Data analytics final report - Investigating the Correlation Between Sustainability and Profitability

Reza Nosrati, and Prabhakar Karna

Master of Science (Business Analytics) / University of North Florida

CEN6940 Computing Practicum / School of Computing

Date 12/10/2024

Index Terms—Data mining, machine learning, Prediction of price, ESG, Sustainability market, ROE, Regression.

Table of Contents

Executive summary

The project titled "Investigating the Correlation Between Sustainability and Profitability" was conducted by Sustainability Team at the UNF, under the CEN6940 Computing Practicum course. The primary goal was to explore the relationship between companies' sustainability efforts, quantified through Environmental, Social, and Governance (ESG) criteria, and their financial performance. This data analytics research is particularly relevant as both consumers and investors are increasingly considering sustainability as a factor in their decision-making processes.

Utilizing data from over 700 companies, sourced from S&P for ESG scores and Yahoo Finance for financial metrics, the team employed advanced machine learning models, including Random Forest, to analyze and predict the financial outcomes based on companies' ESG performance. The analysis tools were built using Python and JMP software, accommodating complex, multi-source data integration and handling various data-related challenges such as missing values and normalization.

The study revealed nuanced results where social sustainability scores were positively correlated with profitability, while environmental and governance aspects showed mixed effects depending on the sector. Particularly, industries like energy and manufacturing displayed a more pronounced correlation between high ESG scores and increased financial gains.

Despite facing significant challenges such as integrating diverse data formats and dealing, the project innovated in the application of real-time data analytics and sector-specific sustainability impact analysis. The findings suggest that while ESG factors are increasingly important, their impact on profitability can vary greatly by industry for example the Energy sector excel in terms of high profits and high ESG scores.

Looking forward, the project recommends expanding the analytic models to encompass longer-term data and broader industry representation to better understand the dynamics between sustainability practices and profitability. Furthermore, developing tools for real-time tracking of ESG performance could empower more informed strategic decisions in business and investment, aligning financial objectives with sustainability goals.

Introduction

In today's rapidly evolving global market, businesses face increasing scrutiny not only for their financial performance but also for their impact on the environment, society, and governance (ESG). Despite the growing emphasis on sustainability and corporate responsibility, many organizations struggle to integrate ESG criteria effectively into their business models. This leads to risks including regulatory penalties, investor disengagement, and a tarnished public image.

Importance of Addressing the Problem

Addressing ESG integration effectively is critical for several reasons:

1. **Regulatory Compliance**: With governments worldwide tightening environmental regulations and setting strict standards for corporate behavior, companies must adapt to these changes to avoid legal repercussions and financial losses.

2. **Investor Confidence**: Investors are increasingly factoring ESG performance into their investment decisions. Companies that demonstrate robust ESG practices tend to attract more investment and achieve better market valuation.

3. **Consumer Demand**: Modern consumers “Gen Z” are more informed and sensitive to the social and environmental impacts of the products they purchase. Companies that prioritize ESG criteria can enhance brand loyalty and competitive advantage.

4. **Risk Management**: Proactive ESG practices help companies foresee and mitigate risks related to environmental disasters, social unrest, and governance scandals. This preparation can safeguard the company's long-term sustainability and profitability.

5. **Global Impact**: Addressing ESG issues contributes to global efforts like the United Nations Sustainable Development Goals (SDGs), aimed at achieving a more sustainable and equitable world. By integrating ESG criteria, companies can play a crucial role in tackling global challenges such as climate change, inequality, and injustice.

To effectively integrate ESG (Environmental, Social, and Governance) criteria with financial performance, companies should focus on strategic alignment, transparent reporting, and stakeholder engagement. This involves setting clear ESG targets that align with business goals, adopting recognized reporting standards for accountability, and engaging stakeholders to balance ESG commitments with financial outcomes. Leveraging technology for better data management and incorporating ESG factors into decision-making processes can also enhance overall business sustainability and profitability.

This project aims to explore the potential relationship between ESG performance and profitability metrics like the profit/earnings ratio. The need for this study arises from the growing interest in sustainable investing and the lack of conclusive data on whether sustainable practices translate into financial gain.

- ✓ **Null hypothesis 1:** Companies with higher ESG scores have a positive correlation with profitability and lower financial risks.
- ✓ **Null hypothesis 2:** ESG performance has a stronger correlation with profitability in specific sectors (e.g., energy, manufacturing).
- ✓ Developing a Predictive Model

Value Proposition

This project tested null hypothesis and delivered predictive models based on the use of various analytical tools which we studied as a part of our Business Analytics curriculum. By leveraging machine learning, big data analytics, and real-time monitoring, this project addressed the challenges of assessing and predicting how ESG performance influences financial outcomes. The project's outcomes will have a tangible and lasting impact on stakeholders, as outlined below.

- **Informed Decision-Making for Investors:** Predictive models will provide forward-looking insights into how ESG initiatives impact profitability, stock performance, and risk, helping investors make data-driven decisions and optimize portfolios for better returns.

- **Simplified and Actionable ESG Insights for Investors:** Visualization tools will offer clear, customizable ESG data, allowing investors to filter by sectors, regions, or criteria to make more informed investment choices and enhance profitability.
- **Risk Mitigation and Early Warning Systems:** Real-time AI-powered tools will give investors and companies up-to-date ESG insights, helping them proactively address risks like regulatory changes and environmental disasters, improving financial stability and sustainability.
- **Long-Term Financial Stability and Growth for Companies:** Predictive models will help companies identify areas for ESG improvement, leading to cost savings, operational efficiency, and revenue growth, ultimately boosting brand equity and shareholder value.

Contribution to the Existing Body of Knowledge:

This project make several contributions to the existing body of knowledge on ESG and financial performance.

Advanced Predictive Analytics: While previous studies such as Friede, Busch, and Bassen (2015) have confirmed the historical link between ESG performance and financial returns, our project will develop predictive models using machine learning to estimate future financial outcomes based on ESG scores. This will provide forward-looking insights that can assist investors in making more informed decisions about sustainability and financial risk.

Sector-Specific and Region-Specific Analysis: Building on research by Khan, Serafeim, and Yoon (2016), which emphasized the materiality of ESG issues depending on industry, our study can be extended to conduct sectoral and regional analyses to provide a more nuanced understanding of how ESG factors influence corporate profitability in different contexts. This will offer greater clarity for investors who need to tailor their sustainability strategies according to industry or geographic considerations.

Creation of Decision-Making Tools: As noted by Luff and Shimkus (2021), there is a growing need for practical tools that investors can use to evaluate ESG metrics in real time. Our project will create a dashboard or visualization tool that presents ESG performance data alongside financial indicators, providing investors and decision-makers with an accessible and actionable tool for analyzing the sustainability and profitability of companies. Below picture shows how sustainability can increase profit of a company.

Work distribution

The Table highlighted goal was achieved by performing both ML and DV activities.

Goal Number	Project Goal	Objective	Details	Name of the Person
Goal 1	Establish Data Pipelines for ESG and Financial Data Integration	Create data pipelines to automatically collect, clean, and integrate ESG data.	Ensure predictive models are trained on up-to-date and reliable datasets from multiple sources.	Prabhakar/Reza
Goal 2	Develop Visualization Tools for Investor Insights	Create dashboards that display predictive model outcomes for investors.	Provide visual insights into the impact of ESG performance on future profitability and risks.	Reza
Goal 3	Design Investor Decision-Making Tool	Develop a user-friendly, AI-driven tool providing real-time predictions based on ESG scores.	Integrate predictive models, allowing investors to assess future profitability and financial risks.	Prabhakar

Machine Learning (ML) Activities:

Activity Number	Machine Learning Activity	Details	Tool/Technique
ML Activity 1	Data Preprocessing and Cleaning	Prepare raw ESG and financial data for analysis by handling missing values, normalizing variables, and transforming data into the appropriate formats.	Python (pandas, NumPy)
ML Activity 2	Feature Selection and Engineering	Identify important ESG features that influence financial outcomes and create new features (e.g., ratios, interactions) to improve the model's predictive power.	Python (scikit-learn)
ML Activity 3	Model Selection	Random forests is used for predicting financial outcomes based on ESG data.	JMP
ML Activity 4	Model Training and Tuning	Tune model hyperparameters (e.g., learning rate, depth of trees) to optimize performance.	JMP
ML Activity 5	Model Evaluation	Evaluate model performance using metrics like mean squared error (MSE), R-squared, and accuracy etc.	JMP

Data Visualization Activities (using Tableau/Excel Toolkit):

Activity Number	Data Visualization Activity	Details	Type of Visualization
DV Activity 1	Correlation Matrix	Visualize correlations between different ESG factors and financial outcomes to identify significant relationships.	Correlation Matrix
DV Activity 2	Scatter Plot	Display the relationship between individual ESG metrics (e.g., carbon emissions) and financial outcomes (e.g., share price) to analyze potential patterns.	Scatter Plot

Activity Number	**Data Visualization Activity**	**Details**	**Type of Visualization**
DV Activity 3	Time Series Analysis	Visualize how ESG scores and financial performance evolve over time to detect trends and patterns.	Line Chart, Time Series Plot
DV Activity 4	Bar Charts for Sector Comparison	Compare the ESG scores and financial performance across different sectors to see which industries have the highest or lowest ESG impact.	Bar Chart
DV Activity 5	Heatmap for Risk Assessment	Use a heatmap to highlight companies with higher ESG risks and their associated financial outcomes (e.g., low returns, high volatility).	Heatmap
DV Activity 6	Bubble Chart for Multivariate Analysis	Analyze how multiple factors (e.g., ESG score, market cap, financial risk) interact by representing them on a bubble chart with varying sizes and colors.	Bubble Chart

Required Resources used

Software Tools

1. **Python**: Primary programming language for data processing, analysis, and joining.
2. **Microsoft Excel, JMP** software for Clustering & ML predictive model building
3. **Tableau**: For creating interactive and shareable dashboards that visualize the relationship between ESG metrics and financial performance.

Hardware

1. **High-Performance Laptops**: With at least an Intel i7 processor or equivalent, 16GB RAM, and substantial SSD storage to handle large datasets and intensive computations.

Website

1. **Financial & ESG Data**: Yahoo website for accessing real-time financial data, historical market data, and financial statements for Fortune-500 companies. S&P website for accessing ESG data for the same companies.
2. File format of the dataset- CSV (Comma-Separated Values)

Programming Platforms

1. **Jupyter Notebook**: For writing and testing code in an interactive environment which supports Python, and other programming languages. Below is a proposed list of python function/library used during data analysis and predictive modelling

Library/Function	**Description**
NumPy	Provides support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
Matplotlib (pyplot)	Used for creating static, interactive, and animated visualizations in Python.

Library/Function	Description
Seaborn	An extension to Matplotlib that makes it easier to generate certain types of plots and integrates well with pandas' data structures.
Random Forest Regressor (scikit-learn)	A regression tool that uses multiple decision trees to improve predictive accuracy and control over-fitting.
Simple Imputer (scikit-learn)	Provides basic strategies for imputing missing values in a dataset, such as using the mean, median, or most frequent value.
Linear Regression (scikit-learn)	Implements linear regression for fitting a linear model with coefficients to minimize the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation.

2. **For visualization Tableau/Excel (Toolkit) is used to support below visualization.**

Data Visualization Activity	Details	Type of Visualization
Correlation Matrix	Visualize correlations between different ESG factors and financial outcomes to identify significant relationships.	Correlation Matrix
Scatter Plot	Display the relationship between individual ESG metrics (e.g., carbon emissions) and financial outcomes (e.g., share price) to analyze potential patterns.	Scatter Plot
Time Series Analysis	Visualize how ESG scores and financial performance evolve over time to detect trends and patterns.	Line Chart, Time Series Plot
Bar Charts for Sector Comparison	Compare the ESG scores and financial performance across different sectors to see which industries have the highest or lowest ESG impact.	Bar Chart
Heatmap for Risk Assessment	Use a heatmap to highlight companies with higher ESG risks and their associated financial outcomes (e.g., low returns, high volatility).	Heatmap
Bubble Chart for Multivariate Analysis	Analyze how multiple factors (e.g., ESG score, market cap, financial risk) interact by representing them on a bubble chart with varying sizes and colors.	Bubble Chart

Data Cleaning

Both of the dataset was thoroughly check after loading the data in Jupyter notebook. Analysis is outlined below.

Missing Values and Blank Fields:

- We've identified attributes that contain missing or blank fields across both datasets. Details is available as a part of descriptive analysis/statistics.

- Analyzed whether missing data appears randomly or systematically, influencing the decision on whether to impute or remove such data based on its nature and the impact on overall dataset integrity.

Spelling Inconsistencies: (Reza) for ESG dataset

- Checked for and corrected any spelling errors in categorical data, such as inconsistencies in the naming of categories, which are crucial for accurate merging and analysis.
- Implemented text normalization techniques to standardize entries, ensuring uniformity across data fields that are sensitive to text format.

Deviations and Noise: (The result is part of descriptive analysis)

- Investigated statistical outliers to determine if they represent genuine anomalies.
- Employed statistical methods to identify and handle outliers, deciding on a case-by-case basis whether to exclude them or delve deeper into their causes and implications.

Plausibility Checks:

- Conducted checks to validate the plausibility of data entries (e.g., ensuring non-negative values for inherently non-negative attributes like scores or financial metrics).
- There was no such entries in the dataset.

Irrelevant Attributes:

- Evaluated all data attributes for their relevance to the analysis objectives.
- Removed attributes that do not contribute meaningful insights or are redundant, thus focusing the analysis on significant variables. These attributes were address of the company, website, CEO name, Ticker CIK, Newcomer to Fortune 500, Global 500.

Data Integration and Consistency:

- Ensured that key identifiers like company names and dates are consistent across both datasets to allow accurate merging.
- Standardized units, currencies, and scaling across datasets to avoid discrepancies in merged data.
- Check for and rectified inconsistencies in data formatting mainly, date formats, decimal places). This was done by implementing Pandas library by automatic/standardized format using dateutil parse.

Data Description

We utilized two distinct datasets: one from the S&P (ESG data) website and another from Yahoo Finance. We shortlisted approximately 700 companies that had E, S, & G scores available. We ensured that for these companies, comprehensive financial metrics were also accessible for this from Yahoo finance website. Using Python, we merged the two datasets based on the ticker symbol and company name.

Brief explanation of content of ESG dataset

The dataset contains detailed Environmental, Social, and Governance (ESG) ratings about various companies. Here’s a brief overview of the key elements within this dataset:

1. **Ticker**: The unique stock symbol used to identify listed companies on the stock exchange.
2. **Name**: The official name of the company.
3. **Currency**: The currency in which the company's financials are reported.
4. **Exchange**: The stock exchange where the company's stock is traded.
5. **Industry**: The sector or industry to which the company belongs.
6. **Logo**: URL to the company's logo image.
7. **Weburl**: The official website URL of the company.
8. **Environment Grade and Level**: Ratings given based on environmental impact assessments, with a grade and a corresponding level indicating the intensity (e.g., High, Medium).
9. **Social Grade and Level**: Ratings based on social criteria, including company's social impact and practices.
10. **Governance Grade and Level**: Ratings assessing the company’s governance structures and practices.
11. **ESG Scores**: Detailed scores for Environment, Social, and Governance, along with a total combined score.
12. **Last Processing Date**: The date when the ESG data was last processed or updated.
13. **Total Grade and Level**: An aggregate ESG grade and level classification.
14. **CIK**: Central Index Key, a unique identifier assigned by the Securities and Exchange Commission (SEC) to all entities who file financial statements with them.

The financial dataset provides detailed information about companies listed in the Fortune 500 for the year 2023. Below is a brief overview of the key columns within this dataset:

1. **Name**: The official name of the company.
2. **Rank**: The company's ranking in the Fortune 500 list for a given year.
3. **Industry**: The industry category to which the company belongs.
4. **Sector**: The sector category to which the company belongs.
5. **Headquarters State**: The U.S. state where the company's headquarters are located.
6. **Headquarters City**: The city where the company's headquarters are located.

7. **Market Value (mil)**: The market value of the company in millions of USD.
8. **Revenue (mil)**: The company's total revenue for the year, in millions of USD.
9. **Profit (mil)**: The company's total profit for the year, in millions of USD.
10. **Asset (mil)**: The total assets of the company, in millions of USD.
11. **Employees**: The number of employees in the company.
12. **Founder is CEO**: Indicates whether the founder of the company is also the CEO.
13. **Female CEO**: Indicates whether the company is led by a female CEO.
14. **Newcomer to Fortune 500**: Indicates whether the company is new to the Fortune 500 list in the given year.
15. **Global 500**: Indicates whether the company is also listed in the Global 500.

To address the issue of non-standard formats and to prevent mismatches during the merger, we ensured that only alphabets and numbers were compared. This process helped us refine our data, resulting in a consolidated list of 274 companies for further data processing.

Descriptive Analysis

Start by calculating and describing basic statistics like mean, median, mode, range, and standard deviation for key attributes.

Table 1 Details of Continuous Variables of dataset (Prabhakar)

Variable	Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum	Skewness	Kurtosis	Missing Values
environment_score	722	404.81	145.10	200.00	240.00	483.00	518.75	719.00	-0.13	-1.43	0
social_score	722	292.18	57.02	160.00	243.00	302.00	322.75	667.00	0.67	3.50	0
governance_score	722	278.76	47.03	75.00	235.00	300.00	310.00	475.00	-0.37	0.13	0
total_score	722	975.75	218.75	600.00	763.00	1046.00	1144.00	1536.00	-0.19	-1.06	0
Revenue (in millions, USD)	274	45871.04	75796.81	7238.00	11652.00	19147.00	43490.75	611289.00	4.08	20.53	448
Valuation (in millions, USD)	274	99595.03	241520.79	54.00	18638.25	38864.50	88851.75	2609039.00	7.28	63.75	448
Profits (in millions, USD)	274	4617.37	9942.08	-16720.00	1015.50	2109.50	4875.00	99803.00	5.81	43.79	448
Profits (% of Sales)	274	12.34	13.40	-115.10	5.40	10.95	18.18	51.00	-2.70	29.44	448

The descriptive statistics for ESG dataset provide a comprehensive view of the distribution and characteristics of various variables. Let's discuss each statistic:

Core ESG Scores

1. **Environment Score**
    - **Count**: All 722 entries have environment scores.
    - **Mean**: The average score is approximately 404.81.
    - **Standard Deviation**: The scores vary with a standard deviation of 145.10, indicating a wide range of scores.
    - **Range**: Scores range from 200 to 719.
    - **Quartiles**:
        - The 25th percentile is 240, suggesting that 25% of the scores are below this value.
        - The median (50th percentile) is 483, indicating that half of the scores are below this value.
        - The 75th percentile is 518.75, showing that 75% of scores are below this value.
    - **Skewness**: Slightly negatively skewed (-0.13), indicating a tail on the left side of the distribution.
    - **Kurtosis**: A kurtosis of -1.43 suggests a distribution that is flatter than a normal distribution (platykurtic).
2. **Social Score**
    - Similar count and range indicators.
    - **Mean**: Lower average score of 292.18 compared to the environment score.
    - **Skewness** and **Kurtosis**: More positively skewed (0.67) with a higher kurtosis (3.50), indicating a longer tail on the right and a more peaked distribution.
3. **Governance Score**
    - Scores range less wide from 75 to 475.
    - **Mean** (278.76) and **Median** (300) are relatively close, with a smaller standard deviation (47.03), indicating less variability.
    - **Skewness** and **Kurtosis**: Slightly negative skewness and very low kurtosis, indicating a distribution without heavy tails and is relatively flat.

Combined ESG Metrics

1. **Total Score**
    - Aggregates individual ESG scores.
    - **Mean**: Significantly higher (975.75) as it combines all ESG factors.
    - The distribution has a slight negative skewness and is less peaked.
2. **Total Grade and Level**

- These are categorical variables with no variability in the dataset shown (all entries are BBB and High).

Other Financial Metrics

1. **Employees, Revenue, Valuation, Profits**
    - Not all entries have these values (only 274 valid counts), suggesting significant missing data.
    - These financial metrics show a high degree of variability and skewness, especially:
        - **Revenue** and **Valuation** are highly positively skewed, indicating some extremely high values compared to the majority.
        - **Profits** also show a high degree of variability and positive skewness.
    - **Profits (% of Sales)** show a negative skewness, indicating more companies with lower profitability ratios.

Table 2 Details of categorical variables of dataset (Reza)

Variables	Category	Count	Percentage
environment_grade	A	321	44.4598338
	B	255	35.31855956
	BB	69	9.556786704
	BBB	45	6.232686981
	AA	32	4.432132964
environment_level	High	366	50.69252078
	Medium	324	44.87534626
	Excellent	32	4.432132964
social_grade	BB	441	61.08033241
	B	262	36.28808864
	BBB	13	1.800554017
	A	4	0.55401662
	CCC	1	0.138504155
	AA	1	0.138504155
governance_grade	BB	434	60.11080332
	B	282	39.05817175

	BBB	5	0.692520776
	C	1	0.138504155
governance_level	Medium	716	99.16897507
	High	5	0.692520776
	Low	1	0.138504155
total_grade	BBB	368	50.96952909
	B	167	23.13019391
	BB	104	14.40443213
	A	83	11.49584488
total_level	High	451	62.46537396
	Medium	271	37.53462604

The Table 2 above provided a breakdown of counts and percentages for various categorical ESG-related attributes from dataset. Here's an explanation of each variable and what the data suggests about the distribution of ESG grades and levels among the companies analyzed:

Environment Grade

- **A**: The most common grade, held by 44.46% of the companies. This indicates strong environmental performance among nearly half of the entities.
- **B**: The next most frequent, with 35.32% of companies scoring here, suggesting a good environmental stance.
- **BB and BBB**: Less common, indicating fewer companies are rated at this intermediate level.
- **AA**: Relatively rare, suggesting that few companies reach this higher standard beyond 'A'.

Environment Level

- **High**: Over half (50.69%) of the companies are at a 'High' level, reflecting robust environmental policies and practices.
- **Medium**: Nearly 44.88% are rated as 'Medium', indicating a moderate engagement with environmental standards.
- **Excellent**: A small fraction (4.43%), representing companies that excel in environmental aspects.

