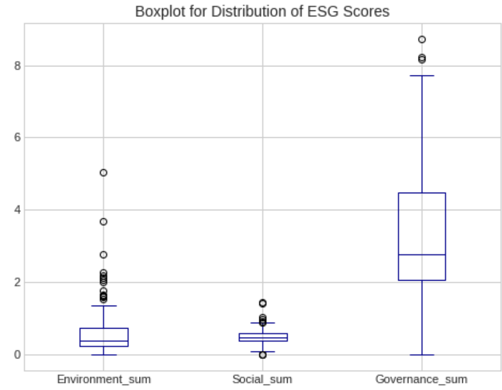


AverageDocumentLengthperCluster(Hierarchical,2011)

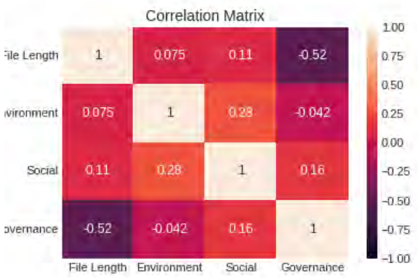


ClusterDistributionofIndustries(Hierarchical,2011)

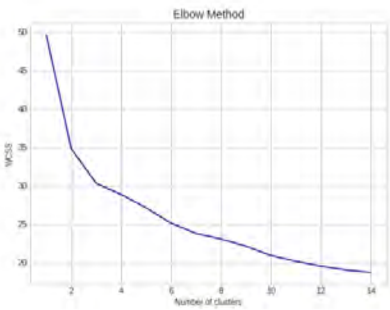
2012



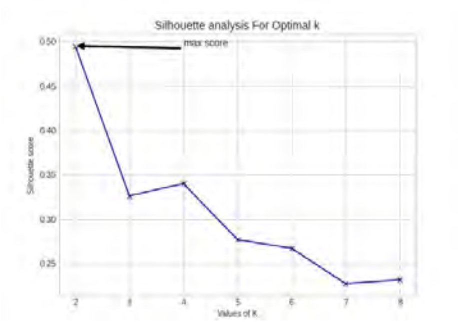
BoxplotforDistributionofESGScores2012



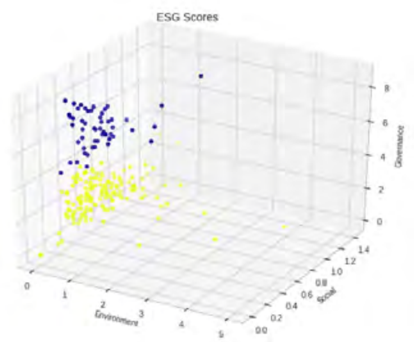
CorrelationMatrix2012



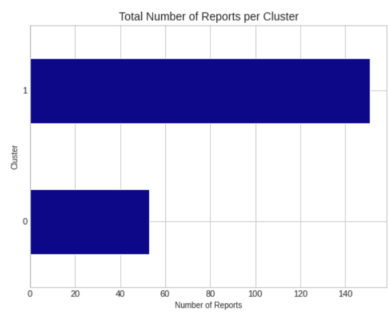
ElbowMethod(K-Medoids,2012)



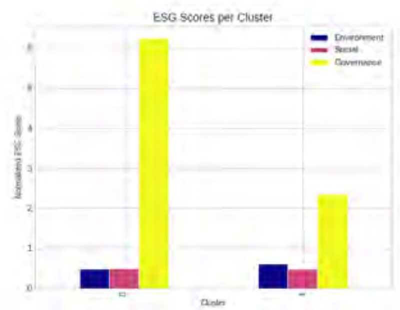
SilhouettePlot((K-Medoids,2012)



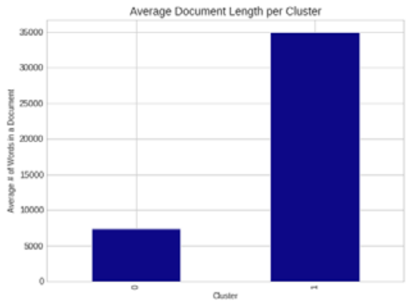
3-dplotofdatapoints(K-Medoids,2012)



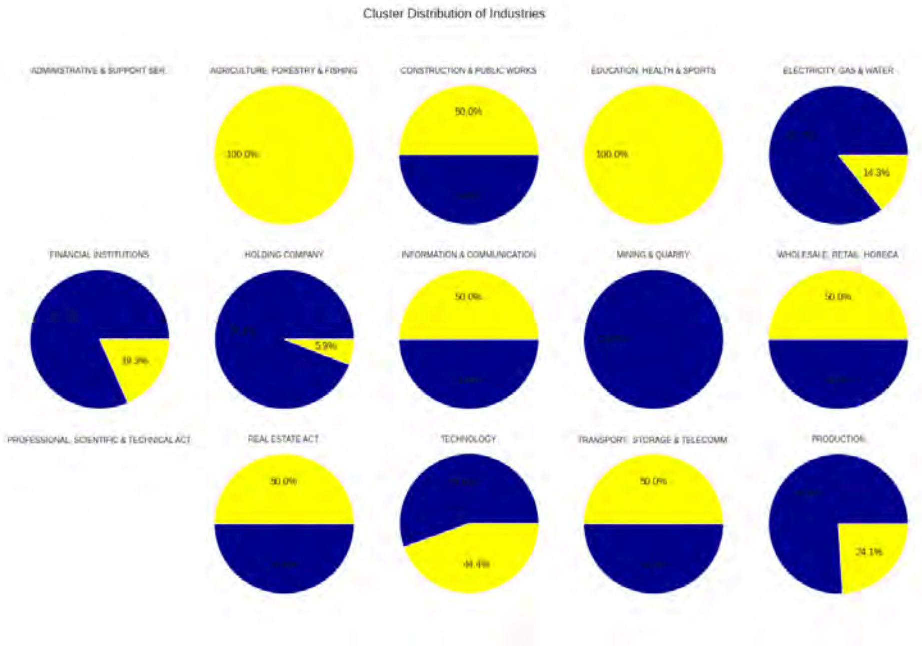
TotalNumberofReportsperCluster(K-Medoids,2012)



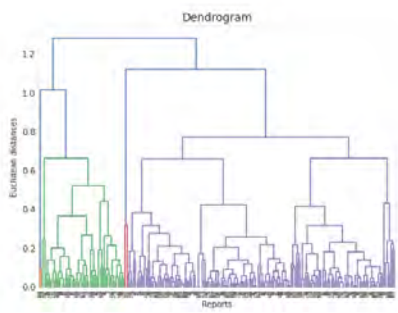
ESGScoresPerCluster(K-Medoids,2012)



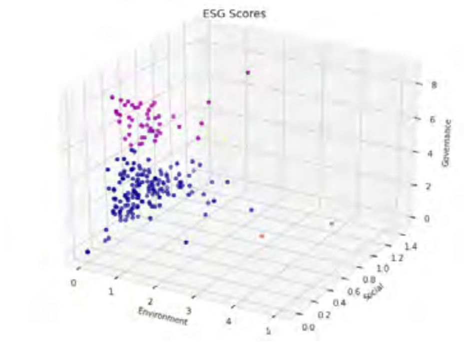
AverageDocumentLengthperCluster(K-Medoids,2012)



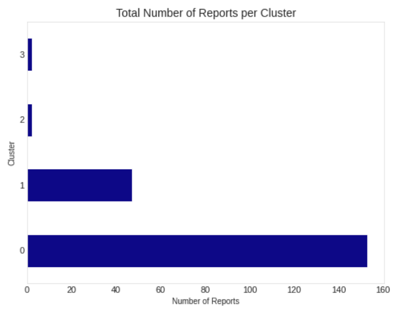
ClusterDistributionofIndustries(K-Medoids,2012)



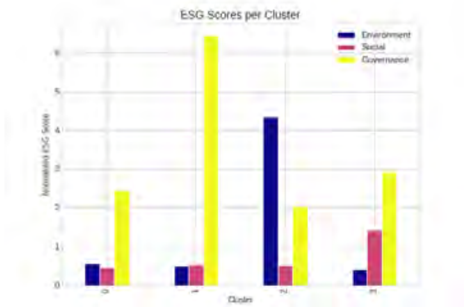
Dendrogram(Hierarchical,2012)



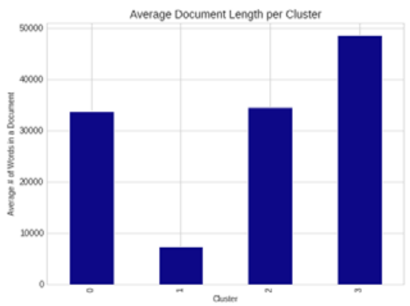
3-dplotofdatapoints(Hierarchical,2012)



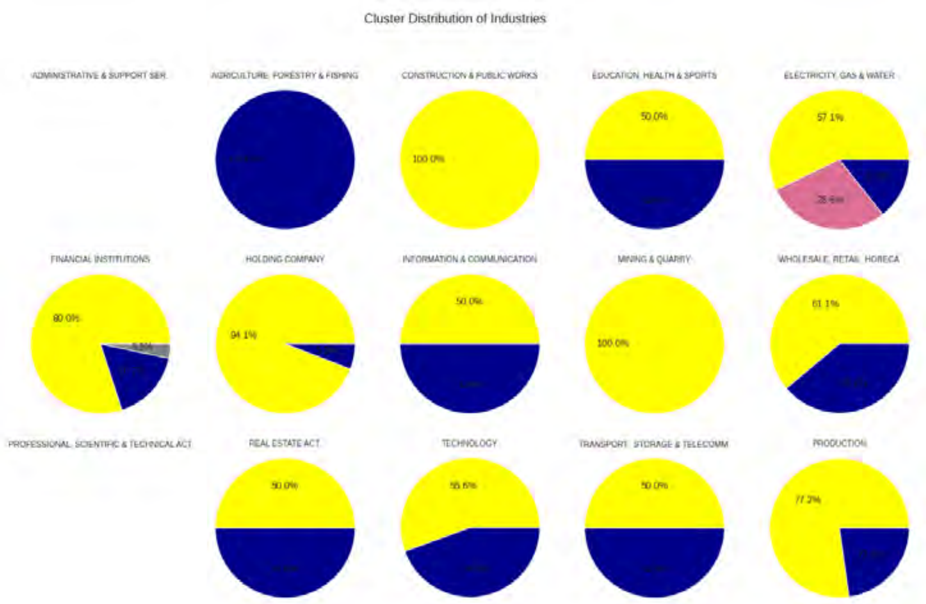
TotalNumberofReportsperCluster(Hierarchical,2012)



ESGScoresPerCluster(Hierarchical,2012)

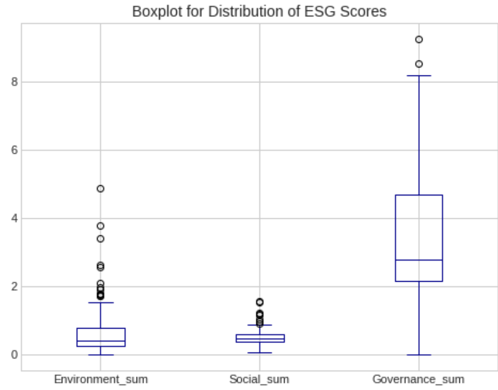


AverageDocumentLengthperCluster(Hierarchical,2012)

