OULUN AMMATTIKORKEAKOULU

COMPARATIVE ANALYSIS OF CLUSTERING TECHNIQUES FOR STOCK SELECTION IN FINNISH STOCK MARKETS USING COMMON FINANCIAL METRICS.

Timo Rautio

COMPARATIVE ANALYSIS OF CLUSTERING TECHNIQUES FOR STOCK SELECTION IN FINNISH STOCK MARKETS USING COMMON FINANCIAL METRICS.

Timo Rautio
Final projects
Spring 2024
Data Analytics and Project Management
Oulu University of Applied Sciences

ABSTRACT

Oulu University of Applied Sciences
Master's degree, Data Analytics and Project Management

Author(s): Timo Rautio
Title of the thesis: Comparative analysis of clustering techniques for stock selection in Finnish stock markets using common financial metrics.
Thesis examiner(s): Ilpo Virtanen
Term and year of thesis completion: Spring 2024 Pages: 68

This thesis aimed to study the efficacy of unsupervised machine learning techniques, specifically clustering algorithms, in the context of stock selection. The goal was to find out are the clustering methods suitable for stock selection and whether they were able to outperform explicit rule-based strategies utilizing favourable financial ratios. Additionally, this study aimed to examine the generalizability of clustering results for profitable, risk-tolerant investment decisions.

The dataset reflected the state of the markets on the final trading day of 2022 and included some standard financial metrics, such as return on equity, price-to-earnings ratio, price-to-book ratio, debt-to-equity ratio, dividend yield, earnings yield, and earnings per share.

Removing missing values and outlier removal were applied as pre-processing measures, and the dataset was fed to different clustering algorithms including K-Means, Hierarchical clustering, and Gaussian Mixture Model. These clustering methods were selected due their distinct clustering approach and the ability to specify the number of clusters. The clustering results were compared using internal evaluation methods including silhouette score, Davies Boulding index, and Dunn index. Results were also analysed using annual price fluctuation between 2022 and 2023.

The result of this research indicated that Hierarchical clustering outperforms the K-Means and GMM based on internal evaluation methods measuring the compactness and separation of clusters. However, the differences were not notably different and rather marginal. Analysis of the best-performing cluster revealed an average annual stock price growth of 36.76%, but as an investment strategy, it presented a higher risk compared to the approach based on favourable financial metrics. These findings suggested that while hierarchical clustering can offer the best performance in some cases, its higher risk profile may limit its usage in investment strategies.

As the results indicated a poor suitability of using clustering algorithms used alone to stock selection, future research should explore different clustering algorithms like DBSCAN and neural network models. The future research should also focus on longer-term analysis to gain better understanding over economic cycles as well to study and take in consideration the industry-specific differences.

Keywords: Machine learning, Clusters, stock analysis, K-Means, Classification problem

CONTENTS

CONTENTS 4
1 INTRODUCTION 6
1.1 Background and motivation 6
1.2 Research question 6
2 THEORETICAL BACKGROUND 8
2.1 Machine learning 8
2.1 Unsupervised learning 9
2.2 Clustering algorithms 9
2.2.1 K-Means clustering 9
2.2.2 Hierarchical Clustering 11
2.2.3 Gaussian Mixture Model 12
2.3 Clustering evaluation methods 13
2.4 Internal evaluation methods 14
2.4.1 Silhouette Score 15
2.4.2 Elbow Method 15
2.4.3 Bayesian Information Criterion 16
2.4.4 Davies-Bouldin Index 17
2.4.5 Dunn Index 18
3 FINANCIAL RATIOS 19
3.1 Stock valuation indicators 19
3.1.1 Return on equity (ROE) 19
3.1.2 Price/earnings ratio (P/E ratio) 20
3.1.3 Price to book (P/B) 21
3.1.4 Debt-to-equity (D/E) ratio 21
3.1.5 Dividend yield 22
3.1.6 Earnings Yield 23
3.1.7 Earnings per share EPS 24
4 DATA 25
4.1 Source of data 25
4.2 Data collection 25
4.3 Managing missing data 27
4.4 Dataset overview 28

4.4.1 P/B ratio 29
4.4.2 P/E ratio 31
4.4.3 Dividend yield 31
4.4.4 Dept to equity ratio 33
4.4.5 Return of equity (ROE) 34
4.4.6 Earnings per share (EPS) 35
4.4.7 Earnings yield 36
4.5 Balancing the dataset 37
5 RESULTS 40
5.1 K-means Clustering 40
5.1.1 Choosing the k-value 40
5.1.2 K-Means Clustering Results 41
5.2 Hierarchical Clustering 46
5.2.1 Defining the number of clusters for hierarchical clustering 46
5.2.2 Hierarchical clustering Results 47
5.3 Gaussian Mixture Model 51
5.3.1 Choosing the n-components for GMM 51
5.3.2 Gaussian Mixture Model (GMM) Results 52
5.4 Comparison between clustering results 57
5.4.1 Internal evaluation 58
5.4.2 Business metrics evaluation 59
5.4.3 Principal Component Analysis 60
6 DISCUSSION AND CONCLUSION 62
REFERENCES 66

1 INTRODUCTION

1.1 Background and motivation

I have gained increasing interest towards investing and the stock market, but then had a lack of technical skills when comes to analytics. However, things changed in 2020 after I participated in a basic data analysis course in Haaga-Helia. One of the assignments involved using publicly listed stock information from Yahoo Finance for some basic analysis. This assignment helped me realize that data collection for such financial data could be done in a relatively easy way.

I noticed an opportunity to utilize and straightforward the process for constructing datasets from stock market information and realized that this would enable me to conduct comprehensive datasets for analyses that were otherwise too time-consuming to collect. Streamlining the data collection and analysis process would allow me to explore stock market and financial data in greater depth and customization than traditional investor sites.

This set the foundation for a growing interest in the data analysis techniques, financial markets, and machine learning algorithms. This thesis aims to show how analytical tools can reveal patterns and trends in stocks. This research also supports my long-term aim of deepening my comprehension of data-driven decision-making specifically in stock market investments.

My thesis aims to assess the feasibility of comparing clustering algorithms by using common value-investing financial ratios. The objective is to evaluate different clustering methods and determine their effectiveness in grouping companies based on value investing principles.

1.2 Research questions

The central focus of this thesis is to address the following questions:

- How do clustering algorithms perform when applied to dataset constructed with value investing financial ratios and the investment horizon is short, a maximum of 12 months?

- Are the clustering methods able to outperform the stocks picked using explicit rules based on the favourable financial ratios?
- Are the clustering results generalizable in such a way that it is possible to make profitable risk-tolerant investment decision based on them?

2 THEORETICAL BACKGROUND

In this chapter, the theoretical background is outlined. The technical background of the methods utilized for clustering and classification will be introduced.

2.1 Machine learning

The goal of machine learning is generalization. In the field of computing, the traditional method typically involves programmers creating specific rules for a computer program, enabling it to transform input data into the desired output. Machine learning represents a whole different approach where the machine examines the input data alongside corresponding outputs to deduce the underlying rules. This means that a machine learning system is trained rather than explicitly programmed. This training process entails presenting the system with a plethora of examples pertinent to a task, enabling it to identify statistical structures within these examples, which ultimately leads to the formulation of rules for automating the task. (Chollet, 2021).

Consider the example of categorizing vacation photographs. By providing a machine learning system with many images already tagged by humans, it learns statistical rules to link specific pictures with specific tags. Machine learning, a branch of Artificial Intelligence (AI), has seen remarkable growth since the 1990s, propelled by advancements in computational power and the availability of larger datasets. (Chollet, 2021).

While machine learning shares connections with mathematical statistics, it diverges in several notable ways. This divergence is comparable to the relationship between medicine and chemistry; while related, medicine addresses distinct systems with their unique properties. Machine learning, unlike traditional statistics, often tackles large and complex datasets. For instance, it might deal with millions of images, each consisting of tens of thousands of pixels, where classical statistical methods such as Bayesian analysis would be unfeasible. As a result, machine learning, and especially deep learning, tends to have comparatively limited mathematical theory perhaps controversially so and is fundamentally an engineering discipline. (Chollet, 2021).

Machine learning is an empirical field, driven by practical findings and heavily reliant on advances in software and hardware. Its nature is more pragmatic and application-oriented, as opposed to being purely theoretical, aligning more with engineering rather than fields like theoretical physics or mathematics. (Chollet, 2021).

2.1 Unsupervised learning

In unsupervised learning, the focus is on understanding how systems can autonomously acquire representations of specific input patterns that mirror the statistical structure inherent in the entire set of input patterns. Unlike supervised learning that encompasses the problem set of having a labelled dataset that can be used to either classify or fit a regression line on, unsupervised learning does not involve explicit target outputs or environmental evaluations linked to each input (Johnston, Jones, Kruger, 2019). Instead, the unsupervised learner relies on predefined biases to discern which aspects of the input's structure should be reflected in the output. (Dayan, 2023.)

2.2 Clustering algorithms

This section presents an overview of the clustering algorithms employed in this study. The objective was to evaluate clustering algorithms that offer the capability for users to specify the number of clusters, thus enabling straightforward comparison of results. In accordance with this objective, three distinct types of clustering algorithms were selected and assessed for their suitability in the context of unsupervised learning tasks for stock classification: K-Means clustering, Hierarchical clustering and Gaussian Mixture Model (GMM).

2.2.1 K-Means clustering

The K-means algorithm can be considered one of the most powerful and used clustering algorithms in the research field (Ahmed 2020). The K-Means clustering algorithm clusters data by separating samples in n groups of equal variances by minimising the inertia or within-cluster sum-of-squares. K-Means algorithms require that the number of clusters is specified. The benefit of K-Means is its good scalability to a large number of samples which means it can be used in a large range of application areas in many different fields. (scikit learn, 2024).

The k-means algorithm segregates a collection of N samples, denoted as X, into K distinct clusters labelled as C. Each cluster is characterized by its mean, denoted as μ_j, which represents the average of the samples within that cluster. These means are commonly referred to as the cluster "centroids." Notably, these centroids are often distinct from points within the original set X, despite existing in the same spatial context. The K-means algorithm aims to choose centroids that minimise the inertia, or within-cluster sum-of-squares criterion. (scikit learn, 2024).

One of the main limitations of the K-means algorithm is its potential convergence to a local optimum. This issue largely hinges on the selection of initial prototypes. To mitigate this limitation, the algorithm is typically executed multiple times with varying random initializations, adhering to a trial-and-error methodology. The optimal solution is then chosen based on the criterion of the minimum sum of squared errors. Furthermore, the sum of squared error's objective function presents another challenge. This function's nature means that patterns situated at substantial distances can disproportionately attract the cluster centers towards them, rendering the algorithm highly sensitive to outliers. (Albalate, Minker 2011).

An additional constraint of the K-means algorithm pertains to the detection of the shape and size of clusters. Given the algorithm's partitioning of the dataset into Voronoi regions, it predominantly identifies convex or globular clusters. However, it may falter in recognizing clusters of non-standard shapes or those that vary in size, especially when such clusters are near one another. (Albalate, Minker 2011).

Lastly, the K-means algorithm necessitates the pre-determination of the number of clusters, denoted as 'k.' Accurately identifying the optimal number of clusters within a dataset can be a complex process, generally achieved through the evaluation of cluster partitions obtained from multiple iterations of the algorithm with varying values of k. (Albalate, Minker 2011).

An example of K-Means clustering is illustrated in Figure 1, where the left plot depicts the scenario prior to clustering and the right plot reveals the clusters. The figure demonstrates clear clustering results with four distinct clusters. The red 'X' marks the cluster centroid.

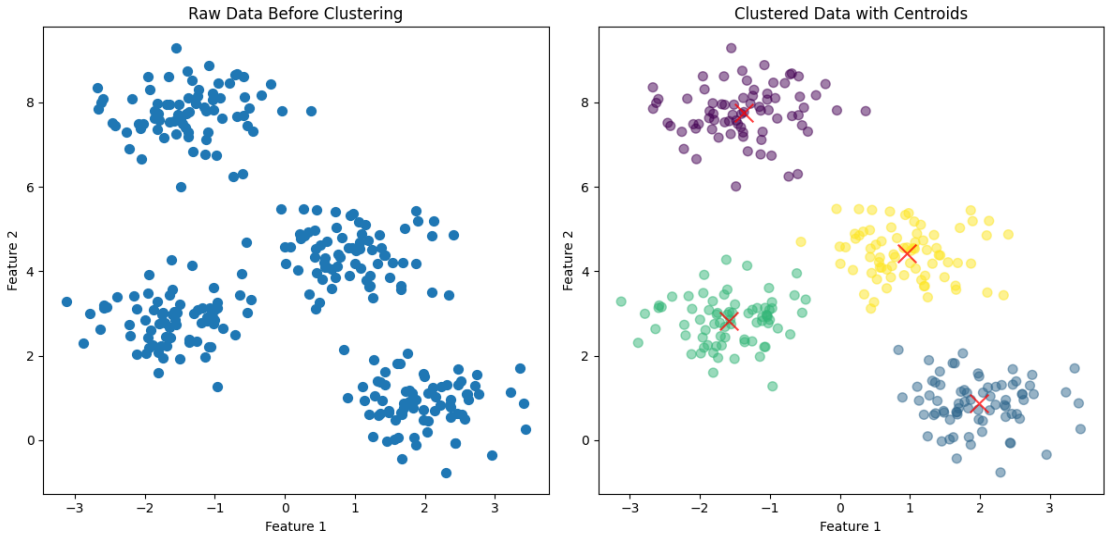

Figure 1 Example of cluster formation. Unclustered data on the left and K-means clustered data with visible cluster centroids on the right (k=4)

2.2.2 Hierarchical Clustering

Hierarchical clustering constructs nested clusters by either merging or splitting them step by step. This hierarchical arrangement is visually represented as a dendrogram, a tree-like diagram where the root symbolises a single cluster containing all data points, and the leaves correspond to clusters with individual data points (scikit-learn, 2024). This process and its visualisation are demonstrated in figure 2.

In hierarchical clustering, the Agglomerative clustering method uses a bottom-up approach. In the beginning, each data point is treated as its own cluster. As the algorithm progresses, clusters gradually merge based on a predefined linkage criterion. The choice of linkage criterion influences the merging strategy. The merging strategy used in this thesis is Ward's method which minimises the sum of squared differences within all clusters, aiming to reduce variance within clusters. This approach is similar to the k-means objective function but applies an agglomerative hierarchical methodology. (scikit-learn, 2024).

Other linkage criteria include complete linkage, which minimises the maximum distance between observations in different clusters; average linkage, which minimises the average distance between all pairs of observations in different clusters; and single linkage, which minimises the distance

between the closest observations in different clusters. Although Agglomerative Clustering can scale to large datasets when used with a connectivity matrix, it remains computationally demanding without connectivity constraints because it must consider all possible merges at each step. (scikit-learn, 2024).

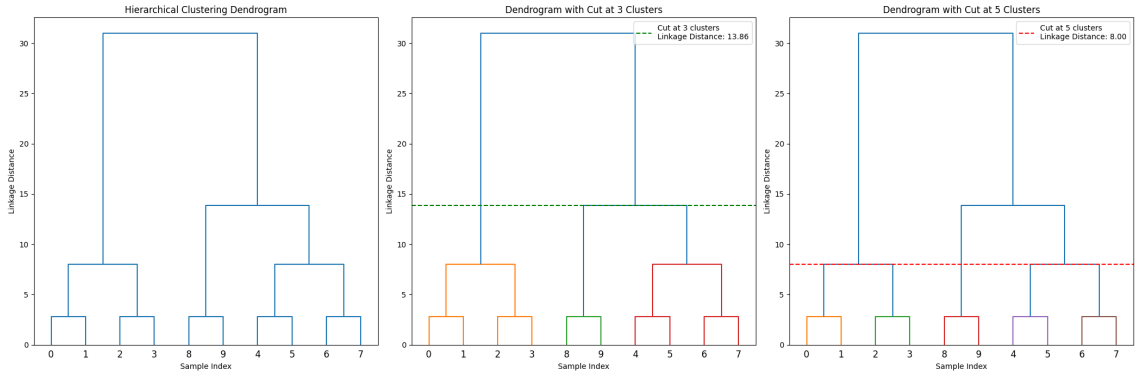

Figure 2 Example of Hierarchical Clustering dendrograms with cuts at 3 and 5 clusters, separated by colour.

2.2.3 Gaussian Mixture Model

According to scikit-learn documentation (scikit-learn, 2024), a Gaussian mixture model is a machine learning method used to determine the probability each data point belongs to a given cluster. It is a type of probabilistic model that supposes that all data points come from a combination of a limited number of Gaussian distributions with unknown parameters. This means that mixture models can be considered as an extension of the k-means clustering algorithm, which includes information about the covariance structure of the data in addition to the centers of the hidden Gaussians.

The theoretical framework of GMM revolves around the assumption that every sample in the dataset can be represented as a mixture of multiple Gaussian components. Each component is characterized by parameters such as mean and covariance, contributing to the overall shape and spread of the distribution. Importantly, GMMs employ a set of mixture weights that sum to one, ensuring that each component's contribution to the model is appropriately scaled. (Reynolds, 2024).

One of the fundamental processes in the application of GMMs is parameter estimation, typically achieved through algorithms such as Expectation-Maximization (EM) or Maximum A Posteriori (MAP). These algorithms facilitate the iterative refinement of parameters, ensuring that the model

closely approximates the underlying data distribution. The EM algorithm, for instance, alternates between estimating the latent variables and updating the model parameters to maximize the likelihood of the data given the model. (Reynolds, 2024).

2.3 Clustering evaluation methods

One of the most important issues in cluster analysis is evaluating clustering results to find the partition that best fits the underlying data. This process is the central focus of clustering validation. (Halkidi, Batistakis, Vazirgiannis, 2001)

Clustering enables the identification of groups of similar data within a dataset. However, many clustering methods like k-means for example, do not inherently specify the optimal number of clusters. Consequently, it becomes imperative to ascertain the optimal number of clusters and evaluate their quality independently. Failing to do so could potentially result in clustering that misguides decision-making processes. (Dancker 2022).

The objective is to achieve high intra-cluster similarity and low inter-cluster similarity, essentially aiming for clusters that are compact yet well-separated. High intra-cluster similarity indicates that data points within the same cluster share similar properties or features, while low inter-cluster similarity ensures that data points in different clusters possess distinct properties or features. (Dancker 2022). This concept is illustrated in figure 3, which highlights desirable dense clusters with high inter-cluster distances on the left. Undesirable clusters with low inter-cluster and high intra-cluster distances are present on the right side of the figure.