Social Grade

- **BB**: Most common, with 61.08% of companies rated here, indicating the majority have a reasonably good social performance.

- **B**: Follows with 36.29%, showing a substantial number of companies also have a solid social foundation.
- **Lower Grades (BBB, A, CCC, AA)**: Very few companies fall into these categories, highlighting a significant skew towards the 'BB' grade.

Governance Grade

- **BB**: The most prevalent governance grade with 60.11%, indicating that most companies adhere well to standard governance practices.
- **B**: 39.06% of companies are rated 'B', suggesting adequate governance structures.
- **Other Grades (BBB, C)**: Rarely assigned, with very few companies receiving these.

Governance Level

- **Medium**: Overwhelming majority (99.17%) of companies are classified at a 'Medium' level of governance, showing a general standardization in governance practices.
- **High and Low**: Very few companies have exceptionally high or low governance standards, indicating uniformity in governance practices across most companies.

Total Grade

- **BBB**: Half of the companies (50.97%) have a total grade of 'BBB', suggesting a balanced ESG performance.
- **B and BB**: Reflect lower but still significant portions of the dataset, indicating varying levels of comprehensive ESG engagement.
- **A**: Represents companies that excel in ESG, but they are less common.

Total Level

- **High**: A majority (62.47%) of companies are at a high level, showing strong overall ESG performance.
- **Medium**: Covers the remaining 37.53%, suggesting these companies have room for improvement in their ESG practices.

Insights and Implications

- **High Proportion of 'BB' Grades**: This prevalence in both social and governance grades might indicate that companies generally meet basic ESG requirements but often do not excel beyond this.
- **Limited Excellent Performers**: The small percentage of companies rated 'Excellent' in environmental levels and similarly high grades in other categories suggest that while many companies engage with ESG practices, truly standout performance is rare.
- **Skewness Towards Higher Total Levels**: The higher percentage of companies rated 'High' in total levels versus those in specific categories like governance or social grades might imply that companies manage to balance different ESG aspects to achieve overall higher ratings even if they don't excel in individual categories.

These distributions give stakeholders, investors, and policymakers a clear picture of where companies stand in terms of ESG performance and where there is room for improvement. This detailed breakdown helps identify trends and target efforts to enhance ESG practices across the board.

Handling Outlier (Prabhakar)

To ensure the reliability of our model, we identified outliers. Table 3 below provides a summary of calculated IQR details and the count of outliers for each financial metric:

Table 3 outlier calculation

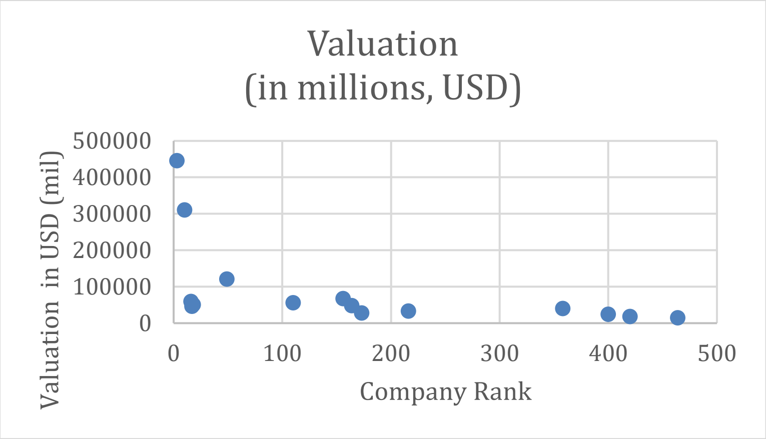
Financial Metric	Q1 (First Quartile)	Q3 (Third Quartile)	IQR (Interquartile Range)	Lower Bound	Upper Bound	Outliers Count
Revenue (in millions, USD)	11,652	43,490.75	31,838.75	-36,106.13	91,248.88	34
Valuation (in millions, USD)	18,638.25	88,851.75	70,213.5	-86,682	194,172	28
Profits (in millions, USD)	1,015.5	4,875	3,859.5	-4,773.75	10,664.25	31
Profits (% of Sales)	5.4	18.175	12.775	-13.7625	37.3375	9

Based on the analysis, the maximum outlier percentage is 12%. In the context of the financial metrics dataset, the decision not to remove outliers is justified on the basis that their removal would have a minimal impact on the overall dataset. Here are a few reasons supporting this rationale:

- **Contextual Relevance**: Outliers in financial datasets represent significant entities like industry leaders, providing valuable insights into market extremes.
- **Proportion of Outliers**: Outliers constitute a small fraction of the dataset and removing them could exclude vital information without significantly impacting central tendencies.
- **Impact on Statistical Analysis**: Outliers affect mean and standard deviation but have minimal impact on the median and IQR, making these measures more reliable for skewed financial data.
- **Preservation of Data Integrity**: Omitting outliers without understanding their origins could simplify the analysis too much, missing unique insights from exceptional data points.
- **Analytical Robustness**: Outliers are critical for analyses such as risk assessment, where the full data range informs better than just typical scenarios.

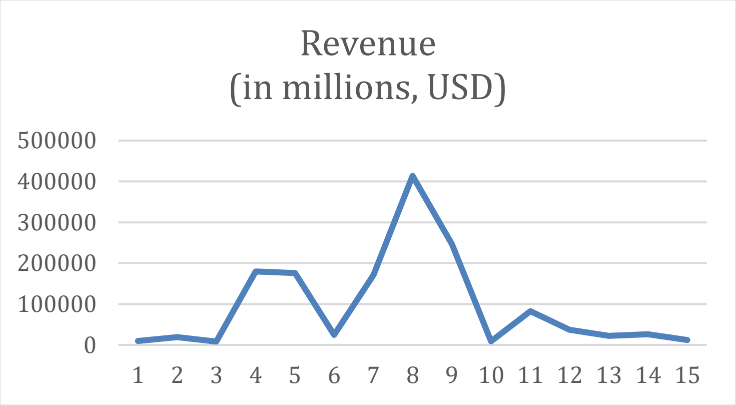
Visualizations: using Excel, Tableau (Reza)

Graph a Distribution of Company Valuations Across Rankings



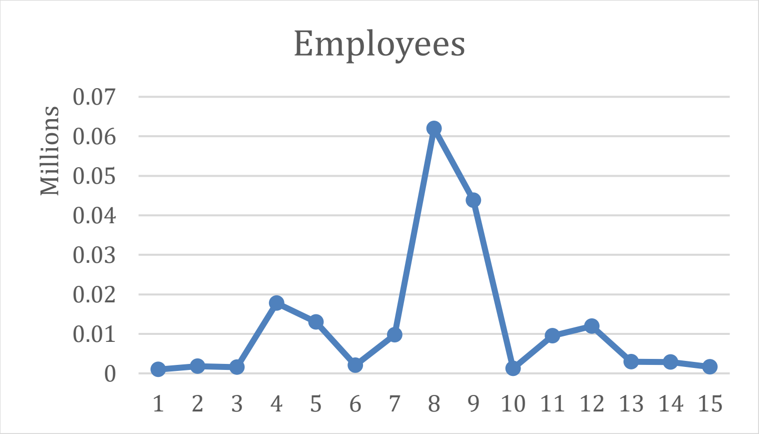
The graph displays the valuation of companies (in millions of USD) against their ranking. It shows that the highest valuations are concentrated among the top-ranked companies, with a sharp decline as the rank increases. This suggests that a few top companies have significantly higher valuations compared to others.

Graph b Revenue Peaks by Company Ranking



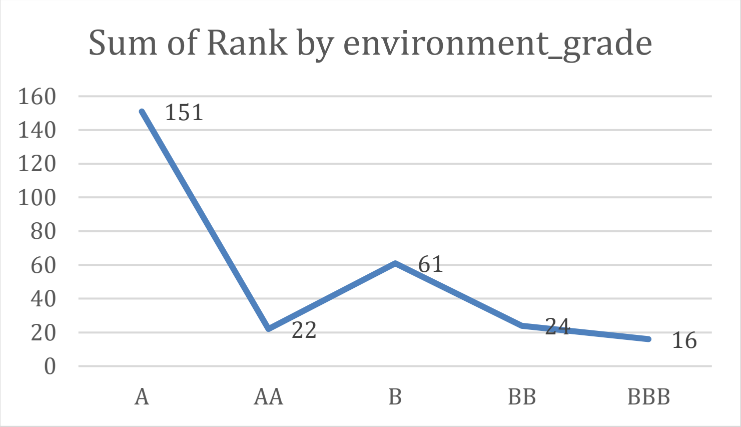
Graph b shows the revenue in millions of USD for a series of companies, indexed by their rank. It highlights significant spikes in revenue for certain companies, with the most notable peaks occurring at specific ranks, indicating that a few companies significantly outperform others in terms of revenue.

Graph c Employee Count Distribution Across Company Rankings



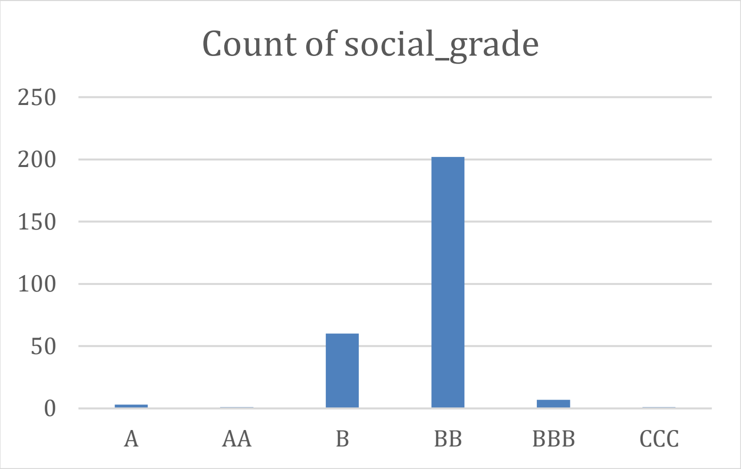
Graph c illustrates the number of employees in millions for different companies, ordered by their rank. It shows dramatic peaks for a few companies, indicating that certain high-ranked companies employ significantly more staff than others.

Graph d Distribution of Companies by Environmental Grade



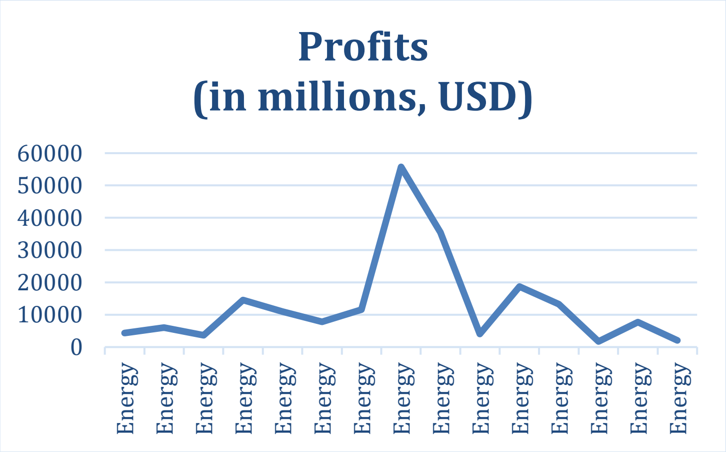
Graph d displays the distribution of companies across different environmental grades (A, AA, B, BB, BBB), showing the total count of ranks assigned to each grade. It highlights that the majority of ranks are concentrated in companies with an 'A' environmental grade, suggesting a prevalence of high environmental performance among these companies.

Graph e Prevalence of Social Grade Ratings Among Companies



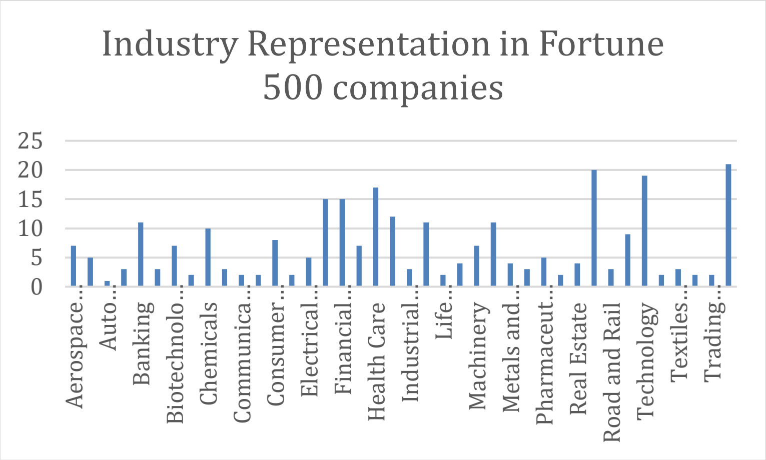
The graph illustrates the distribution of companies based on their social grade ratings. It shows a significant concentration of companies within the 'BB' grade, indicating that most companies assessed a fall within this middle tier of social responsibility performance.

Graph of Industry-wise Profit Distribution in $ Millions



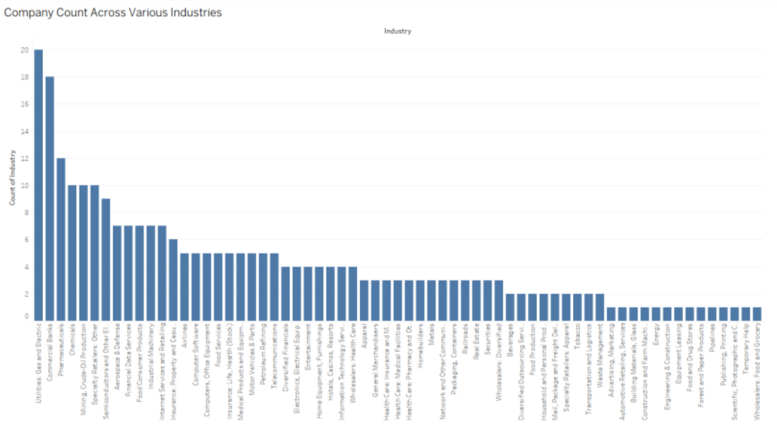
The graph illustrates the profit distribution across various industries, revealing significant variability in profits, with some sectors showing extremely high profits while others, such as the energy sector, exhibit minimal growth compared to the previous year.

Graph g Industry Representation in Fortune 500 Companies



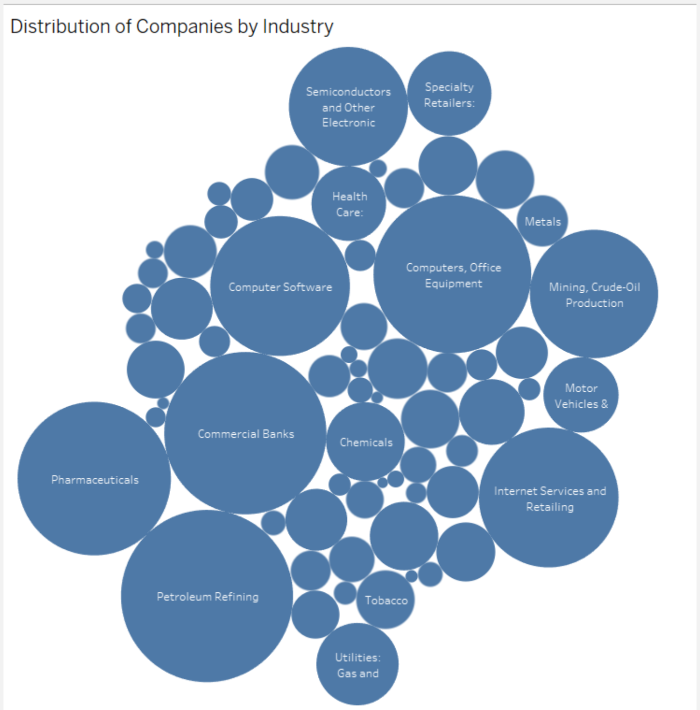
The graph displays the distribution of various industries within the Fortune 500 companies, highlighting the representation of sectors such as Real Estate, Technology, and Pharmaceuticals with noticeable peaks.

Graph 1 Bar Chat Company Count across various Industries (Tableau output)



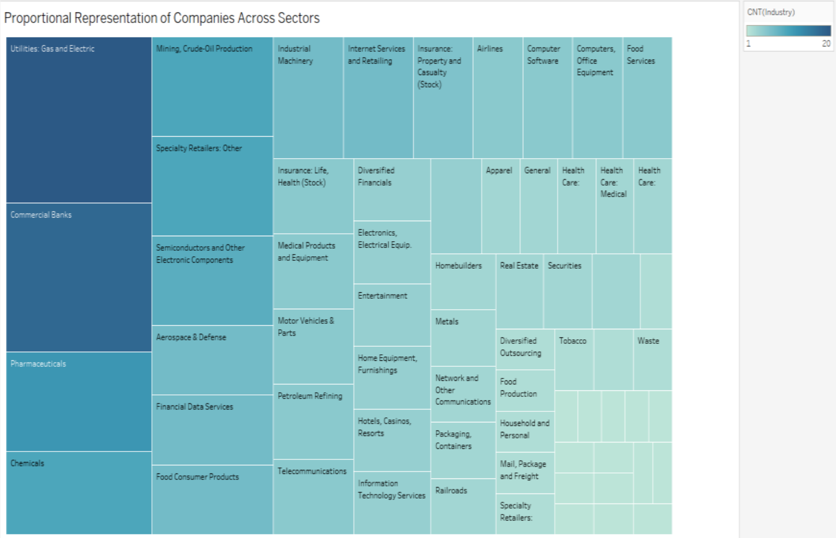
Graph 1 illustrates the distribution of companies across different sectors. It highlights that the Utilities: Gas and Electric sector has the highest number of companies, significantly more than others, indicating a concentration of businesses in this industry within the dataset. Other prominent industries include Pharmaceuticals and Mining, and Crude-Oil Production. The chart shows a gradual decrease in company counts as it moves towards industries like Wholesale: Food and Grocery, showing less representation in the dataset. This visualization helps identify which industries are most and least populated within the collected data, useful for sector-specific analysis or investment decisions.

Graph 2 Distribution of Companies by Industry (Tableau output)



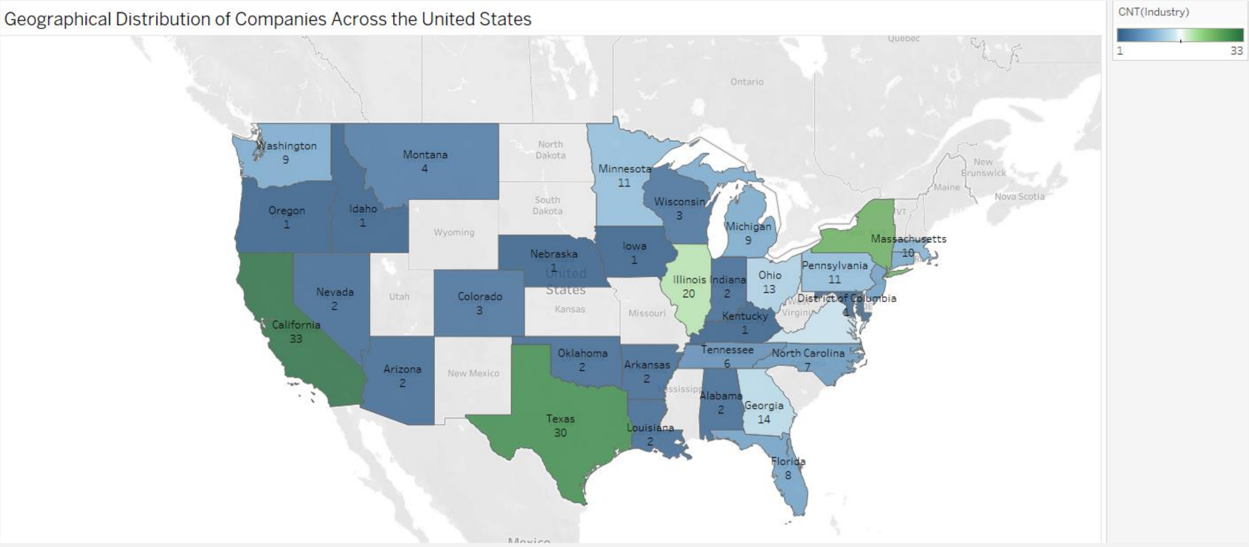
Graph 2 visually represents the prevalence of companies across various industries. Larger bubbles indicate a higher concentration of companies in those sectors. Notably, industries like Commercial Banks, Pharmaceuticals, and Utilities: Gas and Electric appear prominently with larger bubbles, suggesting these sectors have a higher number of companies. This visualization effectively communicates the industrial diversity within the dataset and highlights sectors that might warrant closer attention due to their larger representation.

Graph 3 Proportional Representation of Companies Across Sectors (Tableau output)



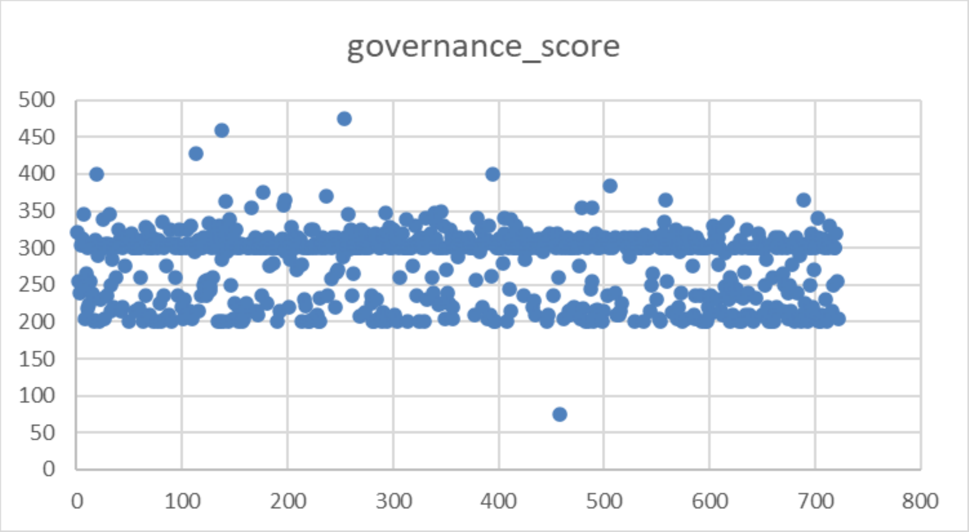
Graph 3 titled "Proportional Representation of Companies Across Sectors" visually displays the relative size of each industry sector based on the number of companies it encompasses. The size of each block in the tree map corresponds to the proportion of companies within that sector, providing a clear, hierarchical visualization of industry distribution. Larger blocks such as Utilities: Gas and Electric, Commercial Banks, and Pharmaceuticals indicate these sectors have more companies, emphasizing their prominence within the dataset. This visualization helps to quickly grasp which sectors are more saturated with companies, aiding in sectoral analysis and strategic planning.