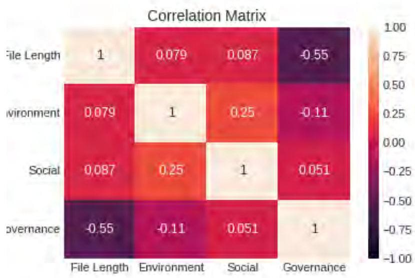


ClusterDistributionofIndustries(Hierarchical,2012)

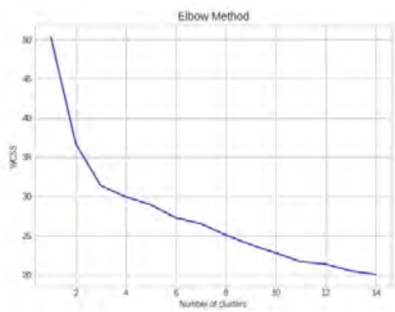
2013



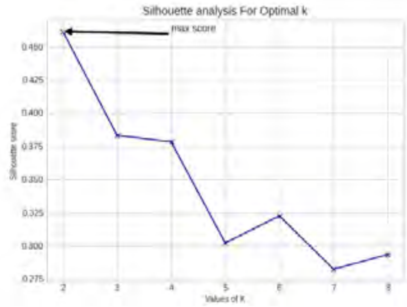
BoxplotforDistributionofESGScores2013



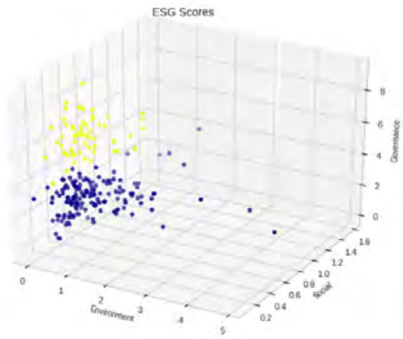
CorrelationMatrix2013



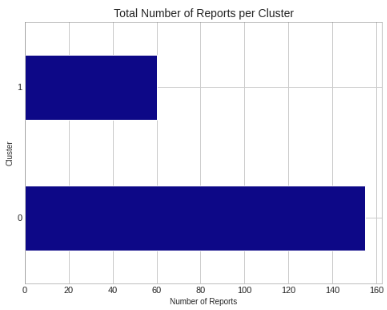
ElbowMethod(K-Medoids,2013)



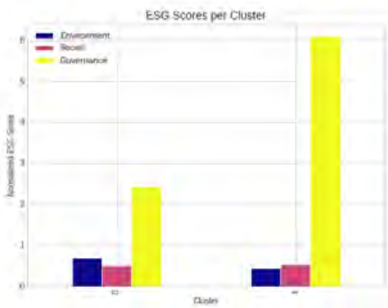
SilhouettePlot((K-Medoids,2013)



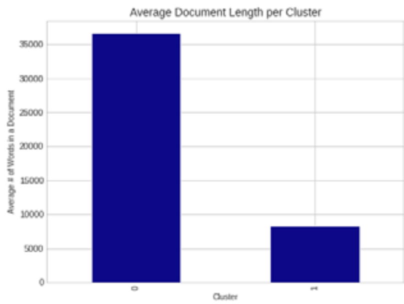
3-dplotofdatapoints(K-Medoids,2013)



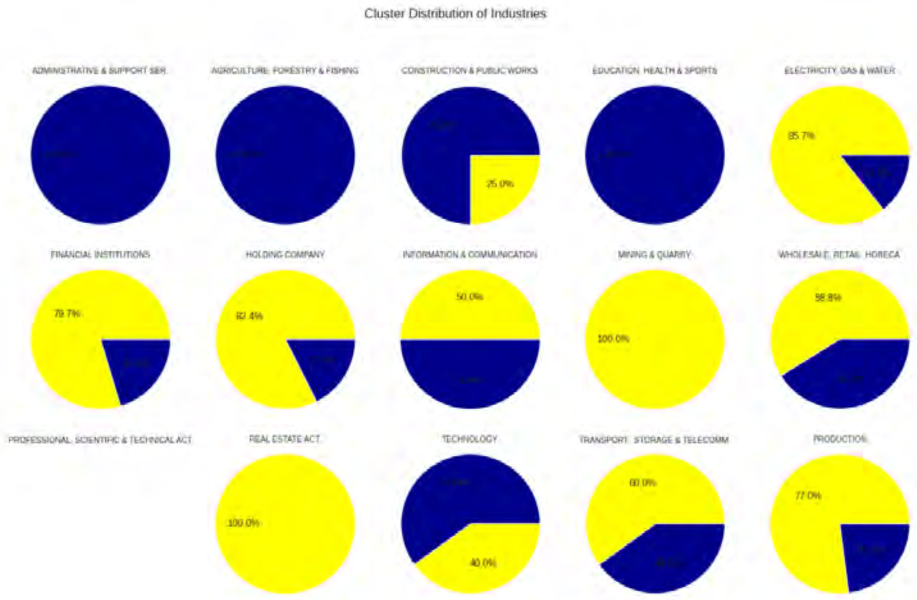
TotalNumberofReportsperCluster(K-Medoids,2013)



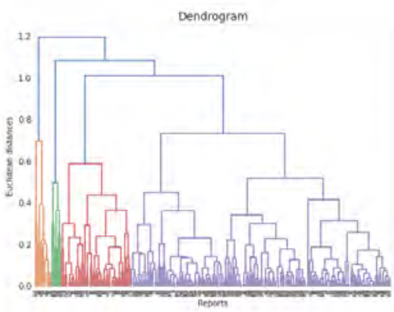
ESGScoresPerCluster(K-Medoids,2013)



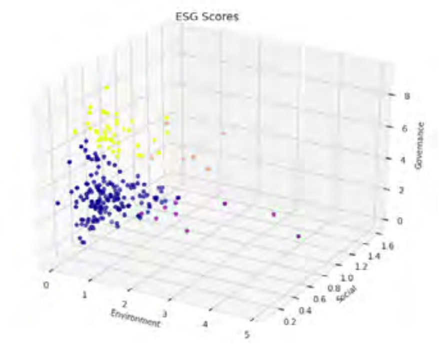
AverageDocumentLengthperCluster(K-Medoids,2013)



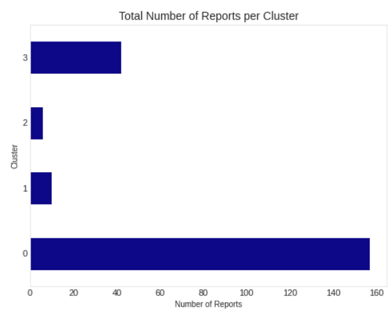
ClusterDistributionofIndustries(K-Medoids,2013)



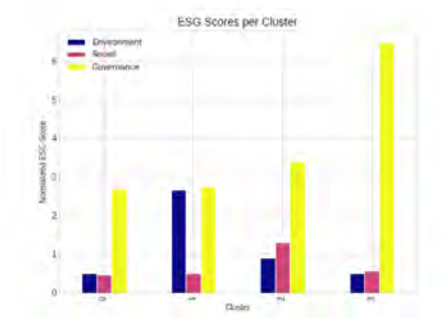
Dendrogram(Hierarchical,2013)



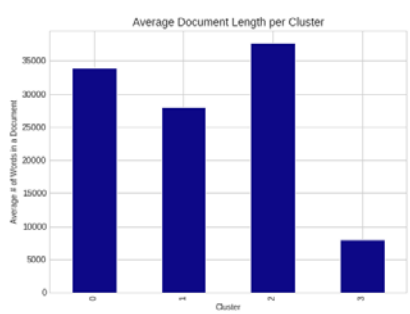
3-dplotofdatapoints(Hierarchical,2013)



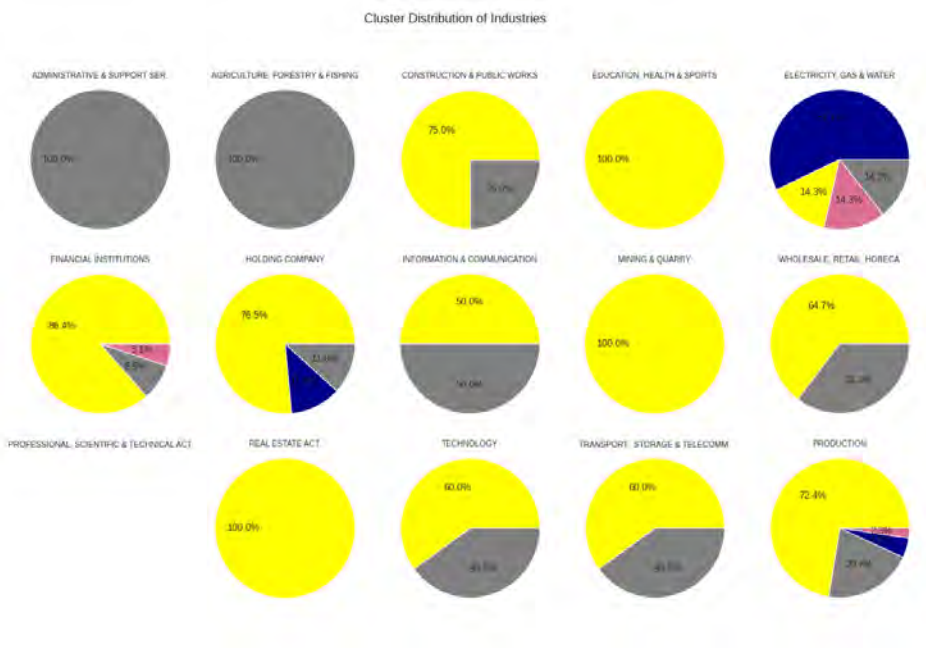
TotalNumberofReportsperCluster(Hierarchical,2013)



ESGScoresPerCluster(Hierarchical,2013)

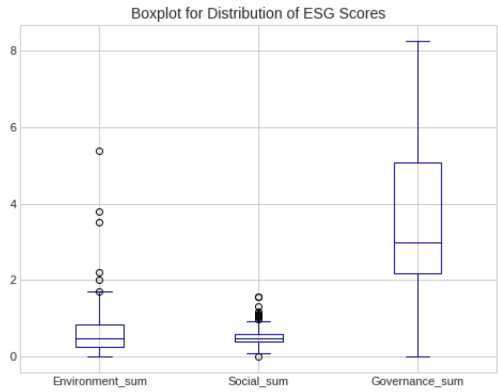


AverageDocumentLengthperCluster(Hierarchical,2013)

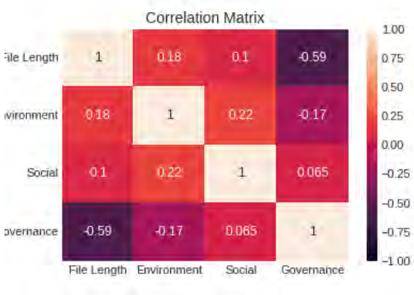


ClusterDistributionofIndustries(Hierarchical,2013)

