UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

CAMILA CALEONES DE FIGUEIREDO

ANALYZING STOCK MARKET DATA USING UNSUPERVISED LEARNING TECHNIQUES

RIO DE JANEIRO
2024

CAMILA CALEONES DE FIGUEIREDO

ANALYZING STOCK MARKET DATA USING UNSUPERVISED LEARNING TECHNIQUES

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Profa. Giseli Rabello Lopes
Coorientadora: Profa. Juliana Vianna Valério

RIO DE JANEIRO
2024

CIP - Catalogação na Publicação

F475a

Figueiredo, Camila Caleones de
Analyzing stock market data using unsupervised learning techniques / Camila Caleones de Figueiredo. -- Rio de Janeiro, 2025.
89 f.

Orientadora: Giseli Rabello Lopes.
Coorientador: Juliana Vianna Valério.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Computação, Bacharel em Ciência da Computação, 2025.

1. Inteligência artificial. 2. Mercado de ações. 3. Aprendizado de máquina. 4. Aprendizado não supervisionado. 5. Clusterização hierárquica. I. Lopes, Giseli Rabello, orient. II. Valério, Juliana Vianna, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

CAMILA CALEONES DE FIGUEIREDO

ANALYZING STOCK MARKET DATA USING UNSUPERVISED LEARNING TECHNIQUES

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 14 de Fevereiro de 2025

BANCA EXAMINADORA:

gov.br
Documento assinado digitalmente
GISELI RABELLO LOPES
Data: 29/05/2025 18:46:22-0300
Verifique em https://validar.iti.gov.br

Giseli Rabello Lopes
D.Sc. (Instituto de Computação - UFRJ)

gov.br
Documento assinado digitalmente
JULIANA VIANNA VALERIO
Data: 02/06/2025 17:12:52-0300
Verifique em https://validar.iti.gov.br

Juliana Vianna Valério
D.Sc. (Instituto de Computação - UFRJ)

gov.br
Documento assinado digitalmente
ANTONIO REVAIL ALVES PEREIRA
Data: 03/06/2025 12:23:14-0300
Verifique em https://validar.iti.gov.br

Antonio Revail Alves Pereira
M.Sc. (Instituto de Computação - UFRJ)

gov.br
Documento assinado digitalmente
JOAO CARLOS PEREIRA DA SILVA
Data: 03/06/2025 19:19:15-0300
Verifique em https://validar.iti.gov.br

João Carlos Pereira da Silva
D.Sc. (Instituto de Computação - UFRJ)

To Natália, Liliane, Eduardo Júnior, Lezi, Ivone, Eduardo and the whole family, whose unwavering support has been with me every step of the way during this important stage of my life.

To my grandfather, José Nascimento (in memoriam), who was always there and guide me along the way.

To my partner, Luan, and to all my friends who shared this journey with me, thank you for being by my side.

ACKNOWLEDGEMENTS

I would like to thank all the people who helped me, not only with this project, but also the ones I met during my graduation. Special thanks to Giseli Rabello and Juliana Valério, who embraced this research with me and for every meeting, correction, and time invested. Thanks to João Paixão and Laura de Oliveira for the support when I started this project and for being the foundation of this research.

I would also like to express my gratitude to my guidance counselor, Maria Luiza Campos, for guiding me through this journey. Additionally, thanks to all the teachers with whom I had the chance to share these experiences.

Finally, thanks to all my friends who helped me, making this time of my life happier. Thanks for every lunch, games, jokes, and for always being there for me.

“Tudo ia se ajeitar, o tempo nunca falha em suas habilidades.”

Carla Madeira

RESUMO

Compreender o mercado financeiro pode ser desafiador, especialmente para investidores que precisam lidar com uma grande quantidade de informações, termos técnicos e variações nos preços dos ativos. Este projeto busca facilitar esse processo ao analisar como diferentes indicadores financeiros – como volatilidade, capitalização de mercado e rendimento de dividendos – podem ajudar na categorização de empresas. Utilizando técnicas como a Análise de Componentes Principais (PCA) para redução de dimensionalidade e métodos de clusterização, como o K-Means, o estudo identifica padrões e semelhanças entre os ativos. Isso permite agrupar empresas com características financeiras semelhantes, tornando a avaliação de risco e a busca por oportunidades de investimento mais intuitivas. Por exemplo, é possível identificar clusters de empresas mais voláteis e com menor retorno, que podem representar investimentos de maior risco, assim como empresas com retornos mais estáveis e menor volatilidade, que podem interessar a perfis mais conservadores. A PCA contribui para a simplificação da análise ao destacar os fatores mais relevantes, ajudando a filtrar ruídos e priorizar as informações mais úteis para a tomada de decisão. Embora essas ferramentas não ofereçam garantias sobre o desempenho futuro dos investimentos, elas ajudam a estruturar os dados de forma mais clara, permitindo que investidores tomem decisões mais informadas. O objetivo principal do projeto é tornar a análise do mercado mais acessível, principalmente para iniciantes, ao oferecer uma visão mais organizada e visual dos dados. Dessa forma, tanto investidores experientes quanto novos participantes do mercado podem usar essas informações para entender melhor os riscos, identificar oportunidades e desenvolver estratégias com mais confiança.

Palavras-chave: inteligência artificial; mercado de ações; aprendizado de máquina; aprendizado não supervisionado; k-means; modelo de mistura gaussiana; clusterização hierárquica.

ABSTRACT

Understanding the financial market can be challenging, especially for investors who must navigate vast amounts of information, technical jargon, and fluctuations in asset prices. This project aims to simplify this process by analyzing how different financial indicators—such as volatility, market capitalization, and dividend yield—can assist in categorizing companies. By applying techniques like Principal Component Analysis (PCA) for dimensionality reduction and clustering methods such as K-Means, the study identifies patterns and similarities among assets. This allows companies with similar financial characteristics to be grouped together, making risk assessment and investment opportunity identification more intuitive. For example, it becomes possible to identify clusters of highly volatile companies with lower returns, which may represent higher-risk investments, as well as companies with stable returns and lower volatility, which might appeal to more conservative investors. PCA helps simplify the analysis by highlighting the most relevant factors, filtering out noise, and prioritizing useful information for decision-making. While these tools do not guarantee future investment performance, they help structure data more clearly, enabling investors to make more informed decisions. The primary goal of this project is to make market analysis more accessible, particularly for beginners, by providing a more structured and visual representation of financial data. In this way, both experienced investors and newcomers can leverage these information to better understand risks, identify opportunities, and develop strategies with greater confidence.

Keywords: artificial intelligence; stock market; machine learning; unsupervised learning; k-means; gaussian mixture model; hierarchical clustering.

LIST OF FIGURES

Figure 1 – Elbow Curve 31
Figure 2 – Dendrogram of the Hierarchical Model 31
Figure 3 – Boxplot of Clusters 32
Figure 4 – Correlation Matrix 39
Figure 5 – PCA Variance 40
Figure 6 – Elbow Curve for all dataframes 43
Figure 7 – Base evaluation indicators 44
Figure 8 – PCA Base Assessment Indicators 45
Figure 9 – Dendrogram of Hierarchical clustering 47
Figure 10 – Number of Tickers per cluster by clustering Method 48
Figure 11 – Dispersion for Each Pair of K-Means Attributes 51
Figure 12 – Distribution of Sectors by cluster 55
Figure 13 – Dispersion for Each Pair of GMM Attributes 84

LIST OF TABLES

Table 1 – Clustering Performance Metrics for Different Algorithms 49
Table 2 – Classification Intervals for Attributes 52
Table 3 – cluster Characteristics Based on K-Means 52
Table 4 – Company Name with associated Ticker 63
Table 5 – Cluster Characteristics Based on Hierarchical Clustering 83
Table 6 – Cluster Characteristics Based on GMM 85
Table 7 – Clusters with associated Tickers by K-Means 87
Table 8 – Clusters with associated Tickers by Hierarchical 88
Table 9 – Clusters with associated Tickers by GMM 89

LIST OF ABBREVIATIONS AND ACRONYMS

GMM	Gaussian Mixture Model
PCA	Principal Component Analysis
PC	Principal Component
SSE	Sum of Squared Errors
P/E Ratio	Price-to-earning Ratio
EPS	Earnings per Share
S&P 500	Standard & Poor's 500
NYSE	New York Stock Exchange

CONTENTS

1 INTRODUCTION . . . 13

2 FUNDAMENTALS AND RELATED WORKS . . . 15
2.1 FINANCIAL MARKET . . . **15**
2.2 MACHINE LEARNING . . . **15**
2.3 EVALUATION METRICS . . . **21**
2.3.1 Inertia . . . 21
2.3.2 Silhouette Score . . . 21
2.3.3 Calinski-Harabasz Index . . . 22
2.4 PRINCIPAL COMPONENT ANALYSIS . . . **23**
2.5 RELATED WORKS . . . **25**

3 INITIAL INVESTIGATION . . . 27
3.1 DATA COLLECTION, PREPROCESSING AND DIMENSIONALITY REDUCTION FOR CLUSTERING . . . **29**
3.2 COMPANY CLUSTERING . . . **30**
3.3 PATTERN IDENTIFICATION AND DISCUSSION OF RESULTS . . . **32**

4 DATASET . . . 35
4.1 CREATION OF THE UTILIZED DATASET . . . **38**
4.2 ANALYSIS OF ATTRIBUTES . . . **38**
4.3 DIMENSIONALITY REDUCTION . . . **40**

5 CLASSIFICATION . . . 42
5.1 ANALYSIS OF THE CLASSIFIED DATASETS . . . **43**
5.2 APPLYING UNSUPERVISED LEARNING METHODS . . . **46**
5.3 CLUSTERING SELECTION: PROCEEDING WITH K-MEANS . . . **50**
5.4 INVESTMENT RECOMMENDATIONS BASED ON CLUSTERS . . . **53**
5.4.1 Conservative Profile . . . 53
5.4.2 Moderate Profile . . . 53
5.4.3 Risky Profile . . . 54
5.5 INVESTOR PROFILE MATCHING BASED ON CLUSTER CHARACTERISTICS . . . **54**
5.6 ANALYSIS OF FINANCIAL METRICS BY SECTOR . . . **55**

6 CONCLUSION . . . 57
6.1 CHALLENGES FACED . . . **57**

6.2 FUTURE WORK **58**

BIBLIOGRAPHY **61**

APPENDIX A – TABLE WITH COMPANY NAME WITH ASSOCIATED TICKER **63**

APPENDIX B – ANALYSIS OF TICKER CLASSIFICATION INTO CLUSTERS FOR EACH METHOD **83**

1 INTRODUCTION

In today's financial market, various tools and platforms help individuals of different backgrounds understand investing. However, gaining a clear understanding of investment strategies and market dynamics remains a challenge. Many people want to learn more about the stock market, not necessarily to invest immediately, but to better grasp how companies are evaluated and how financial decisions are made.

One key aspect of investing is aligning strategies with different risk profiles. Generally, there are three types of investors: conservative investors, who prioritize stability and avoid risk even if it means lower returns; moderate investors, who accept some level of risk while setting limits on potential losses; and aggressive investors, who are willing to take on higher risks in pursuit of greater returns. Understanding these profiles can help anyone interested in finance recognize how different investment strategies work and how risk tolerance shapes decision-making.

To provide investors with simplified information about the financial health of a company, data from Yahoo Finance (KUHN, 2024) was used. This data was used to cluster companies based on their volatility, beta - a measure of a stock's volatility in relation to the overall market -, price-to-earnings (P/E ratio), dividend yield, market cap, revenue growth, earnings-per-share (EPS) and debt-to-equity ratio, which will be explained in the next chapters.

This study considered using Principal Component Analysis (PCA) (MACKIEWICZ; RATAJCZAK, 1993) to reduce the dimensionality of the attributes. Furthermore, by applying clustering methods to PCA results, it is possible to provide investors with clusters of companies that exhibit similar behaviors. This clustering helps investors better understand the risk and potential return associated with companies within each identified group. For clustering, in this work, K-means (JIN; HAN, 2010), Hierarchical clustering (HALKIDI, 2009) and Gaussian mixture models (REYNOLDS, 2009) were used.

Therefore, this study seeks to explore the following research questions:

(i) Is it possible to group companies that exhibit similar behavior in the stock market?

(ii) Can clustering companies help investors understand patterns?

The remainder of this work is structured as follows: In Chapter 2, the fundamental principles and relevant previous work in this area were explored. Chapter 3 presents the previous work that led to this project. In Chapter 4, an analysis of the data set used in this study was provided, including the collection and pre-processing of Web data. Chapter 5 discusses clustering models and their application in the context of companies and discussed the results obtained, examining the findings in relation to the study objectives. Finally, Chapter 6 offers suggestions for future research that could expand and enhance the results presented.

2 FUNDAMENTALS AND RELATED WORKS

The purpose of this chapter is to present the theoretical foundation for the topics discussed in this work, providing essential background information for a better understanding of the concepts explored. Additionally, it will introduce related studies, highlighting previous research and methodologies that contribute to the development of this study.

2.1 FINANCIAL MARKET

The financial market is an environment composed of various institutions, such as banks, brokerage firms, companies and others. In this system, it is possible to carry out buying and selling operations of assets — goods, stocks, commodities, reserves, accounts, among others. Thus, in this exchange, institutions, investors, regulatory bodies and resource takers are involved, i.e., those who will both sell and buy (SANTANDER, 2024).

Investors buy and sell assets through brokerage firms, which are intermediaries in the market. They are typically used to give investors the ability to trade items on the stock exchange, but also offer additional services such as advisory and technical support. When referring to assets, it is a centralized market governed by regulatory bodies, but there are several brokerage firms interacting with the same market.

In this context, it is important to define the three investor profiles that will be used throughout the work (ANBIMA, 2023). The first is the conservative investor, who prefers to avoid significant risks and preserve the capital invested. Therefore, this profile typically opts for investments that offer stable returns, adopting a safer and longer-term strategy. The moderate investor, on the other hand, allocates part of their resources in slightly more volatile investments, with higher return potential in the medium and long term, understanding that this level of risk is moderate. Finally, the aggressive or risk-taking investor is willing to accept higher risks in search of higher returns, adjusting their strategy based on market fluctuations and believing that, in the long run, these changes will yield positive results, acknowledging that volatility is part of the investment strategy.

2.2 MACHINE LEARNING

The main topic of machine learning used in this work is clustering. The details of the techniques, implementations and choices will be explained throughout this work. However, in general terms, as the name suggests, it is the task of grouping data into clusters or groups, such that each item within the group is more similar to each other than to items in other groups.

These algorithms aim to identify patterns or underlying structures in the data without predefined labels, seeking to discover natural groups. Some of the methods used in this work will include: K-Means, Hierarchical clustering and Gaussian mixture model (GMM).

The first is K-Means (IBM, 2024), one of the most popular techniques. It is an iterative algorithm that seeks to minimize the sum of the distances between the data points and the centroids of the clusters, which in this work will use the Euclidean distance. Thus, the data point closest to a centroid will be grouped in the same cluster.

The centroid can be found using:

$$\mu_k = \left(\frac{X_1^{(1)} + X_2^{(1)} + \cdots + X_n^{(1)}}{n}, \frac{X_1^{(2)} + X_2^{(2)} + \cdots + X_n^{(2)}}{n}, \ldots, \frac{X_1^{(d)} + X_2^{(d)} + \cdots + X_n^{(d)}}{n} \right) \tag{2.1}$$

Where:

- μ_k is the centroid (mean point) of cluster k.
- $X_i^{(j)}$ represents the j-th coordinate of the i-th point in the cluster.
- d is the number of dimensions of the points.
- n is the number of points in the cluster.

and the distance will be calculated using the Euclidean distance in N dimensions:

$$d_{ij} = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2} \tag{2.2}$$

Where:

- d_{ij} represents the Euclidean distance between points i and j in N-dimensional space.
- x_{ik} is the k-th coordinate of point i.
- x_{jk} is the k-th coordinate of point j.
- N is the number of dimensions.

The first step is to determine an ideal number of clusters, k, which can be obtained using the Elbow Method. This method is a graphical way to find the optimal value for K. It is based on the observation that as the number of clusters increases, the internal variability of each group, measured by the SSE (Sum of Squared Errors), tends to decrease.

The SSE can be expressed as:

$$SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\| x_i^{(k)} - \mu_k \right\|^2 \tag{2.3}$$

where:

- $x_i^{(k)}$ represents the data points in cluster k.
- μ_k is the centroid (mean point) of cluster k.
- n_k is the number of points in cluster k.
- K is the total number of clusters.
- $\| \cdot \|^2$ denotes the squared Euclidean distance.

The SSE measures the total variance within clusters by summing the squared distances between each data point and its corresponding cluster centroid. A lower SSE indicates that the data points are closer to their centroids, meaning the clusters are more compact and internally homogeneous. As more clusters are added, SSE tends to decrease because the data points are divided into smaller, more similar groups, reducing the overall within-cluster variability.

Therefore, the elbow identified on the graph represents an ideal K, as from this point on, adding more clusters results in only a marginal decrease in the SSE, suggesting that the model may become overly complex. Thus, by selecting the K corresponding to the elbow, the goal is to obtain a model that maintains simplicity while adequately representing the underlying structure of the data.

The second algorithm is Hierarchical clustering (HALKIDI, 2009), which is a grouping technique that creates a hierarchy of groups, allowing the visualization of relationships between the formed groups. According to Noble (2024), there are different forms of hierarchical clustering. This work follows the agglomerative approach, which is the most common. In this approach, each data point is initially considered as an individual cluster, and clusters are progressively merged based on a similarity measure.

The idea behind the algorithm is to represent the structure of the data through a dendrogram, a graph that illustrates the merging or division of groups at different levels of similarity. From this dendrogram, it is possible to determine the ideal number of groups based on a horizontal cut that separates the desired groups.

The hierarchical clustering algorithm calculates the distance between clusters using different linkage criteria, such as Single Linkage (SL), Complete Linkage (CL), or Average Linkage (AL). The distance between two clusters C_i and C_j is defined as follows:

- **Single Linkage**: The distance between two clusters is defined as the minimum distance between any pair of points, one from each cluster:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{2.4}$$

  where:

  - $d(C_i, C_j)$ is the distance between clusters C_i and C_j.

- x and y are individual points belonging to clusters C_i and C_j, respectively.
- $d(x, y)$ represents the Euclidean distance between points x and y.

- **Complete Linkage**: The distance between two clusters is defined as the maximum distance between any pair of points, one from each cluster:

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y) \tag{2.5}$$

where:

- $d(C_i, C_j)$ is the distance between clusters C_i and C_j.
- x and y are individual points belonging to clusters C_i and C_j, respectively.
- $d(x, y)$ represents the Euclidean distance between points x and y.

- **Average Linkage**: The distance between two clusters is defined as the average distance between all pairs of points, one from each cluster:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y) \tag{2.6}$$

where:

- $d(C_i, C_j)$ is the distance between clusters C_i and C_j.
- $|C_i|$ and $|C_j|$ represent the number of points in clusters C_i and C_j, respectively.
- x and y are individual points belonging to clusters C_i and C_j, respectively.
- $d(x, y)$ represents the Euclidean distance between points x and y.

In this work, the single linkage criterion will be used, as it preserves the hierarchical structure of the data and is computationally efficient.

In the agglomerative approach, the algorithm begins by treating each data point as a separate cluster. Then, at each step, the two closest clusters are merged, progressively building a tree-like structure. This process continues until all points are merged into a single cluster, resulting in a hierarchical structure that can be represented visually by a dendrogram.

The last algorithm is the Gaussian Mixture Model (GMM) (REYNOLDS, 2009), a more advanced approach to grouping that assumes that the data is generated from a combination of several Gaussian distributions. This technique is based on the concept that a dataset can be described as a mixture of several normal distributions, each representing a distinct group.

According to Genaro e Astorino (2022), GMM is a probabilistic model that provides a flexible representation of the data structure, allowing each group to have its own mean and covariance. This flexibility is especially useful when groups are not spherical or have different shapes and sizes. Furthermore, GMM not only provides the allocation of points to groups but also the probability of each point belonging to each group, offering a richer understanding of the data structure.

The GMM algorithm uses maximum likelihood to adjust the parameters of the Gaussian distributions, which are the weights, means and covariances. The fitting process is done iteratively through the Expectation-Maximization (EM) algorithm, which alternates between two steps: the Expectation (E) step, where the probabilities of each point belonging to each group are calculated and the Maximization (M) step, where the model parameters are updated based on these probabilities.

The probability of a data point x_i belonging to a particular cluster k is given by the following equation:

$$P(k|x_i) = \frac{\pi_k \mathcal{N}(x_i|\theta_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\theta_j, \Sigma_j)} \tag{2.7}$$

where:

- π_k is the weight of the k-th Gaussian component,
- $\mathcal{N}(x_i \mid \theta_k, \Sigma_k)$ is the probability density function of the multivariate normal distribution with mean θ_k and covariance matrix Σ_k,
- K is the total number of clusters
- the denominator is the sum of the weighted probabilities across all clusters.

