Multivariate Analysis of USA Companies

Multivariate Analysis Final Project

Anna Monsó Rodríguez, Walter José Troiani Vargas, Joan Acero Pousa

Contents

1. Introduction **2**

2. Exploratory Data Analysis **2**

Dataset Preprocessing 2

Variable Description 3

Variable Transformation 5

3. Multivariate Analysis **5**

Principal Component Analysis (PCA) 5

Factor Analysis 6

Multidimensional Scaling (MDS) 7

Correspondence Analysis (CA) 9

Cluster Analysis Profiling 11

Discriminant Analysis 12

4. Conclusions **14**

5. Bibliography **14**

Appendix **15**

1. Introduction

This project aims to integrate all the methods covered in the Multivariate Analysis course within the same context, providing practical experience with each technique. It follows the complete pipeline of an analysis process, from data acquisition and preparation to a comprehensive application of the multivariate methods introduced during the course.

The primary objective of our specific case is to analyze various aspects of companies based in the United States. To achieve this, the first step was to develop a Python web-scraping script to retrieve data from Yahoo Finance, a referent in the economic sector. Using its API, we collected data from over 5,000 companies worldwide. For the purpose of this project, we filtered the dataset using the Pandas library to include just 666 companies based in the USA. Additionally, we selected 14 variables out of the original 153, focusing on those most relevant to our analysis objectives.

After completing the data collection and preprocessing steps, we obtained our starting point dataset (*USACompanies*), which contains 666 observations and 14 variables: 10 numerical and four categorical (including the company name, which is the observation's ID).

In the following Exploratory Data Analysis section, we explain and assess all variables one by one to comprehend them before starting to apply the following analysis methods in the Multivariate Analysis part: Principal Component Analysis, Factor Analysis, Multidimensional Scaling, Correspondence Analysis, Cluster Analysis Profiling, and Discriminant Analysis.

2. Exploratory Data Analysis

This section provides insights into the data and ensures it is prepared for proper analysis. We conduct a general analysis of the data as well as an individual examination of each variable. Moreover, we evaluate the distributions of the variables along with their missing values and perform an imputation process and validation.

Dataset Preprocessing

Although initial preprocessing was performed during data retrieval from Yahoo Finance, we identified a critical issue that required resolution before conducting the multivariate analysis.

Specifically, we observed significant magnitude differences among certain numeric variables (*marketCap*, *enterpriseValue*, *totalCash*, *totalDebt*, and *totalRevenue*), which could bias the analysis. To address this, we applied logarithmic transformations to rescale these variables, ensuring a more comparable scale across the dataset. This adjustment resulted in distributions that are now better suited for the subsequent analyses (see Figure A.1 and Figure A.2 in the Appendix).

The next step was to assess missing values. First, variables *overallRisk* and *dividendRate* were removed because they had 140 (21%) and 243 (36%) missing values, respectively. Then, we imputed missing values on

the remaining variables - *marketCap* (1), *enterpriseValue* (32), *profitMargins* (56), *totalCash* (29), *totalCashPerShare* (41), *totalDebt* (37), *totalRevenue* (31), and *revenueGrowth* (50) - using the mice package. We checked the distributions of the variables after imputations to make sure that this process was not biasing our dataset. Comparing Figure A.3 and Figure A.4 in the Appendix, one can see that the distributions remain nearly the same, and thus the imputation results are valid.

Variable Description

In this section, we provide a brief explanation of each variable, including an assessment of their distribution for numerical variables.

Categorical variables

- *shortName:* Identifies the official name of a company, and it is the main supplementary categorical variable. All values are unique.
- *state* (39 levels)*:* Identifies the state of the company. There are 43 unique state values. It follows an exponential distribution that is almost Zipfian, with 'CA', 'NY', and 'TX' accounting for the vast majority of the companies
- *sector* (11 levels): The sector in which the enterprise works. The most common industries are Financial Services (124 companies), Technology (101), and Healthcare (87).
- *recommendationKey* (5 levels)*:* Yahoo Finance financial recommendation: "Buy" (391), "Hold" (167), "None" (94), “Strong_buy”(10), “Underperform”(4)

Numerical variables

Given that the data originates from a real-world source, it is challenging to identify perfectly defined distributions among the variables, and none of the numerical variables exhibit a normal distribution. All referenced plots discussed in the subsequent analysis are provided in the Appendix, spanning Figures A.5 to A.20.

- *marketCap:* Log-transformed market value (mean: 22.38, median: 23.59). The histogram shows a right-skewed distribution, and the Q-Q plot indicates an apparent deviation from normality. Moreover, Shapiro and Kolmogorov-Smirnov are rejected. Thus, we treat this variable as not normally distributed.
- *enterpriseValue:* The enterprise's log-transformed valuation has a mean of 21.88 and a median of 23.81. The corresponding histogram reveals a highly skewed distribution characterized by extreme negative and positive values. Additionally, a noticeable gap with no observations is present between the values of -10 and 10.
- *profitMargins:* Net income-to-revenue ratio (mean: 0.13, median). The histogram exhibits a right-skewed distribution, with most values concentrated near zero and a long tail extending toward higher values. This pattern highlights the presence of extreme outliers. The Q-Q plot further confirms substantial deviations from normality, particularly in the tails. Both the Shapiro-Wilk and Kolmogorov-Smirnov tests reject normality.

- *totalCash:* Log-transformed cash values (mean: 19.58, median: 20.22). The histogram suggests a slightly right-skewed distribution, with most values concentrated around the central range but a few smaller values extending the left tail. The Q-Q plot demonstrates deviations from normality, particularly in the lower tail, reflecting these extreme values. Both the Shapiro-Wilk and Kolmogorov-Smirnov tests reject the hypothesis of normality.
- *totalCashPerShare:* Cash per share (USD) (mean: 1.64, median: 1.46). The histogram clearly shows a non-normal distribution.
- *totalDebt:* Log-transformed debt (mean: 21.14, median: 21.93). The histogram, together with the Q-Q plot, shows a right-skewed distribution.
- *totalRevenue:* Log-transformed revenue (mean: 21.51, median: 22.43). The distribution includes three negative values and a gap between -15 and 5. The histogram indicates a slightly right-skewed pattern, with the majority of values clustered around the center and a few smaller values stretching into the left tail.
- *revenueGrowth:* Daily revenue growth percentage (mean: 0.07, median: 0.04). The histogram reveals a non-normal distribution, with the majority of values heavily concentrated around 0.

Relations between variables

The correlation matrix reveals key relationships between numerical variables. Market capitalization (*marketCap*) is positively correlated with *enterpriseValue* (0.51), *totalCash* (0.81), *totalDebt* (0.82), and *totalRevenue* (0.74), indicating that larger companies tend to have more cash, higher debt, and greater revenues. However, it shows a weak negative correlation with *profitMargins* (-0.10).

EnterpriseValue has moderate positive correlations with *totalCash* (0.32), *totalDebt* (0.48), and *totalRevenue* (0.52), suggesting that higher enterprise value companies also tend to have more cash, debt, and revenue. *ProfitMargins* is negatively correlated with *totalCash* (-0.15) and *totalRevenue* (-0.51), implying that more profitable companies may have lower cash reserves and revenues.

TotalCash is strongly correlated with *totalDebt* (0.77) and *totalRevenue* (0.68), suggesting that companies with larger cash reserves also have more debt and higher revenues. It shows a moderate correlation with *totalCashPerShare* (0.65).

TotalDebt exhibits strong positive correlations with *marketCap* (0.82), *totalCash* (0.77), and *totalRevenue* (0.71), indicating that companies with higher debt levels tend to have larger market capitalizations, more cash reserves, and greater revenues. It shows a weak negative correlation with *profitMargins* (-0.11).