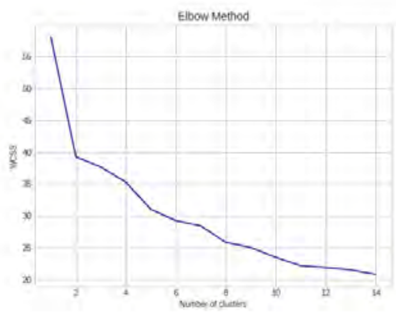
2014



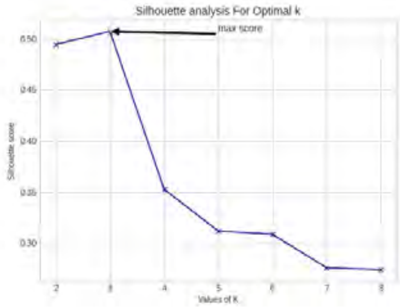
BoxplotforDistributionofESGScores2014



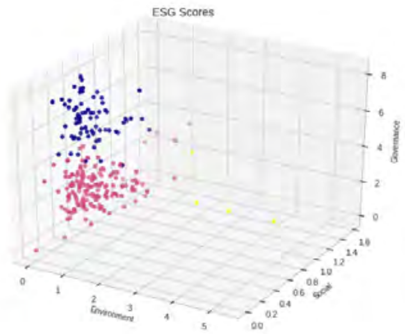
CorrelationMatrix2014



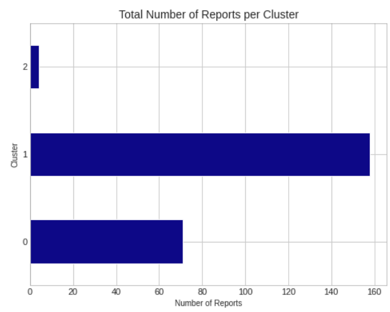
ElbowMethod(K-Medoids,2014)



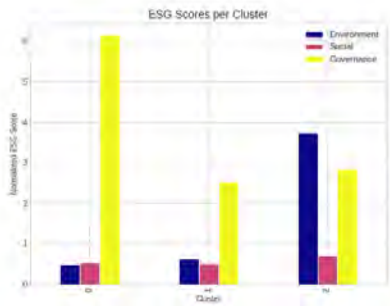
SilhouettePlot((K-Medoids,2014)



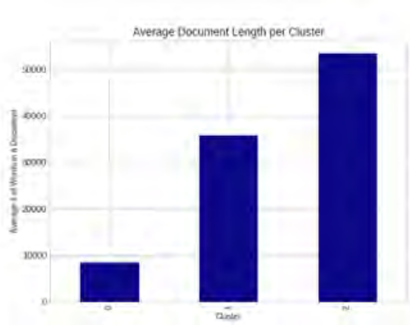
3-dplotofdatapoints(K-Medoids,2014)



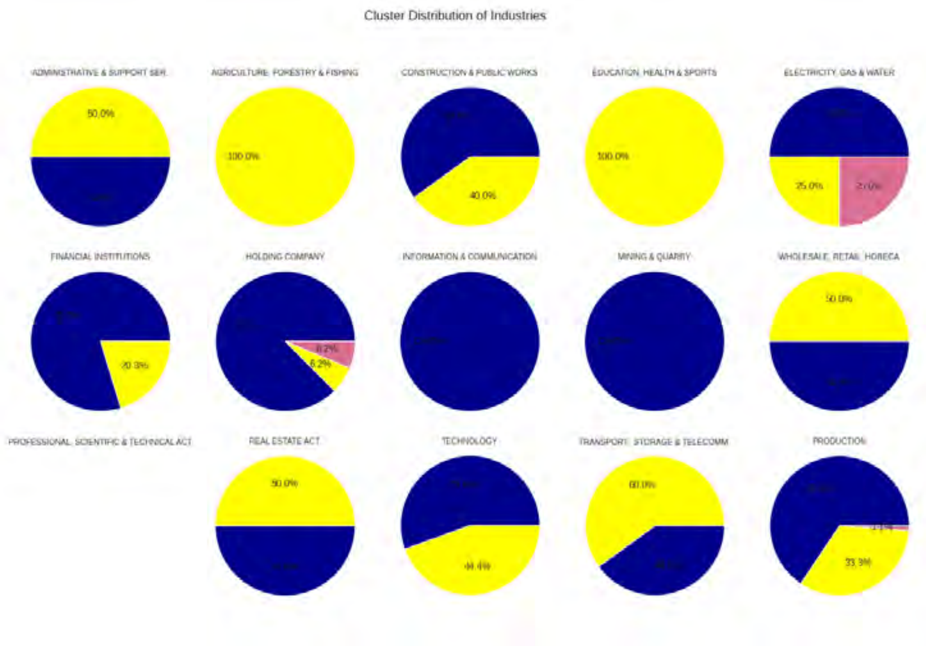
TotalNumberofReportsperCluster(K-Medoids,2014)



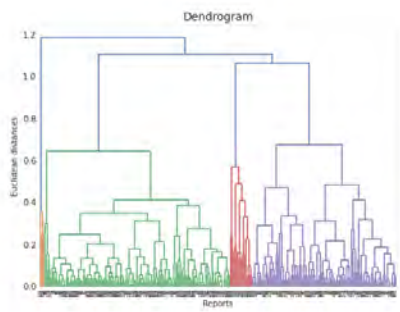
ESGScoresPerCluster(K-Medoids,2014)



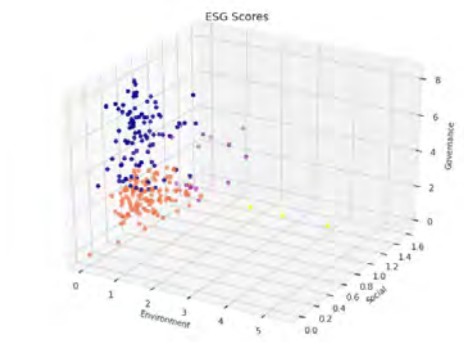
AverageDocumentLengthperCluster(K-Medoids,2014)



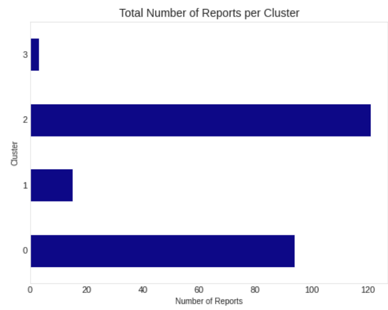
ClusterDistributionofIndustries(K-Medoids,2014)



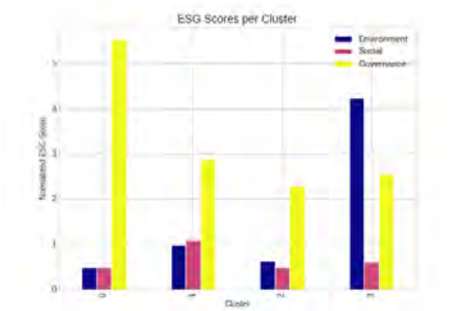
Dendrogram(Hierarchical,2014)



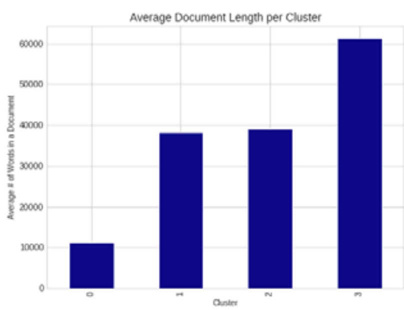
3-dplotofdatapoints(Hierarchical,2014)



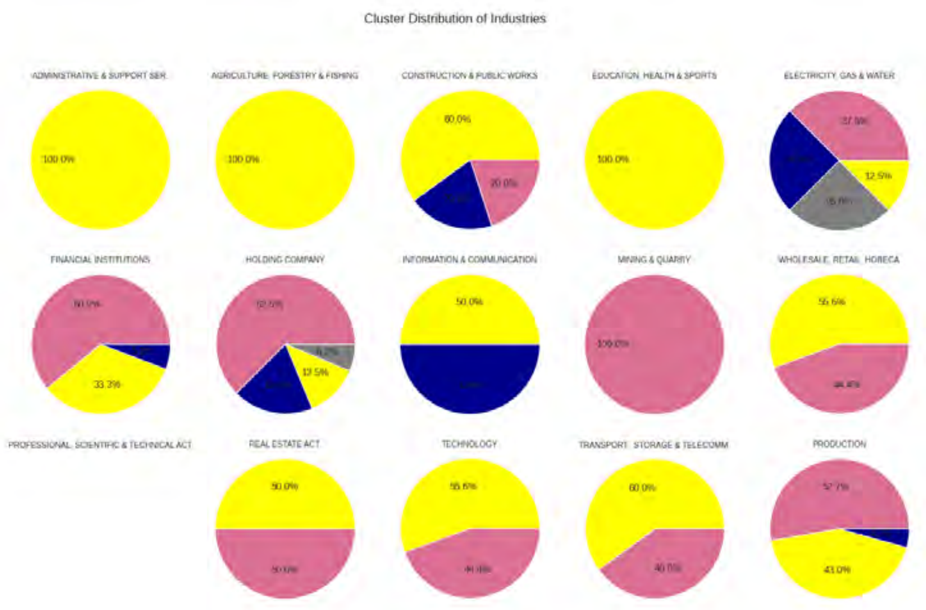
TotalNumberofReportsperCluster(Hierarchical,2014)



ESGScoresPerCluster(Hierarchical,2014)

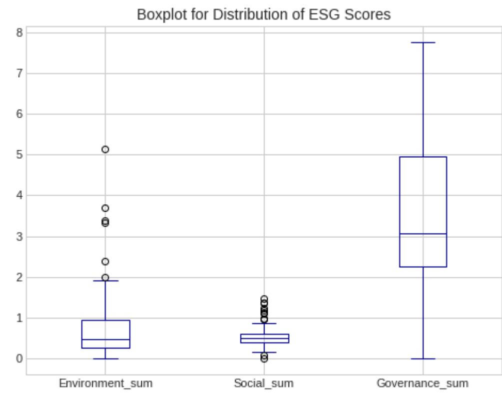


AverageDocumentLengthperCluster(Hierarchical,2014)

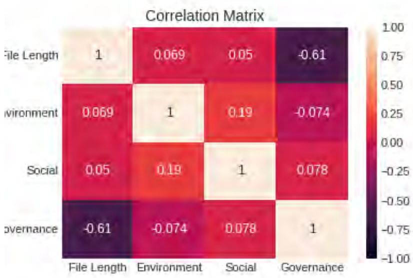


ClusterDistributionofIndustries(Hierarchical,2014)