The algorithm:

Algorithm 1 Expectation-Maximization (EM) Algorithm for Gaussian Mixture Model (GMM)

Input: Data points $X = \{x_1, x_2, \ldots, x_n\}$, number of clusters K
Initialize: Mixture weights π_k, means μ_k, covariances Σ_k for each cluster k
repeat
  E-step:
  for each data point x_i **do**
    for each cluster k **do**
      Calculate the responsibility γ_{ik} for each point x_i and cluster k:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

    end for
  end for
  M-step:
  for each cluster k **do**
    Update the mixture weight π_k:

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}$$

    Update the mean μ_k:

$$\mu_k = \frac{\sum_{i=1}^{n} \gamma_{ik} x_i}{\sum_{i=1}^{n} \gamma_{ik}}$$

    Update the covariance Σ_k:

$$\Sigma_k = \frac{\sum_{i=1}^{n} \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} \gamma_{ik}}$$

  end for
until convergence criterion is met (e.g., parameters change by a small amount)
Output: The estimated parameters π_k, μ_k, Σ_k for each cluster

The EM algorithm is an iterative method used to estimate the parameters of GMM. It alternates between two main steps: the **Expectation (E) step** and the **Maximization (M) step**. In the E-step, the algorithm computes the responsibilities, which represent the probability that each data point belongs to each cluster, based on the current parameter estimates. In the M-step, the parameters of the model (mixture weights, means, and covariances) are updated to maximize the likelihood of the data given these responsibilities. This process repeats until convergence, typically defined as the point where the parameter updates become sufficiently small. The EM algorithm is widely used in clustering and density estimation tasks where the data is assumed to be generated from a mixture of Gaussian distributions.

2.3 EVALUATION METRICS

After explaining all three methods, it is essential to evaluate each one based on some metrics. The first method, K-Means (JIN; HAN, 2010), can be analyzed using inertia, silhouette score and the Calinski-Harabasz index. The second and third methods will be evaluated using the silhouette score and the Calinski-Harabasz index.

2.3.1 Inertia

This metric measures how well a dataset has been grouped by calculating the sum of squared distances between each point and its centroid. The use of inertia is justified because it provides a simple and effective way to evaluate the compactness of clusters. By minimizing inertia, this ensure that the data points are close to their respective centroids, which is a desirable property for well-formed clusters. Additionally, inertia can be used to assess different clustering configurations, such as choosing the optimal number of clusters, with the goal of finding a balance between the number of clusters and the compactness of the resulting groups. Inertia is defined as:

$$I = \sum_{i=1}^{n} \left| x_i - \mu_{k(i)} \right|^2 \tag{2.8}$$

where:

- n is the total number of data points
- x_i represents a data point
- $\mu_{k(i)}$ is the centroid of the cluster to which x_i belongs.

2.3.2 Silhouette Score

The silhouette score evaluates the quality of clustering by considering both the cohesion (density within clusters) and the separation (distance between clusters). The score ranges from -1 to 1, where a value close to 1 indicates well-clustered points that are far from other clusters, while a negative value suggests that points may be misclassified. The silhouette score for a data point i is calculated as:

$$s(i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))} \tag{2.9}$$

where:

- $a(x_i)$ is the average distance from point x_i to all other points in the same cluster.
- $b(x_i)$ is the average distance from point x_i to all points in the nearest neighboring cluster.

- $\max(a(x_i), b(x_i))$ represents the maximum value between $a(x_i)$ and $b(x_i)$, ensuring that the score $s(i)$ remains between 0 and 1. This normalization helps to standardize the silhouette score, making it comparable across different clusters.

2.3.3 Calinski-Harabasz Index

The Calinski-Harabasz index evaluates clustering quality by measuring the ratio between the dispersion of points within clusters and the dispersion between clusters. A higher value indicates better cluster separation and compactness. According to Calinski e Harabasz (1974), the Calinski-Harabasz index is defined as:

$$\text{CH} = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{n - K}{K - 1} \tag{2.10}$$

where:

- $\text{Tr}(B_k)$ is the trace of the between-cluster dispersion matrix. The trace of a matrix is the sum of the diagonal elements of that matrix. It is defined as:

  $$\text{Tr}(B_k) = \sum_{i=1}^{p} \lambda_i$$

  where λ_i represents the eigenvalues of the matrix B_k (the between-cluster dispersion matrix).

- $\text{Tr}(W_k)$ is the trace of the within-cluster dispersion matrix. Similarly, this is the sum of the diagonal elements of the within-cluster dispersion matrix. It is defined as:

  $$\text{Tr}(W_k) = \sum_{i=1}^{p} \mu_i$$

  where μ_i represents the eigenvalues of the matrix W_k (the within-cluster dispersion matrix).

- n is the total number of data points.

- K is the number of clusters.

In this case, $\text{Tr}(B_k)$ and $\text{Tr}(W_k)$ represent the total dispersion between clusters and within clusters, respectively. The trace operation helps to aggregate these dispersions into scalar values, allowing for the calculation of the Calinski-Harabasz index, which is used to evaluate the quality of clustering.

2.4 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) (MACKIEWICZ; RATAJCZAK, 1993) is a widely used statistical technique for dimensionality reduction and data exploration. The main goal of PCA is to transform a set of potentially correlated variables into a smaller set of uncorrelated variables, called principal components. These components are ordered such that the first component captures the largest portion of the variance in the data, the second component captures the second-largest portion, and so on.

PCA is especially useful in situations where the number of variables is significantly larger than the number of observations, which makes it difficult to visualize and interpret the data. When dealing with high-dimensional datasets, visualizing relationships and structures can become increasingly complex. By reducing the dimensionality, PCA facilitates the visualization, interpretation, and analysis of patterns in the data while preserving as much of the original information as possible. This reduction also helps mitigate the curse of dimensionality, where increasing the number of features may lead to overfitting and poor generalization.

In PCA, the first principal component represents the direction of greatest variance in the dataset. The second principal component is orthogonal to the first and captures the next highest variance, and so on. These components are orthogonal to each other, meaning they are uncorrelated, and they are ordered in decreasing order of their corresponding eigenvalues, which indicate the amount of variance each principal component captures.

The mathematical process of PCA involves several key steps:

1. **Standardization**: Since PCA is affected by the scale of the data, it is common to standardize the data (zero mean, unit variance) to ensure that all variables contribute equally.

2. **Covariance Matrix Calculation**: After standardization, the covariance matrix is computed, which measures how much the dimensions vary from the mean with respect to each other. It is calculated as:

$$\Sigma = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X}) \tag{2.11}$$

   where X is the matrix of standardized data (each row is a data point), $\bar{X}$ is the matrix where each row is the mean vector of the dataset, and n is the number of data points.

3. **Eigenvalue and Eigenvector Computation**: The eigenvalues and eigenvectors of the covariance matrix are then computed. The eigenvectors represent the directions of maximum variance (principal components), and the eigenvalues represent the magnitude of variance captured by each component.

4. **Selecting Principal Components**: A subset of principal components is selected based on the eigenvalues, with those corresponding to larger eigenvalues being more significant. The number of components chosen depends on how much variance is desired to be retained in the data.

5. **Projection onto Principal Components**: The data is then projected onto the selected principal components, resulting in a lower-dimensional representation of the data.

Therefore, given its relevance in various fields of data science, this work will use PCA to investigate feature selection, evaluating how this approach can impact data classification. PCA is a dimensionality reduction technique that transforms the original features into a new set of uncorrelated variables, called principal components. This transformation is achieved by computing the eigenvalues and eigenvectors of the covariance matrix of the data, where the eigenvectors represent the directions of maximum variance (the principal components), and the eigenvalues indicate the amount of variance captured by each component. This relationship is expressed by the following equation:

$$\mathbf{C}\mathbf{v} = \lambda \mathbf{v} \tag{2.12}$$

where:

- $\mathbf{C}$ is the covariance matrix of the data.
- $\mathbf{v}$ is an eigenvector (the direction of maximum variance).
- λ is the corresponding eigenvalue (the amount of variance explained by the eigenvector).

By selecting a subset of principal components, PCA allows for the retention of the most significant patterns in the data while reducing noise and redundancy, which can ultimately improve the performance of classification models.

The ability of PCA to remove redundancy from the data makes it particularly useful in feature selection. By selecting only the most important components, reduce the dimensionality of the dataset without losing significant information, which can improve the computational efficiency and accuracy of classification models. Moreover, by removing noise and irrelevant features, PCA can help enhance the generalization capability of the model, reducing overfitting and improving performance on unseen data.

A common practice when applying PCA is to retain a specified percentage of the total variance in the data, typically around 95%, 90%, or 80%. Retaining 95% of the variance ensures that most of the original information is preserved while reducing the dimensionality. Retaining 90% or 80% of the variance is often used when a further reduction in dimensionality is desired, at the cost of losing some of the original information. The choice of how much variance to retain depends on the specific needs of the application, balancing between dimensionality reduction and the retention of important data characteristics.

2.5 RELATED WORKS

This section presents the theoretical framework used as the foundation for the development of the work, utilizing machine learning techniques and data analysis. To this end, several studies were reviewed to understand the approaches used in classification and unsupervised learning algorithms. Among these, several works stand out, providing support for the development of the present research.

Three key studies are directly relevant to the classification of companies in the stock market, offering valuable information and methodologies that support the goals of this research. They provide essential knowledge on the use of clustering techniques and financial indicators, improving the understanding of market trends and informing investment strategies.

The first study, conducted by D et al. (2024), presents a company recommender system designed for investment decision-making. This research explores various machine learning techniques, with a particular focus on long short-term memory (LSTM) neural networks, to analyze large volumes of temporal data. The ability of LSTMs to capture long-term dependencies in time series is essential for modeling stock price evolution and financial market trends. The current work aligns with this study by aiming to cluster companies based on financial behavior, leveraging similar temporal attributes such as earnings per share (EPS), market capitalization and price-to-earnings (P/E) ratio. By identifying patterns in these financial indicators, this study provides meaningful classifications that assist investors in making informed decisions. Thus, the work of D et al. (2024) directly supports the development of clustering models that enhance market segmentation and investment strategies.

The second relevant study was performed by Pedriali e Dester (2021), which focuses on clustering companies based on their financial attributes. This research is particularly relevant to the present work as it demonstrates how clustering techniques can uncover hidden patterns within extensive financial databases. The study applies clustering algorithms such as K-means and hierarchical clustering to identify groups of companies with similar financial characteristics, which is instrumental in risk assessment and portfolio diversification. Similarly, current research applies clustering techniques to segment the stock market by analyzing financial metrics, providing a comprehensive view of the financial health and potential of different market sectors. The study of Pedriali e Dester (2021) serves as the basis for the clustering methodology used in this work, reinforcing the importance of grouping companies to facilitate strategic decision making.

Finally, the study developed by Zhu et al. (2020) contributes significantly by combining dimensionality reduction techniques, such as Principal Component Analysis (PCA), with machine learning approaches for stock index prediction. This work is particularly relevant to the present research as it highlights the benefits of dimensionality reduction in handling the complexity of financial data. By applying PCA, it becomes possible to transform high-dimensional financial datasets into a smaller set of meaningful variables, preserving critical information while improving model efficiency. This approach is instrumental in the current study effort to classify companies into distinct groups, ensuring that clustering is performed on a simplified yet informative dataset. Moreover, the integration of PCA with machine learning models such as Support Vector Machines (SVM)(CRISTIANINI; RICCI, 2008) and neural networks provides a robust framework for financial analysis, helping to mitigate issues like overfitting. The findings of Zhu et al. (2020) offer a solid theoretical basis for applying dimensionality reduction techniques to improve the precision and interpretability of the clustering models in this investigation.

In summary, these three studies provide crucial support for the classification methodology adopted in this research by providing valuable information on machine learning applications, clustering techniques and dimensionality reduction. The combination of these approaches enables a more comprehensive and effective strategy for classifying companies based on their financial characteristics, ultimately contributing to improved investment decision-making and market analysis.

Other research works serve as complementary foundations to analyze additional techniques and applications of machine learning in the financial market scenario and other fields. For example, Gambim et al. (2023) proposes a strategy for stock portfolio allocation using machine learning algorithms combined with fuzzy rules. This approach stands out for integrating probabilistic and linguistic techniques to improve decision-making in investments, providing an additional perspective to the use of algorithms in finance in this work.

Another relevant study (MANCHEV; MIRCHEV; MISHKOVSKI, 2024), explores the use of Graph Neural Networks for company classification. This research applies neural network models to graphs, an interesting field for analyzing interactions and business relationships in complex networks. Thus, the application of these techniques has broadened the scope of the analysis and the feasibility of their use in the present work.

Finally, Kiersztyn et al. (2022) uses fuzzy systems to classify companies based on innovation levels. This fuzzy logic-based approach allows for handling the subjectivity associated with business innovation, contributing to a more flexible and accurate analysis of companies in dynamic and uncertain environments. Therefore, combining machine learning with fuzzy techniques emerges as a robust alternative to deal with the complexity and variability of the financial market. However, the application of fuzzy logic was not applied in the current work, but it will serve as a foundation for future studies.

3 INITIAL INVESTIGATION

The work was developed as the final project for the course *Scientific Computing and Data Analysis* taught in the second semester of 2023 at the Federal University of Rio de Janeiro by Professor João Antonio Recio da Paixão. The project involved the classification of the 500 companies included in the S&P 500 index (FIGUEIREDO, 2023) based on their daily returns and average volatility of their stock prices. The goal was to help ease the decision-making process for investors, especially those who are inexperienced, by grouping companies according to their characteristics. The project was further refined and turned into a poster that was submitted to CNMAC (FIGUEIREDO et al., 2025), which was accepted.

The purpose of this study was to apply a matrix dimension reduction method and a clustering method to cluster companies in the financial market, taking into account return and volatility. These methods, when used together, allowed for a mathematical analysis to define a way to quantify similarities between companies, facilitating the interpretation of possible patterns present in the data analyzed.

The methodology applied in this work included the following steps:

1. **Data Collection**: Historical data on daily opening and closing prices of the selected companies were collected. The collection carried out in this work included data on E companies over an interval of D days.

2. **Data Preprocessing**: The daily return for the considered interval was calculated using the percentage change between the opening and closing prices:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

where P_t is the asset price at day t and P_{t-1} is the asset price on the previous day. The average daily return was then computed as:

$$\bar{R} = \frac{1}{D} \sum_{t=1}^{D} R_t$$

where D is the total number of days in the considered period. Next, the volatility of these companies was computed as the standard deviation of the daily returns:

$$\sigma_{\text{daily}} = \sqrt{\frac{1}{D-1} \sum_{t=1}^{D} (R_t - \bar{R})^2}$$

Normalization was then applied to both return and volatility values to ensure comparability. The normalization technique used scales the data within a standardized range, improving the robustness of subsequent analyses.

3. **Dimensionality Reduction**: PCA was applied to reduce the dimensionality of the data referring to the daily return history, as the original matrix contained 1032 days, leading to high computational costs. The PCA input consisted of a matrix with the return values of the companies per day ($E \times 1032$). As a result, the reduced matrix had PC columns of principal components ($E \times PC$), where the dimensionality PC was determined based on a variance analysis, ensuring a balance between information retention and computational efficiency.

4. **Company clustering**: A vector corresponding to the volatility of each company was added to the reduced matrix resulting from the application of PCA, considering the entire period analyzed (D days), making the matrix with dimensionality $E \times (PC+1)$. The addition of the volatility vector helped in a more detailed analysis to understand how each company would behave not only in terms of its return, but also how unstable or stable that company was. The resulting matrix was used as input for the K-Means and Hierarchical clustering methods, which were applied to cluster companies based on the return and volatility criteria of the financial stocks of the companies under analysis. At this stage, the clustering method with the best performance among those evaluated and the most appropriate number of clusters (C, where $C = K$ in the case of K-Means) were decided.

5. **Pattern Identification**: The clustering results were analyzed to understand the patterns in the data associated with the companies belonging to the C clusters, obtained from the previous step and how they might be related to the success/failure of the companies in the financial market. Additionally, the distribution of clusters was analyzed considering companies by sector of activity.

This approach enabled a semi-automated analysis of companies' performance in the financial market, significantly reducing human bias in the clustering process. The analysis started with the data collection, where financial data from the 500 companies was gathered. Once the data was collected, Preprocessing steps were performed to clean the data, including the removal of duplicates and non-numeric values. The features were then normalized to ensure each contributed equally to the analysis. After preprocessing, PCA was applied to reduce the dimensionality of the dataset. This reduction in complexity made the subsequent clustering process more efficient and effective.

Clustering techniques such as K-Means or Hierarchical Clustering were employed to group companies based on their financial characteristics. This approach allowed for a more accurate and consistent analysis, making it easier to identify patterns and similarities in company performance without the influence of subjective biases.

3.1 DATA COLLECTION, PREPROCESSING AND DIMENSIONALITY REDUCTION FOR CLUSTERING

Before starting the project, a review of similar works was conducted, as well as a search for potential datasets on Kaggle to find the best way to obtain stock data. However, after analyzing the available resources, it was concluded that it would be more efficient to collect data directly from Yahoo Finance using the *yfinance* library available in Python (KUHN, 2024), thus creating a custom dataset, as none of the available datasets contained all the desired features.

Talking about data collection, in this step, the data were extracted from the Web page List of S&P 500 companies(MAGALHAES, 2024). After that, in the web page there is a table containing the tickers, sectors and other relevant information of the listed companies. Using this data, a dataset was constructed containing the history of daily opening and closing prices of the shares of each of the $E = 503$ companies in the period from 02/01/2020 to 12/17/2023, corresponding to a total of $D = 1{,}032$ days, extracted from the Yahoo Finance.

Based on the data collected from Web, according to the previous step, the daily return of the $E = 503$ largest companies on the stock exchange was calculated. In addition, the volatility of each company was also calculated, through the standard deviation of each of them using the interval of $D = 1{,}032$ days.

To calculate the annualized return and volatility of a financial asset from a time series of prices, daily percentage variations are used. First, the daily return is computed as:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

where P_t represents the asset price at day t and P_{t-1} is the asset price on the previous day. The average daily return is then obtained as:

$$\bar{R} = \frac{1}{D} \sum_{t=1}^{D} R_t$$

where D is the number of days in the considered period. The annualized return is calculated by multiplying the average daily return by the number of business days in a year (N, typically 252):

$$R_{\text{annual}} = \bar{R} \times N$$

Next, the volatility is computed as the standard deviation of the daily returns:

$$\sigma_{\text{daily}} = \sqrt{\frac{1}{D-1}\sum_{t=1}^{D}(R_t - \bar{R})^2}$$

Finally, the annualized volatility is obtained by scaling the daily volatility by the square root of the number of business days:

$$\sigma_{\text{annual}} = \sigma_{\text{daily}} \times \sqrt{N}$$

This approach provides a standardized way to evaluate and compare the returns and risks of financial assets over different time horizons.

When preparing the data, it is important to start by checking for possible inconsistencies, for example, null or empty data and therefore perform data cleaning. At this stage, inconsistencies are treated and normalized using the Z-Score, that is, normalizing each column so that it has a mean of zero (0) and a variance of one (1). In this way, it is possible to compare companies that work with different orders of magnitude (for example, a national company vs. a multinational company), but that have related behaviors, since normalization allows for relative comparison of growth (both increased or decreased in relation to the average of the companies).

3.2 COMPANY CLUSTERING

In this step, it is important to select a clustering model that is capable of grouping the companies given the matrix to which the PCA aggregated to the volatility vector was applied. Two models were chosen to be evaluated in the clustering of the companies, both widely recognized in the literature: K-means and Hierarchical clustering. The final choice between the two will depend on the specific characteristics of the data used and the desired results.

The first method chosen was K-Means. Using this algorithm, it is important to identify the inflection point, commonly known as the elbow, in the graph related to the number of clusters, as the first step in choosing the appropriate number of clusters. Thus, Figure 1 presents this graph where the y-axis represents the Sum of Squared Errors (SSE). Therefore, it is necessary to choose a case where there is an inflection and in this curve, it is noted when there are 3 clusters. This is a significant indication that the increase in the explanation of variability decreases considerably after this point. Therefore, the study proceed with the analysis using 3 clusters in K-Means.

After applying Hierarchical Clustering, the algorithm groups the data points into clusters based on their similarities, starting by treating each data point as its own cluster and progressively merging the closest clusters. This process is represented visually in a

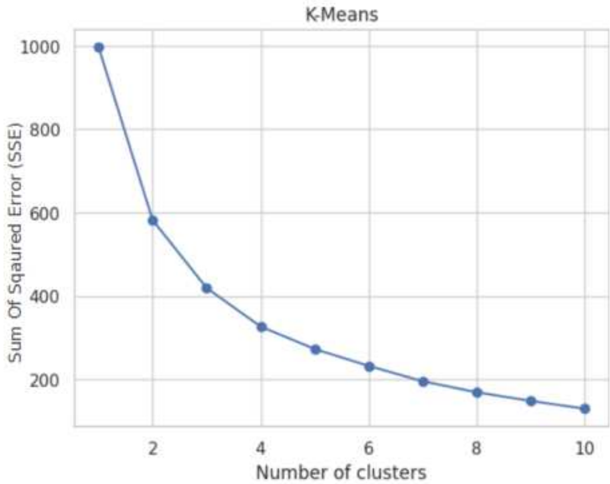


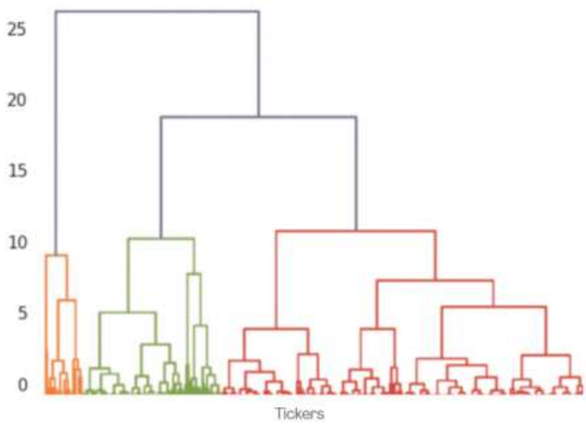
Figure 1 – Elbow Curve

Source: By the Author, 2024

dendrogram,as shown in Figure 2, where the height of each merge indicates the distance or dissimilarity between clusters.

To determine the optimal number of clusters, the algorithm typically look for a "cut" in the dendrogram, where the distances between merged clusters are relatively large, suggesting that additional merging would result in less meaningful groupings. In this case, based on the structure of the dendrogram, the cut was made to form 3 clusters, as these appeared to provide a meaningful and distinct classification of the data.