TotalRevenue has a positive correlation with *marketCap* (0.74), *enterpriseValue* (0.52), *totalCash* (0.68), and *totalDebt* (0.71), suggesting that companies with higher revenues also tend to have larger market capitalizations, higher enterprise values, more cash, and greater debt. It also shows a notable negative correlation with *profitMargins* (-0.51), indicating that companies with higher revenues may experience diminishing returns in terms of profitability. Finally, *revenueGrowth* shows weak correlations with all other variables.

Variable Transformation

We performed logarithmic, square root, and Box-Cox transformations on all numeric variables to evaluate whether rescaling improved their adherence to a normal distribution. None of the original variables exhibited normality, but *totalCashPerShare* and shifted *totalCash* demonstrated distributions that closely approximate normality post-transformation. For these variables, the Kolmogorov-Smirnov test failed to reject the null hypothesis of normality or did so marginally at a 99% confidence level. The Q-Q plots revealed distributions that align well with a normal shape in their central regions, although deviations were observed in the tails.

A notable characteristic of the dataset is the concentration of most observations around a positive value, coupled with a smaller cluster of negative-valued observations at a considerable distance from the majority of the data. While removing these negative values might result in distributions that more closely approximate normality, we opted to retain them due to their importance. These observations represent enterprises with negative values in critical metrics such as market value, profit margins, revenue growth, total cash, or total revenue, making them essential for understanding the dataset's overall structure.

3. Multivariate Analysis

Principal Component Analysis (PCA)

This part of the analysis focuses on reducing the dimensionality of the dataset while retaining the maximum variability possible. We are going to conduct the Principal Component Analysis (PCA) based on the correlation matrix to identify patterns and relationships between the variables before deriving the principal components.

By applying the Kaiser criterion, the first two components should be retained, as their eigenvalues exceed 1. In the scree plot observed in Figure A.26 (elbow method), there is a sharp drop in variance after Component 1, followed by another notable drop after the second component, where the curve begins to flatten. Based on these results, it is appropriate to retain and extract two components for further analysis.

Dim 1	*marketCap* = 0.90	*totalDebt* = 0.88	*totalCash* = 0.88	*totalRevenue* = 0.88
Dim 2	*totalCashPerShare* = 0.69	*enterpriseValue* = -0.62	*profitMargins*= 0.44	*totalCash*= 0.32

Dimension 1 → “Company size and financial scale” : size-related financial metrics, large companies with high revenues, cash, and market caps will be strongly positive on this dimension. Medium-sized companies will occupy a moderate positive position, and small companies like startups with limited resources will be in the negative.

Dimension 2 → “Profitability and shareholder efficiency”: captures the balance between profitability and cash efficiency versus growth-focused valuation strategies. Companies with high *profitMargins* (profits relative to revenue) and *totalCashPerShare* (cash reserves per share) are positioned positively, reflecting

financial stability and efficiency. On the other hand, companies with high *enterpriseValue* (total company valuation, including debt) are positioned negatively, as they prioritize reinvestment and growth over immediate profitability or cash reserves. This dimension helps to distinguish between companies focused on expansion and future growth versus those focused on current financial efficiency and stability.

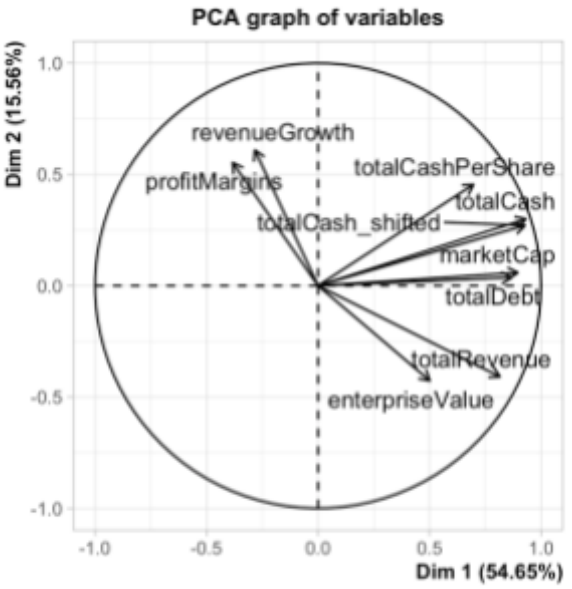

Figure 1: PCA graph of variables

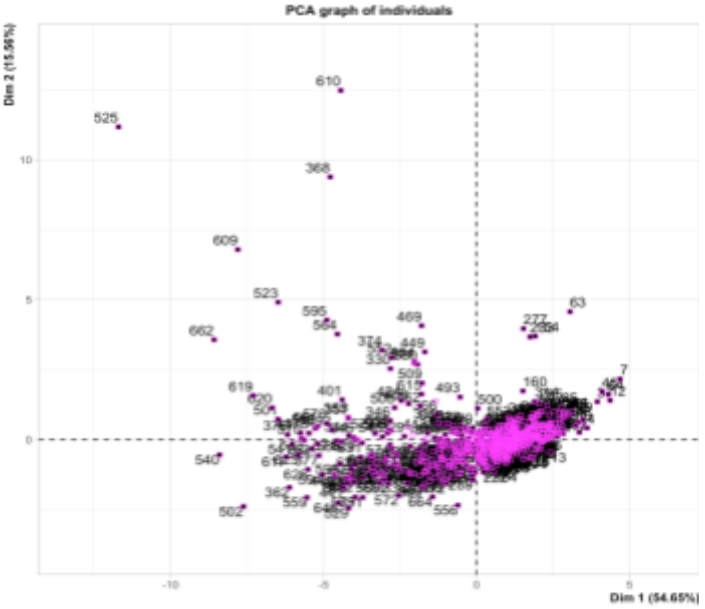

Figure 2: PCA graph of individuals

Types of companies:

1. Top-right quadrant: Large companies with high financial scale focused on stability and efficiency. (e. g. JP Morgan Chase & CO, (id 7) or Goldman Sachs Group, Inc (id 63))
2. Top-left quadrant: Small companies that are still highly efficient and profitable, like niche companies. (e. g. Mentor Capital (id 609) or Endurance Exploration Group Inc. (id 525))
3. Bottom-left quadrant: Small companies with low financial scale, low profitability, and limited cash reserves. These are likely startups, struggling companies, or in the early stages. (e.g Banneker Inc (id 540) or Pharmaroth labs Inc (id 502))
4. Bottom-right quadrant: Large companies that are less profitable or cash-efficient. Likely growth-focused companies that reinvest profits heavily (e.g, Walmart (id 13) or Apple Inc (id 1)

Factor Analysis

To initiate the Factor Analysis (FA), we conducted preliminary tests to assess the suitability of the data. The Bartlett test, used to determine whether the variables in the dataset are sufficiently correlated, yielded a p-value of $2.23\times10-162.23\times10-16$, indicating statistically significant correlations among the variables. Additionally, the Kaiser-Meyer-Olkin (KMO) test produced a value of 0.79, exceeding the acceptable threshold of 0.6. These results confirm that the data is appropriate for Factor Analysis.

The analysis initially retained four factors (MR1, MR2, MR3, and MR4); however, based on the SS loadings, only three factors—MR1 (2.42), MR3 (1.91), and MR2 (1.07)—are considered relevant, as they collectively account for 67% of the cumulative variance. MR1 is strongly associated with *marketCap* (0.86), *totalDebt* (0.72), and *totalRevenue* (0.66), capturing the financial size of the company. MR3 is primarily defined by *totalCashPerShare* (0.84) and *totalCash_shifted* (0.81), representing liquidity or cash reserves. Finally, MR2 is dominated by *profitMargin* (0.90), which is the only strongly associated variable representing profitability.