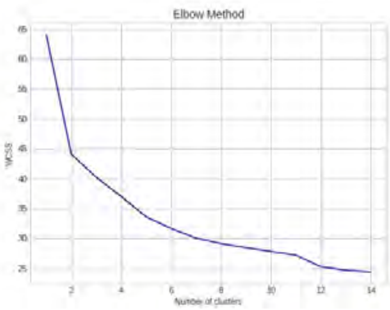
2015



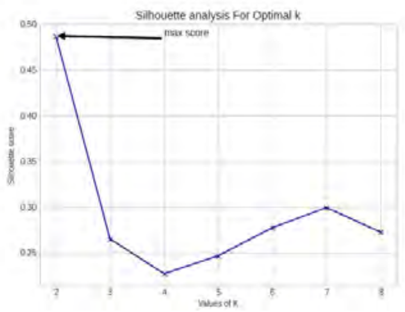
BoxplotforDistributionofESGScores2015



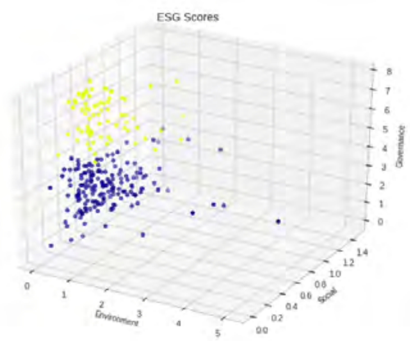
CorrelationMatrix2015



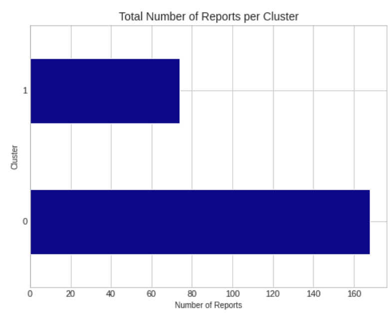
ElbowMethod(K-Medoids,2015)



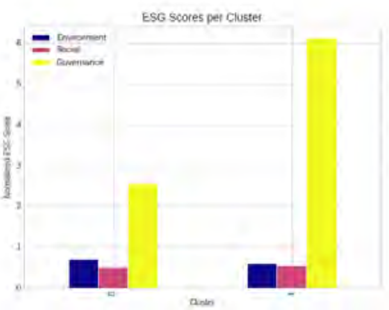
SilhouettePlot((K-Medoids,2015)



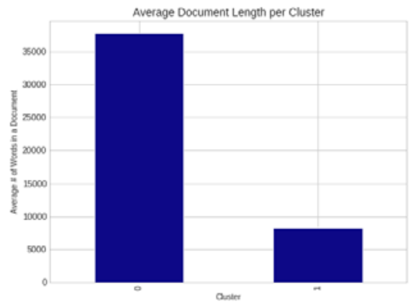
3-dplotofdatapoints(K-Medoids,2015)



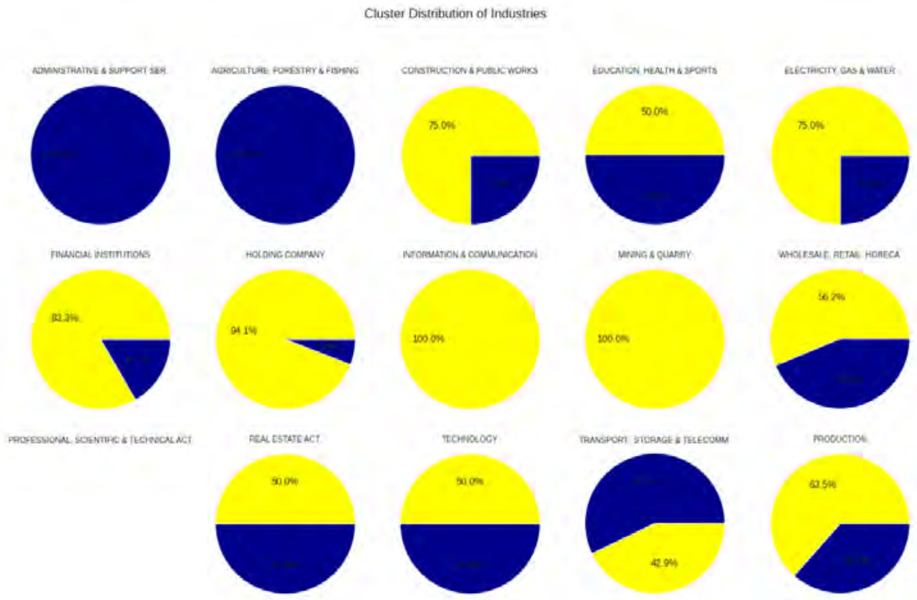
TotalNumberofReportsperCluster(K-Medoids,2015)



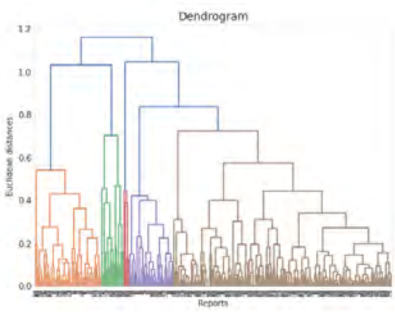
ESGScoresPerCluster(K-Medoids,2015)



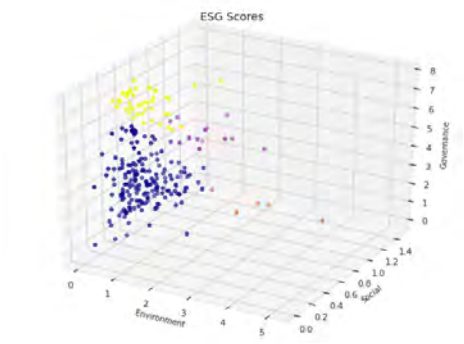
AverageDocumentLengthperCluster(K-Medoids,2015)



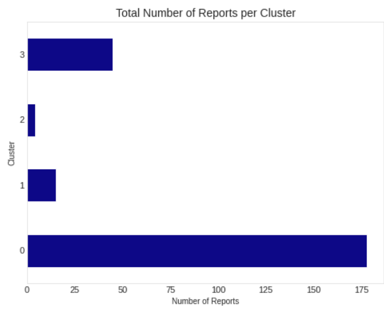
ClusterDistributionofIndustries(K-Medoids,2015)



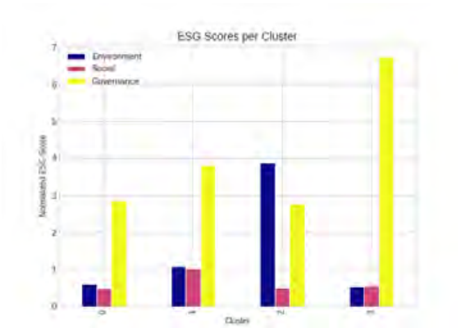
Dendrogram(Hierarchical,2015)



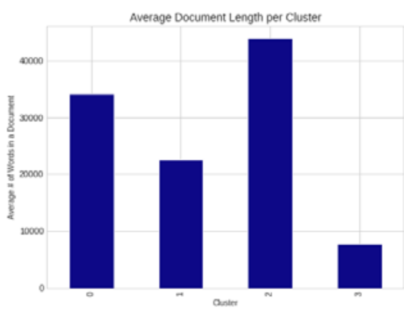
3-dplotofdatapoints(Hierarchical,2015)



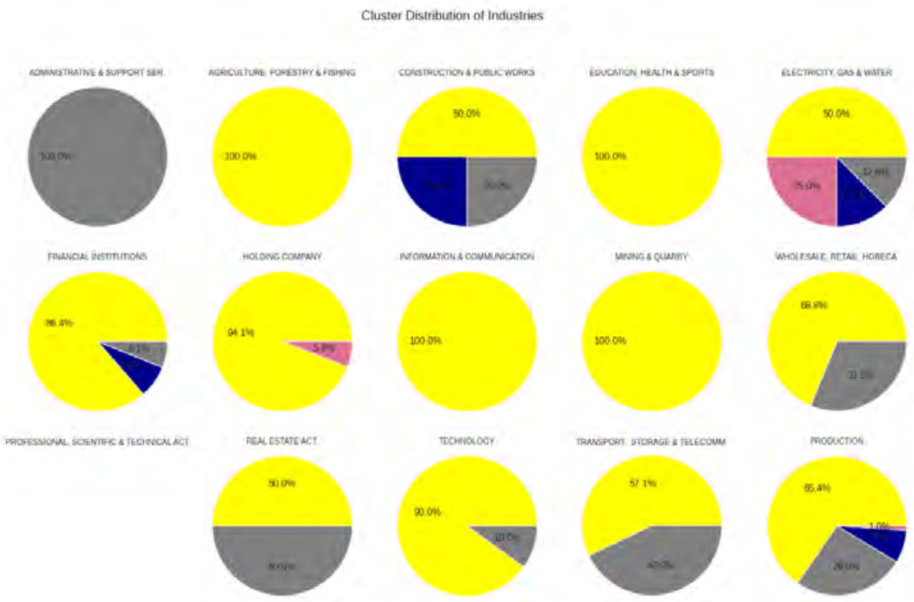
TotalNumberofReportsperCluster(Hierarchical,2015)



ESGScoresPerCluster(Hierarchical,2015)

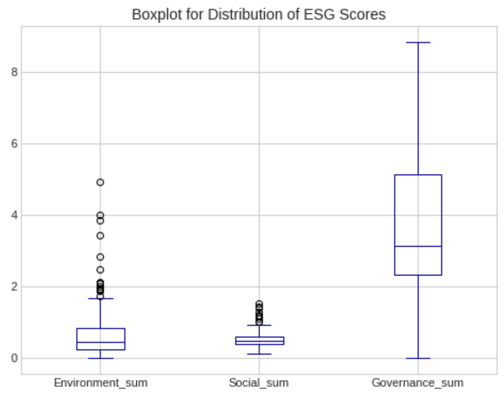


AverageDocumentLengthperCluster(Hierarchical,2015)

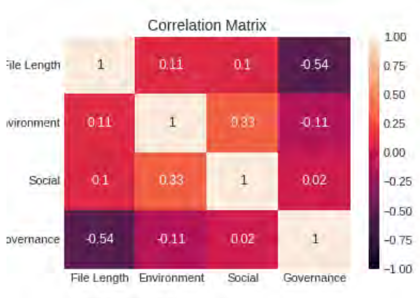


ClusterDistributionofIndustries(Hierarchical,2015)

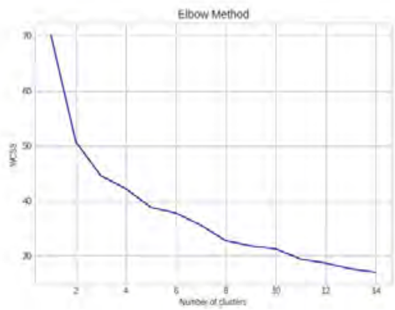
2016



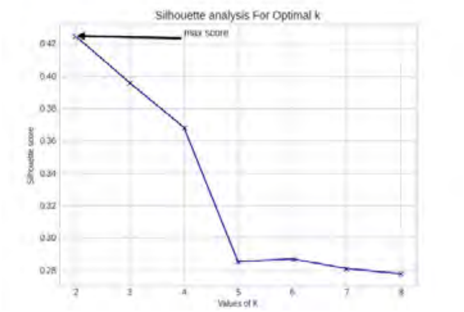
BoxplotforDistributionofESGScores2016



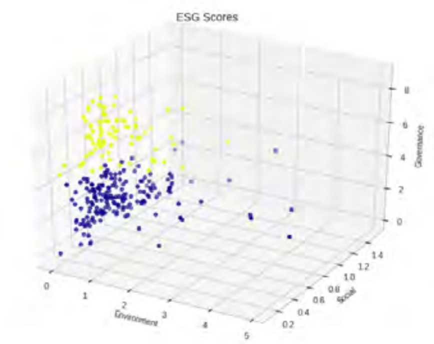
CorrelationMatrix2016



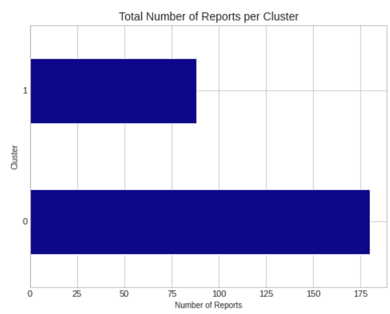
ElbowMethod(K-Medoids,2016)