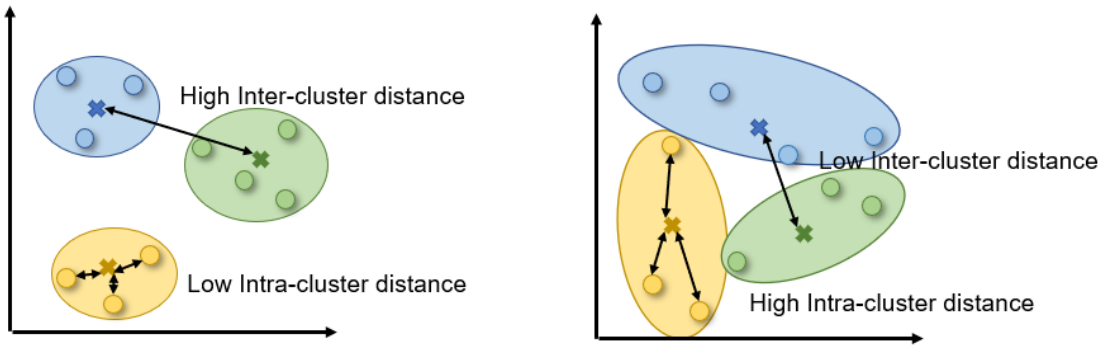

Figure 3 Good (left) vs. bad (right) clustering based on the Inter-cluster and Intra-cluster distance (Dancker 2022, cited 22.4.2024)

To determine the optimal number of clusters, it is essential to assess the quality of clusters in terms of their compactness, connectedness, and separation. This can be achieved through three primary approaches:

1. Internal Cluster Validation: Utilizes internal data from the clustering process, such as the within-cluster sum of squares, to evaluate cluster quality.
2. Relative Cluster Validation: Involves adjusting clustering parameters, such as the number of clusters, to observe changes and determine optimal settings.
3. External Cluster Validation: Compares clustering results with externally known outcomes, such as predefined labels, to validate the accuracy of the clustering method.

For measuring and validating the quality of clusters, methods are generally categorized into direct and statistical approaches. Direct methods, including the Elbow Curve, Silhouette Score, and Davies-Bouldin Index, focus on measuring within-cluster and between-cluster similarities. Meanwhile, statistical methods like the Gap Statistic assess the clustering effectiveness against a null hypothesis. Typically, internal and relative validations are combined to ascertain the optimal number of clusters, while external validation is used to select the most suitable clustering method. (Dancker 2022).

In the context of this thesis, the focus is set on internal cluster validation methods and the overall comparison is made by employing the average Gains%-value across the clusters. This decision is made primarily by the nature of the data and the objectives of the clustering process. The dataset used in this thesis does not come with known structures or predefined labels. This absence of external ground truth makes it impractical to employ external evaluation methods.

2.4 Internal evaluation methods

Internal evaluation methods assess the quality of clustering by examining the data clustering structure without reference to external information. These methods primarily focus on measuring the compactness and separation of the clusters formed by the algorithm.

2.4.1 Silhouette Score

The Silhouette Score measures both the mean intra-cluster distance and the mean nearest-cluster distance for each data point, effectively quantifying the variation within and between clusters. The difference between these distances is normalized by the maximum of these distances, and the process is averaged across all data points to get the overall Silhouette score. The score ranges from -1, indicative of incorrect clustering, to +1, which signifies highly dense clustering. A score of zero suggests that the clusters overlap. Therefore, higher Silhouette Scores, which are indicative of clusters that are both dense and well-separated, reflect a more distinct and appropriately segmented data structure (Dancker 2022). This relationship is visually represented in Figure 4, where an example of the optimal number of clusters is identified based on the highest Silhouette score.

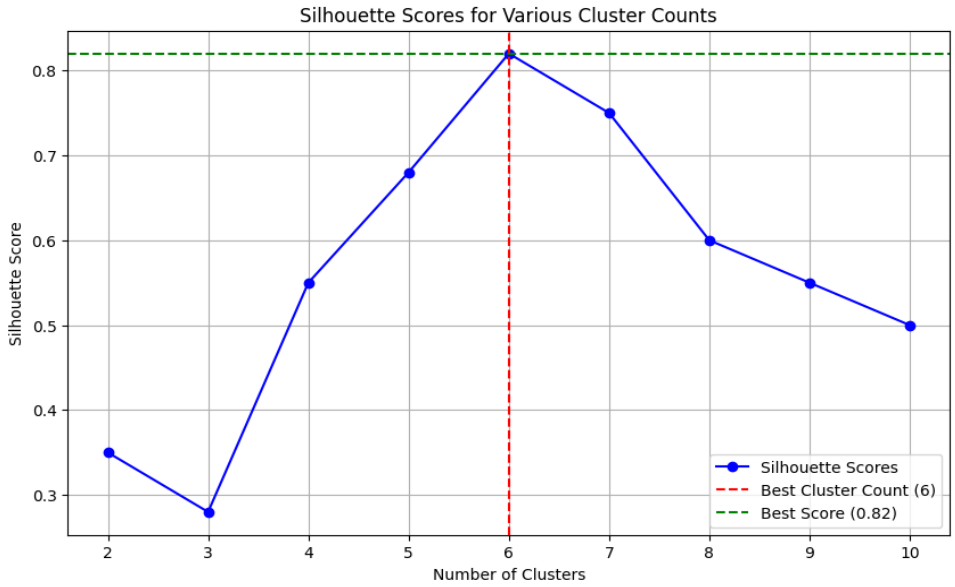

Figure 4 Example of Silhouette scores for various cluster counts

2.4.2 Elbow Method

The Elbow Method involves plotting the sum of squared distances (inertia) from each point to its assigned centre for different numbers of clusters. As the number of clusters increases, the inertia decreases. In Figure 5, you can see an illustration that highlights the contrast between a clear and an unclear elbow. The optimal cluster size is determined by looking for the clear elbow. The point on the graph where the inertia begins decreasing at a slower rate, forming an 'elbow' shape, is considered the ideal number of clusters. This method helps to balance between the number of

clusters and the distance of the data points from the centroids. The smaller the inertia, the denser the clusters. (Dancker, 2022).

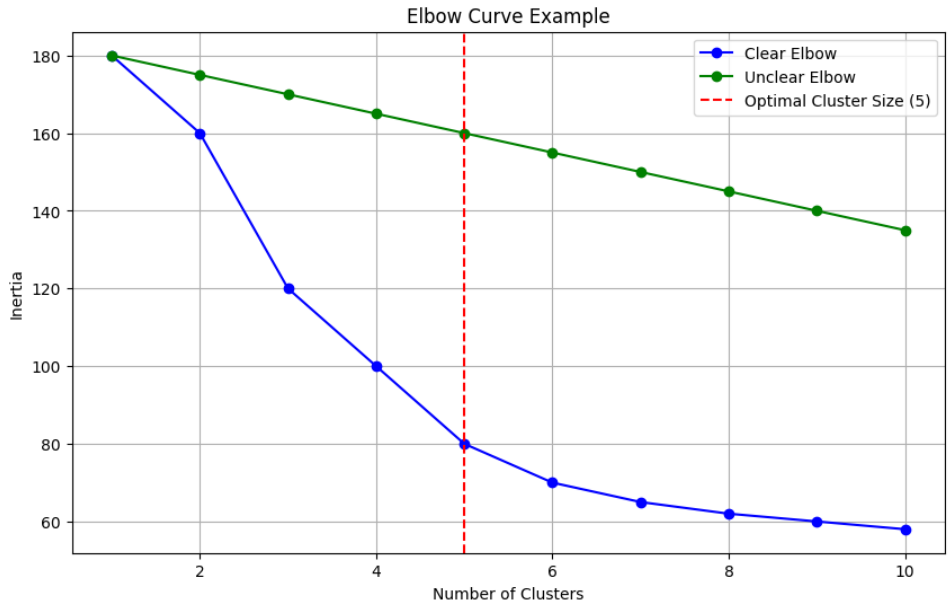

Figure 5 Example of elbow curve with clear(blue) and unclear(green) elbow

2.4.3 Bayesian Information Criterion

Bayesian information criterion (BIC) is a scoring system for model comparison in classical statistics dealing with models with different numbers of free parameters (Schwarz, 1978). The Bayesian Information Criterion offers a measure of how well a Gaussian Mixture Model predicts the available data. A lower BIC indicates a better-fitting model for accurately predicting the data, which also implies its suitability for approximating the true, unknown distribution. To prevent overfitting, this criterion imposes a penalty on models with a large number of clusters. (Lavorini, 2022).

As shown in Figure 6, BIC incorporates a penalty for models that use a larger number of clusters, which helps prevent overfitting. Ideally, a lower BIC suggests a better model fit. The BIC values are plotted against different cluster counts, revealing potential shifts in slope. The optimal number of clusters is typically where the BIC curve levels off, indicating diminishing returns on model improvement with additional clusters. This method ensures a balance between model complexity and fitting accuracy, to minimize BIC while avoiding excessive complexity.

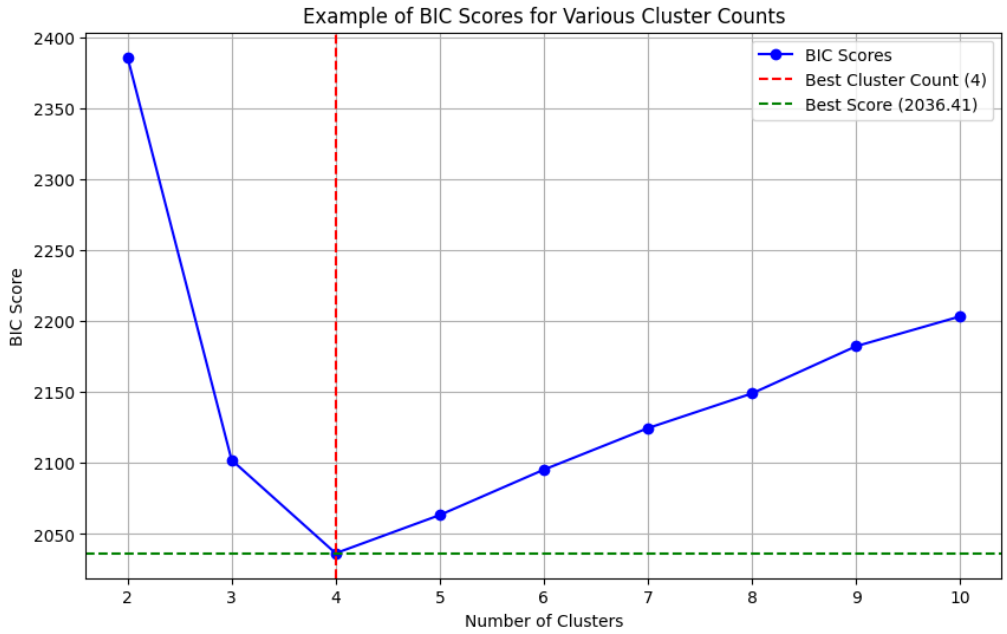

Figure 6 Example of BIC scores for various cluster counts.

2.4.4 Davies-Bouldin Index

The Davies-Bouldin Index (DBI) is a technique for assessing the internal validity of a cluster analysis. Internal validity measures how effectively clustering has been performed by analysing the quantities and features derived from the datasets. (Fahmi, Suprapto, Wirawan, 2016).

DBI helps assess how well a clustering model performs. A lower score indicates a clearer separation between the clusters formed by the model. DBI represents the average 'similarity' between clusters. Similarity is a measure that compares the distance between clusters with the size of the clusters themselves. (scikit-learn, 2024).

The score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Clusters which are farther apart and less dispersed will result in a better score. Theoretically, the best possible score is zero and the lower values indicate better clustering. (scikit-learn, 2024).

2.4.5 Dunn Index

The Dunn Index evaluates clustering quality by considering two factors: compactness and separation. Compactness is defined as the maximum distance between data points within the same cluster, while separation refers to the minimum distance between different clusters. The index is particularly useful for determining the optimal number of clusters in a dataset. A higher Dunn Index value indicates the best partitioning, characterized by well-separated clusters with tightly grouped data points. (Ramos, 2024).

3 FINANCIAL RATIOS

3.1 Stock valuation indicators

In this thesis, the chosen valuation indicators represent widely recognized metrics that usually capture an investor's interest. The selection of these indicators was primarily based on their widespread use and effectiveness. They are relatively straightforward to calculate, and the data is easily available. These indicators serve as a swift reference point for assessing the condition of companies. Significantly, they contribute to the creation of a standardized dataset. This standardization is crucial, as it facilitates the comparative analysis of diverse companies. Without such scaling and standardization measures, the comparison would be rendered impractical, impeding the suitability of the data for clustering or any other analysis purposes.

3.1.1 Return on equity (ROE)

$$ROE = \left(\frac{Net\ income\ for\ period\ t}{Average\ shareholder's\ equity\ t}\right) * 100$$

Return on equity is a common measure indicating company profitability that many analysts consider as one of the most important financial ratios for investors. It measures how much profit a company is generating from the money shareholders have invested. ROE is most useful when comparisons are made between similar companies in the same sector. (Marr, 2012).

Companies exhibiting a high Return on Equity (ROE) possess the capability to expand their operations without necessitating substantial capital investments. This enables managers to reinvest capital into enhancing business processes, thus optimizing operational efficiency, all without requiring additional capital injections from the business owners. Furthermore, a robust ROE signifies reduced reliance on external debt financing, mitigating the need to borrow funds from external sources (Marr,2012).

Calculating ROE involves a simple computation using data readily accessible from financial systems and statements. To calculate ROE, we must first calculate the shareholders equity. This is achieved by deducting total liabilities from total assets, with the values sourced from the balance sheet. (Marr,2012).

Return on Equity (ROE) serves as a crucial indicator of profitability and efficiency, with higher values generally preferred as they signify a greater return on investment. As a rough guideline, an ROE falling between 15% and 20% is commonly regarded as favourable, suggesting robust financial performance and efficient use of shareholder equity. (Marr,2012).

3.1.2 Price/earnings ratio (P/E ratio)

$$P/E = \frac{Current\ price\ per\ share}{Earnings\ per\ share}$$

The P/E ratio is considered as a cornerstone of stock market analysis. It examines the connection between a company's stock price and its earnings, serving as a fundamental gauge for investors. The P/E ratio examines historical performance and quantifies the price an investor pays per unit of a company's earnings. Alternatively, it can be construed as the duration required for investors to earn back their initial investment if the company sustains its previous year's earnings. A higher P/E ratio means that investors are paying more for each unit of net income, rendering the stock comparatively expensive in contrast to one with a lower P/E ratio. (Marr,2012).

Data for P/E ratio is sourced from the company's accounting information and the current stock market value. It can be calculated eighter by dividing the current price per share with earnings per share or by dividing current market capitalisation with earnings for the company for the year. (Marr,2012).

There is not any specific target or benchmark that says some P/E -values are better than another. Some things to keep in mind are that different industries have different P/E ranges that are considered as normal and P/E values may vary depending on the way they are calculated. It's important to use the same formula when comparing P/E values between companies. Usually, investors are

looking for the forward P/E that is calculated by dividing share price with expected earnings for the year. (Marr,2012).

3.1.3 Price to book (P/B)

$$P/B\ Ratio = \frac{Market\ Price\ per\ Share}{Book\ Value\ per\ Share}$$

The price-to-book (P/B) ratio is one the most common metrics used by value investors when seeking undervalued stocks. The price-to-book ratio is significant as it aids investors in determining if a company's market price appears justified in relation to its balance sheet. By investing in undervalued stock, investors anticipate being rewarded when the market recognizes the stock's undervaluation and adjusts its price accordingly, aligning with the investor's analysis. Typically, the company's market value is higher than the book value. P/B ratios below 1.0 are generally regarded as favourable investments by value investors, but it is important to notice that a "good" P/B ratio is relative to a business and its industry. (Fernando 2022).

Data needed for calculating the P/B ratio can be acquired from the balance sheet. To calculate the P/B ratio, we first need to get the book value per share (BVPS). This can be calculated by dividing the total shareholder equity by the total number of outstanding shares. After the BVPS is calculated, we can obtain the P/B ratio by dividing the market price per share by the book value per share (BVPS). (Fernando 2022).

3.1.4 Debt-to-equity (D/E) ratio

$$Dept - to - equity\ ratio = \frac{Total\ liabilities}{Total\ equity}$$

Debt-to-equity (D/E) ratio is a crucial measure used to provide insight between debt and shareholders' equity, as the funding for a company's operations can be sourced from either shareholder investments (equity) or debt acquisition. Borrowing is not necessarily a bad thing, as it can facilitate growth. If a company has more debt than equity meaning high leverage or gearing, it typically indicates an aggressive approach to growth through borrowing. Taking on debt can amplify

earnings beyond the level achievable solely with shareholder capital. Consequently, if a company can enhance earnings by a greater margin than the costs and interest payments associated with the debt, it is generating value. (Marr,2012).

Too careless borrowing could potentially result in business challenges and expose earnings to unpredictable interest expenses. Businesses with extensive leverage ratios are especially vulnerable during a recession or turndown. Usually, investors are looking for a lower Debt-to-Equity (D/E) ratio especially if they are interested on reducing the risk of potential loss in case of liquidation. A low debt-to-equity (D/E) ratio could suggest that a company is not utilizing its capacity for debt, which can potentially boost its profits. (Marr,2012).

The Debt-to-Equity (D/E) ratio is calculated using data from company's balance sheet. In simple terms, it's dividing a company's total liabilities by its shareholders' equity. The formula gives a result of percentage score. Above 1 score means that funding by debt outweighs funding by equity. For example, if the D/E-ratio is 1,5 it means that for every euro of company X owned by shareholder, it owes 1,5 euro to creditors. (Marr,2012).

A good Debt-to-Equity (D/E) ratio -value strongly depends on the industry and the circumstances. The number should always be compared with the industry average. It is important to keep in mind that there are many ways to calculate Debt-to-Equity (D/E) ratio, and therefore It's crucial to be aware what types of debt and equity are being used in the calculation. (Marr,2012).