Figure 2 – Dendrogram of the Hierarchical Model



Source: By the Author, 2024

For a more in-depth evaluation and an effective comparison between the K-Means and Hierarchical model, the silhouette metric will be employed. This technique is used to analyze the internal cohesion and separation between the clusters, with the aim of identifying potential areas for improvement and validating the consistency of the generated clusters.

K-Means clustering is favored for its computational efficiency and scalability, making it suitable for large datasets with well-defined and separated clusters. By iteratively assigning data points to the nearest cluster centroid and updating centroids based on mean values, K-Means efficiently minimizes the within-cluster sum of squares.

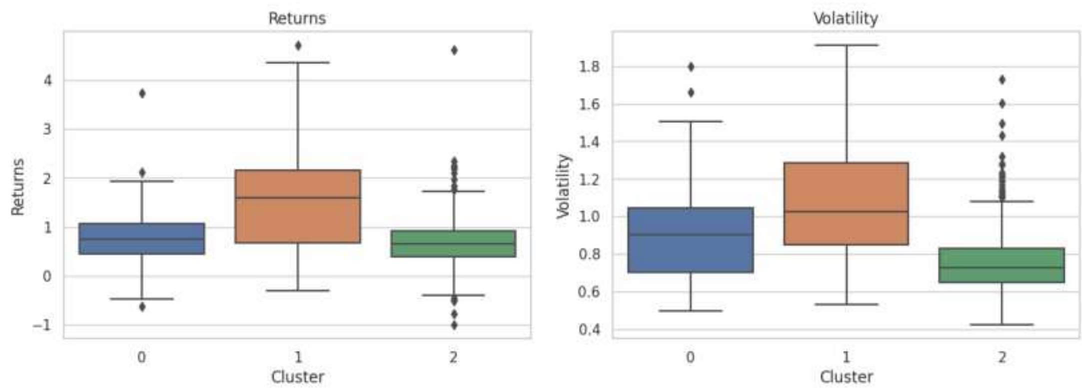
On the other hand, hierarchical clustering offers versatility in capturing hierarchical relationships and handling complex cluster structures. It does not require a predefined number of clusters and can reveal nested clusters within the data. However, hierarchical clustering may be more computationally intensive, especially with larger datasets.

For K-Means, a value of 0.42 was found for the silhouette metric and for the hierarchical method, a value of 0.41. This indicates good cohesion among the points within the clusters formed by these methods and a good division due to the method.

Given that the silhouette metric values obtained for both clustering methods evaluated are very close, for the scenario in question, the decision was to use K-Means in subsequent analyses.

3.3 PATTERN IDENTIFICATION AND DISCUSSION OF RESULTS

Figure 3 – Boxplot of Clusters



Source: By the Author, 2024

For a more detailed analysis of the distributions of the return and volatility values of the companies belonging to the Clusters, a *boxplot* graphic is presented in Figure 3. In this, it can be observed that both the blue cluster (0) and the green cluster (2) present similar distributions in relation to the mean and variance, since they are around 1 and present equal variances. This means a balanced distribution of values both above and below the median. In contrast, the orange cluster (1) displays an asymmetric distribution, indicating a greater concentration of values at one end of the scale, which makes these stocks unstable due to their group distribution.

All three clusters exhibit asymmetric distributions in relation to volatility, so there is no exact mean in these values. It is important to highlight that the green cluster (2) stands out for presenting the smallest asymmetry of the median when compared to the other clusters.

The analysis of the different characteristics of each cluster based on the period of the 1032 days, in terms of return and volatility, provides information about the behavior patterns of companies that can provide subsidies for supporting the decision-making of potential investors, as shown below:

- The green cluster (2) presents a positive return and low volatility, suitable for ***conservative investors***, since they do not like more relaxed investments, without so many variations.

- The blue cluster (0) has a slightly higher return than the green cluster, but with high volatility, attracting ***moderate investors***, those who have a greater interest in return but do not want to take so much risk.

- The orange cluster (1) offers high returns and high volatility and is recommended for more ***risky investors***, those who seek higher returns and do not mind the risks.

An intriguing behavior is observed in the blue cluster (0), which presents a return as low as the green cluster (2), but demonstrates a volatility almost as high as the orange cluster (1). This observation suggests that the blue cluster (0) may represent an interesting and unique case, characterized by an undesirable balance between a low return and a medium volatility.

Regarding returns, the orange cluster (1) stands out for presenting a higher return compared to the blue (0) and green (2). While the blue (0) and green (2) exhibit nearly equivalent median returns, it is important to note that the green cluster (2) stands out for having a lower return compared to the others.

These types of analyses presented make it easier for inexperienced investors to understand the characteristics of groups of companies. Thus, they can be used as subsidies for recommending investments accordingly, taking into account the profile of each investor.

In the previous research for the Congresso Nacional de Matemática Aplicada e Computacional (CNMAC) (FIGUEIREDO et al., 2025), the main focus was on clustering companies based on return and volatility, utilizing dimensionality reduction techniques like PCA to reduce data complexity. This allowed for an initial understanding of patterns and similarities between companies, offering information into potential investment opportunities. However, this approach primarily focused on historical financial data and did not account for a broader range of financial indicators or incorporate more advanced clustering and machine learning methods.

Building on this foundation, the present work expands the scope by incorporating additional financial attributes, such as market capitalization, dividend yields and other key performance indicators, alongside volatility. The introduction of these variables allows for a more comprehensive analysis, enriching the clustering results and enabling a deeper understanding of the companies financial behavior. While PCA was considered, the study ultimately focused on K-Means, Hierarchical clustering and GMM for classification, as these methods were deemed more appropriate for handling the enriched dataset. However, PCA will still be revisited to assess whether it remains valuable for dimensionality reduction within the context of these new attributes. The integration of these clustering techniques refined the grouping process, improving both the accuracy and robustness of the analysis.

Furthermore, the work aims to simplify the process for investors by making the financial market more accessible, particularly for novices. Through visual representations and clearer information into financial data, the work seeks to aid investors in assessing risks and identifying valuable opportunities. This progression from a purely technical study to a more user-centered approach reflects the effort to enhance the understanding of financial markets and equip both novice and experienced investors with the tools needed to make better-informed decisions.

4 DATASET

Considering the initial investigation conducted in Chapter 3, several adjustments were made to improve the robustness and applicability of the research. The first adjustment involved replacing the dataset, as the previous one contained numerous outliers, which could distort the analysis and included attributes that were less relevant for the study objectives. Stock market investors often evaluate multiple aspects of a company and the original attributes did not fully capture the range of factors typically considered. Additionally, the new dataset featured a different set of companies compared to the original one, allowing for a more diversified analysis across industries and financial profiles. This adjustment aimed to refine the dataset for better alignment with the research goals and provide a stronger foundation for the clustering process. Lastly, alternative clustering methods were applied to explore different classification perspectives, as will be detailed in Chapter 5.

Regarding the attributes used from this point onwards, they needed to contribute more to the investor analysis, all measured on the same 1-year time scale. The selected attributes are (LIBERTO, 2024):

- Beta: measures a stock sensitivity to market fluctuations; the higher the beta, the greater the stock volatility. Thus, stocks with betas greater than 1 are considered more volatile.

$$\beta = \frac{\mathrm{Cov}(R_{\mathrm{Asset}}, R_{\mathrm{Market}})}{\mathrm{Var}(R_{\mathrm{Market}})} \tag{4.1}$$

  Where:

  - $\mathrm{Cov}(R_{\mathrm{Asset}}, R_{\mathrm{Market}})$ = Covariance between the return of the asset (R_{Asset}) and the return of the market (R_{Market}). The return of the asset represents the percentage change in its value over a period, calculated as the difference between the final and initial price divided by the initial price. Similarly, the return of the market is the percentage change in the market index value over the same period. The covariance measures how the returns of the asset and the market move together, with a positive value indicating they move in the same direction and a negative value indicating they move in opposite directions.
  - $\mathrm{Var}(R_{\mathrm{Market}})$ = Variance of the return of the market

- Volatility: a measure of investment risk that indicates the level of risk exposure an investor faces. This refers to the annualized volatility, which is the volatility adjusted for a one-year period.

$$\text{Volatility} = \sigma\sqrt{T} \tag{4.2}$$

Where:

- σ = Standard deviation of returns
- T = Number of periods in the time frame

- EPS (Earnings Per Share): a measure of a company profitability that indicates how much profit each outstanding common share generated.

$$\text{EPS} = \frac{NI}{S_{\text{Diluted}}} \tag{4.3}$$

Where:

- NI = Net income
- S_{Diluted} = Total number of diluted shares outstanding, which includes not only the shares currently in circulation but also those that could be issued through the conversion of options, convertible bonds, and warrants.

- P/E Ratio: Price/Earnings is determined by dividing the current stock price by the earnings per share reported over a given time period. This index is used to assess the company market value.

$$\text{P/E Ratio} = \frac{P_{\text{Current}}}{\text{EPS}} \tag{4.4}$$

Where:

- P_{Current} = Current stock price
- EPS: measure of a company profitability 4.3

- Dividend Yield: indicates the dividend yield. It is an index created to measure the profitability of a company dividends relative to its stock price.

$$\text{Dividend Yield} = \frac{D_{\text{Share}}}{P_{\text{Share}}} \tag{4.5}$$

Where:

- D_{Share} = Dividend per share
- P_{Share} = Market value per share

- Market Capitalization: an estimate of a company market value based on expectations about future economic and monetary conditions.

$$\text{Market Cap} = P_{\text{Current}} \times S_{\text{Outstanding}} \tag{4.6}$$

  Where:

  - P_{Current} = Current stock price same used at equation (4.4)
  - $S_{\text{Outstanding}}$ = Total number of shares outstanding, which represents the total number of shares currently in circulation, excluding treasury shares.

- Revenue Growth: refers to the increase in a company total revenue over a specific period.

$$\text{Revenue Growth Rate} = \frac{R_{\text{Current}} - R_{\text{Previous}}}{R_{\text{Previous}}} \tag{4.7}$$

  Where:

  - R_{Current} = Current period revenue
  - R_{Previous} = Previous period revenue

- Debt-to-Equity Ratio: A financial index that indicates the relative proportion of equity and debt used to finance a company's assets. Where the equity is t he capital owned by shareholders, representing the difference between a company's assets and liabilities. And Debt is the financial obligations of the company, such as loans, bonds, and other forms of debt.

$$\text{Debt-to-Equity Ratio} = \frac{L_{\text{Total}}}{E_{\text{Total}}} \tag{4.8}$$

- L_{Total} = Total liabilities, which represent all the financial obligations of a company, including both current liabilities (short-term debts) and non-current liabilities (long-term debts).

- E_{Total} = Total shareholders' equity, which represents the difference between a company's total assets and total liabilities, indicating the residual value owned by the shareholders.

After these attribute changes, an attempt was made to gather all the companies from the S&P 500 index. Initially, the dataset contained 503 companies, but after data cleansing, which involved removing missing and duplicate data, the final dataset comprised $E = 307$ companies. However, this number turned out to be less effective, as it was relatively small for the scope of the intended project and analyses.

4.1 CREATION OF THE UTILIZED DATASET

Therefore, a new approach was implemented and a Python script which retrieves all companies listed on NYSE and captures data using the *yfinance* library. After collecting the data, a total of approximately 1.300 companies were obtained. However, the library might lack some information, so after cleaning the data by removing companies with missing values and duplicates, 698 companies remained. Nevertheless, this is still a good number to work with.

After analyzing the collected data, the next step was to normalize it using MinMax scaling, as shown in equation (4.9), since some variables are on different scales. Normalization optimizes certain computational processes during classification. This step adjusts the data so that each feature has a mean close to zero and a standard deviation of one.

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{4.9}$$

Where:

- X_{norm}: Normalized value.
- X: Original value.
- X_{min}: Minimum value of the dataset for the variable.
- X_{max}: Maximum value of the dataset for the variable.

4.2 ANALYSIS OF ATTRIBUTES

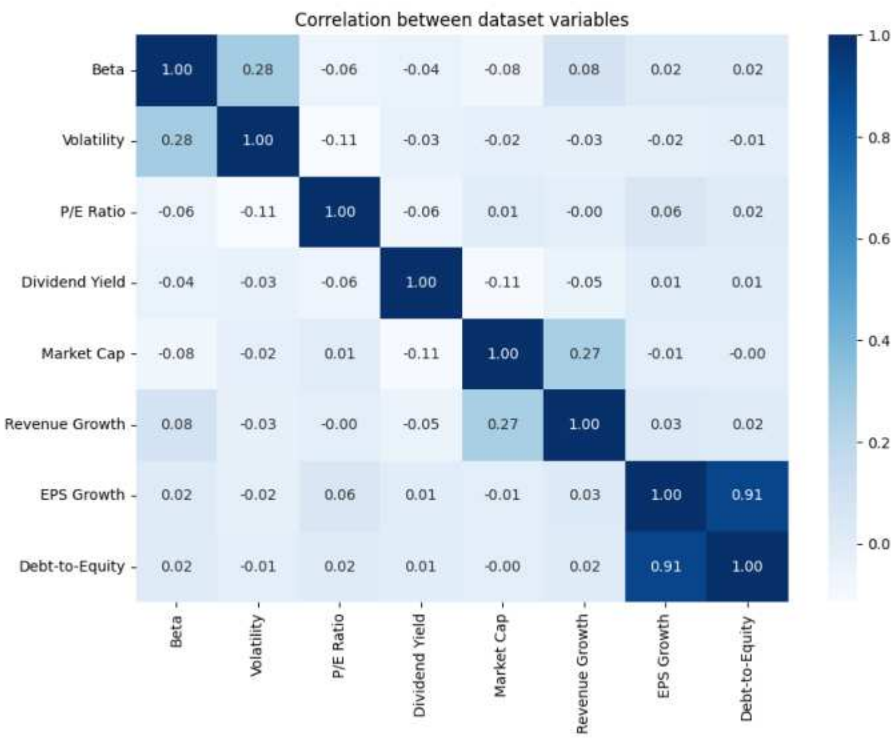
The analysis of the correlation matrix, presented in Figure 4, it is possible to analyse into the strength of relationships between pairs of financial variables. The scale ranges from -1, indicating a perfect negative correlation, to 1, indicating a perfect positive correlation. From Figure 4, three key pairs of attributes are highlighted for deeper analysis:

- **Volatility and Beta:** Beta measures the sensitivity of an asset relative to market movements, while volatility represents the price fluctuation of an asset over time. The moderate positive correlation (0.28) between these two attributes indicates that assets with higher Beta may tend to exhibit slightly higher volatility.
- **Revenue Growth and Market Capitalization:** Market capitalization reflects a company total market value, while revenue growth measures the increase in revenue over time. The positive correlation (0.27) between these two attributes suggests that companies with higher revenue growth tend to have larger market capitalizations. Although this relationship is slightly weaker than that of Volatility and Beta, it highlights that revenue expansion often plays a role in increasing a company's market valuation, which may attract investor interest.

- **EPS Growth and Debt-to-Equity Ratio:** Earnings per share (EPS) growth measures a company's profitability growth, while the Debt-to-Equity ratio indicates the balance between a company's debt and equity. The strong positive correlation (0.91) between these attributes suggests a much closer relationship compared to the other pairs. This significant relationship shows how companies are leveraging their financial structure to support growth and the potential risks associated with their strategies.

In summary, while the correlations between Volatility and Beta (0.28) and Revenue Growth and Market Capitalization (0.27) are moderate and similar in strength, the correlation between EPS Growth and Debt-to-Equity (0.91) stands out as markedly stronger, indicating a much closer and more significant relationship. This correlation suggests that companies with higher EPS growth tend to have a more pronounced alignment between their earnings growth and their capital structure (i.e., the mix of debt and equity used to finance their operations and growth), which could be an important factor for investors to consider.

Figure 4 – Correlation Matrix



Source: By the Author, 2024

With this, it is a good indication that for computational efficiency, it may be beneficial to reduce or combine highly correlated attributes. The high correlation between the identified pairs suggests redundancy in the information, which prompts the consideration of whether applying PCA to reduce the dimensionality of the data would be worthwhile.

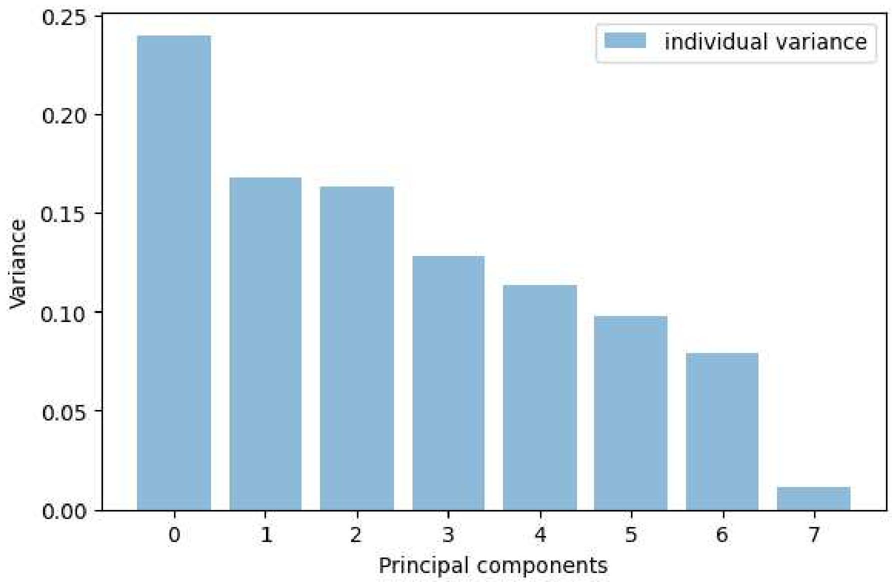
4.3 DIMENSIONALITY REDUCTION

After analyzing the correlation between the attributes, a segmentation is performed where one dataset will undergo dimensionality reduction using PCA and another identical dataset will be used without dimensionality reduction. This approach allows for a comparative evaluation of how dimensionality reduction impacts the clustering results and the overall interpretation of the data. The suspicion is that, in this case, PCA may not be as effective, as it is acting as a dimensionality reduction technique on a set of attributes that are not numerous to begin with. As a result, the benefits of reducing dimensionality might be less pronounced compared to cases with a higher number of features.

Based on the graph presented in Figure 5, it is possible to analyze the variance explained by each principal component, where PC_0 explains the largest portion of this variation, meaning it captures the original fraction of the information and represents most of the data patterns without losing much information. However, in this specific case, PCA will be used to analyze the reduction of attributes rather than the temporal dimensionality of the data, as the study is working with the characteristics of the companies.

In the dataset where PCA is applied, it will be reduced to 6 principal components because, as shown in Figure 5, to retain 95% of the data variance, the ideal number of components is 6. This means that the first 6 components together explain 95% of the total variance in the data, allowing to retain most of the important information while reducing the dimensionality. Additionally, cases were also analyzed with a cumulative variance of 90%, which resulted in a total of 5 components, and 80%, which suggested an ideal number of 4 principal components.

Figure 5 – PCA Variance



Source: By the Author, 2024

Thus, the work will proceed in parallel with four datasets, with non-PCA, PCA-95%, PCA-90%, PCA-80% until the classification step in Chapter 5, where a more in-depth study will be conducted to determine if this matrix reduction is truly necessary and to choose the ideal dataset for the study. This parallel approach ensures that both methods are tested, allowing for a more robust conclusion on the impact of dimensionality reduction in the clustering process.

5 CLASSIFICATION

This chapter of the work presents the issue of whether to use a dataset with or without dimensionality reduction, as well as the implementation and evaluation of unsupervised learning methods. The first analysis involves applying the K-Means algorithm to both the PCA-reduced and non-reduced datasets to better understand how to best utilize the dataset. Furthermore, the decision to use K-Means as the reference algorithm is supported by related works using the method, in addition to personal preference due to its relative simplicity and the previous study in Chapter 3.

However, before applying the method, it is necessary to analyze the ideal number of clusters. A graph was created for the PCA and non-PCA datasets to analyze the inflection curve. In this regard, Figure 6 suggests that the ideal classification lies between 2 and 3 clusters for the non-PCA dataset. While there is a significant reduction in error from 1 to 2 clusters, the decrease continues meaningfully from 2 to 3. The elbow point, where the error rate begins to slow down, is most noticeable around 2 clusters.

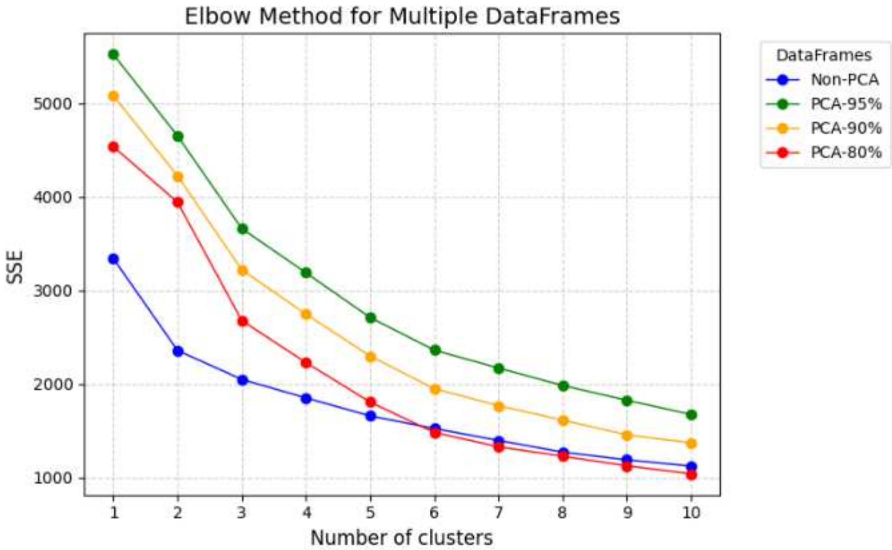
For the dataset without PCA, the error rates show a significant reduction when moving from 1 to 2 clusters. From 2 to 3 clusters, the reduction is less pronounced but remains relevant. Beyond 3 clusters, the reduction in error rates becomes progressively smaller, indicating diminishing returns when increasing the number of clusters.