The results of the FA closely align with insights from the Principal Component Analysis (PCA), providing a consistent and complementary perspective on the structure of the dataset. Both analyses reveal critical dimensions that summarize relationships among the variables and highlight distinct aspects of company performance. MR1 corresponds to PCA's Dimension 1 (Company size and financial scale), emphasizing variables related to size and scale. Similarly, MR2 (Profitability) and MR3 (Liquidity) relate conceptually to PCA's Dimension 2 (Profitability and shareholder efficiency), though FA isolates these dimensions more distinctly.

In conclusion, the results of FA and PCA converge to provide a robust and coherent characterization of the dataset. Both methods underscore financial size, liquidity, and profitability as critical dimensions for analyzing company performance.

Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) aims to reduce dimensionality while preserving as much of the inputted distance between observations, and thus the information, as possible. This is achieved by transforming the original variables into a new set of uncorrelated dimensions. For this process, Euclidean distance will be used for numerical data, and Gower distance will be used for mixed data.

After applying multidimensional scaling using both Euclidean and Gower distances, it becomes apparent that the Euclidean distance explains 87.62% of the variation in two dimensions, compared to 31% for Gower. Despite this, the results are largely comparable. The key distinction is that Gower's approach identifies three clusters instead of two, allowing for more detailed analysis. The clusters in Figure A.25 have been further profiled using statistical tests, including Chi-squared and ANOVA, revealing significant associations with almost all variables, especially recommendationKey.

The green cluster, located in the top-left quadrant, includes prominent companies such as Apple, Nvidia, and Bank of NY. These companies exhibit high values across all numerical metrics, including marketCap, enterpriseValue, totalRevenue, totalCashPerShare, totalCash, and totalDebt. The cluster is dominated by the recommendation "buy," with 98% of all buy recommendations concentrated here. This group also features a strong representation of the Technology, Financial Services, and Healthcare sectors, reflecting a mix of well-established, high-performing companie**s** and promising startups. However, as one move along the first dimension within this cluster, entities like SiNtx Technologies demonstrate decreasing metrics, such as *marketCap* and *totalRevenue*, marking a gradient of performance.

In contrast, the red cluster, situated in the lower portions of the scaling, features companies such as Morgan Stanley, ARMOUR Residential REIT, and Graham Holdings. These entities exhibit negative or very low *enterpriseValue*, *totalRevenue*, and *profitMargins*. Morgan Stanley and ARMOUR are marked by the recommendation "hold," while Graham Holdings reflects "underperform" with minimal revenue growth and high *totalDebt*. The cluster overall has below-average values in metrics like enterpriseValue, totalRevenue, and MarketCap, accompanied by a skewed overrepresentation of recommendations such as "hold," "underperform," and "strong_buy." This group spans sectors including Consumer Defensive, Healthcare, and Technology and likely represents struggling or underperforming companies. If we were to compare a company like ARMOUR to NexImmune, which has similar dimension one values but opposite dimension two values, one would see that the *totalRevenue* is way higher on the second company.

The blue cluster, located in the upper-right quadrant, includes companies like NexImmune Inc. and REVIUM Recovery Inc. This group is characterized by uniformly low values for all numerical metrics, including *marketCap*, *enterpriseValue*, and *totalRevenue*. It has an overrepresentation of the Financial Services sector and a stark underrepresentation of the "buy" or "hold" recommendations. Instead, the predominant recommendation is "none," signaling uncertainty and lack of confidence**.** This cluster likely represents smaller companies with poor performance metrics and limited market visibility. The more we move to the upper right (increment both dimensions) comparing GRAIL Inc and REVIUM RECOVERY, for instance, the more an additive effect it's seen on variables such as *enterpriseValue*, *profitMargins*, and *revenueGrowth*. However, a subtractive effect is observed on *marketCap*, *totalCash*, *totalCashPerShare*, *totalDebt*, and *totalRevenue*.

In summary, the dimensions derived from multidimensional scaling are less intuitive than those derived from methods like PCA or MCA. However, the clustering effect achieved through MDS reduced distances effectively distinguishes groups based on recommendations with remarkable precision.

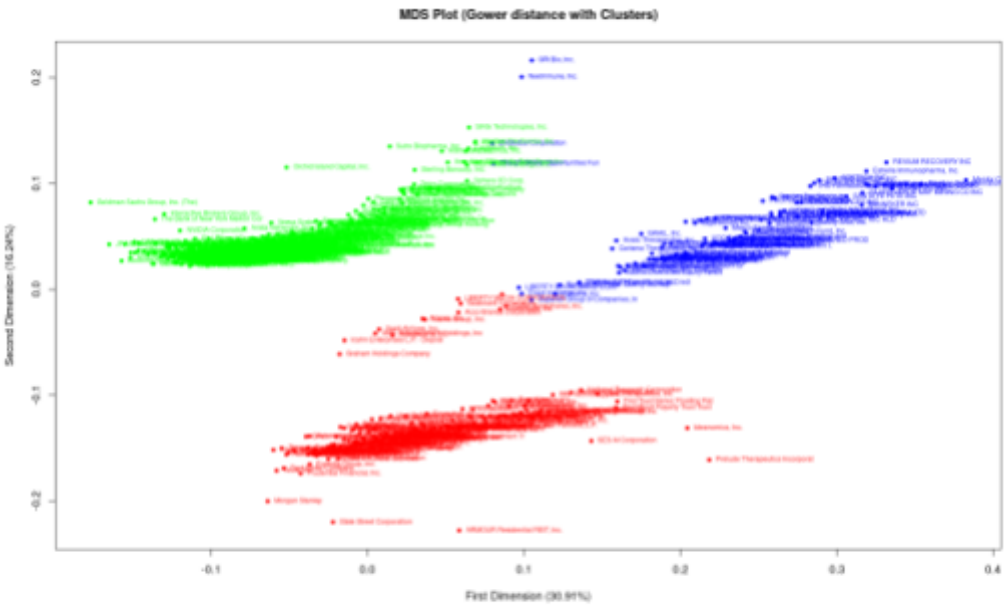

Figure 3. MDS two first dimensions plots

Correspondence Analysis (CA)

The goal of Correspondence Analysis (CA) is to reduce dimensionality by analyzing the relationships between two categorical variables: recommendationKey and Sector. Hence, the relationship between the different categories of these variables can be jointly explained.

Correspondence Analysis (CA) is preferred over Multiple Correspondence Analysis (MCA) when the state is included as a third variable. MCA with state introduces several issues: the first two dimensions explain less than 5% of the total inertia, the state variable adds considerable noise, and its contribution to the analysis is minimal. As a result, the explanatory power is significantly reduced, and most states are clustered near the origin, making the results difficult to interpret. Therefore, CA is better suited for this analysis [3]. The eigenvalue of the new dimension axis has been used to determine the number of dimensions to keep, which is two and explains over 85% of the variance among these factors (See Figure A.25). Glancing at the average eigenvalue, just one dimension should be picked. Still, a second dimension will also be used to ensure a more robust analysis.