3.1.5 Dividend yield

$$Dividend\ yield = \frac{Dividend\ Per\ Share}{Company\ Share\ Price}$$

Dividend yield is an indicator that illustrates the proportion of dividends a company distributes relative to its share price. This is determined by dividing the total annual dividend amount per share by the market price per share. The currency used for calculation (whether pounds, pence, dollars, or cents) is immaterial, as long as consistency is maintained for both amounts. It is important to notice that share prices change constantly, which affects the dividend yield -ratio. When a

company's price increases, its yield decreases, and conversely, when its price decreases, its yield increases. (Michael & Pratt 2024).

The data required for calculating the dividend yield is sourced from the company's financial statements. Dividend yield is calculated by dividing the dividend per share by company share price. (Michael & Pratt 2024).

Various factors influence a company's dividend yield, such as market conditions and corporate performance, but the primary determinant is the company's share price. Typically, as share prices rise, dividend yields decrease unless the company chooses to increase its dividend pay-out. (Michael & Pratt 2024).

Determining what constitutes a 'good' dividend yield has no universal answer, but generally, dividend yields ranging from 2% to 5% are deemed favourable, with yields surpassing this range possibly indicating attractive investments but also carrying inherent risks. (Michael & Pratt 2024).

3.1.6 Earnings Yield

$$Earnings\ Yield = \frac{Earnings\ Per\ Share\ (EPS)}{Company\ share\ price}$$

Earnings yield represents the percentage of a company's earnings per share and is also known as the inverse of the Price-to-Earnings (P/E) ratio. Investment managers frequently utilize the earnings yield to ascertain optimal asset allocations, while investors leverage it to gauge whether assets appear undervalued or overvalued. The earnings yield is calculated by dividing the earnings per share (EPS) for the latest 12-month period by the current market price per share. (Mitchell 2022).

Earnings yield is not as commonly used as the P/E ratio for investment valuation, but it can be useful in assessing the return on investment. However, for equity investors, prioritizing the growth of investment values over periodic income is often paramount. Hence, investors may lean towards value-based metrics like the P/E ratio when making stock investments. Nonetheless, both metrics offer equivalent information, albeit presented differently. (Mitchell 2022).

An overvalued investment tends to reduce earnings yield, whereas an undervalued investment tends to elevate earnings yield. This is because as the stock price rises without a commensurate increase in earnings, the earnings yield declines. Conversely, if stock prices decrease while earnings remain constant or increase, earnings yield rises. Value investors typically pursue the latter situation. (Mitchell 2022).

3.1.7 Earnings per share EPS

$$Earnings\ per\ Share = \frac{Net\ Income - Preferred\ Dividends}{End - of - Period\ Common\ Shares\ Outstanding}$$

EPS is one of the many indicators investors use to pick stocks. By dividing a company's share price by its earnings per share, investors can gauge the stock's value in relation to how much the market is willing to pay for each dollar of earnings. (Fernando, 2023).

EPS indicates the amount of profit generated by a company for each share of its stock and serves as a commonly utilized metric for assessing corporate worth. Earnings per share stands as a crucial metric in assessing company's absolute profitability and plays a significant role in calculating the price-to-earnings (P/E) ratio, with EPS representing the "E" in P/E. (Fernando, 2023).

To calculate a company's EPS, data from the balance sheet and income statement must be acquired. Data needed is the period-end quantity of common shares, dividends distributed on preferred stock (if applicable), and the net income or earnings. The EPS is calculated by dividing the net income by the number of outstanding shares. (Fernando, 2023).

Determining whether an EPS is good depends on factors like the company's recent performance, its competitors' performance, and analysts' expectations. As a rule of thumb, we could state that the higher a company's EPS, the more profitable it is. It's important to note that even if a company reports an increase in EPS, the stock price may still drop if analysts expected higher. Conversely, a decreasing EPS might still cause a price increase if analysts anticipated a worse outcome. It's crucial to assess EPS relative to the company's share price, for example, by examining the company's P/E or earnings yield. (Fernando, 2023).

4 DATA

This chapter outlines the methodology of data acquisition and preparation for the analysis conducted in this thesis. It details the origin of the dataset, the techniques employed to gather data, the approach taken to address missing values, and an introduction to the dataset itself.

4.1 Source of data

The financial data for this thesis was sourced from Yahoo Finance, that is a well-known platform providing comprehensive financial information. A significant factor in selecting Yahoo Finance was its inclusivity of Finnish companies, ensuring that the study's dataset encompasses a comprehensive representation of the Finnish stock market. The crucial values considering this thesis can be found from the financial statement and balance sheet, so it also played a huge role in selecting this approach.

To retrieve financial data from Yahoo Finance, the yfinance library was employed. This library offers a comprehensive access point to an extensive array of financial data, enhancing the efficiency of the data acquisition process through its user-friendly functionalities. Data retrieval via yfinance is done by using ticker symbols.

4.2 Data collection

Collecting data for large datasets, like the one needed for clustering analyses, is challenging if done manually. The required information is usually accessible on the investor relations website of each company. However, extracting information from multiple individual sites would require a significant amount of time, especially if we need to obtain data for more than a few companies.

In this thesis, a systematic approach is used to collect and process financial data from Yahoo Finance, utilizing the yfinance library. This method involves a series of functions each designed to fetch specific financial metrics such as total revenue, price-to-book (P/B) ratio, price-to-earnings (P/E) ratio, dividend yields, debt-to-equity ratio, return on equity (ROE), earnings per share (EPS),

and earnings yield. These functions use the ticker symbols of companies to precisely access and extract the relevant financial data for predefined fiscal years.

The data collection process includes error handling to pinpoint and handle issues related to data accessibility, such as invalid ticker symbols or missing entries. Each function incorporates checks to validate the availability of necessary data before conducting any computations. If data for a required field is missing or the ticker symbol is incorrect, the function logs these errors, facilitating effective troubleshooting and ensuring the integrity of the data collection process.

The financial data retrieved through these functions is then systematically organized into a structured dataframe that provides a foundation for financial analysis of company performance. This methodological approach ensures the reliability and accuracy of the data and enhances the reproducibility of the research.

The collected financial metrics represent the state of the company as of the final market day of 2022. The 'Gains %' metric calculates the change in share price over the observation period, using the final market day's share price of 2023 as the reference point. This metric is used to capture the price fluctuation of the stock within the specified timeframe. It is important to note that this thesis only focuses on the variation in stock prices, and therefore any dividends are excluded from the calculation of the yearly gains percentage.

The data was obtained using a script that processed a list of 199 Finnish stock tickers. The script effectively managed to retrieve most of the necessary data across these tickers. Successfully compiled values are presented in a bar graph in Figure 7 as non-null count of each metric. Given the comprehensive nature of the data retrieved, achieving total coverage of all possible data points was not essential for the clustering task and therefore not pursued.

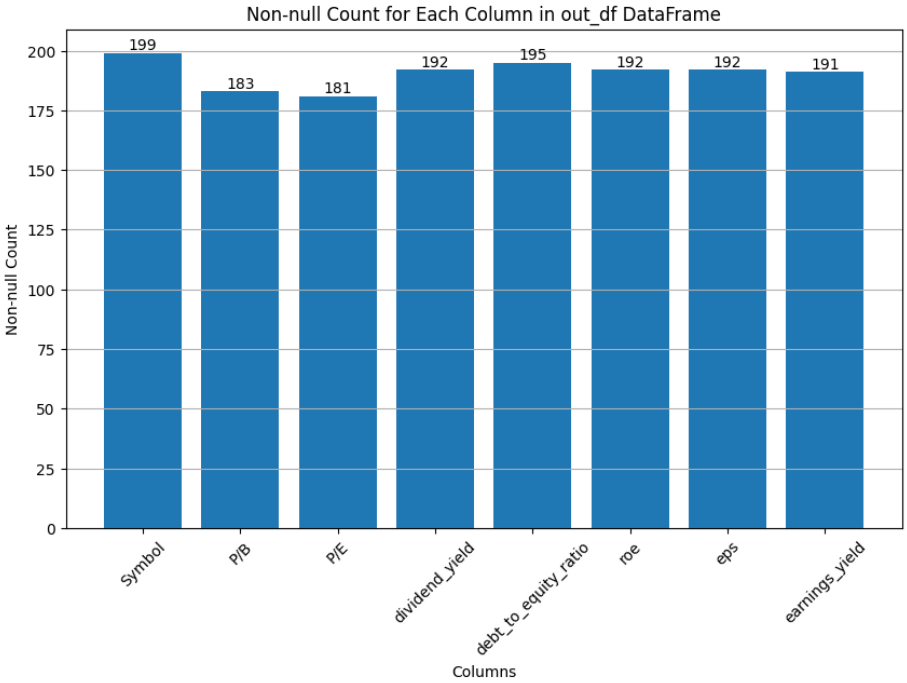

FIGURE 7 Non-null count of the collected parameters. n=199

4.3 Managing missing data

At this stage the dataset contains values from 199 companies. However, the dataset includes missing values as we can see from the figure 7. Before the dataset is suitable for further analysis and clustering, the missing values needs to be handled.

In financial datasets, ensuring data accuracy and precision is critical, as using methods that incorporate estimated values might lead to a distortion of actual financial indicators. For this reason, complete case analysis or listwise deletion is used. This method, as outlined by Liu (2016), involves the exclusion of all records from the analysis that contain any missing values. Such an approach ensures that only complete data is used in the analysis, thus avoiding the introduction of variability that might arise from imputation methods.

To manage the missing values in the dataset, the dropna-function was employed, resulting in the exclusion of 20 entries leaving the total number of 179 companies. This constituted a reduction of approximately 10% of the total dataset, optimizing the data for the clustering task.

Table 1 provides an overview of the central tendencies, dispersion, and shape of the dataset's distribution after removing the missing values. Values like count, mean, standard deviation, min,

max and the quartiles provide a quick way to identify potential anomalies or outliers that may require further investigation.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield
count	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000
mean	2.550195	40.268646	0.043018	1.247411	0.066575	0.256294	-0.150248
std	3.728573	323.648726	45.516576	8.529193	0.636918	1.601251	1.386480
min	-20.415521	-413.500023	-406.962578	-104.384166	-5.889117	-10.570000	-17.714286
25%	1.075855	-1.244207	0.000000	0.632847	-0.015447	-0.050000	-0.018206
50%	1.717504	11.631747	2.393066	1.136730	0.088755	0.230000	0.039156
75%	2.967175	20.481243	4.631612	1.880421	0.167672	0.830000	0.076045
max	20.509224	4145.841570	237.326176	17.941663	3.071733	3.970000	2.364865

TABLE 1 Summary statistics of the data frame

4.4 Dataset overview

After removing the missing values, the dataset consists of 179 observations. The summary of the dataset statistics is introduced in the table 1. Dataset is also described in figure 8 in a box plot chart. The chart is flat and not very informative because of the extreme outliers in the dataset.

Before any other pre-processing measures, it is important to gain more understanding of the data and pinpoint the problems in it. Therefore, a deeper analysis of each financial metric is presented in the next following subchapters.

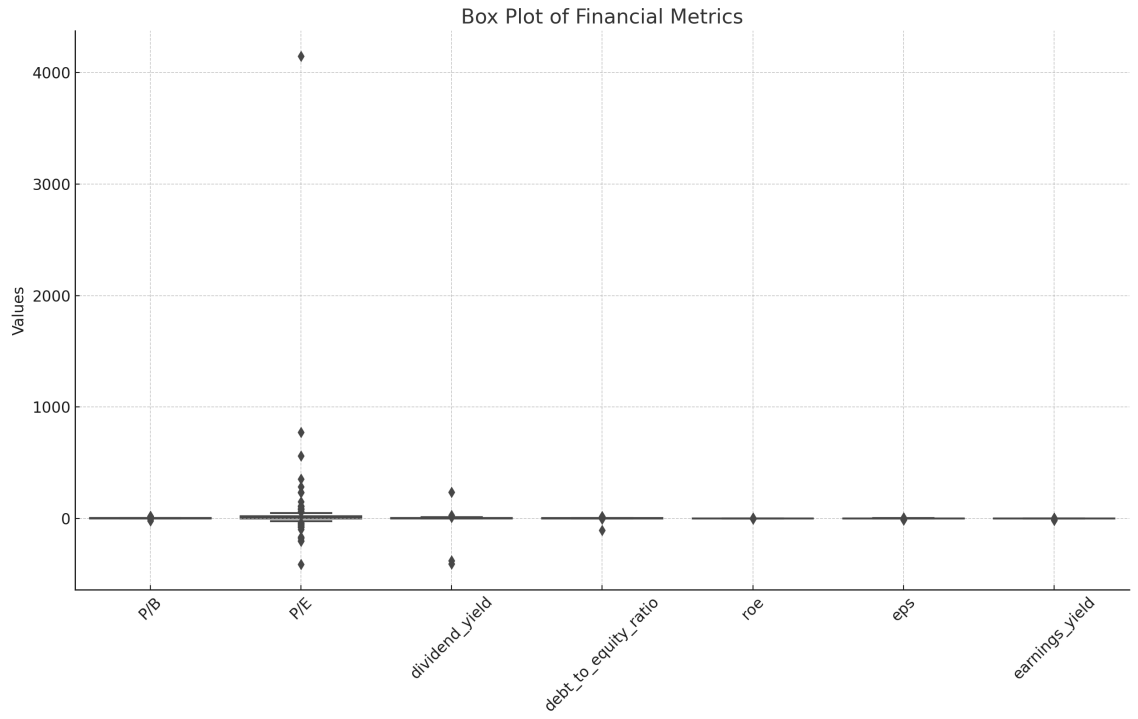

Figure 8 Boxplot of the unbalanced dataset, n=179

4.4.1 P/B ratio

The dataset consists of 179 observations of the Price to Book (P/B) ratio. On average, the P/B ratio across all companies is approximately 2.55, suggesting that the market price is typically about 2.55 times the book value of the company's equity. The standard deviation at 3.73 indicates a substantial spread of P/B ratios across the dataset.

The P/B ratio values range from -20.42 to 20.51. The negative minimum value can indicate companies with a market price below their stated book value, which might occur during financial distress or when the market perceives that the book value is overstated. P/B ratio greater than one, particularly the high maximum value, suggests that some companies are valued by the market at a premium above their book value, often reflecting expectations of future growth or profitability.

Figure 9 provides a visualisation of the P/B ratio. The boxplot on the right indicates a dataset with a notable range of values. The outliers reveal that some companies have exceptionally high or low P/B ratios compared to the broader dataset. The positioning of the median suggests a certain skewness in the distribution, indicating a tendency towards higher or lower P/B values among most companies. The spread between the first and third quartiles highlights the variability in P/B ratios,

emphasizing differences in market valuation relative to book value across companies. This variability and outliers suggest diverse financial health and market perceptions among the companies in the dataset.

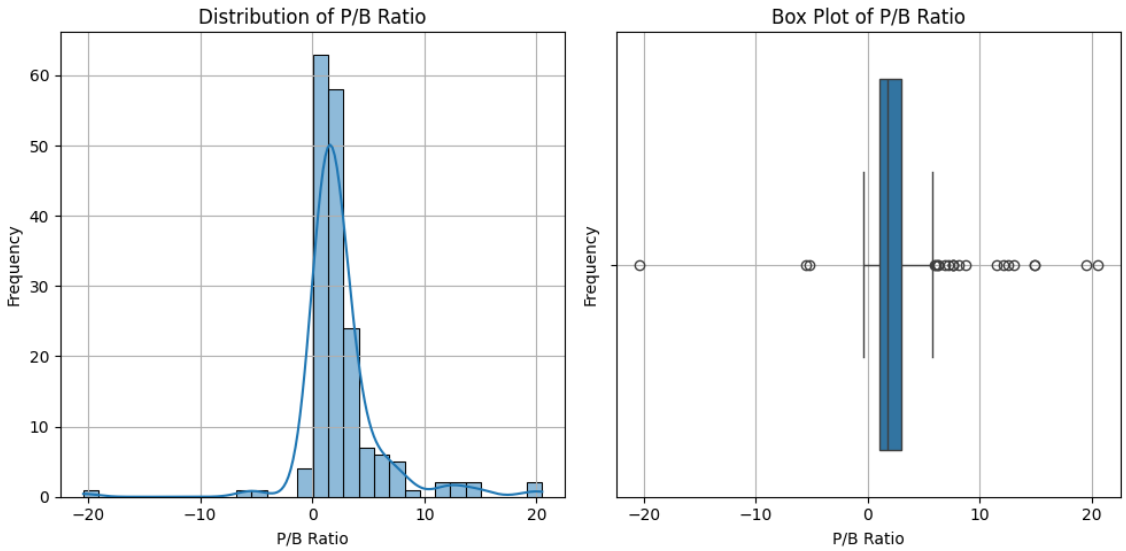

FIGURE 9 Distribution of the P/B ratio in unbalanced dataset, n=179

The histogram in the figure 9 reinforces the perception made earlier from the statistical overview and the box plot by revealing the right-skewed distribution in the P/B-ratio more clearly. Most of the data points are concentrated on the lower side of the scale, with a long tail extending towards higher P/B ratio. The analysis reveals that a few companies have exceptionally high ratios, as well some companies with negative ratios.

From the peak of the distribution, we can clearly see that most of the companies lie in the lower range of the P/B ratio. This suggest that the mode of the P/B ratio is relatively low. Due the skewness, the mean is likely to be higher than the median.

The long tail towards the higher P/B ratios suggests that there are some companies which the market values much more in terms of price compared to their book value. These might be companies with significant intangible assets, strong brand value, or expected future growth. The presence of a long tail in the distribution also indicates that outliers and extreme values significantly influence the overall distribution.

4.4.2 P/E ratio

As we could already state from the figure 8, the P/E ratio has a broad range of values, from highly negative (-413,50) to exceptionally high positive (4145,84). The box plot in Figure 10 shows a significant spread in P/E ratios with several outliers. The histogram in Figure 10 provides a detailed view of the distribution.