Thus, using 3 clusters achieves a good balance between simplicity and capturing the underlying structure of the data. While 2 clusters could be an alternative for a simpler division, the study proceed with 3 clusters for now, as it provides a more detailed and complex representation of the data. To further support the selection of the ideal number of clusters, the numerical values of the error rates for each cluster count can be analyzed.

Furthermore, by observing the elbow curves for each of the PCA variations, it is evident from Figure 6 that the ideal number of clusters is also 3. For more explained details, the PCA-80% dataset (blue curve), the optimal number of clusters is 3, as the elbow point is most evident there. Similarly, for the PCA-90% dataset (orange curve), the elbow point also suggests 3 clusters, with a possible alternative at 4 clusters. For the PCA-95% dataset (green curve), the elbow method indicates 3 clusters as the best choice, although 4 clusters could also be considered depending on the application.

These findings indicate that while most datasets suggest 3 clusters as the optimal configuration, the non-PCA dataset shows a preference for 2 clusters. This preference for 2 clusters could be attributed to the nature of the normalization process, which adjusts the scale of the data, there by affecting how the algorithm perceives the structure of the dataset. By reducing the effect of extreme values or large variations in the attributes, normalization may lead to a clearer distinction between two natural groupings within the data.

Figure 6 – Elbow Curve for all dataframes



Source: By the Author, 2024

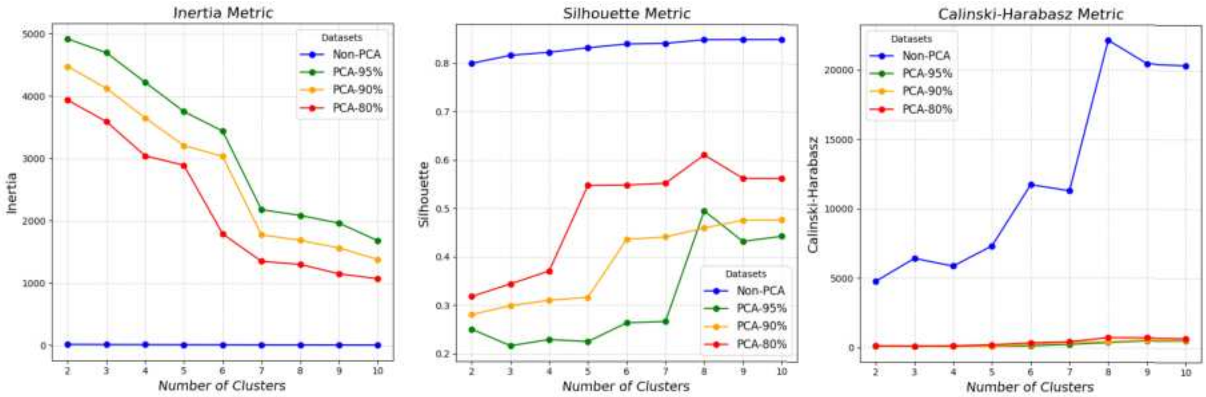
Moreover, the choice of the number of clusters is not purely data-driven; it also depends on the algorithm sensitivity to initial conditions, such as the random initialization of centroids in K-Means or other clustering techniques. As a result, the number of clusters identified as optimal can vary based on both the preprocessing applied and the inherent patterns within the data. Further testing and validation techniques, such as silhouette analysis or the elbow method, can be used to refine the selection and ensure that the final clustering configuration provides the most meaningful information for decision-making.

5.1 ANALYSIS OF THE CLASSIFIED DATASETS

This section presents a comparison of dataset variations based on the K-Means metrics discussed in Chapter 2. Figure 7 illustrates the performance of four configurations: the original dataset without PCA and the datasets reduced using PCA while retaining 95%, 90%, and 80% of the variance. The comparison is conducted using inertia, the silhouette index, and the Calinski-Harabasz index considering different numbers of clusters. The decision to analyze more clusters was made out of curiosity, aiming to understand how the model's performance varied with different cluster numbers, as well as to assess the influence of dimensionality reduction on the evaluation metrics.

The inertia graph shows that the non-PCA dataset consistently exhibits the lowest inertia values, indicating tighter clusters compared to the PCA-reduced datasets. Although all configurations demonstrate a reduction in inertia as the number of clusters increases, the original dataset maintains a clear advantage in terms of cluster compactness. The PCA-reduced datasets, in contrast, display higher inertia values, suggesting a slight loss of compactness as a result of dimensionality reduction.

Figure 7 – Base evaluation indicators



Source: By the Author, 2024

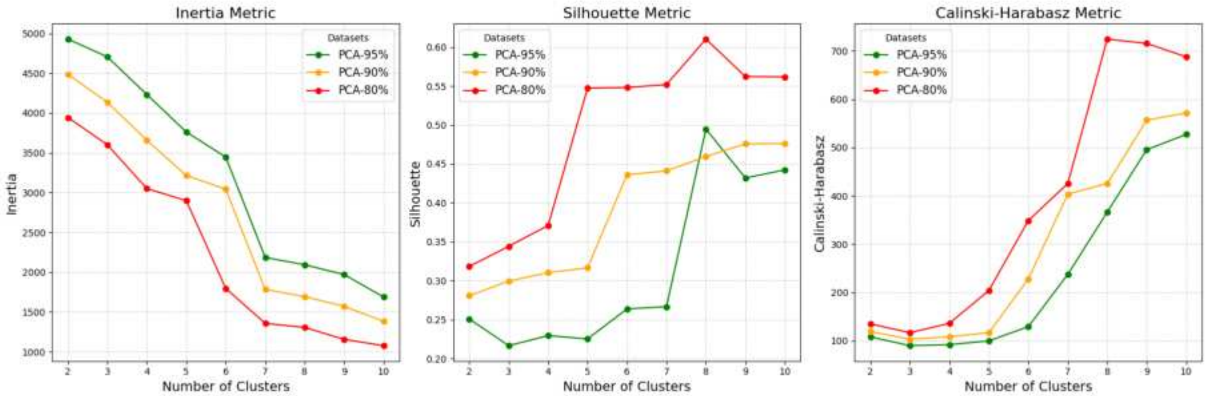
The silhouette index provides further evidence of the superior clustering quality achieved by the non-PCA dataset. The silhouette score indicates that using 3 clusters provides a good balance between intra-cluster cohesion and inter-cluster separation, highlighting this configuration as an optimal choice for clustering. While it may not yield the highest silhouette score across all configurations, it represents a practical and well-rounded option, ensuring effective clustering while maintaining model simplicity. Among the PCA-reduced datasets, the version retaining 80% of the variance performs slightly better than the others, though it still falls short of the results obtained with the original dataset. This pattern underscores the impact of dimensionality reduction on clustering performance, as the separation and cohesion of clusters are reduced.

The Calinski-Harabasz index corroborates these findings, with the non-PCA dataset achieving significantly higher values across all numbers of clusters. The peak value is observed at 3 clusters, further supporting this configuration as the most appropriate for the original dataset. In contrast, the PCA-reduced datasets exhibit lower index values, reflecting diminished separation and compactness of clusters.

While the metrics indicate that increasing the number of clusters beyond 3 might slightly improve separation and compactness, the gains are minimal and do not justify the additional complexity. Therefore, 3 clusters strike an appropriate balance between simplicity and effectively capturing the underlying structure of the data. For the subsequent analysis, the choice will focus on the non-PCA dataset with 3 clusters, ensuring a robust and interpretable clustering solution.

Figure 8 provides a detailed comparison of the performance of K-Means clustering across multiple datasets that have been subjected to dimensionality reduction using PCA. The datasets were reduced to varying levels of retained variance, namely 95%, 90% and 80%, in order to evaluate how different levels of dimensionality affect clustering outcomes. By observing the clustering results at these different variance retention levels, we can assess the trade-offs between maintaining the original data structure and reducing computational complexity. This analysis not only highlights the differences in clustering performance but also offers information into how PCA impacts the effectiveness and efficiency of the K-Means algorithm when applied to financial data.

Figure 8 – PCA Base Assessment Indicators



Source: By the Author, 2024

In terms of inertia, the metric decreases as the number of clusters increases for all PCA-reduced datasets, which is expected since adding clusters reduces intra-cluster variance. Among the PCA versions, the dataset retaining 95% of the variance shows the highest inertia values, indicating less compact clusters. Conversely, the dataset retaining 80% of the variance achieves the lowest inertia values, reflecting tighter clusters compared to the other PCA configurations.

For the silhouette index, which evaluates the balance between intra-cluster cohesion and inter-cluster separation, the PCA-80% dataset achieves the highest silhouette score, peaking at 8 clusters with a value of approximately 0.55. The PCA-90% and PCA-95% datasets perform worse, with lower silhouette values that gradually increase up to 6 or 7 clusters but never surpass the PCA-80% dataset. This result suggests that reducing dimensionality further may better capture separable patterns in the data.

The Calinski-Harabasz index follows a similar trend to the silhouette index, increasing as the number of clusters grows. Once again, the PCA-80% dataset achieves the best performance, peaking at 8 clusters, indicating higher clustering quality at this level. In contrast, the PCA-95% dataset consistently yields the lowest Calinski-Harabasz values, indicating less distinct clustering.

A key observation is the discrepancy between the elbow method and silhouette analysis. The elbow method, as derived from the inertia graph, suggests that 3 clusters provide a good balance between simplicity and performance. However, the silhouette analysis indicates that 8 clusters are optimal, as this configuration maximizes the silhouette score across all PCA-reduced datasets. This discrepancy arises because the elbow method focuses solely on reducing intra-cluster variance, while the silhouette index accounts for both intra-cluster cohesion and inter-cluster separation, offering a more comprehensive evaluation of clustering quality.

In summary, while the elbow curve suggests 3 clusters for simplicity, the silhouette and Calinski-Harabasz indices favor 8 clusters, particularly for the PCA-80% dataset. However, it was observed that the PCA-reduced datasets lost all comparisons, confirming the suspicion that PCA would not be effective in this case. The reduced dimensionality did not provide significant improvements in clustering performance, reinforcing the idea that PCA may not be beneficial when dealing with a dataset that already has a manageable number of attributes.

5.2 APPLYING UNSUPERVISED LEARNING METHODS

The selected dataset was the non-PCA one, as it achieved the best evaluation results. Machine learning methods will now be applied, starting with K-Means, followed by hierarchical clustering, and finally, GMM.

First, K-Means was applied to the dataset with three clusters, as discussed earlier in the chapter, as this choice balances optimal performance and interpretability. Regarding the method, it is necessary to evaluate the classification performance based on the metrics discussed in Chapter 2.

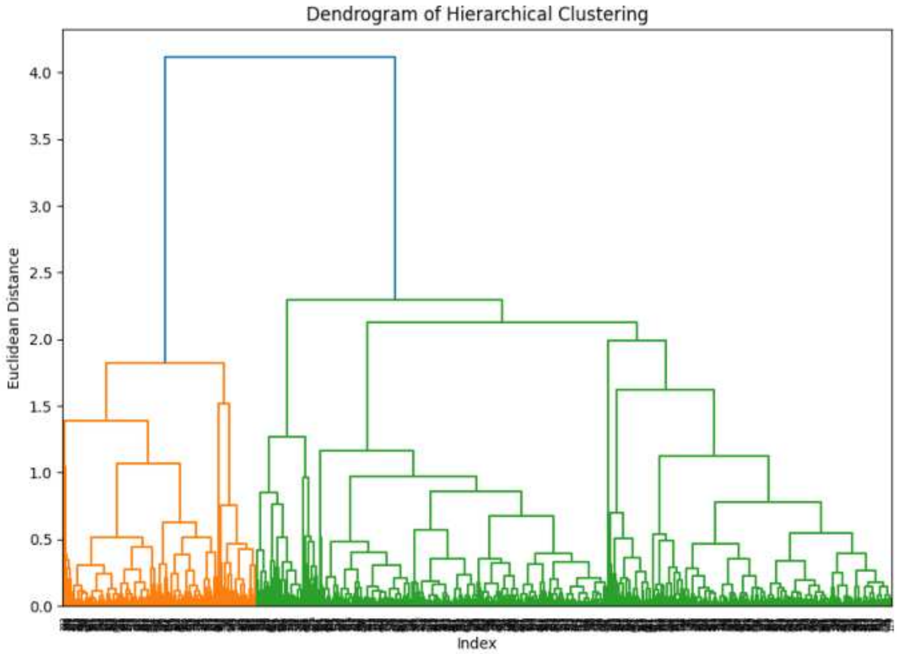
The Figure 7 suporte that the first metric analyzed, the inertia, obtained a value of 20.515. This value should be considered in relation to the number of clusters and the total variability of the data. An inertia value of 20.5 indicates that the points are relatively close to their centroids, which is a good sign, suggesting that the data is well clustered.

Next, the silhouette analysis resulted in a value of 0.815, indicating excellent separation between the clusters. This value suggests that the points are mostly close to the center of their respective clusters and distant from other clusters, which is highly positive for the quality of the classification.

Finally, the Calinski-Harabasz index, which evaluates the quality of the clusters based on the dispersion between and within the clusters, presented a value of 6424.757. This high value suggests that the clusters are well separated and have good internal cohesion.

The second method applied was Hierarchical Clustering and the Figure 9 presents the result of the agglomeration returned by the algorithm. From this, it is possible to observe that the data points were grouped into several clusters with varying distances. The dendrogram indicates that, initially, the points form smaller clusters that merge into larger groups as the distance increases.

Figure 9 – Dendrogram of Hierarchical clustering



Source: By the Author, 2024

In Figure 9, the green branches represent smaller, more cohesive clusters, while the orange branch shows the distance at which these clusters were merged. The height of the branches in the dendrogram corresponds to the Euclidean distance between points or clusters when they are combined, with higher branches representing larger distances. By cutting the dendrogram at a distance level that corresponds to the desired number of clusters, 2 distinct groups were identified.

Thus, it is possible to see that hierarchical clustering also provides clear groupings, though with a slightly different structure compared to the K-Means clustering. The dendrogram can help assess the closeness of clusters and validate the choice of the number of groups by examining the merging process and distances.

After defining the groups using the hierarchical algorithm, it is necessary to evaluate the classification performance using the same metrics applied in K-Means, except the inertia. The silhouette metric, for example, obtained a value of 0.33, which suggests weak separation between the formed groups. Therefore, the low silhouette value indicates that the clusters are not well-defined or separated, suggesting that the hierarchical algorithm was not effective in dividing the data. This is because the algorithm may have grouped points that, although close in terms of distance, do not have sufficiently similar characteristics to justify classification into the same group.

Additionally, the Calinski-Harabasz index presented a value of 236.06, which is a low value. This suggests that the separation between the clusters is not ideal, corroborating the result from the silhouette and indicating that the hierarchical algorithm failed to effectively segment the data.

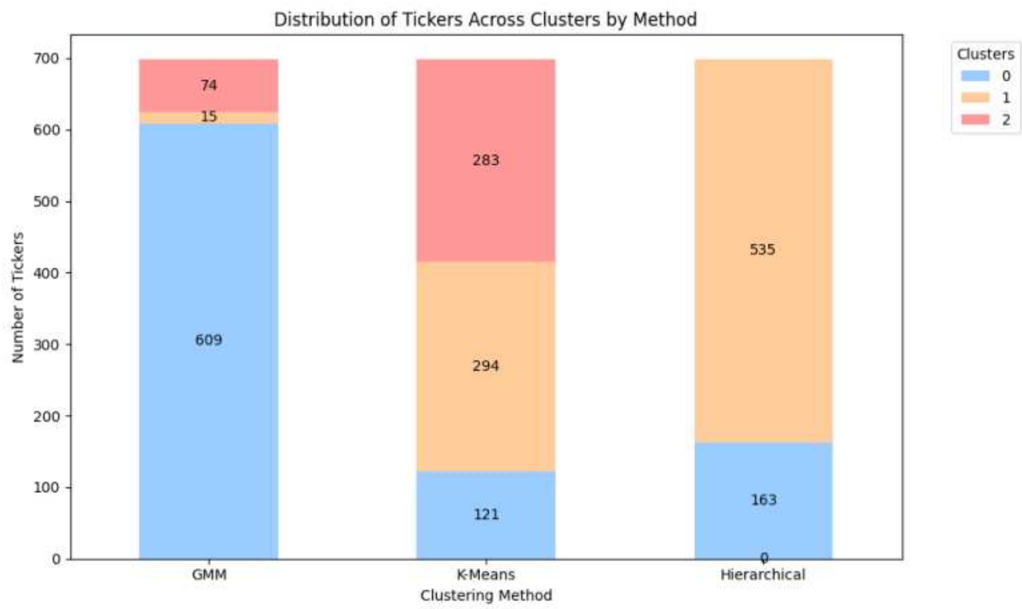
Lastly, regarding the application of the methods, the GMM was chosen with 3 clusters and the model was evaluated using two metrics: the silhouette index and the Calinski-Harabasz index. The silhouette index, which achieved a value of 0.767, indicates that the model was able to form well-defined, cohesive and distinct clusters, which is a good sign of quality in data classification. Additionally, the Calinski-Harabasz index achieved a value of 2890.510, reflecting a high dispersion between the clusters compared to the internal dispersion.

This high value suggests that the GMM effectively separated the clusters, reinforcing the idea that the formed groups are not only well-defined but also well-distinguished from each other. Therefore, both the silhouette index and the Calinski-Harabasz index indicate that the GMM performed well in the clustering task, with well-separated and cohesive clusters.

Furthermore, Figure 10 provides a comparative analysis of the three clustering methods applied: GMM, K-Means, and Hierarchical. The results demonstrate distinct patterns of cluster distribution for each technique.

It is important to note that the clusters labeled as 0, 1, and 2 in each method do not necessarily correspond to the same risk profiles, such as conservative, moderate, or aggressive. These labels were assigned purely for visualization purposes to analyze the distribution of tickers across methods. For a more detailed interpretation of each cluster's characteristics and their financial attributes, refer to the appendix 6.2.

Figure 10 – Number of Tickers per cluster by clustering Method



Source: By the Author, 2024

The GMM method shows a highly uneven distribution, with cluster 0 containing the majority of tickers (609), while clusters 1 and 2 contain substantially fewer tickers (15 and 74 respectively). This suggests that the GMM algorithm identified a dominant pattern that grouped most tickers into a single cluster.

In contrast, the K-Means method presents a more balanced distribution between clusters 1 and 2, containing 294 and 283 tickers respectively, while cluster 0 has a smaller representation with 121 tickers. This distribution suggests that the algorithm captured underlying patterns in the data, grouping companies with similar financial characteristics while maintaining a meaningful segmentation.

The Hierarchical method shows a notably different pattern, with cluster 1 containing the majority of tickers (535), followed by cluster 0 with 163 tickers, while cluster 2 shows minimal presence (as indicated by the small red segment). This pattern aligns with the dendrogram shown in Figure 9, suggesting that the hierarchical clustering identified one dominant group while maintaining a secondary substantial cluster.

These comparative results highlight how each clustering technique approaches the data structure differently, with GMM favoring a single dominant cluster, K-Means providing more balanced groupings, and Hierarchical clustering identifying a two-cluster structure in the data.

To advance in the company classification section, it is essential to first compare the three clustering methods evaluated, which is presented in Table 1. Using performance metrics, this comparison provides a quantitative assessment of each method's ability to separate the data into meaningful clusters. K-Means emerged as the best clustering method, achieving the highest silhouette score and a solid Calinski-Harabasz value, which indicates its effectiveness in producing well-defined clusters and also a more evenly distributed characteristics.

Table 1 – Clustering Performance Metrics for Different Algorithms

Method	**Inertia**	**Silhouette**	**Calinski-Harabasz**
K-Means	20.515	0.815	6424.757
Hierarchical	-	0.33	236.06
GMM	-	0.767	2890.510

In contrast, while GMM demonstrated satisfactory performance, it was less effective in distinctly separating the clusters compared to K-Means. Hierarchical clustering, on the other hand, yielded the poorest results based on both the silhouette score and the Calinski-Harabasz metric, indicating that it was not well-suited for this clustering task due to the difficulty in separating the data. Consequently, K-Means was identified as the most appropriate method for clustering the data in this context.

5.3 CLUSTERING SELECTION: PROCEEDING WITH K-MEANS

After evaluating multiple clustering methods, including K-Means, GMM, and Hierarchical Clustering, a quantitative comparison was conducted using performance metrics such as the silhouette score and Calinski-Harabasz index. As demonstrated in Table 1, K-Means outperformed the other methods by providing well-defined and cohesive clusters, making it the most suitable choice for this analysis.

This section will focus on the interpretation of the K-Means results, detailing the composition of each cluster and its implications for investment decision-making. The classification of companies, sectoral distribution, and investor profile recommendations will be based on the clusters generated by this method. Although other methods were analyzed, the K-Means algorithm was run 10 times, and the results obtained are presented in Appendix 6.2.