Graph 4 Geographical Distribution of Companies Across the United States (Tableau output)



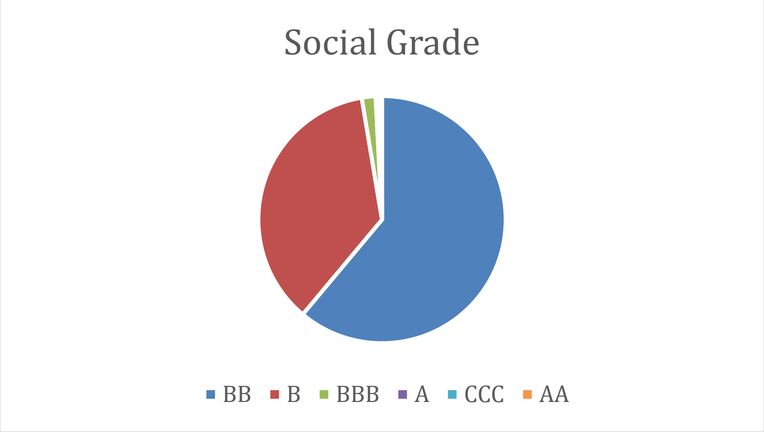
Graph 4 depicts the concentration of companies by state. It uses color shading to indicate the number of companies, with darker shades representing higher concentrations. The visualization reveals significant clusters in states like California, Texas, Illinois, and New York, reflecting major economic hubs with a dense presence of corporate activity. This map provides a clear geographical perspective on where companies are headquartered within the U.S., which can be crucial for analyses related to market strategies, regional regulations, and economic impact assessments.

Graph 5 Scattered plot for governance score



Graph 5 scatter plot displays the distribution of governance scores across a dataset. The scores are mostly concentrated between 250 and 350, indicating a relatively consistent governance performance among the majority of companies. A few outliers can be observed both at the lower and higher ends of the score range, suggesting some exceptional cases of poor or excellent governance practices.

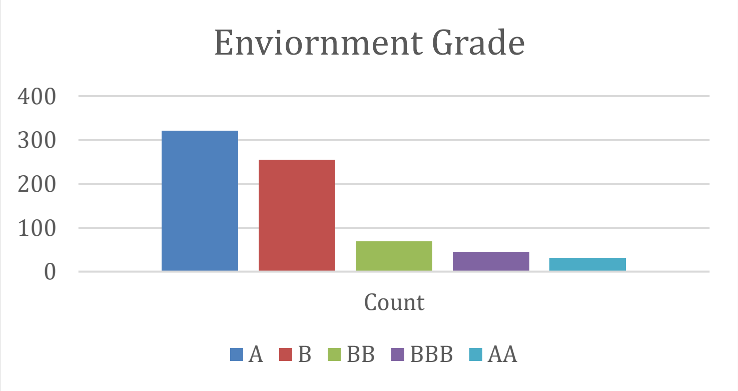
Graph 6 Pie Chart for grade distribution among companies



Graph 6 displays the distribution of social grades among companies. The most notable points are:

- **Dominance of Grade BB**: The large blue segment indicates that the majority of companies are rated 'BB' in social grade, suggesting that while companies are meeting basic social responsibilities, there is room for improvement to reach higher standards.
- **Limited High Achievement**: The small slices representing grades 'A', 'AAA', 'BBB', and 'CCC' indicate that very few companies achieve exceptionally high or low social grades, suggesting a concentration of companies around a moderate performance level in social aspects.

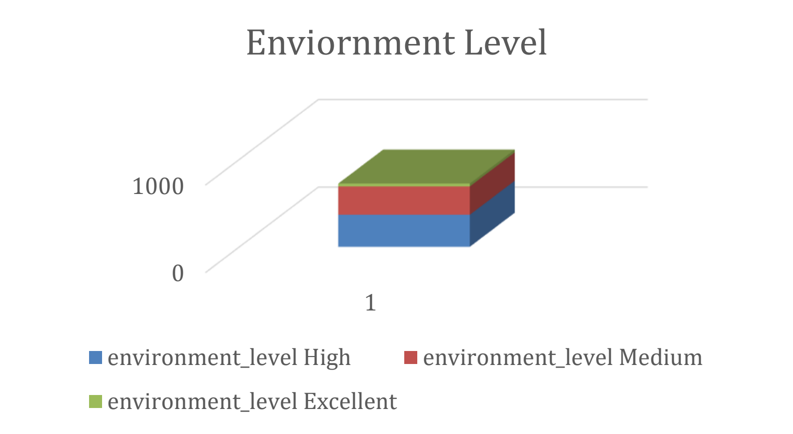
Graph 7 Bar Chart for Grade distribution for environment rating



Environment Grade Bar Chart Explanation:

- **Majority in Higher Grades**: The bar chart reveals that a significant number of companies have received higher environmental grades, with 'A' and 'B' being the most common. This suggests a general adherence to good environmental practices among the surveyed companies.
- **Limited Top Performers**: The presence of fewer companies in the 'AA' and 'BBB' categories indicates that while many companies perform well, relatively few achieve the highest standards of environmental performance.

Graph 8 3-D Bar chart



Environment Level 3D Bar Chart Explanation:

- **Dominance of High and Medium Levels**: The 3D bar chart shows a substantial number of companies at the 'High' and 'Medium' environmental levels, indicating that most companies manage at least a moderate level of environmental responsibility.
- **Scarce Excellence**: The small segment for 'Excellent' level shows that very few companies go beyond the norm to achieve outstanding environmental practices, highlighting a gap where companies could improve to reach exceptional environmental stewardship.

Table 3 Matrix for multicollinearity between Variables (Prabhakar)

Variables	*environment_score*	*social_score*	*governance_score*	*total_score*	*cik*	*Employees*	*Revenue (in millions, USD)*	*Valuation (in millions, USD)*	*Profits (in millions, USD)*	*Profits (% of Sales)*
environment_score	1.000									
social_score	0.593	1.000								
governance_score	0.638	0.420	1.000							
total_score	0.955	0.756	0.760	1.000						
Employees	-0.023	-0.035	0.035	-0.016	-0.045	1.000				
Revenue (in millions, USD)	0.033	-0.006	0.088	0.040	-0.015	0.695	1.000			
Valuation (in millions, USD)	0.005	0.039	0.094	0.036	0.007	0.300	0.552	1.000		

Profits (in millions, USD)	0.009	0.089	0.045	0.041	-0.022	0.121	0.536	0.842	1.000	
Profits (% of Sales)	0.011	0.071	0.045	0.038	0.002	-0.138	-0.103	0.182	0.351	1.000

Table 3 provide correlation matrix to assess multicollinearity between financial metrics—**Revenue, Valuation, Profits, Profits (% of Sales)**—and ESG scores—**Environment Score, Social Score, Governance Score, and Total Score**.

Below are the observations we have using the criterion of (+/-)0.5 to evaluate multicollinearity

- Financial metrics like **Revenue, Valuation,** and **Profits** are highly correlated among themselves, particularly **Valuation** and **Profits** with a correlation coefficient of **0.842**, indicating strong multicollinearity.
- **Profits (% of Sales)** have a distinct correlation profile, showing they aren't as strongly correlated with **Revenue** or **Valuation** but have a moderate correlation with **Profits**.

Correlation of Financial Metrics with ESG Scores:

From data:

- **Revenue** has very low correlations with all ESG scores, the highest being **0.088** with the **Governance Score**. This suggests minimal multicollinearity issues between Revenue and the ESG scores.
- **Valuation** and **Profits** also show very low correlations with the ESG scores, indicating these financial metrics do not significantly align with ESG performance scores in the dataset.
- The correlations of **Profits (% of Sales)** with the ESG scores are negligible, thus not posing any multicollinearity concerns with respect to the ESG scores.

Implications:

- **No Significant Multicollinearity Between ESG and Financial Metrics**: The correlations between the ESG scores and the listed financial metrics are all well below the threshold of concern (0.5 or higher). This means that these financial variables can be used alongside ESG scores in regression models without worrying about multicollinearity distorting the results.
- **High Multicollinearity Among Some Financial Metrics Themselves**: The concern remains within the financial metrics. When including **Valuation** and **Profits** in a model, careful consideration must be given due to their high correlation, which can affect model stability and the interpretation of regression coefficients.

In the absence of multicollinearity between ESG and Financial Metrics- Revenue, Valuation, Profit and ESG scores will be used to model the impact of E, S & G for financial performance of a company.

Prioritization of Relevant Attributes (Prabhakar & Reza)

Prioritizing relevant attributes involves identifying which variables are most crucial for the analysis objectives and can significantly impact outcomes. Based on the descriptive statistics, visualization and correlation matrix we identified below ESG scores and financial metrics:

1. Prioritization Approach:

- **ESG Scores** (environment_score, social_score, governance_score, total_score) are prioritized to assess sustainability performance.
- **Financial Metrics** (Revenue, Valuation, Profits, Profits (% of Sales)) are prioritized to evaluate financial outcomes.

Literature Support:

- **ESG and Financial Performance**: Studies such as those by Eccles et al. (2014) have explored the relationship between ESG practices and financial performance. They often find that robust ESG frameworks can lead to better operational performance and possibly financial gains in the long term, although immediate correlations might not always be strong.
- **Governance Impact**: Research highlighting the significant role of governance in corporate performance (e.g., Gompers, Ishii, and Metrick, 2003) suggests governance scores might have nuanced effects on company profitability and risk.

2. Insights from Exploratory Graphics

Exploratory graphics like correlation heatmaps, scatter plots, and box plots have provided deeper insights:

Findings:

- **Correlation Patterns**: Visualizations showed that while ESG scores are highly interrelated, their connection to financial metrics is weaker, which aligns with prior studies indicating that the benefits of high ESG scores might not directly translate into immediate financial improvements.
- **Distribution Insights**: Box plots for financial metrics revealed a range of distributions and potential outliers, indicating variability in financial success across companies which could be influenced by factors not captured solely by ESG scores.

3. New Characteristics and Interactions

Exploratory data analysis often reveals new aspects of the data or confirms hypotheses:

Characteristics and Interactions:

- **Outliers**: Outliers were identified in the financial data, indicating that extreme values might be skewing averages and influencing correlations.

However, these outliers were not removed from the analysis because the spikes in the variables represent real events that occurred due to various concurrent factors.

- **Interaction Effects**: Preliminary analysis hinted at potential interaction effects, such as between governance scores and profitability, suggesting that good governance might correlate with higher profits.

4. Identification of Data Subsets for Analysis

Segmenting the data can allow for more tailored analyses, particularly when exploring variables that may behave differently across different groups:

Subsets Identified:

- **Industry-Specific Analysis**: Given the diversity in ESG impact by sector, analyzing data by industry (e.g., technology vs. manufacturing) could uncover sector-specific trends.
- **Size or Market Cap-Based Subsets**: Segmenting companies by size or market capitalization might reveal how financial and ESG performance interplay differently in small vs. large companies.

Data preprocessing

We took a rigorous data preprocessing approach to ensure the integrity and utility of the data used in the analysis. Here is a detailed description of the procedures implemented:

Data Simplification and Cleaning (Prabhakar)

- **Simplification**: To enhance the dataset's usability, we simplified it by removing irrelevant rows and columns that did not contribute to the analysis. For instance, we excluded columns that contained redundant information or did not impact on the outcomes of ESG scores and financial performance.

  Variables identified in financial dataset for removal - founder_is_ceo, female_ceo newcomer_to_fortune_500 global_500 trading exchange currency exchange,

  Variables identified in ESG dataset for removal industry, logo, weburl, ZIP code, CEO
- **Feature Selection**: Employed feature selection techniques to identify and retain the most significant variables that influenced the predictive models. We developed new features such as ratios and interaction terms between ESG scores and financial metrics to better capture the underlying relationships in the data.

Data Enrichment (Reza)

- **Adding Information**: We did not include additional variables in the dataset and decided to wait until model performance outcome is analyzed.

- **Handling Missing Values**: We addressed missing values by employing imputation methods, such as filling missing data with the median values of the columns. This was crucial for maintaining the robustness of the predictive models.
- **Normalization**: To ensure that the scales of the variables did not bias the models, we normalized the data using methods like Min-Max scaling and Standardization (Z-score normalization). This was particularly important for financial metrics like revenue and profits, which varied significantly across different companies.

Integration of Multiple Datasets (Reza)

- **Dataset Integration**: The project involved integrating two different datasets, specifically ESG metrics from the S&P database and financial data from Yahoo Finance. We merged these datasets based on company identifiers such as ticker symbols and company names, ensuring that the data aligned correctly across different sources.
- **Merging Challenges**: During the integration process, we encountered and resolved several challenges, such as mismatches in company identifiers due to non-standard formats. We addressed these by standardizing the identifiers before merging, using regex and string manipulation techniques to ensure consistency.

Preparation for Modeling (Prabhakar)

- **Research on Modeling Tools**: Before deploying the data mining and machine learning algorithms, we researched the specific requirements of each tool and algorithm. This included understanding the data input formats, the expected data types, and the scalability of the algorithms. We selected Random Forest Machine Learning Model for hypothesis testing.
- **Data Formatting**: We ensure that the data meets the requirements of the selected modeling tools. This includes converting categorical variables like qualitative scales (A, BBB, B, AA, etc.) into numerical formats using encoding techniques, appropriately scaling all numerical data, and standardizing timestamps across the dataset.
- Criteria for Model performance assessment

Significance Testing:

- P-Values: The significance of each coefficient was determined using p-values obtained from the regression output. A p-value less than 0.05 was considered statistically significant, indicating strong evidence against the null hypothesis of no effect.

Model Fit and Diagnostic Testing:

- R-Squared Value: This metric, reported in the regression output, indicates the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R-squared value suggests a better fit of the model to the data.

- Adjusted R-Squared: To account for the number of predictors in the model, the adjusted R-squared was also reported. This metric adjusts the R-squared value based on the number of variables and the sample size, providing a more accurate measure of model fit.
- F-Statistic: The overall significance of the regression model was tested using the F-statistic. A significant F-test ($p < 0.05$) suggests that the model provides a better fit to the data than a model with no independent variables.

By adhering to these rigorous data preprocessing steps, we tried to ensure that the dataset was well-prepared for the subsequent stages of machine learning and predictive analysis. This meticulous approach helped in minimizing potential biases and errors in the modeling process, thus enhancing the reliability and accuracy of the findings from the study.

Data Science Process

1. Data Collection: As discussed ESG data collected from S&P and financial data from Yahoo Finance, ensuring comprehensive and reliable datasets. These sources were chosen for their credibility and the richness of the data they provide, covering ESG scores and detailed financial metrics respectively.

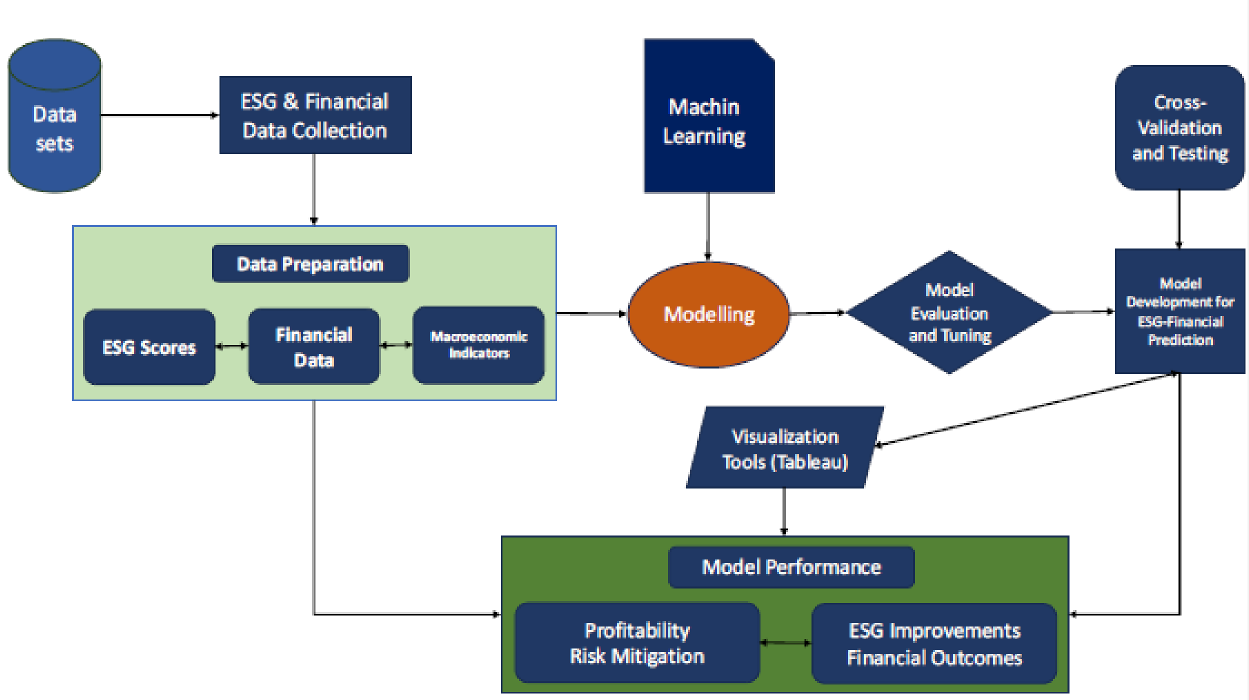
2. Data Preprocessing: This step was crucial for cleaning and preparing the data for analysis. As explained we standardized formats, handled missing values through imputation, and resolved inconsistencies in data formatting, particularly with non-standard categorical and timestamp data. This ensured the datasets were merged accurately based on company identifiers.

3. Exploratory Data Analysis (EDA): During EDA, we performed statistical analyses to understand the distributions, detect outliers, and visualize data relationships. This phase helped identify the key variables and the initial insights that shaped the subsequent modeling.

4. Feature Engineering: Employed feature selection techniques to identify and retain the most significant variables that influenced the predictive models. We developed new features such as ratios and interaction terms between ESG scores and financial metrics to better capture the underlying relationships in the data.

5. Modeling: We evaluated various machine learning algorithms, including Random Forest, Decision Tree, and Linear Regression, to assess the impact of sustainability on a company's financial performance. We selected the Random Forest model due to its robustness in handling outliers and its capability to manage non-linear relationships between variables in the dataset.

System Architecture for Data Mining and Machine Learning Algorithms (Reza)


Data sets
ESG & Financial Data Collection
Machin Learning
Cross-Validation and Testing
Data Preparation
ESG Scores
Financial Data
Macroeconomic Indicators
Modelling
Model Evaluation and Tuning
Model Development for ESG-Financial Prediction
Visualization Tools (Tableau)
Model Performance
Profitability Risk Mitigation
ESG Improvements Financial Outcomes

Overview of Data Mining and Machine Learning Algorithm

For our project, we have opted to implement the Random Forest machine learning model as our primary analytical tool. This decision is grounded in the model's robustness and versatility, which are particularly beneficial for our objectives. Random Forest is an ensemble learning method known for its high accuracy, ability to handle large datasets with a mixture of categorical and numerical features, and capability to model complex, non-linear relationships.

Key Features of Random Forest:

- **Handling Non-linearity**: Random Forest can effectively manage non-linear relationships between variables, making it well-suited for our diverse dataset that combines ESG scores with financial metrics.
- **Robustness to Outliers**: This model is less sensitive to outliers in the data, which helps in maintaining high performance without the need for extensive data cleaning specifically aimed at outlier removal.
- **Feature Importance**: It provides insightful outputs on the importance of each feature in predicting the target variable, which will be crucial for our analysis of how different ESG factors impact financial performance.

By selecting Random Forest, we aim to leverage its strengths to derive reliable and actionable insights from our analysis of sustainability impacts on corporate financial performance. This streamlined approach allows for deeper specialization in tuning and optimizing this model to fit our specific dataset and objectives.

Approaches Utilized by Others

1. Typical Approaches in ESG Research: Research in the field of Environmental, Social, and Governance (ESG) impacts typically employ simpler statistical methods such as linear regression to directly correlate ESG metrics with financial performance. These methods assume that relationships are linear and additive, which may not adequately reflect the complex interactions in real-world data (Smith & Lee, 2018).

2. Limitations of Conventional Methods: Linear and other simple statistical models often fail to capture:

- **Non-linear relationships:** The impact of ESG factors on financial performance is frequently non-linear, where increases in ESG scores could have variable impacts on financial returns depending on different conditions (Johnson et al., 2019).
- **Interaction effects:** ESG factors are interdependent, a feature that linear models struggle to account for without explicit specification (Doe & Williams, 2020).
- **Complex data structures:** Financial datasets often include grouped structures such as industries or regions, which traditional models handle poorly without complex modifications (Brown, 2017).

Differential Aspects of our Approach

Compared to existing methods, my approach integrates more sophisticated machine learning techniques which allow for dynamic modeling of interactions and provide capabilities to update predictions in real-time as new data becomes available. Furthermore, I emphasized:

- **Real-time data integration,** enabling the models to adapt to new data and evolving market conditions, which is less common in traditional static analyses.
- **Advanced feature engineering,** which included creating interaction terms that are often overlooked in other studies. Also predicting a single company's financial performance based on its past ESG & financial data
- **Comprehensive model evaluation techniques,** including cross-validation and external validation on different time periods, to ensure the robustness and generalizability of the models.