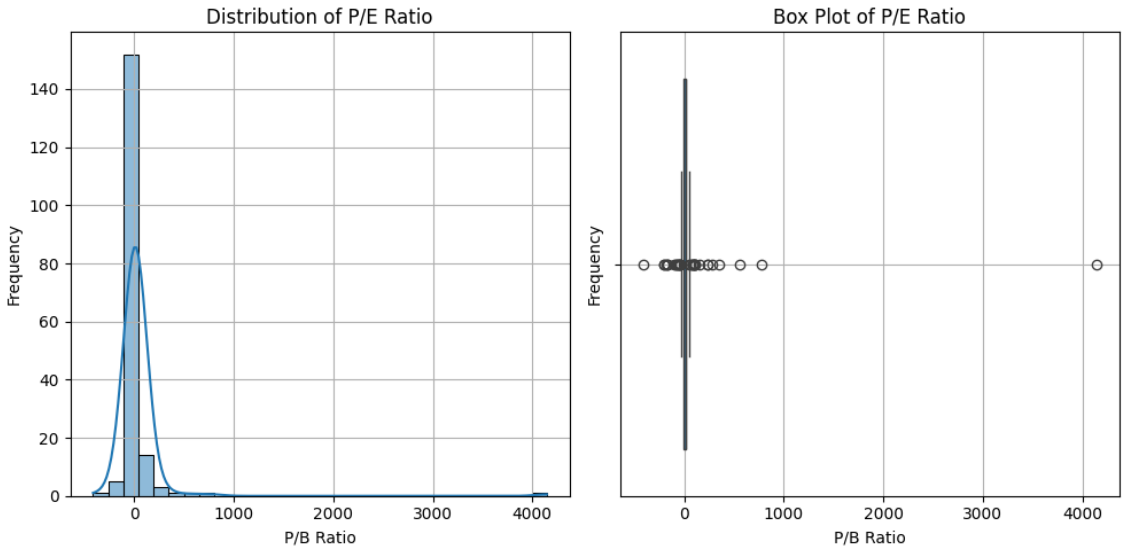

FIGURE 10 Distribution of the P/E ratio in unbalanced dataset, n=179

4.4.3 Dividend yield

Like we learned from the chapter 3, the Dividend Yield is calculated as the ratio of yearly dividends per share to the stock price, expressed as a percentage. The average Dividend Yield across all companies is 0.043%, suggesting that the typical company in the dataset returns 0.043% of the stock price as dividends to shareholders annually. The standard deviation, 45.51, indicates a considerable spread of values around this mean, pointing to diverse dividend policies across the dataset. Dividend Yield values are spanning from -406 to 237. A negative yield is unusual as it suggests that dividends may have been distributed during a period where the share price was exceptionally low or negative earnings were reported, leading to a negative ratio when annualized. The high maximum of 237 might reflect special dividends or instances where the stock price is quite low, amplifying the yield percentage or even an error in the data collection.

Regarding the interquartile range of 4.63, it signifies that 50% of the companies have dividend yields from 0 to 4.63, with a median of 2.39. The zero or near-zero values at the 25th percentile indicate that a quarter of the companies do not pay dividends, or pay very minimal dividends, in relation to their stock price. We can visually inspect the comparison between dividend paying and non-dividend paying companies in the figure 11.

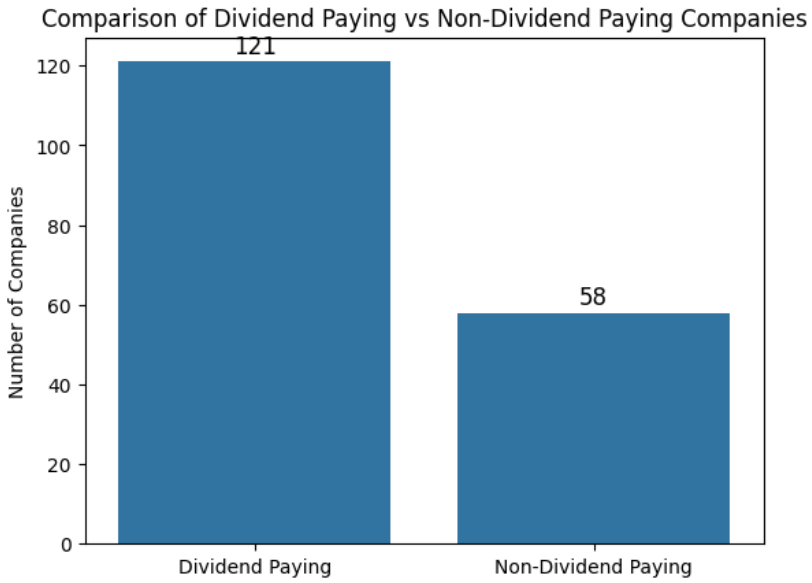

FIGURE 11 Dividend paying vs non-dividend paying companies, n=179

The figure 12 demonstrates the presence of outliers on both the lower and higher ends. The presence of negative and extremely high dividend yields is noteworthy. Negative yields could result from companies paying dividends while reporting negative earnings, or they may indicate data recording issues. The high yields might be attributed to special dividend payments, smaller base share prices, or exceptional company policies.

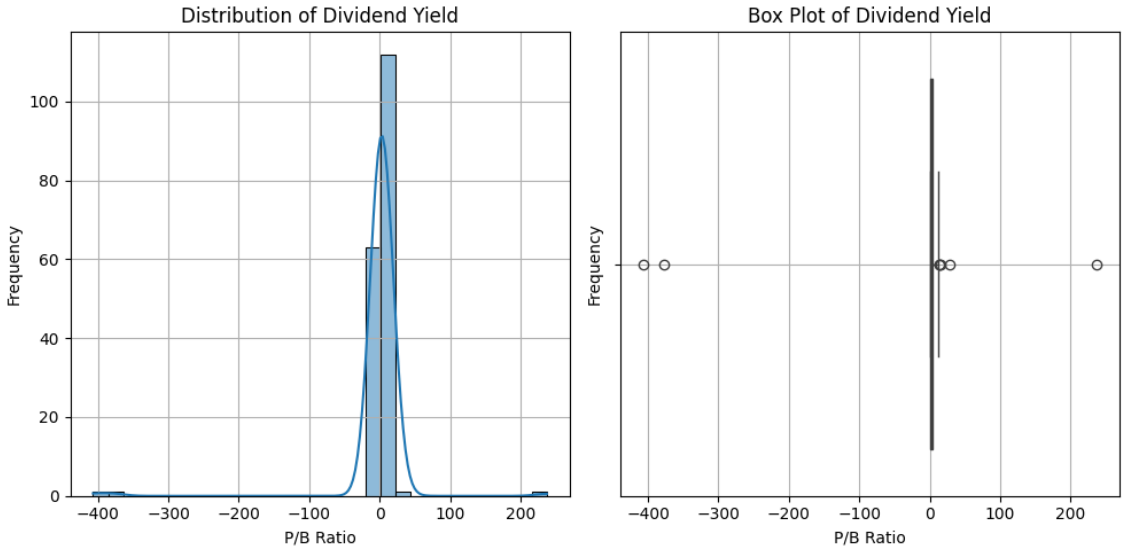

FIGURE 12 Distribution of the Dividend Yield ratio in unbalanced dataset, n=179

4.4.4 Dept to equity ratio

The average Debt to Equity Ratio across all companies is approximately 1.247, indicating that on average, companies have about 1.25€ of debt for every euro of equity. The standard deviation is 8.529, which points to a considerable variation in the Debt-to-Equity Ratio among the companies. This wide variation is further highlighted by the range of values, stretching from a minimum of -104.38 to a maximum of 17.94. Such a negative minimum suggests that some companies have more equity than debt, potentially indicating negative debt or accounting adjustments.

At the 25th percentile, the Debt-to-Equity Ratio is about 0.633, indicating that the bottom quarter of companies have 0.63 of debt for every euro of equity or less. The median (50th percentile) is approximately 1.137, and the 75th percentile is around 1.880, which implies that half of the companies have a Debt-to-Equity Ratio between 0.633 and 1.880.

The dept to equity ratio in the dataset indicates a diverse range of financial leverage among the companies. Most of the companies maintain a balanced or moderately leveraged structure, but there are also some exceptions with extremely high debt levels. As we can observe from the figure 13, the presence of negative values and outliers warrants further investigation to understand the specific circumstances of those companies.

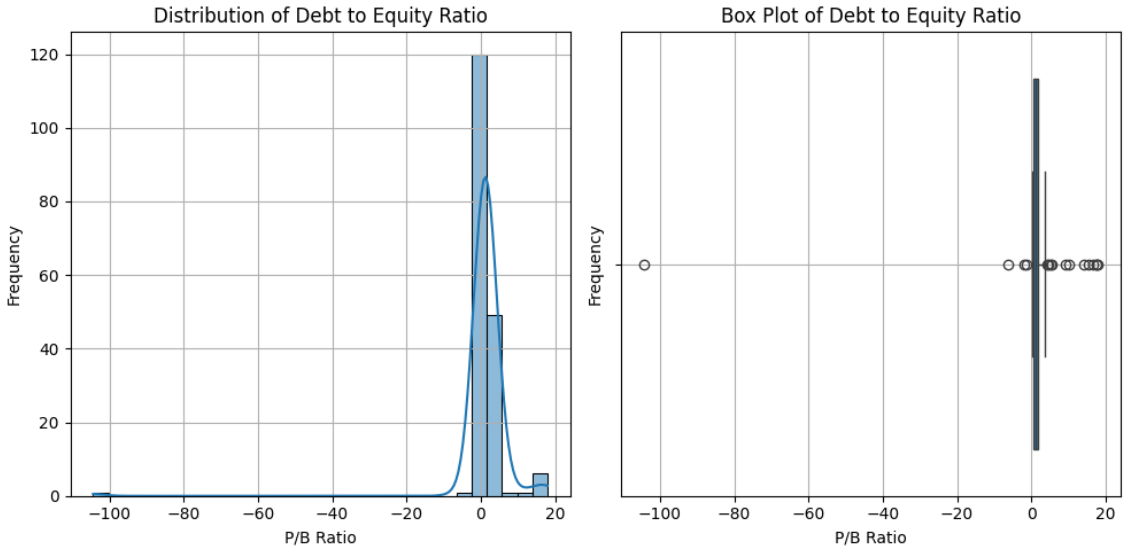

FIGURE 13 Distribution of the Debt-to-Equity Ratio in unbalanced dataset, n=179

The histogram in the figure 13 shows the distribution of the dept to equity ratios for the companies in the dataset. The histogram leans towards the left side, which means most companies in the dataset have lower dept to equity ratios. This indicates that most of these companies are not heavily reliant on debt compared to their equity. The figure also shows that while most companies have lower ratios, there are some companies with much higher ratios. These are not as common, but they stand out because their ratios are significantly different from the majority.

4.4.5 Return of equity (ROE)

The average ROE across all companies is 0,067, indicating that on average, companies generate a 6,7% return on shareholders' equity. The standard deviation 0.637, points some variation in ROE among the companies. This is further evidenced by the range of ROE values, stretching from a minimum of -5.889 to a maximum of 3.072. The 25th percentile is about -0.015, the median (50th percentile) is approximately 0.089, and the 75th percentile is around 0.168. This implies that half of the companies have an ROE between -1.5% and 16.8%.

The histogram in the figure 14 shows a left skewed distribution. This skewness is primarily due to a few companies with extremely low ROE values. The presence of several points beyond the whiskers of the box plot in the figure 14 indicates outliers in the ROE data. These outliers are companies with unusually high or low ROE values compared to the rest.

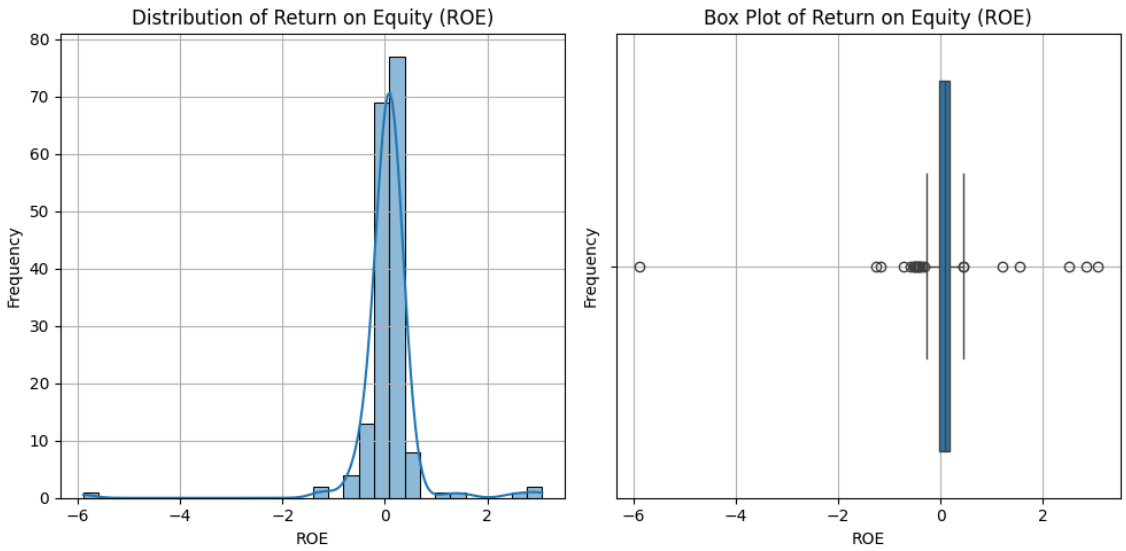

FIGURE 14 Distribution of the Return on Equity (ROE) Ratio in unbalanced dataset, n=179

4.4.6 Earnings per share (EPS)

On average, companies have an EPS of 0,256, indicating that typically, companies in the dataset are earning about 0,256€ per share. The standard deviation is 1,601, revealing substantial variation in EPS among the companies. The range of EPS values is considerable, extending from a minimum of -10,57 to a maximum of 3,97. At the 25th percentile, the EPS is approximately -0,05, the median EPS is 0,23, and the 75th percentile is around 0,83. This suggests that half of the companies have an EPS between -0,05€ and 0,83€.

The histogram for the EPS, displayed in figure 15, reveals a right-skewed distribution, which indicates that the bulk of the companies have lower EPS values with a tail extending towards higher EPS values. This skewness is primarily due to several companies with notably high EPS values. Conversely, the presence of several data points beyond the whiskers of the box plot in Figure 15 signifies outliers in the EPS data. Companies within the observed group show considerable variation in profitability, with some having exceptionally high or low EPS values compared to the dataset median, indicating the presence of outliers.

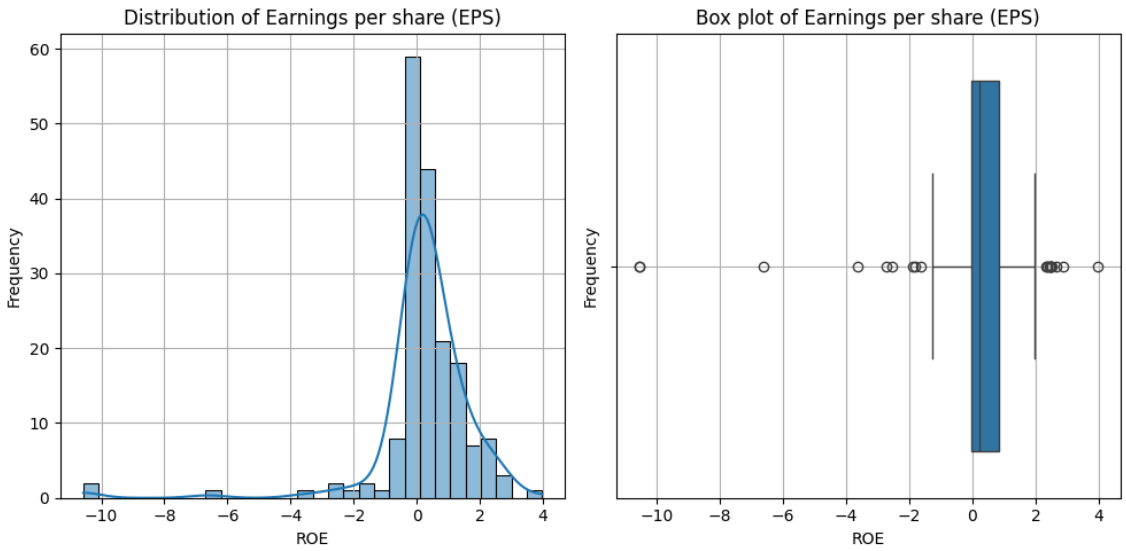

FIGURE 15 Distribution of the Earnings per share (EPS) in unbalanced dataset, n=179

4.4.7 Earnings yield

On average, companies have an earnings yield of approximately -0.15, which may suggest that, typically, companies in the dataset have a negative yield when it comes to the earnings generated per unit of share price in the inspected year. The standard deviation of 1.38 reveals a substantial variation in earnings yield among the companies. The range of earnings yield values is considerable, extending from a minimum of about -17.71 to a maximum of 2.36. This range indicates that there are some companies with negative earnings relative to their share price.

At the 25th percentile, the Earnings yield is around -1.82%, the median earnings yield is approximately 3.92%, and the 75th percentile is about 7.60%. This suggests that half of the companies have an earnings yield between -1.82% and 7.60%.

The histogram for the earnings yield in the figure 16 reveals a distribution with a significant left skew, given the negative average and the presence of extreme negative values. This skewness is primarily due to several companies with notably low negative earnings yields. Conversely, the presence of several data points beyond the whiskers of the boxplot in the figure 17 signify outliers in the Earnings Yield data. These outliers represent companies with exceptionally high or low Earnings Yields when compared to the median of the dataset, illustrating the substantial variation in company performance within the observed group.

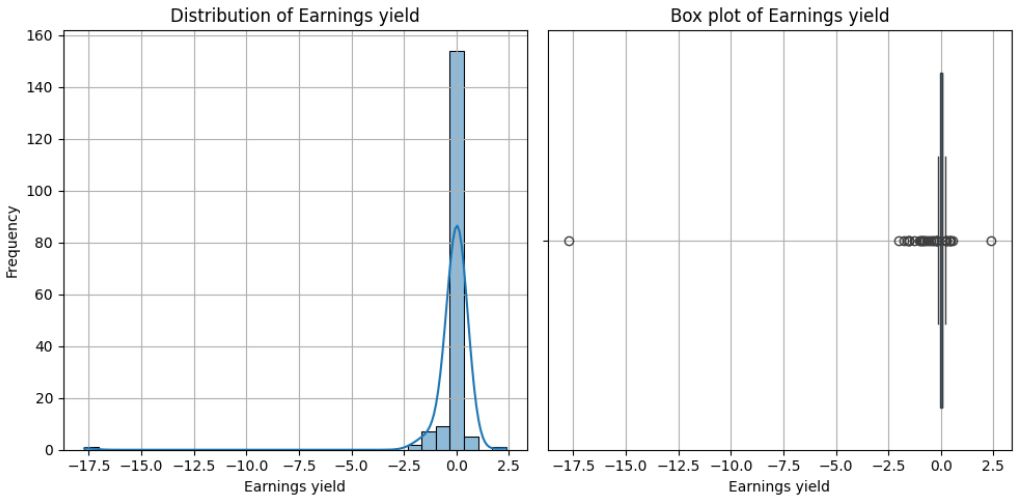

FIGURE 16 Distribution of the Earnings yield in unbalanced dataset, n=179

4.5 Balancing the dataset

This chapter describes the approach for balancing the dataset by outlier removal. The analysis in the previous chapter indicated the need for balancing. This step is essential in preparing the dataset for clustering analysis, as the effectiveness of clustering algorithms like K-means is affected by the presence of outliers, as outlined in Zhang (2021). Extreme values can disproportionately affect cluster formation, leading to clusters that may not accurately capture the actual financial characteristics of the dataset. Removing these outliers enables the algorithms to discern more meaningful and representative clusters based on general financial trends.