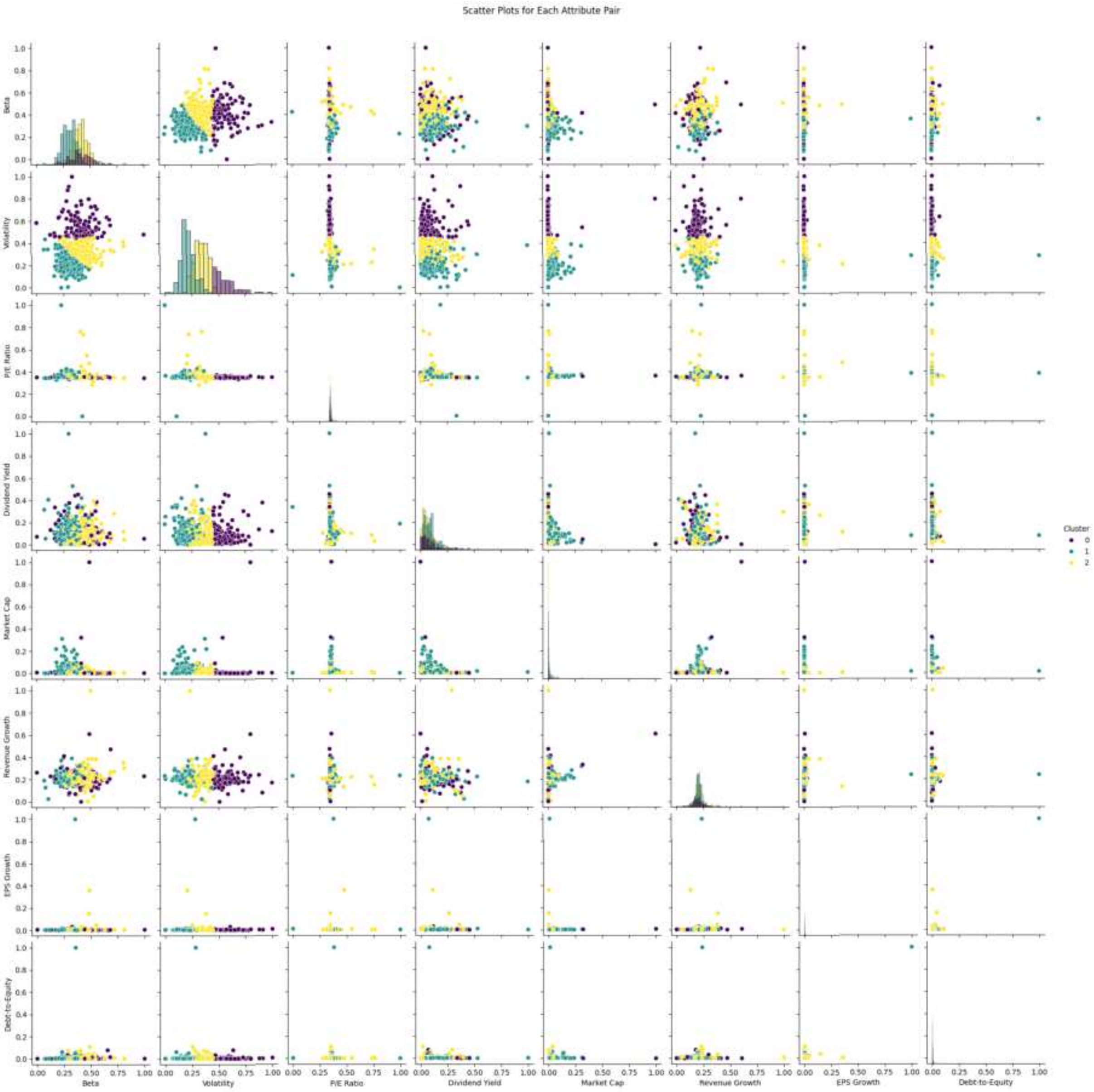
The scatter plot matrix provided in Figure 11 serves as a critical visualization tool for understanding the relationships between various financial indicators across different clusters identified in your analysis. This type of plot is particularly valuable because it enables the simultaneous examination of pairwise relationships between multiple variables, providing an overview of how these variables interact and contribute to cluster formation. Despite its limitations, it plays a fundamental role in preliminary data exploration and cluster validation.

Despite its limitations, the scatter plot matrix remains a valuable resource in cluster analysis. It facilitates the visual inspection of data distributions within clusters and highlights patterns or separations associated with specific financial indicators. Moreover, it supports the evaluation of the clustering model by revealing whether clusters are well-separated or overlapping in the projected dimensions. For instance, clusters that are distinct across multiple variable pairs suggest that the clustering algorithm, such as K-Means or GMM, effectively identifies meaningful groupings. Conversely, significant overlap between clusters may indicate the need for refining the model or incorporating additional features.

The diagonal elements of the scatter plot matrix show histograms for each variable, which illustrate the univariate distribution of data within the clusters. Some variables exhibit distinct distributions among clusters, suggesting that these variables play a key role in differentiating the groups. In contrast, other variables display significant overlap, indicating that they contribute less to the clustering process.

Certain pairs of variables, such as Beta and Volatility or EPS Growth and Revenue Growth, display clear separations between clusters. This observation highlights the importance of these variables in defining the clusters. However, other combinations show substantial overlap, suggesting that those variables are less effective in distinguishing between groups. Such overlaps could indicate areas where the clustering model may need

Figure 11 – Dispersion for Each Pair of K-Means Attributes



Source: By the Author, 2024

refinement or where additional features could improve the analysis.

The scatter plot matrix also reveals potential non-linear relationships between variables within clusters. For instance, the relationship between Beta and Volatility exhibits a non-uniform dispersion, suggesting a possible non-linear trend between market risk and price fluctuations. Similarly, the interaction between Debt-to-Equity and EPS Growth indicates a more complex relationship, implying that variations in capital structure may influence earnings growth in a non-linear manner. These patterns highlight the need for potential transformations or interactions between variables to better capture the underlying dynamics in the data, ultimately improving the model's ability to analyze complex financial structures.

After the analysis, it is possible to classify the identified clusters. Based on the terminology and definitions provided in Fernando (2024), the thresholds for each attribute were established to ensure a consistent and accurate classification into categories such as low, moderate or high. These thresholds are detailed in Table 2 to provide a clear and systematic approach for cluster interpretation.

Table 2 – Classification Intervals for Attributes

Attribute	**Low ($\leq$)**	**Moderate (] ,])**	**High ($\geq$)**
Beta	$[0, 0.3]$	$]0.3, 0.7]$	$]0.7, \infty[$
Volatility	$[0, 0.15]$	$]0.15, 0.3]$	$]0.3, \infty[$
P/E Ratio	$[0, 15]$	$]15, 25]$	$]25, \infty[$
Dividend Yield	$[0, 0.02]$	$]0.02, 0.4]$	$]0.4, \infty[$
Market Cap	$[0, 1.10^9]$	$]1.10^9, 1.10^{10}]$	$]1.10^{10}, \infty[$
Revenue Growth	$[0, 0.05]$	$]0.05, 0.15]$	$]0.15, \infty[$
EPS Growth	$[0, 0.05]$	$]0.05, 0.15]$	$]0.15, \infty[$
Debt-to-Equity Ratio	$[0, 0.5]$	$]0.5, 1.0]$	$]1.0, \infty[$

After defining the limits and analyzing the distribution, the clusters generated by K-Means exhibit distinct characteristics based on various financial attributes, as detailed in Table 3. K-Means is used to define characteristics and facilitate the classification process, given the large number of companies in the dataset.

The table categorizes companies into three clusters based on key financial metrics. For example, Beta, which measures a company's market risk, shows moderate values in cluster 0 (0.41) and cluster 2 (0.42), while cluster 1 has a lower Beta (0.30). Volatility, representing fluctuations in stock prices, is high in cluster 0 (0.56) and cluster 2 (0.33), but moderate in cluster 1 (0.21). Additionally, all clusters have similar P/E Ratios (around 35-36), indicating that the companies within each cluster have relatively similar price-to-earnings ratios.

Table 3 – cluster Characteristics Based on K-Means

Attribute	**cluster 0**	**cluster 1**	**cluster 2**
Beta	Moderate (0.41)	Low (0.30)	Moderate (0.42)
Volatility	High (0.56)	Moderate (0.21)	High (0.33)
P/E Ratio	High (35)	High (36)	High (36)
Dividend Yield	Moderate (0.10)	Moderate (0.10)	Moderate (0.08)
Market Cap	Low ($1 \cdot 10^9$)	Moderate ($2 \cdot 10^9$)	Low ($1 \cdot 10^9$)
Revenue Growth	High (0.21)	High (0.21)	High (0.21)
EPS Growth	Low (0.004)	Low (0.007)	Low (0.006)
Debt-to-Equity Ratio	Low (0.005)	Low (0.008)	Low (0.005)

All clusters show a consistent Revenue Growth of 0.21, indicating that companies within each cluster are experiencing similar growth rates. However, EPS Growth is low across all clusters, suggesting that while the companies are expanding their revenues, their profit growth remains relatively modest. Additionally, the Debt-to-Equity Ratio is consistently low, ranging from 0.005 to 0.008, which indicates that the companies are not highly leveraged. The detailed classification of each company can be found in Appendix 6.2, where the results of each clustering method are compared. This appendix provides a breakdown of the specific companies assigned to each cluster, offering further overviews into the characteristics of the companies within each grouping.

5.4 INVESTMENT RECOMMENDATIONS BASED ON CLUSTERS

In this section, the investment recommendations will be made based on the cluster analysis using the K-Means method. The cluster characteristics highlighted in earlier sections serve as the basis for matching the investment profiles of conservative, moderate, and risky investors. Each investor type will be matched with a cluster that aligns with their risk tolerance and investment goals, providing them with a more tailored approach to their investment strategies. This section will present a detailed analysis of each cluster, explaining the rationale behind the suggested investments based on the investor profile.

5.4.1 Conservative Profile

The conservative investor is recommended to consider investing in **cluster 1**. This cluster is appropriate because of the following factors, such as the Beta of 0.30, which aligns with the conservative investor preference to minimize exposure to market fluctuations. The moderate volatility of 0.21 suggests that the stock prices of the companies in cluster 1 are not excessively unstable, which is suitable for conservative investors seeking stability. A P/E ratio of 36 implies that the companies are reasonably valued, which may signal stability, making it attractive to conservative investors. Additionally, a moderate dividend yield of 0.10 provides a steady income, which is particularly appealing to conservative investors who prioritize security and steady returns. The low market capitalization of 2.10^9 indicates that the companies are small in size, which may be viewed as an opportunity for growth but with additional risks, something that the conservative investor should carefully consider. The high revenue growth of 0.21 indicates that the companies in cluster 1 are expanding, but conservative investors may weigh this against the potential risks. The low EPS growth of 0.007 and the low debt-to-equity ratio of 0.008 suggest that the companies are financially stable, making them a relatively safe option for conservative investors.

5.4.2 Moderate Profile

The moderate investor is recommended to consider investing in **cluster 0**. This cluster is suitable because of the following factors as the Beta of 0.41 suggests moderate market risk, which is appropriate for an investor willing to accept some risk in exchange for potential returns. A higher volatility of 0.56 can be acceptable for a moderate investor who is comfortable with larger price fluctuations in exchange for greater potential returns. The low P/E ratio of 35 suggests that the companies are reasonably valued, which could provide potential for growth. The moderate dividend yield of 0.10 is attractive for a moderate investor seeking a balance between growth and income. The low market cap of 1.10^9 indicates that the companies are small-sized, which may offer growth opportunities, though they also come with additional risks, which a moderate investor may find acceptable. The high revenue growth of 0.21 suggests that the companies in cluster 0 are expanding, which is appealing to moderate investors seeking a balance between risk and return. The low EPS growth of 0.004 and low debt-to-equity ratio of 0.005 indicate that while growth may be slow, the companies are financially stable, making the cluster suitable for moderate investors.

5.4.3 Risky Profile

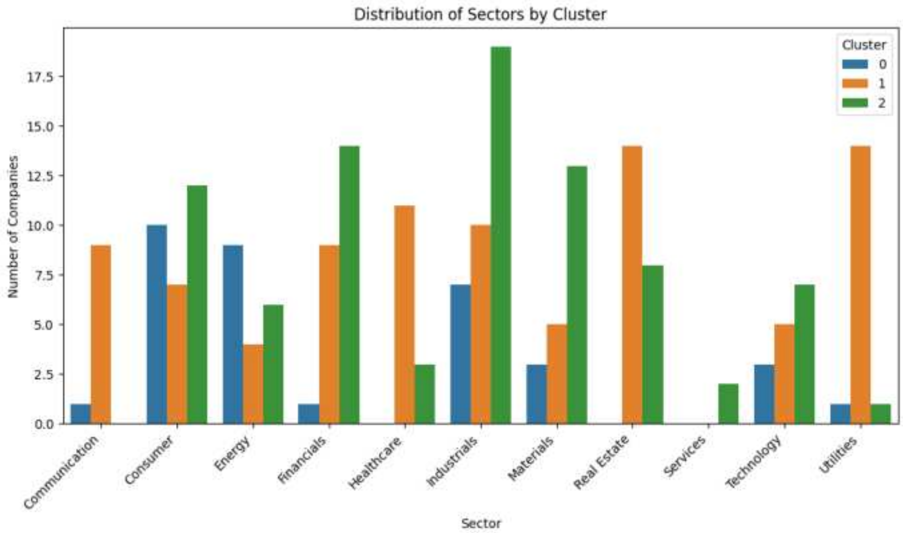
The risky investor is recommended to consider investing in **cluster 2**. This cluster is suitable for the following reasons the Beta of 0.42 indicates moderate market risk, but it is higher than that in cluster 1, which may be suitable for an investor seeking higher returns despite increased risks. The higher volatility of 0.33 suggests larger price fluctuations, which are acceptable to a risky investor willing to tolerate more risk in exchange for potential rewards. The low P/E ratio of 36 may indicate undervaluation, which is appealing to a risky investor seeking companies with high growth potential. A moderate dividend yield of 0.08 is acceptable for a risky investor who prioritizes capital appreciation over income generation. The low market capitalization of 1.10^9 suggests higher growth potential, but also higher risk, which aligns with the profile of a risky investor. The high revenue growth of 0.21 signals that the companies in cluster 2 are in an expansion phase, which could lead to substantial returns over time, appealing to a risky investor. Despite the low EPS growth of 0.006, the low debt-to-equity ratio of 0.005 suggests that the companies are not highly leveraged, providing a relatively stable foundation for risky investors who focus on growth potential.

5.5 INVESTOR PROFILE MATCHING BASED ON CLUSTER CHARACTERISTICS

In conclusion, the conservative investor should invest in **cluster 1**, due to the moderate risk, controlled volatility and stable dividend yield. The moderate investor should invest in **cluster 0**, which offers a balance between risk and return, with high revenue growth and moderate volatility. The risky investor should consider **cluster 2**, due to the higher potential for growth, despite the higher volatility and risks.

Each cluster exhibits characteristics that align with specific investor profiles based on trade-offs between risk, reward and other financial indicators, allowing investors to choose

Figure 12 – Distribution of Sectors by cluster



Source: By the Author, 2024

the cluster that best suits their preferences and risk tolerance.

5.6 ANALYSIS OF FINANCIAL METRICS BY SECTOR

To analyze the distribution of sectors within each cluster, the sectors of the companies were identified by inputting their respective tickers and names into ChatGPT, as detailed in Appendix 6.2. This process did not involve creating a new dataset; rather, it complemented the existing data by adding sector information to the companies already grouped by the K-Means clustering method. Adjustments were made to simplify the sector names for easier grouping and analysis. For instance, detailed subsector names such as "Technology AI" were generalized to "Technology," ensuring that only the primary sectors were retained.

This simplification aimed to streamline the clustering process by reducing the complexity of sectoral classifications, focusing on broader categories rather than detailed subsectors. The resulting dataset allowed for a clearer interpretation of sectoral trends within each cluster, facilitating a more straightforward and intuitive analysis of the clustering results.

With that, the chart on Figure 12 illustrates the distribution of companies across different market sectors for three distinct clusters derived from the clustering algorithm. Each cluster (0, 1 and 2) groups companies with similar financial characteristics based on the selected features. The x-axis represents market sectors, including Communication, Consumer, Energy, Financials, Healthcare, Industrials, Materials, Real Estate, Services, Technology and Utilities, while the y-axis indicates the number of companies in each sector.

The analysis of sector distribution across clusters reveals distinct patterns that highlight how clustering effectively separates companies based on their financial and operational characteristics. cluster 0 exhibits a relatively balanced distribution among several sectors, with particular prominence in the Consumer and Energy industries. It also shows moderate representation in sectors such as Healthcare, Real Estate and Financials. However, its presence is minimal in more specialized sectors, such as Communication, Technology and Utilities. This distribution suggests that companies in cluster 0 are more evenly spread across traditional and consumer-driven industries, but have limited representation in niche or specialized sectors.

In contrast, cluster 1 is notable for its strong concentration in the Real Estate and Technology sectors. This indicates that companies in this cluster share financial characteristics commonly associated with these industries, which are often growth-oriented and innovation-driven. Nevertheless, cluster 1 has relatively low representation in sectors such as Industrials, Materials and Utilities, underscoring its focus on sectors with high potential for growth rather than those tied to industrial production or essential goods and services.

cluster 2, on the other hand, emerges as the largest and most dominant cluster in several key sectors, including Industrials, Materials and Utilities. This highlights a concentration of companies with financial attributes characteristic of resource-driven and industrial sectors. Furthermore, cluster 2 demonstrates significant representation in the Energy and Financial sectors, reflecting its broad coverage of industries tied to essential goods and services. However, its presence in Real Estate and Technology is limited, which stands in stark contrast to the focus observed in cluster 1.

In summary, the distribution of sectors across clusters reveals clear trends that align with specific financial and operational profiles. cluster 0 tends to have a balanced presence across various sectors, cluster 1 emphasizes growth-oriented industries and cluster 2 dominates resource-driven and industrial sectors. This information provides valuable guidance for investors looking to identify sector-specific patterns and refine their investment strategies accordingly.

6 CONCLUSION

This study set out to explore the concepts of machine learning in the financial market context. The primary objective was to develop a summary of companies and their characteristics using unsupervised learning, providing a reference for stock market investors. The analysis successfully divided companies into clusters based on their financial and market attributes, offering valuable informations for investment strategies.

One of the key findings was that companies do not exhibit significant variation in cluster membership, contrary to initial expectations. This suggests a certain stability in the grouping patterns, which enhances the interpretability and reliability of the clusters. Among the methods tested, K-means emerged as the most effective algorithm, demonstrating superior clustering performance with strong evaluation scores.

The results confirm that unsupervised learning can effectively identify meaningful patterns in financial data, allowing investors to make more informed decisions. These findings not only highlight the potential of clustering techniques for financial analysis but also pave the way for future studies to explore more complex or dynamic features in the financial market.

Overall, the algorithms employed showed quite promising results, with a particular emphasis on K-Means, which, although not perfect, demonstrated good cohesion and coherence in identifying patterns among assets. However, even within the financial market context, it is important to consider that there are simplifications in problem modeling that, in a real-world environment, would lead to additional costs and greater complexities, depreciating the results obtained. One of these challenges is the impact that news from the press and actions of individuals with significant economic power have on stock prices. Another example is uncontrollable and large-scale events, such as the COVID-19 pandemic. Furthermore, issues such as unfair competition between companies can also affect the final results.

Based on the studies conducted throughout this work, it is concluded that it is possible to utilize unsupervised learning techniques in the financial market context and achieve interesting results. However, it is clear that in real-world environments, the challenges are greater, requiring additional effort to experiment with more robust techniques.

6.1 CHALLENGES FACED

One of the main challenges encountered during the project was the execution environment and the machine used. Google Colab is an internet-connected environment, so processing a large amount of data often took about 1h hours to download and there was no guarantee of uninterrupted execution. As a result, data loading was frequently interrupted during execution, requiring the environment to be restarted.

Another challenge was finding a reliable data source with the desired number of companies. Considerable time was spent searching for available datasets from sources like Kaggle and others, with approximately one week dedicated to identifying and selecting datasets that met the requirements for this study. However, many of them were not user-friendly or lacked relevant information. Consequently, Yahoo Finance was chosen as the primary data source. While it is a good library, many tickers were not found due to name discrepancies, such as Brazilian tickers that often end with numbers, while others do not.

The concept of machine learning relies heavily on the exhaustive exploration of data to draw conclusions that are faithful to real-world scenarios. Therefore, the larger the dataset, over 500 companies, the greater the chances of recognizing classification patterns and accurately reflecting reality.

6.2 FUTURE WORK

An interesting idea for future work would be to use more robust techniques, such as fuzzy logic. Fuzzy logic is particularly well-suited for problems involving uncertainty and ambiguity, as it allows for the representation of partial membership in clusters rather than forcing data points into strictly defined categories. This characteristic makes it a technically superior choice in scenarios where boundaries between groups are not clearly defined, such as in financial markets where company characteristics often overlap.

Moreover, fuzzy clustering methods, like Fuzzy C-Means (FCM)(GUPTA, 2021), provide additional information by assigning membership probabilities to each cluster, enabling a more nuanced interpretation of the data. These qualities make fuzzy logic a powerful tool for understanding complex systems with inherent vagueness.

However, due to limited exposure to this methodology during undergraduate studies and the complexity of implementing such algorithms within the scope and timeline of this project, fuzzy logic was not explored. Nevertheless, its potential to improve clustering accuracy and interpretability makes it a promising avenue for future research.

Another point to consider is the possibility of modeling the problem with greater depth. Depth in this context refers to incorporating additional layers of analysis and information that go beyond traditional numerical or financial metrics. This could involve leveraging sentiment analysis of news related to stock market companies, capturing qualitative aspects such as public perception and market sentiment. These information add context to the data, enriching the analysis by including external factors that influence stock performance.

Furthermore, techniques for predicting attributes, such as forecasting future earnings or estimating volatility, can enhance the dataset by providing a forward-looking perspective. An example of deep modeling would be analyzing Twitter posts containing stock tickers to assess their sentiment and their potential impact on stock prices, as was done in Brolesi e Bueno (2022). By integrating these complex and dynamic data sources, the resulting model becomes more holistic, offering a more complete and realistic representation of the factors affecting the stock market.

