Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Financial Economics

Deciphering the True Nature of Investment Factors: Employing Advanced Clustering Techniques to Isolate Firm-Level Predictors in Asset Pricing

Tein Baaijens (474073)

Supervisor:	dr. Amar A Soebhag
Second assessor:	dr. Jan Lemmen
Date final version:	1st February 2024

CONTENTS

Abstract

This thesis investigates the impact of applying advanced clustering techniques, such as K-Means, Hierarchical Clustering, and DBSCAN, on the behavior of 22 investment factors in the US stock market. Specifically, it examines how neutralizing these factors from their clusters affects their ability to predict stock returns. The study primarily focuses on contrasting cluster-neutral investment factors with traditional ones. Through the application of panel regression and decile portfolio sorting methods, the research uncovers significant changes in the predictive power of well-known factors like the probability of bankruptcy and market capitalization after they are neutralized from their clusters. Additionally, it reveals interesting shifts in the behavior of other factors, such as the growth in net operating assets and the volume of shares traded, under cluster-neutral conditions. The effectiveness and reliability of each clustering technique are assessed, showing how they categorize firms based on various financial indicators. One of the key findings is the diverse effects of cluster-neutralization on different factors. Some factors demonstrate enhanced predictive power, while others show consistent patterns, highlighting the nuanced and context-dependent nature of financial factor behavior. The study acknowledges certain limitations, including the issue of multicollinearity and the potential for exploring different clustering arrangements. Future research paths are suggested, including the examination of various clustering approaches, addressing multicollinearity challenges and implementing dynamic clustering. Overall, this thesis contributes to the field by demonstrating how advanced clustering techniques can alter the interpretation and significance of financial factors in stock market analysis. These insights are valuable for developing more informed investment strategies and advancing financial market research.

1 INTRODUCTION

The main aim of this thesis is to analyze how firm-level stock predictors perform independently when separated from similar firms' influences. I use advanced clustering techniques to group firms by common characteristics, enabling a clearer analysis of each predictor's unique impact by adjusting for the average traits of their group.

1.1 *Problem Statement*

In the world of asset pricing, a central quest is to understand what drives stock returns. This thesis introduces a nuanced approach to this challenge, analogous to the nature versus nurture debate in broader scientific inquiry. Just as the nature versus nurture debate scrutinizes the relative impact of genetics and environment on individual traits, this research focuses on distinguishing the inherent qualities of factors ('nature') from the external influences of their financial environment ('nurture'). By applying advanced clustering techniques, I aim to isolate and examine the intrinsic, cluster-neutral characteristics of assets, effectively separating them from the myriad of external factors within closely-knit financial groups.

This methodology builds on the following thought experiment: Imagine a scenario within the financial markets where a group of stocks, identical in every aspect except for one distinct risk exposure or factor, is examined. According to the fundamental principles of asset pricing, stocks with similar risk exposures are expected to yield similar expected returns (Fama and MacBeth, 1973 and Fama and French, 1993). Thus, isolating stocks that differ only in one

risk factor, while keeping other factors constant, allows for the attribution of any observed differences in returns primarily to this singular variable risk exposure. Therefore, observed variances in returns within this controlled group can be directly linked to the unique differing risk factor. This hypothesis acts as a metaphor for the study's methodological approach, utilizing clustering to explore the nuanced behaviors of investment factors once they are separated from the confounding influences of their financial environment

Investment factors, much like individuals, do not exist in isolation. They are continually influenced by a complex network of other factors, market conditions, and economic indicators (Seyfi, 2022). Traditional models in finance, such as the Capital Asset Pricing Model (CAPM) and Fama-French models (Fama and MacBeth, 1973,Fama and French, 1993 and Fama and French, 2015), have provided foundational insights into these relationships. However, they often treat factors as if they operate in a vacuum, overlooking the nuanced interdependencies and cluster effects that can skew asset pricing. My approach challenges this perspective, seeking to untangle these interdependencies and provide a clearer understanding of the true nature of asset returns.

By employing three sophisticated clustering techniques, K-Means clustering (MacQueen et al., 1967), Hierarchical Clustering (Johnson, 1967) and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester et al., 1996), this research aims to neutralize the 'nurture' aspect - the external influences and interconnectedness among firms - thereby allowing a clearer view of the 'nature' of each investment factor.

1.2 *Research Question*

Thus, to address the main research objective of this thesis, I ask the following research question:

> *How does the behavior of cluster-neutral investment factors, derived using advanced clustering techniques, diverge from traditional investment factors in financial markets, in relationship to stock returns?*

And expand the main research objective with the following sub-research questions:

> **Sub-RQ1:** *In what ways do cluster-neutral investment factors, as identified by advanced clustering techniques, exhibit distinct behaviors from traditional factors in explaining stock returns when analyzed through panel regression models?*

This method will explore how cluster-neutral investment factors, impact stock returns over time and across different firms. Panel regression is ideal for this investigation as it can capture the dynamic nature of these factors, highlighting their consistency and variability in influencing returns in a multi-dimensional dataset (Haugen and Baker, 1996a).

> **Sub-RQ2:** *How do the performance and characteristics of cluster-neutral investment factors, isolated using advanced clustering methods, differ from those of traditional factors when evaluated using a decile portfolio sorting approach?*

The decile portfolio sorting approach complements panel regression by offering a detailed examination of how cluster-neutral investment factors impact stock performance across varying

levels of factor intensity. Unlike the broader aggregate analysis provided by panel regression, this method reveals nuanced insights into the performance gradients and risk-reward profiles of stocks, enabling a deeper understanding of the predictive power and investment utility of these factors (Fama and French, 1993).

> **Sub-RQ3:** *Which advanced clustering technique (e.g., K-Means, Hierarchical Clustering, DBSCAN) most effectively distinguishes the behavioral patterns of cluster-neutral investment factors from those of traditional factors in financial market data?*

Each clustering method employs a distinct methodology and makes different assumptions about the data. Exploring these clustering methods side by side will allow for a more robust examination of clustering in financial data.

1.3 *Societal and Academic Relevance*

Academically, as explained earlier, this thesis contributes to the field of financial economics by enhancing the understanding of asset pricing. Traditional models like CAPM and the Fama-French models (Fama and MacBeth, 1973 and Fama and French, 1993) have laid the groundwork in this area, but they often treat investment factors in isolation. This research challenges and extends these models by incorporating the influence of firm clusters, thereby providing a more nuanced understanding of asset returns. The use of sophisticated machine learning clustering techniques to analyze cluster-neutral investment factors represents an innovative approach in financial research, potentially leading to new insights and methodologies in the field.

From a societal perspective, this research has implications for investors, financial analysts, and policymakers. By offering a deeper understanding of how individual firm-level predictors influence stock returns, this study can aid in more informed investment decision-making and risk management strategies. For individual and institutional investors, this translates to potentially better investment performance and risk mitigation. For policymakers and regulatory bodies, the insights gained can inform the development of more robust financial regulations and policies, contributing to the stability and efficiency of financial markets.

2 RELATED WORK

Investment factors are essential for understanding and predicting asset returns in financial markets. This field of study was pioneered by the Capital Asset Pricing Model (CAPM), introduced by (Sharpe, 1964), which brought the beta factor into focus as a measure of market risk. Building on this, the Fama and French, 1993 three-factor model further enriched this understanding. Introduced in 1993, it incorporated size (small vs. large capitalization) and value (high vs. low book-to-market ratio) as additional determinants of stock returns. The evolution continued with Carhart's four-factor model in 1997 (Carhart, 1997), which added momentum to acknowledge the influence of persistent trends in stock performance. More recently, the research by Green et al., 2017 highlighted a plethora of characteristics affecting stock returns.

As the complexity of models increases with the addition of more predictors, sophisticated methods are used to understand the interplay between stock returns and these predictors. This

includes cross-sectional regression analyses (Haugen and Baker, 1996a; Lewellen, 2014) and advanced deep learning models (Gu et al., 2020). In addition to examining the relationship between stock returns and firm-level characteristics, the literature is increasingly focusing on intricate relationships among firms themselves.

As noted in the paper "Neighbouring Assets" by Seyfi, 2022, there are various types of relationships among firms, including connections through common principal customers (Cohen and Frazzini, 2008), suppliers (Menzly and Ozbas, 2010), linkages within the same industry (Moskowitz and Grinblatt, 1999), similarities in technology (Lee et al., 2016), and overlaps in analyst coverage (Ali and Hirshleifer, 2020). Besides these connections, Seyfi, 2022 also mentions that firms can be linked by sharing similar characteristics. For example, Fama and French, 1995 have documented that the earnings of firms with similar size and book-to-market ratios are influenced by common factors. Seyfi, 2022 further argues that firms can also be connected through a plethora of similar firm-level characteristics.

Seyfi, 2022 then demonstrates that by employing a sophisticated method, to identify these connections, based on a wide array of similar firm-level characteristics, it's possible to better predict stock returns. This method involves defining 'neighbouring assets' as those sharing the most similar characteristics and using their past performance to forecast the future returns of a given asset. Seyfi's analysis, incorporating a comprehensive set of 94 characteristics, reveals that assets with similar traits tend to exhibit correlated performance patterns.

While the correlations among firms' performances have been shown to be of significant importance in the asset pricing literature, even leading to the development of more sophisticated models such as deep learning approaches that capture this network of firms (notably, graph neural networks as discussed by Uddin et al., 2023), it's important to note that this approach also emphasizes a crucial aspect of asset pricing: the need for detailed scrutiny of each firm at an individual level.

The interconnectedness and clustering of firms present a significant challenge in evaluating firm factors individually, potentially leading to biases in asset pricing models. When firms are closely connected or clustered based on certain characteristics, their performances tend to exhibit correlation, making it difficult to isolate the impact of individual firm factors. This correlation, as explored in Seyfi, 2022 methodology, can result in a phenomenon akin to a 'herding effect', where the performance of one firm is influenced by its peers within the same cluster. As a result, traditional factor models may overestimate the influence of shared characteristics, overlooking the unique aspects and idiosyncratic risks of individual firms. For instance, two companies in the same industry might have different risk profiles due to their management strategies or product diversification, but a model emphasizing industry-based clustering might fail to capture these nuances. Recognizing and adjusting for these interfirm correlations is thus essential for a more accurate and unbiased analysis of individual firm factors and their impact on stock returns.

While approaches to neutralize the effect of clusters have been explored, such as factor tilting[1] (Fama and French, 2012), risk-parity methods[2] (Maillard et al., 2010), quantile regres-

[1] Factor tilting involves adjusting the weights of factors in a portfolio to reduce the influence of certain clusters, particularly useful for addressing overexposure to specific market segments.

[2] Risk-parity is a portfolio construction strategy that allocates capital based on the risk contribution of each asset or factor, aiming to achieve balanced risk distribution across different components.

sion[3] (Koenker and Hallock, 2001), covariance matrix adjustment[4] (Ledoit and Wolf, 2003), residualization of factors[5] (Fama and French, 1993, and portfolio constraints[6] (Jagannathan and Ma, 2003), none of these techniques utilize the strong linkage between firms and their array of characteristics, as identified by Seyfi, 2022.

My thesis contributes to the literature by exploring these linkages, employing sophisticated machine learning clustering techniques to make the factors cluster neutral. Furthermore, cluster-neutral analysis distinguishes itself from traditional models by paying special attention to the specific characteristics of each stock cluster. Unlike conventional approaches that may apply broad generalizations, this method acknowledges the unique qualities within homogeneous stock groups. This focus allows for a more detailed and relevant analysis of factor effects, ensuring that the peculiarities of each cluster are considered.

Which clustering techniques are most suitable for this purpose? (Seyfi, 2022) utilized K-Means (MacQueen et al., 1967) clustering, demonstrating it to be a robust and easily interpretable method. However, my methodology differs from that of Seyfi, 2022. While he utilizes the clustered nature of stock data to predict asset returns, my approach focuses on deconstructing this clustered nature to examine individual factors more closely. I aim to explore other clustering techniques which have different assumptions about the dataset. K-Means assumes spherical clusters and linear relationships, an assumption that has been challenged. For instance, Gu et al., 2020 have shown that stock data often exhibits non-linear relationships.

There are numerous clustering methods that accommodate non-linear relationships in their clustering approach, including Spectral Clustering (Von Luxburg, 2007)[7], DBSCAN (Density-Based Spatial Clustering of Applications with Noise)[8] (Ester et al., 1996), Gaussian Mixture Models (GMM)[9] (Reynolds et al., 2009), and Affinity Propagation[10] (Frey and Dueck, 2007). However, for my analysis, I focus on Hierarchical Clustering[11] (Johnson, 1967) and DBSCAN. These methods were chosen for their distinctive capabilities: Hierarchical Clustering provides a detailed exploration of market structures, essential for understanding multi-layered stock relationships, while DBSCAN offers robustness against the irregularities and non-linearities prevalent in financial data, discerning clusters based on density.

By combining these methods with the simpler K-Means clustering method, a more nuanced view of the intricate grouping patterns within stocks is achieved, along with a deeper understanding of cluster-neutral factors. In addition, since asset pricing fundamentally revolves

[3] Quantile regression is a statistical technique that estimates the relationship between variables for different quantiles of the dependent variable, providing a more comprehensive view of this relationship across the entire distribution.

[4] Covariance matrix adjustment involves modifying the covariance estimates between stocks, often to reflect a higher degree of correlation within clusters, thereby impacting portfolio diversification strategies.

[5] Residualization of factors refers to the process of isolating the unique impact of individual factors by removing the common component of returns attributable to clustering effects.

[6] Portfolio constraints are specific limitations set during portfolio construction, such as capping the maximum weight allocated to stocks within the same cluster, to reduce cluster-driven risks and biases.

[7] Spectral Clustering uses eigenvalues of a similarity matrix to reduce dimensionality and is effective in capturing complex and non-linear structures within data.

[8] DBSCAN identifies clusters with arbitrary shapes and efficiently handles outliers, making it suitable for non-linear and irregular data patterns.

[9] GMM assumes data is generated from a mixture of several Gaussian distributions, offering flexibility in capturing diverse and non-linear relationships.

[10] Affinity Propagation identifies clusters by exchanging messages between data pairs, adaptively finding cluster centers and handling complex relationships.

[11] While Hierarchical clustering may not directly allow for non-linearity, its ability to reveal nested structures and Hierarchical organization in the stock market is invaluable for exploratory data analysis and understanding multi-layered stock relationships.

around returns, it's critical to examine how these newly identified cluster-neutral factors differ from their traditional counterparts in explaining returns. In this thesis, this is achieved by employing well-established techniques such as cross-sectional regression (Haugen and Baker, 1996a; Lewellen, 2014) an portfolio sorting (Fama and French, 1993).

In conclusion, this literature review has traced the evolution of investment factor models, highlighting the shift from traditional models to a nuanced understanding of inter-firm connections. My thesis builds upon this foundation, employing advanced clustering techniques to dissect and neutralize these connections, thus offering a fresh perspective in asset pricing. By contrasting traditional factor models with a more detailed, cluster-neutral approach, this research aims to provide a deeper understanding of individual firm factors, refining both the theoretical and practical aspects of asset pricing.

3 METHODOLOGY

As shown in the workflow (Figure 1) below, this section initially details the data utilized. Then it begins with the overarching framework, clarifying the concept of cluster-neutral factors. Subsequently, it discusses the four methods to achieve neutrality, where the first three are cluster-based—K-Means, Hierarchical, DBSCAN—and the latter is regression-based. Following this, the metrics used to assess the adequacy of cluster fitting are presented. Concurrently, two different methods — panel regression and decile portfolio sorting — are employed to statistically evaluate the differences between the cluster-neutral and original factors, along with corresponding statistical measures. Additionally, various robustness checks are performed to bolster confidence in the findings.

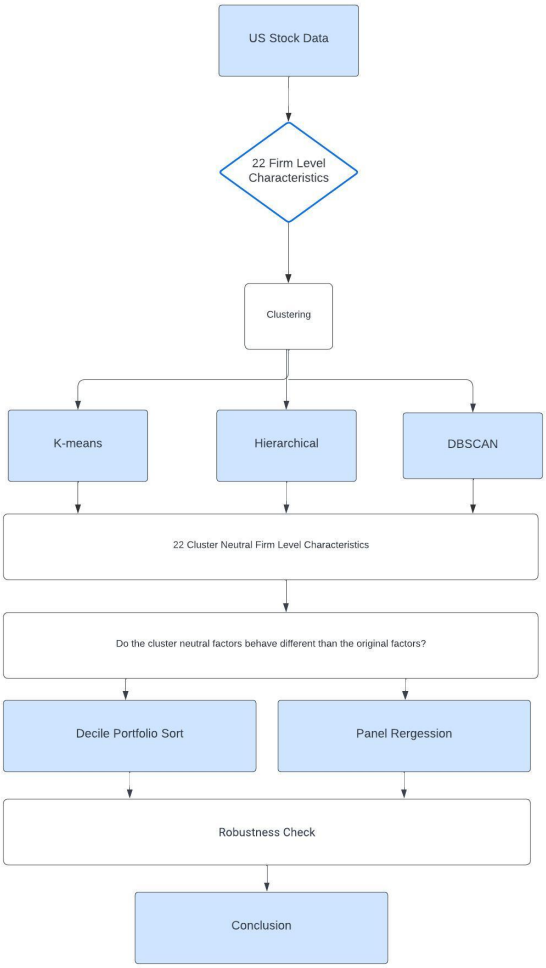

Figure 1: Workflow:

Here the general outline of the methodology is shown in the form of a workflow, underlining the step-by-step process from data collection to analysis of the results

Furthermore, all methods are elucidated in a manner that does not require a background in machine learning or external resources for comprehension. Nonetheless, a more detailed explanation of both the mathematical foundations of the model and the different potential designs of the models is provided in Appendix B.

3.1 *Data*

In this thesis, data on US equities, recorded on a monthly basis from January 1971 to December 2021, are analyzed. This data, obtained from the Center for Research in Security Prices CRSP, 2023,ncludes approximately 27,000 unique firms and over 3 million observations listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and NASDAQ, providing a comprehensive view of market behaviors and trends over five decades. Complementing the CRSP data, 22 firm-specific characteristics from Open Asset Pricing (Chen and Zimmermann, 2022) are gathered. These characteristics are selected based on their prominence in financial literature, as highlighted in several key publications. This includes the Fama

and French factors (Fama and French, 2015 and Fama and French, 1993), momentum factors as described by Carhart Carhart, 1997, Q-factors, and all sub-components of the mispricing model from Stambaugh et al. Stambaugh et al., 2012. Additionally, factors demonstrating robust out-of-sample performance on stock returns, as reported Gu et al., 2020, are included. The specifics of these characteristics, along with their detailed explanations, are provided in Appendix A.

To ensure the robustness of the analysis, the dataset undergoes refinement, focusing on data integrity and relevance. Smaller firms, specifically those in the bottom 5 percent in terms of lagged market capitalization, are excluded to mitigate the effects of illiquidity and incomplete data. This step is crucial in ensuring that the findings are not disproportionately influenced by smaller, more volatile market entities. Regarding missing data within the dataset, alignment with the approach of Gu et al., 2020 occurred, and missing values were imputed using the cross-sectional median. This technique aids in maintaining the continuity and integrity of the dataset. To facilitate comparability across various metrics and prevent any single characteristic from skewing the analysis due to scale differences, data normalization occurred through cross-sectional demeaning and unit variance scaling. Outliers, which can potentially distort the analysis, were managed by winsorizing the dataset at the 1st and 99th percentile levels. This method is effective in reducing the impact of extreme data points.

Furthermore, the study incorporates additional data such as the risk-free rate, market returns, and the Fama-French factors like size, value, operating profitability, and investment factors, sourced from the Kenneth R. French Library (French, 2023). These additional metrics provide a more nuanced understanding of market dynamics and firm performance.

Through this data collection and preparation process, the goal is to deliver a thorough and detailed examination of the US equities market across a significant historical span.

3.2 *Framework for Creating Cluster-Neutral Factors*

Achieving cluster neutrality involves a two-step process: subtracting the cluster mean from individual factor values, known as *centering*, and excluding a specific factor f during the clustering phase. Each step plays a critical role in ensuring an unbiased analysis of factor contributions.

Centering adjusts stock factor values within a cluster. It involves removing cluster-wide trends by subtracting the average value of the cluster from each stock's factor value. This step helps in isolating each stock's performance compared to others in the same group. It standardizes factor values so that zero represents the cluster average, positive values indicate better-than-average performance, and negative values show below-average performance. This normalization makes it easier to compare stocks across different clusters and focuses on how each stock stands out from its cluster average in terms of specific factors.