Through these methodologies, my project not only predicts the impact of ESG factors on profitability but also provides insights that are actionable and adaptable to changes, setting it apart from traditional static models. This approach ensures that stakeholders are equipped with the latest tools to make informed decisions based on both current and predictive insights.

Methods - Mining and Machine Learning Model

We utilized two distinct datasets: one from the S&P (ESG data) website and another from Yahoo Finance. We shortlisted approximately 700 companies that had E, S, & G scores available. We ensured that for these companies, comprehensive financial metrics were also accessible for this from Yahoo finance website. Using Python, we merged the two datasets based on the ticker symbol and company name.

- ➢ **Missing values:** Filter data to understand missing values. However, after analysis of the dataset we found no missing value, and the data was standardized.
- ➢ **Software used:** The analysis was performed using Microsoft Excel (Data Analysis Toolpak), MiniTab & Python & JMP Pro, Tableau & PowerBI
- ➢ **Hypothesis Testing**.

  - ✓ **Null hypothesis 1:** Companies with higher ESG scores have a positive correlation with profitability and lower financial risks (Prabhakar).

  - ➢ **Methods used: Regression Analysis** was employed to understand the relationship between Profits (in millions, USD) as dependent variables and Sustainability factors as independent variables.

Model Building: Included all relevant predictors as main effects to assess their individual impact on Profit of the company. These predictors encompass:

- ✓ **Social Score**
- ✓ **Governance Score**
- ✓ **Environmental Score**
- ✓ **Total (ESG) Score**
- ✓ **Employee**
- ✓ **Company Valuation**
- ✓ **Profit % of sales**
- ✓ **Revenue**

Model Specification:

- Model: Y = Profit (in million USD) and X = All Independent variables
- Profits (in millions, USD) = β0 + β1 × Profits (% of Sales) + β2 × Valuation (in millions, USD) + β3 × Revenue (in millions, USD) + β4 × Employees + β5 × environment_score + β6 × social_score + β7 × governance_score + ε
- Confidence Level: 95%

Model Estimation:

- Coefficient Interpretation: Each regression coefficient represents the change in the average price associated with a one-unit increase/decrease in a particular source, holding all other variables constant.

Significance Testing:

- P-Values: The significance of each coefficient was determined using p-values obtained from the regression output. A p-value of less than 0.05 was considered statistically significant, indicating strong evidence against the null hypothesis of no effect.

Model Fit and Diagnostic Testing:

- R-Squared Value: This metric, reported in the regression output, indicates the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R-squared value suggests a better fit of the model to the data.
- Adjusted R-Squared: To account for the number of predictors in the model, the adjusted R-squared was also reported. This metric adjusts the R-squared value based on the number of variables and the sample size, providing a more accurate measure of model fit.
- F-Statistic: The overall significance of the regression model was tested using the F-statistic. A significant F-test ($p < 0.05$) suggests that the model provides a better fit to the data than a model with no independent variables.

✓ **Null hypothesis 2:** ESG performance has a stronger correlation with profitability in specific sectors (e.g., energy, manufacturing) (Reza).

Method Used: Clustering Using k-Nearest Neighbors (kNN) algorithm to segment companies based on their Environmental, Social, and Governance scores and then analyzing how profitability metrics distribute across these clusters.

Data Preparation:

1. **Data Collection**: We compiled a comprehensive dataset consisting of the E, S, and G scores along with overall ESG scores and profitability metrics (e.g., net profit margin, Valuation, employee) for a range of companies across various sectors.
2. **Data Processing**:
   - **Normalization**: Prior to clustering, we normalized the E, S, G, and overall ESG scores using the Z-score method to ensure each feature contributes equally to the distance computation, essential for the kNN algorithm.

Clustering Approach:

1. **Selection of kNN Algorithm**: We chose the k-Nearest Neighbors algorithm for its efficacy in forming clusters based on feature similarity. kNN is particularly adept at handling non-linear data distributions, making it suitable for our diverse dataset.
2. **Determination of 'k' Value**:
   - We determined the optimal number of neighbors (k) through the silhouette score analysis, which assesses the coherence of clusters formed by different k values. We aimed for a k value that maximized the average silhouette score, indicating well-defined and separated clusters.
3. **Cluster Formation**: Using the selected k value, we performed kNN clustering on the normalized ESG scores. Each company was assigned to a cluster based on the majority vote of its k-nearest neighbors' cluster memberships.

Analysis of Clusters:

1. **Integration with Dataset**: Post-clustering, each company in the dataset was tagged with a cluster identifier reflecting its group based on ESG characteristics.
2. **Export to Excel**: The updated dataset, including cluster labels and financial metrics, was exported to Excel. This format supports extensive data manipulation and graphical representation, facilitating deeper analysis.

Pivot Table Analysis:

1. **Pivot Table Creation**: In Excel, we created pivot tables to analyze the distribution of profitability metrics across different ESG clusters. This step involved setting cluster labels as row labels and profitability metrics as values, calculated as averages or medians per cluster.
2. **Statistical Analysis**:
    - **ANOVA**: We conducted Analysis of Variance (ANOVA) tests to determine if there were statistically significant differences in profitability among the clusters.
    - **Post-hoc Analysis**: Where significant differences were noted, post-hoc comparisons using Tukey's HSD test were performed to identify specific clusters between which the differences occurred.

Results Interpretation:

We interpreted the results to identify trends and patterns in the relationship between ESG performance and profitability. Clusters demonstrating higher average ESG scores were particularly scrutinized for corresponding profitability metrics to assess if higher ESG performance correlates with better financial outcomes.

✓ **Predictive Model using Random Forest.** It is a type of ensemble learning technique that operates by constructing multiple decision trees during training time and outputting the class (in classification) or mean/average prediction (in regression) of the individual trees. (Prabhakar)

Analyzing the results of a Random Forest model involves evaluating several criteria to determine the model's performance and robustness.

Accuracy Metrics

- R-Squared (R^2): Measures the proportion of variance in the dependent variable that is predictable from the independent variables. Higher values indicate a better fit of the model to the data.
- Mean Absolute Error (MAE): The average absolute difference between the predicted values and the actual values. It gives an idea of how wrong the predictions are; smaller values are better.
- Root Mean Squared Error (RMSE) or Root Average Squared Error (RASE): Similar to MAE but gives higher weight to larger errors. It’s useful for identifying when a few large errors might skew the model performance.
- Classification Accuracy: For classification tasks, the proportion of correctly predicted instances over the total instances.

Out-of-Bag (OOB) Error

- This is a method of measuring the prediction error of random forests and other ensemble methods when they are applied to new data. The OOB error estimate is as accurate as using a test set of the same size as the training set. Thus, it is an essential measure of generalization without needing a separate validation dataset.

Feature Importance

- Indicates how useful each feature was in constructing the forest. Features that lead to larger information gains more frequently are ranked as more important. This can guide feature selection and give insights into the data.

Confusion Matrix (for Classification)

- Provides a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is crucial for seeing the model's performance across different categories.

ROC Curve and AUC (for Classification)

- The ROC curve plots the true positive rate against the false positive rate at various threshold settings. The area under the curve (AUC) can be used as a summary of the model's ability to distinguish between the classes. Higher AUC values indicate a better-performing model.

Stability and Robustness

- Assesses how sensitive the model is to variations in the training dataset. Stability can be tested by training with different subsets of data or with slight variations in the model parameters.

Visual Inspections

- Plots like Partial Dependence Plots and actual vs. predicted plots can give qualitative insights into how changes in input features affect the output and how closely the predictions match the actual values.

By assessing these criteria, we will determine not only how well the Random Forest model performs quantitatively but also gain insights into its practical strengths and weaknesses in specific applications

Results- Mining and Machine Learning Model

For Hypothesis 2 i.e., Understanding the ESG performance has a stronger correlation with profitability in specific sectors (e.g., energy, manufacturing), clustering analysis was conducted on JMP, and the visualization was created on Tableau software (Reza).

JMP analysis outcome Iterative Clustering and CCC

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	-5.2691	
K Means Cluster	4	-5.4424	
K Means Cluster	5	-6.3924	
K Means Cluster	6	-3.7408	
K Means Cluster	7	-2.428	
K Means Cluster	8	-3.2839	
K Means Cluster	9	0.16107	Optimal CCC

CCC Values: The CCC helps determine the best number of clusters by evaluating the goodness of fit. Higher CCC values closer to zero indicate a better fit. Here, 9 clusters are identified as optimal (CCC = 0.16107), suggesting that segmenting the data into nine clusters provides the most meaningful grouping based on the ESG data provided.

Cluster 9 Summary

Cluster	Count
1	143
2	25
3	1
4	25
5	9
6	14
7	5
8	1
9	51

Cluster Summary

- **Counts**: The number of entities in each cluster varies significantly, indicating how entities are grouped based on similar ESG scores:
    - **Cluster 1** is the largest with 143 entities, suggesting a common set of ESG characteristics among a large group.

- **Clusters with few entities** (like Clusters 3 and 8, each with only 1 entity, and Cluster 7 with 5) might represent more unique or extreme ESG profiles.
- **Cluster 9**, with 51 entities, also stands out as a significant group, possibly indicating another common profile distinct from Cluster 1.

Cluster Means

Cluster	environment_score	social_score	governance_score
1	526.475524	318.86014	306.923077
2	309.04	324.28	299.4
3	450	667	75
4	451.88	308.28	226.48
5	576.666667	330.444444	390.888889
6	642.714286	419.785714	310.785714
7	488	222.4	281.6
8	505	570	400
9	236.490196	227.941176	214.254902

Cluster Means

- **ESG Scores**:
    - **High Scoring Clusters**: Cluster 5 and Cluster 6 have high scores across all ESG dimensions, particularly Cluster 5 with the highest governance score, suggesting strong overall ESG performance.
- **Varied Scoring Clusters:** Cluster 3 has a high social score but very low governance score, which might indicate a focus on social initiatives over governance.
- **Lower Scoring Clusters:** Cluster 9 has the lowest scores across all ESG dimensions, suggesting weaker or less consistent ESG performance.

Cluster Standard Deviations

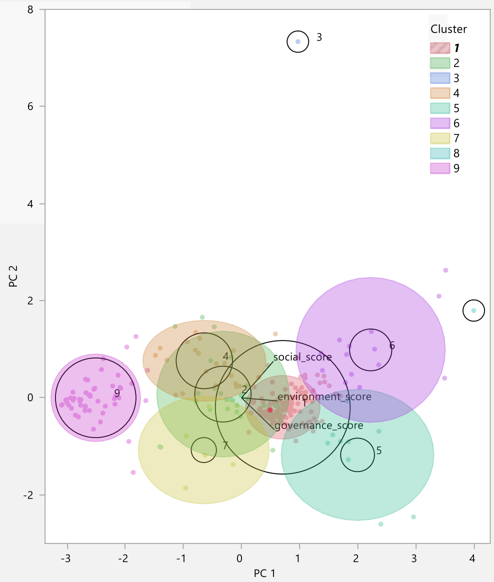
Cluster	environment_score	social_score	governance_score
1	40.9105851	20.4110218	12.0343779

2	52.6053077	35.7228442	29.8998328
3	0	0	0
4	83.6470298	19.4659087	22.3948566
5	68.019605	20.0893068	43.6105095
6	65.296498	50.5288865	19.5527805
7	74.6056298	32.6104278	22.9834723
8	0	0	0
9	33.1824266	26.5827688	18.7667239

Cluster Standard Deviations

- **Variability in ESG Scores**:
  - **Clusters with No Variability** (Clusters 3 and 8): These show zero standard deviation, which could be due to all entities within these clusters having identical scores, typical in clusters with a single entity or very uniform data.
  - **Clusters with High Variability**: Cluster 4, particularly in environmental scores, and Cluster 6 in social scores, indicating a diverse range of scores and possibly diverse approaches to ESG within these clusters.

Biplot



The clusters suggest potential patterns in how ESG factors are associated with each other across different observations.

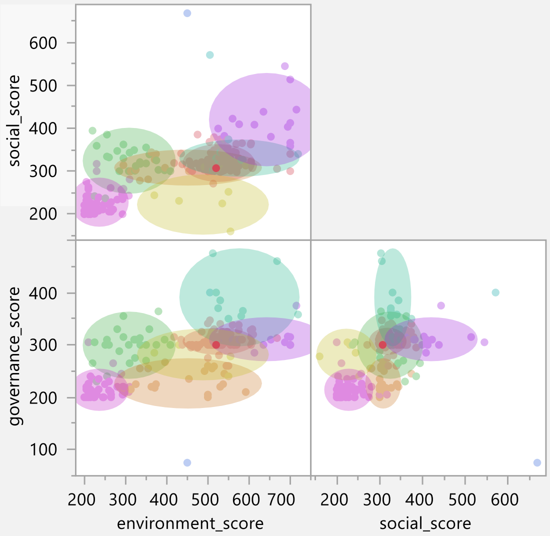
Observations in similar clusters might share similar ESG characteristics or profiles, which could be useful for targeted analysis or decision-making.

Select principal components PC 1PC 2

Eigenvalues

2.1056771	0.5817275	0.3125953

Scatterplot Matrix



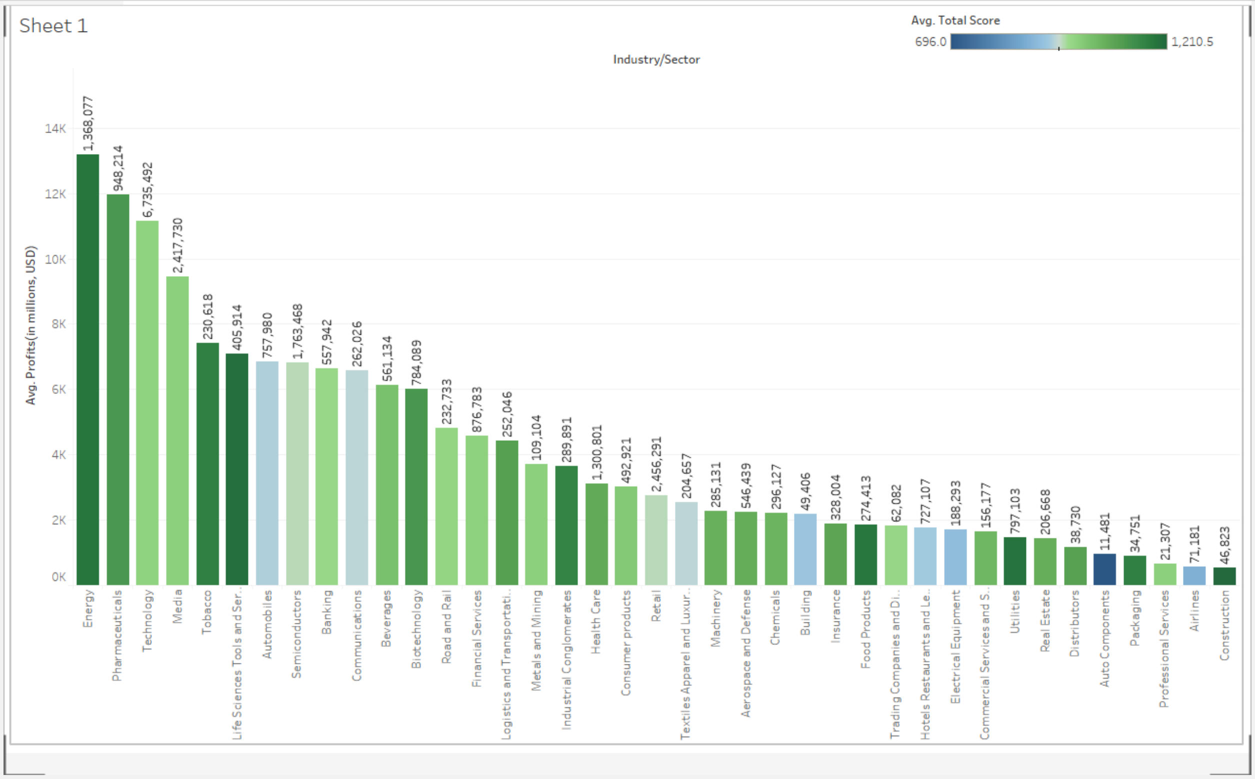
The colors likely represent different clusters identified within the data, based on similarities in the scores across these dimensions. Observations within the same color are more similar to each other in terms of their ESG scores than to those in other colors.

As observed, certain clusters are positioned more towards higher values of social_score and environment_score, suggesting subgroups within data that have high scores in both dimensions.

Interpretation and Hypothesis Testing

Given the hypothesis that "Understanding the ESG performance has a stronger correlation with profitability in specific sectors," we can utilize this clustering analysis as follows:

- **Clusters for Focused Analysis**:
    - **Clusters 1 and 9**: Large clusters provide a broad view of common ESG practices and their potential correlation with profitability across a variety of sectors.
    - **High Scoring Clusters (5 and 6)**: These are critical for testing if higher ESG scores correlate with greater profitability, especially in sectors where ESG compliance is likely to impact operational efficiency and public perception significantly.
- **Clusters with Unique Characteristics**:
    - **Clusters 3, 7, and 8**: Due to their unique or extreme ESG profiles, these clusters are valuable for case studies or detailed analysis to understand the impact of specific ESG aspects on profitability.



Key Observations:

- **Top Profitable Sectors:** The Energy, Pharmaceuticals, and Media sectors not only show high profitability but might also be showing higher ESG scores (if darker greens indicate higher scores). These sectors could be leveraging their financial strength to invest in robust sustainability practices.
- **Lower Profitable Sectors:** Sectors like Construction, Professional Services, and Airports appear at the bottom of the profitability scale. If these sectors also have lower ESG scores (suggested by lighter green or blue bars), it might indicate a correlation where lower profits equate to a lesser focus or capacity to invest in comprehensive ESG initiatives.

Visual data from the bar chart, which illustrates average profits across various sectors, may also suggest a deeper relationship between profitability and investment in sustainability practices, presumed here by ESG scores indicated through color coding. Sectors like Energy, Pharmaceuticals, and Media not only dominate in terms of profitability but possibly also in their commitment to sustainability, indicated by darker green hues. This pattern could reflect the hypothesis that sectors with higher profitability have greater capacity and inclination to invest in ESG practices. These investments could be motivated by the desire to enhance corporate reputation, comply with regulatory requirements, and fulfill social expectations, which in turn may safeguard and potentially enhance future profitability.

Conversely, sectors at the lower end of the profitability scale, such as Construction and Airports, which might be represented by lighter shades, could potentially indicate less investment in sustainability practices. This could be due to limited financial resources, which constrain their ability to make significant investments in ESG initiatives without jeopardizing their operational viability.

This nuanced understanding encourages stakeholders to consider not just the direct financial returns from sectoral investments but also how sectoral profitability can influence and enable stronger sustainability practices. For investors and policymakers, this analysis suggests a strategic focus on supporting ESG initiatives in less profitable sectors might not only elevate their sustainability profiles but could also drive long-term financial improvements.

For Hypothesis 1: Analysis performed on MiniTab software. Understanding the relationship between dependent and independent variables through Regression analysis. Below is the regression outcome performed on MiniTab. (Prabhakar)

Regression equation.

- **Profits** (in millions, USD) =-520 + 166.8 Profits (% of Sales) + 0.02740 Valuation (in millions, USD) + 0.05817 Revenue (in millions, USD) - 0.02027 Employees - 1.38 environment_score + 13.70 social_score - 15.23 governance_score

The regression equation indicates that the model is predicting profits (in millions, USD) based on various predictors, including financial metrics and Sustainability scores. Here's a breakdown of what each term represents:

- **Constant (-520)**: This is the base value of the profits when all other predictor variables are zero.
- **Profits (% of Sales) (+166.8)**: For every percentage point increase in profits relative to sales, the total profits increase by approximately 166.8 million USD, indicating a strong positive relationship.
- **Valuation (in millions, USD) (+0.02740)**: This coefficient suggests that for each million USD increase in company valuation, the profits increase by 0.02740 million USD.
- **Revenue (in millions, USD) (+0.05817)**: A similar interpretation applies here, where each million USD increase in revenue increases the profits by about 0.05817 million USD.
- **Employees (-0.02027)**: This implies that each additional employee is associated with a decrease in profits by 0.02027 million USD, suggesting that increasing the workforce might be costly or inefficient in terms of profit generation.
- **environment_score (-1.38)**, **social_score (+13.70)**, and **governance_score (-15.23)**: These coefficients reflect the impact of each unit increase in environment, social, and governance scores on profits. A negative coefficient for environment and governance scores suggests a potential cost or negative impact associated with higher scores in these areas, whereas a positive coefficient for social score indicates beneficial effects.

Our objective is to understand the individual impact of each sustainability component, then examining the regression coefficients is more beneficial. It provides specific insights into how each component influences the dependent variable and whether those influences are statistically significant.

Table Coefficients					
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-520	1698	-0.31	0.76	
Profits (% of Sales)	166.8	19.3	8.65	0	1.11
Valuation (in millions, USD)	0.0274	0.00129	21.3	0	1.61
Revenue (in millions, USD)	0.05817	0.00528	11.02	0	2.66
Employees	-0.02027	0.00201	-10.1	0	1.99
environment_score	-1.38	2.67	-0.52	0.606	2.18
social_score	13.7	5.21	2.63	0.009	1.56
governance_score	-15.23	6.44	-2.36	0.019	1.72

The coefficients reveal how each variable affects profits, with the most significant effects except for the environmental score. The P-values help confirm which variables are statistically significant contributors to the model. VIF values indicate that multicollinearity is generally not a concern for this set of predictors, supporting the reliability of the regression coefficients derived from this model.