Another example of system modeling that could be refined is the use of a different time interval. It is very common for real-world investors to use additional information beyond what was used in this project to make decisions, such as analyzing a specific time window like a 3-month period or even over a year. This new interval could be incorporated into the observation of attributes to provide more context for decision-making.

Exploring more complex attributes would also be beneficial. Complex attributes, such as macroeconomic indicators (e.g., GDP growth, inflation rates), company-specific financial ratios (e.g., debt-to-equity ratio, earnings growth) and sectoral trends, provide a more nuanced understanding of the factors influencing stock performance. These attributes often capture intricate dynamics that simpler metrics, like return and volatility, may overlook. For example, incorporating sentiment analysis from news or social media can reveal market sentiment, while analyzing supply chain data can provide description of the operational risks.

This aspect was examined to a certain extent in the project, as the initial investigation contained different information than the final version. However, a deeper study would involve not only identifying these attributes but also understanding their interactions and relative importance within the modeling framework. For instance, this could mean performing feature importance analysis, conducting experiments with derived features (e.g., trend indicators, seasonality), or integrating time-series data to capture temporal dependencies.

By investigating these aspects in greater depth, the modeling process could move beyond surface-level patterns, enabling a more comprehensive representation of real-world complexities and improving the accuracy and reliability of the results. This deeper analysis can uncover hidden relationships and nuances, allowing for more informed decision-making and better predictions in practical scenarios.

Finally, a valuable direction for future work would be to develop an interactive visualization platform that allows for a more intuitive exploration of the data and clustering analysis results. This platform could include dynamic charts and customizable dashboards, enabling investors to visually examine patterns and trends, making it easier to understand the outcomes.

A promising idea for future work is integrating a Large Language Model (LLM) (SERVICES, 2024) into the platform to enable interactive conversations about the data. Users could ask natural language questions, such as “Which companies are in the most stable cluster?” or “What factors influenced these clusters?” The LLM would provide detailed explanations, visualize data and suggest investment opportunities based on analyses. Finetuning the model with domain-specific data and designing workflows to interact with clustering algorithms would ensure reliable and user-friendly outputs, making advanced analytics accessible to all investors.

BIBLIOGRAPHY

ANBIMA. **Perfil de investidor: o que é e como descobrir o seu**. 2023. Accessed: 2024-09-02. Available on: https://comoinvestir.anbima.com.br/noticia/perfil-de-investidor-o-que-e-e-como-descobrir-o-seu/.

BROLESI, F. F.; BUENO, A. C.

Análise de sentimentos e impacto de postagens do Twitter na bolsa de valores brasileira — Universidade de São Paulo (USP), 2022. Trabalho apresentado para obtenção do título de especialista em Data Science e Analytics.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics - Theory and Methods**, v. 3, n. 1, p. 1–27, 1974.

CRISTIANINI, N.; RICCI, E. Support vector machines. In: ______. **Encyclopedia of Algorithms**. Boston, MA: Springer US, 2008. p. 928–932. ISBN 978-0-387-30162-4. Disponível em: https://doi.org/10.1007/978-0-387-30162-4_415.

D, L. et al. Stock recommendation system for better investment plan. In: **2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)**. [S.l.: s.n.], 2024. p. 1–6.

FERNANDO, J. **Dictionary**. 2024. Investopedia, atualizado em 18 set. 2024. Accessed: 10 nov. 2024: Available on: https://www.investopedia.com/financial-term-dictionary-4769738.

FIGUEIREDO, C. C. et al. Uma análise preliminar da dinâmica volatilidade-retorno de ações das grandes empresas utilizando técnicas de aprendizado não-supervisionado. In: **Proceedings of CNMAC 2024**. [S.l.]: SBMAC, 2025. v. 11, n. 1, p. Resumos.

FIGUEIREDO, C. C. de. **Clusterização de Empresas Utilizando PCA**. 2023. Last updated: 2023-12-14, Accessed: 2024-09-09. Available on: https://colab.research.google.com/drive/18PzCR2wWbXq6Lm8n5HB2H46x9bLyPRSH?usp=sharing.

GAMBIM, M. et al. Uma estratégia para alocação de carteira de ações usando algoritmos de aprendizado de máquina e regras fuzzy. In: **Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)**. [S.l.: s.n.], 2023. p. 1195–1209. ISSN 2763-9061.

GENARO, A. d.; ASTORINO, P. **Um Tutorial sobre o Método Generalizado dos Momentos (GMM) em Finanças**. [S.l.]: Associação Nacional de Pós-Graduação e Pesquisa em Administração, 2022. e210287 p.

GUPTA, A. Fuzzy c-means clustering (fcm) algorithm. **Geek Culture**, jun. 2021. Available on: https://geekculture.com/fuzzy-c-means-clustering-fcm-algorithm.

HALKIDI, M. Hierarchial clustering. In: ______. **Encyclopedia of Database Systems**. Boston, MA: Springer US, 2009. p. 1291–1294. ISBN 978-0-387-39940-9. Disponível em: https://doi.org/10.1007/978-0-387-39940-9_604.

IBM. **K-Means Clustering**. 2024. Published: 26 June 2024, Contributors: Eda Kavlakoglu, Vanna Winland. Disponível em: https://www.ibm.com/topics/k-means-clustering.

JIN, X.; HAN, J. K-means clustering. In: _____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 563–564. ISBN 978-0-387-30164-8.

KIERSZTYN, A. et al. Classification of companies based on fuzzy levels of innovation. In: **2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)**. [S.l.: s.n.], 2022. p. 1–5.

KUHN, B. **yfinance**. 2024. Accessed: 2024-10-17. Available on: https://pypi.org/project/yfinance/.

LIBERTO, D. Investopedia glossary. **Investing Quantitative Analysis**, 2024. Updated June 06, 2024. Available on: https://www.investopedia.com/financial-term-dictionary-4769738.

MACKIEWICZ, A.; RATAJCZAK, W. Principal components analysis (pca). **Computers & Geosciences**, v. 19, n. 3, p. 303–342, 1993. ISSN 0098-3004.

MAGALHAES, W. **S&P 500: o que é e como funciona esse índice norte-americano**. 2024. Accessed: 2024-09-02. Available on: https://www.remessaonline.com.br/blog/sep-500-o-que-e/.

MANCHEV, J.; MIRCHEV, M.; MISHKOVSKI, I. Classification of companies using graph neural networks. In: **2024 47th MIPRO ICT and Electronics Convention (MIPRO)**. [S.l.: s.n.], 2024. p. 1087–1092.

NOBLE, J. **Hierarchical Clustering**. 2024. Published: 05 August 2024, Available on: https://www.ibm.com/think/topics/hierarchical-clustering.

PEDRIALI, R. A.; DESTER, P. S. Clusterização de empresas da bovespa. In: **XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT 2021**. [S.l.: s.n.], 2021.

REYNOLDS, D. Gaussian mixture models. In: _____. **Encyclopedia of Biometrics**. Boston, MA: Springer US, 2009. p. 659–663. ISBN 978-0-387-73003-5.

SANTANDER. **Mercado financeiro: o que é e como funciona**. 2024. Accessed: 2024-09-02. Available on: https://www.santander.com.br/blog/mercado-financeiro-o-que-e-como-funciona.

SERVICES, A. W. **What is a Large Language Model?** 2024. Accessed: 2024-01-30. Available on: https://aws.amazon.com/what-is/large-language-model/.

ZHU, S. et al. Stock index prediction based on principal component analysis and machine learning. In: **2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)**. [S.l.: s.n.], 2020. p. 246–249.

APPENDIX A – TABLE WITH COMPANY NAME WITH ASSOCIATED TICKER

Table 4 – Company Name with associated Ticker

Company Name	Symbol
Arizona Public Service	PNW
Park Electrochemical Corporation	PKE
Danaher Corporation	DHR
Teck Resources Limited	TECK
Cummins	CMI
Apollo Tactical Income Fund Inc.	AIF
Eli Lilly and Company	LLY
Chevron Corporation	CVX
Darden Restaurants	DRI
Clorox	CLX
Argan, Inc.	AGX
Ambev S.A.	ABEV
Mueller Water Products, Inc.	MWA
Host Hotels & Resorts, Inc.	HST
Equity Lifestyle Properties, Inc.	ELS
Myers Industries, Inc.	MYE
W. P. Carey Inc.	WPC
Zimmer Biomet Holdings, Inc.	ZBH
Omnicom Group Inc.	OMC
Armada Hoffler Properties, Inc.	AHH
Trane Technologies plc	TT
AbbVie Inc.	ABBV
Ameren Corporation	AEE
IDT Corporation	IDT
Bunge	BG
RPM International	RPM
Movado Group, Inc.	MOV
Crown Castle	CCI
Standex International Corporation	SXI
DTE Energy	DTE
Minerals Technologies Inc.	MTX

Company Name	**Symbol**
Fidelity National Financial Inc.	FNF
Shutterstock	SSTK
Packaging Corporation Of America	PKG
United Rentals, Inc.	URI
Costamare	CMRE
Ardmore Shipping Corporation	ASC
Canadian National Railway	CNI
Illinois Tool Works Inc.	ITW
Vistra Corp.	VST
Avnet Inc	AVT
ONE Gas, Inc.	OGS
Global Partners LP	GLP
Advance Auto Parts, Inc.	AAP
Prologis, Inc.	PLD
Sherwin-Williams	SHW
Avangrid Inc.	AGR
Getty Realty Corp.	GTY
SunCoke Energy, Inc.	SXC
Pembina Pipeline	PBA
Assured Guaranty Ltd.	AGO
Meritage Homes Corporation	MTH
Kadant	KAI
Motorola Solutions, Inc.	MSI
Accenture plc	ACN
Watsco, Inc.	WSO
Harmony Gold Mining Company Limited	HMY
Safe Bulkers Inc.	SB
Huntsman Corporation	HUN
APA Corporation	APA
Telkom Indonesia	TLK
Telekomunikasi Indonesia (Persero) Tbk	TLK
CVR Partners	UAN
Ingersoll-Rand plc	IR
GlaxoSmithKline	GSK
Vipshop Holdings Limited	VIPS
Zoetis Inc.	ZTS
Sensient Technologies	SXT

Company Name	**Symbol**
MPLX LP	MPLX
TFI International Inc.	TFII
Unitil Corporation	UTL
Entergy Corporation	ETR
Camden Property Trust	CPT
First American Financial Corporation	FAF
Gerdau S.A.	GGB
Invesco Ltd.	IVZ
D. R. Horton	DHI
Lincoln National Corporation	LNC
Ladder Capital Corp	LADR
UWM Holdings Corporation	UWMC
Apple Hospitality REIT Inc.	APLE
Atmos Energy Corporation	ATO
Taiwan Semiconductor Manufacturing Company Ltd.	TSM
Commercial Metals Company	CMC
Brunswick Corporation	BC
Honeywell International, Inc.	HON
World Kinect Corporation	WKC
H&R Block Inc.	HRB
Coca-Cola FEMSA, S.A.B. De C.V.	KOF
U.S. Physical Therapy, Inc.	USPH
Reinsurance Group of America	RGA
Regency Centers Corporation	REG
Smith & Nephew	SNN
The Cooper Companies	COO
Wipro Limited	WIT
Eaton Corporation plc	ETN
Williams-Sonoma, Inc.	WSM
POSCO	PKX
Terreno Realty Corporation	TRNO
EPR Properties	EPR
TransUnion	TRU
Crown Holdings	CCK
Procter & Gamble	PG
Cameco	CCJ
Newell Rubbermaid	NWL

Company Name	**Symbol**
Dolby Laboratories	DLB
International Game Technology	IGT
CNA Financial	CNA
UniFirst Corporation	UNF
Civeo Corporation	CVEO
Unilever PLC	UL
CSX Corporation	CSX
Peabody Energy	BTU
Ameriprise Financial, Inc.	AMP
WEC Energy Group, Inc.	WEC
Oxford Industries Inc.	OXM
Ashland Inc.	ASH
Ametek Inc.	AME
Trinity Industries Inc.	TRN
Becton Dickinson	BDX
Whitestone REIT	WSR
Snap-on	SNA
Apollo Global Management, LLC	APO
Leidos	LDOS
American Tower Corporation	AMT
Sonic Automotive	SAH
Helmerich & Payne Inc.	HP
Eletrobras	EBR
Colgate-Palmolive	CL
Ultrapar Participações S.A.	UGP
Weyerhaeuser Company	WY
Telefônica Brasil S.A.	VIV
FactSet Research Systems Inc.	FDS
Starwood Property Trust, Inc.	STWD
Vici Properties Inc.	VICI
Eagle Materials Inc.	EXP
CMS Energy	CMS
Masco Corporation	MAS
Kimco Realty	KIM
Harley-Davidson, Inc.	HOG
Adecoagro S.A.	AGRO
Baytex Energy	BTE

Company Name	Symbol
Bristol-Myers Squibb	BMY
Allison Transmission Holdings, Inc.	ALSN
Rogers Communications	RCI
ITT Educational Services Inc.	ESI
Radian Group	RDN
ARC Document Solutions, Inc.	ARC
Global High Income Fund, Inc.	GHI
Philippine Long Distance Telephone Company	PHI
Knight Transportation	KNX
National Oilwell Varco	NOV
Sun Life Financial	SLF
Thor Industries, Inc.	THO
Brady Corporation	BRC
Royal Caribbean Group	RCL
Equifax Inc.	EFX
Comfort Systems USA	FIX
Phillips 66	PSX
Deluxe Corporation	DLX
RPC Inc.	RES
Oshkosh Corporation	OSK
Idacorp Inc.	IDA
Saul Centers Inc.	BFS
Scorpio Tankers	STNG
Group 1 Automotive Inc.	GPI
Pentair, Ltd.	PNR
Ares Management, L.P.	ARES
Intercontinental Exchange, Inc.	ICE
Telus Corporation	TU
Martin Marietta Materials Inc.	MLM
Artisan Partners Asset Management Inc.	APAM
Sensata Technologies Holding N.V.	ST
Orange S.A.	ORAN
Thermo Fisher Scientific Inc.	TMO
Alliant Energy Corporation	LNT
Morgan Stanley	MS
Alibaba Group Holding Ltd.	BABA
Rayonier	RYN

Company Name	**Symbol**
Goldman Sachs Group Inc.	GS
The Timken Company	TKR
Steris Corporation	STE
Camping World	CWH
Caterpillar Inc.	CAT
PVH Corp.	PVH
Xenia Hotels & Resorts, Inc.	XHR
Sociedad Química y Minera	SQM
WESCO International, Inc.	WCC
Nike, Inc.	NKE
Black Hills Corporation	BKH
ConocoPhillips	COP
Sealed Air	SEE
Arch Coal Inc	ARCH
Carriage Services	CSV
Atkore International Group Inc.	ATKR
Evercore Partners Inc.	EVR
American Homes 4 Rent	AMH
Corning Inc.	GLW
New Mountain Capital	NMFC
CRH plc	CRH
Portland General Electric Company	POR
Marriott Vacations Worldwide Corporation	VAC
Agnico Eagle Mines Limited	AEM
Nelnet	NNI
SM Energy	SM
AerCap Holdings N.V.	AER
General Dynamics Corporation	GD
Abbott Laboratories	ABT
Hyatt Hotels Corporation	H
Targa Resources Corp.	TRGP
Gildan Activewear Inc.	GIL
Acadia Realty Trust	AKR
Xylem Inc.	XYL
Textron Inc	TXT
Quest Diagnostics Incorporated	DGX
Miller Industries Inc.	MLR

Company Name	Symbol
Quanta Services Inc.	PWR
DiamondRock Hospitality Company	DRH
Halliburton Company	HAL
Murphy Oil Corporation	MUR
Marathon Oil Corporation	MRO
Infosys Limited	INFY
Devon Energy	DVN
Mueller Industries, Inc.	MLI
IDEX Corporation	IEX
NextEra Energy	NEE
Iron Mountain Inc.	IRM
U.S. Steel Corporation	X
International Paper Co.	IP
Service Corporation International	SCI
American Financial Group Inc.	AFG
Antero Midstream Partners LP	AM
Telefónica S.A.	TEF
Murphy USA Inc.	MUSA
Crane Co.	CR
Ecopetrol S.A.	EC
Juniper Networks	JNPR
Wyndham Hotels & Resorts, Inc.	WH
Century Communities	CCS
Belden	BDC
MetLife, Inc.	MET
Life Time Fitness	LTM
ArcelorMittal	MT
ONEOK, Inc.	OKE
Mid America Apartment Communities Inc.	MAA
EOG Resources, Inc.	EOG
Allegion Public Limited Company	ALLE
Installed Building Products, Inc.	IBP
RLI Corp	RLI
Genpact Limited	G
Hexcel Corporation	HXL
Redwood Trust Inc	RWT
PennyMac Loan Services	PFSI

Company Name	**Symbol**
Eastgroup Properties Inc.	EGP
Kroger	KR
Praxair, Inc.	PX
Highwoods Properties Inc.	HIW
Physicians Realty Trust	DOC
Westlake Chemical Partners LP	WLKP
American Electric Power Company	AEP
Chunghwa Telecom	CHT
Dr. Reddy's Laboratories	RDY
SandRidge Energy	SD
J. M. Smucker Company	SJM
American States Water Co.	AWR
Aramark	ARMK
Chimera Investment Corporation	CIM
Canadian Pacific Railway	CP
RenaissanceRe Holdings Ltd.	RNR
Hess Corporation	HES
Amphenol Corporation	APH
Employers Holdings, Inc.	EIG
TransAlta Corporation	TAC
NRG Energy	NRG
Chemed Corporation	CHE
CubeSmart	CUBE
Natural Resource Partners LP	NRP
Vulcan Materials Company	VMC
Tanger Factory Outlet Centers, Inc.	SKT
AFLAC Incorporated	AFL
Nucor	NUE
Charles Schwab Corporation	SCHW
Molson Coors Brewing Company	TAP
Greif, Inc.	GEF
Advanced Semiconductor Engineering, Inc.	ASX
Roper Industries	ROP
Union Pacific Corporation	UNP
Universal Insurance Holdings, Inc.	UVE
Ellington Financial LLC	EFC
Newmont Corporation	NEM

Company Name	**Symbol**
Ford Motor Company	F
Target Corporation	TGT
Republic Services	RSG
Prudential plc	PUK
Ralph Lauren Corporation	RL
TriNet Group, Inc.	TNET
Pfizer Inc.	PFE
Hormel Foods Corporation	HRL
X Financial	XYF
Sanofi	SNY
TotalEnergies SE	TTE
UnitedHealth Group Incorporated	UNH
Honda Motor Co., Ltd.	HMC
Olin Corporation	OLN
Dillard's	DDS
Vontier Corporation	VNT
Church & Dwight	CHD
United Microelectronics Corporation	UMC
Dana Holding Corporation	DAN
CNO Financial Group	CNO
Fomento Economico Mexicano, S.A.B. De C.V.	FMX
LabCorp	LH
Zurn Elkay Water Solutions Corporation	ZWS
Hannon Armstrong Sustainable Infrastructure Capital Inc	HASI
MAXIMUS, Inc.	MMS
Ventas, Inc.	VTR
Principal Financial Group	PFG
MasterCard Incorporated	MA
Alexander & Baldwin, Inc.	ALEX
Hyster-Yale Materials Handling, Inc.	HY
Delta Air Lines	DAL
SJW Corp.	SJW
The Coca-Cola Company	KO
Avery Dennison Corporation	AVY
DHT Holdings, Inc.	DHT
Terex Corporation	TEX
Quaker Chemical Corporation	KWR

Company Name	**Symbol**
AstraZeneca Group plc	AZN
Vishay Intertechnology, Inc.	VSH
Polaris Industries	PII
Gold Fields Limited	GFI
B&G Foods	BGS
MSA Safety Incorporated	MSA
Southern Company	SO
Waste Connections, Inc.	WCN
Alexandria Real Estate Equities Inc.	ARE
Donaldson Company	DCI
Core Laboratories	CLB
Tegna Inc.	TGNA
Frontline Ltd.	FRO
Penske Automotive Group	PAG
ResMed	RMD
Teleflex Inc.	TFX
Stantec	STN
NorthWestern Corporation	NWE
Teekay Corporation	TK
ALLETE, Inc.	ALE
Lennox International	LII
Matador Resources Company	MTDR
Cenovus Energy	CVE
Korn Ferry	KFY
EMCOR Group, Inc.	EME
Alexanders Inc.	ALX
Northrop Grumman	NOC
Westinghouse Air Brake Technologies Corporation	WAB
Graphic Packaging Holding Company	GPK
American Express Company	AXP
Lear Corporation	LEA
Navigator Holdings Ltd.	NVGS
Orion Engineered Carbons S.A.	OEC
Coterra	CTRA
Public Storage	PSA
Grupo Aeroportuario del Sureste, S.A.B. de C.V.	ASR
Vale S.A.	VALE

Company Name	**Symbol**
Noah Holdings	NOAH
MDU Resources Group, Inc.	MDU
Arthur J.Gallagher & Co.	AJG
ABM Industries Incorporated	ABM
PennyMac Mortgage Investment Trust	PMT
Eastman Chemical Co.	EMN
Exelon Corporation	EXC
Dover Corporation	DOV
Franklin Resources Inc.	BEN
St. Joe Company	JOE
AptarGroup Inc.	ATR
ENI S.p.A.	E
EnLink Midstream, LLC	ENLC
Essent Group Ltd.	ESNT
Wabash National Corporation	WNC
Chatham Lodging Trust	CLDT
Graco Inc.	GGG
LTC Properties Inc.	LTC
TE Connectivity Ltd.	TEL
Extra Space Storage, Inc.	EXR
TIM S.A.	TIMB
CVR Energy, Inc.	CVI
KB Home	KBH
Enterprise Products Partners L.P.	EPD
Autoliv Inc.	ALV
Sunstone Hotel Investors, Inc.	SHO
Vertiv Holdings Co	VRT
Tennant Company	TNC
Marathon Petroleum Corporation	MPC
USA Compression Partners, LP	USAC
Western Midstream Partners, LP	WES
Turning Point Brands, Inc.	TPB
H.B. Fuller Company	FUL
Msc Industries Direct Co Inc.	MSM
Occidental Petroleum Corporation	OXY
Orix Corporation	IX
Discover Financial	DFS