Excluding Factor f during clustering adds another layer of depth. By not considering factor f when forming clusters, the resulting groups are not influenced by this factor. This exclusion leads to clusters based on other shared characteristics, ensuring a fairer grouping. It also allows for a clearer analysis of factor f across different stocks, as any variation in f reflects genuine differences in stock characteristics, not just a byproduct of how the clusters

were formed. Moreover, it provides insights into how factor f interacts with other factors, which might be missed if f were included in the clustering.

Following this conceptual foundation, the mathematical formulation is as follows:

Consider a dataset $\mathbf{X}$ consisting of various factors for a set of stocks indexed by i over time periods indexed by t. Let $\mathbf{X}$ be structured such that $x_{i,t,f}$ represents the value of factor f for stock i at time t. For each factor f and each time period t, perform clustering on the set of stocks based on all factors except f. This results in a set of clusters $\mathcal{C}_{t,-f}$ for each time period and each factor. For a stock i in cluster $c_{j,t} \in \mathcal{C}_{t,-f}$, adjust the factor value $x_{i,t,f}$ to obtain the cluster-neutral value $x'_{i,t,f}$ as follows:

$$x'_{i,t,f} = x_{i,t,f} - \bar{x}_{c_{j,t},f} \tag{1}$$

where $\bar{x}_{c_{j,t},f}$ is the mean value of factor f in cluster $c_{j,t}$ at time t.

Construct a new matrix $\mathbf{X}'$ where each element $x'_{i,t,f}$ represents the cluster-neutral value of factor f for stock i at time t. This matrix forms the basis for subsequent empirical analysis, ensuring that factor values are adjusted for cluster effects while maintaining their time-specific characteristics.

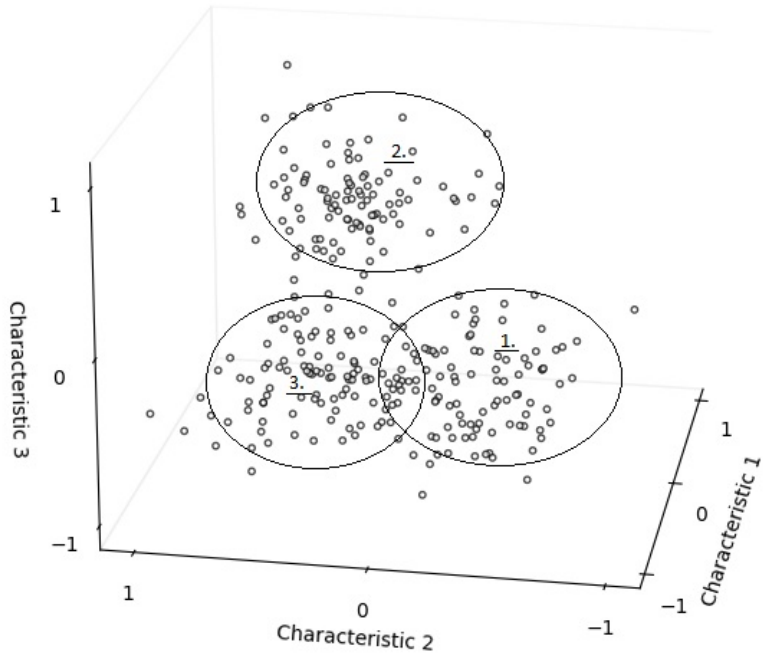

Figure 2: Synthesized Visualization of Stock Clusters.

This figure illustrates three distinct groups of stocks, each represented as a cluster in a simplified, synthesized dataset for clarity. Cluster 1 (at [0, 0, 0]) shows stocks with average values across all financial characteristics, embodying a baseline performance. Cluster 2 (at [1, 1, 1]) represents stocks scoring above average in all characteristics, indicating higher performance or risk. Cluster 3 (at [0, 1, 0]) highlights stocks with distinct values in one characteristic but average in others, showcasing variability in specific financial aspects.

To visually demonstrate the concept of clusters in financial datasets, I have synthesized [12] a dataset and depicted it in Figure 2. This dataset represents three hypothetical groups of stocks, created using a random multivariate distribution centered around different positions. Each group is characterized by its unique mean and variance along three dimensions, which correspond to different financial characteristics such as Beta, Book-to-Market (BM) ratio, and market capitalization. In the following sections, it becomes clear how the three different clustering techniques dissect this synthetic dataset, making different assumptions about the data.

3.3 *Clustering Techniques*

3.3.1 *K-Means Clustering*

K-Means clustering (MacQueen et al., 1967) is a widely recognized and fundamentally important method in data analysis, particularly useful as a benchmark in the exploration of more intricate clustering techniques. In the multifaceted and high-dimensional world of financial data, where each stock can be characterized by a multitude of factors, K-Means provides a relatively simple yet effective way to discern patterns and group similar stocks.

At its core, K-Means clustering aims to partition the dataset into k distinct clusters. This is achieved by assigning each stock to the cluster whose centroid—the mean value of points in the cluster—is closest to it. The process iteratively adjusts the positions of centroids and reassigns stocks to clusters, with the goal of minimizing the within-cluster variance. In simpler terms, it seeks to ensure that stocks within each cluster are as similar to each other as possible, while maximizing the dissimilarity between different clusters.

In Figure 3 the results of the clustering can be seen for k = 3, you can clearly see three distinct clusters as was expected given the design of the synthetic dataset. At each green cross, a centroid is placed, representing the average position of the stocks within each cluster. The stocks are marked as points, color-coded to indicate their respective clusters.

[12] As the data used in this thesis is high-dimensional with many observations, the synthesized dataset in Figure 2 serves as a simplified representation for illustrative purposes. This abstraction allows for a clearer visualization and understanding of the concept of clusters in financial datasets.

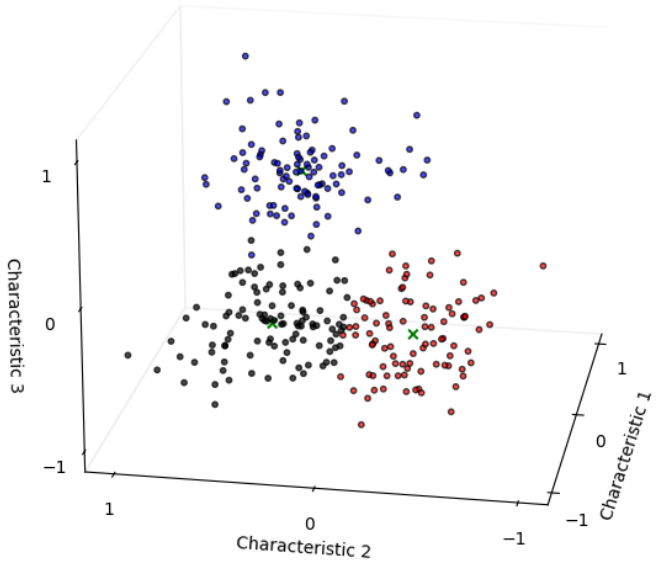

Figure 3: K-Means Clustering.

The figure illustrates the application of K-Means clustering in financial data. Centroids of each cluster are marked with green crosses, representing the mean position of stocks within those clusters. Individual stocks are depicted as color-coded points, with each color corresponding to a different cluster. The three axes in the figure represent various financial metrics or factors that characterize the stocks.

To make the factors cluster neutral in the context of the framework, K-Means is applied separately for each factor f and each time period t. Specifically, it clusters stocks based on all factors except the current factor f, forming clusters $\mathcal{C}_{t,-f}$. This clustering is crucial for computing the cluster-neutral adjustments of factor f, aligning with the objective of my analysis to neutralize cluster-specific biases in factor values.

The model for K-Means in the context of this study is defined by the relationship:

$$\text{Minimize} \sum_{i=1}^{N} \sum_{j=1}^{k} \|x_{i,t,-f} - \mu_{j,t,-f}\|^2 \tag{2}$$

where $x_{i,t,-f}$ represents the data point (stock i at time t excluding factor f) and $\mu_{j,t,-f}$ is the centroid of cluster j. This model seeks to minimize the sum of squared distances between stocks and their respective cluster centroids.

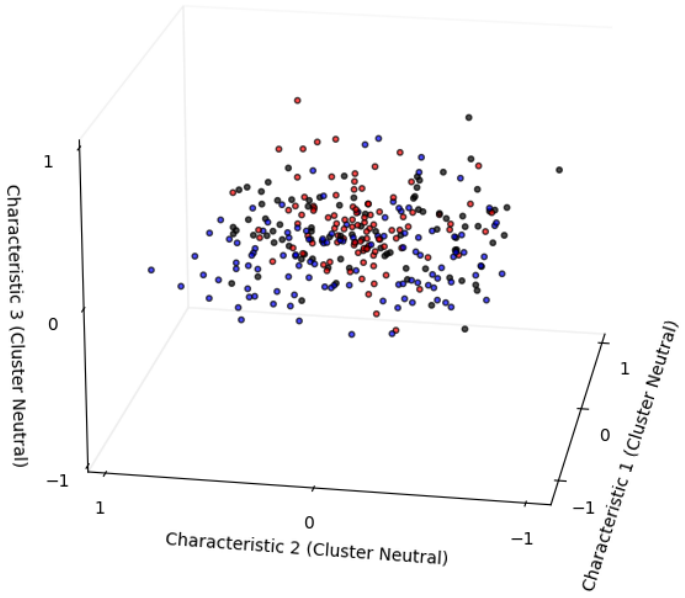

Figure 4: K-Means Cluster Neutralization.

The plot provides an view into how the data points are distributed and clustered in a three-dimensional space after neutralizing the influence of each characteristic within its respective cluster, highlighting the underlying structure and relationships within the data set.

As can be seen in Figure 4, making the factors cluster neutral with K-Means has led to the firms' characteristic values being more centered around one mean. This centralization indicates a reduction in the variance within each cluster, suggesting that the K-Means algorithm effectively minimizes within-cluster variability while maximizing between-cluster differences. However, the effectiveness of this approach is highly dependent on the choice of k, the number of clusters. In this example, it was clear that three is a good value for k. However, this is not the case with actual stock data.

The choice of k is pivotal, as it directly influences the structure and granularity[13] of the resulting clusters. A smaller k may lead to overly broad clusters, potentially grouping together firms with significantly different characteristics. Conversely, a larger k can result in excessively fine clusters, capturing noise and overfitting the data. The challenge lies in finding a balance where clusters are neither too general nor too specific.

To ensure the robustness of the analysis, various k values are examined. Additionally, multiple metrics are utilized to assess the fit of clusters, as detailed in Section 3.5.3. These steps are crucial to validate the findings and enhance the reliability of the cluster analysis in financial datasets.

In conclusion, K-Means clustering stands out for its simplicity and efficiency, particularly in large datasets. Its ability to quickly converge and its ease of implementation make it a popular choice for various applications. However, K-Means has its limitations, such as the requirement

[13] Granularity, in this context, refers to the level of detail or specificity in the clustering. High granularity means more, smaller clusters, each capturing finer distinctions among stocks, while low granularity results in fewer, larger clusters, each encompassing a broader range of stocks.

to specify the number of clusters beforehand, which can be challenging in datasets where the natural cluster count is not known. Additionally, it is sensitive to the initial choice of centroids and can struggle with non-spherical clusters, unlike DBSCAN which do not assume cluster shapes and can discover clusters with complex structures. Despite these drawbacks, K-Means remains a valuable tool, especially when combined with other techniques to overcome its inherent weaknesses.

3.3.2 *Hierarchical Clustering*

While K-Means clustering is a fundamental and widely-used technique in data analysis, it comes with its inherent limitations, most notably the prerequisite to predefine the number of clusters (k). This requirement can be a significant constraint, especially in complex datasets where the natural grouping is not apparent. To address these limitations and explore a more nuanced approach, Hierarchical Clustering (Johnson, 1967) is introduced.

One of the defining features of Hierarchical Clustering is its ability to construct a dendrogram, a tree-like diagram that illustrates the arrangement of the clusters formed at each stage of the analysis. This dendrogram provides a visual and intuitive understanding of the data's Hierarchical structure, showcasing how individual stocks group together at various levels of similarity.

In Figure 5, a dendrogram illustrates the dissimilarities between firms based on firm-level characteristics, with firms on the x-axis and their distances on the y-axis. A notable feature in the dendrogram is the natural division in the data observed between distances of three and four, suggesting this range as an optimal value for the Maximum Dissimilarity Threshold (DMax). DMax, a critical hyperparameter in this clustering algorithm, influences the clustering granularity. A lower DMax value typically yields a larger number of smaller clusters, whereas a higher value results in fewer, larger clusters.

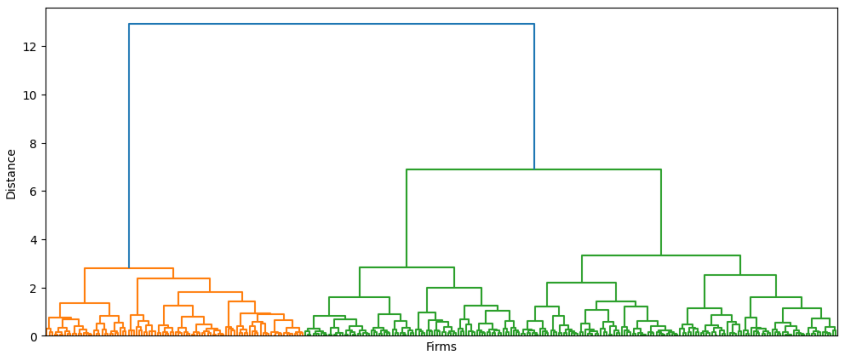

Figure 5: Dendrogram Showcasing Firm Clustering.

Firms are aligned along the x-axis, with Hierarchical clustering distances on the y-axis, illustrating the natural data division and optimal DMax for clustering analysis.

Applying Hierarchical Clustering with DMax set at three revealed four distinct clusters (Figure 6), differing from the three clusters identified by K-Means. Conversely, increasing the DMax value to around seven (Figure 7) led to broader groupings, with fewer clusters.

This variability illustrates the fine-tuning capability of Hierarchical Clustering: capturing finer distinctions at lower DMax values and presenting a more generalized view of the dataset at higher values.

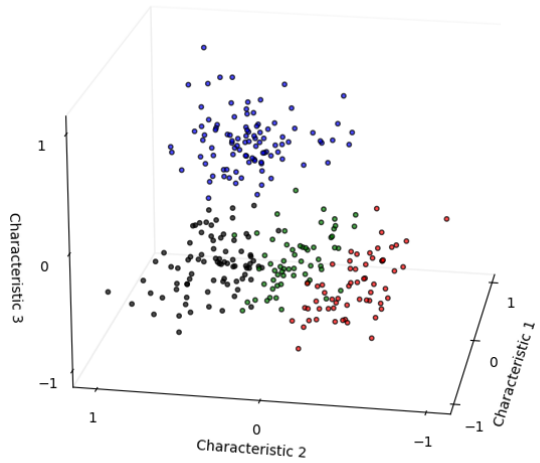

Figure 6: Hierarchical Clustering with DMax=3.

Demonstrating the formation of four distinct clusters, highlighting the method's sensitivity to finer data structures.

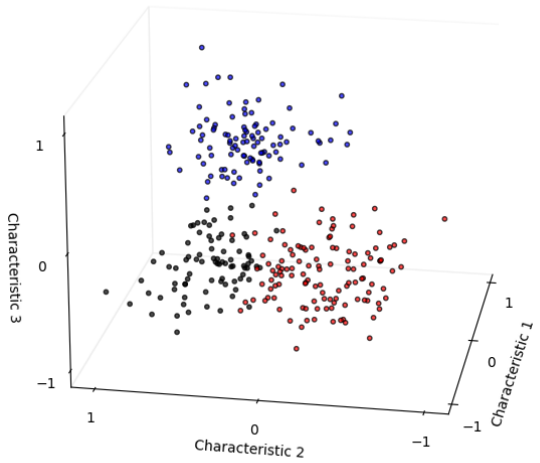

Figure 7: Hierarchical Clustering with DMax=7.

Showcasing broader groupings with fewer clusters, reflecting the method's adaptability to higher DMax values.

Besides the choice of the max distance metric, Hierarchical clustering encompasses various linkage methods, which are strategies to cluster initial data points. These methods, each with their unique approach to defining the closeness of data points, are elaborated in the appendix B. Nonetheless in this thesis, I have employed the Ward method. The Ward method is particularly effective in minimizing the total within-cluster variance (similarly to K-Means). It operates by iteratively merging the pair of clusters that result in the least increase in total within-cluster variance at each step. This method is especially adept at identifying groups with similar variances, which is a significant advantage when analyzing financial data. In such contexts, clusters often correspond to different market sectors or investment styles, and the Ward method's propensity for creating homogenous clusters makes it highly suitable for discerning these intricate patterns.

By using the Ward method the mathematical formulation, following the main framework, is applied separately for each factor f and each time period t. The method clusters stocks based on all factors except the current factor f, resulting in clusters $\mathcal{C}_{t,-f}$. This clustering is crucial for computing the cluster-neutral adjustments of factor f, aligning with the objective to neutralize cluster-specific biases in factor values.

The model for Hierarchical clustering using the Ward method in this study is defined by the following objective:

$$\text{Minimize}\,\Delta(\text{Var}) = \sum_{c \in C_{\text{new}}} \text{Var}(c) - \sum_{c \in C_{\text{old}}} \text{Var}(c) \tag{3}$$

Where:

- $\Delta(\text{Var})$ represents the increase in total variance due to merging clusters, which the method aims to minimize.
- C_{old} and C_{new} denote the sets of clusters before and after merging, respectively.
- $\text{Var}(c)$ is the variance within a cluster c.

This approach ensures that stocks are grouped into clusters with minimal internal variance, setting the stage for the subsequent adjustment of factor values to be cluster-neutral.

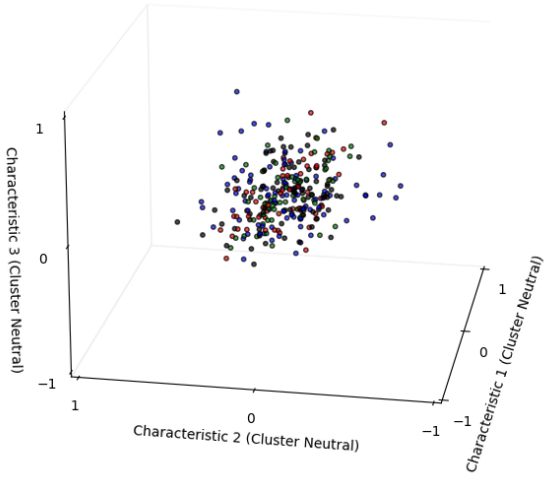

Figure 8: Cluster Neutral Factors with DMax=3.

Illustrating a central concentration in four initial clusters, reflecting the impact of a lower DMax on cluster neutral characteristics.

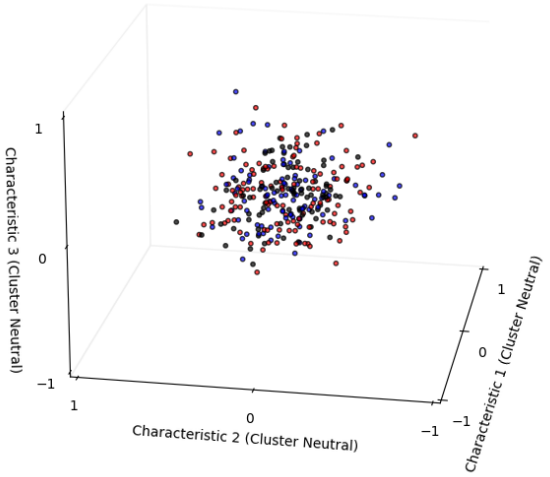

Figure 9: Cluster Neutral Factors with DMax=7.

Displaying a spread in three clusters, showcasing the effect of a higher DMax on the distribution of cluster neutral characteristics.

The resulting cluster neutral factors, as seen in Figure 8 and 9, exhibit significant changes in means, variance, and distribution. In Figure 8, the four initial clusters demonstrate data that is more centrally concentrated around a single point, compared to the three clusters shown in Figure 9. This highlights how varying the DMax threshold can distinctly affect the cluster neutral factors, emphasizing the need for careful selection of this threshold. Although it does not directly appear in the variance minimization formula (3), DMax influences the clustering process by setting a threshold for the maximum permissible dissimilarity between clusters. Once the dissimilarity between any pair of clusters exceeds this max distance, the algorithm ceases to merge further clusters.