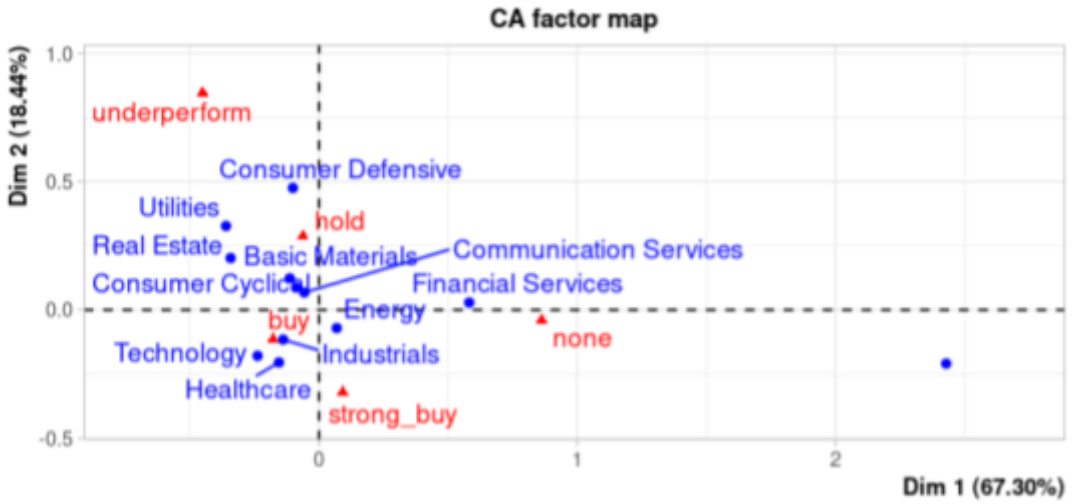

Figure 4. Correspondence Analysis factor maps

Dimension 2 shows a strong relationship with recommendations: lower values are associated with "strong_buy," while higher values align with "underperform." Dimension 1, in conjunction with Dimension 2, plays a significant role in separating or clustering sectors. For instance, as seen in the plot, the healthcare and technology industries are closely linked to the "buy" recommendation and are somewhat near "strong_buy." Conversely, the financial services sector is positioned closer to "none."

The Consumer Defensive sector stands out as it is positioned near both "underperform" and "hold" recommendations, suggesting a stronger association with these categories. On the other hand, sectors such as Energy, Communication Systems, and Consumer Cyclical are positioned close to the origin. Their proximity to the origin indicates low variation, making them harder to characterize. This is because these

sectors have a relatively even distribution of recommendations across categories or are very close to the mean row profiles.

Simplifying these results, sectors near the origin provide little actionable information regarding whether to buy, hold, or sell. Sectors higher on the plot tend to perform worse and align with a "hold" recommendation. In contrast, sectors lower on the plot correspond to a higher likelihood of being recommended as a "buy." The horizontal axis also reflects a gradient of certainty, with sectors on the left associated with more decisive recommendations and sectors on the right with greater uncertainty, often aligning with the "none" category.

Cluster Analysis Profiling

This section's objective is to detect, for each cluster generated from the application of hierarchical clustering to the PCA results, which modalities of the categorical explanatory variables and which continuous explanatory variables deviate significantly from the overall values, thus characterizing the enterprises in the clusters.

As already stated in the PCA section, the dimensions can be summarized as the Company size and financial scale (X-axis) and Profitability and shareholder efficiency (Y-axis). Note that the factor map and the dendrogram in Figures 5 and 6 show a central separation of the enterprises into two main clusters, which contribute an inertia gain of 3.09 and 0.9, respectively. The criteria for choosing the number of clusters are based on the inertia gain and the proposed number of clusters in the HCPC method [2].

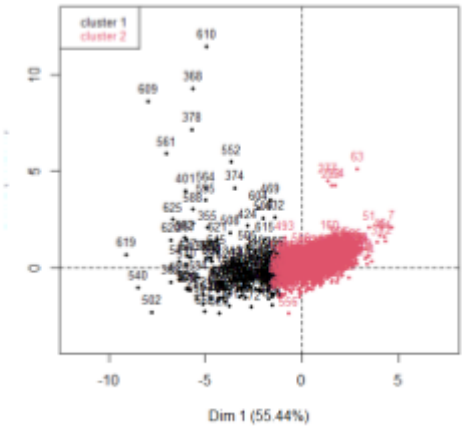

Figure 5: Factor map of the clusters.

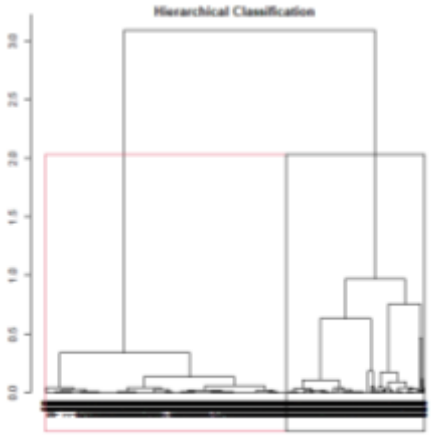

Figure 6: Hierarchical classification dendrogram

Regarding the relation between the categorical variables and the clusters, the p-values from the chi-squared tests indicate a significant relationship at a 95% confidence level in all of them: *recommendationKey* (p = 7.66e-64), *sector* (p = 7.65e-05), and *state* (p = 5.40e-03). Among these, the relationship is particularly strong for the *recommendationKey* variable. Given by z-tests for proportions, Cluster 1 is primarily characterized by a high prevalence of the recommendationKey=none category, the sector Financial Services, and certain states such as Maryland and Florida, with significant underrepresentation of states like Washington and Georgia,

the technology sector, and the hold and buy recommendation keys. In contrast, Cluster 2 is marked by an overrepresentation of recommendationKey=buy, sector_Technology, and states such as Minnesota, Illinois, Georgia, and Washington. Additionally, states like Florida and Maryland are underrepresented in Cluster 2. All the chi-square p-values are highly significant, confirming the strong relationship between the clusters and the categorical variables, with the v-test values further reinforcing the strength of these associations.

Regarding the relationship between the quantitative variables and the clusters, the ANOVA and t-test-based analysis show significant associations at a 95% confidence level for all variables, with high eta-squared values and extremely low p-values. Variables such as *marketCap* (p = 8.15e-85), *totalDebt* (p = 1.15e-82), *totalCash* (p = 1.22e-76), and *totalRevenue* (p = 1.56e-80) demonstrate particularly strong relationships with the clusters.

Cluster 1 is characterized by significantly lower values for variables like *marketCap*, *totalDebt*, *totalRevenue*, *totalCash*, *totalCashPerShare*, and *enterpriseValue*, indicating smaller companies with weaker financial metrics. Notably, *profitMargins* (p = 2.23e-14) and *revenueGrowth* (p = 2.53e-08) are higher in Cluster 1, suggesting that while these companies are smaller, they tend to have stronger profitability and growth rates.

In contrast, Cluster 2 consists of larger, financially stronger companies with significantly higher values for these metrics. Key variables such as *marketCap*, *totalDebt*, *totalRevenue*, *totalCash*, and *enterpriseValue* show much higher values in Cluster 2, with very low p-values (all < 1e-37), indicating a strong relationship with cluster membership. These companies are more capitalized, have more debt, generate more revenue, and hold larger cash reserves.

Overall, Cluster 1 consists of smaller companies with lower financial metrics, including marketCap, totalDebt, totalCash, totalRevenue, and enterpriseValue, but with higher profitMargins and revenueGrowth, indicating stronger profitability and growth despite weaker financial strength. It is predominantly characterized by sectors like Financial Services and recommendationKey=none, with an overrepresentation of companies from states like Maryland and Florida. In contrast, Cluster 2 includes larger, more financially robust companies with significantly higher values for marketCap, totalDebt, totalCash, totalRevenue, and enterpriseValue. It is marked by sectors such as Technology and recommendationKey=buy, with overrepresentation in states like Minnesota, Illinois, and Georgia.

Discriminant Analysis

The objective of this section is to use the discriminant analysis technique to construct a model which classifies the observations to any of the “buy”, “hold”, or “none” categories of the *recomentadionKey* variable given a set of explanatory numerical variables. The original variable contains “strong_buy” and “underperform” labels, which we remove since there are only 14 observations that provide noise to the model.