Company Name	Symbol
Science Applications International Corporation	SAIC
Old Republic International Corporation	ORI
Diageo	DEO
Ubiquiti Inc.	UI
Stifel	SF
Urban Edge Properties	UE
Freeport-McMoRan Inc.	FCX
Norfolk Southern Railway	NSC
MGIC Investment Corporation	MTG
Texas Pacific Land Corporation	TPL
Stepan Company	SCL
Constellation Brands	STZ
Applied Industrial Technologies, Inc.	AIT
Toll Brothers Inc.	TOL
Medtronic Inc.	MDT
Assurant, Inc.	AIZ
Evertec, Inc.	EVTC
Kennametal	KMT
Walker & Dunlop, Inc.	WD
Genie Energy Ltd.	GNE
Brown & Brown	BRO
Takeda Pharmaceutical Company Limited	TAK
Parker Hannifin Corporation	PH
Standard Motor Products	SMP
Westlake Corporation	WLK
The Western Union Company	WU
Fresh Del Monte Produce Inc.	FDP
Verizon Communications Inc.	VZ
Valmont Industries, Inc.	VMI
Tenaris S.A.	TS
Raymond James Financial	RJF
Winnebago Industries, Inc.	WGO
Yiren Digital Ltd.	YRD
General Motors Company	GM
Las Vegas Sands	LVS
AdvanSix Inc.	ASIX
Stewart Information Services Corporation	STC

Company Name	**Symbol**
Matson, Inc.	MATX
ZTO Express (Cayman) Inc.	ZTO
Archrock Inc.	AROC
Rexford Industrial Realty, Inc.	REXR
Pearson PLC	PSO
Simon Property Group	SPG
Lockheed Martin	LMT
Yum China Holdings, Inc.	YUMC
Valero Energy Corporation	VLO
TJX Companies Inc.	TJX
Vaalco Energy, Inc.	EGY
Federal Signal Corporation	FSS
Koppers	KOP
Public Service Enterprise Group	PEG
KE Holdings	BEKE
Sabesp	SBS
Ormat Technologies, Inc.	ORA
Nordic American Tankers Limited	NAT
Curtiss-Wright	CW
Benchmark Electronics	BHE
Main Street Capital Corporation	MAIN
PPL Corporation	PPL
National Presto Industries	NPK
Griffon Corporation	GFF
Simpson Manufacturing Co., Inc.	SSD
Boise Cascade	BCC
Dollar General	DG
Equity Residential	EQR
Aon Corporation	AON
Prudential Financial, Inc.	PRU
The Williams Companies, Inc.	WMB
Regional Management Corp.	RM
FirstEnergy Corp	FE
Ship Finance International Limited	SFL
Nordstrom	JWN
American Eagle Outfitters, Inc.	AEO
FMC Corporation	FMC

Company Name	Symbol
Edison International	EIX
Sempra Energy	SRE
Watts Water Technologies, Inc.	WTS
Grupo Aeroportuario del Pacifico, S.A.B de C.V.	PAC
Avista Corporation	AVA
Novo Nordisk	NVO
Federal Agricultural Mortgage Corporation	AGM
Southwest Gas	SWX
Badger Meter	BMI
Brixmor Property Group	BRX
Novartis	NVS
West Pharmaceutical Services, Inc.	WST
MFS Charter Income Trust	MCR
OGE Energy Corp.	OGE
HCA Holdings, Inc.	HCA
Whirlpool Corporation	WHR
Acushnet Holdings Corp.	GOLF
Ennis, Inc.	EBF
Cohen & Steers	CNS
Voya Financial, Inc.	VOYA
Best Buy	BBY
GeoPark Limited	GPRK
La-Z-Boy	LZB
Haverty Furniture Companies Inc.	HVT
The Buckle	BKE
Plains GP Holdings, L.P.	PAGP
Moody's Corporation	MCO
Tempur Sealy International, Inc.	TPX
WPP plc	WPP
Exxon Mobil Corporation	XOM
The Interpublic Group of Companies Inc.	IPG
Celanese	CE
Alamo Group Inc.	ALG
The Greenbrier Companies, Inc.	GBX
Archer Daniels Midland Co.	ADM
Cabot Corporation	CBT
Pacific Gas and Electric Company	PCG

Company Name	**Symbol**
Tecnoglass Inc.	TGLS
Flowserve Corporation	FLS
National Retail Properties	NNN
Johnson Controls, Inc.	JCI
Agree Realty Corporation	ADC
Salesforce.com	CRM
Toyota Motor Corporation	TM
Cementos Pacasmayo	CPAC
Genuine Parts Company	GPC
The Blackstone Group	BX
Air Products and Chemicals, Inc.	APD
Boston Properties	BXP
ITT Corporation	ITT
Restaurant Brands International	QSR
Lindsay Manufacturing	LNN
STMicroelectronics	STM
Waste Management, Inc.	WM
AT&T Inc.	T
Universal Health Services, Inc.	UHS
Advanced Drainage Systems Inc.	WMS
Select Medical	SEM
Vail Resorts, Inc.	MTN
Progressive Corporation	PGR
Boyd Gaming	BYD
Enbridge, Inc.	ENB
Kinder Morgan	KMI
Teekay Tankers Ltd.	TNK
Global Ship Lease, Inc	GSL
Baker Hughes	BKR
Kilroy Realty Corporation	KRC
STAG Industrial, Inc.	STAG
Berry Plastics	BERY
Southwest Airlines	LUV
Huntington Ingalls Industries, Inc.	HII
Pulte Homes	PHM
California Water Service Group	CWT
Robert Half International	RHI

Company Name	Symbol
Build-A-Bear Workshop	BBW
Reliance Steel & Aluminum Co.	RS
Materion Corporation	MTRN
Plains All American Pipeline	PAA
Tootsie Roll Industries Inc.	TR
Visa Inc.	V
Acuity Brands, Inc.	AYI
John Deere	DE
A. O. Smith Corporation	AOS
UDR, Inc.	UDR
WisdomTree, Inc.	WT
John Bean Technologies Corporation	JBT
Western Asset Intermediate Muni Fund Inc.	SBI
ManpowerGroup	MAN
Manulife Financial Corporation	MFC
CF Industries	CF
Humana Inc.	HUM
Tencent Music Entertainment Group	TME
Douglas Dynamics, Inc.	PLOW
Jabil Circuit Inc.	JBL
CenterPoint Energy	CNP
DRDGOLD Limited	DRD
HEICO Corporation	HEI
Tapestry, Inc.	TPR
Arbor Realty Trust, Inc.	ABR
Stryker Corporation	SYK
Carter's, Inc.	CRI
NextEra Energy Partners	NEP
Autohome Inc.	ATHM
Kimberly-Clark	KMB
SAP SE	SAP
Cousins Properties	CUZ
Piper Sandler	PIPR
Walmart Inc.	WMT
CTS Corporation	CTS
One Liberty Properties, Inc.	OLP
LIN Media	LIN

Company Name	**Symbol**
Home Depot, Inc.	HD
Global Payments Inc.	GPN
Ryman Hospitality Properties	RHP
Cemex	CX
Broadridge Financial Solutions	BR
Chesapeake Utilities	CPK
Kinross Gold	KGC
Omega Healthcare Investors, Inc	OHI
Telecom Argentina S.A.	TEO
Hartford Financial Services Group Inc.	HIG
BorgWarner	BWA
Carlisle Companies	CSL
Essex Property Trust, Inc.	ESS
Federal Realty Investment Trust	FRT
Site Centers	SITC
Dick's Sporting Goods	DKS
FS KKR Capital Corp.	FSK
Spectrum Brands	SPB
Unum Group	UNM
Copa Holdings	CPA
Fresenius Medical Care AG & Co. KGAA	FMS
Armstrong World Industries, Inc.	AWI
Kohl's	KSS
Kellanova	K
A10 Networks, Inc.	ATEN
First Industrial Realty Trust, Inc.	FR
Loews Corporation	L
American Assets Trust, Inc.	AAT
Emerson Electric Co.	EMR
LyondellBasell	LYB
Canadian Natural Resources	CNQ
Axis Capital Holdings Limited	AXS
National Health Investors Inc.	NHI
Quanex Building Products Corporation	NX
Anheuser-Busch Inbev SA/NV	BUD
TC Energy Corporation	TRP
Dow Chemical Company	DOW

Company Name	**Symbol**
Owens Corning	OC
Ingredion Incorporated	INGR
Nomura Holdings	NMR
El Paso Electric Co.	EE
Marsh & McLennan Companies Inc.	MMC
Chubb Limited	CB
Danaos Corporation	DAC
Brink’s	BCO
Ethyl Corporation	NEU
PPG Industries, Inc.	PPG
Silvercorp Metals	SVM
Suncor Energy	SU
SK Telecom	SKM
HCC Insurance Holdings, Inc.	HCC
Warrior Met Coal, Inc.	HCC
Digital Realty	DLR
Johnson & Johnson	JNJ
Empire State Realty Trust, Inc.	ESRT
Synnex	SNX
TD SYNNEX Corporation	SNX
The Hershey Company	HSY
Sysco	SYY
Rio Tinto Group	RIO
Schlumberger	SLB
United Parcel Service, Inc.	UPS
BlackRock	BLK
Bell Canada	BCE
American Water Works Company, Inc.	AWK
Pebblebrook Hotel Trust	PEB
Copel	ELP
PNM Resources	PNM
Air Lease Corporation	AL
CEMIG	CIG
Albany International Corp	AIN
Two Harbors Investment Corp.	TWO
Vitamin Cottage Natural Grocers	NGVC
Rockwell Automation	ROK

Company Name	Symbol
Cigna	CI
Booz Allen Hamilton	BAH
International Flavors & Fragrances Inc.	IFF
Consolidated Edison	ED
Lithia Motors	LAD
International Business Machines Corporation	IBM
Rollins Inc.	ROL
EnerSys	ENS
CVS Health	CVS
General Mills, Inc.	GIS
McCormick & Company, Inc.	MKC
Welltower Inc.	WELL
Cato Corporation	CATO
RLJ Lodging Trust	RLJ
Wolverine World Wide, Inc.	WWW
Insperity, Inc.	NSP
Hercules Technology Growth Capital, Inc.	HTGC
Magna International Inc.	MGA
Advent Claymore Convertible Securities and Income Fund	AVK
North American Energy Partners Inc.	NOA
Dominion Resources	D
Dorian LPG Ltd.	LPG
Arcos Dorados Holdings Inc.	ARCO
NiSource	NI
Nvidia Corporation	NVDA
National Grid plc	NGG
Esco Technologies Inc.	ESE
Affiliated Managers Group	AMG
Steelcase	SCS
Energy Transfer LP	ET
Carpenter Technology Corporation	CRS
Sonoco	SON
Thomson Reuters Corporation	TRI
Agilent Technologies Inc.	A
Wheaton Precious Metals Corp.	WPM
Ritchie Bros. Auctioneers	RBA
Primerica, Inc.	PRI

Company Name	**Symbol**
AvalonBay Communities, Inc.	AVB
KKR	KKR

APPENDIX B – ANALYSIS OF TICKER CLASSIFICATION INTO CLUSTERS FOR EACH METHOD

The analysis of ticker classification across different clustering methods provides an overview into how companies are grouped based on their financial attributes. This appendix presents a detailed evaluation of the classification results for each method, highlighting the differences in cluster assignments.

Furthermore, the classification of each cluster by the Hierarchical Clustering can be analyzed in Table 5. The table presents the characteristics of the clusters generated by hierarchical clustering, focusing on key financial attributes. cluster 0 and cluster 1 exhibit moderate Beta values, with cluster 0 slightly higher (0.44) than cluster 1 (0.35), indicating moderate market risk. Volatility is higher in cluster 0 (0.50), while cluster 1 shows moderate volatility (0.27), suggesting that cluster 0 contains more unstable companies. Both clusters have high P/E Ratios, around 35-36, indicating that the companies in both clusters are generally valued highly relative to their earnings.

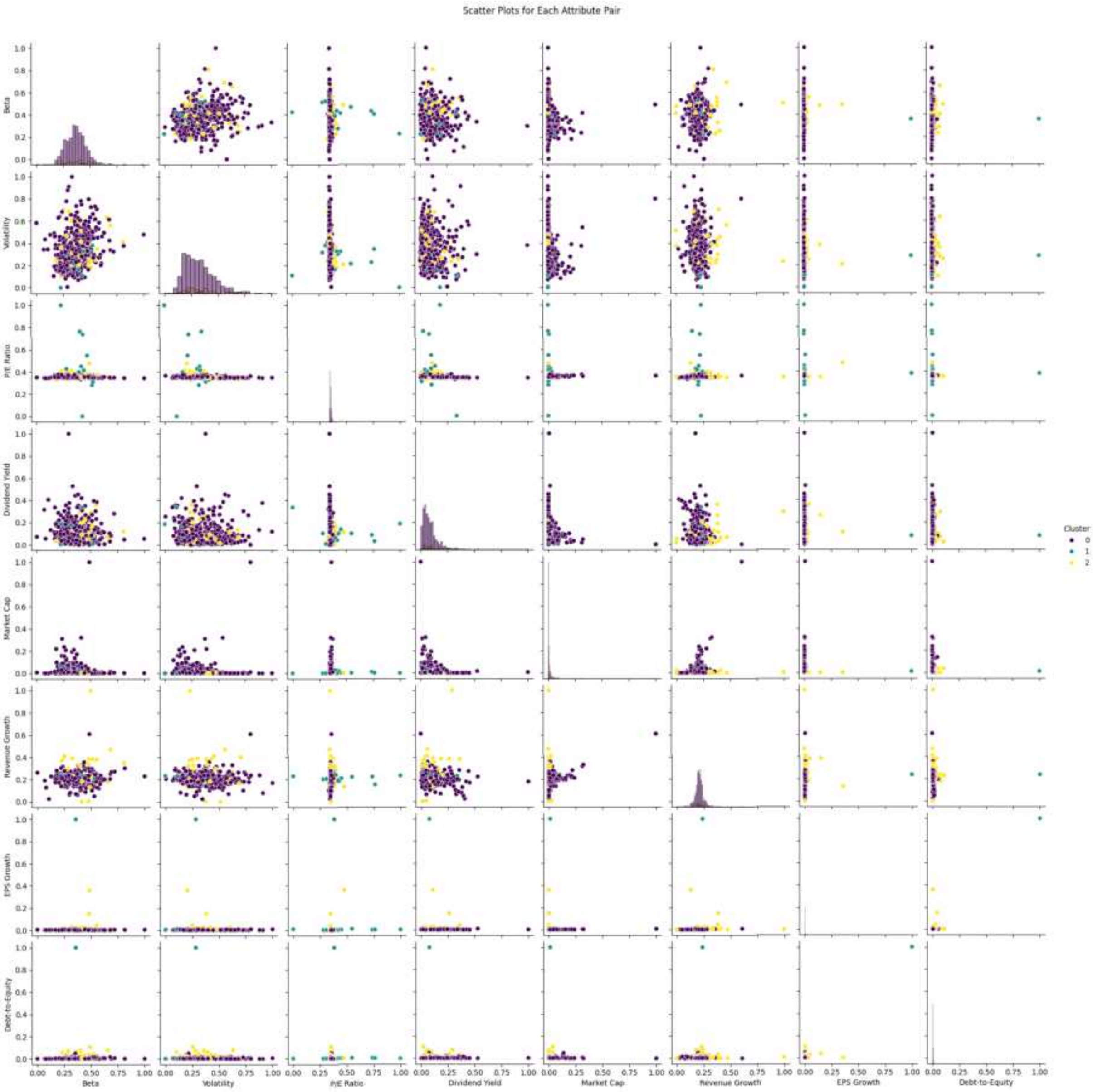
Both clusters demonstrate similar Revenue Growth (0.21), indicating that the companies are growing at the same rate. However, EPS Growth remains low in both clusters, signaling that profit growth is minimal. Finally, Debt-to-Equity Ratios are low in both clusters, pointing to companies that are not highly leveraged. This breakdown shows the main financial traits of companies in each cluster, helping to understand how these clusters are grouped based on financial performance.

Table 5 – Cluster Characteristics Based on Hierarchical Clustering

Attribute	**cluster 0**	**cluster 1**
Beta	Moderate (0.44)	Moderate (0.35)
Volatility	High (0.50)	Moderate (0.27)
P/E Ratio	High (35)	High (36)
Dividend Yield	Moderate (0.07)	Moderate (0.10)
Market Cap	Low ($1 \cdot 10^9$)	Moderate ($2 \cdot 10^9$)
Revenue Growth	High (0.21)	High (0.21)
EPS Growth	Low (0.004)	Low (0.007)
Debt-to-Equity Ratio	Low (0.005)	Low (0.007)

The third method applied was the GMM with three clusters. As shown in Figure 13, the scatter plot matrix displays the dispersion of attributes for the GMM clustering method. The visualization suggests that although some attributes—such as Beta and Volatility, as well as Beta and Dividend Yield—show minor distinctions between clusters, there is considerable overlap across most attribute pairs. Notably, attributes like Market Cap and Debt-to-Equity Ratio exhibit substantial overlap, indicating that GMM has difficulty in creating well-separated clusters based on these features. This lack of clear separation may limit the method's effectiveness in providing distinct groupings for investment strategies.

Figure 13 – Dispersion for Each Pair of GMM Attributes



Source: By the Author, 2024

This analysis underscores potential limitations in the dataset's ability to achieve clear segmentation. The lack of well-defined clusters suggests that additional feature engineering or dimensionality reduction techniques may be necessary to enhance the clustering process and improve classification accuracy.

In this sense, the classification of the clusters by the GMM can be observed in Table 6. Therefore, cluster 0 and cluster 1 are similar in attributes such as Beta, Volatility and Revenue Growth, with moderate values for most attributes. cluster 2, however, stands out with slightly higher Volatility and Revenue Growth, alongside a notable difference in Market Cap, where cluster 0 has high Market Cap and cluster 1 and 2 exhibit lower values. The Debt-to-Equity ratio and EPS Growth also show variation, with cluster 1 having a notably higher Debt-to-Equity ratio.

Table 6 – Cluster Characteristics Based on GMM

Attribute	cluster 0	cluster 1	cluster 2
Beta	Moderate (0.37)	Moderate (0.38)	Moderate (0.39)
Volatility	High (0.32)	Moderate (0.24)	High (0.33)
P/E Ratio	High (35)	High (45)	High (36)
Dividend Yield	Moderate (0.09)	Moderate (0.10)	Moderate (0.09)
Market Cap	High ($2 \cdot 10^9$)	Low ($1 \cdot 10^9$)	Low ($1 \cdot 10^9$)
Revenue Growth	High (0.21)	High (0.21)	High (0.24)
EPS Growth	Low (0.003)	Moderate (0.072)	Low (0.015)
Debt-to-Equity Ratio	Low (0.004)	Low (0.069)	Low (0.015)

The comparison of clustering methods highlights distinct differences in the classification of companies based on their financial attributes. Each method—K-Means, GMM, and Hierarchical Clustering—presents variations in the number of companies assigned to each cluster and the characteristics defining each group.

In the K-Means clustering results based on Table 3, Cluster 1 corresponds to the **Conservative** profile, characterized by low market risk and volatility, indicating stable investment opportunities with lower exposure to market fluctuations. Cluster 0 represents the **Moderate** profile, with a higher Beta and significant volatility, suggesting exposure to market fluctuations while maintaining balanced risk levels. Cluster 2 is associated with the **Risky** profile, exhibiting high volatility, making it suitable for investors willing to take on greater risk for potential returns.

The GMM results on Table 6 present a slightly different classification. Cluster 1 aligns with the **Conservative** profile, as it groups companies with moderate market risk and lower volatility. Cluster 0 represents the **Moderate** profile, with moderate market risk but higher volatility, exposing investors to a higher degree of price fluctuations. Cluster 2 corresponds to the **Risky** profile, characterized by moderate market risk and high volatility, resembling the K-Means classification for this category.

Hierarchical Clustering on Table 5 results differ by producing only two clusters. Cluster 1 represents the **Conservative** profile, featuring companies with moderate market risk and moderate volatility. Cluster 0 merges characteristics from both the **Moderate** and **Risky** profiles, containing companies with moderate market risk and significantly higher volatility. The lack of a clear separation between these two profiles suggests that Hierarchical Clustering is less effective at distinguishing investment categories.

After completing the work, a more detailed analysis of the tickers was performed to identify the position of each one in the clusters formed by each method used throughout the research. The clustering results obtained from K-Means on Table 7, Hierarchical clustering based on Table 8, and GMM at Table 9 reveal significant differences in how each algorithm interprets the structure of the data, which can be explained by the underlying principles of each method.

K-Means, as explained in 2, is a centroid-based algorithm that partitions data into a predefined number of k clusters by iteratively assigning each point to the nearest cluster centroid and updating the centroids to minimize intra-cluster variance. This method assumes clusters are spherical and have similar sizes, which may explain why it produced three distinct groups: cluster 0 (TECK, GLP), cluster 1 (AIF, ELS), and cluster 2 (MWA, ENLC). The rigid assignment of points to the nearest centroid may overlook potential relationships among elements that do not conform to spherical shapes, leading to the observed differences in clustering compared to other methods.