Another crucial aspect of outlier removal is its role in facilitating comparative analysis across different entities. Outliers can obscure true financial comparisons by introducing significant disparities. By standardizing the dataset through the exclusion of these extreme data points, we enhance its uniformity, enabling more accurate and reliable comparative studies.

Balancing is done by defining a function for removing outliers in the data, which is then applied to multiple columns to reduce the impact of extreme values on statistical models. Outliers are defined quantitatively as observations that lie outside the 5th and 95th percentiles of the data distribution in each column. This threshold ensures that the most extreme deviations are excluded while preserving the dataset's overall integrity and diversity. By focusing on this range, the function effectively narrows the data to the central 90% of observations, excluding the most extreme 5% at both ends of the spectrum.

The implementation of this function involves iterating over a list of columns intended for outlier removal. For each column, the function calculates the lower and upper quantiles based on the 0.05 and 0.95 thresholds, respectively. The data is then filtered to include only those rows where the column values fall within the specified quantile range. This filtering is done sequentially for each column listed, which means that the criteria are applied one after another, potentially leading to a significant reduction in data size depending on the distribution of values across multiple dimensions.

The presence of outliers can significantly skew the data distribution as we noticed in previous chapter. By eliminating these extreme values, we can reduce the skewness and ensure that the data is more accurately representing the central tendencies and variations. This leads to a more concentrated and presumably more representative subset of the original dataset, albeit at the cost of a reduced sample size.

From statistical perspective, balancing further reduced the number of observations from 179 to 87 making the final sample size n=87. With a sample size of n=87 and assuming a conservative approach with a 95% confidence level, the margin of error is 10.51%. This high margin of error highlights the trade-off between data quality (reduced influence of outliers) and statistical precision (wider confidence intervals).

This reduction in dataset might seem dramatical, but it was an expected outcome due to the nature of the filtering process, which is designed to remove extremes systematically across several dimensions of the data. This leads to a more concentrated and presumably more representative subset of the original dataset. The results of the balancing can be observed in figure 17 in a pairwise comparison before and after balancing. The unbalanced dataset is on the left side colored in red and the balanced dataset in the right with green color.

FIGURE 17 Pair-wise comparison of distribution in financial metrics before and after data balancing

5 RESULTS

The objective of this thesis was to compare and investigate the suitability of unsupervised learning and clustering algorithms in stock classification using common value investing metrics. The valuation metric for examination of the “goodness” of clustering results is Gains% which is derived from the Lastprice2022 and lastprice2023 as percentage values. A higher value is better.

In the following chapters, each clustering algorithm is examined. These chapters begin by presenting the overview principles of the data processing pipeline, followed by a justification for the selection of a specific number of clusters. Subsequently, the results of the clustering results of each method are detailed. The final chapter undertakes an evaluation and comparison of clustering outcomes together. The suitability of these algorithms for identifying financial characteristics within the data that could potentially forecast short-term future returns is also assessed.

5.1 K-means Clustering

In this thesis, K-Means clustering was applied to the balanced dataset to identify distinct groups based on various financial metrics. The balancing process involved removing extreme values from the dataset, particularly focusing on the 5th and 95th percentiles of key financial indicators. This method ensured a more representative and less skewed dataset for clustering.

5.1.1 Choosing the k-value

To determine a suitable number of clusters (k-value) for K-Means clustering on the balanced dataset, I used the Elbow Method and the Silhouette Score. The Elbow Method plot in the figure 18 shows how the inertia (sum of squared distances from each point to its assigned centre) decreases as the number of clusters increases. The 'elbow' point in the graph is where the rate of decrease sharply changes. This point suggests the optimal number of clusters for the dataset.

The Silhouette Score plot in figure 18 indicates the average silhouette score for different numbers of clusters. A higher silhouette score suggests that the data points are well matched to their own

cluster and distinct from other clusters. The optimal number of clusters typically corresponds to the number of clusters with the highest silhouette score.

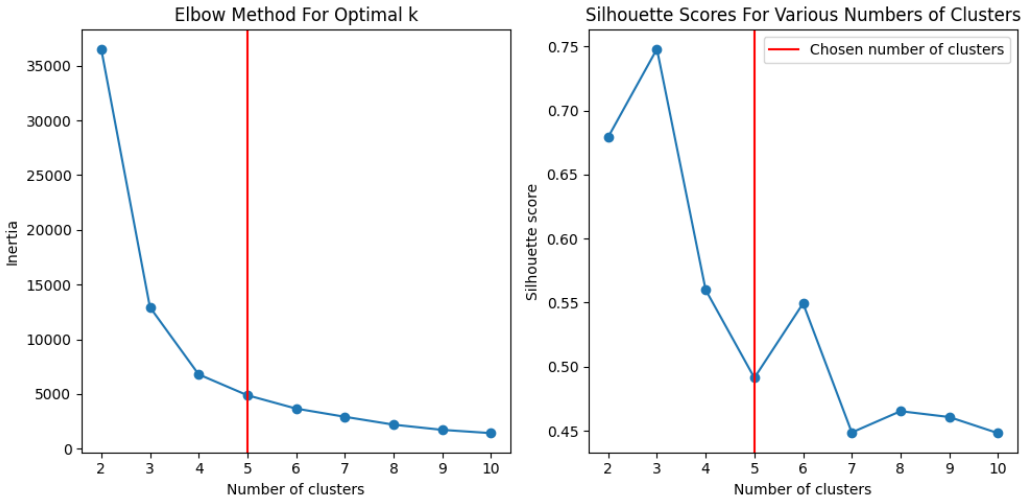

FIGURE 18 Elbow method for optimal k-value and Silhouette scores

After examining the Elbow Method and Silhouette Score plots, I chose to use 5 clusters for the balanced dataset, because as can be seen from figure 18, the decrease in inertia becomes less pronounced around 5 clusters. While the 'elbow' isn't sharply defined, this point represents a balance between a significant reduction in inertia and avoiding too many clusters. The silhouette score for 5 clusters is also among the higher values, though not the absolute highest. This suggests that the clusters formed with 5 clusters have reasonable cohesion and separation from each other. Choosing 5 clusters represents a compromise between the Elbow Method and Silhouette Score indications, and it should provide meaningful clustering results.

5.1.2 K-Means Clustering Results

The K-Means clustering algorithm revealed distinct groups with unique characteristics from the dataset. Figure 19 includes two key visualisations: a bar plot showing the distribution of data points across the K-Means clusters and a boxplot detailing the distribution of 'Gains%' within each cluster. The bar plot reveals the number of companies in each cluster, highlighting the diversity in cluster sizes and the prevalence of certain financial characteristics. In contrast, the boxplot of 'Gains%' offers insights into the financial performance of companies in each cluster from 2022 to 2023, visualizing the range, median, and outliers of stock performance metrics across the clusters.

The disparity in cluster sizes is significant, as it highlights the diversity of the financial profiles captured by the K-Means clustering algorithm. Some clusters encompass a larger number of stocks, suggesting more common financial characteristics among these companies, while other clusters are smaller, potentially indicating more unique or specific financial profiles. This initial view sets the stage for a deeper exploration into the distinctive features and investment profiles inherent within each cluster.

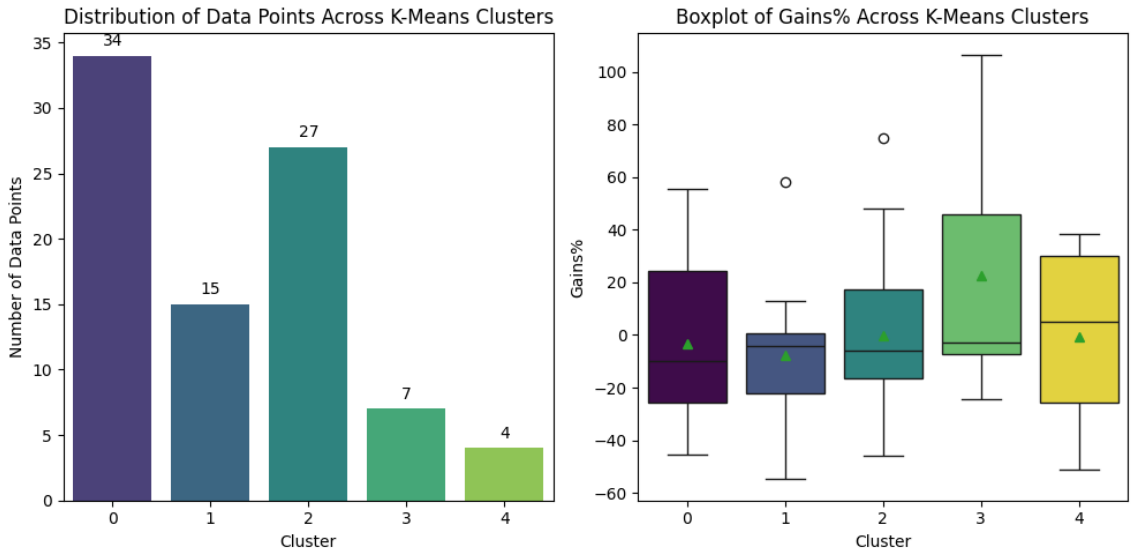

FIGURE 19 Stock count and gains% in each cluster using K-Means with k=5, n=87

Table 2 illustrate the characteristics of Cluster 0, which included 34 companies. This cluster displayed an average Price to Book (P/B) ratio of 1.88 and a Price to Earnings (P/E) ratio of 37.00. Additionally, Table 2 details an average dividend yield of 2.36%, a debt-to-equity ratio of 1.09, and an average stock price decrease of -3.34%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.0	34.000000
mean	1.875443	37.004451	2.357497	1.088020	0.068606	0.296029	0.044170	0.0	-3.339348
std	1.028594	27.066538	2.082567	0.555099	0.043878	0.280940	0.026517	0.0	30.526002
min	0.522064	9.216007	0.000000	0.256057	0.010006	0.019000	0.007028	0.0	-45.592948
25%	1.084256	19.236393	0.000000	0.530847	0.030753	0.049250	0.024703	0.0	-25.616907
50%	1.716202	26.760004	2.050933	1.164937	0.063000	0.265000	0.041914	0.0	-10.037597
75%	2.444285	45.664641	3.624818	1.484022	0.095884	0.377500	0.054058	0.0	24.342784
max	5.227362	108.225745	7.878799	1.973371	0.151050	1.330000	0.117820	0.0	55.519095

TABLE 2 Descriptive statistics for companies in cluster 0 Identified by K-Means

Table 3 presents a statistical overview of Cluster 1, which comprises 15 companies. This cluster exhibits an average Price to Book (P/B) ratio of 5.18 and a Price to Earnings (P/E) ratio of 21.77. The average dividend yield for these companies stands at 1.84%, with a debt-to-equity ratio of 1.16. The cluster recorded an average stock price decline of 7.84%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.0	15.000000
mean	5.180139	21.772624	1.839873	1.158848	0.257336	0.886271	0.067276	1.0	-7.837483
std	1.850113	10.308325	1.671751	0.526252	0.070732	0.655387	0.041325	0.0	24.987094
min	2.481016	7.500000	0.000000	0.431671	0.131845	0.080000	0.030201	1.0	-54.853803
25%	3.319137	13.522626	0.390820	0.764075	0.195066	0.327036	0.035809	1.0	-22.261620
50%	5.426953	20.412484	1.660847	1.033105	0.270190	0.780000	0.053309	1.0	-3.971782
75%	6.553775	29.016783	3.145585	1.577472	0.318406	1.345000	0.075802	1.0	0.446591
max	8.084153	43.485186	4.435484	2.171254	0.344501	2.330000	0.168067	1.0	58.167104

TABLE 3 Descriptive statistics for companies in cluster 1 Identified by K-Means

Table 4 provides a statistical summary of Cluster 2, which consists of 27 companies. This cluster exhibits an average Price to Book (P/B) ratio of 1.78 and a Price to Earnings (P/E) ratio of 11.93. The companies in this cluster have a higher dividend yield of 5.30%, a debt-to-equity ratio of 1.29, and experienced nearly break-even stock price changes with an average of -0.26%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	27.000000	27.000000	27.000000	27.000000	27.000000	27.000000	27.000000	27.0	27.000000
mean	1.775616	11.927653	5.304800	1.293195	0.152741	1.050211	0.096863	2.0	-0.255630
std	0.626443	3.299771	1.814280	0.590435	0.050240	0.552702	0.037522	0.0	26.939933
min	0.654634	6.348980	2.114767	0.441682	0.063286	0.090000	0.042999	2.0	-46.002957
25%	1.341201	9.865082	4.224232	0.866123	0.121585	0.575000	0.071296	2.0	-16.428862
50%	1.720406	11.897023	5.075537	1.167339	0.142922	1.100000	0.085237	2.0	-6.104531
75%	2.184020	13.466488	5.910478	1.810072	0.173183	1.520000	0.122929	2.0	17.242159
max	3.166677	19.829695	9.027587	2.405441	0.278703	1.970000	0.170588	2.0	74.955607

TABLE 4 Descriptive statistics for companies in cluster 2 Identified by K-Means

Table 5 presents the characteristics of Cluster 3, consisting of 7 companies. This cluster has an average Price to Book (P/B) ratio of 1.22 and a negative Price to Earnings (P/E) ratio of -41.69. The dividend yield for these companies is 5.02%, the debt-to-equity ratio stands at 1.52, and the cluster achieved the highest average stock price gains at 22.42%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.0	7.000000
mean	1.221543	-41.694003	5.024182	1.516423	-0.041089	-0.145585	-0.037500	3.0	22.417955
std	0.624994	21.940856	2.964245	0.488983	0.037524	0.096043	0.036267	0.0	48.445675
min	0.447067	-68.891048	0.000000	0.930788	-0.095007	-0.270000	-0.093750	3.0	-24.210522
25%	0.835994	-56.693735	3.708528	1.143097	-0.065943	-0.215000	-0.056976	3.0	-7.354259
50%	1.170226	-45.990081	4.736842	1.353717	-0.029795	-0.150000	-0.020204	3.0	-2.662092
75%	1.591638	-25.759035	7.056595	1.980963	-0.016613	-0.079548	-0.013264	3.0	45.920079
max	2.078243	-12.071350	8.902184	2.082333	0.002294	-0.010000	-0.008065	3.0	106.666660

TABLE 5 Descriptive statistics for companies in cluster 3 Identified by K-Means

Table 6 details the characteristics of Cluster 4, which includes 4 companies. This cluster exhibits an average Price to Book (P/B) ratio of 2.32 and a Price to Earnings (P/E) ratio of 39.89. The companies in Cluster 4 have a dividend yield of 2.48%, the highest debt to equity ratio observed at 4.24, and experienced slight average declines in stock prices, recorded at -0.67%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.0	4.000000
mean	2.324675	39.887520	2.483278	4.238120	0.094047	0.250000	0.045133	4.0	-0.665183
std	1.198850	34.285915	1.756502	1.092226	0.078873	0.152534	0.029463	0.0	41.354644
min	1.564534	20.285715	0.000000	2.720849	0.018293	0.070000	0.007919	4.0	-51.231246
25%	1.616858	20.942366	1.882864	3.968240	0.061172	0.160000	0.029887	4.0	-25.571898
50%	1.817082	24.070712	3.015806	4.455433	0.076463	0.255000	0.048308	4.0	4.991116
75%	2.524898	43.015866	3.616220	4.725313	0.109337	0.345000	0.063554	4.0	29.897831
max	4.100002	91.122941	3.901500	5.320763	0.204969	0.420000	0.076000	4.0	38.588280

TABLE 6 Descriptive statistics for companies in cluster 4 Identified by K-Means

The boxplot visualizations in Figure 20 align with the detailed descriptions of each cluster, revealing variations in financial metrics across clusters.