To ensure that DMax operates at its optimum, I adopt a strategy akin to K-Means. This involves starting with a broad spectrum of DMax values as indicated by the dendrogram, and then iteratively assessing the fit of the clusters using the metrics outlined in section 3.5.3.

In conclusion, Hierarchical clustering distinguishes itself with its flexibility and depth, particularly in handling complex datasets. Unlike K-Means, Hierarchical clustering does not require pre-specifying the number of clusters, making it suitable for datasets where the natural

cluster count is unclear. It also excels in identifying clusters with complex structures, as it does not assume spherical cluster shapes like K-Means. While it may be computationally more intensive, especially for large datasets, its ability to create a dendrogram provides valuable insights into the data's structure. Despite these strengths, Hierarchical clustering, like any method, has its limitations, but it remains an indispensable tool in the clustering toolkit, often complementing other techniques to provide a comprehensive understanding of the underlying patterns in diverse data scenarios

3.3.3 *DBSCAN Clustering*

While Hierarchical Clustering offers a more intricate view of data relationships compared to the centroid-based spherical clusters typically generated by K-Means, its effectiveness can be limited when dealing with varying densities and noise in datasets. This is where Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) becomes highly relevant. DBSCAN, like Hierarchical Clustering, does not require the pre-specification of the number of clusters, but it uniquely excels in identifying clusters based on 'dense'[14] areas of data points. This feature is particularly effective for financial datasets with their complex, non-linear relationships (Gu et al., 2020).

DBSCAN distinguishes itself with two key parameters: ϵ (epsilon) and MinPts. Epsilon defines the search radius around each data point to find neighbors, while MinPts is the minimum number of points required to form a cluster. Its ability to discover clusters of arbitrary shapes and sizes, coupled with its efficacy in managing outliers, makes DBSCAN a powerful tool for revealing intricate and varied patterns in stock market data. This versatility and robustness against noise make DBSCAN a preferred choice in many complex real-world clustering scenarios.

In Figure 10, the results of applying DBSCAN to the example dataset are illustrated, using an epsilon (ϵ) value of 0.2 (remember, the data is scaled between -1 and 1) and a minimum samples (MinPts) threshold of 5. As depicted in the figure, two primary clusters have been identified: one represented in red and the other in blue. Additionally, there is a third category, marked in white, which represents the outliers. This showcases the unique capabilities of DBSCAN of handling noisy data, like stock data.

[14] In the context of DBSCAN, 'dense' refers to areas of the dataset where data points are closely packed together. The density of an area is determined by the number of points within a specified radius. This concept is crucial for DBSCAN, as it forms the basis of identifying clusters - a cluster is defined as a region of high density surrounded by a region of low density.

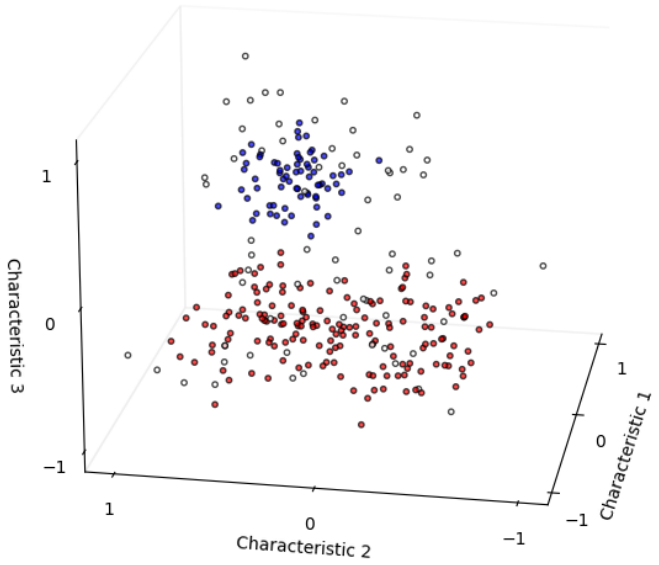

Figure 10: DBSCAN Cluster Analysis.

Using $\epsilon = 0.2$ and MinPts=5, two main clusters (red and blue) and outliers (white) are identified, demonstrating DBSCAN's effectiveness in distinguishing core, border, and noise points.

The mathematical formulation is then also quite different from K-Means or Hierarchical clustering: For each attribute f and each time point t, we identify clusters of stocks. A cluster is formed around a stock based on two parameters: ε (epsilon) and MinPts. Mathematically, this is expressed as:

$$\text{For each } x_{i,t,-f}, \text{ form a cluster } c_{j,t} \text{ if } \left|N_\varepsilon(x_{i,t,-f})\right| \geq \text{MinPts} \tag{4}$$

Here, $N_\varepsilon(x_{i,t,-f})$ represents the set of stocks within the ε-neighborhood of stock i for all attributes except f. A stock is part of N_ε if it's within an ε distance from stock i, considering all attributes except f.

After identifying clusters, the conceptual framework is followed by subtracting the mean to get the cluster neutral values:

$$x'_{i,t,f} = x_{i,t,f} - \text{mean}(c_{j,t}, f) \tag{5}$$

where $\text{mean}(c_{j,t}, f)$ represents the mean of attribute f in cluster $c_{j,t}$ at time t.

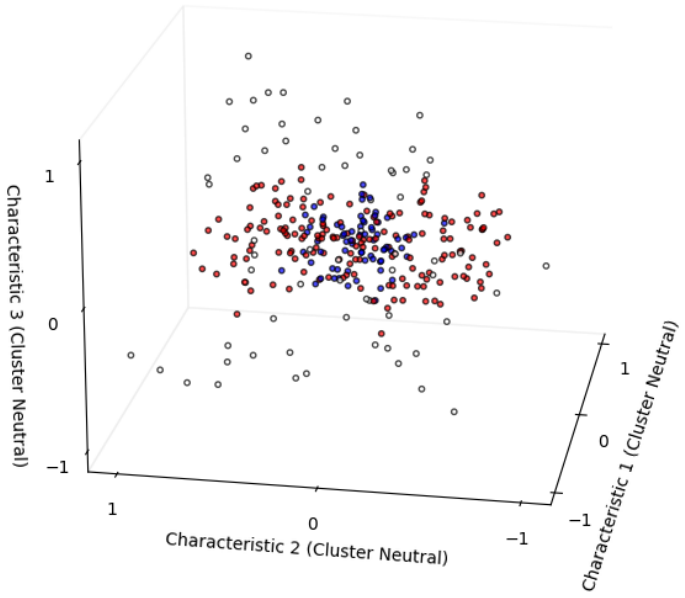

Figure 11: DBSCAN's Impact on Cluster Neutral Factors.

Demonstrating variations in cluster characteristics and the exclusion of outliers, highlighting the influence of epsilon and MinPts settings on factor cluster neutral formation and outlier identification.

The application of DBSCAN to the analysis leads to significant changes in the cluster's neutral factors, as shown in Figure 11. The different shapes identified by DBSCAN correspond to variations in these factors. Additionally, since outliers are excluded from the cluster neutral process, they remain relatively dispersed, resulting in some information loss. The choice of epsilon and MinPts parameters directly affects the shape of the clusters and determines which firms are considered outliers.

To optimize the settings of epsilon and MinPts, a strategy similar to the one used in K-Means is employed. This involves experimenting with a range of values for these parameters, guided by preliminary data analysis, and iteratively assessing the clustering results. The goodness of fit for the clusters can be evaluated using metrics discussed in section 3.5.3, ensuring that the chosen parameters effectively capture the underlying data structure.

In conclusion, DBSCAN stands out for its ability to identify clusters based on data density and to handle outliers effectively. Unlike K-Means, DBSCAN does not require pre-specification of the number of clusters and is capable of identifying clusters of arbitrary shapes, making it particularly suitable for complex datasets. This flexibility, coupled with its robustness to outliers, makes DBSCAN a valuable tool in the clustering toolkit. While DBSCAN is less sensitive to the shape of clusters compared to K-Means and Hierarchical clustering, its reliance on density parameters means careful calibration is essential for optimal performance. Despite these considerations, DBSCAN's ability to reveal intricate patterns in diverse datasets makes it an indispensable technique in data analysis scenarios.

3.4 *OLS-Based Factor Neutralization*

To benchmark[15] the clustering techniques, Ordinary Least Squares (OLS) regression is employed for factor neutralization. This involves regressing a target factor, denoted as f, against other factors (collectively referred to as $-f$) in the dataset. The residual (ε) from this regression represents the cluster-neutral value of f. This value captures the unique aspect of f that isn't explained by the other factors.

Mathematically, the regression for each factor f and each time period t is expressed as:

$$f = \beta_0 + \beta_1 \times factor\_1 + \beta_2 \times factor\_2 + \ldots + \beta_n \times factor\_n + \varepsilon \tag{6}$$

Here, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients, and ε is the residual, providing the cluster-neutral factor values. This approach helps isolate the specific influence of f on stock behavior, separating it from the effects of correlated factors.

While the OLS regression approach focuses on quantifying and removing the influences of correlated factors to isolate the unique contribution of the target factor f, cluster-neutral factors derived from clustering techniques like DBSCAN emphasize adjusting data based on intra-cluster characteristics. In clustering, the adjustment aims to reflect the common properties or behaviors within each cluster, whereas OLS regression isolates the specific influence of f independent of other factors. Thus, while both methods aim to neutralize factors, they do so from different perspectives and with distinct outcomes.

3.5 *Evaluation*

Having developed the cluster-neutral factors, it is imperative to undertake a comprehensive evaluation of the differences between these and the original factors. This evaluation will be methodically executed through two principal approaches: an analysis of portfolio performance and comparative panel regression studies

3.5.1 *Panel Regression*

This section explains a panel regression to explore the influence of various factors on stock returns. It focuses on understanding the role of both original factors and cluster-neutral factors.

The analysis proceeds through several stages. Initially, a panel regression incorporating all available (original) factors is conducted to establish a baseline for their impact on stock returns. This is followed by a regression with cluster-neutral factors, allowing for the assessment of their distinct contributions. Subsequently, a combined analysis using both sets of factors is performed to compare their relative effects. The final model includes cluster fixed effects, addressing unobserved heterogeneity between clusters. Throughout these stages, time fixed effects are consistently included to mitigate time-specific influences.

For robust statistical inference, the study employs clustered standard errors in the estimation of t-values for regression coefficients. This approach is vital in financial data analysis to account for potential cross-sectional and temporal correlations.

[15] This approach is exclusive to the portfolio sorting section because the OLS method does not generate cluster fixed effects. This complicates comparison.

The panel regression model is formulated as follows:

$$Y_{it} = \alpha + \beta_1 X_{1,it} + \beta_2 X_{2,it} + \ldots + \beta_n X_{n,it} + \mu_i + \gamma_t + \varepsilon_{it} \tag{7}$$

where:

- Y_{it} represents the dependent variable, namely the stock return for stock i at time t.
- $X_{1,it}, \ldots, X_{n,it}$ denote the independent variables, encompassing all factors under consideration.
- μ_i indicates the cluster fixed effects.
- γ_t denotes the time fixed effects.
- ε_{it} is the error term in the model.

This structured approach to panel regression analysis, incorporating both original and cluster-neutral factors with robust standard error estimation, provides a comprehensive understanding of the factors influencing stock returns. The inclusion of cluster effects and the comparison across different model setups contribute to a nuanced analysis of these impacts.

3.5.2 *Decile Portfolio Sort*

This process involves sorting stocks into deciles based on various factors, both original and cluster-neutral. For each factor, and on a monthly basis, stocks are ranked. An average is then calculated across the entire dataset, and the dataset is divided into deciles. The 1st decile contains stocks with the lowest factor values, and the 10th decile contains those with the highest. Average future returns for each decile are value-weighted, ensuring that larger stocks have a proportionately greater impact on the average return for each decile.

In addition to average future returns for each decile, this methodology includes the calculation of a long-short portfolio return (10th decile minus the 1st decile). Alphas are calculated using the Fama and French five-factor model plus Carhart's momentum.

Additionally, the Sharpe Ratio, a measure adjusted for risk, is calculated using the formula:

$$\text{Sharpe Ratio} = \frac{\bar{R}_p - R_f}{\sigma_p} \tag{8}$$

where $\bar{R}_p$ is the average return of the portfolio, R_f is the risk-free rate, and σ_p is the standard deviation of the portfolio's excess return. In this study, the Sharpe Ratio is annualized assuming monthly data.

This methodology allows for a comprehensive comparison between original and cluster-neutral factors by analyzing the sorted portfolios' returns, Sharpe Ratios, and alphas across different models. The results offer insights into the efficacy of cluster-neutral factors in predicting stock performance.

3.5.3 *Robustness*

The integrity of the results critically depends on the effectiveness of the clustering methods in identifying meaningful groups within the data. It is imperative to meticulously evaluate

how each clustering approach performs under different configurations and assumptions. This aspect of the thesis, nuanced and pivotal, diverges from traditional analyses that primarily focus on end goals like out-of-sample returns (as seen in Seyfi, 2022). My interest lies not solely in predictive power but in understanding how factors behave independently in the cross-section when isolated from their clusters. Whether these cluster-neutral factors are more or less predictive of stock returns, the outcome is equally significant.

The aim is to ensure the stability, consistency, and meaningfulness of the resulting clusters, verifying they accurately reflect the data's underlying structure.

Detailed mathematical explanations and formulations for these assessments are provided in Appendix B. Additionally, sensitivity analyses for various hyperparameter settings, vital for panel regression and portfolio sorting, are detailed in Appendix C. This approach guarantees a comprehensive evaluation of the methods, blending theoretical precision with practical applicability.

Therefore, the evaluation employs specific metrics and analyses to assess not only the stability and consistency of the clusters but also their quality and fit.

- **Silhouette Score:** Measures how similar an object is to its own cluster compared to other clusters. A higher Silhouette Score indicates better-defined clusters.
- **Davies-Bouldin Index:** Evaluates the cluster separation and compactness. A lower Davies-Bouldin Index suggests better clustering.
- **Calinski-Harabasz Index:** Also known as the Variance Ratio Criterion, this index measures the dispersion between and within clusters. A higher value indicates better clustering.
- **Stability Metric:** Quantifies the stability of the clusters by assessing the changes in cluster assignments for each stock over time. A lower average number of changes implies higher stability of the clusters.
- **Correlation with Cluster Labels:** Calculates the correlation between each factor and its corresponding cluster label to understand the relationship between the factor values and the cluster assignments.

Essentially, clusters are created for each factor (Formula 1), so the metrics are averaged across all clusters to allow for easier comparison.

4 RESULTS

This section presents the key findings of the thesis, starting with panel regression, continuing with portfolio sorting, and concluding with an examination of cluster fit and the robustness of the results. It begins by discussing the outcomes of the DBSCAN model in the panel regression analysis, then moves on to Hierarchical clustering and K-Means. This sequence differs from the methodology section, where DBSCAN was introduced last. Considering DBSCAN's robustness (see Robustness section (4.3)), this order allows for a comparative analysis of the other models against the most effective one tested.

4.1 *Panel Regression*

4.1.1 *DBSCAN*

Column	(1)	(2)	(3)	(4)
Dependent Variable	Excess Return at t + 1			
Constant	0.74*** (0.24)	0.74*** (0.24)	0.74*** (0.24)	0.68** (0.26)
Accruals	0.06*** (0.02)	-	0.01 (0.74)	-0.14 (0.79)
AssetGrowth	0.20*** (0.03)	-	0.89** (0.45)	1.50 (0.89)
BM	0.80*** (0.07)	-	1.55*** (0.45)	1.28** (0.60)
Beta	-0.15* (0.07)	-	0.09 (0.40)	0.38 (0.52)
CompEquIss	0.12*** (0.03)	-	1.80** (0.90)	2.11* (1.16)
DolVol	1.56*** (0.13)	-	2.13*** (0.47)	1.97*** (0.42)
GP	0.21*** (0.04)	-	-0.13 (0.41)	-0.28 (0.42)
IndMom	0.26*** (0.05)	-	1.14 (0.68)	1.44** (0.69)
InvestPPEInv	-0.002 (0.03)	-	1.06 (0.56)	3.50*** (1.08)
Investment	0.07*** (0.01)	-	0.91 (1.27)	0.88 (1.30)
MaxRet	-0.02 (0.06)	-	0.09 (0.45)	0.60 (0.56)
Mom12m	0.36*** (0.05)	-	0.85 (0.62)	0.38 (0.67)
Mom6m	-0.15** (0.06)	-	-0.65 (0.65)	-1.00 (0.71)
Mom1m	-0.76*** (0.09)	-	1.80** (0.82)	1.74** (0.84)
OScore	-0.10*** (0.02)	-	-0.53 (0.34)	-0.73 (0.84)
OperProf	0.07*** (0.01)	-	0.49 (0.93)	0.18 (1.02)
RoE	-0.08*** (0.02)	-	0.34 (0.75)	0.61 (0.79)
ShareIss5Y	0.03 (0.02)	-	0.50 (1.13)	1.11 (1.69)
ShareVol	-0.18*** (0.02)	-	-1.16 (0.68)	-1.01 (0.73)
dNoa	0.04 (0.03)	-	-1.15* (0.55)	-2.55** (1.07)
MarketCap	0.05* (0.02)	-	3.05 (2.28)	2.42 (2.39)
roaq	0.16*** (0.03)	-	0.35 (0.30)	0.40 (0.31)
	Cluster Neutral Factors			
Accruals	-	0.05* (0.02)	0.04 (0.74)	0.20 (0.78)
AssetGrowth	-	0.21*** (0.03)	-0.67 (0.43)	-1.27 (0.87)
BM	-	0.75*** (0.07)	-0.73** (0.40)	-0.48 (0.57)
Beta	-	-0.14** (0.06)	-0.24 (0.37)	-0.53 (0.50)
CompEquIss	-	0.10** (0.03)	-1.66* (0.89)	-1.99 (1.15)
DolVol	-	1.50*** (0.12)	-0.56 (0.39)	-0.41 (0.38)
GP	-	0.19*** (0.04)	0.34 (0.40)	0.48 (0.41)
IndMom	-	0.25*** (0.05)	-0.88 (0.65)	-1.18 (0.66)
InvestPPEInv	-	-0.0021 (0.03)	-1.05* (0.55)	-3.47*** (1.07)
Investment	-	0.08*** (0.01)	-0.84 (1.27)	-0.82 (1.30)
MaxRet	-	-0.001 (0.06)	-0.10 (0.42)	-0.60 (0.53)
Mom12m	-	0.32*** (0.05)	-0.50 (0.58)	-0.03 (0.63)
Mom6m	-	-0.16* (0.06)	0.49 (0.60)	0.70 (0.63)
Mom1m	-	-0.78*** (0.08)	-2.57*** (0.77)	-2.51*** (0.79)
OScore	-	-0.001 (0.03)	0.39 (0.30)	0.60 (0.81)
OperProf	-	0.07*** (0.01)	-0.42 (0.93)	-0.11 (1.02)
RoE	-	-0.08*** (0.02)	-0.42 (0.74)	-0.69 (0.78)
ShareIss5Y	-	0.07*** (0.02)	-0.46 (1.12)	-1.48 (1.72)
ShareVol	-	-0.08*** (0.02)	1.17 (0.66)	0.72 (0.72)
dNoa	-	0.05* (0.03)	1.17** (0.54)	2.56** (1.06)
MarketCap	-	0.03 (0.02)	-3.01 (2.28)	-2.41 (2.39)
roaq	-	0.15*** (0.03)	-0.18 (0.28)	-0.22 (0.29)
Time Fixed Effects	X	X	X	X
Cluster Fixed Effects	-	-	-	X

Table 1: Panel Regression For DBSCAN Model.

This table presents the results of panel regression analyses for factor neutralization utilizing the most robust DBSCAN model (epsilon = 4 and min samples = 50; see Section 4.3) for "Return at $t+1$" across four models. Model (1) utilizes original factors, Model (2) applies cluster-neutral factors, and Models (3) and (4) integrate these factors, with Model (4) additionally incorporating cluster fixed effects. Coefficients and standard errors (presented in parentheses) are provided for each variable, with significance levels denoted as: '***' ($p < 0.01$), '**' ($p < 0.05$), '*' ($p < 0.1$). Time-fixed effects are included in all models, while cluster fixed effects are unique to Model (4). Moreover, all factors are standardized through cross-sectional demeaning and unit variance scaling, meaning that an increase of one standard deviation in a coefficient results in a corresponding increase in the dependent variable by the coefficient's amount, which is expressed in percentages. The results highlight the relative influence of various factors on future stock returns for 22 firm level characteristics, accounting for potential cross-sectional and temporal correlations with clustered standard errors.