Given the lack of homoscedasticity in our data, we will use Quadratic Discriminant Analysis (QDA), which is more flexible than Linear Discriminant Analysis (LDA) in the sense that it does not assume the equality of variance/covariance. In other words, for QDA, the covariance matrix can be different for each class [1].

Moreover, QDA is recommended if the training set is extensive so that the variance of the classifier is not a significant issue or if the assumption of a common covariance matrix for the K classes is clearly untenable [4].

As stated in the Variable Transformation section, only *totalCash* and *totalCashPerShare* variables follow a normal distribution. Therefore, incorporating additional variables may reduce the model's effectiveness or accuracy, as Quadratic Discriminant Analysis (QDA) presumes that predictor variables within each class follow a multivariate normal distribution.

After dividing the data into training (80%) and test sets, followed by standardization, we observed notable outcomes during the variable selection process. Using the stepwise "backward" selection method, which started with all numerical variables, the final model retained only the totalCash and totalCashPerShare variables. This result is particularly interesting, as these are the only variables that conform to a normal distribution, which is a key assumption for QDA. The model, based on these selected variables, achieved a classification accuracy of 0.70 on the training set and 0.71 on the test set.

Alternatively, assuming homoscedasticity for totalCash and totalCashPerShare (see Figures 7 and Figure 8), we applied LDA and obtained a model with a classification accuracy of 0.72 on the test set.

Figure 7: *totalCash* Variance among *recommendationKey* categories

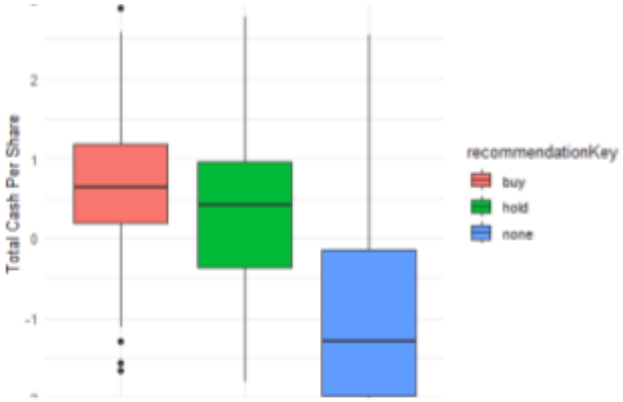

Figure 8: *totalCashPerShare* Variance among *recommendationKey* categories

However, when we looked deeper at the models, we noticed that despite high classification accuracy for the "buy" and "none" categories in both the QDA and LDA models, the "hold" category was entirely misclassified, indicating that the models failed to adequately capture the relationship between the explanatory variables and this category.

To improve the model, we balanced the dataset to ensure equal representation of each category (since the "buy" category was previously overrepresented). Following this adjustment, the stepwise backward selection identified additional variables (*marketCap*, *enterpriseValue*, *profitMargins*, *totalDebt*, and *totalRevenue*) that may introduce noise due to non-normal distributions. The resulting QDA model showed a 70% classification accuracy and improved the identification of "hold" observations (buy: 0.94, hold: 0.38, none: 0.78).

The discriminant's importance in this model was tested using a Q-statistic, yielding Qstat = 1734. With a critical Chi-squared value of 9.21 (for k-1 degrees of freedom), we reject the null hypothesis, concluding that the discriminant function is significant and not based on random classification.

Despite not fully adhering to its assumptions, the LDA model under this framework produced a classification accuracy of 0.66 on the test set, offering balanced accuracy across all categories: buy (61%), hold (50%), and none (89%). This model is arguably the best, as it classifies each category with more than 50% accuracy, outperforming random chance. The Q-statistic for this model was Qstat = 1536. With the same critical value of 9.21, we reject the null hypothesis, confirming that the discriminant function is meaningful and not classifying by chance.

In conclusion, we have seen how modeling financial data, particularly for enterprises, presents significant challenges due to the sector's complexity and variability. The discriminant analysis results demonstrate that techniques like QDA and LDA can struggle to accurately classify categories, such as the "hold" category, which highlights the difficulty of capturing the full scope of the market.

4. Conclusions

This project analyzed the U.S.-based companies using data scraped by the authors from Yahoo Finance. It focused on extracting meaningful insights from a multivariate dataset of 666 companies. Despite significant challenges, the findings provide valuable perspectives for understanding company performance.

The dataset revealed significant imbalances in the categorical variable, recommendationKey, with "Buy" overrepresented. This imbalance, coupled with the lack of normality in most variables, posed challenges for analytical techniques. However, such distributional characteristics are typical for financial data, where skewed metrics often reflect the inherent volatility and variability of the stock market.

Through Principal Component Analysis (PCA) and Factor Analysis (FA), two key dimensions emerged:

1. Financial Size—capturing metrics like market capitalization, revenue, and debt, representing a company's scale.
2. Profitability and Shareholder Efficiency—reflecting profitability margins and cash efficiency relative to growth strategies.

These dimensions provided a robust framework for distinguishing companies based on size and profitability.

The cluster analysis highlights significant differences in company size, financial strength, profitability, and geographic distribution. Smaller companies (Cluster 1) offer opportunities to leverage higher profitability and growth, while larger firms (Cluster 2) benefit from robust financial metrics and market presence.

Multidimensional Scaling (MDS) highlighted the complexity of defining inter-company distances due to the mixed nature of the data and the limitations of Gower distance. While MDS identified clusters that corresponded to performance and recommendations, deeper business knowledge is necessary for even more precise interpretations.

Correspondence Analysis was used to analyze the relationship between recommendationKey and Sector, providing clearer insights than MCA when state is included. The analysis identifies patterns where some sectors align closely with specific recommendations while others near the origin show low variability and limited insight. The vertical axis reflects performance trends, and the horizontal axis captures a certainty gradient in recommendations.

Discriminant Analysis (DA) illustrated the difficulty of fitting effective classification models due to the complexity of financial data. Even though the best variables for the DA were chosen by using stepwise selection (instead of MANOVA), both Linear (LDA) and Quadratic Discriminant Analysis (QDA) struggled to classify underrepresented categories like "Hold". The application of more complex machine learning models could better handle such challenges, capturing more complex relationships in the data. More data and more balance in the categories of the target variable would also help DA work better.

In summary, this analysis highlighted the inherent complexities of financial datasets, including intricate inter-variable relationships and the challenges associated with model fitting. Traditional multivariate

techniques have provided valuable foundational insights, paving the way for future investigations to explore advanced modeling approaches.

5. Bibliography

[1] "Discriminant Analysis Essentials in R - Articles - STHDA." *Sthda.com*, 11 Mar. 2018, www.sthda.com/english/articles/36-classification-methods-essentials/146-discriminant-analysis-essentials-in-r/. Accessed 23 Dec. 2024.

[2] Husson F. "Clustering with FactoMineR." *YouTube*, 18 Apr. 2013, www.youtube.com/watch?v=4XrgWmN9erg&list=PLnZgp6epRBbTsZEFXi_p6W48HhNyqwxIu&index=8. Accessed 23 Dec. 2024.

[3] Husson, F. (n.d.). *Multivariate Data Analysis*. Retrieved [22/12/2024], from https://francoishusson.wordpress.com/

[4] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated.