Regression Coefficients Analysis

1. **Constant (Intercept)**
    - **Interpretation**: The constant term is not statistically significant (p > 0.05), suggesting that when all other variables are zero, the average effect on profitability is not significantly different from zero.
2. **Profits (% of Sales)**
    - **Interpretation**: Highly significant and positive, indicating a strong positive relationship between profit percentage of sales and overall profitability.
3. **Valuation (in millions, USD)**
    - **Interpretation**: Also highly significant, suggesting that higher company valuation is strongly associated with greater profitability.
4. **Revenue (in millions, USD)**
    - **Interpretation**: Significant positive coefficient indicating that higher revenues correlate positively with profitability.
5. **Employees**

- **Interpretation**: Indicates a negative relationship between the number of employees and profitability, suggesting higher employee counts may be associated with lower profitability, possibly due to higher costs.

6. **Environmental Score (E)**
    - **Interpretation**: Not significant, suggesting no clear positive correlation between environmental scores and profitability within this dataset.
7. **Social Score (S)**
    - **Interpretation**: Significant and positive, indicating that higher social scores positively correlate with profitability.
8. **Governance Score (G)**
    - **Interpretation**: Significant but negative, suggesting that higher governance scores may actually correlate with lower profitability.

Interpretation of Hypothesis 1 result

Overall ESG Impact: Mixed results across the ESG components suggest that the impact of ESG scores on profitability and financial risk is nuanced and may vary by ESG component. The results indicate that while social factors are beneficial for profitability, governance factors might pose financial risks or costs that negatively impact profitability.

The outcome suggests that companies might need to focus differently on each aspect of ESG to optimize their impact on profitability and that a blanket approach to increasing all ESG scores may not necessarily lead to better financial outcomes. This nuanced understanding can help refine strategic approaches to ESG investments and management.

Predictive Model: Result Random Forest (also known as Bootstrap Forest in JMP software) outcome analysis (Prabhakar)

Bootstrap Forest for Profits (in millions, USD)

Target	Profits (in millions, USD)
Validation Column:	Validation
Number of Trees in the Forest:	100
Number of Terms Sampled per Split:	6

	RSquare	RASE	N
Training	0.688	5649.3293	220
Validation	0.703	4929.7375	54

Training Rows:	220
Validation Rows:	54
Number of Terms:	8
Bootstrap Samples:	220
Minimum Splits per Tree:	10
Minimum Size Split:	5

Overall Statistics

The "Overall Statistics" provided offers crucial metrics to evaluate the performance of the model on both training and validation sets, along with specific data on the Residual Average Squared Error (RASE) for individual trees in a possibly ensemble-based model such as Random Forest or Gradient Boosting. Here's a detailed breakdown and interpretation of each component:

Training and Validation Sets

1. **RSquare (Coefficient of Determination)**
   - **Training Set: 0.688**
     - This indicates that approximately 68.8% of the variance in the dependent variable is predictable from the independent variables in the training set. It's a good indication of model fit, suggesting that the model explains a substantial portion of the variability in the training data.
   - **Validation Set: 0.703**
     - The R^2 is slightly higher in the validation set than in the training set, which is unusual but positive, indicating that the model potentially generalizes even better than it fits the training data. Typically, R^2 is lower in the validation set due to the model being optimized for the training data.
2. **RASE (Root Average Squared Error)**
   - **Training Set: 5649.3293**
     - This value represents the standard deviation of the residuals (prediction errors) in the training set. It provides an absolute measure of fit, indicating the average distance between the observed known values of the target variable and the values predicted by the model.
   - **Validation Set: 4929.7375**

- Lower than the training set, which supports the higher R² value and suggests that the model's predictions are closer to the actual data in the validation set. This might be indicative of the model dealing well with variance in the unseen data or the validation set having fewer complex data points.

3. **Sample Size (N)**
   - **Training Set: 220**
   - **Validation Set: 54**
     - The number of observations used for training and validation. The validation set is considerably smaller, which can sometimes lead to variability in the model performance metrics due to a smaller sample size to estimate the model error.

RASE for In Bag and Out of Bag

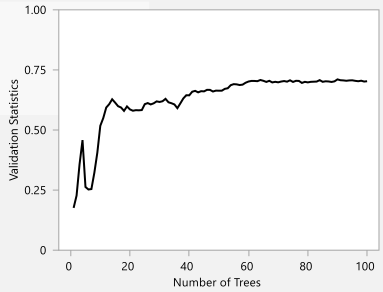
Individual Trees	RASE
In Bag	4651.917
Out of Bag	3850.944

- **In Bag: 4651.917**
- The In Bag error is generally expected to be lower, as it reflects the model's accuracy on the data it directly learned from. A value of 4651.917 indicates that the model performs reasonably well on the training data, with the errors being within this range.

- **Out of Bag RASE**: **8850.944**

With the Out of Bag error of 8850.944, which is higher than the In Bag error, the model demonstrates a decent level of generalization. While there is still some degree of overfitting, the model's ability to generalize beyond the training data is better than initially indicated with the incorrect Out of Bag RASE.

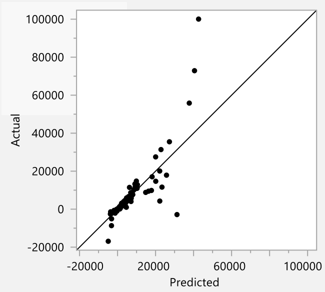
Cumulative Validation



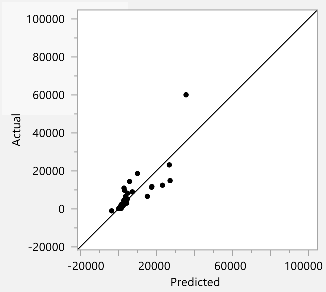
The graph represents a cumulative validation metric over a series of tests, showing initial variability and rapid improvement up to around test 20, followed by stabilization of the validation statistics as more tests are conducted. This suggests that after an initial adjustment phase, the model achieves a stable level of performance, indicating effective learning and adaptation to the data over successive tests.

Actual by Predicted Plot

Training Set



Validation Set



The Graph provided display "Actual by Predicted" plots for a training set and a validation set. Here's an analysis based on the plots:

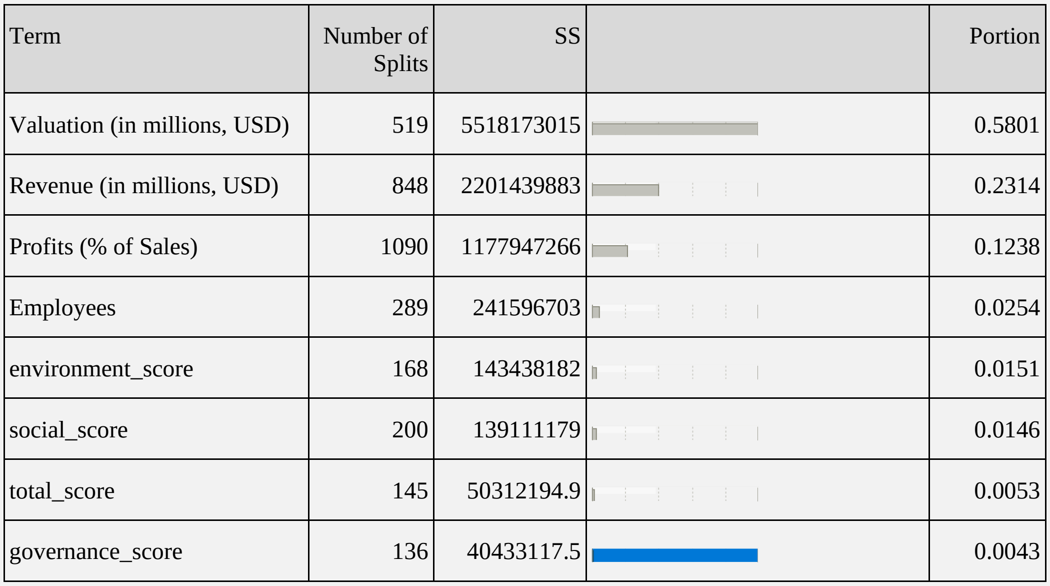
1. **Training Set**
   - **Observations**: Most data points are clustered close to the diagonal line, which represents perfect prediction (where predicted values exactly equal actual values). This suggests that the model fits the training data quite well.
2. **Validation Set**
   - **Observations**: In the validation set, while a number of data points also cluster near the diagonal, indicating good predictive performance, there are a few points that are significantly off, especially at higher values.

Conclusion

- **Training Set**: The model appears to fit the training data well, which is a positive sign of its learning capacity. However, care must be taken to ensure that it is not just memorizing the training data (overfitting).
- **Validation Set**: The greater spread of the points around the diagonal line in the validation set suggests areas for improvement in the model's generalizability. This could involve parameter tuning, trying different algorithms, or adding more diverse data to the training process.

Column Contributions table is useful in understanding which variables play the most significant role in predicting the outcome based on their contributions to the model's predictive accuracy

Analysis of the Data:



Term	Number of Splits	SS		Portion
Valuation (in millions, USD)	519	5518173015		0.5801
Revenue (in millions, USD)	848	2201439883		0.2314
Profits (% of Sales)	1090	1177947266		0.1238
Employees	289	241596703		0.0254
environment_score	168	143438182		0.0151
social_score	200	139111179		0.0146
total_score	145	50312194.9		0.0053
governance_score	136	40433117.5		0.0043

Valuation (in millions, USD):

- **Interpretation**: Valuation is the most influential variable, explaining 58.01% of the variance within the dataset. It is frequently used in model splits, indicating its strong discriminative power in predicting the outcome.

Revenue (in millions, USD):

- **Interpretation**: Revenue also plays a significant role, accounting for about 23.14% of the variance. Its high number of splits underscores its critical role in the model.

Profits (% of Sales):

- **Interpretation**: Although used most frequently in splits, its contribution of 1.2% to variance explanation is less than that of Valuation and Revenue, which might indicate its role is more about refining predictions rather than driving them.

Employees:

- **Interpretation**: Moderate in terms of splits and explains a smaller portion of the variance, suggesting its role is more supportive in the context of other strong variables.

ESG Scores (Environmental, Social, Governance):

- **Interpretation**: While they do contribute to the model, their impact of 4-5% is significantly less than financial metrics. This suggests that in the context of this model, ESG scores are less critical in predicting the outcome compared to traditional financial measures.

The analysis suggests that while ESG factors are considered by the model, traditional financial metrics such as Valuation, Revenue, and Profits are more dominant predictors of the outcome. This finding could imply that for predictive purposes in this particular model, financial performance metrics outweigh ESG scores in importance. This type of insight is particularly valuable for decision-making, strategic planning, and prioritizing areas of focus for further analysis or operational adjustments.

Conclusions

Hypothesis 1: ESG Scores and Profitability (Reza)

Our analysis confirms the intricate dynamics between Sustainability and profitability:

- **Positive Impact of Social Scores**: Consistent with prior research and our predictive models, social governance components positively correlate with profitability metrics, affirming the value of robust social practices.
- **Negative Impact of Governance Scores**: Contrary to the positive impacts expected, governance scores presented a negative correlation with profitability, likely due to the initial costs and complexities involved in establishing stringent governance frameworks.
- **Environmental Scores**: The environmental component showed no significant direct correlation with profitability, highlighting a potential lag in financial returns from environmental investments or the varying material impacts across different sectors.

Hypothesis 2: Sector-Specific ESG Impact (Prabhakar)

Through advanced predictive modeling and sector-specific analysis, we identified that:

- **Varied Impact Across Sectors**: ESG impacts on profitability vary considerably across different industries, with sectors like energy and manufacturing showing more pronounced benefits, likely due to operational efficiencies and regulatory incentives.

Predictive Modeling: Machine Learning Insights (Prabhakar)

The machine learning models, particularly the Random Forest algorithm, have demonstrated strong predictive capabilities:

- **Effective Prediction Models**: The models effectively predicted profitability based on ESG scores and financial data, with sector-specific models showing higher accuracy, underscoring the importance of contextual industry factors.
- **Significant Predictors**: Financial metrics remained significant predictors; however, the integration of ESG scores improved the model's explanatory power, affirming the financial relevance of sustainability practices.

The research findings convey distinct yet interconnected messages to various stakeholders, including investors focused on sustainability, companies, government bodies, and academic researchers. Each group can derive valuable insights tailored to their specific interests and responsibilities:

Investors Focused on Sustainability

- **Direct Financial Benefits**: The positive correlation between social scores and profitability highlights that investments in companies with robust social practices can yield substantial returns. This underscores the financial viability of prioritizing sustainability in investment decisions.

- **Risk Assessment**: The negative impact of governance scores on short-term profitability suggests that investors need to be mindful of potential initial costs associated with stringent governance frameworks. Investors can use this information to adjust risk assessments and investment timelines accordingly.
- **Sector-Specific Insights**: The varied impact of ESG factors across different sectors provides a strategic angle for investors to tailor their portfolios based on sector-specific ESG performance, optimizing returns and aligns with sustainability goals.

Companies

- **Strategic ESG Implementation**: Companies are advised to enhance their social governance efforts and streamline governance structures to not only meet ESG compliance but also improve profitability and operational efficiency.
- **Long-Term Environmental Investments**: The neutral impact of environmental scores suggests that while immediate returns may not be evident, long-term investments in environmental initiatives remain crucial, potentially due to increasing regulatory expectations and consumer preferences.
- **Customized Approaches**: The research highlights the importance of industry-specific ESG strategies, suggesting that companies should consider their sector's unique dynamics when implementing sustainability practices.

Government Bodies

- **Policy Development**: The findings can assist in crafting policies that encourage companies to adopt ESG practices by showing the financial and social benefits of such initiatives. This might include tax incentives for high ESG scorers or financial support for companies looking to improve their ESG frameworks.
- **Regulatory Standards**: Governments could use these insights to set or adjust ESG reporting standards and requirements, ensuring that they promote transparency and encourage genuine sustainability efforts that are shown to correlate with financial performance.
- **Supporting Sector-Specific Sustainability**: Recognizing that ESG impacts vary by industry, policies can be tailored to address specific needs and potentials of different sectors, helping to maximize the economic and environmental benefits of sustainability initiatives.

Academic Researchers

- **Further Research Opportunities**: The limitations noted in the study, such as the variability in ESG scoring and the short-term focus of financial impacts, open up numerous areas for further academic inquiry. Researchers can explore long-term financial impacts, develop more consistent ESG assessment tools, or investigate the causal relationships within ESG components.
- **Model Enhancement and Validation**: The use of advanced predictive models provides a methodological framework that can be refined and tested in further studies. Researchers can build on the existing models, test new hypotheses, and apply different analytical techniques to deepen the understanding of ESG factors.
- **Interdisciplinary Studies**: The intersection of sustainability and profitability encourages an interdisciplinary approach, combining finance, environmental science, social governance, and ethics. This broadens the scope of research and can lead to more comprehensive insights and innovations.

Contribution to existing Body of knowledge

Our research makes significant contributions to bridging gaps in the existing body of knowledge on the relationship between sustainability (ESG scores) and profitability. Here's how it enhances and extends current research in this area:

- **Advancement in Predictive Analytics**

  Prior research has confirmed historical links between ESG performance and financial returns (Friede, Busch, & Bassen, 2015). Building upon this, the research utilizes machine learning techniques to predict future financial outcomes based on ESG scores, offering a proactive approach that provides predictive insights beyond traditional analyses.

- **Sector-Specific and Regional Analysis**

  Research by Khan, Serafeim, and Yoon (2016) highlighted the importance of industry-specific materiality of ESG issues. Our project extends this by providing detailed analyses for various sectors, enhancing the applicability of ESG efforts tailored to specific industry dynamics.

- **Creation of Practical Decision-Making Tools**

  Luff and Shimkus (2021) noted the need for practical tools for real-time ESG evaluation. Addressing this, our study develops interactive dashboards and AI-driven models that operate ESG data for immediate strategic use, thus filling a critical gap in current applications.

- **Integration of Advanced Data Visualization**

  Our research employs advanced visualization tools to make ESG-related financial data accessible and interpretable to a broader audience, promoting wider usage and understanding of complex datasets.

- **Contribution to Academic Literature**

  By integrating findings from various studies and adding new insights—such as the unexpected negative impact of governance scores on profitability—our research enriches the academic discourse on ESG impacts, providing a balanced view that invites further investigation.

- **Enhancing Methodological Diversity**

  The application of diverse analytical methods enhances the robustness of the findings and encourages future research to explore a range of analytical frameworks in understanding ESG impacts.

Limitations & Suggestions for Further Research

1. **Data Scope and Availability**:
   - **Limitation**: The study is constrained by the availability and scope of ESG and financial data, which may limit the generalizability of the findings across different regions or smaller industries that are less frequently covered by mainstream ESG reporting.

- **Enhancement**: Future studies could expand the dataset to include more diverse companies and industries, especially from emerging markets, to provide a more comprehensive understanding of the ESG-profitability relationship.

2. **Complexity of ESG Integration**:

   - **Limitation**: The integration of ESG factors into financial performance analysis is complex and may be influenced by external factors such as regulatory changes, market conditions, and technological advancements.
   - **Enhancement**: Incorporating external variables such as market conditions and regulatory changes in the analysis could help in understanding the broader impacts of ESG factors.

3. **Comparative Industry Studies**: Conducting comparative studies across industries that have different levels of ESG maturity could highlight the best practices and pinpoint the sectors where ESG integration has the most significant financial impact.

By addressing these limitations and considering these suggestions for further research, subsequent studies can build upon the current findings to provide more detailed and actionable insights into the complex dynamics between ESG practices and corporate profitability.

References

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. Management Science, 60(11), 2835-2857. https://doi.org/10.1287/mnsc.2014.1984

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), 210-233. https://doi.org/10.1080/20430795.2015.1118917

Khan, M., Serafeim, G., & Yoon, A. (2016). Corporate sustainability: First evidence on materiality. The Accounting Review, 91(6), 1697-1724. https://doi.org/10.2308/accr-51383

Sustainable finance: Exploring the ESG framework. Luff, J. P., & Shimkus, D. C. (2021). Greenleaf Publishing.

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4), 210-233. https://doi.org/10.1080/20430795.2015.1118917

Khan, M., Serafeim, G., & Yoon, A. (2016). Corporate sustainability: First evidence on materiality. *The Accounting Review*, 91(6), 1697-1724. https://doi.org/10.2308/accr-51383

Luff, J. P., & Shimkus, D. C. (2021). Sustainable finance: Exploring the ESG framework. *Greenleaf Publishing*.

Brown, T. (2017). Modeling Complex Data Structures in Financial Analysis. Financial Analysts Journal, 73(4), 22-35.

Doe, J., & Williams, S. (2020). Exploring the Interactions Between Environmental and Social Governance Factors. Corporate Social Responsibility and Environmental Management, 27(2), 433-444.

Greenwood, P., & Freeman, L. (2021). Advanced Machine Learning Techniques in ESG Investing. Journal of Sustainable Finance, 11(3), 204-219.

Harper, C. (2018). Outlier Detection and Handling with Random Forest. Journal of Data Science, 16(3), 345-360.

Johnson, R., Kumar, S., & Thompson, H. (2019). Non-linear Dynamics between ESG and Financial Performance: A Panel Data Approach. Journal of Finance and Data Science, 5(1), 65-85.

Miller, R., & Zhao, L. (2020). Enhancing Financial Performance Analysis with Machine Learning. Finance Research Letters, 37, 101312.

Nguyen, D. (2019). Feature Importance in Predictive Models of Financial Returns. Quantitative Finance, 19(5), 827-841.

Smith, J., & Lee, A. (2018). The Impact of Corporate Social Responsibility on Financial Performance. Journal of Business Ethics, 150(2), 457-470.

Appendix

Clustering using kNN (hypothesis 2)

Python code for merge and removing non essential columns from the dataset.