In Table 1, the results for the DBSCAN model are presented. This addresses the main objective of this thesis: to investigate whether cluster neutral factors behave differently in relation to stock returns compared to their traditional counterparts.

In columns (1) and (2), the two different factors are regressed separately. Concluding that most factors showed no initial difference after cluster neutralization. However, a moderate change is evident. The probability of bankruptcy, as represented by the O-Score, shifts from a significant negative impact in column (1) (coefficient -0.10***)[16] to an insignificant effect in column (2) (coefficient -0.01). This change suggests that the influence of the O-Score on stock returns is, in part, attributable to the clusters in which firms are situated. The growth in net operating assets (dNOA) shows the opposite effect, becoming significant after cluster neutralization (from 0.04 in column (1) to 0.05* in column (2)). Lastly, the size of the firm (MarketCap) proves to be an insignificant predictor after factors neutralization (from 0.05* in column (1), to 0.03 in column (2)).

In addition, columns (3) and (4) examine the significance of the coefficient shifts observed in columns (1) and (2) by concurrently incorporating traditional and cluster-neutral factors. This assessment aims to determine whether the impact, whether positive or negative, of cluster neutralization is significant. Furthermore, the inclusion of cluster fixed effects in column (4) aims to capture broader sector-wide influences, contrasting the more specific insights of cluster-neutral factors within a cluster.

To provide specific explanations for each factor, the effects of the Book to Market (BM) ratio can be described as follows:

- **Column (1) - Traditional Analysis:** The traditional BM ratio shows a positive relationship with stock returns, as evidenced by a coefficient of 0.80***.
- **Column (2) - BM (Cluster-Neutral):** In this column, the BM ratio, adjusted for intra-cluster variance (i.e., variations within similar groups), shows a slightly reduced but still positive effect (0.75***). This suggests that even when accounting for the unique characteristics within the same group, the BM ratio maintains its positive relationship with returns, albeit slightly diminished.
- **Column (3) - Combining Factors:** When both traditional and cluster-neutral BM factors are included, there's a notable shift. The traditional BM's coefficient increases significantly to 1.55***, indicating a stronger positive effect. Conversely, the cluster-neutral BM turns negative, when controlled for intra-cluster variance and alongside the traditional factor, BM's effect on stock returns might be less positive(-0.73**).
- **Column (4) - Including Inter-cluster Variance:** In this model, which controls for both intra-cluster (within the same group) and inter-cluster (across different groups) variance, the cluster-neutral BM remains negative (-0.48), Meanwhile, the traditional BM is still positively related to stock returns (1.28**), indicating that when controlled for other factors and there cluster neutral counterpart, together with cluster fixed effects. The relation ship with stock returns increases significantly.

Following this explanation some additional factors show significant results:

[16] Since the data is standardized, a positive coefficient of 1.00***, for example, would indicate that a one standard deviation increase in a factor corresponds to a 1 percent increase in tomorrow's excess return.

In the regression analysis for Composite Equity Issuance (CompEquIss), a significant change is observed in the traditional factor's coefficient, which increases from 0.12*** in column (1) to 2.11* in column (4). Conversely, the cluster-neutral version of CompEquIss shifts from a positive coefficient of 0.10** in column (2) to a negative -1.99 in column (4), though this change is not statistically significant. These variations highlight differing perceptions of CompEquIss under different analytical conditions.

Dollar Trading Volume (DolVol) shows a consistent pattern in the regression. Initially, DolVol (traditional factor) emerges as a significant predictor in column (1) with a coefficient of 1.56***. This significance slightly decreases in column (2) but remains notable. In column (4), DolVol (traditional factor) maintains its significance with an increased coefficient of 1.97**, despite the introduction of controls for inter-cluster variance. The cluster-neutral variant of DolVol, however, does not show significant results in either column (3) or column (4).

Industry Momentum (IndMom) initially shows a positive effect in columns (1) and (2), with a coefficient of 0.26*** and 0.25*** respectively. In columns (3) and (4), no significant difference is found between the cluster-neutral factor and the traditional factors. However, the traditional IndMom becomes significant and larger in these columns when adjusted for cluster effects and other variables.

Investment in Property, Plant, and Equipment (InvestPPEInv) undergoes a remarkable transformation across the regression. In column (1), its coefficient is negligible and statistically insignificant at -0.002. However, in column (4), the coefficient for InvestPPEInv dramatically increases to 3.50*** for the traditional factor and -3.48*** for the cluster-neutral version, suggesting a significant, yet complex, impact on stock returns under different conditions.

Short term momentum (Mom1m) initially shows a negative coefficient of approximately -0.76** in columns (1) and (2), indicating a reversal effect. However, in columns (3) and (4), the traditional variant of Mom1m shifts to a positive coefficient, registering 1.80** in column (3) and 1.74** in column (4), while the cluster-neutral variant becomes more negative, reaching -2.57*** in column (3) and -2.51*** in column (4). Finally, the growth in net operating assets (dNOA) presents a varied pattern. In column (1), dNOA is statistically insignificant, but in column (2), it becomes significantly positive. In columns (3) and (4), the positive impact of the cluster-neutral dNOA becomes more pronounced with coefficients of 1.17** and 2.56**, respectively. In contrast, the traditional dNOA shows a negative relationship in these columns, with coefficients of -1.17* and -2.60**. Besides multiple factors showing significant differences in column (4) compared to columns (1) and (2), there are also factors that do not show much change in their coefficients, such as the other momentum factors (Mom6m and Mom12m), the CAPM Beta, or the return on equity (RoE). This suggests these factors are strong predictors on their own and are not significantly influenced by other factors in their close proximity.

In summary, column (4) isolates the effect of cluster neutralization by controlling for cluster fixed effects and traditional factors. For all 22 factors, the impact is opposite to their traditional counterparts. This indicates that firms with similar characteristics within a group enhance the relationship of the factors with stock returns. When this group influence is removed through factor cluster neutralization, the effect becomes less pronounced.

In addition, the drop in the constant from 0.74*** to 0.68***(in column (4)) upon adding cluster fixed effects, with factors standardized to have a mean of 0, means that after accounting

for the inherent differences between clusters, the baseline expected excess return (when all other variables are at their mean) slightly decreases.

Overall, these findings underscore the complexities in stock market behavior, with some factors showing influence from group dynamics and external conditions, while others maintain strong predictive power independently.

4.1.2 *Hierarchical*

Column	(1)	(2)	(3)	(4)
Dependent Variable		Return at $t+1$		
Constant	0.74*** (0.24)	0.74*** (0.24)	0.74** (0.24)	0.18 (0.59)
Accruals	0.06** (0.02)	-	0.18 (0.24)	0.20 (0.25)
AssetGrowth	0.20*** (0.03)	-	0.19*** (0.06)	0.14 (0.09)
BM	0.80*** (0.07)	-	1.22*** (0.27)	1.12*** (0.28)
Beta	-0.15* (0.07)	-	0.10 (0.28)	0.05 (0.29)
CompEquIss	0.12*** (0.03)	-	0.31* (0.18)	0.33 (0.21)
DolVol	1.56*** (0.13)	-	1.36*** (0.32)	1.44*** (0.31)
GP	0.21*** (0.04)	-	-0.01 (0.21)	-0.16 (0.23)
IndMom	0.26*** (0.05)	-	0.99 (0.59)	1.01 (0.59)
InvestPPEInv	-0.002 (0.03)	-	-0.04 (0.07)	-0.04 (0.07)
Investment	0.07*** (0.01)	-	0.95** (0.39)	1.04** (0.40)
MaxRet	-0.02 (0.06)	-	-0.35 (0.35)	-0.36 (0.35)
Mom12m	0.36*** (0.05)	-	0.57* (0.32)	0.54* (0.30)
Mom6m	-0.15** (0.06)	-	-0.61 (0.36)	-0.63 (0.35)
Mom1m	-0.77*** (0.09)	-	1.27 (1.00)	1.38 (0.95)
OScore	-0.10*** (0.02)	-	-0.39* (0.22)	-0.43* (0.22)
OperProf	0.07*** (0.01)	-	0.16*** (0.06)	0.17*** (0.05)
RoE	-0.08*** (0.02)	-	-0.10 (0.06)	-0.10 (0.06)
ShareIss5Y	0.03 (0.02)	-	0.03 (0.19)	0.07 (0.19)
ShareVol	-0.18*** (0.02)	-	-0.60** (0.28)	-0.60** (0.29)
dNoa	0.04 (0.03)	-	-0.05 (0.06)	-0.07 (0.06)
MarketCap	0.05* (0.02)	-	0.00 (0.02)	0.06 (0.04)
roaq	0.16*** (0.03)	-	0.14 (0.22)	0.01 (0.23)
		Cluster Neutral Factors		
Accruals	-	0.06*** (0.02)	-0.13 (0.23)	-0.15 (0.24)
AssetGrowth	-	0.19*** (0.02)	0.004 (0.04)	0.03 (0.05)
BM	-	0.74*** (0.06)	-0.42* (0.24)	-0.34 (0.25)
Beta	-	-0.18** (0.06)	-0.27 (0.26)	-0.22 (0.27)
CompEquIss	-	0.07* (0.03)	-0.19 (0.15)	-0.21 (0.19)
DolVol	-	1.11*** (0.07)	0.17 (0.22)	0.14 (0.21)
GP	-	0.20*** (0.04)	0.22 (0.21)	0.36 (0.23)
IndMom	-	0.23*** (0.04)	-0.73 (0.56)	-0.75 (0.56)
InvestPPEInv	-	0.05* (0.02)	0.03 (0.05)	0.03 (0.06)
Investment	-	0.06*** (0.01)	-0.89** (0.38)	-0.98** (0.40)
MaxRet	-	-0.03 (0.05)	0.35 (0.31)	0.36 (0.31)
Mom12m	-	0.26*** (0.05)	-0.21 (0.29)	-0.19 (0.27)
Mom6m	-	-0.13*** (0.05)	0.45 (0.32)	0.47 (0.32)
Mom1m	-	-0.77*** (0.07)	-2.04** (0.95)	-2.16** (0.90)
OScore	-	-0.04 (0.03)	0.34* (0.20)	0.37* (0.20)
OperProf	-	0.03** (0.01)	-0.09* (0.05)	-0.09** (0.05)
RoE	-	-0.07*** (0.02)	0.01 (0.05)	0.02 (0.05)
ShareIss5Y	-	0.10*** (0.02)	0.01 (0.17)	-0.02 (0.18)
ShareVol	-	0.02 (0.02)	0.49** (0.28)	0.47* (0.30)
dNoa	-	0.09*** (0.02)	0.07* (0.04)	0.08* (0.04)
MarketCap	-	0.01 (0.02)	0.05 (0.04)	0.00 (0.04)
roaq	-	0.14*** (0.03)	0.04 (0.20)	0.16 (0.21)
Time Fixed Effects	X	X	X	X
Cluster Fixed Effects	-	-	-	X

Table 2: Panel Regression Results For Hierarchical Clustering.

This table presents the results of panel regression analyses for factor neutralization utilizing the most robust Hierarchical clustering model (Dmax = 80; see Section 4.3). (See Table 1 for extensive description of the table)

The panel regression results for the Hierarchical Clustering model are presented in Table 2.

Upon comparing these results with those obtained from the DBSCAN clustering model, it is observed that several factors exhibit similar effects in both models: BM, DolVol, dNoa, MarketCap, Oscore and Mom1m. However, the analysis reveals a divergence in the significance of three other factors when compared between the two models: CompEqIss, IndMom and investPPEInv.

In addition, tree more conclusions can be drawn from the table:

In the panel regression analysis, capital expenditure (Investment) shows a significant impact in both its traditional and cluster-neutral forms, as observed in columns (3) and (4). Specifically, the traditional factor of Investment demonstrates a coefficient of 1.04** in column (4). This indicates an increased positive correlation between the traditional investment factor and stock returns, especially when accounting for other factors and controlling for variance across clusters. A notable shift is seen with the cluster-neutral counterpart. It exhibits a negative and significant coefficient at -0.98** in column (4), giving evidence of effect of factor neutralization.

Similarly, operating profitability (OperProf) displays a distinct pattern when analyzed alongside its cluster-neutral counterpart. The traditional factor for OperProf shows a positive coefficient of 0.17*** in column (4). In contrast, the cluster-neutral factor for operating profitability turns negative and significant at -0.09** in column (4).

Furthermore, the analysis of the number of traded shares (ShareVol) reveals significant results. In the initial regression (column (1)), ShareVol exhibits a negative coefficient of -0.18***, suggesting a potential correlation with market skepticism or uncertainty. This negative correlation becomes positive in column (2) with a coefficient of 0.02. Moving to columns (3) and (4), the traditional ShareVol factor shows a further decrease to -0.60**, indicating a consistent negative perception associated with high trading volumes. Conversely, the cluster-neutral ShareVol factor turns positive in columns (3) and (4), with a coefficient of 0.47*.

Finally, there is also a significant drop in the constant from 0.74** to 0.18 in the Hierarchical clustering model, upon introducing cluster fixed effects, highlights that once the model accounts for the inherent differences between clusters, the baseline expected return is notably lower. This indicates that much of what was previously attributed to the general market conditions (captured by the constant) is actually due to specific cluster characteristics.

4.1.3 *K-Means*

Variable	(1)	(2)	(3)	(4)
Dependent Variable	Dependent Variable: Return at $t+1$			
Constant	0.74*** (0.24)	0.74*** (0.24)	0.74*** (0.24)	0.29 (0.59)
Accruals	0.06** (0.02)	-	-0.14 (0.28)	-0.18 (0.25)
AssetGrowth	0.20*** (0.03)	-	0.34*** (0.09)	0.35*** (0.07)
BM	0.80*** (0.07)	-	1.24*** (0.22)	1.22*** (0.22)
Beta	-0.15* (0.07)	-	-0.26 (0.38)	-0.29 (0.23)
CompEquIss	0.12*** (0.03)	-	0.43** (0.64)	0.42 (0.22)
DolVol	1.56*** (0.13)	-	1.44*** (0.32)	1.43*** (0.26)
GP	0.21*** (0.04)	-	0.24 (0.33)	0.28 (0.24)
IndMom	0.26*** (0.05)	-	1.18** (0.58)	1.17** (0.50)
InvestPPEInv	-0.002 (0.03)	-	-0.18** (0.09)	-0.17* (0.08)
Investment	0.07*** (0.01)	-	0.78 (0.81)	0.82 (0.54)
MaxRet	-0.02 (0.06)	-	-0.52 (0.41)	-0.46 (0.36)
Mom12m	0.36*** (0.05)	-	0.13 (0.40)	0.09 (0.23)
Mom6m	-0.15** (0.06)	-	-0.34 (0.39)	-0.32 (0.23)
Mom1m	-0.76*** (0.09)	-	0.78 (1.09)	0.81 (0.73)
OScore	-0.10*** (0.02)	-	-0.31 (0.32)	-0.28 (0.17)
OperProf	0.07*** (0.01)	-	0.13** (0.09)	0.12** (0.05)
RoE	-0.08*** (0.02)	-	-0.06 (0.10)	-0.06 (0.06)
ShareIss5Y	0.03 (0.02)	-	-0.04 (0.51)	-0.02 (0.18)
ShareVol	-0.18*** (0.02)	-	-0.67*** (0.24)	-0.68*** (0.20)
dNoa	0.04 (0.03)	-	-0.03 (0.09)	-0.02 (0.06)
MarketCap	0.05* (0.02)	-	0.34 (2.29)	0.24 (2.06)
roaq	0.16*** (0.03)	-	0.09 (0.24)	0.12 (0.23)
Cluster Neutral Factors				
Accruals	-	0.08*** (0.02)	0.19 (0.27)	0.22 (0.24)
AssetGrowth	-	0.16*** (0.02)	-0.10** (0.07)	-0.11** (0.04)
BM	-	0.70*** (0.06)	-0.43** (0.34)	-0.41** (0.18)
Beta	-	-0.16** (0.07)	0.06 (0.35)	0.09 (0.20)
CompEquIss	-	0.05* (0.02)	-1.22* (0.62)	-0.31 (0.20)
DolVol	-	0.99*** (0.08)	0.16 (0.23)	0.10 (0.16)
GP	-	0.17*** (0.04)	-0.07 (0.31)	-0.07 (0.22)
IndMom	-	0.20*** (0.04)	-0.91** (0.56)	-0.91** (0.47)
InvestPPEInv	-	0.07*** (0.02)	0.14** (0.07)	0.14** (0.06)
Investment	-	0.08*** (0.01)	-0.75 (0.80)	-0.75 (0.53)
MaxRet	-	-0.01 (0.05)	0.46 (0.37)	0.48 (0.31)
Mom12m	-	0.23*** (0.05)	0.17 (0.36)	0.24 (0.19)
Mom6m	-	-0.15*** (0.04)	0.14 (0.35)	0.17 (0.18)
Mom1m	-	-0.78*** (0.07)	-1.60** (1.04)	-1.60** (0.67)
OScore	-	-0.00 (0.02)	0.15 (0.30)	0.16 (0.11)
OperProf	-	0.00** (0.00)	-0.06 (0.09)	-0.06 (0.04)
RoE	-	-0.06*** (0.02)	-0.02 (0.10)	-0.02 (0.05)
ShareIss5Y	-	0.12*** (0.02)	0.07 (0.50)	0.06 (0.16)
ShareVol	-	0.03* (0.02)	0.59*** (0.24)	0.57*** (0.19)
dNoa	-	0.07*** (0.02)	0.05 (0.07)	0.05 (0.04)
marketCap	-	0.02 (0.02)	-0.17 (0.38)	-0.18 (0.38)
roaq	-	0.14*** (0.02)	0.06 (0.22)	0.08 (0.20)
Time Fixed Effects	X	X	X	X
Cluster Fixed Effects	-	-	-	X

Table 3: Panel Regression Analysis For K-Means Clustering Model.

This table presents the results of panel regression analyses for factor neutralization utilizing the most robust K-Means model (k=10; see Section 4.3).(See Table 1 for extensive description of the table)

Finally, in Table 3, the results obtained through the application of the K-Means method are shown.

Firstly, these results notably enhance the previously moderate and insignificant effects observed by the DBSCAN for BM and IndMom. These factors now demonstrate significant influences after controlling for cluster effects, as evidenced by their considerable coefficients at -0.39* and -0.91** in column (4), respectively. Furthermore, the analysis validates the

findings for CompEqIss and InvestPPEInv, which were also identified in the DBSCAN's factor neutralization. There is also consistency with the observations for OperProf and ShareVol, as highlighted in the Hierarchical clustering model.

In addition to the factors mentioned, the K-Means clustering model unveils a significant change in the impact of AssetGrowth. While it displayed moderate influence in other models, AssetGrowth becomes more prominent in the K-Means approach. In column (1), AssetGrowth shows a positive effect on stock returns (0.20***), suggesting that an increase in assets typically correlates with higher future returns. This effect is even more marked when considering cluster-fixed effects and cluster-neutral variables, at 0.34***. However, when adjusting for traditional factors alongside cluster-fixed effects, the cluster-neutral variant of AssetGrowth reveals a negative effect at -0.11**.

Lastly, also the constant decrease in column (4) significantly to 0.29, the same as DBSCAN and K-Means.

4.1.4 *Time period Split*

In Table 4, the dataset is divided into two periods: the first 25 years starting in 1971 (Period 1) and the next 25 years ending in 2021 (Period 2). This split allows for a detailed analysis of potential changes in results over time. Table 4 focuses on the outcomes presented in the last column (column (4)).

A notable observation is the increased significance in the second period for both traditional factors and their cluster-neutral counterparts using the DBSCAN model (Column (1 and 2)). Of the seven factors showing significant results across the entire timeline in earlier analyses, only one, Mom1m, maintains significant results in Period 1 for its cluster-neutral counterpart (-2.28**).

In contrast, both K-Means and Hierarchical clustering show similar levels of significance in Period 1 and Period 2. Interestingly, the effect identified by K-Means for InvestPPeInv differs from that found by DBSCAN in Period 2 (Column (1) and (6)). Additionally, factors such as CAPM Beta, the maximum monthly return (MaxRet), return on equity (RoE), and Carhart's momentum factors (Mom6m and Mom12m), show significance in the first period but not in the second.