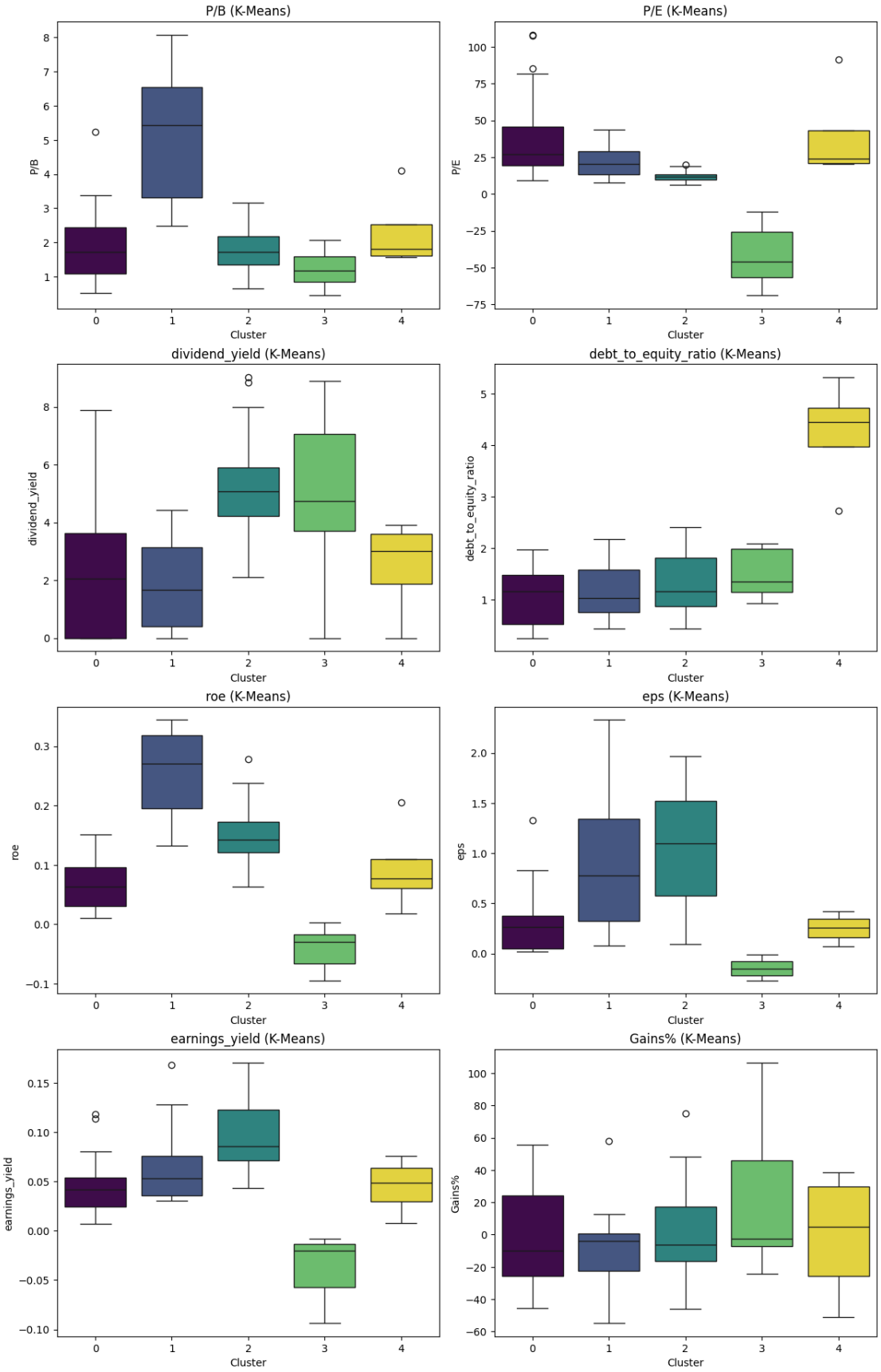

FIGURE 20 Distribution of financial metrics across clusters defined by K-Means Clustering

5.2 Hierarchical Clustering

5.2.1 Defining the number of clusters for hierarchical clustering

Number of clusters was set at five also with hierarchical clustering. This decision was based on analyses involving the elbow method and silhouette scores that are illustrated in the Figure 21. These methods indicated that five clusters provide an optimal balance for our dataset, capturing a significant reduction in within-cluster variation while maintaining model simplicity. It was also a strategic consideration for choosing five clusters, because I wanted to maintain consistency with the number of clusters determined in earlier sections using K-Means and Gaussian Mixture Models (GMM). By aligning the number of clusters across different methods, it simplifies comparative analysis, allowing for clearer insights into how each clustering approach segments the data under similar conditions.

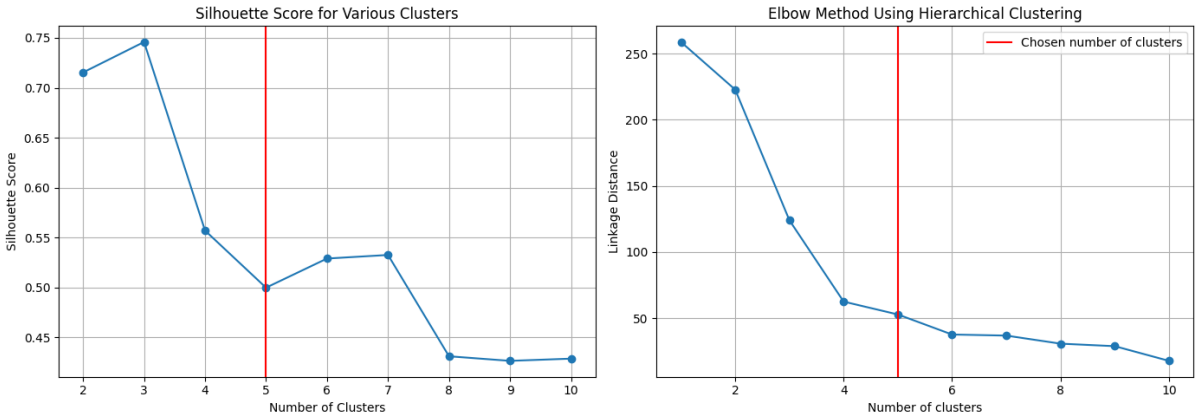

FIGURE 21 Elbow method for optimal number of clusters and Silhouette scores

5.2.2 Hierarchical clustering Results

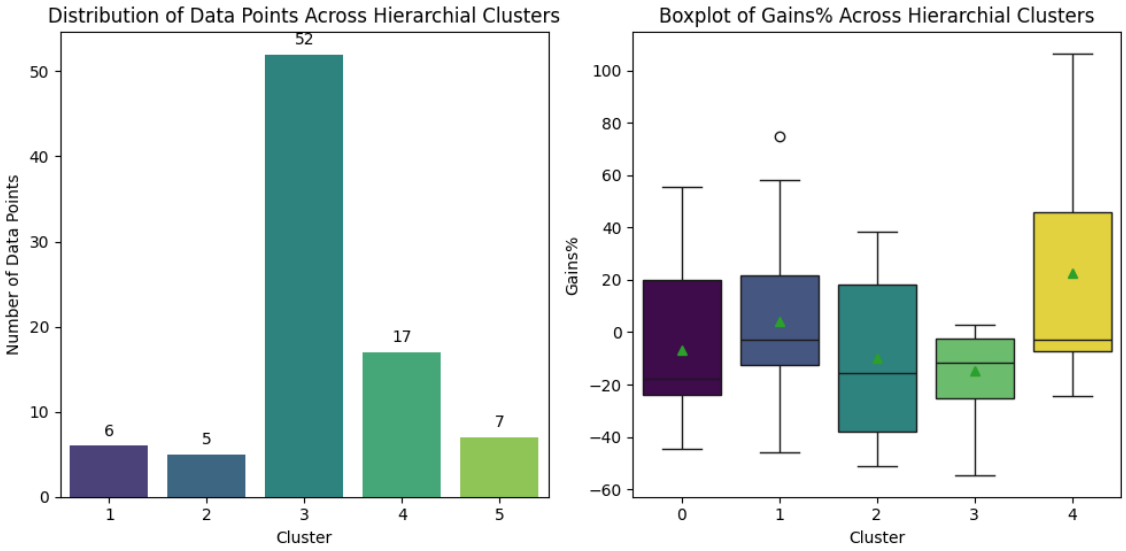

FIGURE 22 Stock count and gains% in each cluster using Hierarchical clustering with k=5, n=87

Cluster 0 consists of companies with an average Price to Book (P/B) ratio of 2.30 and a Price to Earnings (P/E) ratio of 92.45. The average dividend yield is 2.15%, and the debt-to-equity ratio stands at 1.46. The Return on Equity (ROE) averages at 0.024, with an average Earnings Per Share (EPS) of 0.19 and an earnings yield of 1.18%. The average gains percentage is -0.60%. The detailed descriptive statistics for Cluster 0 are presented in table 7.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.0	6.000000
mean	2.297141	92.451816	2.152264	1.455491	0.024437	0.186667	0.011809	0.0	-0.595876
std	1.473788	12.409861	1.365957	1.608751	0.012291	0.167053	0.003583	0.0	27.923010
min	1.405536	81.047058	0.000000	0.262457	0.015638	0.030000	0.007028	0.0	-22.056257
25%	1.503029	82.568036	1.583370	0.381344	0.018258	0.040000	0.008888	0.0	-20.467527
50%	1.675902	88.220443	2.171100	0.970026	0.018275	0.150000	0.012837	0.0	-15.285021
75%	2.183528	103.290089	3.104446	1.606375	0.025808	0.335000	0.014748	0.0	20.672089
max	5.227362	108.225745	3.768365	4.526829	0.047863	0.390000	0.015198	0.0	38.588280

TABLE 7 Descriptive statistics for companies in cluster 0 Identified by Hierarchical clustering

Cluster 1 features companies with an average P/B ratio of 1.23 and a negative average P/E ratio of -53.14. The dividend yield averages at 4.71%, and the debt-to-equity ratio is 1.44. The ROE is -

0.021 on average, and the EPS is -0.10, with an earnings yield of -1.67%. The average gains percentage is 36.76%. The detailed descriptive statistics for Cluster 1 are presented in table 8.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.0	5.000000
mean	1.232672	-53.142519	4.705806	1.435781	-0.020768	-0.101819	-0.016728	1.0	36.759660
std	0.764343	12.172738	3.485673	0.513660	0.017337	0.073100	0.008718	0.0	50.621770
min	0.447067	-68.891048	0.000000	0.930788	-0.043111	-0.190000	-0.028846	1.0	-7.417295
25%	0.526518	-60.000002	3.167056	1.052559	-0.029795	-0.150000	-0.020204	1.0	-7.291224
50%	1.170226	-53.387469	4.250000	1.233634	-0.021866	-0.110000	-0.018061	1.0	20.392099
75%	1.941307	-45.990081	7.209793	1.885576	-0.011360	-0.049096	-0.008467	1.0	71.448058
max	2.078243	-37.443997	8.902184	2.076350	0.002294	-0.010000	-0.008065	1.0	106.666660

Table 8 Descriptive statistics for companies in cluster 1 Identified by Hierarchical clustering

Cluster 2 includes companies with an average P/B ratio of 2.22 and a P/E ratio of 12.73. The dividend yield is 3.87% on average, with a debt-to-equity ratio of 1.37. The average ROE is 0.156, EPS is 0.80, and the earnings yield is 7.91%. The average gains percentage is 0.99%. The detailed descriptive statistics for Cluster 2 are presented in table 9.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	52.000000	52.000000	52.000000	52.000000	52.000000	52.000000	52.000000	52.0	52.000000
mean	2.215209	12.733105	3.874428	1.368533	0.156346	0.804784	0.079074	2.0	0.992431
std	1.335871	6.695890	2.434870	0.890836	0.092976	0.612165	0.050315	0.0	28.649943
min	0.598292	-14.074074	0.000000	0.296487	-0.095007	-0.270000	-0.093750	2.0	-54.853803
25%	1.346018	10.170454	2.346439	0.899612	0.106609	0.350000	0.055133	2.0	-19.714970
50%	1.821859	12.572972	4.224232	1.167146	0.144163	0.632351	0.073497	2.0	-4.146314
75%	2.660862	17.319622	5.468595	1.720438	0.209330	1.210000	0.103772	2.0	21.656144
max	6.833962	21.161249	9.027587	5.320763	0.344501	2.330000	0.170588	2.0	74.955607

Table 9 Descriptive statistics for companies in cluster 2 Identified by Hierarchical clustering

Cluster 3 is comprised of companies with the highest average P/B ratio of 3.14 and a P/E ratio of 28.52. The average dividend yield is 2.56%, and the debt-to-equity ratio is 1.40. The ROE averages at 0.107, EPS at 0.49, and the earnings yield is 4.54%. The average losses in gains percentage are -15.03%. The detailed descriptive statistics for Cluster 3 are presented in table 10.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	17.000000	17.000000	17.000000	17.000000	17.000000	17.000000	17.000000	17.0	17.000000
mean	3.143324	28.516870	2.559176	1.404224	0.107388	0.491529	0.045363	3.0	-15.032040
std	2.502738	3.364750	1.961738	0.618409	0.081427	0.507439	0.014443	0.0	25.039987
min	0.541302	22.778080	0.000000	0.285317	0.019712	0.019000	0.024453	3.0	-51.231246
25%	1.318047	26.473685	1.132675	0.860967	0.045296	0.190000	0.034667	3.0	-37.938741
50%	1.999864	28.181820	1.995539	1.439127	0.075465	0.270000	0.043478	3.0	-15.155931
75%	4.370476	31.225663	4.435484	1.790133	0.153251	0.720000	0.052000	3.0	-1.302924
max	8.084153	34.600001	6.399277	2.720849	0.270190	1.600000	0.076000	3.0	33.996023

Table 10 Descriptive statistics for companies in cluster 3 Identified by Hierarchical clustering

Cluster 4 has companies with an average P/B ratio of 2.67 and a P/E ratio of 48.37. The dividend yield averages at 2.10%, and the debt-to-equity ratio is the lowest at 0.84. The average ROE is 0.058, EPS is 0.13, and the earnings yield is 2.49%. The average gains percentage is -8.57%. The detailed descriptive statistics for Cluster 4 are presented in table 11.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster	Gains%
count	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.0	7.000000
mean	2.669238	48.367548	2.095158	0.844528	0.058124	0.125714	0.024859	4.0	-8.574555
std	1.926808	9.459478	2.819297	0.592039	0.043467	0.111334	0.006466	0.0	29.404176
min	0.522064	38.076870	0.000000	0.256057	0.010006	0.040000	0.014245	4.0	-41.856302
25%	1.285778	41.367594	0.000000	0.422384	0.021367	0.045000	0.021701	4.0	-32.291019
50%	2.765679	47.802854	1.660847	0.476938	0.070310	0.100000	0.025455	4.0	-17.817221
75%	3.275888	52.623780	2.563229	1.289749	0.075986	0.150000	0.028390	4.0	19.232263
max	6.273588	64.710367	7.878799	1.754434	0.131845	0.350000	0.034130	4.0	25.769148

Table 11 Descriptive statistics for companies in cluster 4 Identified by Hierarchical clustering.

The boxplot visualizations in the figure 23 support the detailed descriptions provided for each cluster by revealing the variations in financial metrics across clusters.

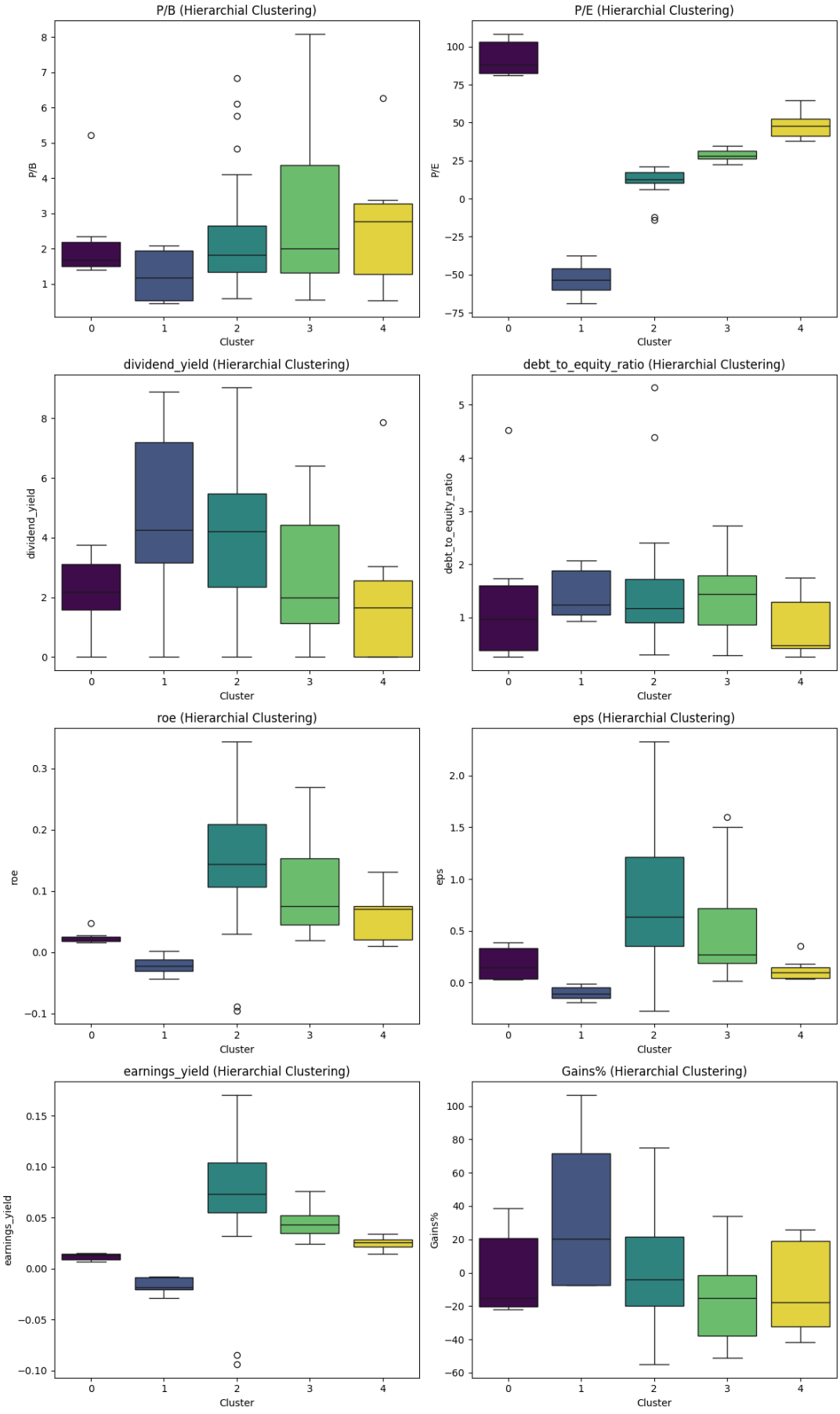

FIGURE 23 Distribution of financial metrics across clusters defined by hierarchical clustering

5.3 Gaussian Mixture Model

Gaussian Mixture Models (GMM) is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. It's well-suited for identifying clusters that have different sizes and correlation structures.

5.3.1 Choosing the n-components for GMM

As we learned previously, the Silhouette Score measures the similarity of an object to its own cluster compared to other clusters. A high Silhouette Score indicates well-clustered data. We computed these scores for different numbers of components, ranging from 2 to 10. The aim was to identify several components that maximized the Silhouette Score, indicating clear and distinct clustering.

Figure 24 illustrates the Silhouette Score and Bayesian Information Criterion (BIC) for different cluster counts. While the highest Silhouette Score of 0.229 was achieved with seven clusters, indicating optimal separation and distinctiveness at this level, the thesis employs a strategic decision to utilize five clusters. This choice, with a Silhouette Score of 0.21, represents a compromise that maintains consistency across comparative analyses of clustering results.