Appendix

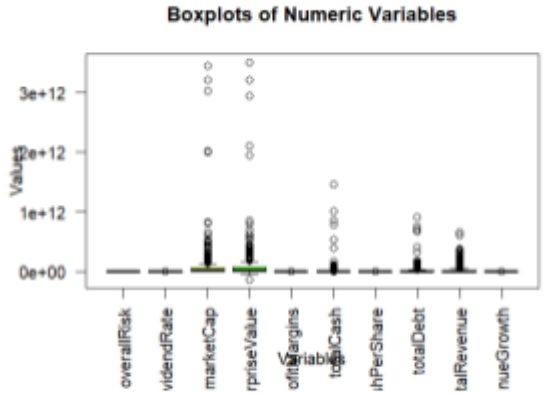

Figure A.1: Numerical variables boxplots before transformations

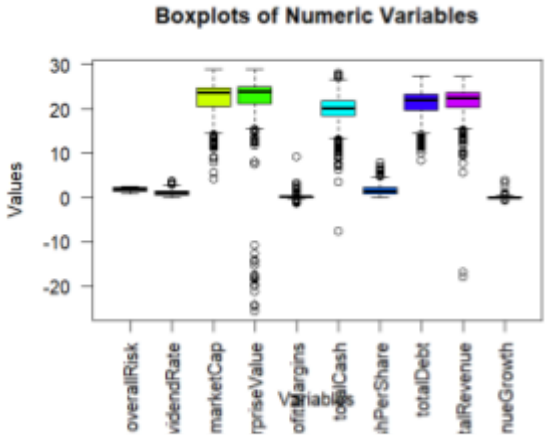

Figure A.2: Numerical variables boxplots after transformations

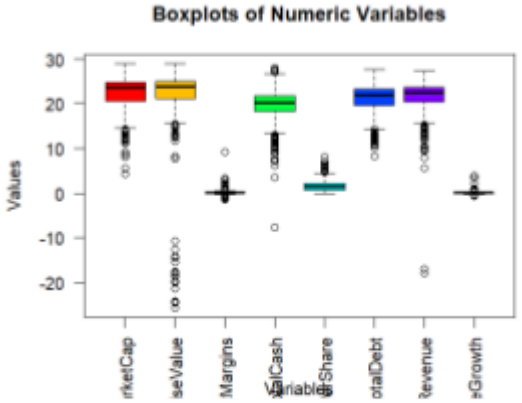

Figure A.3: Numerical variables boxplots before imputation

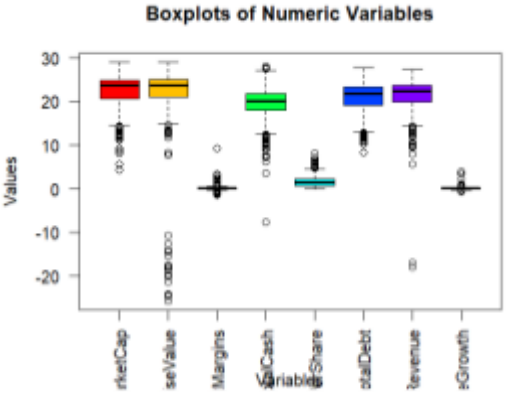

Figure A.4: Numerical variables boxplots after imputation

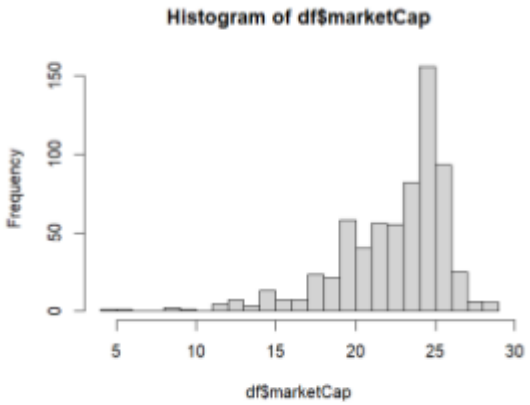

Figure A.5: Histogram of marketCap variable

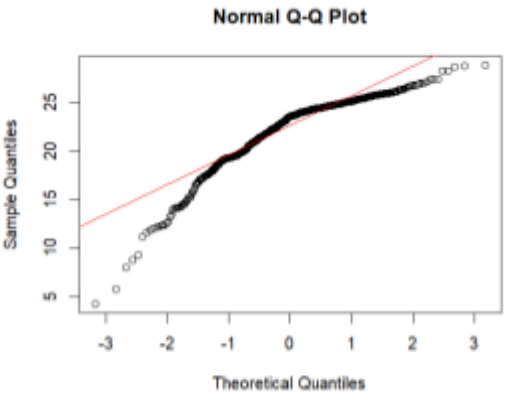

Figure A.6: marketCap Q-Q plot

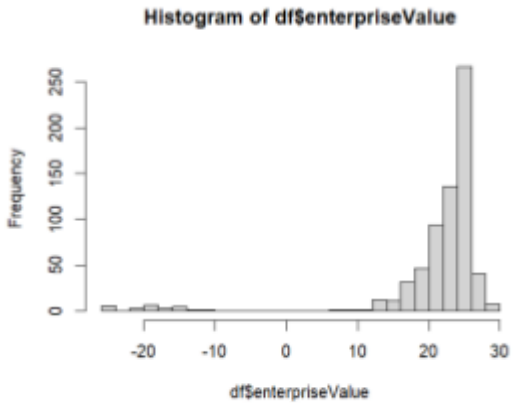

Figure A.7: Histogram of enterpriseValue variable

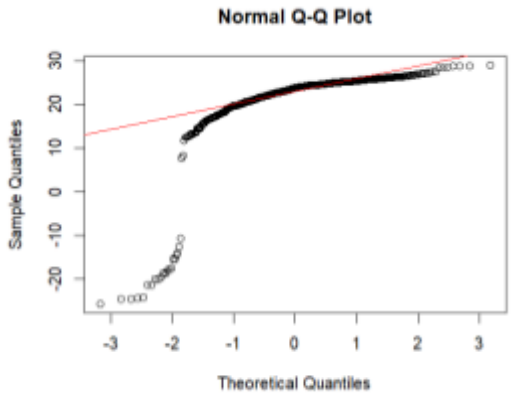

Figure A.8: enterpriseValue Q-Q plot

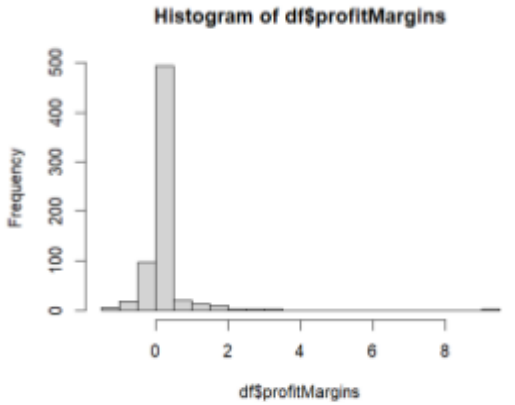