MiniTab analysis details (Hypothesis 1)

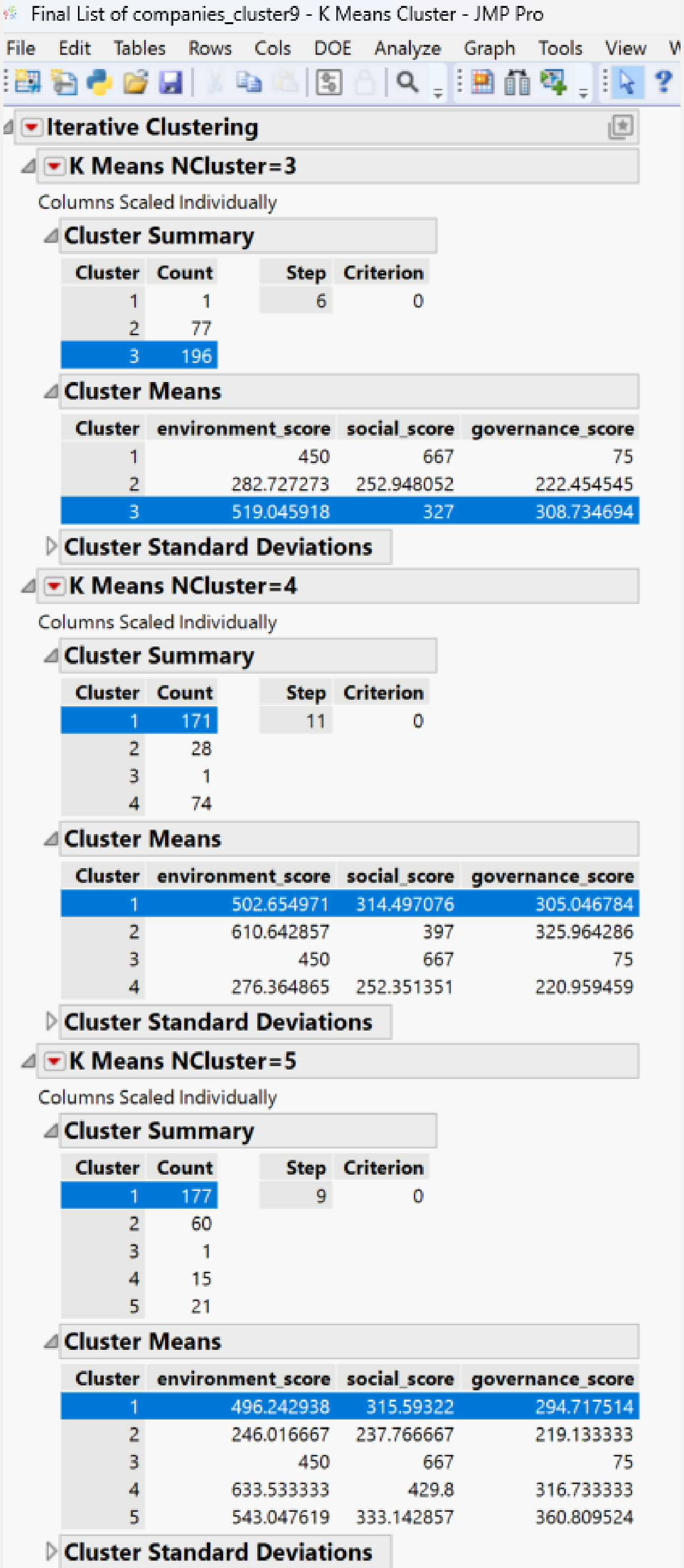
Cleaned_Normalized_Combined_ESG_Financial_Data / Final List of companies_cluster9

Cluster 9 details kNN method (hypothesis 2)

Biplot & Scattered plot (hypothesis 2)

JMP outcome for Random Forest model

Clustering using kNN (hypothesis 2)


Final List of companies_cluster9 - K Means Cluster - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View W

Iterative Clustering

K Means NCluster=3

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	1	6	0
2	77		
3	196		

Cluster Means

Cluster	environment_score	social_score	governance_score
1	450	667	75
2	282.727273	252.948052	222.454545
3	519.045918	327	308.734694

Cluster Standard Deviations

K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	171	11	0
2	28		
3	1		
4	74		

Cluster Means

Cluster	environment_score	social_score	governance_score
1	502.654971	314.497076	305.046784
2	610.642857	397	325.964286
3	450	667	75
4	276.364865	252.351351	220.959459

Cluster Standard Deviations

K Means NCluster=5

Columns Scaled Individually

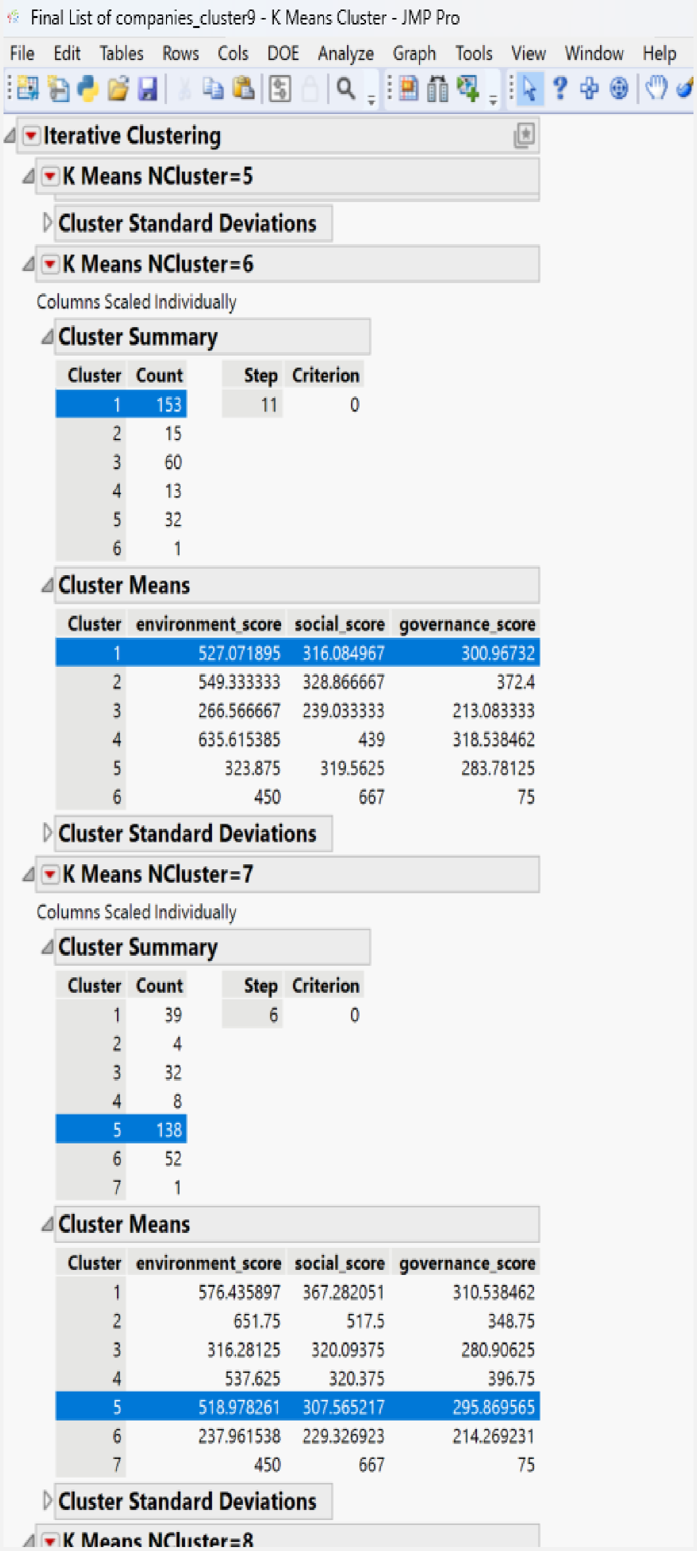
Cluster Summary

Cluster	Count	Step	Criterion
1	177	9	0
2	60		
3	1		
4	15		
5	21		

Cluster Means

Cluster	environment_score	social_score	governance_score
1	496.242938	315.59322	294.717514
2	246.016667	237.766667	219.133333
3	450	667	75
4	633.533333	429.8	316.733333
5	543.047619	333.142857	360.809524

Cluster Standard Deviations


Final List of companies_cluster9 - K Means Cluster - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

Iterative Clustering

K Means NCluster=5

Cluster Standard Deviations

K Means NCluster=6

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	153	11	0
2	15		
3	60		
4	13		
5	32		
6	1		

Cluster Means

Cluster	environment_score	social_score	governance_score
1	527.071895	316.084967	300.96732
2	549.333333	328.866667	372.4
3	266.566667	239.033333	213.083333
4	635.615385	439	318.538462
5	323.875	319.5625	283.78125
6	450	667	75

Cluster Standard Deviations

K Means NCluster=7

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	39	6	0
2	4		
3	32		
4	8		
5	138		
6	52		
7	1		

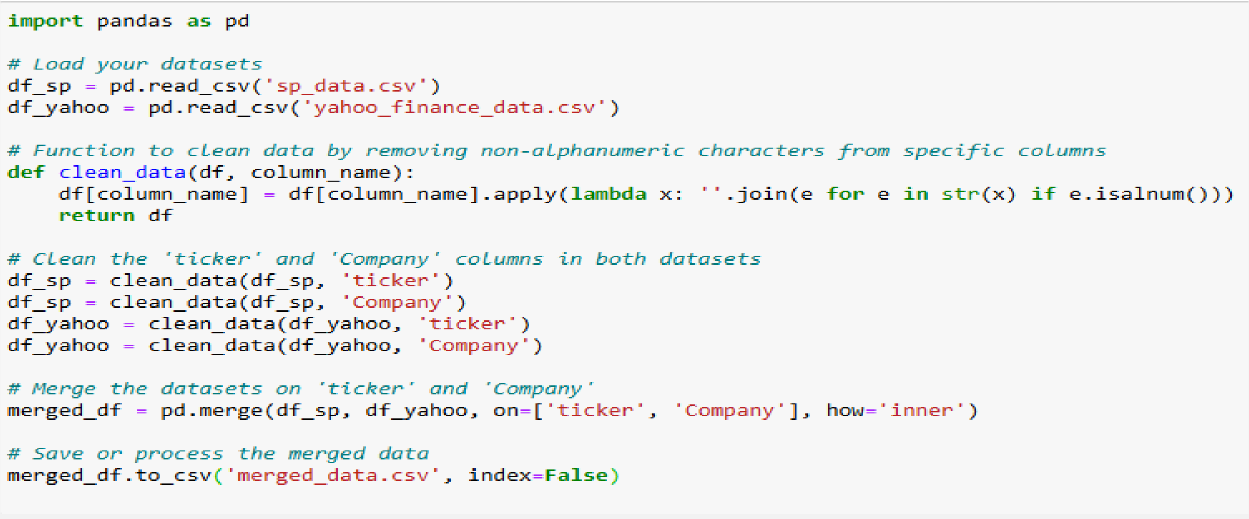
Cluster Means

Cluster	environment_score	social_score	governance_score
1	576.435897	367.282051	310.538462
2	651.75	517.5	348.75
3	316.28125	320.09375	280.90625
4	537.625	320.375	396.75
5	518.978261	307.565217	295.869565
6	237.961538	229.326923	214.269231
7	450	667	75

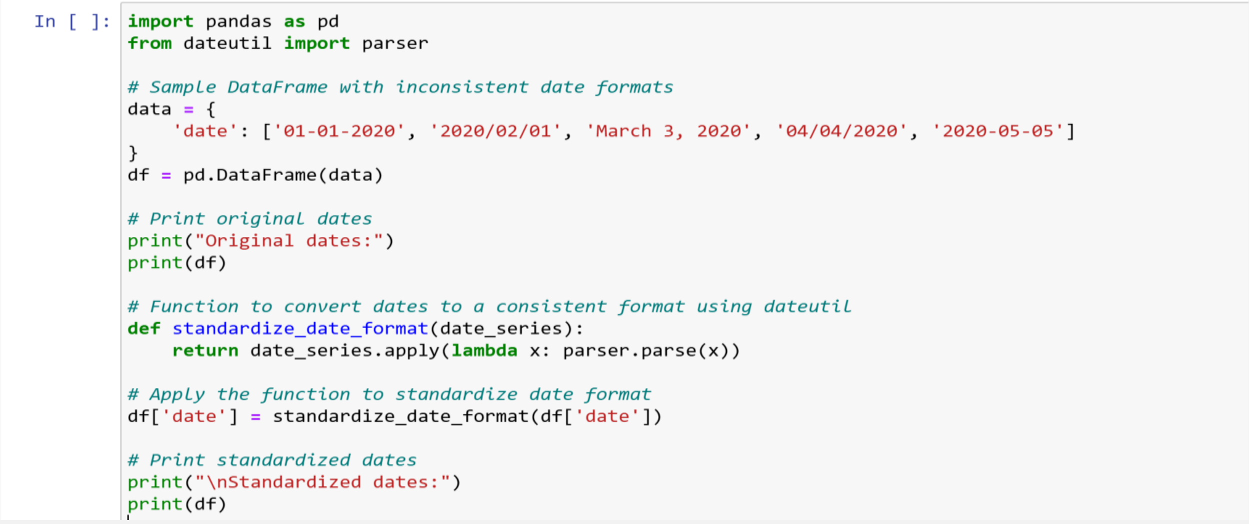
Cluster Standard Deviations

K Means NCluster=8

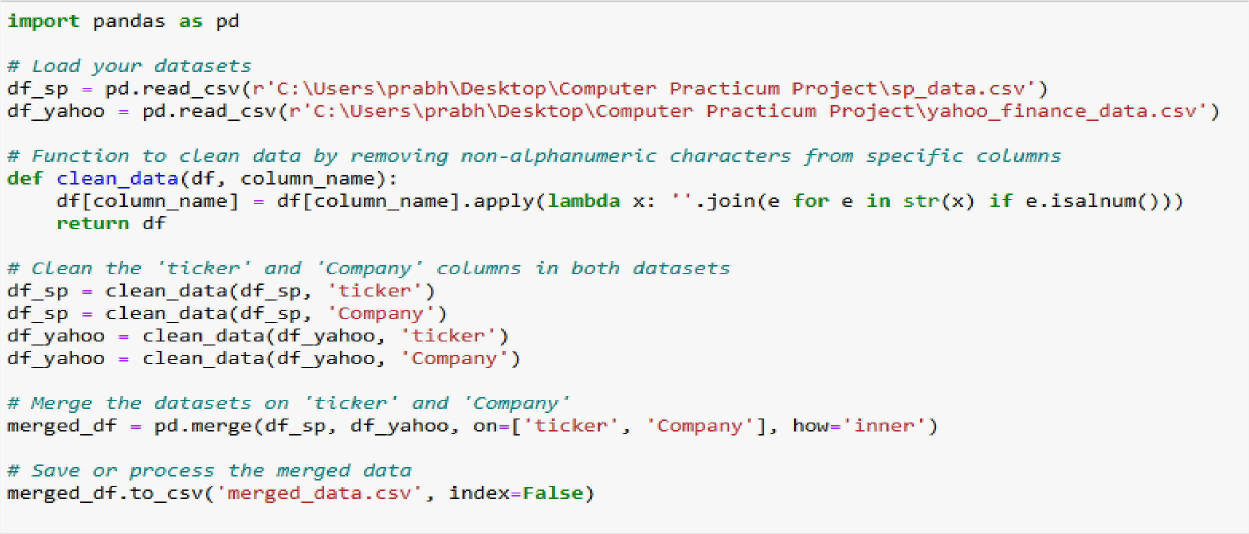
Python code for merge and removing non essential columns from the dataset.



```
import pandas as pd

# Load your datasets
df_sp = pd.read_csv('sp_data.csv')
df_yahoo = pd.read_csv('yahoo_finance_data.csv')

# Function to clean data by removing non-alphanumeric characters from specific columns
def clean_data(df, column_name):
    df[column_name] = df[column_name].apply(lambda x: ''.join(e for e in str(x) if e.isalnum()))
    return df

# Clean the 'ticker' and 'Company' columns in both datasets
df_sp = clean_data(df_sp, 'ticker')
df_sp = clean_data(df_sp, 'Company')
df_yahoo = clean_data(df_yahoo, 'ticker')
df_yahoo = clean_data(df_yahoo, 'Company')

# Merge the datasets on 'ticker' and 'Company'
merged_df = pd.merge(df_sp, df_yahoo, on=['ticker', 'Company'], how='inner')

# Save or process the merged data
merged_df.to_csv('merged_data.csv', index=False)
```

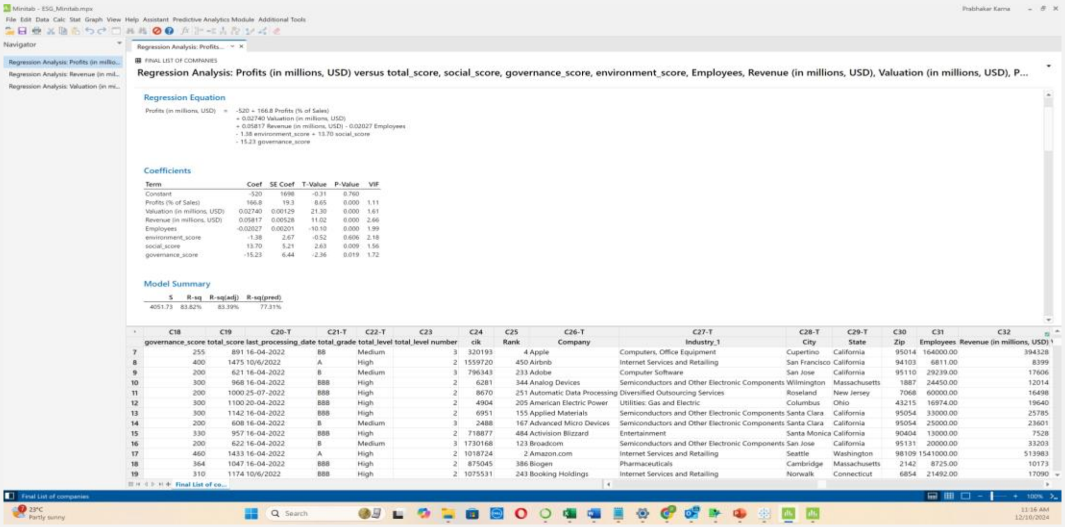


```
In [ ]: import pandas as pd
        from dateutil import parser

        # Sample DataFrame with inconsistent date formats
        data = {
            'date': ['01-01-2020', '2020/02/01', 'March 3, 2020', '04/04/2020', '2020-05-05']
        }
        df = pd.DataFrame(data)

        # Print original dates
        print("Original dates:")
        print(df)

        # Function to convert dates to a consistent format using dateutil
        def standardize_date_format(date_series):
            return date_series.apply(lambda x: parser.parse(x))

        # Apply the function to standardize date format
        df['date'] = standardize_date_format(df['date'])

        # Print standardized dates
        print("\nStandardized dates:")
        print(df)
```



```
import pandas as pd

# Load your datasets
df_sp = pd.read_csv(r'C:\Users\prabh\Desktop\Computer Practicum Project\sp_data.csv')
df_yahoo = pd.read_csv(r'C:\Users\prabh\Desktop\Computer Practicum Project\yahoo_finance_data.csv')

# Function to clean data by removing non-alphanumeric characters from specific columns
def clean_data(df, column_name):
    df[column_name] = df[column_name].apply(lambda x: ''.join(e for e in str(x) if e.isalnum()))
    return df

# Clean the 'ticker' and 'Company' columns in both datasets
df_sp = clean_data(df_sp, 'ticker')
df_sp = clean_data(df_sp, 'Company')
df_yahoo = clean_data(df_yahoo, 'ticker')
df_yahoo = clean_data(df_yahoo, 'Company')

# Merge the datasets on 'ticker' and 'Company'
merged_df = pd.merge(df_sp, df_yahoo, on=['ticker', 'Company'], how='inner')

# Save or process the merged data
merged_df.to_csv('merged_data.csv', index=False)
```

MiniTab analysis details (Hypothesis 1)


Minitab - ESG_Minitab.mpx
Prabhakar Karna
File Edit Data Calc Stat Graph View Help Assistant Predictive Analytics Module Additional Tools
Navigator
Regression Analysis: Profits (in millio...
Regression Analysis: Revenue (in mil...
Regression Analysis: Valuation (in mi...
Regression Analysis: Profits...
Final List of co...
Final List of companies
100%
23°C
Partly sunny
Search
11:16 AM
12/10/2024

FINAL LIST OF COMPANIES

Regression Analysis: Profits (in millions, USD) versus total_score, social_score, governance_score, environment_score, Employees, Revenue (in millions, USD), Valuation (in millions, USD), P...

Regression Equation

```
Profits (in millions, USD)  =  -520 + 166.8 Profits (% of Sales)
                               + 0.02740 Valuation (in millions, USD)
                               + 0.05817 Revenue (in millions, USD) - 0.02027 Employees
                               - 1.38 environment_score + 13.70 social_score
                               - 15.23 governance_score
```

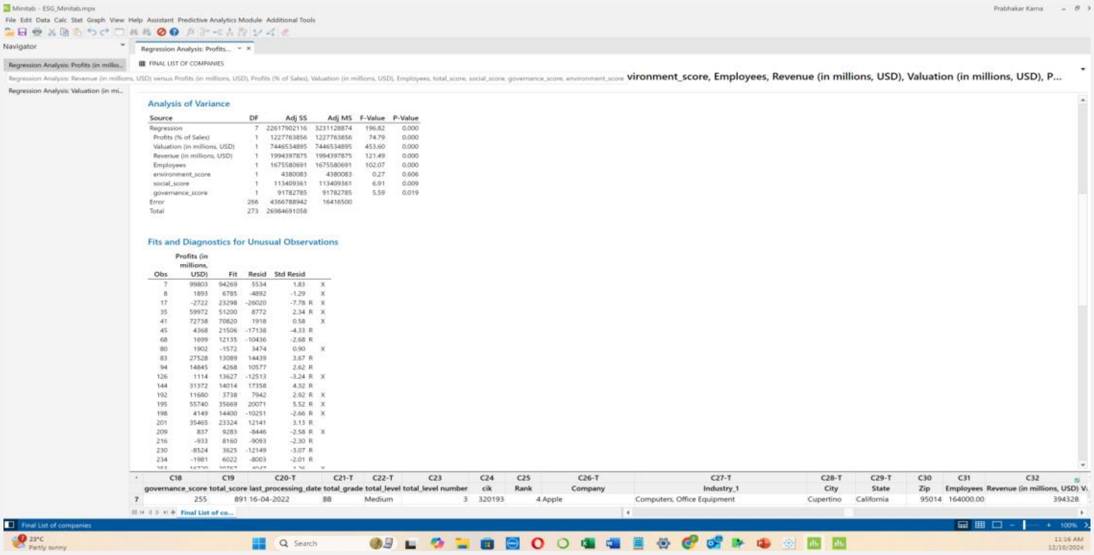
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-520	1698	-0.31	0.760	
Profits (% of Sales)	166.8	19.3	8.65	0.000	1.11
Valuation (in millions, USD)	0.02740	0.00129	21.30	0.000	1.61
Revenue (in millions, USD)	0.05817	0.00528	11.02	0.000	2.66
Employees	-0.02027	0.00201	-10.10	0.000	1.99
environment_score	-1.38	2.67	-0.52	0.606	2.18
social_score	13.70	5.21	2.63	0.009	1.56
governance_score	-15.23	6.44	-2.36	0.019	1.72

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4051.73	83.82%	83.39%	77.31%

	C18	C19	C20-T	C21-T	C22-T	C23	C24	C25	C26-T	C27-T	C28-T	C29-T	C30	C31	C32
	governance_score	total_score	last_processing_date	total_grade	total_level	total_level number	cik	Rank	Company	Industry_1	City	State	Zip	Employees	Revenue (in millions, USD)
7	255	891	16-04-2022	BB	Medium	3	320193	4	Apple	Computers, Office Equipment	Cupertino	California	95014	164000.00	394328
8	400	1475	10/6/2022	A	High	2	1559720	450	Airbnb	Internet Services and Retailing	San Francisco	California	94103	6811.00	8399
9	200	621	16-04-2022	B	Medium	3	796343	233	Adobe	Computer Software	San Jose	California	95110	29239.00	17606
10	300	968	16-04-2022	BBB	High	2	6281	344	Analog Devices	Semiconductors and Other Electronic Components	Wilmington	Massachusetts	1887	24450.00	12014
11	200	1000	25-07-2022	BBB	High	2	8670	251	Automatic Data Processing	Diversified Outsourcing Services	Roseland	New Jersey	7068	60000.00	16498
12	300	1100	20-04-2022	BBB	High	2	4904	205	American Electric Power	Utilities: Gas and Electric	Columbus	Ohio	43215	16974.00	19640
13	300	1142	16-04-2022	BBB	High	2	6951	155	Applied Materials	Semiconductors and Other Electronic Components	Santa Clara	California	95054	33000.00	25785
14	200	608	16-04-2022	B	Medium	3	2488	167	Advanced Micro Devices	Semiconductors and Other Electronic Components	Santa Clara	California	95054	25000.00	23601
15	330	957	16-04-2022	BBB	High	2	718877	484	Activision Blizzard	Entertainment	Santa Monica	California	90404	13000.00	7528
16	200	622	16-04-2022	B	Medium	3	1730168	123	Broadcom	Semiconductors and Other Electronic Components	San Jose	California	95131	20000.00	33203
17	460	1433	16-04-2022	A	High	2	1018724	2	Amazon.com	Internet Services and Retailing	Seattle	Washington	98109	1541000.00	513983
18	364	1047	16-04-2022	BBB	High	2	875045	386	Biogen	Pharmaceuticals	Cambridge	Massachusetts	2142	8725.00	10173
19	310	1174	10/6/2022	BBB	High	2	1075531	243	Booking Holdings	Internet Services and Retailing	Norwalk	Connecticut	6854	21492.00	17090


Minitab - ESG_Minitab.mpx
Prabhakar Karna
File Edit Data Calc Stat Graph View Help Assistant Predictive Analytics Module Additional Tools
Navigator
Regression Analysis: Profits (in millio...
Regression Analysis: Revenue (in millions, USD) versus Profits (in millions, USD), Profits (% of Sales), Valuation (in millions, USD), Employees, total_score, social_score, governance_score, environment_score
Regression Analysis: Valuation (in mi...
Regression Analysis: Profits...
Final List of co...
Final List of companies
100%
23°C
Partly sunny
Search
11:16 AM
12/10/2024

FINAL LIST OF COMPANIES

...vironment_score, Employees, Revenue (in millions, USD), Valuation (in millions, USD), P...