Operprof emerges as a significant factor not identified by either Hierarchical clustering or K-Means across the entire timeline. In the second period, the impacts of many factors are more pronounced. Factors like DolVol, Indmom, InvestPPeINV, Mom1m, and dNoa continue to exhibit the same or increased effects. However, CompEquIss and BM do not maintain these effects.

In summary, the factors showing the greatest consistency and robustness across different periods and methodologies are DolVol, Mom1m, InvestPPeINV, OperProf, and Indmom. This consistency highlights their significance even after cluster neutralization, indicating their critical importance amidst temporal and methodological changes.

Dependent Variable	Return at $t+1$					
Model	DBSCAN		Hierarchical		K-Means	
Column	(1)	(2)	(3)	(4)	(5)	(6)
Period	**(1971-1996)**	**(1996-2021)**	**(1971-1996)**	**(1996-2021)**	**(1971-1996)**	**(1996-2021)**
Constant	0.49 (0.36)	0.90* (0.40)	0.29 (0.78)	0.36 (0.93)	-0.26 (0.74)	0.85 (0.94)
Accruals	-0.55 (0.98)	2.03 (1.93)	-0.09 (0.19)	0.71 (0.63)	-0.33 (0.22)	0.07 (0.50)
AssetGrowth	1.48 (1.63)	1.41 (1.30)	0.007 (0.08)	0.31 (0.13)	0.22 (0.07)	0.40 (0.12)
BM	1.47 (0.83)	0.65 (0.85)	0.78 (0.22)	1.60 (0.57)	1.21 (0.20)	1.34 (0.41)
Beta	0.08 (0.69)	-0.29 (0.82)	-0.54 (0.27)	0.33 (0.41)	-0.93 (0.23)	0.05 (0.34)
Beta	0.08 (0.69)	-0.29 (0.82)	-0.54 (0.27)*	0.33 (0.41)	-0.93 (0.23)**	0.05 (0.34)
CompEquIss	1.88 (1.53)	1.06 (1.95)	0.25 (0.24)	0.46 (0.33)	0.06 (0.27)	0.82 (0.35)*
DolVol	1.17 (1.58)	3.28 (0.75)***	1.43 (0.24)***	2.23 (0.63)***	1.17 (0.24)***	2.50 (0.51)***
GP	0.37 (0.48)	-0.32 (0.57)	0.10 (0.36)	-0.16 (0.37)	0.49 (0.28)*	0.53 (0.47)
IndMom	0.18 (0.77)	2.46 (1.07)*	0.36 (0.33)	1.64 (1.01)	0.39 (0.32)	1.68 (0.83)*
InvestPPEInv	1.84 (1.65)	5.12 (1.74)***	0.14 (0.09)	-0.09 (0.14)	0.03 (0.08)	-0.29 (0.16)*
Investment	-0.71 (1.96)	2.05 (1.98)	0.47 (0.34)	2.36 (0.90)**	-0.08 (0.44)	1.88 (1.14)
MaxRet	-0.23 (0.63)	1.16 (0.93)	-0.65 (0.32)**	-0.29 (0.59)	-0.78 (0.34)*	-0.15 (0.67)
Mom12m	0.88 (0.91)	-0.49 (0.90)	0.62 (0.32)*	0.41 (0.43)	0.58 (0.19)***	-0.42 (0.36)
Mom6m	-1.20 (0.71)*	-0.78 (1.06)	-0.89 (0.30)***	-0.46 (0.64)	-0.47 (0.20)*	-0.21 (0.42)
Mom1m	1.29 (0.81)	1.65 (1.22)	0.65 (0.70)	1.94 (1.50)	0.72 (0.54)	0.88 (1.23)
OScore	-0.74 (1.78)	-1.22 (1.17)	-0.01 (0.22)	-0.61 (0.38)	-0.16 (0.17)	-0.43 (0.29)
OperProf	2.18 (1.10)*	-2.18 (1.64)	0.17 (0.08)**	0.18 (0.07)**	0.17 (0.08)**	0.11 (0.07)
RoE	0.49 (0.87)	1.78 (1.52)	-0.15 (0.09)*	-0.05 (0.08)	-0.17 (0.09)*	0.01 (0.09)
ShareIss5Y	1.88 (2.19)	1.21 (2.94)	0.32 (0.22)	0.07 (0.30)	0.36 (0.16)*	-0.24 (0.33)
ShareVol	-1.02 (1.22)	-4.25 (2.66)*	-0.03 (0.22)	-2.88 (1.11)**	-0.11 (0.15)	-2.88 (0.82)***
dNoa	-0.77 (2.28)	-3.69 (1.44)**	-0.02 (0.08)	-0.10 (0.12)	0.03 (0.08)	-0.06 (0.10)
market$_c$*ap*	2.48 (3.81)	5.38 (4.07)	0.36 (0.44)	-1.15 (0.85)	0.09 (0.38)	-0.38 (0.72)
roaq	0.07 (0.38)	0.13 (0.49)	-0.25 (0.35)	0.06 (0.33)	0.27 (0.30)	0.12 (0.31)
Cluster Neutral Factors						
Accruals	0.63 (0.98)	-1.98 (1.93)	0.15 (0.18)	-0.67 (0.62)	0.38 (0.20)*	-0.03 (0.48)
AssetGrowth	-1.27 (1.60)	-1.20 (1.26)	0.15 (0.05)***	-0.14 (0.13)	-0.06 (0.04)	-0.18 (0.08)*
BM	-0.51 (0.79)	0.18 (0.79)	0.12 (0.20)	-0.76 (0.51)	-0.26 (0.14)*	-0.48 (0.33)
Beta	-0.32 (0.65)	0.17 (0.79)	0.28 (0.24)	-0.48 (0.38)	0.63 (0.18)***	-0.25 (0.29)
CompEquIss	-1.77 (1.53)	-0.91 (1.93)	-0.15 (0.21)	-0.28 (0.30)	0.05 (0.23)	-0.67 (0.33)*
DolVol	0.16 (1.62)	-0.80 (0.59)	-0.07 (0.14)	0.30 (0.40)	0.13 (0.13)	0.0045 (0.27)
GP	-0.13 (0.47)	0.51 (0.55)	0.12 (0.36)	0.35 (0.36)	-0.25 (0.27)	-0.32 (0.42)
IndMom	0.02 (0.76)	-2.14 (1.02)*	-0.15 (0.32)	-1.32 (0.95)	-0.19 (0.30)	-1.35 (0.77)*
InvestPPEInv	-1.72 (1.63)	-5.15 (1.72)***	-0.03 (0.06)	0.02 (0.14)	0.05 (0.05)	0.18 (0.13)
Investment	0.76 (1.96)	-1.97 (1.98)	-0.42 (0.33)	-2.28 (0.89)**	0.12 (0.43)	-1.80 (1.13)
MaxRet	0.17 (0.62)	-1.16 (0.86)	0.57 (0.29)*	0.31 (0.53)	0.72 (0.29)*	0.25 (0.58)
Mom12m	-0.41 (0.89)	0.71 (0.85)	-0.16 (0.29)	-0.20 (0.38)	-0.11 (0.15)	0.58 (0.28)*
Mom6m	0.90 (0.69)	0.68 (0.98)	0.63 (0.28)*	0.35 (0.56)	0.21 (0.16)	0.11 (0.33)
Mom1m	-2.28 (0.78)***	-2.30 (1.12)*	-1.64 (0.66)**	-2.59 (1.41)*	-1.70 (0.49)***	-1.54 (1.12)
OScore	0.64 (1.74)	1.03 (1.12)	0.004 (0.20)	0.53 (0.33)	0.02 (0.13)	0.18 (0.16)
OperProf	-2.15 (1.25)	2.29 (1.63)	-0.13 (0.07)*	-0.06 (0.07)	-0.13 (0.07)*	-0.07 (0.07)
RoE	-0.55 (0.86)	-1.89 (1.51)	0.08 (0.08)	0.07 (0.08)	0.10 (0.08)	-0.12 (0.08)
ShareIss5Y	-1.81 (2.19)	-1.21 (2.93)	-0.22 (0.21)	-0.09 (0.27)	-0.26 (0.14)*	0.25 (0.30)
ShareVol	0.88 (1.24)	3.87 (2.61)	-0.13 (0.20)	2.79 (1.11)**	0.08 (0.13)	2.71 (0.80)***
dNoa	0.79 (2.25)	3.70 (1.42)**	0.04 (0.05)	0.11 (0.12)	0.01 (0.07)	0.06 (0.10)
MarketCap	-2.47 (3.81)	-5.35 (4.06)	-0.32 (0.42)	1.21 (0.83)	-0.08 (0.37)	0.41 (0.69)
roaq	0.23 (0.38)	-0.13 (0.45)	0.52 (0.33)	-0.04 (0.28)	0.04 (0.28)	-0.08 (0.27)
Time Fixed Effects	X	X	X	X	X	X
Cluster Fixed Effects	X	X	X	X	X	X

Table 4: Differential Impact of Cluster Neutral Factors on Stock Returns Across Two Distinct Periods.

This table presents the results of panel regression analyses for factor neutralization utilizing the most robust models on "Return at $t+1$". The table divides the dataset into two time periods, the first spanning the initial 25 years starting in 1971 (Period 1) and the second covering the last 25 years ending in 2021 (Period 2), to investigate the temporal robustness of factor impacts on future stock returns. Coefficients and standard errors (presented in parentheses) are provided for each variable, with significance levels denoted as: '***' ($p < 0.01$), '**' ($p < 0.05$), '*' ($p < 0.1$). Time-fixed effects and Cluster Fixed Effects are included in all models. Moreover, all factors are standardized through cross-sectional demeaning and unit variance scaling, meaning that an increase of one standard deviation in a coefficient results in a corresponding increase in the dependent variable by the coefficient's amount, which is expressed in percentages. The results highlight the relative influence of various factors on future stock returns for 22 firm level characteristics, accounting for potential cross-sectional and temporal correlations with clustered standard errors.

4.2 *Portfolio Sort*

4.2.1 *Decile Portfolios*

Table 21 plays a crucial role in the analysis by comparing the performance of long-short (LS) portfolios created through traditional methods and a novel cluster-neutral approach. It emphasizes the most effective configurations for each machine learning clustering technique and the ordinary least squares (OLS) method.

The discussion mainly targets factors that demonstrated significant results in the panel regression analysis, alongside notable insights from other factors. For a comprehensive view of the portfolio sorts across all models and their variations, refer to Appendix C.

The study uncovers varied impacts of specific factors across different econometric models. Initially, the Dollar Volume (DolVol) factor exhibits a shift from marginally negative Long-Short (LS) returns (-0.01) in traditional models to positive outcomes in cluster-neutral setups, notably peaking at a 0.76 LS return within the OLS framework. Despite these gains, DolVol's performance is tempered by consistently low or negative Sharpe Ratios (SRs) and Alphas, particularly in the DBSCAN model where it posts a positive LS return of 0.06 but struggles with a low SR (-0.25) and a negative Alpha (-1.67), hinting at heightened risk.

Transitioning to the Maximum Daily Return of the Last Month (MaxRet), there is a movement from insignificance in traditional approaches to a positive stance in cluster-neutral models. This shift is especially pronounced in DBSCAN, where MaxRet achieves an LS return of 0.38, SR of 0.27, and Alpha of 1.10, indicating a tilt towards momentum factor behavior.

The Probability of Bankruptcy (OScore) exhibits variability across models, with a remarkable positive adjustment in the DBSCAN framework, achieving an LS return of 0.19. Conversely, the OLS model records an LS return of 0.45 for OScore but is marred by a lower SR and negative Alpha, suggesting mixed performance.

The metric for Investing in Property, Plant, and Equipment (InvestPPEInv) reveals significant impacts in the DBSCAN model, pointing to industry-specific ramifications. Meanwhile, the book-to-market ratio (BM) transitions from a slight negative influence in traditional models (-0.08) to a more pronounced negative effect in DBSCAN (-0.46).

Compeqissu, notable within the K-Means configuration, experiences a considerable dip in LS returns, moving from 0.83 to 0.35. The relationship of Sharevol with stock returns shifts towards a more positive correlation, especially under the DBSCAN model's scrutiny.

Lastly, Operating Profitability (OperProf) maintains stable returns in both K-Means and Hierarchical models but faces a decline in the DBSCAN and OLS frameworks, underscoring the diversity of factor performances across models. This varied performance across financial metrics and models underscores the complexity and dynamic nature of financial markets, highlighting the importance of nuanced analysis in understanding market behaviors.

However, there are also some factors such as growth in operating assets (dNOA), short term momentum (MOm1m) and capital expenditures (Investment) that do not show much change, implying that the effect of cluster neutralization is rather nuanced on decile portfolio sorts.

					Cluster Neutral											
Factor	Decile	Normal			K = 10			DMax =80			ε = 4, M= 50			OLS		
		R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$
BM	(1)	1.49	0.25	0.36	1.40	0.25	0.33	1.41	0.25	0.40	1.50	0.27	0.43	1.31	0.20	-0.05
	(10)	1.41	0.14	-0.66	1.48	0.18	-0.64	1.52	0.18	-0.69	1.04	0.07	-0.74	1.13	0.14	-0.40
	(LS)	-0.08	-0.12	-1.02	0.08	-0.06	-0.97	0.11	-0.07	-1.09	-0.46	-0.20	-1.16	-0.19	-0.06	-0.34
CompEquIss	(1)	0.76	0.09	-0.02	1.23	0.18	0.12	1.01	0.14	0.05	0.77	0.08	-0.14	1.26	0.20	0.23
	(10)	1.59	0.32	0.61	1.57	0.31	0.56	1.59	0.31	0.58	1.59	0.32	0.62	1.43	0.22	0.28
	(LS)	0.83	0.23	0.63	0.35	0.13	0.44	0.57	0.17	0.53	0.82	0.23	0.76	0.17	0.02	0.04
DolVol	(1)	1.12	0.39	0.93	0.94	0.24	0.83	0.95	0.28	0.88	1.12	0.39	0.93	0.93	0.18	0.53
	(10)	1.11	0.12	-0.78	1.24	0.16	-0.58	1.24	0.15	-0.63	1.18	0.14	-0.74	1.69	0.28	-0.42
	(LS)	-0.01	-0.27	-1.72	0.30	-0.08	-1.41	0.29	-0.13	-1.50	0.06	-0.25	-1.67	0.76	0.10	-0.95
IndMom	(1)	0.74	0.12	-0.13	0.96	0.16	-0.11	0.86	0.15	-0.13	0.80	0.13	-0.14	1.03	0.18	-0.11
	(10)	1.13	0.18	0.05	1.07	0.16	-0.01	1.15	0.18	0.04	1.09	0.17	0.05	1.01	0.15	-0.09
	(LS)	0.38	0.06	0.18	0.11	0.00	0.10	0.30	0.03	0.17	0.29	0.04	0.19	-0.01	-0.02	0.02
InvestPPEInv	(1)	1.23	0.21	0.11	1.20	0.20	0.08	1.25	0.21	0.10	1.08	0.18	0.14	1.56	0.25	-0.05
	(10)	1.21	0.17	-0.23	1.13	0.15	-0.12	1.23	0.18	-0.16	1.32	0.16	-0.43	0.99	0.14	0.00
	(LS)	-0.02	-0.04	-0.34	-0.07	-0.04	-0.21	-0.02	-0.03	-0.25	0.24	-0.02	-0.57	-0.57	-0.11	0.05
Investment	(1)	0.95	0.19	0.35	0.94	0.18	0.33	0.89	0.17	0.34	0.94	0.18	0.35	0.99	0.19	0.31
	(10)	1.72	0.21	0.08	1.61	0.22	0.09	1.58	0.20	0.09	1.67	0.21	0.08	1.71	0.23	0.09
	(LS)	0.78	0.03	-0.27	0.67	0.04	-0.24	0.69	0.03	-0.24	0.72	0.03	-0.27	0.72	0.05	-0.22
MaxRet	(1)	1.06	0.02	-1.10	1.08	0.06	-0.96	0.94	0.05	-1.00	0.90	0.01	-1.05	1.37	0.12	-0.78
	(10)	0.99	0.40	0.23	1.40	0.27	-0.10	1.17	0.26	-0.01	1.27	0.29	0.05	1.47	0.18	-0.43
	(LS)	-0.08	0.38	1.32	0.33	0.21	0.86	0.23	0.21	0.99	0.38	0.27	1.10	0.10	0.06	0.35
Mom1m	(1)	0.65	-.01	1.01	0.81	0.03	-0.87	0.69	0.01	-0.96	0.79	0.02	-1.01	1.44	0.13	-0.75
	(10)	1.70	0.23	-0.09	1.80	0.24	-0.19	1.58	0.21	-0.12	1.62	0.22	-0.08	1.26	0.13	-0.30
	(LS)	1.05	0.23	0.92	0.99	0.21	0.69	0.89	0.20	0.84	0.84	0.19	0.93	-0.19	0.00	0.46
OScore	(1)	1.09	0.08	-0.59	1.39	0.22	-0.09	1.55	0.25	-0.14	1.27	0.11	-0.52	1.03	0.14	-0.24
	(10)	0.82	0.19	0.71	1.04	0.06	-0.62	1.17	0.10	-0.25	1.46	0.21	0.05	1.48	0.10	-0.86
	(LS)	-0.28	0.11	1.30	-0.35	-0.17	-0.53	-0.39	-0.14	-0.12	0.19	0.11	0.57	0.45	-0.04	-0.62
OperProf	(1)	1.07	0.17	0.13	1.02	0.16	0.18	1.04	0.15	0.09	1.29	0.16	-0.09	1.51	0.22	0.01
	(10)	1.20	0.34	0.73	1.17	0.23	0.38	1.17	0.25	0.39	1.18	0.31	0.66	1.16	0.15	-0.34
	(LS)	0.13	0.16	0.60	0.15	0.08	0.20	0.14	0.10	0.30	-0.11	0.15	0.75	-0.35	-0.07	-0.35
ShareVol	(1)	1.39	0.18	-0.19	1.81	0.33	-0.05	1.73	0.30	-0.04	1.54	0.23	-0.16	1.69	0.27	-0.22
	(10)	0.64	0.14	0.61	1.22	0.17	-0.03	1.26	0.16	0.04	1.42	0.19	-0.27	1.26	0.16	0.08
	(LS)	-0.75	-0.04	0.80	-0.59	-0.16	0.02	-0.34	-0.10	0.08	-0.11	-0.04	-0.10	-0.42	-0.11	0.30
dNoa	(1)	1.06	0.15	0.10	1.03	0.14	0.06	1.01	0.14	0.07	0.86	0.12	0.12	0.83	0.10	-0.15
	(10)	1.50	0.19	-0.28	1.38	0.18	-0.15	1.20	0.15	-0.19	1.54	0.18	-0.47	1.36	0.19	0.12
	(LS)	0.44	0.04	-0.39	0.35	0.04	-0.21	0.19	0.01	-0.26	0.68	0.06	-0.59	0.53	0.09	0.27

Table 5: Comparison of Long-Short (LS) Portfolio Returns Across Different Clustering Models and Traditional Analysis.

This table delineates the performance metrics of long-short portfolios constructed using traditional factors and cluster-neutral approaches across various econometric models. Each model is evaluated based on decile performance of selected financial factors, highlighting the shift from traditional to cluster-neutral analysis. For each factor, the table presents monthly average returns (R) in percentages, Annualized Sharpe Ratios (SR), and Fama-French 5-factor model alpha ($FF5^{\alpha}$) across the first decile (1), the tenth decile (10), and the long-short (LS) strategy, which involves buying the first decile and selling the tenth decile. The cluster-neutral models are differentiated by the most robust models, including K=10 for K-Means, DMax = 80 for Hierarchical Clustering, Epsilon ε = 4 and Minpts. M= 50 for DBSCAN, and OLS for an approach based on OLS regression.

4.2.2 *Cumulative Returns*

The examination of cumulative returns offers a dynamic perspective on the efficacy of investment strategies derived from factor analysis. Focusing on the actual implementation of going long in the highest decile and short in the lowest decile each month allows for visual tracking of strategy performance over time.

The graphical representation of these cumulative returns, is showcased in the subsequent figures (12).

The O-Score, which measures bankruptcy risk, displays varying trends between its traditional and cluster-neutral versions when analyzed through K-Means and DBSCAN algorithms during financial crises such as the dot.com bubble, the 2008 financial crisis, and the COVID-19 pandemic. Specifically, the traditional O-Score's cumulative returns decrease during these crises, while the cluster-neutral O-Score's returns do not follow this downward trend.