In contrast, Hierarchical clustering builds a hierarchy of clusters by either iteratively merging smaller clusters, using a chosen distance metric and linkage criterion (e.g., single, complete, or average linkage). This approach tends to capture nested relationships within the data and does not require a predefined number of clusters. In the given results, it formed only two clusters: one grouping TECK and GLP, and another containing AIF, ELS, and ENLC. The hierarchical approach likely considered the overall pairwise similarity between elements, leading to a broader grouping structure compared to the more rigid partitions of K-Means.

On the other hand, GMM is a probabilistic model that assumes data is generated from a mixture of Gaussian distributions, each with its own mean and covariance. Unlike K-Means, which assigns each point to a single cluster, GMM provides a soft clustering approach, where each data point has a probability of belonging to multiple clusters. In this case, GMM grouped all elements into a single cluster (0), with no elements assigned to the remaining two clusters, suggesting that the algorithm perceived a single underlying distribution encompassing all data points. This result may indicate that the dataset lacks distinct, well-separated subgroups or that the variance structure is too complex for rigid clustering techniques like K-Means to capture effectively.

Overall, these differences highlight how the assumptions and mechanisms of each clustering method influence their results. K-Means enforces distinct, non-overlapping clusters with equal variance, Hierarchical clustering emphasizes nested relationships based on chosen distance criteria, and GMM allows for flexible, probabilistic cluster assignments that can capture overlapping distributions.

In Tables 7, 8 and 9 , it is possible to observe how each company was classified into the different clusters by each method. These tables provide a detailed overview of the distribution of companies across the clusters, highlighting how each clustering technique categorizes the companies based on various financial metrics. By comparing the classification results, one can gain a deeper understanding of how each method identifies patterns and groupings in the data, which can help make informed decisions or analysis.

Table 7 – Clusters with associated Tickers by K-Means

cluster	Ticker
0	TECK, AGX, MYE, SSTK, URI, ASC, VST, GLP, AAP, SXC, MTH, HMY, VIPS, UWMC, TSM, WIT, WSM, CCJ, NWL, IGT, TRN, SAH, HP, HOG, BTE, FIX, DLX, RES, CWH, PVH, SQM, WCC, ARCH, ATKR, VAC, SM, X, CCS, IBP, PX, SD, CIM, ASX, UVE, NEM, F, XYF, DDS, DAN, HASI, SCCO, HY, TEX, GFI, BGS, CLB, FRO, RMD, TK, OEC, NOAH, WNC, CVI, VRT, UI, TPL, TOL, GNE, YRD, ASIX, ZTO, YUMC, EGY, KOP, BEKE, BHE, GFF, BCC, DG, RM, JWN, AEO, FMC, PAC, WST, WHR, LZB, HVT, ADM, TGLS, CRM, STM, WMS, SEM, LUV, BBW, MTRN, TME, PLOW, JBL, DRD, ABR, NEP, ATHM, CX, KGC, TEO, DKS, KSS, ATEN, NX, NMR, EE, SVM, HCC, NGVC, CATO, WWW, LPG, NVDA, CRS
1	PNW, PKE, DHR, CMI, AIF, LLY, CVX, DRI, CLX, ABEV, ELS, WPC, ZBH, OMC, AHH, ABBV, AEE, IDT, BG, RPM, CCI, DTE, PKG, CNI, ITW, OGS, SHW, AGR, GTY, PBA, MSI, TLK, GSK, SXT, MPLX, UTL, ETR, CPT, APLE, ATO, HON, RGA, REG, SNN, TRNO, PG, DLB, CNA, UL, CSX, AMP, WEC, AME, BDX, SNA, LDOS, AMT, CL, VIV, FDS, VICI, CMS, BMY, RCI, GHI, PHI, SLF, BRC, IDA, BFS, STNG, ICE, TU, MLM, ORAN, TMO, LNT, RYN, STE, BKH, COP, AMH, NMFC, POR, NNI, GD, ABT, XYL, DGX, INFY, IEX, NEE, IRM, SCI, AFG, TEF, MUSA, EC, MET, MAA, ALLE, RLI, EGP, KR, WLKP, AEP, CHT, RDY, SJM, AWR, CP, RNR, EIG, CHE, CUBE, NRP, VMC, AFL, TAP, GEF, ROP, UNP, RSG, PFE, HRL, SNY, TTE, UNH, HMC, CHD, SCM, FMX, LH, MMS, PFG, MA, ALEX, SJW, KO, AVY, DHT, AZN, MSA, SO, WCN, DCI, TGNA, STN, NWE, ALE, ALX, NOC, GPK, CTRA, PSA, VALE, MDU, AJG, EXC, DOV, ATR, E, ESNT, GGG, LTC, EXR, TIMB, EPD, SHO, MSM, SAIC, ORI, DEO, SF, STZ, MDT, AIZ, BRO, TAK, WU, FDP, VZ, RJF, REXR, PSO, LMT, TJX, PEG, SBS, ORA, NAT, CW, MAIN, PPL, NPK, EQR, AON, WMB, FE, SFL, EIX, SRE, WTS, AVA, NVO, SWX, NVS, MCR, OGE, EBF, VOYA, PAGP, MCO, XOM, IPG, PCG, NNN, ADC, TM, CPAC, GPC, QSR, WM, T, PGR, ENB, KMI, TNK, STAG, HII, CWT, TR, V, DE, UDR, SBI, MFC, CNP, HEI, SYK, KMB, WMT, LIN, HD, BR, CPK, OHI, HIG, ESS, FRT, FSK, UNM, K, FR, L, LYB, AXS, NHI, BUD, TRP, DOW, INGR, MMC, CB, NEU, PPG, SKM, DLR, JNJ, HSY, SYY, RIO, BLK, BCE, AWK, ELP, PNM, CI, BAH, ED, IBM, ROL, GIS, MKC, WELL, AVK, D, NI, NGG, ET, SON, TRI, PRI, AVB
2	MWA, HST, TT, MOV, SXI, MTX, FNF, CMRE, AVT, PLD, AGO, KAI, ACN, WSO, SB, HUN, APA, UAN, IR, ZTS, TFII, FAF, GGB, IVZ, DHI, LNC, LADR, CMC, BC, WKC, HRB, KOF, USPH, COO, ETN, PKX, EPR, TRU, CCK, UNF, CVEO, BTU, OXM, ASH, WSR, APO, EBR, UGP, WY, STWD, EXP, MAS, KIM, AGRO, ALSN, ESI, RDN, ARC, KNX, NOV, THO, RCL, EFX, PSX, OSK, GPI, PNR, ARES, APAM, ST, MS, BABA, GS, TKR, CAT, XHR, NKE, SEE, CSV, EVR, GLW, CRH, AEM, AER, H, TRGP, GIL, AKR, TXT, MLR, PWR, DRH, HAL, MUR, MRO, DVN, MLI, IP, AM, CR, JNPR, WH, BDC, LTM, MT, OKE, EOG, G, HXL, RWT, PFSI, HIW, DOC, ARMK, HES, APH, TAC, NRG, SKT, NUE, SCHW, EFC, TGT, PUK, RL, TNET, OLN, VNT, UMC, CNO, ZWS, VTR, DAL, KWR, VSH, PII, ARE, PAG, TFX, LII, MTDR, CVE, KFY, EME, WAB, AXP, LEA, NVGS, ASR, ABM, PMT, EMN, BEN, JOE, ENLC, CLDT, TEL, KBH, ALV, TNC, MPC, USAC, WES, TPB, FUL, OXY, IX, DFS, UE, FCX, NSC, MTG, SCL, AIT, EVTC, KMT, WD, PH, SMP, WLK, VMI, TS, WGO, GM, LVS, STC, MATX, AROC, SPG, VLO, FSS, SSD, PRU, AGM, BMI, BRX, HCA, GOLF, CNS, BBY, GPRK, BKE, TPX, WPP, CE, ALG, GBX, CBT, FLS, JCI, BX, APD, BXP, ITT, LNN, UHS, MTN, BYD, GSL, BKR, KRC, BERY, PHM, RHI, RS, PAA, AYI, AOS, WT, JBT, MAN, CF, HUM, TPR, CRI, SAP, CUZ, PIPR, CTS, OLP, GPN, RHP, BWA, CSL, SITC, SPB, CPA, FMS, AWI, AAT, EMR, CNQ, OC, DAC, BCO, SU, ESRT, SNX, SLB, UPS, PEB, AL, CIG, AIN, TWO, ROK, IFF, LAD, ENS, CVS, RLJ, NSP, HTGC, MGA, NOA, ARCO, ESE, AMG, SCS, A, WPM, RBA, KKR

Table 8 – Clusters with associated Tickers by Hierarchical

cluster	Ticker
0	TECK, AGX, MWA, MYE, SSTK, URI, CMRE, VST, AAP, MTH, KAI, HMY, APA, VIPS, DHI, UWMC, BC, USPH, WSM, PKX, TRU, CCJ, NWL, IGT, BTU, OXM, TRN, APO, SAH, HP, STWD, EXP, HOG, BTE, ESI, NOV, THO, RCL, EFX, FIX, DLX, RES, GPI, ST, CWH, PVH, SQM, WCC, NKE, ATKR, VAC, SM, MLR, PWR, X, CR, CCS, BDC, IBP, HXL, PFSI, SD, NRG, ASX, UVE, NEM, F, RL, TNET, XYF, DDS, DAN, HASI, SCCO, HY, DAL, TEX, PII, GFI, BGS, CLB, RMD, MTDR, CVE, EME, OEC, NOAH, WNC, KBH, VRT, WES, DFS, UI, FCX, TPL, TOL, WD, VMI, WGO, YRD, ASIX, AROC, EGY, KOP, BHE, GFF, SSD, BCC, DG, RM, JWN, AEO, FMC, PAC, WST, WHR, CNS, BBY, LZB, HVT, TPX, GBX, CBT, TGLS, CRM, STM, WMS, SEM, BYD, LUV, PHM, BBW, MTRN, AYI, JBT, TME, PLOW, JBL, DRD, TPR, CX, KGC, TEO, BWA, DKS, SPB, ATEN, NX, OC, EE, SVM, HCC, PEB, AL, NGVC, ROK, LAD, WWW, NOA, NVDA, SCS, CRS, KKR
1	PNW, PKE, DHR, CMI, AIF, LLY, CVX, DRI, CLX, ABEV, HST, ELS, WPC, ZBH, OMC, AHH, TT, ABBV, AEE, IDT, BG, RPM, MOV, CCI, SXI, DTE, MTX, FNF, PKG, ASC, CNI, ITW, AVT, OGS, GLP, PLD, SHW, AGR, GTY, SXC, PBA, AGO, MSI, ACN, WSO, SB, HUN, TLK, UAN, IR, GSK, ZTS, SXT, MPLX, TFII, UTL, ETR, CPT, FAF, GGB, IVZ, LNC, LADR, APLE, ATO, TSM, CMC, HON, WKC, HRB, KOF, RGA, REG, SNN, COO, WIT, ETN, TRNO, EPR, CCK, PG, DLB, CNA, UNF, CVEO, UL, CSX, AMP, WEC, ASH, AME, BDX, WSR, SNA, LDOS, AMT, EBR, CL, UGP, WY, VIV, FDS, VICI, CMS, MAS, KIM, AGRO, BMY, ALSN, RCI, RDN, ARC, GHI, PHI, KNX, SLF, BRC, PSX, OSK, IDA, BFS, STNG, PNR, ARES, ICE, TU, MLM, APAM, ORAN, TMO, LNT, MS, BABA, RYN, GS, TKR, STE, CAT, XHR, BKH, COP, SEE, ARCH, CSV, EVR, AMH, GLW, NMFC, CRH, POR, AEM, NNI, AER, GD, ABT, H, TRGP, GIL, AKR, XYL, TXT, DGX, DRH, HAL, MUR, MRO, INFY, DVN, MLI, IEX, NEE, IRM, IP, SCI, AFG, AM, TEF, MUSA, EC, JNPR, WH, MET, LTM, MT, OKE, MAA, EOG, ALLE, RLI, G, RWT, EGP, KR, PX, HIW, DOC, WLKP, AEP, CHT, RDY, SJM, AWR, ARMK, CIM, CP, RNR, HES, APH, EIG, TAC, CHE, CUBE, NRP, VMC, SKT, AFL, NUE, SCHW, TAP, GEF, ROP, UNP, EFC, TGT, RSG, PUK, PFE, HRL, SNY, TTE, UNH, HMC, OLN, VNT, CHD, SCM, UMC, CNO, FMX, LH, ZWS, MMS, VTR, PFG, MA, ALEX, SJW, KO, AVY, DHT, KWR, AZN, VSH, MSA, SO, WCN, ARE, DCI, TGNA, FRO, PAG, TFX, STN, NWE, TK, ALE, LII, KFY, ALX, NOC, WAB, GPK, AXP, LEA, NVGS, CTRA, PSA, ASR, VALE, MDU, AJG, ABM, PMT, EMN, EXC, DOV, BEN, JOE, ATR, E, ENLC, ESNT, CLDT, GGG, LTC, TEL, EXR, TIMB, CVI, EPD, ALV, SHO, TNC, MPC, USAC, TPB, FUL, MSM, OXY, IX, SAIC, ORI, DEO, SF, UE, NSC, MTG, SCL, STZ, AIT, MDT, AIZ, EVTC, KMT, GNE, BRO, TAK, PH, SMP, WLK, WU, FDP, VZ, TS, RJF, GM, LVS, STC, MATX, ZTO, REXR, PSO, SPG, LMT, YUMC, VLO, TJX, FSS, PEG, BEKE, SBS, ORA, NAT, CW, MAIN, PPL, NPK, EQR, AON, PRU, WMB, FE, SFL, EIX, SRE, WTS, AVA, NVO, AGM, SWX, BMI, BRX, NVS, MCR, OGE, HCA, GOLF, EBF, VOYA, GPRK, BKE, PAGP, MCO, WPP, XOM, IPG, CE, ALG, ADM, PCG, FLS, NNN, JCI, ADC, TM, CPAC, GPC, BX, APD, BXP, ITT, QSR, LNN, WM, T, UHS, MTN, PGR, ENB, KMI, TNK, GSL, BKR, KRC, STAG, BERY, HII, CWT, RHI, RS, PAA, TR, V, DE, AOS, UDR, WT, SBI, MAN, MFC, CF, HUM, CNP, HEI, ABR, SYK, CRI, NEP, ATHM, KMB, SAP, CUZ, PIPR, WMT, CTS, OLP, LIN, HD, GPN, RHP, BR, CPK, OHI, HIG, CSL, ESS, FRT, SITC, FSK, UNM, CPA, FMS, AWI, KSS, K, FR, L, AAT, EMR, LYB, CNQ, AXS, NHI, BUD, TRP, DOW, INGR, NMR, MMC, CB, DAC, BCO, NEU, PPG, SU, SKM, DLR, JNJ, ESRT, SNX, HSY, SYY, RIO, SLB, UPS, BLK, BCE, AWK, ELP, PNM, CIG, AIN, TWO, CI, BAH, IFF, ED, IBM, ROL, ENS, CVS, GIS, MKC, WELL, CATO, RLJ, NSP, HTGC, MGA, AVK, D, LPG, ARCO, NI, NGG, ESE, AMG, ET, SON, TRI, A, WPM, RBA, PRI, AVB

Table 9 – Clusters with associated Tickers by GMM

cluster	**Ticker**
0	PNW, PKE, DHR, TECK, CMI, AIF, LLY, CVX, DRI, ABEV, MWA, HST, ELS, MYE, WPC, ZBH, OMC, TT, ABBV, AEE, IDT, BG, RPM, MOV, DTE, MTX, FNF, SSTK, PKG, URI, CMRE, ASC, CNI, ITW, VST, AVT, OGS, GLP, AAP, PLD, SHW, AGR, GTY, SXC, PBA, AGO, MTH, KAI, ACN, WSO, HMY, SB, HUN, TLK, UAN, IR, GSK, VIPS, ZTS, SXT, MPLX, TFII, UTL, ETR, FAF, GGB, IVZ, DHI, LADR, UWMC, APLE, ATO, TSM, CMC, BC, HON, KOF, USPH, RGA, REG, SNN, COO, WIT, ETN, WSM, PKX, TRU, CCK, PG, NWL, DLB, IGT, CNA, UNF, UL, CSX, BTU, AMP, WEC, OXM, ASH, AME, BDX, WSR, SNA, LDOS, AMT, SAH, HP, EBR, UGP, WY, VIV, FDS, VICI, EXP, CMS, KIM, HOG, AGRO, BMY, ALSN, RDN, ARC, GHI, PHI, KNX, NOV, SLF, THO, BRC, RCL, EFX, FIX, PSX, DLX, RES, OSK, IDA, BFS, STNG, GPI, PNR, ARES, ICE, TU, MLM, APAM, ST, ORAN, TMO, LNT, MS, BABA, GS, TKR, STE, CAT, PVH, SQM, WCC, NKE, BKH, COP, ARCH, CSV, ATKR, EVR, GLW, NMFC, CRH, POR, VAC, AEM, NNI, SM, AER, GD, ABT, TRGP, GIL, XYL, TXT, DGX, MLR, PWR, DRH, HAL, MUR, MRO, INFY, DVN, MLI, IEX, NEE, X, IP, SCI, AFG, AM, TEF, MUSA, CR, EC, JNPR, WH, CCS, BDC, MET, MT, OKE, MAA, EOG, ALLE, IBP, RLI, G, HXL, PFSI, EGP, KR, WLKP, AEP, CHT, RDY, SD, SJM, AWR, ARMK, CIM, CP, APH, EIG, TAC, NRG, CHE, CUBE, NRP, VMC, SKT, AFL, NUE, SCHW, TAP, GEF, ASX, ROP, UNP, UVE, F, TGT, RSG, PUK, RL, PFE, HRL, XYF, SNY, TTE, UNH, HMC, OLN, DDS, VNT, CHD, SCM, UMC, DAN, CNO, FMX, LH, ZWS, HASI, MMS, SCCO, PFG, MA, ALEX, HY, DAL, SJW, KO, AVY, DHT, TEX, KWR, AZN, VSH, PII, GFI, BGS, MSA, SO, WCN, ARE, DCI, CLB, TGNA, FRO, PAG, RMD, TFX, STN, NWE, TK, ALE, LII, MTDR, CVE, KFY, EME, ALX, NOC, WAB, GPK, AXP, LEA, NVGS, OEC, CTRA, PSA, ASR, VALE, NOAH, MDU, AJG, ABM, PMT, EMN, EXC, DOV, BEN, ATR, E, ENLC, ESNT, WNC, GGG, LTC, TEL, TIMB, CVI, KBH, EPD, ALV, SHO, VRT, TNC, MPC, WES, TPB, FUL, MSM, OXY, IX, DFS, SAIC, ORI, DEO, SF, FCX, NSC, MTG, TPL, SCL, AIT, TOL, MDT, AIZ, EVTC, KMT, WD, GNE, BRO, TAK, PH, SMP, WLK, WU, FDP, VZ, VMI, TS, RJF, WGO, YRD, GM, LVS, ASIX, STC, MATX, ZTO, AROC, REXR, PSO, LMT, YUMC, VLO, TJX, FSS, KOP, PEG, BEKE, SBS, ORA, NAT, CW, BHE, MAIN, PPL, SSD, BCC, DG, AON, PRU, WMB, RM, FE, SFL, JWN, AEO, EIX, SRE, WTS, PAC, AVA, NVO, SWX, BMI, BRX, NVS, WST, MCR, OGE, WHR, GOLF, EBF, CNS, VOYA, BBY, GPRK, LZB, HVT, BKE, PAGP, MCO, WPP, XOM, IPG, CE, ALG, GBX, ADM, CBT, PCG, TGLS, FLS, NNN, JCI, ADC, CRM, TM, CPAC, GPC, BX, APD, BXP, ITT, QSR, LNN, STM, WM, T, UHS, WMS, SEM, MTN, BYD, ENB, KMI, TNK, GSL, BKR, KRC, STAG, BERY, LUV, HII, PHM, RHI, BBW, RS, MTRN, PAA, V, AYI, DE, AOS, WT, JBT, MAN, MFC, CF, HUM, TME, PLOW, JBL, CNP, DRD, TPR, ABR, SYK, CRI, NEP, ATHM, SAP, WMT, CTS, LIN, HD, GPN, RHP, CX, BR, CPK, KGC, OHI, HIG, BWA, FRT, DKS, FSK, SPB, UNM, FMS, AWI, KSS, ATEN, FR, L, AAT, EMR, LYB, CNQ, AXS, NHI, NX, DOW, OC, INGR, MMC, CB, DAC, NEU, PPG, SVM, SU, SKM, HCC, JNJ, SNX, HSY, RIO, SLB, UPS, BLK, BCE, AWK, ELP, PNM, AL, CIG, AIN, TWO, NGVC, ROK, CI, BAH, ED, LAD, IBM, ROL, ENS, CVS, GIS, MKC, CATO, WWW, NSP, HTGC, MGA, D, LPG, ARCO, NI, NVDA, NGG, ESE, AMG, ET, CRS, SON, TRI, A, WPM, RBA, PRI, AVB, KKR
1	IRM, VTR, JOE, CLDT, UE, PGR, SBI, CUZ, BUD, TRP, DLR, ESRT, PEB, WELL, AVK
2	CLX, AGX, AHH, CCI, SXI, MSI, APA, CPT, LNC, WKC, HRB, TRNO, EPR, CCJ, CVEO, TRN, APO, CL, STWD, MAS, BTE, RCI, ESI, RYN, CWH, XHR, SEE, AMH, H, AKR, LTM, RWT, PX, HIW, DOC, RNR, HES, EFC, NEM, TNET, EXR, USAC, UI, STZ, SPG, EGY, NPK, GFF, EQR, FMC, AGM, HCA, TPX, CWT, TR, UDR, HEI, KMB, PIPR, OLP, TEO, CSL, ESS, SITC, CPA, K, NMR, EE, BCO, SYY, IFF, RLJ, NOA, SCS