SilhouettePlot((K-Medoids,2016)

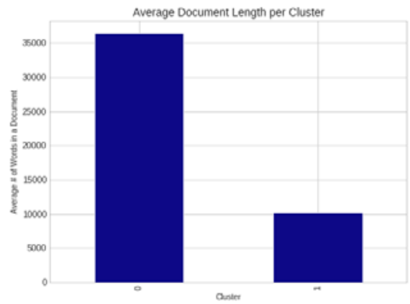


3-dplotofdatapoints(K-Medoids,2016)

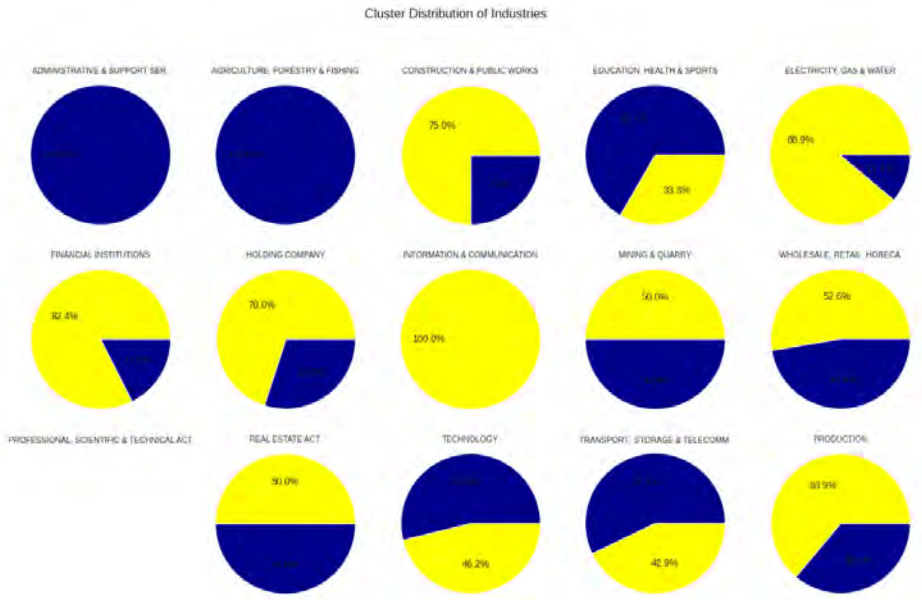


TotalNumberofReportsperCluster(K-Medoids,2016)

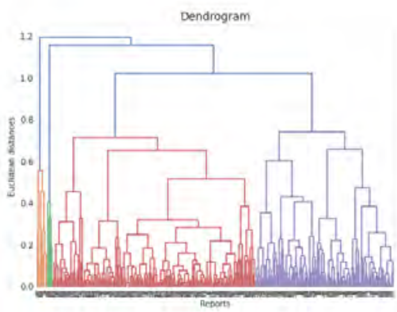
ESGScoresPerCluster(K-Medoids,2016)



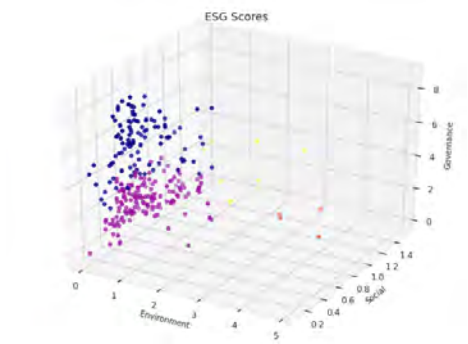
AverageDocumentLengthperCluster(K-Medoids,2016)



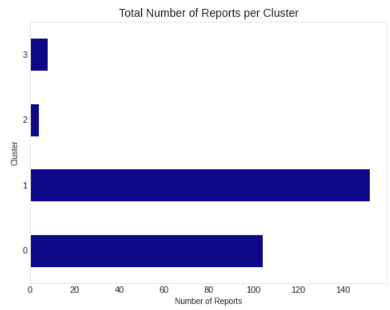
ClusterDistributionofIndustries(K-Medoids,2016)



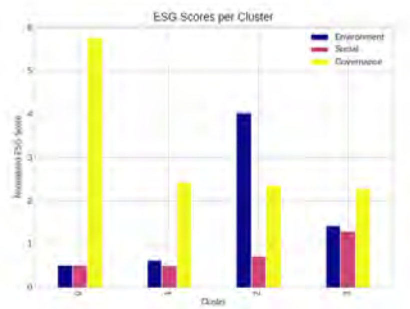
Dendrogram(Hierarchical,2016)



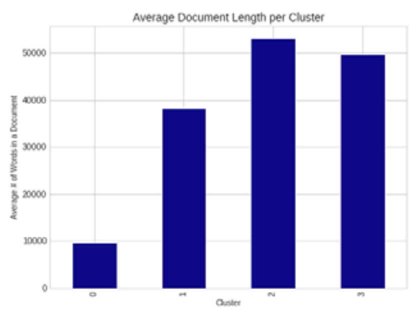
3-dplotofdatapoints(Hierarchical,2016)



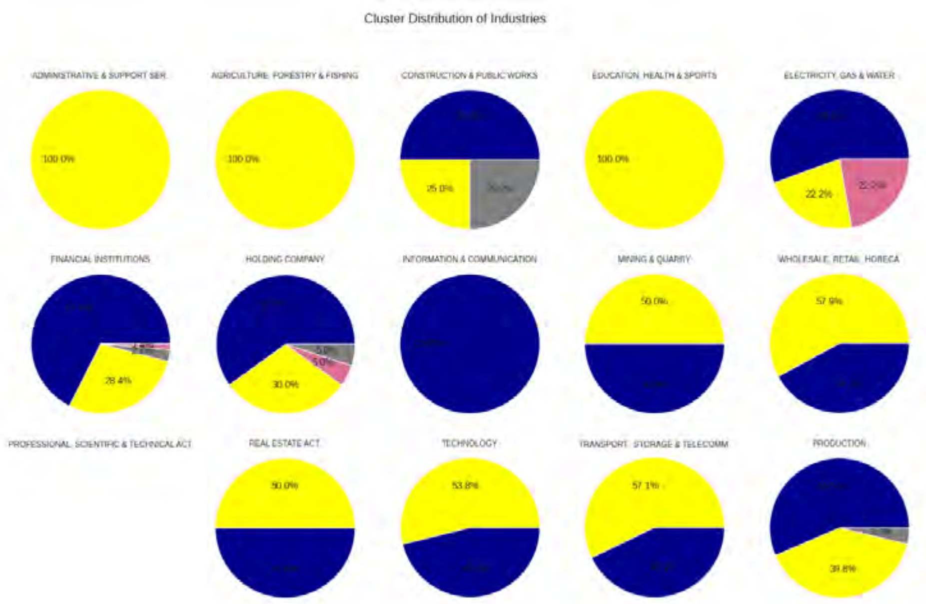
TotalNumberofReportsperCluster(Hierarchical,2016)



ESGScoresPerCluster(Hierarchical,2016)

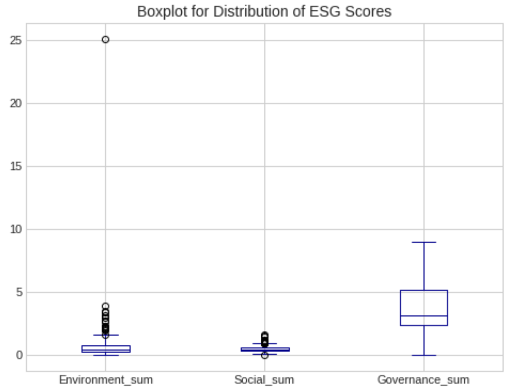


AverageDocumentLengthperCluster(Hierarchical,2016)

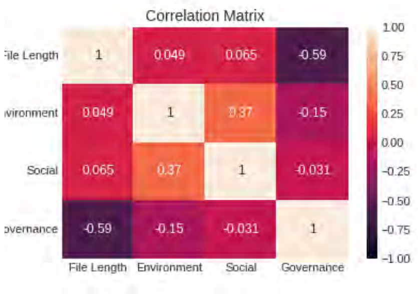


ClusterDistributionofIndustries(Hierarchical,2016)

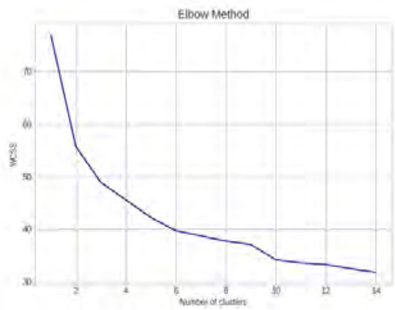
2017



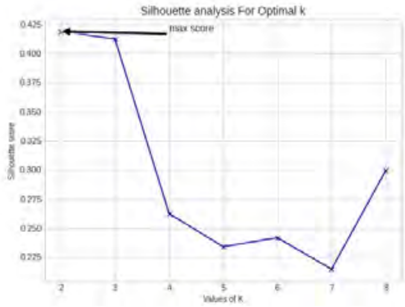
BoxplotforDistributionofESGScores2017



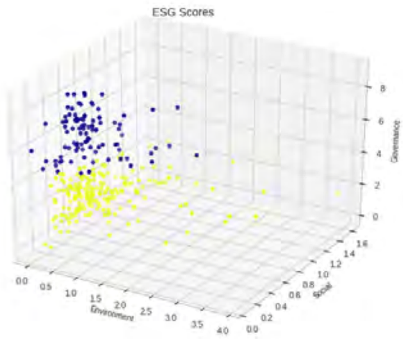
CorrelationMatrix2017



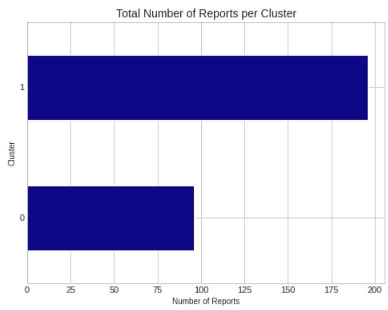
ElbowMethod(K-Medoids,2017)



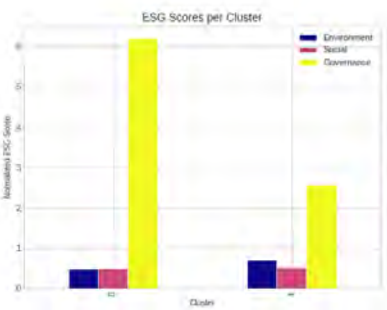
SilhouettePlot((K-Medoids,2017)