For MaxRet, indicating the maximum daily return over the last month, both the traditional and cluster-neutral forms exhibit similar overall trends. However, during financial crises, the cluster-neutral MaxRet shows more significant fluctuations, indicating a variation in response to market conditions compared to the traditional MaxRet.

In the DBSCAN model analysis, dNoa and InvestPPEInv demonstrate a notable increase in volatility, differing from their behavior in other models or when compared to standard factors. This observed volatility is specifically in the context of their returns, indicating a change in how these factors interact with market conditions when a cluster-neutral approach is applied.

In conclusion, the analysis reveals that cluster-neutral versions of financial factors, such as Oscore and MaxRet, provide more nuanced insights for investment strategies, especially during market volatility. However, it also shows that the cumulative returns of the traditional factor and its cluster-neutral counterpart in the long run converge back to each other, indicating that no additional returns can be gained by applying an LS strategy to the cluster-neutral factors.

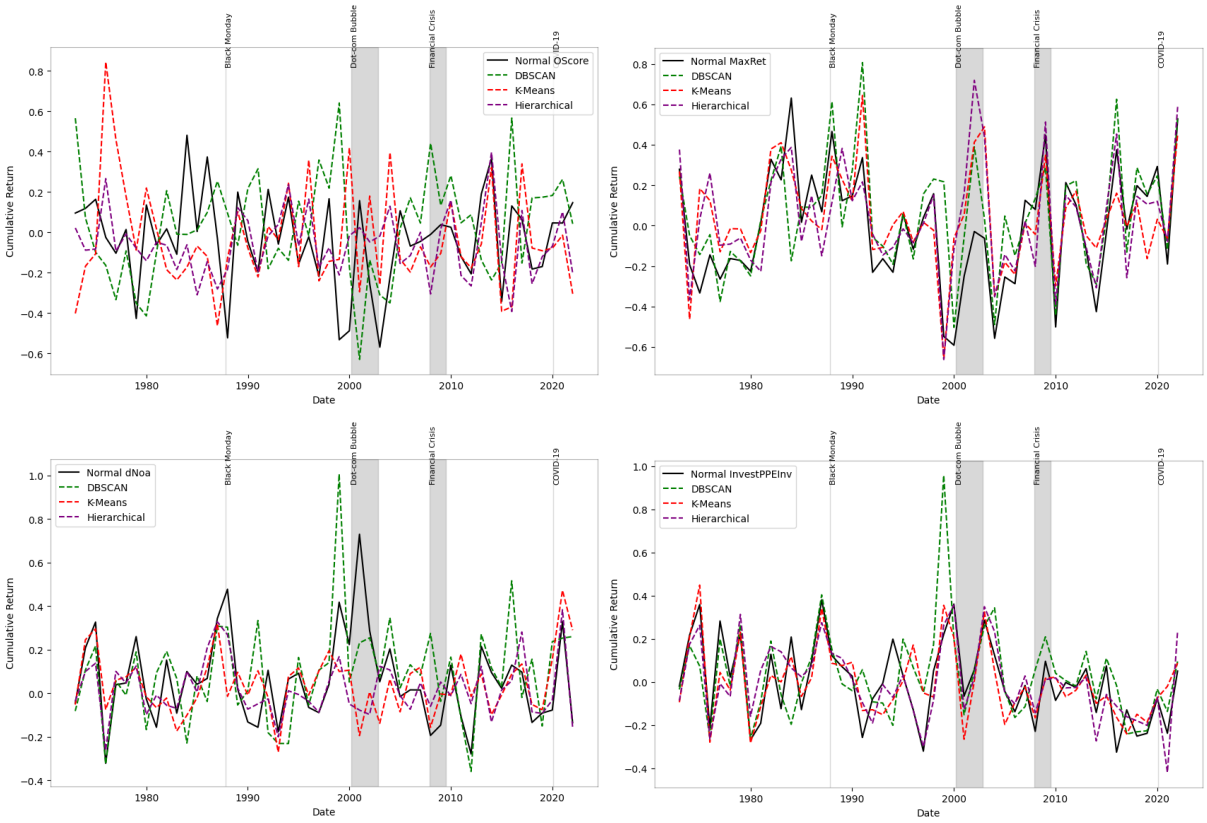

Figure 12: Cumulative Returns Over Time.

The examination of cumulative returns through graphical representations in offers a dynamic view on the efficacy of long-short (LS) investment strategies derived from factor analysis over time. These figures illustrate the trajectory and volatility of LS strategies across four distinct factors, starting at the top left: probability of bankruptcy (Oscore), top right: maximum daily return in the last months (Maxret), bottom left: growth in operating assets (dNoa), and lastly, bottom right: image investment in property, plant, and equipment (InvestPPEinv). The y-axis is cumulative returns in decimals and the x-asis shows the date. Moreover, market crashes are shown with a grey indicator.

4.3 *Cluster Fit And Robustness*

In this section I gauge the robustness of results based on the relative quality of cluster fits across different model configurations.

Table 6 starts with an evaluation of the simplest model, K-Means. Notably, its optimal performance for the stability metric, with a score of 0.71 at k=3, indicates that 71 percent of stocks change clusters monthly. This suggests that the model struggles with the noisy, high-dimensional nature of stock data. The Calinski-Harabasz and Davies-Bouldin indices, which measure cluster separation and compactness, respectively, show a deterioration in scores when the number of clusters exceeds 10. Based on these observations, along with additional robust results from portfolio sorting and panel regression analyses (Appendix C), a configuration of k=10 emerges as the most robust option.

K-Means								
Metric	**k=3**	**k=5**	**k=10**	**k=20**	**k=30**	**k=40**	**k=50**	**k=100**
Calinski-Harabasz	19.02	68.33	284.23	66.47	74.20	40.26	47.87	26.81
Davies-Bouldin	133.07	201.46	73.60	1068.94	4398.89	1884.52	2680.76	3914.85
Silhouette Score	−0.01	−0.04	−0.19	−0.12	−0.17	−0.18	−0.21	−0.28
Stability Metric	0.71	0.80	0.87	0.96	0.96	0.97	0.98	0.99
Clusters	3.00	5.00	10.00	20.00	30.00	40.00	50.00	100.00

This table displays the performance metrics of K-Means clustering models across various cluster sizes (k=3 to k=100). Metrics include the Calinski-Harabasz index (higher values indicate better-defined clusters), the Davies-Bouldin index (lower values suggest minimal intra-cluster dispersion and maximal inter-cluster separation), the Silhouette Score (ranging from -1 to 1, where closer to 1 signifies better clustering), and the Stability Metric (closer to 1 indicates more stable clustering). The number of clusters directly corresponds to the 'k' parameter in each model configuration, illustrating the model's adaptability and effectiveness across different levels of cluster granularity.

Table 6: Performance Metrics of K-Means Models with Different Cluster Sizes

For Hierarchical Clustering, an optimal distance metric (DMax) appears around 80 or higher in Table 7. The dendrogram of the stock data (Figure 13), also suggest this range. With a stability metric of 0.61, Hierarchical Clustering demonstrates more robust clustering over time, though it falls short of K-Means in terms of the Calinski-Harabasz and Davies-Bouldin scores.

Hierarchical Clustering								
Metric	**DMax=20**	**DMax=30**	**DMax=40**	**DMax=60**	**DMax=80**	**DMax=100**	**DMax=125**	**DMax=150**
Calinski-Harabasz	57.08	96.21	121.86	171.89	203.51	203.51	203.51	203.51
Davies-Bouldin	2747.01	1027.10	1193.02	560.08	236.23	236.23	236.23	236.23
Silhouette Score	−0.64	−0.40	−0.41	−0.29	−0.29	−0.29	−0.29	−0.29
Stability Metric	0.97	0.96	0.94	0.88	0.61	0.61	0.61	0.61
Clusters	64.00	38.00	24.00	13.00	7.00	7.00	7.00	7.00

This table outlines the performance metrics for Hierarchical Clustering models across a range of max distance (DMax=20 to DMax=150).

Table 7: Performance Metrics of Hierarchical Models with Different Cluster Sizes

.

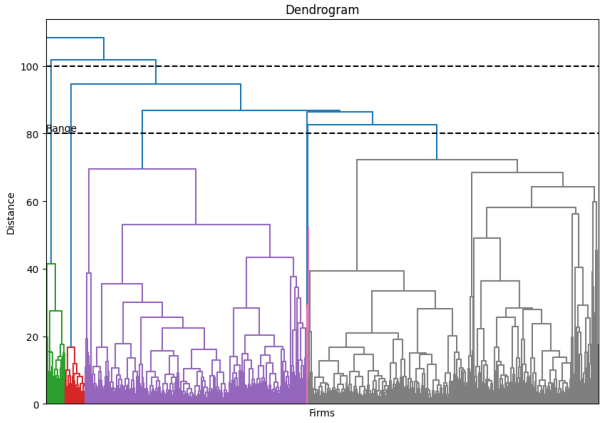

Figure 13: Dendrogram Showcasing Firm Clustering.

Dendrogram Showcasing Firm Clustering: Firms are aligned along the x-axis, with Hierarchical clustering distances on the y-axis, illustrating the natural data division and optimal DMax for clustering analysis. For the actual stock data.

DBSCAN's methodology, dependent on two key parameters (ε and MinPts), prompts a division of the analysis into three distinct tables, each catering to a specific ε value (2, 3, 4), with MinPts spanning from 1 to 2000. The elbow method, depicted in Graph 14, aids in determining a preliminary range for ε .

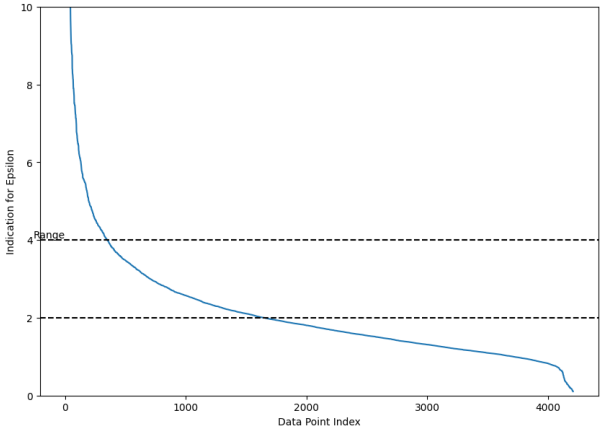

Figure 14: K-distance Plot.

Method to determine the optimal Epsilon for DBSCAN by identifying the "elbow" in the plot, where there is a sharp change in the k-distance to the nearest neighbors. This point indicates a suitable range for Epsilon, balancing between too many small clusters and too few large clusters.

Notably, an additional metric is introduced to account for outliers – a distinctive aspect of the DBSCAN algorithm. This metric represents the proportion of data deemed outliers and thus excluded from the mean subtraction process, a critical step in the cluster-neutral analysis. Given the volatile nature of stock data, a model marking up to 10 percent of data as outliers is considered acceptable. This balance between retaining significant data and leveraging DBSCAN's unique outlier detection is explored in-depth in Appendix C. Subsequent analysis reveals DBSCAN's superior performance across several metrics, especially the stability metric, which suggests that only about 10 percent of stocks transition between clusters on average.

Despite a modest Silhouette Score, which can be attributed to the model's sensitivity to diverse cluster shapes and sizes, DBSCAN's overall performance is impressive. The chosen configuration, with MinPts set at 50 and ε at 4, emerges as the most robust model, effectively balancing outlier identification with cluster stability and quality.

DBSCAN										
ε = 2	**M=1**	**M=3**	**M=5**	**M=10**	**M=25**	**M=50**	**M=100**	**M=500**	**M=1000**	**M=2000**
Calinski-Harabasz	6.93	226.86	523.52	1253.93	1147.07	1639.32	3340.30	4114.31	44.85	*
Davies-Bouldin	336 815.79	209.50	25.91	21.45	46.38	73.58	7.17	6.43	10.40	*
Silhouette Score	−0.99	−0.73	−0.64	−0.41	−0.36	−0.07	0.02	0.05	0.04	*
Stability Metric	0.64	0.16	0.12	0.15	0.15	0.14	0.14	0.15	0.04	*
Clusters	1495.00	19.00	7.50	4.00	4.00	3.00	2.00	2.00	2.00	1.00
Outliers	0.00	0.00	38.05	40.26	43.45	46.44	50.23	63.84	85.85	100
ε = 3	**M=1**	**M=3**	**M=5**	**M=10**	**M=25**	**M=50**	**M=100**	**M=500**	**M=1000**	**M=2000**
Calinski-Harabasz	8.90	307.87	545.43	641.86	1494.87	1395.48	1063.03	2196.35	2452.60	938.71
Davies-Bouldin	129 820.43	36.47	24.83	18.65	15.22	17.24	25.89	7.37	7.27	12.93
Silhouette Score	−0.84	−0.48	−0.60	−0.52	−0.16	−0.15	−0.09	−0.03	−0.02	0.02
Stability Metric	0.48	0.13	0.13	0.12	0.12	0.12	0.12	0.12	0.13	0.25
Clusters	602.00	12.00	7.00	6.00	3.00	3.00	3.00	2.00	2.00	2.00
Outliers	0.00	0.00	14.76	15.99	17.57	18.83	20.68	26.74	30.88	53.56
ε = 4	**M=1**	**M=3**	**M=5**	**M=10**	**M=25**	**M=50**	**M=100**	**M=500**	**M=1000**	**M=2000**
Calinski-Harabasz	12.24	268.09	488.58	743.92	718.08	1342.77	1107.67	1288.66	1290.98	1453.02
Davies-Bouldin	26 176.34	87.14	12.76	13.42	10.84	12.41	13.44	7.81	7.98	7.92
Silhouette Score	−0.88	−0.47	−0.59	−0.47	−0.51	−0.23	−0.20	−0.05	−0.05	−0.04
Stability Metric	0.44	0.10	0.10	0.10	0.10	0.09	0.09	0.10	0.10	0.11
Clusters	299.00	12.00	7.00	5.00	5.00	3.00	3.00	2.00	2.00	2.00
Outliers	0.00	0.00	6.47	7.08	8.12	9.55	10.72	14.14	15.04	17.98

Table 8: Performance Metrics of DBSCAN Models:

This table presents the performance metrics of DBSCAN clustering models across varying ε (Epsilon) values (2, 3, 4) and MinPts (M) ranging from 1 to 2000. Additionally, the table reports the percentage of Outliers detected. Asterisks (*) indicate unavailable or inapplicable data.

4.4 *Factor importance*

As the thesis methodology revolves around the firm level characteristic it is interesting to see which factors contributed the most to defining the clusters. To quantify the importance of various financial metrics in the clustering process, a correlation analysis is performed between these metrics and the cluster centroids generated by each method. Table 9 presents the results of this analysis.

Metric	DBSCAN	K-Means	Hierarchical
Accruals	0.06	0.01	0.04
AssetGrowth	0.21	0.03	0.09
BM	0.19	0.03	0.09
Beta	0.21	0.05	0.03
CompEquIss	0.10	0.02	0.06
DolVol	0.08	0.05	0.11
GP	0.13	0.02	0.05
IndMom	0.02	0.05	0.11
InvestPPEInv	0.15	0.03	0.12
Investment	0.02	0.00	0.01
MaxRet	0.26	0.04	0.07
Mom12m	0.11	0.02	0.04
Mom1m	0.06	0.01	0.04
Mom6m	0.10	0.02	0.05
OScore	0.60	0.08	0.26
OperProf	0.02	0.00	0.02
RET	0.06	0.02	0.04
RoE	0.01	0.00	0.01
ShareIss5Y	0.15	0.03	0.05
ShareVol	0.25	0.04	0.17
dNoa	0.13	0.03	0.08
marketCap	0.04	0.01	0.01
roaq	0.06	0.01	0.02

This table outlines the correlation coefficients between various financial metrics and the centroids of clusters generated by DBSCAN, K-Means, and Hierarchical clustering methods. The coefficients quantify the significance of each financial metric in influencing the formation of clusters by each method. A higher correlation value indicates a stronger relationship between the metric and the clustering configuration, suggesting the metric's prominent role in differentiating between clusters.

Table 9: Correlation between Financial Metrics and Clustering Method Cluster Centroids

Based on the data presented in Table 9, several conclusions about the importance of different financial metrics in relation to the clustering methods used can be made:

DOMINANCE OF CERTAIN METRICS IN DBSCAN The OScore and MaxRet metrics show notably high correlations in the DBSCAN method, with values of 0.60 and 0.26 respectively. This suggests that these metrics are particularly influential in the way DBSCAN clusters firms, indicating that factors related to profitability and maximum return play a significant role.

GENERAL LOW IMPACT IN K-MEANS The correlations for all metrics in the K-Means clustering are relatively low, with the highest being only 0.08 for OScore. This implies that K-Means clustering is less sensitive to variations in individual financial metrics, possibly suggesting a more holistic approach to clustering based on the overall financial profile of the firms.

SELECTIVE INFLUENCE IN HIERARCHICAL CLUSTERING In Hierarchical clustering, certain metrics like OScore (0.26) and ShareVol (0.17) show a moderate level of influence. This could

indicate that Hierarchical clustering is sensitive to specific financial dimensions, particularly those related to profitability and share volume.

OVERALL OBSERVATIONS Across all methods, OScore consistently shows a relatively high level of influence, emphasizing its importance as a financial metric in firm clustering. In contrast, metrics like Investment and RoE show consistently low correlations across all methods, suggesting they have limited influence in the clustering process.

METHOD-SPECIFIC INSIGHTS The differences in metric influences across the methods highlight their unique approaches. DBSCAN appears to be influenced more by profitability and return metrics, Hierarchical clustering shows a selective sensitivity to certain metrics, and K-Means demonstrates a more balanced, less metric-specific approach.

OVERFITTING The dominance of certain factors in the DBSCAN model may also suggest potential overfitting. Specifically, this model might rely too heavily on a limited number of factors

5 DISCUSSION

The primary objective of this thesis was to ascertain whether the behavior of cluster-neutral investment factors in the US stock market, developed through sophisticated clustering techniques, differs from that of traditional investment factors. To this end, I posed three related sub-research questions. The first two concerned evaluating the differential impact of these cluster-neutral factors on stock returns, using panel regression and a decile portfolio sort. The third question examined which of the three advanced clustering techniques (e.g., K-Means, Hierarchical Clustering, DBSCAN) most effectively generated these cluster-neutral factors.

5.1 *Results Discussion*

A key observation from the panel regression is the shift in significance of traditional factors like the probability of bankruptcy (O-score) and Market Capitalization (MarketCap), which become less predictive of stock returns after factor neutralization (across models) (contradicting findings from Fama and French, 1995 and Dichev, 1998). This highlights the context-dependence of these factors on proximal firms and their factors. Conversely, growth in net operating assets (dNoa) consistently emerges as a significant factor post-neutralization across all models, underscoring its potential as an independent predictor of stock returns.

The shift in the number of shares traded (ShareVol) across the K-Means and Hierarchical Clustering models, from negative to positive (-0.19*** to 0.03*), further implies a reevaluation of its intrinsic relationship with stock returns. Additionally, factors such as Book-to-Market (BM), Investment in Property, Plant, and Equipment (InvestPPEInv), Operating Profitability (OperProf), Short Term Momentum (Mom1m), and Industry Momentum (IndMom) underwent significant changes due to cluster neutralization, indicating their interdependency on other factors. Moreover, these results proved to be robust over time, with DBSCAN, in particular, finding more significant results in more recent history. This may suggest that the increasing

connectivity and nonlinear relationships across firms and their factors are intensifying over time, aligning with the findings of Gu et al., 2020.

The long-short (LS) portfolio performance analysis revealed nuanced effects of financial metrics across different econometric models, particularly in the DBSCAN model. Factors like Dollar Volume (DolVol) and Maximum Daily Return of the Last Month (MaxRet) shifted from marginal impacts in traditional models to significant positive returns in cluster-neutral configurations, suggesting their reevaluation under different analytical lenses.

The examination of cumulative returns over time added a dynamic perspective, especially during financial crises. The cluster-neutral versions of factors like OScore and MaxRet provided different insights for investment strategies, especially during market volatility.

In addition, the study's examination of clustering methods-K-Means, Hierarchical Clustering, and DBSCAN—was essential. K-Means exhibited moderate robustness with a stability metric of approximately 0.80, suggesting a reasonable level of consistency in cluster assignments. Hierarchical Clustering showed an increased stability over time compared to K-Means, with a stability metric of 0.61, indicating its effectiveness in capturing the evolving nature of financial data. In contrast, DBSCAN proved to be especially effective due to its outlier detection capability, achieving an average stability metric of 0.10. This means that, on average, only 10 percent of stocks switched clusters, highlighting DBSCAN's ability to identify and maintain coherent groupings even with the presence of outliers. The correlation analysis between financial metrics and cluster centroids revealed distinct patterns: OScore and MaxRet significantly influenced DBSCAN's clustering, indicating a focus on profitability and return metrics. K-Means suggested a more holistic approach, while Hierarchical Clustering displayed selective sensitivity to specific financial metrics.