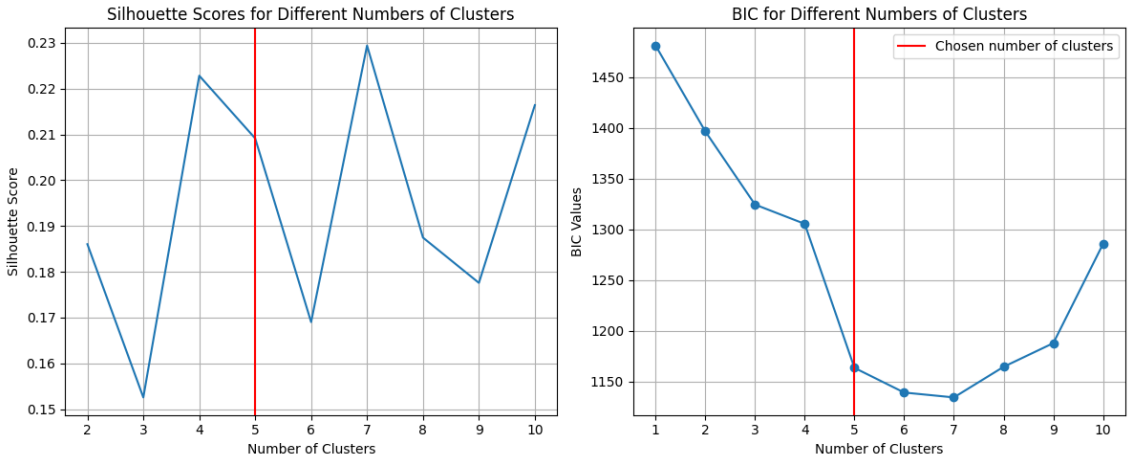

FIGURE 24 Silhouette score and BIC values for different numbers of clusters

Furthermore, the BIC was employed alongside the Silhouette Score to assess model performance, balancing the complexity of the model against its goodness of fit, where a lower BIC value indicates a superior model. The BIC calculations for varying numbers of clusters were integral to identifying the most statistically viable cluster count. Notably, the decision for five clusters was reinforced by

observations of BIC values that began to level off or decrease slowly beyond this point, suggesting diminishing returns on model improvement with additional clusters. This analysis substantiates the selection of five clusters as being statistically and practically justified.

5.3.2 Gaussian Mixture Model (GMM) Results

In the Gaussian Mixture Model (GMM) clustering, a decision was made to use five clusters to provide a direct comparison with other clustering algorithms discussed in this thesis. This deliberate choice ensures that each cluster reflects distinct financial traits and stock market performances, making the results easily comparable across different analyses. A high-level overview of mean values for each cluster is detailed in the subsequent chapters and tables. For a more detailed visual inspection of the clusters, refer to figure 25.

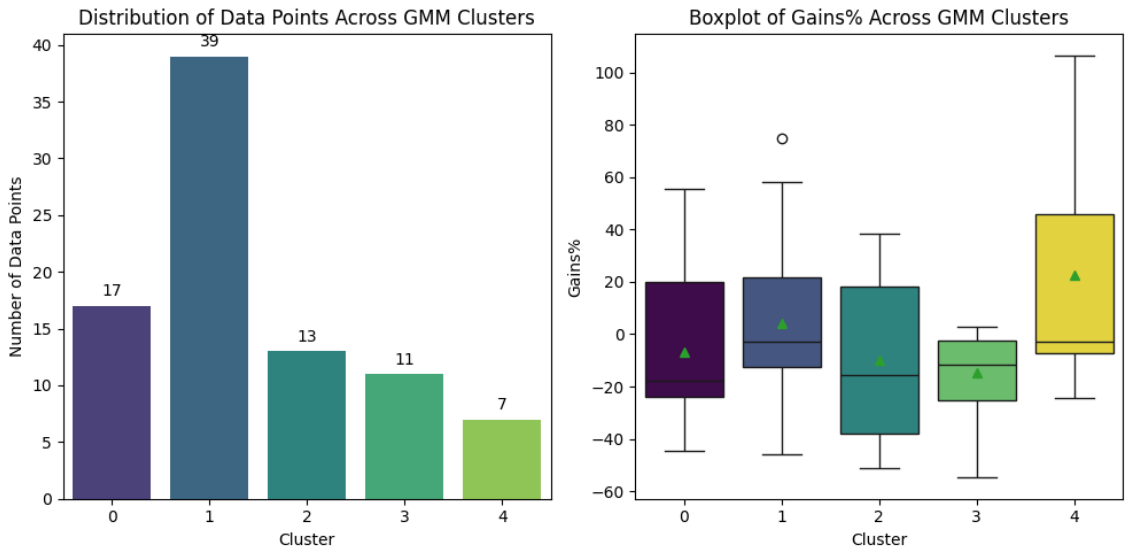

FIGURE 25 Stock count and gains% in each cluster using GMM with 5 clusters, n=87

Table 12 illustrates the characteristics of Cluster 0, which includes companies that exhibit an average Price to Book (P/B) ratio of 2.29 and a Price to Earnings (P/E) ratio of 53.33. The companies in this cluster have an average dividend yield of 1.65% and a debt-to-equity ratio of 1.22. The average Return on Equity (ROE) is 0.19, and the average Earnings Per Share (EPS) is 0.33. The earnings yield for this cluster averages at 2.85%, and there has been an average stock price decrease of -6.82% among these companies.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster_GMM	Gains%
count	17.000000	17.000000	17.000000	17.000000	17.000000	17.000000	17.000000	17.0	17.000000
mean	2.287497	53.332787	1.647024	0.748053	0.051965	0.234059	0.028455	0.0	-6.823432
std	1.583664	29.450241	1.534465	0.516386	0.036865	0.220233	0.015409	0.0	30.100015
min	0.522064	19.901093	0.000000	0.256057	0.010006	0.019000	0.007028	0.0	-44.419744
25%	1.405536	29.722222	0.000000	0.297698	0.018258	0.040000	0.015038	0.0	-23.856572
50%	1.923424	43.485186	1.645135	0.476938	0.047863	0.180000	0.026578	0.0	-17.817221
75%	3.181597	81.047058	2.697065	1.147124	0.080585	0.370000	0.036697	0.0	20.063690
max	6.273588	108.225745	4.672897	1.741150	0.131845	0.720000	0.054745	0.0	55.519095

TABLE 12 Descriptive statistics for companies in cluster 0 Identified by GMM

Cluster 1 is described in the table 13 and includes total of 39 companies with an average P/B ratio of 1.91 and a P/E ratio of 12.74. These companies have a higher dividend yield on average, at 4.22%, and a debt-to-equity ratio of 1.32. The ROE for this cluster averages at 0.19, like Cluster 0, with an EPS of 0.65. The earnings yield for these companies is notably higher at 8.81%, and they have experienced an average stock price gain of 4.24%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster_GMM	Gains%
count	39.000000	39.000000	39.000000	39.000000	39.000000	39.000000	39.000000	39.0	39.000000
mean	1.912137	12.742147	4.223088	1.256789	0.151791	0.924992	0.088067	1.0	4.237389
std	0.724471	3.548851	2.409510	0.551779	0.053449	0.557433	0.036576	0.0	28.704544
min	0.654634	6.348980	0.000000	0.296487	0.063286	0.040000	0.041481	1.0	-46.002957
25%	1.378318	10.412470	2.879834	0.896140	0.117257	0.495000	0.066878	1.0	-12.385295
50%	1.770851	12.235453	4.252082	1.166954	0.143975	0.830000	0.080169	1.0	-2.974181
75%	2.385711	15.483563	5.698952	1.685018	0.167672	1.330000	0.105877	1.0	21.944411
max	3.370205	19.829695	9.027587	2.405441	0.278703	1.970000	0.170588	1.0	74.955607

TABLE 13 Descriptive statistics for companies in cluster 1 Identified by GMM

Table 14 present the characteristics of Cluster 2, where total of 13 companies have an average P/B ratio of 1.80 and a P/E ratio of 32.44. The dividend yield averages at 3.80%, with a relatively lower debt-to-equity ratio of 0.74. The average ROE is 0.07, and the EPS stands at 0.21. The earnings yield is 4.60%, and companies in this cluster have seen an average stock price decrease of -9.87%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster_GMM	Gains%
count	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.0	13.000000
mean	1.798658	32.444575	3.796364	2.370635	0.066802	0.224385	0.046010	2.0	-9.867819
std	1.001610	19.472825	1.989176	1.427197	0.050646	0.119457	0.020788	0.0	29.851711
min	0.811488	19.208908	0.000000	0.913903	0.018293	0.037000	0.007919	2.0	-51.231246
25%	1.068402	22.778080	2.576387	1.439127	0.037168	0.120000	0.034130	2.0	-37.938741
50%	1.536947	25.876035	3.521127	1.754434	0.045296	0.220000	0.047761	2.0	-15.384828
75%	1.999864	31.622550	4.660232	2.720849	0.075465	0.320000	0.059701	2.0	18.400836
max	4.100002	91.122941	7.878799	5.320763	0.204969	0.420000	0.076000	2.0	38.588280

TABLE 14 Descriptive statistics for companies in cluster 2 Identified by GMM

In table 15 Cluster 3 includes total of 11 companies that show a higher average P/B ratio of 5.62 and a P/E ratio of 21.91. This cluster has an average dividend yield of 1.71% and a debt-to-equity ratio of 2.25. Companies in this cluster have an ROE of 0.16 and an EPS of 0.88. The earnings yield averages at 7.18%, with an average stock price decrease of -14.69%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster_GMM	Gains%
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.0	11.000000
mean	5.624011	21.905166	1.714832	1.244927	0.274654	0.885825	0.071844	3.0	-14.694635
std	1.804036	8.893720	1.848236	0.508073	0.065695	0.712753	0.046541	0.0	17.785788
min	2.481016	7.500000	0.000000	0.679148	0.153251	0.080000	0.031078	3.0	-54.853803
25%	4.605899	16.688370	0.000000	0.818611	0.240283	0.327036	0.035809	3.0	-25.053529
50%	5.759272	20.550001	1.278546	1.033105	0.298826	0.780000	0.055742	3.0	-11.545623
75%	7.027872	29.016783	3.110526	1.614596	0.322253	1.370000	0.095364	3.0	-2.319273
max	8.084153	32.951016	4.435484	2.171254	0.344501	2.330000	0.168067	3.0	3.047999

TABLE 15 Descriptive statistics for companies in cluster 3 Identified by GMM

Cluster 4 described in the table 16 represent the smallest cluster containing 4 companies. These companies have an average P/B ratio of 1.22 and a negative average P/E ratio of -41.69. The dividend yield for these companies is higher at 5.02%, and they have an average debt-to-equity ratio of 1.35. Despite the negative P/E ratio, these companies have a positive average ROE of 0.04 and a negative EPS of -0.13. The average earnings yield is -3.75%, yet these companies have experienced a significant average stock price gain of 22.42%.

	P/B	P/E	dividend_yield	debt_to_equity_ratio	roe	eps	earnings_yield	Cluster_GMM	Gains%
count	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.0	7.000000
mean	1.221543	-41.694003	5.024182	1.516423	-0.041089	-0.145585	-0.037500	4.0	22.417955
std	0.624994	21.940856	2.964245	0.488983	0.037524	0.096043	0.036267	0.0	48.445675
min	0.447067	-68.891048	0.000000	0.930788	-0.095007	-0.270000	-0.093750	4.0	-24.210522
25%	0.835994	-56.693735	3.708528	1.143097	-0.065943	-0.215000	-0.056976	4.0	-7.354259
50%	1.170226	-45.990081	4.736842	1.353717	-0.029795	-0.150000	-0.020204	4.0	-2.662092
75%	1.591638	-25.759035	7.056595	1.980963	-0.016613	-0.079548	-0.013264	4.0	45.920079
max	2.078243	-12.071350	8.902184	2.082333	0.002294	-0.010000	-0.008065	4.0	106.666660

TABLE 16 Descriptive statistics for companies in cluster 4 Identified by GMM

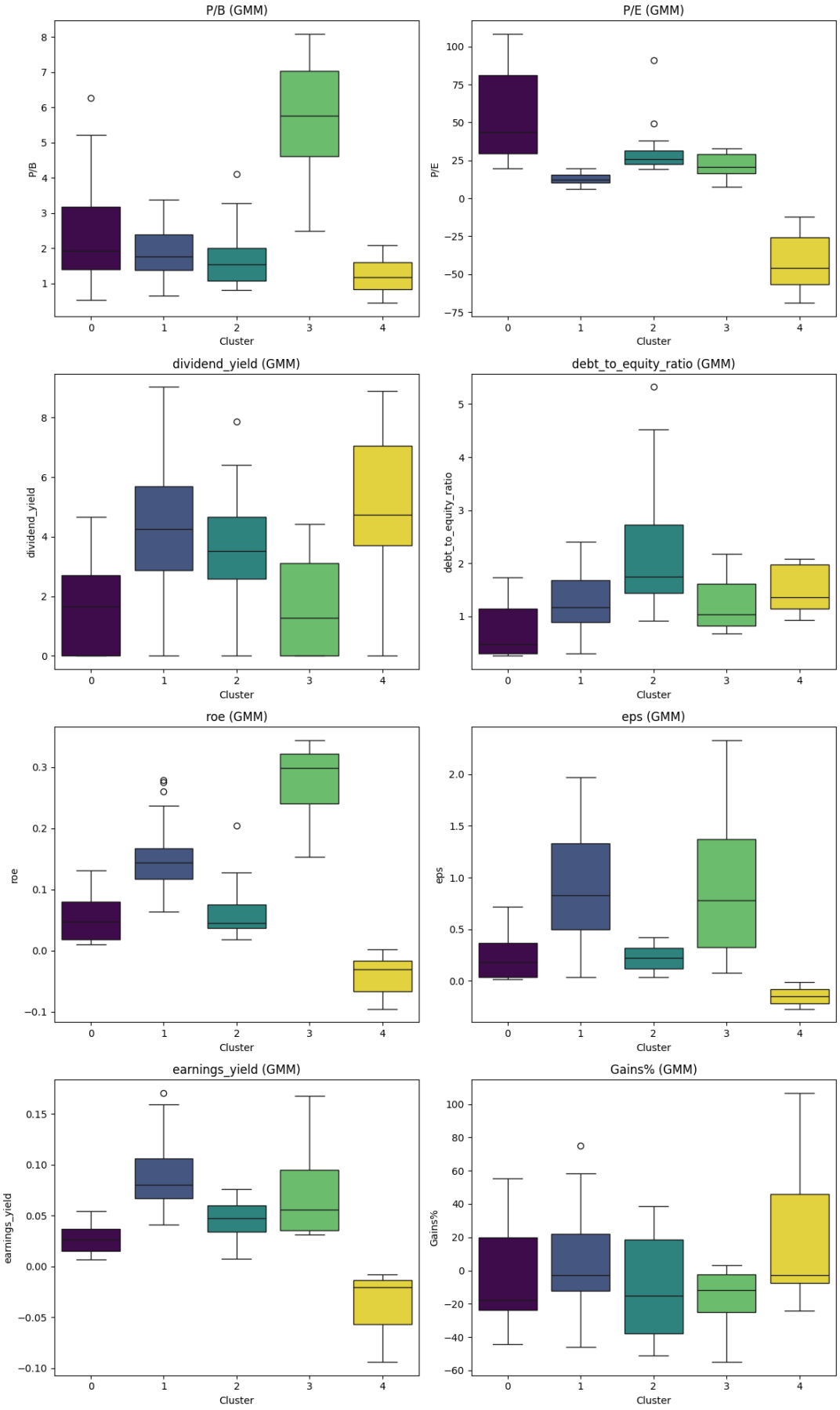

FIGURE 26 Distribution of financial metrics across clusters defined by GMM

5.4 Comparison between clustering results

This section evaluates and compares the performance of three clustering algorithms: K-Means, Hierarchical Clustering (HC), and Gaussian Mixture Model (GMM). The algorithms are evaluated using internal methods such as silhouette score, Davies-Bouldin index, and Dunn index. The results are also compared by utilizing the percentage of average gains per cluster. Principal Component Analysis (PCA) is applied to each clustering method, and the results are visualized to clarify the cluster distribution.

Figure 27 provides a dual perspective by showing both, the count of data points in each cluster and the average gains percentage. In this figure, the bars represent the counts of clusters, while the line graph illustrates the percentage gains achieved by each cluster. The figure shows that the distribution and gain percentages varied significantly among clusters. The clustering methods used in the analysis resulted in different distributions of stocks across the clusters.

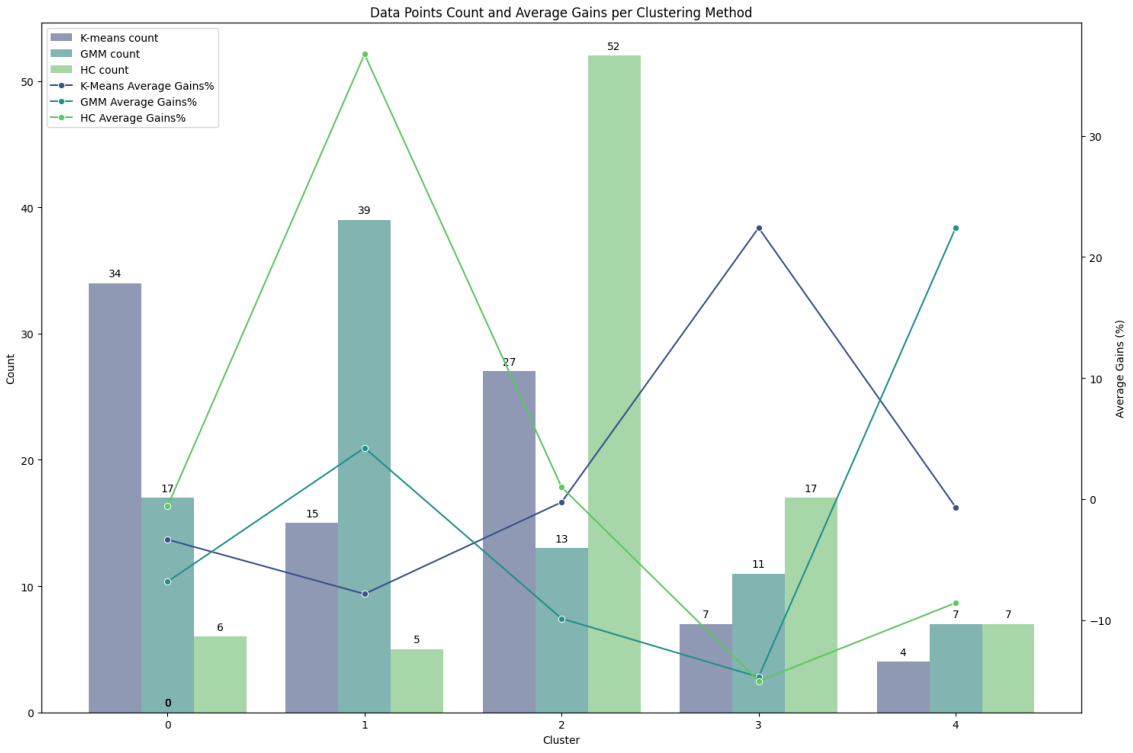

Figure 27 Distribution of stocks between clusters and average gains percentage.