Figure A.9: Histogram of profitMargins variable

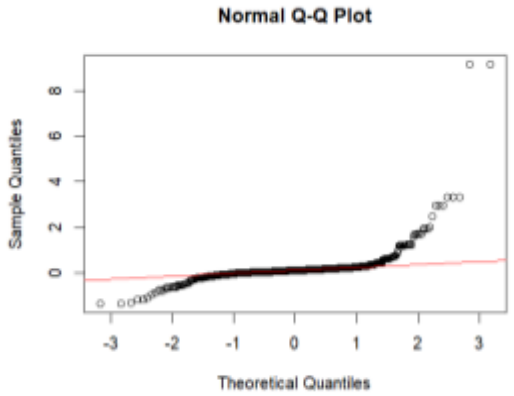

Figure A.10: profitMargins Q-Q plot

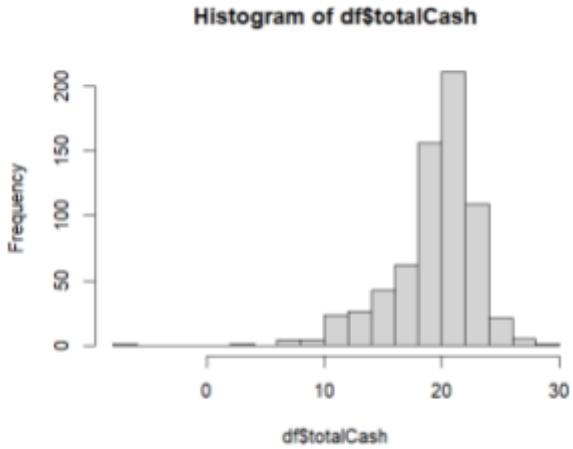

Figure A.11: Histogram of totalCash variable

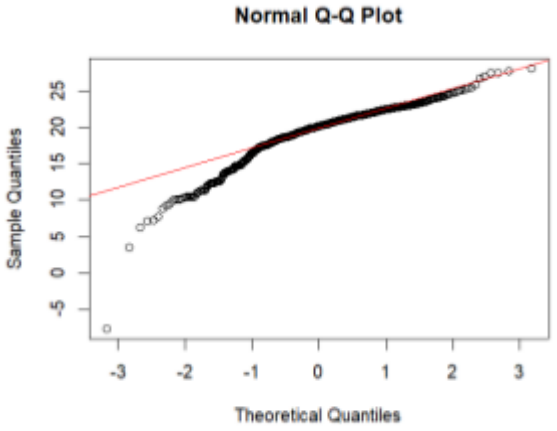

Figure A.12: totalCash Q-Q plot

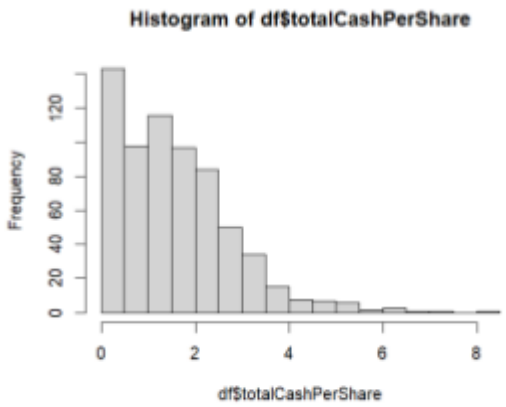

Figure A.13: Histogram of totalCashPerShare variable

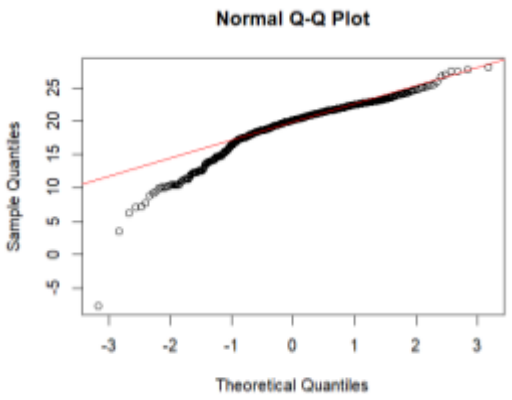

Figure A.14: totalCashPerShare Q-Q plot

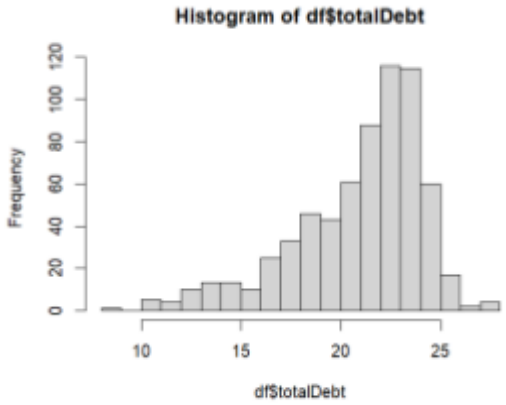

Figure A.15: Histogram of totalDebt variable

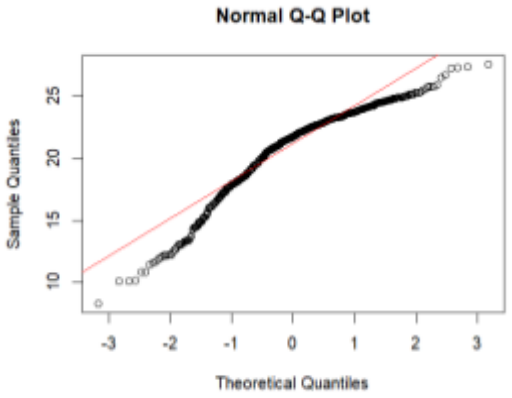

Figure A.16: totalDebt Q-Q plot

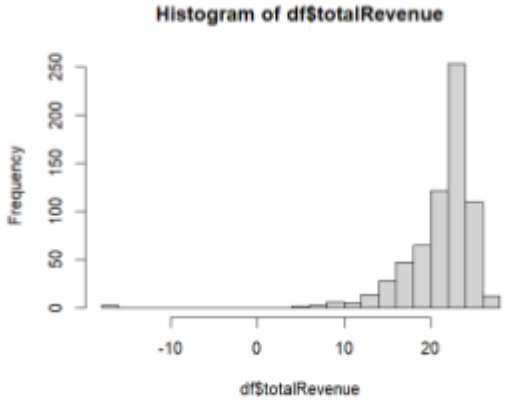

Figure A.17: Histogram of totalRevenue variable

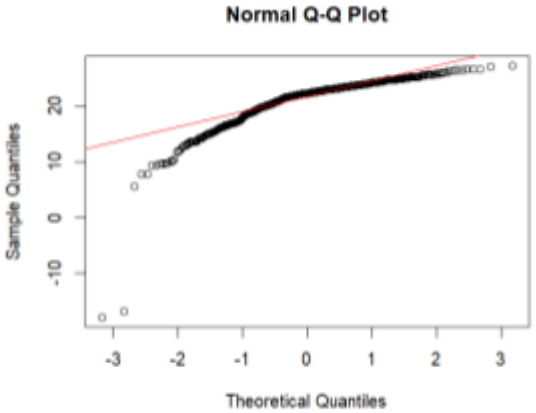

Figure A.18: totalRevenue Q-Q plot

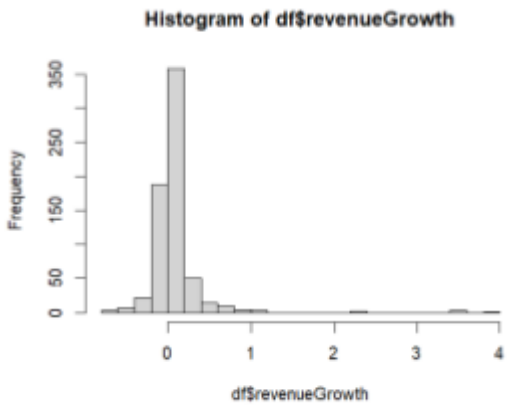