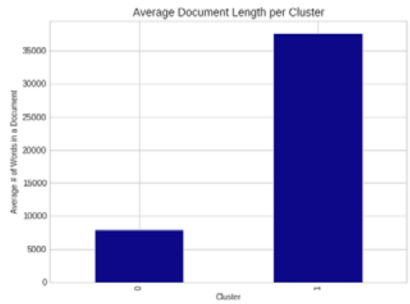
3-dplotofdatapoints(K-Medoids,2017)



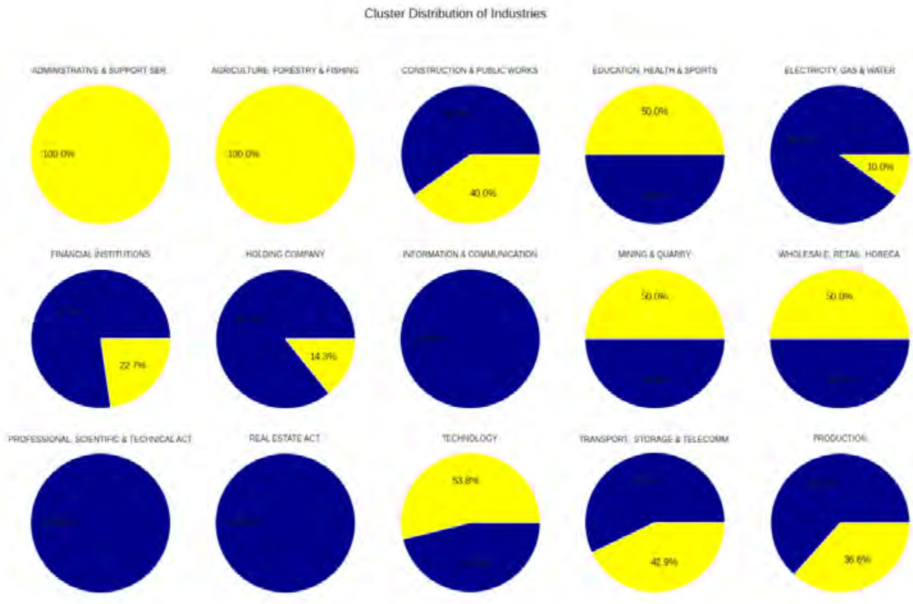
TotalNumberofReportsperCluster(K-Medoids,2017)



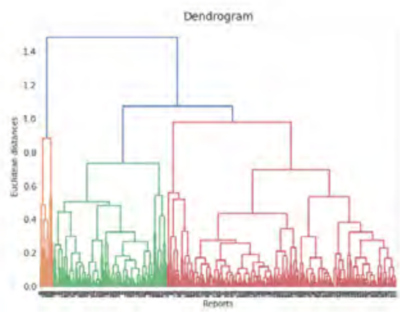
ESGScoresPerCluster(K-Medoids,2017)



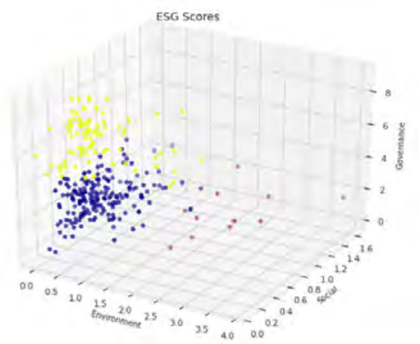
AverageDocumentLengthperCluster(K-Medoids,2017)



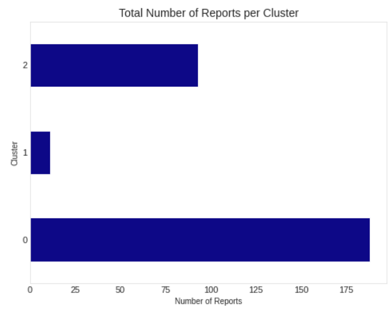
ClusterDistributionofIndustries(K-Medoids,2017)



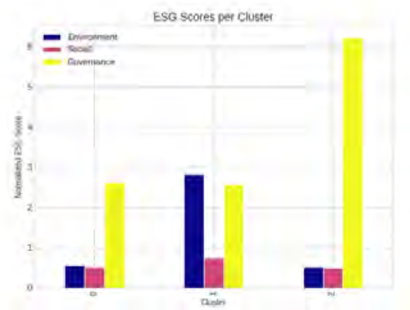
Dendrogram(Hierarchical,2017)



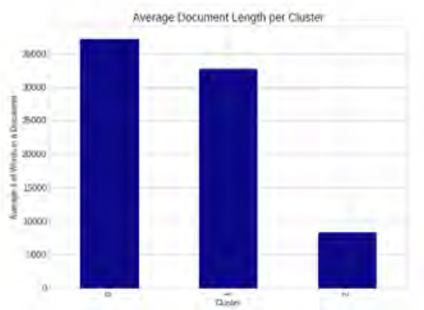
3-dplotofdatapoints(Hierarchical,2017)



TotalNumberofReportsperCluster(Hierarchical,2017)



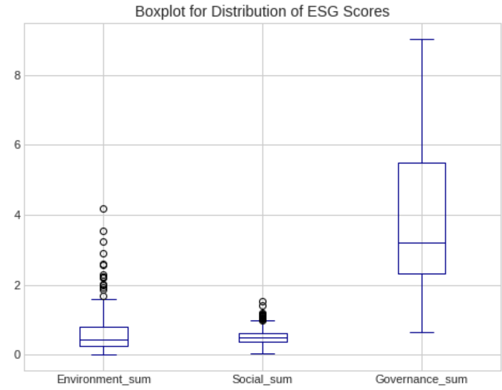
ESGScoresPerCluster(Hierarchical,2017)



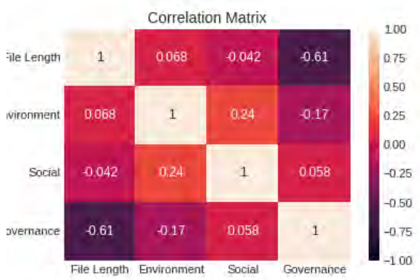
AverageDocumentLengthperCluster(Hierarchical,2017)

ClusterDistributionofIndustries(Hierarchical,2017)

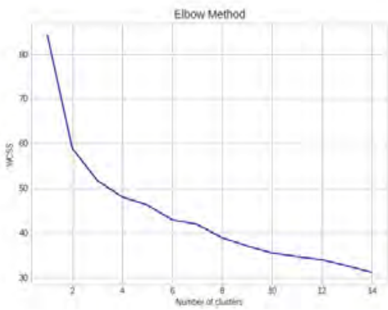
2018



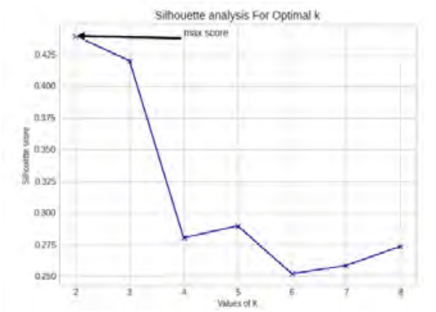
BoxplotforDistributionofESGScores2018



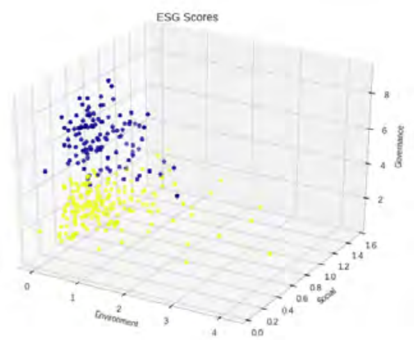
CorrelationMatrix2018



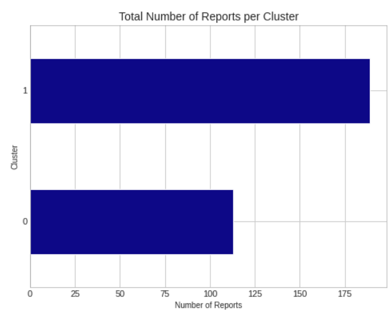
ElbowMethod(K-Medoids,2018)



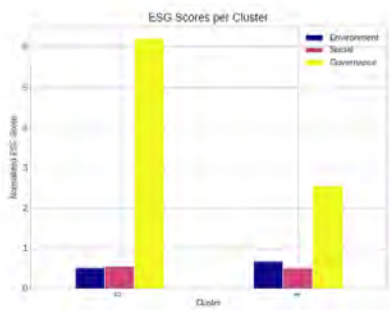
SilhouettePlot((K-Medoids,2018)



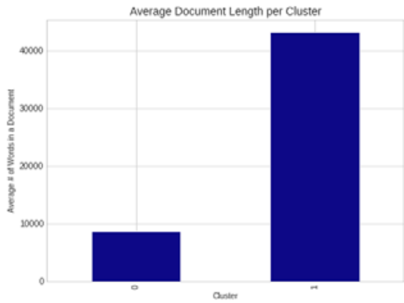
3-dplotofdatapoints(K-Medoids,2018)



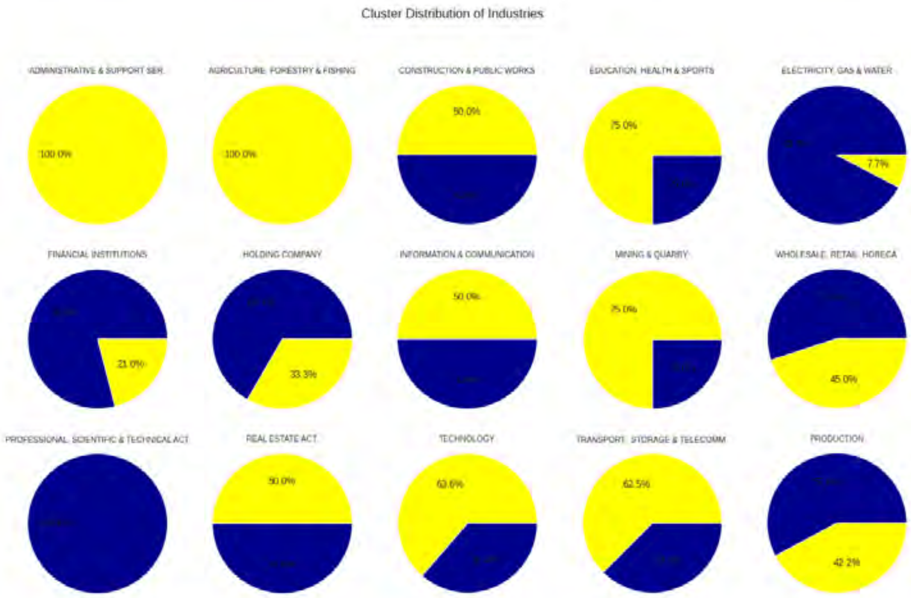
TotalNumberofReportsperCluster(K-Medoids,2018)