5.4.1 Internal evaluation

Each clustering method was evaluated using three internal metrics: the Silhouette Score, the Davies-Bouldin Index, and the Dunn Index, which collectively measure the compactness, separation, and distinctness of the clusters formed. The comparison of these metrics is displayed in figure 28.

The hierarchical clustering approach showed the best performance across all metrics, as shown in figure 28. It achieved the highest Silhouette Score of 0.4997, indicating the most cohesive and well-separated clusters among the methods. It also gained the lowest Davies-Bouldin Index of 0.5057, pointing to compact yet distinct clusters. Furthermore, it recorded the highest Dunn Index of 0.00547, confirming a clearer separation between clusters compared to the other methods.

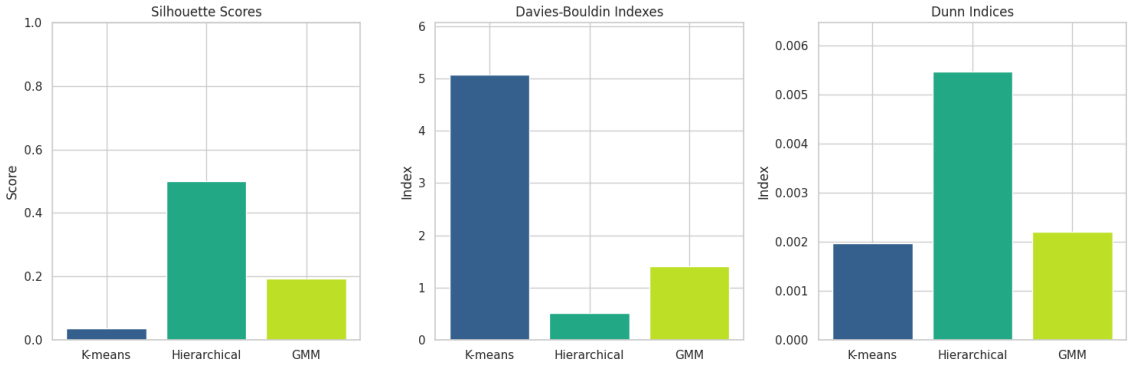

Figure 28 Performance evaluation of clustering methods using Silhouette Score, Davies-Bouldin Index, and Dunn Index.

The GMM method achieved a Silhouette Score of 0.1938 and a Davies-Bouldin Index of 1.4126, indicating a moderate level of cluster separation and compactness. The Dunn Index for GMM was 0.00220, which reflects some separation between clusters, but is notably lower than that of the Hierarchical method, suggesting overlap or less distinct cluster boundaries.

The K-means clustering method, while popular and efficient for many practical applications, did not perform as well in this context. It recorded the lowest Silhouette Score of 0.0348, suggesting that the clusters defined were not as cohesive. Additionally, it had the highest Davies-Bouldin Index of 5.0712, indicating that the clusters were neither compact nor sufficiently separated. The Dunn Index

was also the lowest among the three methods at 0.00197, reinforcing the notion of inadequate separation between the clusters.

5.4.2 Business metrics evaluation

The K-Means clustering method displayed a range of outcomes concerning both the count of stocks per cluster and their respective average gains. The most populated cluster under K-Means had 34 stocks while the least had only 4, indicating a significant variance in stock distribution. This method identified clusters with considerable differences in performance, with the best-performing cluster showing gains of 22.42%, contrasting sharply with the least-performing at -7.84%. The distribution of the K-Means clustering results and gains% is described in figure 19 in chapter 5.1.2.

Hierarchical clustering demonstrated a more balanced distribution of stocks and was notably effective in identifying clusters with high average gains. The cluster with the highest gains under HC exhibited a remarkable 36.76% on average gains, which is significantly higher than the clusters formed by the other methods. The consistency and separation quality shown by Hierarchical clustering was also superior based on the internal evaluation metrics, making it the most reliable method in this dataset for grouping stocks with similar financial characteristics. The distribution of the Hierarchical clustering results and gains% is described in figure 22 in chapter 5.2.2.

GMM's approach yielded a similar diversity in stock counts, with its largest cluster housing 39 stocks and the smallest clusters having 7 stocks each. There was a significant difference in the gains between GMM's clusters. The highest-performing cluster achieved an average of 22.42% yearly gains in price, while the lowest-performing cluster experienced a loss of 14.69%. While GMM successfully identified clusters with strong gains (up to 22.42%), it also grouped stocks that performed poorly (down to -14.69%). The distribution of the GMM clustering results and gains% is described in figure 25 in chapter 5.3.2.

The overall evaluation of these clustering methods in terms of average gains revealed that Hierarchical Clustering outperformed the other methods with an overall average gain of 2.71%, followed by K-Means at 2.06%, and GMM, which exhibited a slight negative overall gain of -0.95%.

5.4.3 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that simplifies high-dimensional datasets into a smaller number of principal components while retaining essential information. PCA provides an efficient representation of the data by creating new variables that capture the maximum variance. (Richardson, 2009).

The first principal component (PC1) captures the direction with the greatest data variance, representing the projected points' overall shape. The more variability PC1 captures, the more information it retains from the original dataset, as no other principal component can hold more variance. The second principal component (PC2) is calculated similarly to PC1 but captures the second-highest variance. PC2 must be orthogonal to PC1, meaning the correlation between PC1 and PC2 equals zero. (Richardson, 2009).

PCA is applied to each clustering result to visualise the clustering outcomes. PCA reduces the high-dimensional features into two principal components, PC1 and PC2, which capture most of the variance. This dimensionality reduction projects the data into a two-dimensional space, making it easier to visualise and understand. For all clustering methods applied, the explained variance for PC1 and PC2 is approximately 32% for PC1 and 22% for PC2.

In figure 29, the scatter plots in side-by-side comparison show how each method clusters the same dataset differently. All the clustering methods reveal five clusters, each with varying degrees of separation. While some clusters are distinctly separated from others, some overlapping exists, indicating that certain financial metrics may not differ significantly between these clusters.

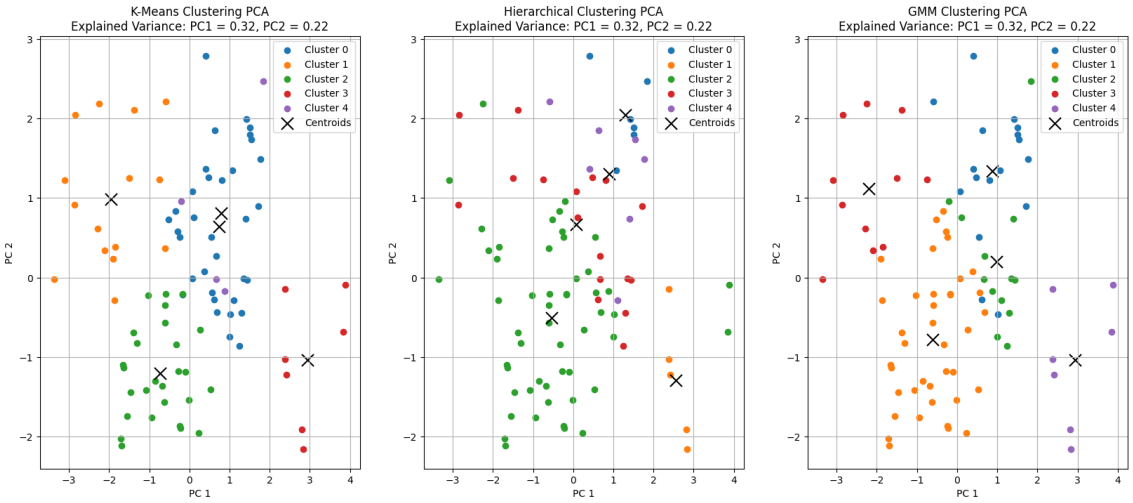

Figure 29 Scatter plot of clustering results.

The clusters in the figure 29 are reasonably well-separated, though some overlap is visible around the origin of the plot. This suggests that some companies share financial characteristics across different clusters. However, Hierarchical Clustering reveals distinct groupings for clusters, providing additional clarity.

6 DISCUSSION AND CONCLUSION

The objective of this thesis was to conduct a comparative analysis of various clustering methods applied to a dataset containing financial ratios of Finnish publicly listed companies. The primary aim was to evaluate the performance of these methods in clustering companies based on value-investment metrics over a 12-month investment period from 2022 to 2023.

K-Means, Hierarchical Clustering, and Gaussian Mixture Model were selected for comparison. These clustering methods are introduced in Chapter 2. The primary rationale for selecting these algorithms lies in their distinct clustering approach and the ability to specify the number of clusters.

The financial ratios used in the clustering task are introduced in Chapter 3 including return on equity, price/earnings, price/book, debt-to-equity, dividend yield, earnings yield, and earnings per share. These financial metrics were selected due to their broad applicability in investment decisions and because they could be calculated using the chosen data collection method. The decision for the methods and the financial metrics was driven mainly by practical considerations, as time constraints made it challenging to conduct a more comprehensive analysis and the thesis had to be delimited somehow sensibly to keep it concise.

The financial metrics are derived from the values available on the last trading day of 2022, which acts as the hypothetical investment day. The observation period concludes on the last trading day of 2023, and the price fluctuation is determined by comparing the values between these two dates. Price fluctuation is calculated by subtracting the 2022 last closing price from the last closing price of 2023, dividing the result by the 2022 closing price, and then multiplying by 100 to obtain the percentage change. The value is denoted in the dataset as ‘Gains%’. The results detailed in this chapter are structured to answer the following key research questions:

- How do clustering algorithms perform when applied to dataset constructed with value investing financial ratios and the investment horizon is short, a maximum of 12 months?
- Are the clustering methods able to outperform the stocks picked using explicit rules based on the favourable financial ratios?

- Are the clustering results generalizable in such a way that it is possible to make profitable risk-tolerant investment decision based on them?

The clustering was implemented by setting the number of clusters to five with all methods. The number of clusters was decided based on the evaluation methods including silhouette score, elbow method and Bayesian information criterion. Evaluation methods are described in chapter 2.4 and the application and results are introduced in chapter 5 under each subchapter. Using the same number of clusters was both a practical choice and a compromise to maintain consistency for comparison purposes.

The clustering results between different methods varied somewhat. Although the variation between results is notable, the differences were not substantial when assessed using the evaluation metrics. This is further evidenced in chapter 5.4 where the results are compared against each other.

The differentiation between clusters was not perfect in any of the three methods as we can observe from the figure 29 in the chapter 5.4.3. The figure 29 reveals a pattern of low inter-cluster and high intra-cluster -distances among clusters as well as some overlap in the clusters. This might be due to the similar characteristics in some financial metrics across companies suggesting that the financial profiles of certain companies might be mixed, sharing similar characteristics that do not allow clear separation, at least with the chosen methods.

Among the clustering methods, hierarchical clustering demonstrated some advantage compared to others. The Hierarchical clustering showed the most effective performance across all evaluation metrics, as described in figure 28 in chapter 5.4.1. It achieved the highest Silhouette Score of 0.4997, indicating that it managed to define the most cohesive and separated clusters among the methods. Additionally, it exhibited the lowest Davies-Bouldin Index of 0.5057, suggesting the most compact and well-separated clusters. The Dunn Index at 0.00547, further confirming the best separation between clusters compared to the other methods. However, it is notable that the difference in performance was not drastically significant across the different methods.

The best-performing cluster in hierarchical clustering was the Cluster 1 containing 5 companies with an average annual stock price growth of 36.76%. A closer analysis reveals that 3/5 of these stocks showed positive gains. When the financial ratios in this cluster are compared to the

theoretical framework in chapter 3, we can see that the ratios are on the opposite side of favourable values. Investing in these stocks would require considerable risk tolerance from the investor.

I also wanted to find out that are the clustering methods able to outperform the stocks picked using explicit rules based on the favourable financial ratios. To do this a set of filter rules was established to dataset based on the favourable financial ratios suggested in chapter 3. This filtering was able to identify 14 different stocks, but only 50% of the stocks showed positive price fluctuation. However, this approach could be useful for investor as it narrows down the larger population for deeper analysis with the favourable financial metrics already applied. This approach also carries less risk compared to the best clustering result.

It's important to recognize certain limitations to ensure a fair and comprehensive interpretation of the results. First of all, the study compared only three clustering methods. This selection was not that comprehensive so comparing other methods, such as density-based methods or advanced machine learning techniques, may reveal different kinds of insights into the dataset. The study was also limited by relying on a limited amount of internal evaluation metrics, which may not fully capture clustering quality from a business perspective. It is also worth notice that the interpretation of the results requires a holistic approach and significant substantive expertise as the impact of global economic trends, regional events, market sentiment and varying market conditions affect to stock market and therefore also the clustering results.

Another critical limitation of this study was its short-term focus which came partly from the practical choice to keep the thesis compact enough and the technical constraints of the selected data collection method. The chosen method permitted data collection only up to four years prior. To better understand the performance of clustering methods over various economic cycles, future research could evaluate how clusters formed by different algorithms fare in the long term. This analysis could reveal patterns that are more suitable for longer-term investment strategies.

The study did not account for detailed industry-specific factors that could significantly influence financial performance. A deeper understanding of the industry context is crucial for interpreting clustering results and improving their applicability.

Market conditions during the data collection period were not fully analyzed, leaving room for external economic factors that may have influenced the results. Changes in global economic trends,

political events, or major shifts in market sentiment could potentially impact the applicability of these clustering outcomes in different time periods.

Future studies should consider expanding the range of clustering techniques beyond those examined here. Methods like DBSCAN (density-based clustering), spectral clustering, or neural network-based models could provide different perspectives on grouping companies by financial metrics. Comparing these with the methods already studied could yield new insights into the strengths and weaknesses of each approach.

Investigating how varying market conditions affect clustering results will be crucial. Future research could assess the impact of global economic trends, regional events, and market sentiment on clustering methods. Understanding these influences would help refine the application of clustering algorithms in volatile investment environments.

REFERENCES

Albalate. A, Minker. W. Semi-Supervised and Unsupervised Machine Learning: Novel Strategies. 2011. Search date 01.4.2024. https://learning.oreilly.com/library/view/semi-supervised-and-unsupervised/9781118586136/

Ahmed, M. Seraj, R. & Islam, S. M. S. 2020. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics 2020, Vol. 9, Page 1295, 9(8), 1295. Search date 21.2.2024. https://doi.org/10.3390/ELECTRONICS9081295

Chollet, F. 2021. Deep Learning with Python, Second Edition. Search date 23.4.2024. https://learning.oreilly.com/library/view/deep-learning-with/9781617296864/

Dayan P, Unsupervised Learning. The MIT Encyclopedia of the Cognitive Sciences. Search date 23.1.2024. https://web.math.princeton.edu/~sswang/developmental-diaschisis-references/dun99b.pdf

Fahmi, Suprapto & Wirawan. 2016. Segmentation and distribution of watershed using K-Modes clustering algorithm and Davies-Bouldin index based on geographic information system (GIS). 2016 International Seminar on Application for Technology of Information and Communication (ISemantic), Semarang, Indonesia, 2016, pp. 235-240. Search date 23.4.2024. https://doi.org/10.1109/ISEMANTIC.2016.7873844.

Fernando, J. 2023. Earnings Per Share (EPS): What It Means and How to Calculate It. Search date 19.2.2024. https://www.investopedia.com/terms/e/eps.asp

Fernando, J. 2022. Price-to-Book (PB) Ratio: Meaning, Formula, and Example. Search date 19.2.2024. https://www.investopedia.com/terms/p/price-to-bookratio.asp

Halkidi, M. Batistakis, Y. Vazirgiannis, M. 2001. On Clustering Validation Techniques. Journal of Intelligent Information Systems 17, 107–145. Search date 25.4.2024. https://doi.org/10.1023/A:1012801612483

Johnston, B., Jones, A., Kruger, C. (2019). Applied Unsupervised Learning with Python. Chapter 1 Unsupervised Learning versus Supervised Learning. Search date 19.2.2024. https://learning.oreilly.com/library/view/applied-unsupervised-learning/9781789952292/

Jonte, D. 2022. A brief introduction to Cluster Validation. Search date 21.4.2024. https://medium.com/@jodancker/a-brief-introduction-to-cluster-validation-ca4215295b06

Lavorini, V. 2022. Gaussian Mixture Model Clusterization: How to Select the Number of Components (Clusters). Search date 2.4.2024 https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4

Liu, X. 2016. Methods for handling missing data, Methods, and Applications of Longitudinal Data Analysis, pp. 441–473.

Marr, B. 2012. Key Performance Indicators (KPI). Pearson Business. Search date 14.2.2024 https://learning.oreilly.com/library/view/key-performance-indicators/9780273750116/ Access required

Michael, A. Pratt, K. 2024. What Is Dividend Yield – And Why Is It Important? Search date 19.2.2024. https://www.forbes.com/uk/advisor/investing/what-is-dividend-yield/

Mitchell, C. 2022. Earnings Yield: Definition, Example, and How To Calculate It. Search date 19.2.2024. https://www.investopedia.com/terms/e/earningsyield.asp

Ramos, J. 2024. Dunn's index. MATLAB Central File Exchange. Search date 7.5.2024. https://www.mathworks.com/matlabcentral/fileexchange/27859-dunn-s-index

Richardson, M. 2009. Principal Component Analysis. Search date 8.5.2024. https://people.duke.edu/~hpgavin/SystemID/References/Richardson-PCA-2009.pdf

Scikit-learn developers. 2024. Gaussian Mixture Models. Search date 19.2.2024. https://scikit-learn.org/stable/modules/mixture.html

Scikit-learn developers. 2024. K-Means. Search date 19.2.2024. https://scikit-learn.org/stable/modules/clustering.html#k-means

Schwarz G. 1978 Estimating the Dimension of a Model. Search date 16.5.2024. https://doi.org/101214/aos/1176344136

Zhang, Z. Q. Feng, J. Huang, Y. Guo, J. Xu, and J. Wang. 2021 A local search algorithm for k-means with outliers q. Search date 1.4.2024. https://doi.org/10.1016/j.neucom.2021.04.028