Figure A.19: Histogram of revenueGrowth variable

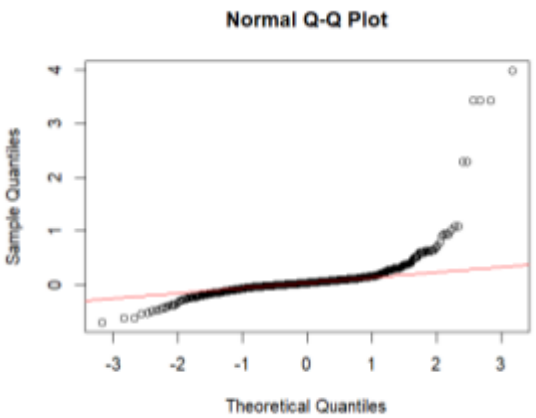

Figure A.20: revenueGrowth Q-Q plot

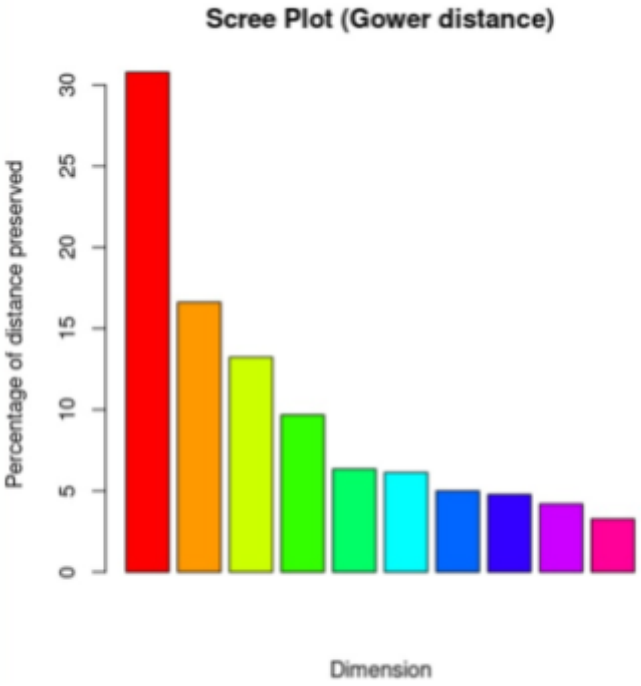

Figure A.21: Distance preserved using Gower distance, MDS

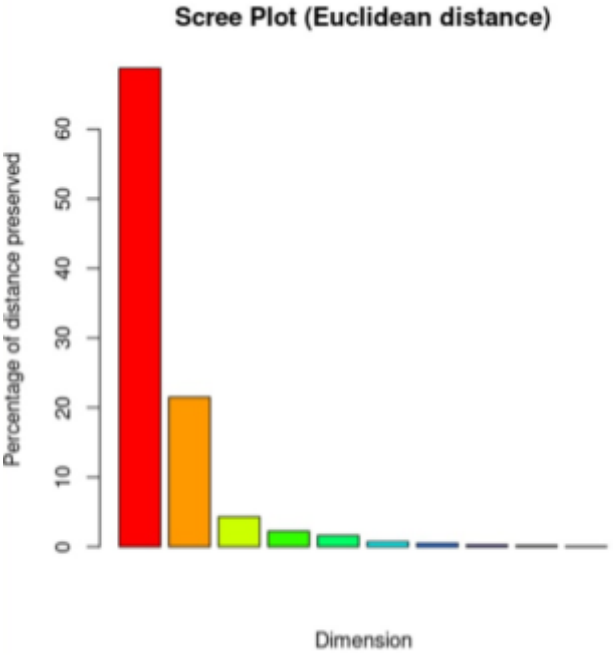

Figure A.22: Distance preserved using Euclidean distance, MDS

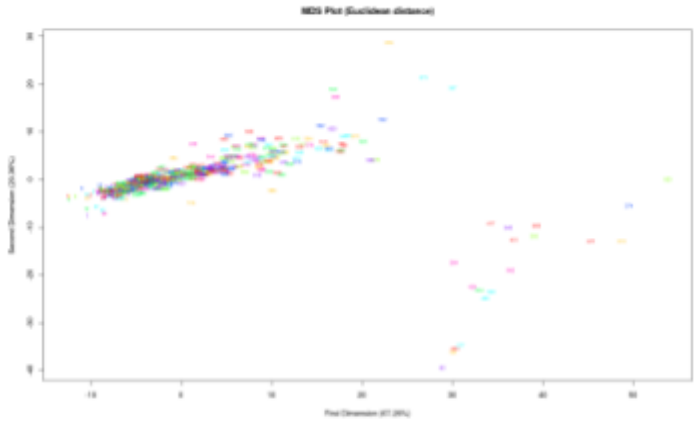

Figure A.23: Results of Euclidean distance based MDS

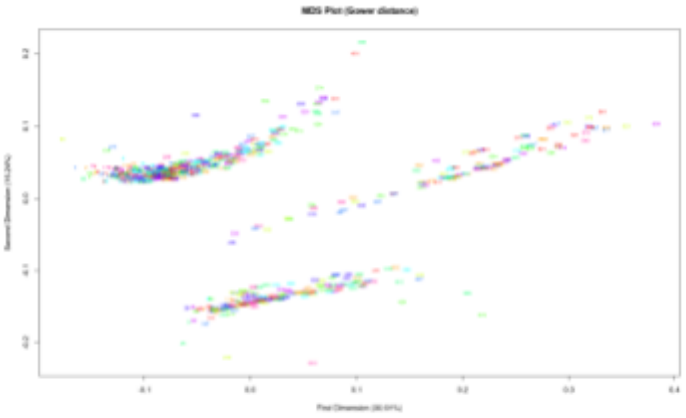

Figure A.24: Results of Gower distance based MDS

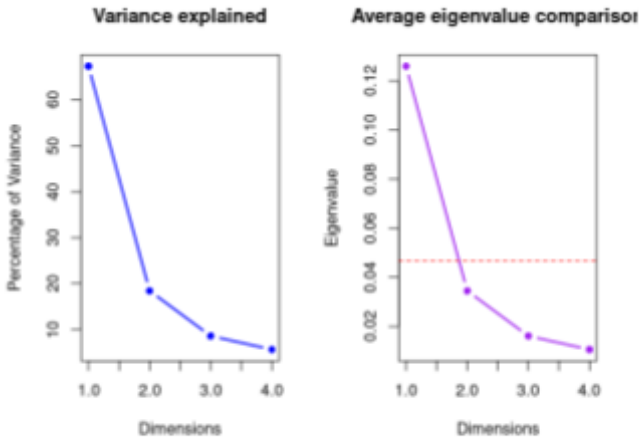

Figure A.25: Variance explained and eigenvalue per dimension in CA

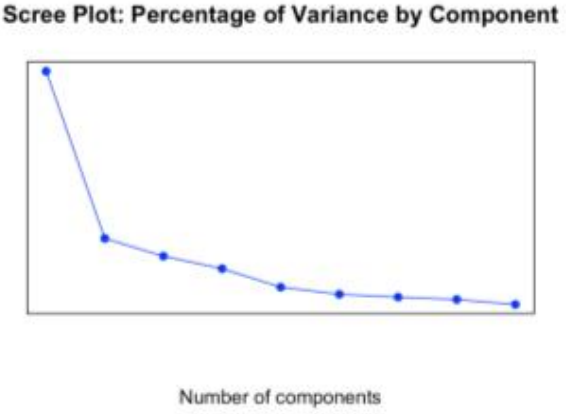

Figure A.26: Scree Plot

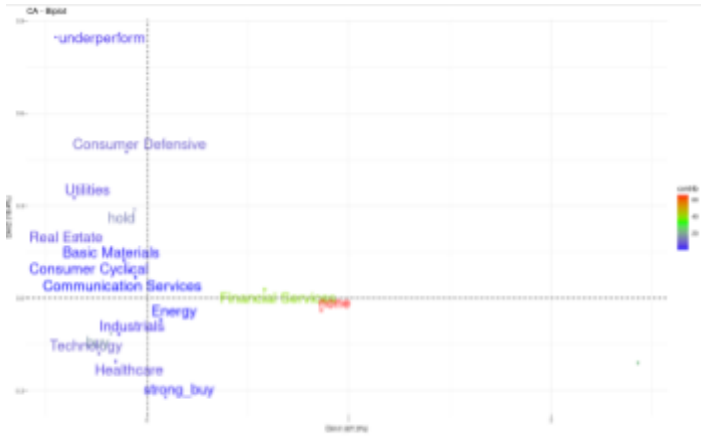

Figure A.27: Correspondence analysis contributions biplot