The findings from this research contribute significantly to our understanding of the behavior of investment factors under cluster-neutral conditions. The study highlights the complexity and context-dependent nature of these factors, the nuanced impact of clustering techniques on factor behavior, and the importance of considering such methodologies in financial analysis. These insights not only inform investment strategies but also pave the way for further research in financial data clustering and factor analysis.

5.2 *Limitations*

This study encounters three primary limitations that merit consideration. Firstly, despite testing various configurations for each model and exploring multiple metrics for cluster fit, the scope for exploring alternative configurations remains vast. For instance, different distance metrics beyond the Euclidean distance could be considered, or alternative linkage metrics other than the Ward method for Hierarchical Clustering. These additional configurations could significantly expand the dimension of hyperparameters. However, due to practical constraints, I opted to select one configuration based on the literature review, recognizing that this choice limits the breadth of the analysis.

Secondly, a notable concern in this study is the high correlation observed between the cluster-neutralized factors and their traditional counterparts. This correlation likely contributed to multicollinearity in the panel regression, potentially inflating standard errors and even causing sign reversals in the regression coefficients. Although the results have been carefully

interpreted, mainly to ascertain the presence of an effect from factor neutralization rather than specifying its nature, the issue of multicollinearity remains pertinent.

lastly, the evidence suggested that the DBSCAN model might be overly reliant on two factors—MaxRet and OScore—for cluster determination, raising concerns about potential overfitting. However, the model's stable cluster fit metrics contradicted this, indicating robustness.

5.3 *Future Research*

The findings and limitations of this study open several avenues for future research, which can further expand our understanding of cluster-neutral investment factors and clustering techniques in financial analysis.

Future studies could delve into different configurations for clustering models, such as experimenting with non-Euclidean distance metrics or alternative linkage methods in Hierarchical Clustering. Given the multicollinearity observed between cluster-neutralized and traditional factors in this study, subsequent research could focus on developing methodologies to mitigate this issue. Advanced statistical techniques or machine learning algorithms could be employed to separate the effects more distinctly, enhancing the precision of factor analysis.

And given the different result across time investigating dynamic clustering approaches that evolve over time could provide valuable insights, especially in rapidly changing market conditions. This could involve real-time clustering or models that adapt to temporal shifts in market dynamics, offering a more agile and responsive analysis of factor behavior. Lastly the evidence of over fitting of the DBSCAN model highlights an area for improvement, suggesting a need to balance model complexity with generalizability to enhance its applicability without compromising on stability.

6 CONCLUSION

This thesis add to the field of Asset Pricing and Financial Economics, by answering the following following research questions:

> *How does the behavior of cluster-neutral investment factors, derived using advanced clustering techniques, diverge from traditional investment factors in financial markets, in relationship to stock returns?*

> **Sub-RQ1:** *In what ways do cluster-neutral investment factors, as identified by advanced clustering techniques, exhibit distinct behaviors from traditional factors in explaining stock returns when analyzed through panel regression models?*

The research demonstrated that cluster neutralizing cluster factors leads to notable, shifts in their impact on stock returns. Key factors such as the Book-to-Market Ratio (BM), Industry Momentum (IndMom), short-term momentum (Mom1m), the probability of bankruptcy (O-Score), Investment in Property, Plant, and Equipment (InvestPPEInv), growth in net operating assets (dNoa), and market capitalization, all tended towards a significant change in there relation ship to stock return. Remarkably, the share volume (ShareVol) underwent a sign change, transitioning from a negative to a positive effect on returns. However, certain predictors

like Beta, and longer term momentum factors were not affected, implying these are robust and strong predictors on their own. These findings were consistent across different models, various time periods, and multiple evaluation methods. This indicates that neutralizing for cluster effects can significantly modify the perceived influence of certain financial factors on stock returns, enhancing our understanding of their true impact in financial markets.

> **Sub-RQ2:** *How do the performance and characteristics of cluster-neutral investment factors, isolated using advanced clustering methods, differ from those of traditional factors when evaluated using a decile portfolio sorting approach?*

The key findings include the Dollar Trading Volume (DolVol) factor transitioning to positive Long-Short (LS) returns in cluster-neutral models. This indicates a sensitivity to clustering and inherent risks. The Maximum Daily Return of the Last Month (MaxRet) shifting to positive in cluster-neutral models, especially in DBSCAN, suggests its potential as a momentum factor. The improved performance of the OScore in DBSCAN aligns with risk-return trade-offs, while the significance of InvestPPEInv in DBSCAN points to industry-specific impacts. The complex performance of BM and the nuanced market dynamics indicated by the Composite Equity Issue (Compeqissu)'s results in K-Means highlight intricate market dynamics. The sensitivity of Share Volume (ShareVol) in DBSCAN and Operating Profitability (OperProf) in Ordinary Least Squares (OLS) models underscore the intricate interactions between volatility, returns, and profitability metrics. The minimal change in factors like the growth in net operating assets (dNOA), Short-term Momentum (Mom1m), and Investment highlights the subtle effects of cluster neutralization. Cumulative returns analysis, especially during market crises, demonstrates the stability and insight provided by cluster-neutral OScore and MaxRet, underscoring their strategic investment value

> **Sub-RQ3:** *Which advanced clustering technique (e.g., K-Means, Hierarchical Clustering, DBSCAN) most effectively distinguishes the behavioral patterns of cluster-neutral investment factors from those of traditional factors in financial market data?*

The comparative analysis of the clustering techniques revealed that each method has its strengths and limitations in isolating and highlighting unique behaviors of investment factors. DBSCAN emerged as a particularly robust model, effectively balancing outlier detection with cluster stability and quality. It showed superior performance in terms of stability metrics and the ability to capture intricate, possibly non-linear, interplays between factors and stock behavior.

In conclusion, this thesis marks a significant step forward in asset pricing research by demonstrating the effectiveness of advanced clustering techniques in unraveling the complex behaviors of investment factors. It paves the way for future investigations into more nuanced and data-driven approaches in financial market analysis.

REFERENCES

Ali, U., & Hirshleifer, D. (2020). Shared analyst coverage: Unifying momentum spillover effects. *Journal of Financial Economics*, *136*(3), 649–675.

Balakrishnan, K., Bartov, E., & Faurel, L. (2010). The predictive value of expenses excluded from 'pro forma' earnings. *Review of Accounting Studies*, *15*(2), 285–307.

Bali, T. G., Cakici, N., & Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, *99*(2), 427–446.

Brennan, M. J., Chordia, T., & Subrahmanyam, A. (1998). Trading volume and cross-autocorrelations in stock returns. *Journal of Finance*, *53*(2), 905–939.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, *52*(1), 57–82.

Chen, A. Y., & Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, *27*(2), 207–264.

Cohen, L., & Frazzini, A. (2008). Economic links and predictable returns. *The Journal of Finance*, *63*(4), 1977–2011.

Cooper, M. J., Gulen, H., & Schill, M. J. (2008). Average returns, bm, and share issuance. *Journal of Finance*, *63*(6), 2971–2995.

CRSP. (2023). Us stock database [Accessed: 2023-02-20].

Daniel, K., & Titman, S. (2006). Market pricing of accruals quality. *Journal of Accounting and Economics*, *42*(1-2), 3–33.

Datar, V., Naik, N. Y., & Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets*, *1*(2), 203–219.

Dichev, I. (1998). Is the risk of bankruptcy a systematic risk? *Journal of Finance*, *53*(3), 1131–1147.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, *96*(34), 226–231.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3–56.

Fama, E. F., & French, K. R. (1995). Size and book-to-market factors in earnings and returns. *The journal of finance*, *50*(1), 131–155.

Fama, E. F., & French, K. R. (2006). Profitability, investment and average returns. *Journal of Financial Economics*, *82*(3), 491–518.

Fama, E. F., & French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of financial economics*, *105*(3), 457–472.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1–22.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607–636.

French, K. R. (2023). Data library [Accessed: 2023-02-20].

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *science*, *315*(5814), 972–976.

Green, J., Hand, J. R., & Zhang, X. F. (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, *30*(12), 4389–4436.

Grinblatt, M., & Moskowitz, T. J. (1999). Do industries explain momentum? *Journal of Finance*, *54*(4), 1249–1290.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Haugen, R. A., & Baker, N. L. (1996a). Commonality in the determinants of expected stock returns. *Journal of financial economics*, *41*(3), 401–439.

Haugen, R. A., & Baker, N. L. (1996b). The expected return on stocks and bonds. *Financial Analysts Journal*, *52*(1), 75–80.

Hirshleifer, D., Hou, K., Teoh, S. H., & Zhang, Y. (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, *38*(1), 297–331.

Jagannathan, R., & Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The journal of finance*, *58*(4), 1651–1683.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, *48*(1), 65–91.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, *15*(4), 143–156.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, *10*(5), 603–621.

Lee, C., Ma, P., & Wang, C. C. (2016). The search for peer firms: When do crowds provide wisdom? *Harvard Business School Accounting & Management Unit Working Paper*, (15-032), 14–46.

Lewellen, J. (2014). The cross section of expected stock returns. *Forthcoming in Critical Finance Review, Tuck School of Business Working Paper*, (2511246).

Lyandres, E., Sun, L., & Zhang, L. (2008). The new issues puzzle: Testing the investment-based explanation. *Review of Financial Studies*, *21*(6), 2825–2855.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*(14), 281–297.

Maillard, S., Roncalli, T., & Teıletche, J. (2010). The properties of equally weighted risk contribution portfolios. *The Journal of Portfolio Management*, *36*(4), 60–70.

Menzly, L., & Ozbas, O. (2010). Market segmentation and cross-predictability of returns. *The Journal of Finance*, *65*(4), 1555–1580.

Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of finance*, *54*(4), 1249–1290.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, *108*(1), 1–28.

Reynolds, D. A., et al. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*(659-663).

Seyfi, S. (2022). Neighbouring assets. *Available at SSRN 4311284*.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425–442.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, *71*(3), 289–315.

Stambaugh, R. F., Yu, J., & Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of financial economics*, *104*(2), 288–302.

Stattman, D. (1980). The relationship between return and market value of common stocks. *Journal of Business*, 97–112.

Titman, S., Wei, K. C. J., & Xie, F. (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*, *39*(4), 677–700.

Uddin, A., Tao, X., & Yu, D. (2023). Attention based dynamic graph neural network for asset pricing. *Global Finance Journal*, *58*, 100900.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*, 395–416.

A APPENDIX A: FIRM LEVEL CHARACTERISTICS

No.	Name	Description	Author(s)
1.	Accruals	Ratio of Net Operating Assets Change to Average Total Assets	Sloan, 1996
2.	Asset Growth	Yearly Growth Rate of Total Assets	Cooper et al., 2008
3.	BM	Logarithm of the Ratio of Book Equity to Market Equity	Stattman, 1980
4.	Beta	Measure of Stock Volatility Based on a 60-Month Rolling Window Regression Against Market Returns	Fama and MacBeth, 1973
5.	CompEquIss	5-Year Growth in Market Value of Equity Adjusted by Stock Returns	Daniel and Titman, 2006
6.	DolVol	Logarithm of the Product of Lagged Volume and Stock Price	Brennan et al., 1998
7.	GP	Ratio of Gross Profit to Total Assets (Excluding Financial Firms)	Novy-Marx, 2013
8.	IndMom	Industry-Weighted 6-Month Average Stock Return	Grinblatt and Moskowitz, 1999
9.	InvestPPEInv	Change in Property, Plant, Equipment, and Inventory Scaled by Lagged Total Assets	Lyandres et al., 2008
10.	Investment	Capital Expenditure as a Ratio of Revenue, Normalized by the Firm's 36-Month Rolling Average	Titman et al., 2004
11.	MaxRet	Highest Daily Stock Return in the Previous Month	Bali et al., 2011
12.	Mom12m	12-Month Stock Return (Excluding the Most Recent Month)	Jegadeesh and Titman, 1993
13.	Mom1m	Stock Return in the Previous Month	Jegadeesh and Titman, 1993
14.	Mom6m	6-Month Stock Return (Excluding the Most Recent Month)	Jegadeesh and Titman, 1993
15.	OScore	A Composite Score Based on Various Financial Ratios and Indicators	Dichev, 1998
16.	OperProf	Operating Profitability (Adjusted for Size)	Fama and French, 2006
17.	RoE	Return on Equity (Excluding Firms with Low Stock Price)	Haugen and Baker, 1996b
18.	ShareIss5Y	5-Year Change in the Number of Shares Outstanding (Adjusted for Splits)	Daniel and Titman, 2006
19.	ShareVol	Binary Indicator of Share Trading Volume Over the Past Three Months	Datar et al., 1998
20.	*dNoa*	Growth in Net Operating Assets Over 12 Months (Scaled by Lagged Total Assets)	Hirshleifer et al., 2004
21.	Market Cap	Market Capitalization	Fama and French, 1993
22.	RoA	Return on Assets	Balakrishnan et al., 2010

Table 10: Appendix A: List of Firm-Specific Characteristics.

This table enumerates 22 key financial factors used in this thesis, they are widely referenced in asset pricing and stock return analysis. Each factor is accompanied by its definition, the original author(s), and the year of publication. These characteristics are integral to various financial models and theories, including the Fama and French factors, momentum factors, Q-factors, and the mispricing model, reflecting their prominence in financial literature. (Source: Chen and Zimmermann, 2022)

B DETAILED MATHEMATICAL UNDERPINNINGS

B.1 *Mathematical Formulation of Cluster Fit Metrics*

SILHOUETTE SCORE For a stock i at time t, the Silhouette Score ($S_{i,t}$) is defined as:

$$S_{i,t} = \frac{b_{i,t} - a_{i,t}}{\max(a_{i,t}, b_{i,t})}$$

where:

- $a_{i,t}$ is the mean distance between stock i at time t and all other stocks in the same cluster.
- $b_{i,t}$ is the smallest mean distance from stock i at time t to stocks in any other cluster.

DAVIES-BOULDIN INDEX The Davies-Bouldin Index (DB) is computed as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left(\frac{\sigma_{C_i,t} + \sigma_{C_j,t}}{d(c_{i,t}, c_{j,t})} \right)$$

where $\sigma_{C_i,t}$ is the average distance of all stocks in cluster C_i at time t to the centroid of C_i, and $d(c_{i,t}, c_{j,t})$ is the distance between centroids of clusters C_i and C_j at time t.

CALINSKI-HARABASZ INDEX The Calinski-Harabasz Index (CH) is given by:

$$CH = \frac{\mathrm{Tr}(B_{k,t})/(k-1)}{\mathrm{Tr}(W_{k,t})/(N-k)}$$

where $B_{k,t}$ and $W_{k,t}$ are the between-group and within-cluster dispersion matrices at time t, respectively.

STABILITY METRIC The stability of clusters over time is quantified as the average number of changes in cluster assignments for each stock i. It is calculated as:

$$\text{Stability} = \frac{1}{N} \sum_{i=1}^{N} \text{Changes}_{i,t}$$

where $\text{Changes}_{i,t}$ represents the number of times the cluster assignment for stock i changes at time t.

CORRELATION WITH CLUSTER LABELS The correlation between each factor and its cluster label at time t is calculated using the Pearson correlation coefficient:

$$\rho_{X_t, Y_t} = \frac{\mathrm{cov}(X_t, Y_t)}{\sigma_{X_t} \sigma_{Y_t}}$$

where X_t and Y_t represent the factor and cluster label at time t, respectively.

B.2 *Mathematical Formulation of K-Means Clustering*

ITERATIVE ALGORITHM K-Means clustering is an iterative algorithm designed to partition a dataset into k distinct clusters. The algorithm minimizes the within-cluster sum of squares (WCSS), which is the sum of squared distances between each data point and the centroid of its cluster.

The distance between two points in Euclidean space is the length of the line segment connecting them. In Cartesian coordinates, if $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ are two points in Euclidean n-space, the Euclidean distance (d) between p and q is given by:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{9}$$

Algorithm 1 K-Means Clustering

Require: Data set $X = \{x_1, x_2, \ldots, x_N\}$, Number of clusters k
Ensure: A set of k clusters that minimizes the WCSS
1: Initialize k centroids $\mu_1, \mu_2, \ldots, \mu_k$ randomly from the data points.
2: **repeat**
3: **Assignment Step:**
4: **for** each point x_i in X **do**
5: Find the nearest centroid μ_j using Euclidean distance.
6: Assign x_i to cluster j.
7: **end for**
8: **Update Step:**
9: **for** each cluster $j = 1$ to k **do**
10: Compute new centroid μ_j as the mean of all points in j.
11: **end for**
12: **until** centroids do not significantly change
13: **Return** clusters and their centroids.

B.3 *Mathematical Formulation of Hierarchical Clustering*

This appendix section delves into the mathematical underpinnings of Hierarchical Clustering, discussing various distance metrics and linkage methods, and explicates the rationale behind the specific choice of Euclidean distance and Ward's method for analyzing stock data.

DISTANCE METRICS IN HIERARCHICAL CLUSTERING In Hierarchical Clustering, the choice of distance metric is pivotal in defining the similarity between data points. The most common metrics include:

- **Euclidean Distance:** Defined as $d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$, it measures the 'straight-line' distance between two points in Euclidean space.
- **Manhattan Distance:** Also known as city block distance, it is defined as $d(p,q) = \sum_{i=1}^{n}|p_i - q_i|$, measuring the distance between two points along axes at right angles.

- **Mahalanobis Distance:** A multivariate distance metric that accounts for correlations between variables, defined as $d(p,q) = \sqrt{(p-q)^T S^{-1}(p-q)}$, where S is the covariance matrix.
- **Cosine Similarity:** Often used in text analysis, it measures cosine of the angle between two vectors.

LINKAGE METHODS IN HIERARCHICAL CLUSTERING The linkage criterion determines the method used to calculate the distance between clusters:

- **Single Linkage:** The shortest distance between any two points in the clusters.
- **Complete Linkage:** The longest distance between any two points in the clusters.
- **Average Linkage:** The average distance between each point in one cluster to every point in the other cluster.
- **Ward's Method:** This method minimizes the increase in total within-cluster variance after merging.

MATHEMATICAL FORMULATION OF WARD'S METHOD Ward's method aims to minimize the total within-cluster variance. The increase in variance when merging clusters U and V is given by:

$$\Delta(\text{Var}) = \text{Var}(U \cup V) - (\text{Var}(U) + \text{Var}(V))$$

RATIONALE FOR EUCLIDEAN DISTANCE AND WARD'S METHOD In the context of financial data analysis:

- **Euclidean Distance** is effective in capturing absolute differences in quantitative features of stocks, assuming these features are commensurable.
- **Ward's Method** tends to create clusters of approximately equal size, which is advantageous when dealing with stocks that vary in scale and characteristics. It is particularly sensitive to subtle similarities within financial data.

COMBINING EUCLIDEAN DISTANCE WITH WARD'S METHOD The combination is particularly effective:

1. **Distance Calculation:** Euclidean distances between all pairs of stocks are computed.
2. **Cluster Formation:** Iteratively merge clusters based on Ward's criterion, focused on minimizing the increase in total within-cluster variance.
3. **Resulting Structure:** The resultant dendrogram from this approach reflects a nuanced and insightful view of the Hierarchical structure within the financial data.

This mathematical exposition and rationale offer insights into why Euclidean distance and Ward's method are particularly suited for clustering stock data, revealing complex relationships in financial datasets.

B.4 *Mathematical Formulation of DBSCAN*

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that identifies clusters as high-density regions separated by regions of low density. Its effectiveness in dealing with varying densities and noise makes it particularly suitable for complex financial datasets.

CORE CONCEPTS DBSCAN's operation is based on two key parameters:

- ε (Epsilon): A distance threshold that defines the neighborhood around a data point.
- MinPts (Minimum Points): The minimum number of points required to form a dense region.

DEFINITIONS

- **Core Point:** A point is a core point if at least MinPts points are within distance ε of it (including the point itself).
- **Border Point:** A point is a border point if it is not a core point but is in the neighborhood of a core point.
- **Noise:** A point is noise if it is neither a core point nor a border point.