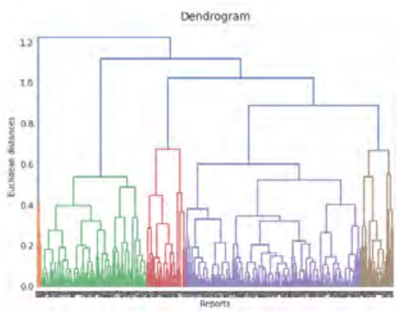
ESGScoresPerCluster(K-Medoids,2018)



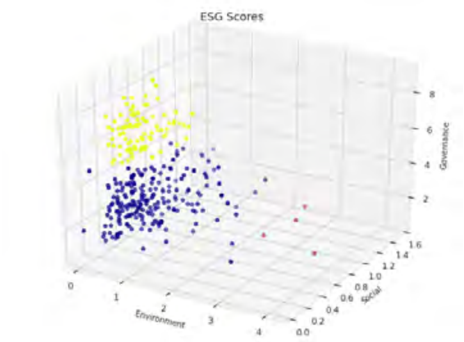
AverageDocumentLengthperCluster(K-Medoids,2018)



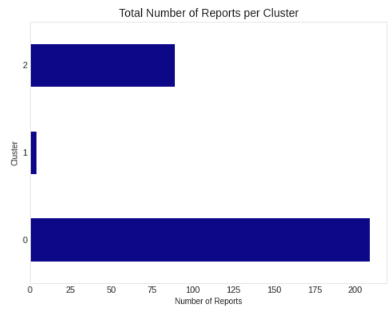
ClusterDistributionofIndustries(K-Medoids,2018)



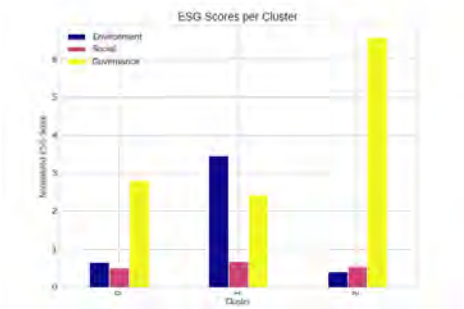
Dendrogram(Hierarchical,2018)



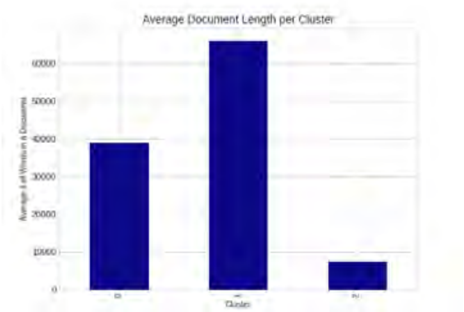
3-dplotofdatapoints(Hierarchical,2018)



TotalNumberofReportsperCluster(Hierarchical,2018)



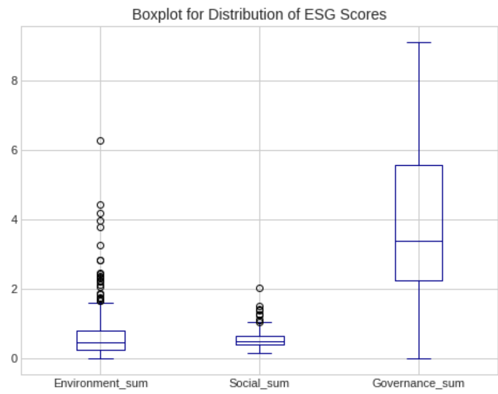
ESGScoresPerCluster(Hierarchical,2018)



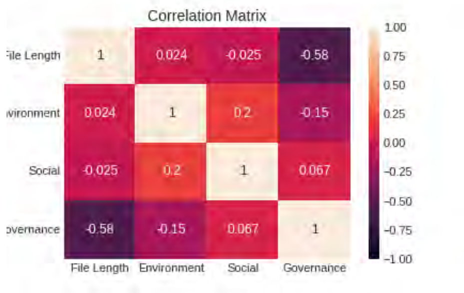
AverageDocumentLengthperCluster(Hierarchical,2018)

ClusterDistributionofIndustries(Hierarchical,2018)

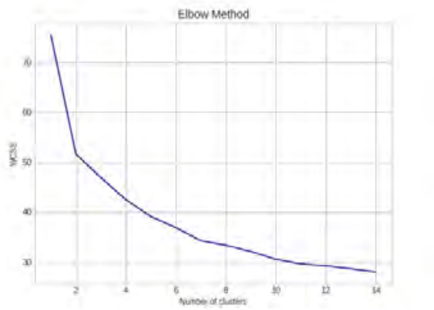
2019



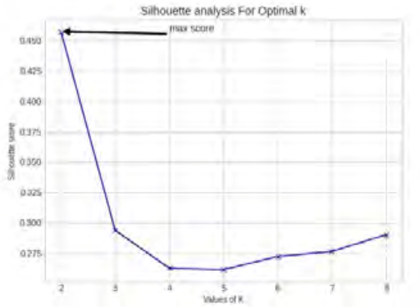
BoxplotforDistributionofESGScores2019



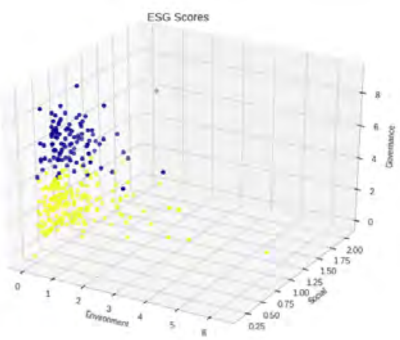
CorrelationMatrix2019



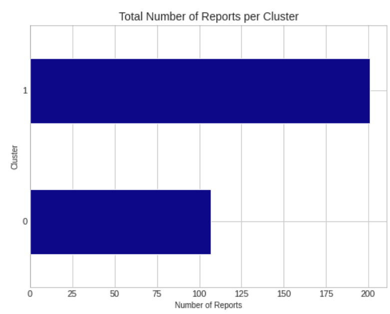
ElbowMethod(K-Medoids,2019)



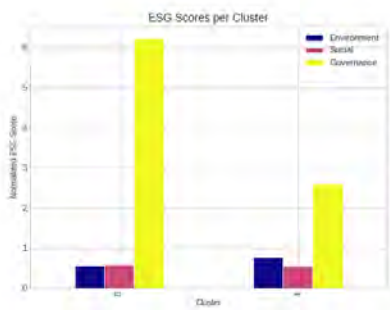
SilhouettePlot((K-Medoids,2019)



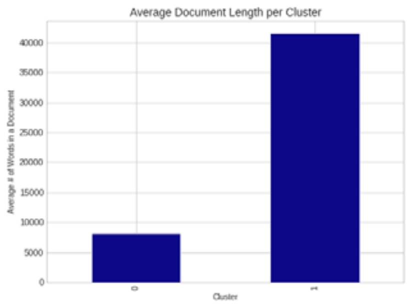
3-dplotofdatapoints(K-Medoids,2019)



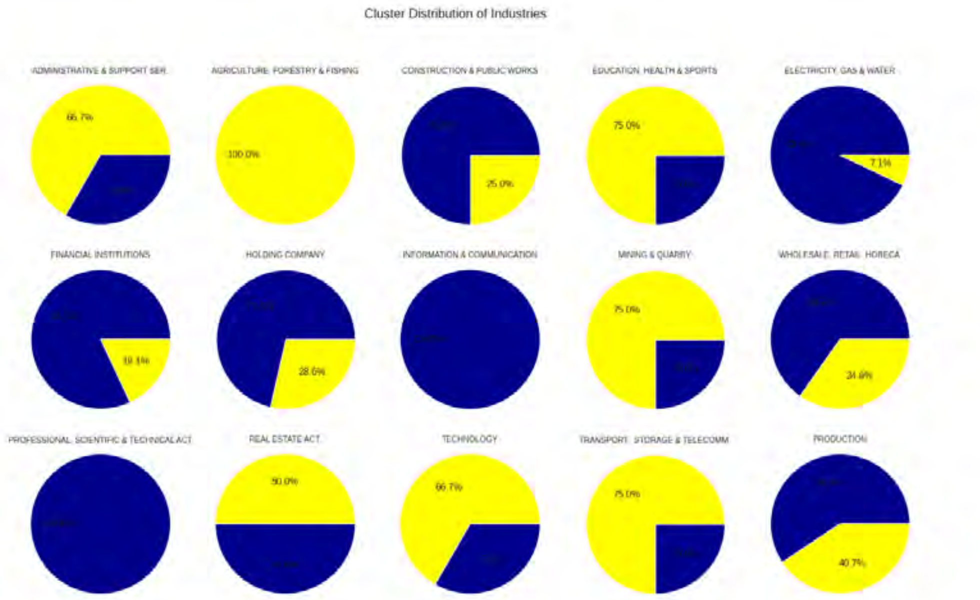
TotalNumberofReportsperCluster(K-Medoids,2019)



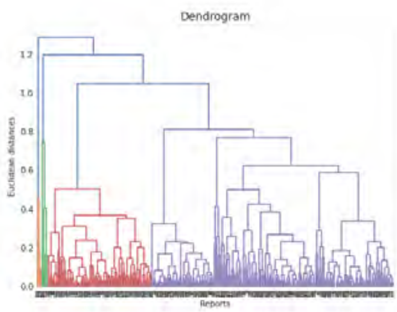
ESGScoresPerCluster(K-Medoids,2019)



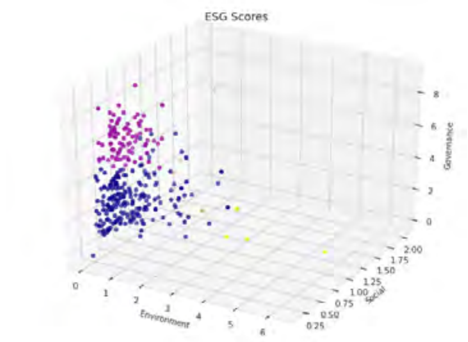
AverageDocumentLengthperCluster(K-Medoids,2019)



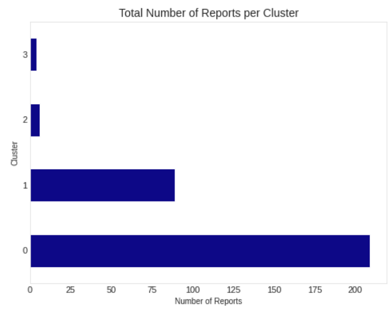
ClusterDistributionofIndustries(K-Medoids,2019)



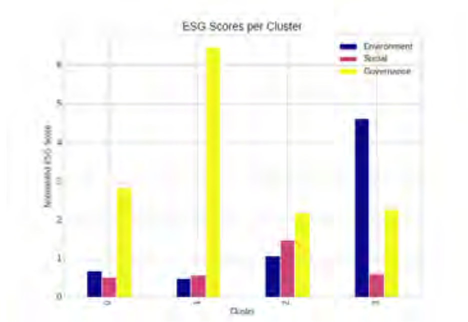
Dendrogram(Hierarchical,2019)