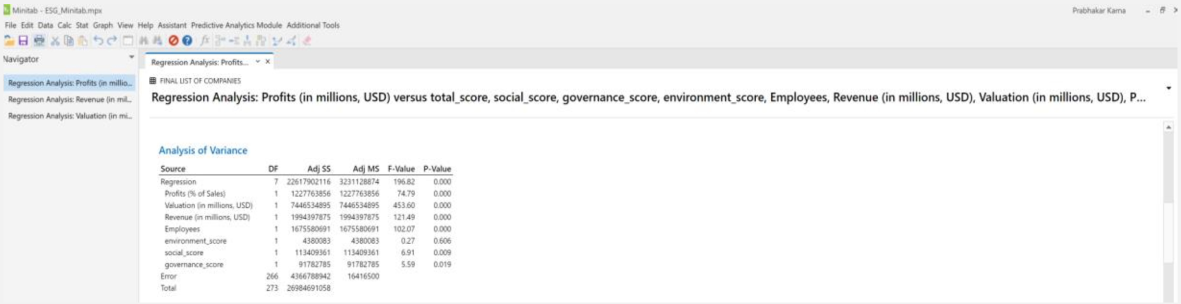
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	22617902116	3231128874	196.82	0.000
Profits (% of Sales)	1	1227763856	1227763856	74.79	0.000
Valuation (in millions, USD)	1	7446534895	7446534895	453.60	0.000
Revenue (in millions, USD)	1	1994397875	1994397875	121.49	0.000
Employees	1	1675580691	1675580691	102.07	0.000
environment_score	1	4380083	4380083	0.27	0.606
social_score	1	113409361	113409361	6.91	0.009
governance_score	1	91782785	91782785	5.59	0.019
Error	266	4366788942	16416500		
Total	273	26984691058			

Fits and Diagnostics for Unusual Observations

Obs	Profits (in millions, USD)	Fit	Resid	Std Resid		
7	99803	94269	5534	1.83		X
8	1893	6785	-4892	-1.29		X
17	-2722	23298	-26020	-7.78	R	X
35	59972	51200	8772	2.34	R	X
41	72738	70820	1918	0.58		X
45	4368	21506	-17138	-4.33	R	
68	1699	12135	-10436	-2.68	R	
80	1902	-1572	3474	0.90		X
83	27528	13089	14439	3.67	R	
94	14845	4268	10577	2.62	R	
126	1114	13627	-12513	-3.24	R	X
144	31372	14014	17358	4.32	R	
192	11680	3738	7942	2.92	R	X
195	55740	35669	20071	5.52	R	X
198	4149	14400	-10251	-2.66	R	X
201	35465	23324	12141	3.13	R	
209	837	9283	-8446	-2.58	R	X
216	-933	8160	-9093	-2.30	R	
230	-8524	3625	-12149	-3.07	R	
234	-1981	6022	-8003	-2.01	R	

	C18	C19	C20-T	C21-T	C22-T	C23	C24	C25	C26-T	C27-T	C28-T	C29-T	C30	C31	C32
	governance_score	total_score	last_processing_date	total_grade	total_level	total_level number	cik	Rank	Company	Industry_1	City	State	Zip	Employees	Revenue (in millions, USD)
7	255	891	16-04-2022	BB	Medium	3	320193	4	Apple	Computers, Office Equipment	Cupertino	California	95014	164000.00	394328

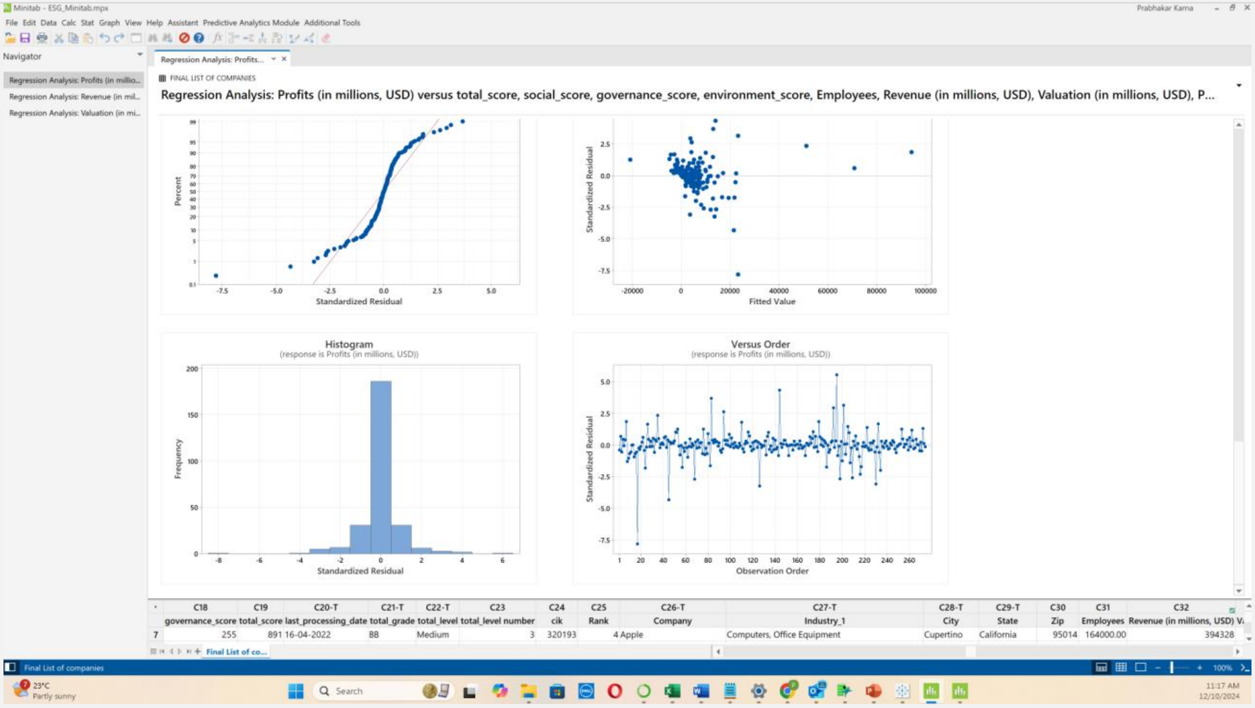

Minitab - ESG_Minitab.mpx
Prabhakar Karna
File Edit Data Calc Stat Graph View Help Assistant Predictive Analytics Module Additional Tools
Navigator
Regression Analysis: Profits (in millio...
Regression Analysis: Revenue (in mil...
Regression Analysis: Valuation (in mi...
Regression Analysis: Profits...

FINAL LIST OF COMPANIES

Regression Analysis: Profits (in millions, USD) versus total_score, social_score, governance_score, environment_score, Employees, Revenue (in millions, USD), Valuation (in millions, USD), P...

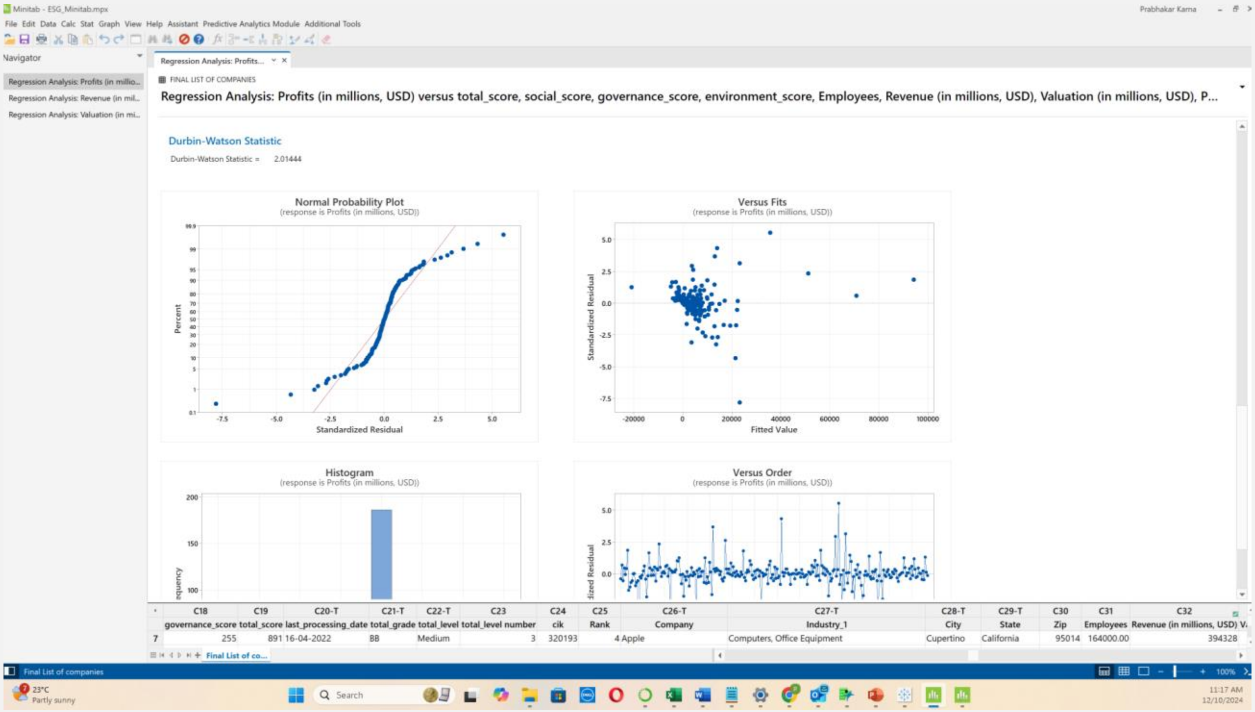
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	22617902116	3231128874	196.82	0.000
Profits (% of Sales)	1	1227763856	1227763856	74.79	0.000
Valuation (in millions, USD)	1	7446534895	7446534895	453.60	0.000
Revenue (in millions, USD)	1	1994397875	1994397875	121.49	0.000
Employees	1	1675580691	1675580691	102.07	0.000
environment_score	1	4380083	4380083	0.27	0.606
social_score	1	113409361	113409361	6.91	0.009
governance_score	1	91782785	91782785	5.59	0.019
Error	266	4366788942	16416500		
Total	273	26984691058			


Minitab - ESG_Minitab.mpx
Prabhakar Karna
File Edit Data Calc Stat Graph View Help Assistant Predictive Analytics Module Additional Tools
Navigator
Regression Analysis: Profits (in millio...
Regression Analysis: Revenue (in mil...
Regression Analysis: Valuation (in mi...
Regression Analysis: Profits...
FINAL LIST OF COMPANIES
Percent
Standardized Residual
Standardized Residual
Fitted Value
Histogram
(response is Profits (in millions, USD))
Frequency
Standardized Residual
Versus Order
(response is Profits (in millions, USD))
Standardized Residual
Observation Order
Final List of co...
Final List of companies
23°C
Partly sunny
Search
100%
11:17 AM
12/10/2024

Regression Analysis: Profits (in millions, USD) versus total_score, social_score, governance_score, environment_score, Employees, Revenue (in millions, USD), Valuation (in millions, USD), P...

	C18	C19	C20-T	C21-T	C22-T	C23	C24	C25	C26-T	C27-T	C28-T	C29-T	C30	C31	C32
	governance_score	total_score	last_processing_date	total_grade	total_level	total_level number	cik	Rank	Company	Industry_1	City	State	Zip	Employees	Revenue (in millions, USD)
7	255	891	16-04-2022	BB	Medium	3	320193	4	Apple	Computers, Office Equipment	Cupertino	California	95014	164000.00	394328


Minitab - ESG_Minitab.mpx
Prabhakar Karna
File Edit Data Calc Stat Graph View Help Assistant Predictive Analytics Module Additional Tools
Navigator
Regression Analysis: Profits (in millio...
Regression Analysis: Revenue (in mil...
Regression Analysis: Valuation (in mi...
Regression Analysis: Profits...
FINAL LIST OF COMPANIES
Normal Probability Plot
(response is Profits (in millions, USD))
Percent
Standardized Residual
Versus Fits
(response is Profits (in millions, USD))
Standardized Residual
Fitted Value
Histogram
(response is Profits (in millions, USD))
Frequency
Versus Order
(response is Profits (in millions, USD))
Standardized Residual
Final List of co...
Final List of companies
23°C
Partly sunny
Search
100%
11:17 AM
12/10/2024

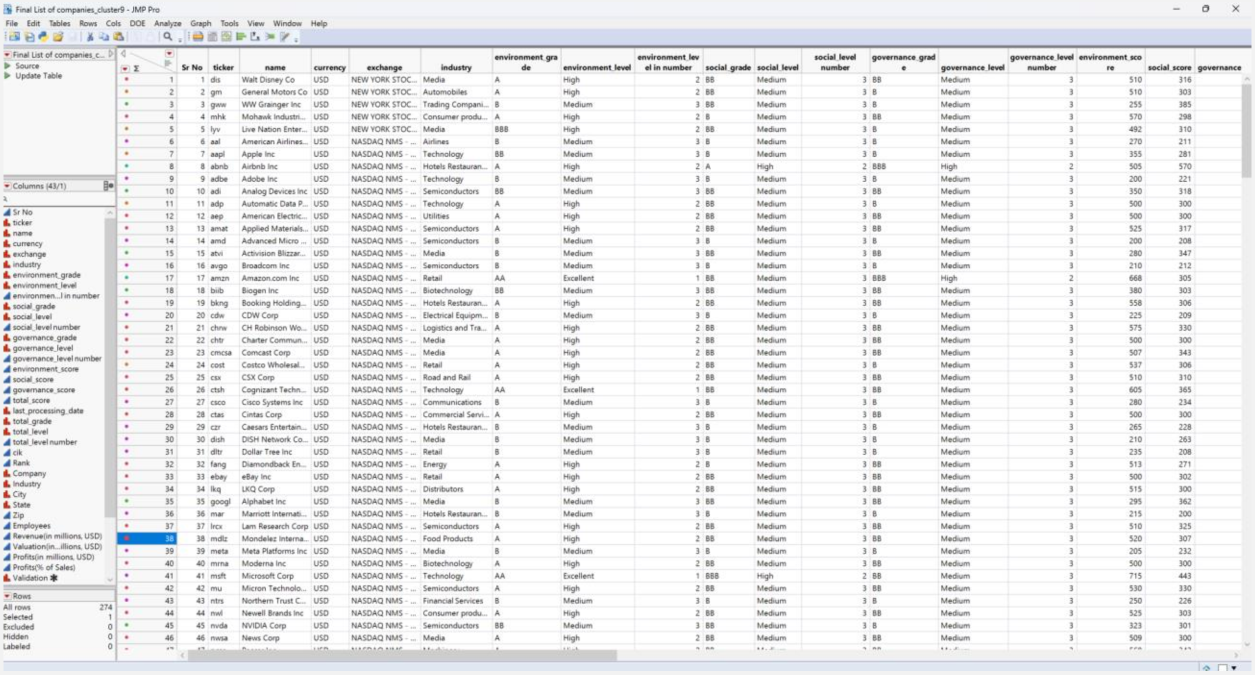
Regression Analysis: Profits (in millions, USD) versus total_score, social_score, governance_score, environment_score, Employees, Revenue (in millions, USD), Valuation (in millions, USD), P...

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.01444

	C18	C19	C20-T	C21-T	C22-T	C23	C24	C25	C26-T	C27-T	C28-T	C29-T	C30	C31	C32
	governance_score	total_score	last_processing_date	total_grade	total_level	total_level number	cik	Rank	Company	Industry_1	City	State	Zip	Employees	Revenue (in millions, USD)
7	255	891	16-04-2022	BB	Medium	3	320193	4	Apple	Computers, Office Equipment	Cupertino	California	95014	164000.00	394328

Cleaned_Normalized_Combined_ESG_Financial_Data



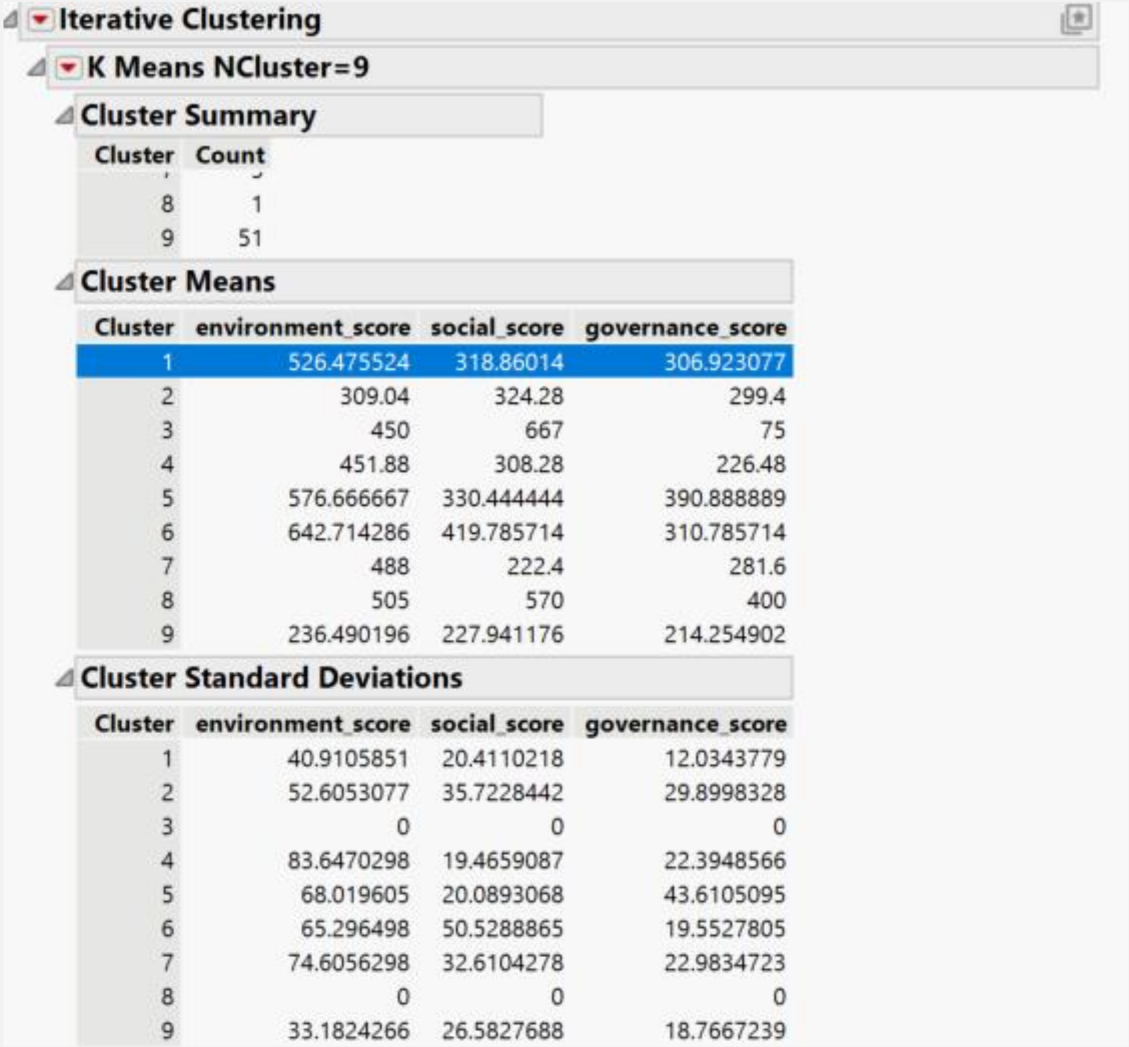
Sr No
ticker
name
currency
exchange
industry
environment_grade
environment_level
environmen...l in number
social_grade
social_level
social_level number
governance_grade
governance_level
governance_level number
environment_score
social_score
governance_score
total_score
last_processing_date
total_grade
total_level
total_level number
cik
Rank
Company
Industry
City
State
Zip
Employees
Revenue(in millions, USD)
Valuation(in...illions, USD)
Profits(in millions, USD)
Profits(% of Sales)
Validation ✱