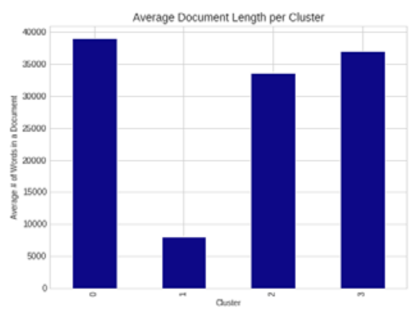
3-dplotofdatapoints(Hierarchical,2019)



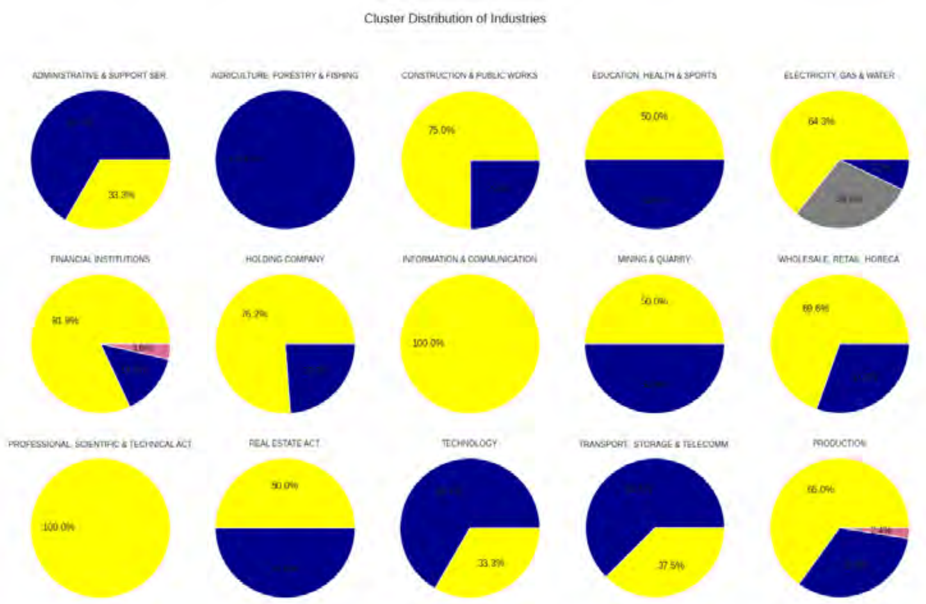
TotalNumberofReportsperCluster(Hierarchical,2019)



ESGScoresPerCluster(Hierarchical,2019)

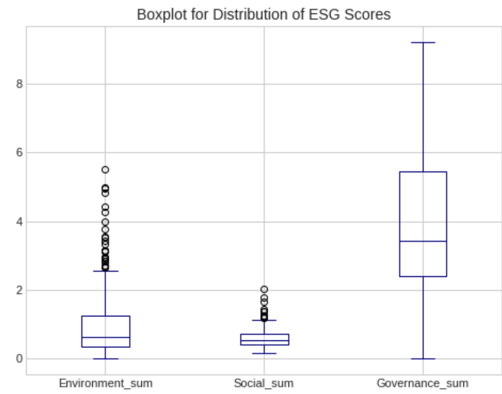


AverageDocumentLengthperCluster(Hierarchical,2019)

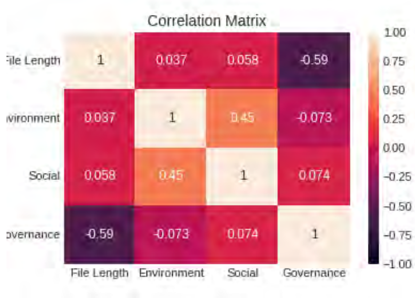


ClusterDistributionofIndustries(Hierarchical,2019)

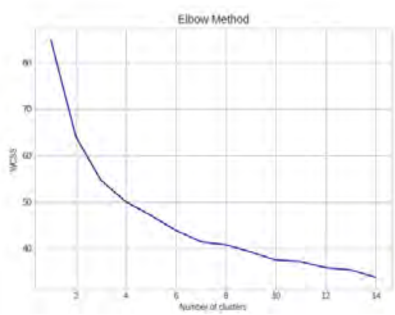
2020



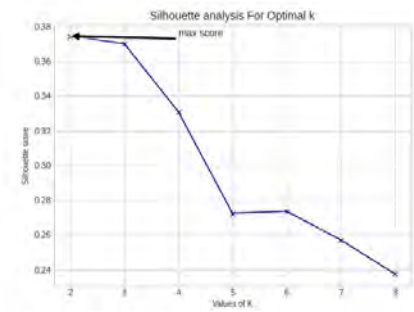
BoxplotforDistributionofESGScores2020



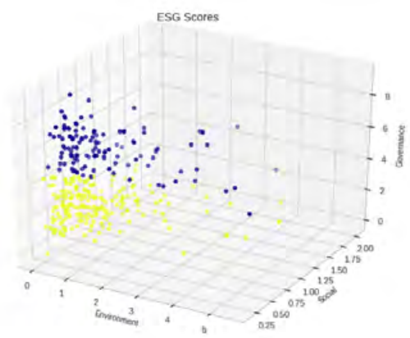
CorrelationMatrix2020



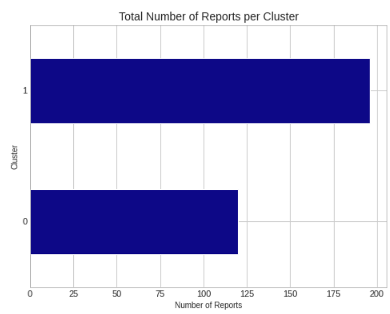
ElbowMethod(K-Medoids,2020)



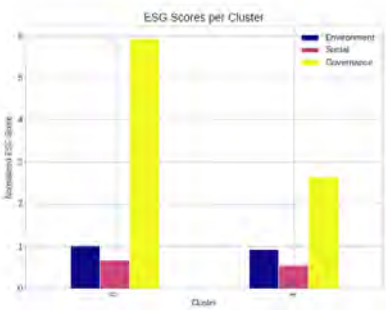
SilhouettePlot((K-Medoids,2020)



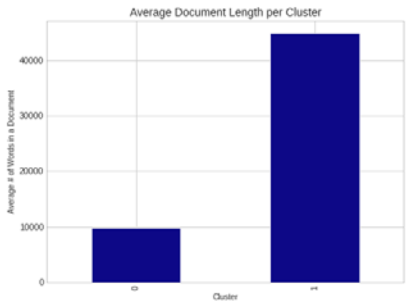
3-dplotofdatapoints(K-Medoids,2020)



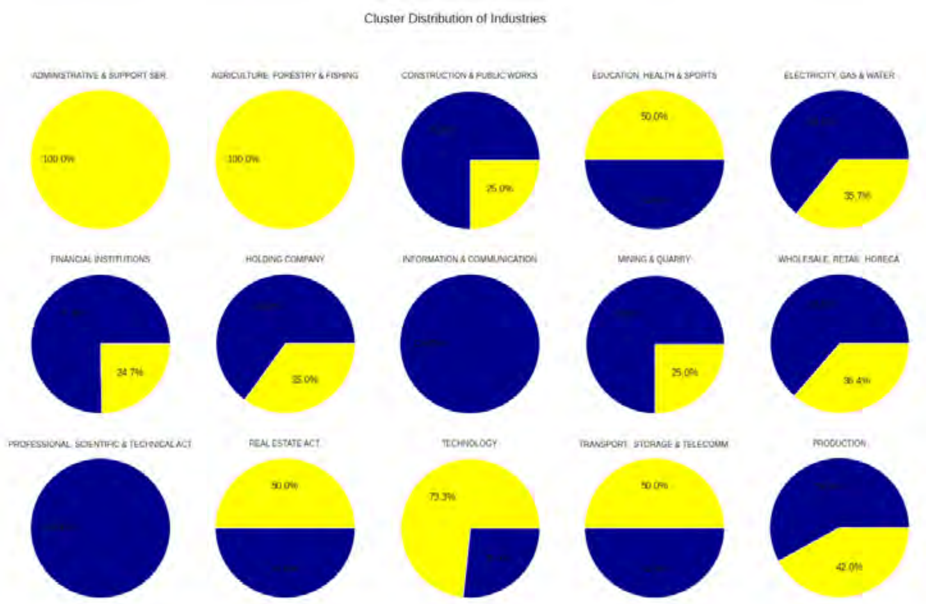
TotalNumberofReportsperCluster(K-Medoids,2020)



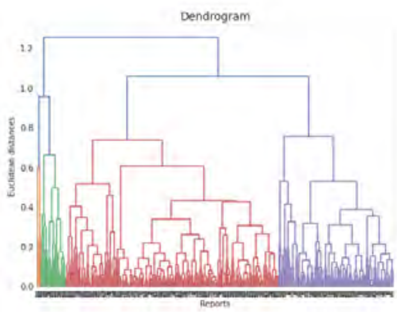
ESGScoresPerCluster(K-Medoids,2020)



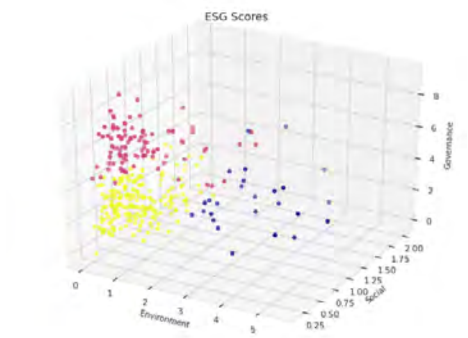
AverageDocumentLengthperCluster(K-Medoids,2020)



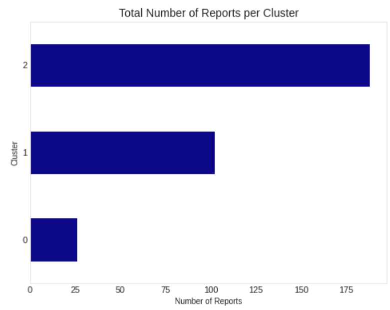
ClusterDistributionofIndustries(K-Medoids,2020)



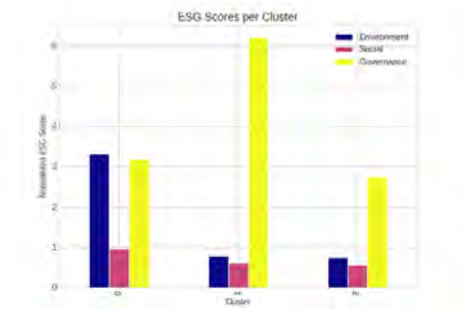
Dendrogram(Hierarchical,2020)



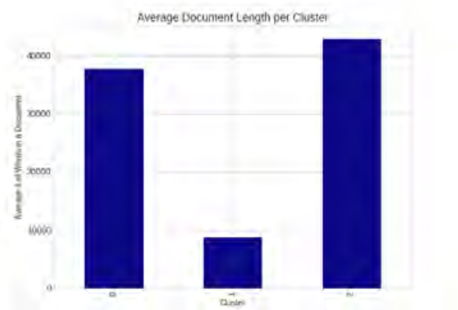
3-dplotofdatapoints(Hierarchical,2020)



TotalNumberofReportsperCluster(Hierarchical,2020)



ESGScoresPerCluster(Hierarchical,2020)



AverageDocumentLengthperCluster(Hierarchical,2020)

ClusterDistributionofIndustries(Hierarchical,2020)