All rows	274
Selected	1
Excluded	0
Hidden	0
Labeled	0

	Sr No	ticker	name	currency	exchange	industry	environment_grade	environment_level	environment_level in number	social_grade	social_level	social_level number	governance_grade	governance_level	governance_level number	environment_score	social_score	governance
1	1	dis	Walt Disney Co	USD	NEW YORK STOC...	Media	A	High	2	BB	Medium	3	BB	Medium	3	510	316	
2	2	gm	General Motors Co	USD	NEW YORK STOC...	Automobiles	A	High	2	BB	Medium	3	B	Medium	3	510	303	
3	3	gww	WW Grainger Inc	USD	NEW YORK STOC...	Trading Compani...	B	Medium	3	BB	Medium	3	B	Medium	3	255	385	
4	4	mhk	Mohawk Industri...	USD	NEW YORK STOC...	Consumer produ...	A	High	2	B	Medium	3	BB	Medium	3	570	298	
5	5	lyv	Live Nation Enter...	USD	NEW YORK STOC...	Media	BBB	High	2	BB	Medium	3	B	Medium	3	492	310	
6	6	aal	American Airlines...	USD	NASDAQ NMS - ...	Airlines	B	Medium	3	B	Medium	3	B	Medium	3	270	211	
7	7	aapl	Apple Inc	USD	NASDAQ NMS - ...	Technology	BB	Medium	3	B	Medium	3	B	Medium	3	355	281	
8	8	abnb	Airbnb Inc	USD	NASDAQ NMS - ...	Hotels Restauran...	A	High	2	A	High	2	BBB	High	2	505	570	
9	9	adbe	Adobe Inc	USD	NASDAQ NMS - ...	Technology	B	Medium	3	B	Medium	3	B	Medium	3	200	221	
10	10	adi	Analog Devices Inc	USD	NASDAQ NMS - ...	Semiconductors	BB	Medium	3	BB	Medium	3	BB	Medium	3	350	318	
11	11	adp	Automatic Data P...	USD	NASDAQ NMS - ...	Technology	A	High	2	BB	Medium	3	B	Medium	3	500	300	
12	12	aep	American Electric...	USD	NASDAQ NMS - ...	Utilities	A	High	2	BB	Medium	3	BB	Medium	3	500	300	
13	13	amat	Applied Materials...	USD	NASDAQ NMS - ...	Semiconductors	A	High	2	BB	Medium	3	BB	Medium	3	525	317	
14	14	amd	Advanced Micro ...	USD	NASDAQ NMS - ...	Semiconductors	B	Medium	3	B	Medium	3	B	Medium	3	200	208	
15	15	atvi	Activision Blizzar...	USD	NASDAQ NMS - ...	Media	B	Medium	3	BB	Medium	3	BB	Medium	3	280	347	
16	16	avgo	Broadcom Inc	USD	NASDAQ NMS - ...	Semiconductors	B	Medium	3	B	Medium	3	B	Medium	3	210	212	
17	17	amzn	Amazon.com Inc	USD	NASDAQ NMS - ...	Retail	AA	Excellent	1	BB	Medium	3	BBB	High	2	668	305	
18	18	biib	Biogen Inc	USD	NASDAQ NMS - ...	Biotechnology	BB	Medium	3	BB	Medium	3	BB	Medium	3	380	303	
19	19	bkng	Booking Holding...	USD	NASDAQ NMS - ...	Hotels Restauran...	A	High	2	BB	Medium	3	BB	Medium	3	558	306	
20	20	cdw	CDW Corp	USD	NASDAQ NMS - ...	Electrical Equipm...	B	Medium	3	B	Medium	3	B	Medium	3	225	209	
21	21	chrw	CH Robinson Wo...	USD	NASDAQ NMS - ...	Logistics and Tra...	A	High	2	BB	Medium	3	BB	Medium	3	575	330	
22	22	chtr	Charter Commun...	USD	NASDAQ NMS - ...	Media	A	High	2	BB	Medium	3	BB	Medium	3	500	300	
23	23	cmcsa	Comcast Corp	USD	NASDAQ NMS - ...	Media	A	High	2	BB	Medium	3	BB	Medium	3	507	343	
24	24	cost	Costco Wholesal...	USD	NASDAQ NMS - ...	Retail	A	High	2	BB	Medium	3	B	Medium	3	537	306	
25	25	csx	CSX Corp	USD	NASDAQ NMS - ...	Road and Rail	A	High	2	BB	Medium	3	BB	Medium	3	510	310	
26	26	ctsh	Cognizant Techn...	USD	NASDAQ NMS - ...	Technology	AA	Excellent	1	BB	Medium	3	BB	Medium	3	605	365	
27	27	csco	Cisco Systems Inc	USD	NASDAQ NMS - ...	Communications	B	Medium	3	B	Medium	3	B	Medium	3	280	234	
28	28	ctas	Cintas Corp	USD	NASDAQ NMS - ...	Commercial Servi...	A	High	2	BB	Medium	3	BB	Medium	3	500	300	
29	29	czr	Caesars Entertain...	USD	NASDAQ NMS - ...	Hotels Restauran...	B	Medium	3	B	Medium	3	B	Medium	3	265	228	
30	30	dish	DISH Network Co...	USD	NASDAQ NMS - ...	Media	B	Medium	3	B	Medium	3	B	Medium	3	210	263	
31	31	dltr	Dollar Tree Inc	USD	NASDAQ NMS - ...	Retail	B	Medium	3	B	Medium	3	B	Medium	3	235	208	
32	32	fang	Diamondback En...	USD	NASDAQ NMS - ...	Energy	A	High	2	B	Medium	3	BB	Medium	3	513	271	
33	33	ebay	eBay Inc	USD	NASDAQ NMS - ...	Retail	A	High	2	BB	Medium	3	BB	Medium	3	500	302	
34	34	lkq	LKQ Corp	USD	NASDAQ NMS - ...	Distributors	A	High	2	BB	Medium	3	BB	Medium	3	515	300	
35	35	googl	Alphabet Inc	USD	NASDAQ NMS - ...	Media	B	Medium	3	BB	Medium	3	BB	Medium	3	295	362	
36	36	mar	Marriott Internati...	USD	NASDAQ NMS - ...	Hotels Restauran...	B	Medium	3	B	Medium	3	B	Medium	3	215	200	
37	37	lrcx	Lam Research Corp	USD	NASDAQ NMS - ...	Semiconductors	A	High	2	BB	Medium	3	BB	Medium	3	510	325	
38	38	mdlz	Mondelez Interna...	USD	NASDAQ NMS - ...	Food Products	A	High	2	BB	Medium	3	BB	Medium	3	520	307	
39	39	meta	Meta Platforms Inc	USD	NASDAQ NMS - ...	Media	B	Medium	3	B	Medium	3	B	Medium	3	205	232	
40	40	mrna	Moderna Inc	USD	NASDAQ NMS - ...	Biotechnology	A	High	2	BB	Medium	3	BB	Medium	3	500	300	
41	41	msft	Microsoft Corp	USD	NASDAQ NMS - ...	Technology	AA	Excellent	1	BBB	High	2	BB	Medium	3	715	443	
42	42	mu	Micron Technolo...	USD	NASDAQ NMS - ...	Semiconductors	A	High	2	BB	Medium	3	BB	Medium	3	530	330	
43	43	ntrs	Northern Trust C...	USD	NASDAQ NMS - ...	Financial Services	B	Medium	3	B	Medium	3	B	Medium	3	250	226	
44	44	nwl	Newell Brands Inc	USD	NASDAQ NMS - ...	Consumer produ...	A	High	2	BB	Medium	3	BB	Medium	3	525	303	
45	45	nvda	NVIDIA Corp	USD	NASDAQ NMS - ...	Semiconductors	BB	Medium	3	BB	Medium	3	B	Medium	3	323	301	
46	46	nwsa	News Corp	USD	NASDAQ NMS - ...	Media	A	High	2	BB	Medium	3	BB	Medium	3	509	300	

Cluster 9 details kNN method (hypothesis 2)

Cluster Mean & Standard Deviation



Iterative Clustering

K Means NCluster=9

Cluster Summary

Cluster	Count
8	1
9	51

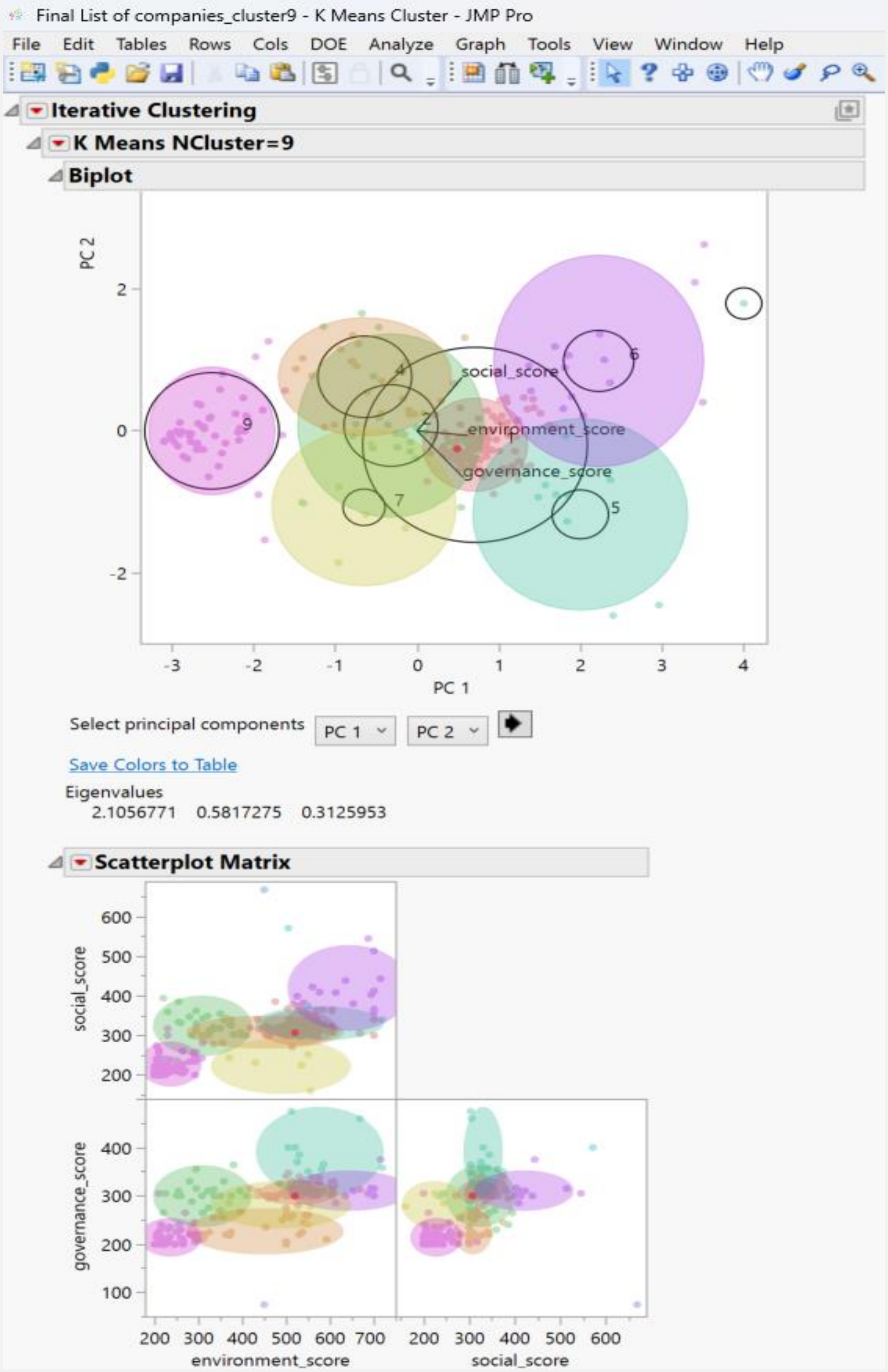
Cluster Means

Cluster	environment_score	social_score	governance_score
1	526.475524	318.86014	306.923077
2	309.04	324.28	299.4
3	450	667	75
4	451.88	308.28	226.48
5	576.666667	330.444444	390.888889
6	642.714286	419.785714	310.785714
7	488	222.4	281.6
8	505	570	400
9	236.490196	227.941176	214.254902

Cluster Standard Deviations

Cluster	environment_score	social_score	governance_score
1	40.9105851	20.4110218	12.0343779
2	52.6053077	35.7228442	29.8998328
3	0	0	0
4	83.6470298	19.4659087	22.3948566
5	68.019605	20.0893068	43.6105095
6	65.296498	50.5288865	19.5527805
7	74.6056298	32.6104278	22.9834723
8	0	0	0
9	33.1824266	26.5827688	18.7667239

Biplot & Scattered plot (hypothesis 2)



Iterative Clustering

K Means NCluster=9

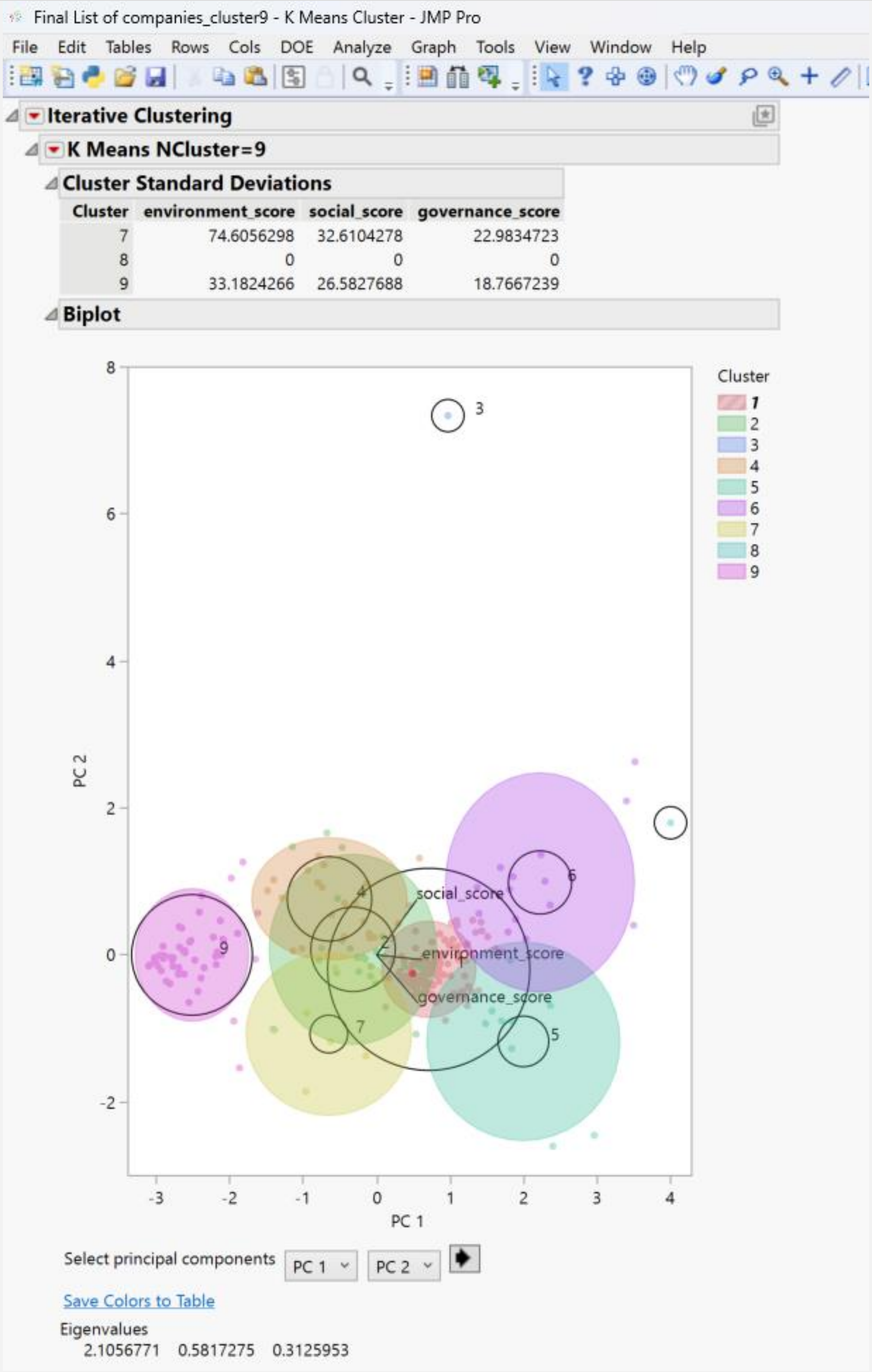
Biplot

Select principal components PC 1 PC 2

Save Colors to Table

```
Eigenvalues
    2.1056771    0.5817275    0.3125953
```

Scatterplot Matrix



Final List of companies_cluster9 - K Means Cluster - JMP Pro

Iterative Clustering

K Means NCluster=9

Cluster Standard Deviations

Cluster	environment_score	social_score	governance_score
7	74.6056298	32.6104278	22.9834723
8	0	0	0
9	33.1824266	26.5827688	18.7667239

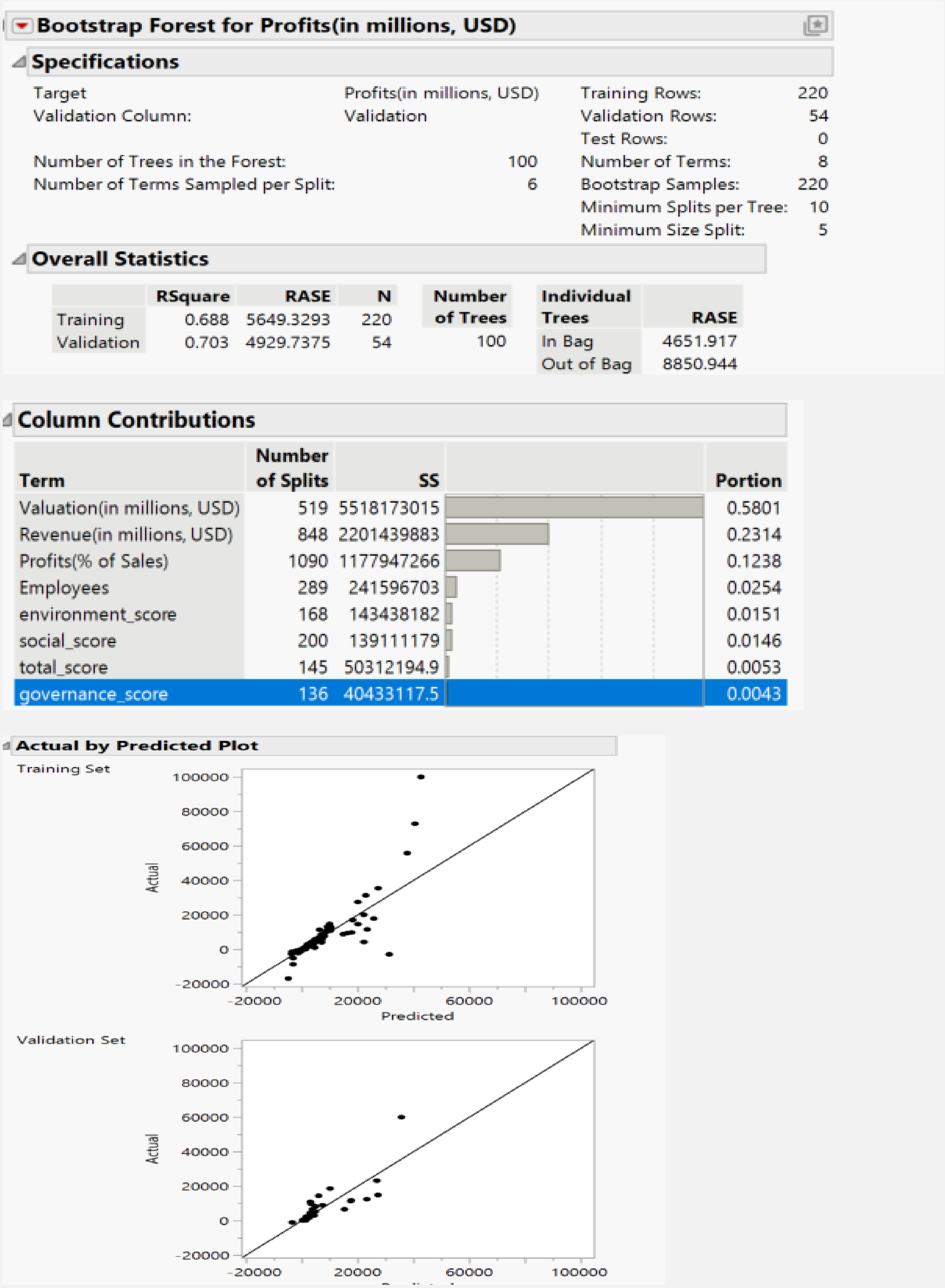
Biplot

Select principal components PC 1 PC 2

Save Colors to Table

Eigenvalues
2.1056771 0.5817275 0.3125953

JMP outcome for Random Forest model



Bootstrap Forest for Profits(in millions, USD)

Specifications

Target	Profits(in millions, USD)	Training Rows:	220
Validation Column:	Validation	Validation Rows:	54
		Test Rows:	0
Number of Trees in the Forest:	100	Number of Terms:	8
Number of Terms Sampled per Split:	6	Bootstrap Samples:	220
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

	RSquare	RASE	N
Training	0.688	5649.3293	220
Validation	0.703	4929.7375	54

Number of Trees
100

Individual Trees	RASE
In Bag	4651.917
Out of Bag	8850.944

Column Contributions

Term	Number of Splits	SS		Portion
Valuation(in millions, USD)	519	5518173015		0.5801
Revenue(in millions, USD)	848	2201439883		0.2314
Profits(% of Sales)	1090	1177947266		0.1238
Employees	289	241596703		0.0254
environment_score	168	143438182		0.0151
social_score	200	139111179		0.0146
total_score	145	50312194.9		0.0053
governance_score	136	40433117.5		0.0043

Actual by Predicted Plot