MATHEMATICAL FORMULATION The mathematical process of DBSCAN can be outlined as follows:

1. **Identifying Core Points:** For each point x_i in the dataset, count the number of points within ε distance of x_i (including x_i itself). If this count is at least MinPts, label x_i as a core point.

$$\text{Core Point if } |N_\varepsilon(x_i)| \geq \text{MinPts} \tag{10}$$

   where $N_\varepsilon(x_i)$ is the ε-neighborhood of x_i.

2. **Forming Clusters:** A cluster is formed around a core point by including all points (core or border) within its ε-neighborhood. This process is recursively applied to all points in the newly formed cluster.
3. **Cluster Expansion:** Expand each cluster by recursively adding all density-reachable points to the cluster. A point y is density-reachable from x if there is a chain of points $x_1, x_2, ..., x_n$, where $x_1 = x$ and $x_n = y$, such that each x_{i+1} is within distance ε from x_i and x_i is a core point.
4. **Handling Noise:** Any point not included in any cluster is labeled as noise.

C APPENDIX C: ROBUSTNESS

C.1 *Panel Regression*

C.1.1 *K-Means*

Table 11: Regression Results per K, for K-Means, to see the difference in results for increasing the cluster amount.

Regressed on Excess Returns At t+1		Cluster Neutral Factors			
Variable	Normal	k=3	k=5	k=10	k=100
const	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***
Accruals	0.06 (0.02)***	0.07 (0.02)***	0.07 (0.02)***	0.08 (0.02)***	0.08 (0.02)***
AssetGrowth	0.20 (0.03)***	0.21 (0.03)***	0.19 (0.02)***	0.16 (0.02)***	0.21 (0.03)***
BM	0.80 (0.07)***	0.77 (0.06)***	0.76 (0.06)***	0.70 (0.06)***	0.65 (0.05)***
Beta	-0.15 (0.07)**	-0.19 (0.07)**	-0.16 (0.07)**	-0.16 (0.06)**	-0.16 (0.05)**
CompEquIss	0.12 (0.03)***	0.08 (0.03)***	0.07 (0.02)***	0.05 (0.02)**	-0.05 (0.02)*
DolVol	1.56 (0.13)***	1.36 (0.10)***	1.18 (0.08)***	0.99 (0.06)***	0.79 (0.04)***
GP	0.21 (0.04)***	0.16 (0.04)***	0.16 (0.04)***	0.17 (0.04)***	0.11 (0.03)***
IndMom	0.26 (0.05)***	0.21 (0.04)***	0.21 (0.04)***	0.20 (0.04)***	0.19 (0.03)***
InvestPPEInv	-0.0016 (0.03)	0.04 (0.02)	0.04 (0.02)*	0.07 (0.02)***	0.09 (0.02)***
Investment	0.07 (0.01)***	0.07 (0.01)***	0.08 (0.01)***	0.08 (0.01)***	0.06 (0.01)***
MaxRet	-0.02 (0.06)	-0.0084 (0.05)	-0.009 (0.05)	-0.01 (0.04)	0.13 (0.03)***
Mom12m	0.36 (0.05)***	0.30 (0.05)***	0.25 (0.05)***	0.23 (0.04)***	0.10 (0.03)***
Mom6m	-0.15 (0.06)**	-0.17 (0.05)***	-0.18 (0.05)***	-0.15 (0.04)***	-0.03 (0.04)
Mom1m	-0.76 (0.09)***	-0.77 (0.07)***	-0.77 (0.07)***	-0.78 (0.06)***	-0.78 (0.05)***
OScore	-0.10 (0.02)***	-0.10 (0.02)***	-0.04 (0.03)	-0.01 (0.02)	-0.02 (0.02)
OperProf	0.07 (0.01)***	0.06 (0.01)***	0.05 (0.01)***	0.03 (0.01)**	0.03 (0.01)**
RoE	-0.08 (0.02)***	-0.07 (0.02)***	-0.08 (0.02)***	-0.06 (0.02)***	-0.05 (0.02)**
ShareIss5Y	0.03 (0.02)	0.04 (0.02)*	0.08 (0.02)***	0.12 (0.02)***	0.13 (0.02)***
ShareVol	-0.18 (0.02)***	-0.13 (0.02)***	-0.05 (0.02)**	0.03 (0.02)*	0.04 (0.02)**
dNoa	0.04 (0.03)	0.07 (0.03)**	0.09 (0.02)***	0.07 (0.02)***	0.12 (0.01)***
MarketCap	0.05 (0.02)**	0.06 (0.02)***	0.05 (0.02)**	0.02 (0.02)	0.01 (0.02)
roaq	0.16 (0.03)***	0.15 (0.03)***	0.14 (0.02)***	0.14 (0.02)***	0.14 (0.02)***
Time Fixed Effects	X	X	X	X	X

This table presents regression results per K, for K-Means clustering. Each column represents a different number of clusters used in the analysis, with 'Normal' referring to the baseline model without clustering. The coefficients and standard errors (in parentheses) are multiplied by 100 for clearer presentation. Significance levels are denoted as follows: '***' for $p < 0.01$, '**' for $p < 0.05$, and '*' for $p < 0.1$. This analysis aims to illustrate how clustering affects the model's coefficients, under various amounts of clusters. Time-fixed effects are included in all models. Moreover, all factors are standardized through cross-sectional demeaning and unit variance scaling, meaning that an increase of one standard deviation in a coefficient results in a corresponding increase in the dependent variable by the coefficient's amount, which is expressed in percentages. Lastly, the panel regression uses clustered standard errors.

C.1.2 *Hierarchical Clustering*

Table 12: Regression Results per maximum distance (DMax), for Hierarchical Clustering, to see the difference in results for increasing the DMax amount

Regressed on Excess Returns At t+1		Cluster Neutral Factors			
Variable	Normal	DMax=20	DMax=40	DMax=60	DMax=80
const	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***
Accruals	0.06 (0.02)***	0.08 (0.02)***	0.08 (0.02)***	0.08 (0.02)***	0.06 (0.02)***
AssetGrowth	0.20 (0.03)***	0.21 (0.03)***	0.19 (0.03)***	0.16 (0.02)***	0.19 (0.03)***
BM	0.80 (0.07)***	0.66 (0.06)***	0.69 (0.06)***	0.70 (0.06)***	0.74 (0.06)***
Beta	-0.15 (0.07)**	-0.16 (0.05)**	-0.16 (0.06)**	-0.16 (0.06)**	-0.18 (0.06)**
CompEquIss	0.12 (0.03)***	-0.03 (0.02)	-0.0025 (0.02)	0.05 (0.02)*	0.07 (0.03)**
DolVol	1.56 (0.13)***	0.82 (0.10)***	0.89 (0.10)***	0.99 (0.10)***	1.12 (0.10)***
GP	0.21 (0.04)***	0.15 (0.04)***	0.17 (0.04)***	0.17 (0.04)***	0.20 (0.04)***
IndMom	0.26 (0.05)***	0.19 (0.04)***	0.20 (0.04)***	0.20 (0.04)***	0.23 (0.04)***
InvestPPEInv	-0.0016 (0.03)	0.09 (0.02)***	0.09 (0.02)***	0.07 (0.02)***	0.05 (0.02)**
Investment	0.07 (0.01)***	0.05 (0.01)***	0.03 (0.01)**	0.08 (0.01)***	0.06 (0.01)***
MaxRet	-0.02 (0.06)	0.10 (0.03)***	0.04 (0.04)	-0.01 (0.04)	-0.03 (0.05)
Mom12m	0.36 (0.05)***	0.13 (0.04)***	0.18 (0.04)***	0.23 (0.04)***	0.26 (0.05)***
Mom6m	-0.15 (0.06)**	-0.05 (0.04)	-0.08 (0.04)*	-0.15 (0.04)***	-0.13 (0.05)**
Mom1m	-0.76 (0.09)***	-0.77 (0.05)***	-0.75 (0.06)***	-0.78 (0.06)***	-0.77 (0.07)***
OScore	-0.10 (0.02)***	-0.03 (0.02)	-0.03 (0.02)	-0.01 (0.02)	-0.03 (0.03)
OperProf	0.07 (0.01)***	0.03 (0.01)**	0.03 (0.01)**	0.03 (0.01)**	0.03 (0.01)**
RoE	-0.08 (0.02)***	-0.05 (0.02)**	-0.04 (0.02)**	-0.06 (0.02)***	-0.07 (0.02)***
ShareIss5Y	0.03 (0.02)	0.14 (0.02)***	0.14 (0.02)***	0.12 (0.02)***	0.10 (0.02)***
ShareVol	-0.18 (0.02)***	0.05 (0.02)**	0.03 (0.02)	0.03 (0.02)*	0.02 (0.02)
dNoa	0.04 (0.03)	0.12 (0.02)***	0.10 (0.02)***	0.07 (0.02)***	0.09 (0.02)***
MarketCap	0.05 (0.02)**	-0.0044 (0.02)	-0.0051 (0.02)	0.02 (0.02)	0.01 (0.02)
roaq	0.16 (0.03)***	0.13 (0.02)***	0.13 (0.02)***	0.14 (0.02)***	0.14 (0.03)***
Time Fixed Effects	X	X	X	X	X

This table presents regression results per maximum distance (DMax), for Hierichical Clustering. Each column represents a different value for DMax used in the analysis, with 'Normal' referring to the baseline model without clustering. The coefficients and standard errors (in parentheses) are multiplied by 100 for clearer presentation. Significance levels are denoted as follows: '***' for $p < 0.01$, '**' for $p < 0.05$, and '*' for $p < 0.1$. This analysis aims to illustrate how clustering affects the model's coefficients, under various amounts of DMax. Time-fixed effects are included in all models. Moreover, all factors are standardized through cross-sectional demeaning and unit variance scaling, meaning that an increase of one standard deviation in a coefficient results in a corresponding increase in the dependent variable by the coefficient's amount, which is expressed in percentages. Lastly, the panel regression uses clustered standard errors.

Table 13: Regression Results For DBSCAB, to see the difference in results for increasing the cluster amount ε (Epsilon) and MinPts (M)

Regressed on Excess Returns At t+1		Cluster Neutral Factors			
Variable	Normal	ε = 3 , M= 3	ε = 4 , M= 5	ε = 4 , M= 25	ε = 4 , M= 50
const	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***
Accruals	0.06 (0.02)***	0.06 (0.02)***	0.05 (0.02)**	0.08 (0.02)***	0.05 (0.02)**
AssetGrowth	0.20 (0.03)***	0.21 (0.03)***	0.21 (0.03)***	0.16 (0.02)***	0.21 (0.03)***
BM	0.80 (0.07)***	0.75 (0.07)***	0.75 (0.07)***	0.70 (0.06)***	0.75 (0.07)***
Beta	-0.15 (0.07)**	-0.15 (0.06)**	-0.14 (0.06)*	-0.16 (0.06)**	-0.14 (0.06)*
CompEquIss	0.12 (0.03)***	0.10 (0.03)***	0.10 (0.03)***	0.05 (0.02)**	0.10 (0.03)***
DolVol	1.56 (0.13)***	1.44 (0.11)***	1.47 (0.12)***	0.99 (0.06)***	1.50 (0.12)***
GP	0.21 (0.04)***	0.20 (0.04)***	0.20 (0.04)***	0.17 (0.04)***	0.19 (0.04)***
IndMom	0.26 (0.05)***	0.25 (0.05)***	0.26 (0.05)***	0.20 (0.04)***	0.25 (0.05)***
InvestPPEInv	-0.0016 (0.03)	-0.0029 (0.02)	-0.0041 (0.03)	0.07 (0.02)***	-0.0020 (0.03)
Investment	0.07 (0.01)***	0.08 (0.01)***	0.08 (0.01)***	0.08 (0.01)***	0.08 (0.01)***
MaxRet	-0.02 (0.06)	-0.01 (0.05)	-0.01 (0.06)	-0.01 (0.04)	-0.01 (0.06)
Mom12m	0.36 (0.05)***	0.32 (0.05)***	0.32 (0.05)***	0.23 (0.04)***	0.32 (0.05)***
Mom6m	-0.15 (0.06)**	-0.15 (0.06)**	-0.16 (0.06)**	-0.15 (0.04)***	-0.16 (0.06)**
Mom1m	-0.76 (0.09)***	-0.77 (0.08)***	-0.77 (0.08)***	-0.78 (0.06)***	-0.78 (0.08)***
OScore	-0.10 (0.02)***	-0.07 (0.02)***	-0.08 (0.03)**	-0.01 (0.02)	-0.01 (0.03)
OperProf	0.07 (0.01)***	0.07 (0.01)***	0.07 (0.01)***	0.03 (0.01)**	0.07 (0.01)***
RoE	-0.08 (0.02)***	-0.08 (0.02)***	-0.08 (0.02)***	-0.06 (0.02)***	-0.08 (0.02)***
ShareIss5Y	0.03 (0.02)	0.06 (0.02)***	0.07 (0.02)***	0.12 (0.02)***	0.07 (0.02)***
ShareVol	-0.18 (0.02)***	-0.04 (0.02)*	-0.08 (0.02)***	0.03 (0.02)*	-0.08 (0.02)***
dNoa	0.04 (0.03)	0.05 (0.03)*	0.05 (0.03)*	0.07 (0.02)***	0.05 (0.03)*
MarketCap	0.05 (0.02)**	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)
roaq	0.16 (0.03)***	0.15 (0.03)***	0.16 (0.03)***	0.14 (0.02)***	0.15 (0.03)***
Time Fixed Effects	X	X	X	X	X

This table presents regression results per ε (Epsilon) values and MinPts (M), for DBSCAN. Each column represents a different configuration, which showed to be the most robust in the in section 4.3 (outliers < 0.10), with 'Normal' referring to the baseline model without clustering. The coefficients and standard errors (in parentheses) are multiplied by 100 for clearer presentation. Significance levels are denoted as follows: '***' for $p < 0.01$, '**' for $p < 0.05$, and '*' for $p < 0.1$. Moreover, all factors are standardized through cross-sectional demeaning and unit variance scaling, meaning that an increase of one standard deviation in a coefficient results in a corresponding increase in the dependent variable by the coefficient's amount, which is expressed in percentages. Lastly, the panel regression uses clustered standard errors.

Table 14: Regression Results For The Most Robust Models.)

Regressed on Excess Returns At t+1		Cluster Neutral Factors			
Variable	Normal	K = 10	DMax = 80	ε = 4 , M= 50	OLS
const	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***	0.74 (0.24)***
Accruals	0.06 (0.02)***	0.08 (0.02)***	0.06 (0.02)***	0.05 (0.02)**	0.13 (0.02)***
AssetGrowth	0.20 (0.03)***	0.16 (0.02)***	0.19 (0.02)***	0.21 (0.03)***	0.57 (0.03)***
BM	0.80 (0.07)***	0.70 (0.06)***	0.74 (0.06)***	0.75 (0.07)***	1.17 (0.07)***
Beta	-0.15 (0.07)**	-0.16 (0.06)**	-0.18 (0.06)**	-0.14 (0.06)*	-0.31 (0.06)***
CompEquIss	0.12 (0.03)***	0.05 (0.02)*	0.07 (0.03)**	0.10 (0.03)***	-0.05 (0.02)*
DolVol	1.56 (0.13)***	0.99 (0.06)***	1.12 (0.07)***	1.50 (0.12)***	1.28 (0.08)***
GP	0.21 (0.04)***	0.17 (0.04)***	0.20 (0.04)***	0.19 (0.04)***	0.10 (0.04)**
IndMom	0.26 (0.05)***	0.20 (0.04)***	0.23 (0.04)***	0.25 (0.05)***	0.14 (0.04)***
InvestPPEInv	-0.0016 (0.03)	0.07 (0.02)***	0.05 (0.02)*	-0.002 (0.03)	0.36 (0.03)***
Investment	0.07 (0.01)***	0.08 (0.01)***	0.06 (0.01)***	0.08 (0.01)***	0.09 (0.01)***
MaxRet	-0.02 (0.06)	-0.01 (0.04)	-0.03 (0.05)	-0.01 (0.06)	0.26 (0.04)***
Mom12m	0.36 (0.05)***	0.23 (0.04)***	0.26 (0.05)***	0.32 (0.05)***	-0.10 (0.06)*
Mom6m	-0.15 (0.06)**	-0.15 (0.04)***	-0.13 (0.05)**	-0.16 (0.06)**	-0.14 (0.06)**
Mom1m	-0.76 (0.09)***	-0.78 (0.06)***	-0.77 (0.07)***	-0.78 (0.08)***	-0.22 (0.04)***
OScore	-0.10 (0.02)***	-0.01 (0.02)	-0.03 (0.03)	-0.01 (0.02)	0.02 (0.03)
OperProf	0.07 (0.01)***	0.03 (0.01)**	0.03 (0.01)**	0.07 (0.01)***	-0.01 (0.01)
RoE	-0.08 (0.02)***	-0.06 (0.02)***	-0.07 (0.02)***	-0.08 (0.02)***	-0.02 (0.02)
ShareIss5Y	0.03 (0.02)*	0.12 (0.02)***	0.10 (0.02)***	0.07 (0.02)***	0.23 (0.02)***
ShareVol	-0.18 (0.02)***	0.03 (0.02)*	0.02 (0.02)	-0.08 (0.02)***	0.14 (0.02)***
dNoa	0.04 (0.03)	0.07 (0.02)***	0.09 (0.02)***	0.05 (0.03)*	0.50 (0.03)***
MarketCap	0.05 (0.02)**	0.02 (0.02)	0.01 (0.02)	0.03 (0.02)	-0.14 (0.02)***
roaq	0.16 (0.03)***	0.14 (0.02)***	0.14 (0.03)***	0.15 (0.03)***	0.07 (0.03)**
Time Fixed Effects	X	X	X	X	X

This table presents regression results for the most robust models based on the cluster fit in section 4.3 together with the OLS model. 'Normal' refers to the baseline model without clustering. The coefficients and standard errors (in parentheses) are multiplied by 100 for clearer presentation. Significance levels are denoted as follows: '***' for $p < 0.01$, '**' for $p < 0.05$, and '*' for $p < 0.1$. Time-fixed effects are included in all models. Moreover, all factors are standardized through cross-sectional demeaning and unit variance scaling, meaning that an increase of one standard deviation in a coefficient results in a corresponding increase in the dependent variable by the coefficient's amount, which is expressed in percentages. Lastly, the panel regression uses clustered standard errors.

C.2 *Portfolio Sort*

The following Appendix shows the comparison of Long-Short (LS) portfolio returns for all 22 firm level characteristic, across different clustering models and configurations. The tables delineate the performance metrics of long-short portfolios constructed using traditional factors and cluster-neutral. Each model is evaluated based on decile performance of selected financial factors, highlighting the shift from traditional to cluster-neutral analysis. For each factor, the table presents monthly average returns (R) in percentages, Annualized Sharpe Ratios (SR), and Fama-French 5-factor model alpha ($FF5^{\alpha}$) across the first decile (1), the tenth decile (10), and the long-short (LS) strategy, which involves buying the first decile and selling the tenth decile.

C.2.1 *K-Means*

Factor	Decile	Normal			K=3			K=5			K=10			K=100		
		R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$
	0	1.05	0.14	0.05	1.08	0.14	0.02	1.03	0.13	0.01	1.05	0.13	-0.01	1.04	0.14	-0.06
Accruals	9	1.48	0.21	-0.23	1.35	0.18	-0.20	1.34	0.18	-0.18	1.33	0.18	-0.17	1.31	0.18	-0.15
	LS	0.43	0.07	-0.28	0.28	0.05	-0.22	0.31	0.05	-0.19	0.28	0.05	-0.16	0.27	0.04	-0.10
	0	1.28	0.21	0.17	1.34	0.21	0.14	1.29	0.20	0.15	1.21	0.18	0.17	1.18	0.19	0.15
AssetGrowth	9	1.30	0.14	-0.52	1.00	0.11	-0.46	1.04	0.10	-0.41	1.15	0.12	-0.38	0.94	0.09	-0.30
	LS	0.02	-0.07	-0.70	-0.34	-0.11	-0.60	-0.25	-0.10	-0.56	-0.06	-0.06	-0.55	-0.23	-0.09	-0.45
	0	1.49	0.25	0.36	1.44	0.25	0.32	1.46	0.26	0.37	1.40	0.25	0.33	1.58	0.26	0.17
BM	9	1.41	0.14	-0.66	1.05	0.10	-0.62	1.33	0.14	-0.67	1.48	0.18	-0.64	1.14	0.15	-0.55
	LS	-0.08	-0.12	-1.02	-0.39	-0.15	-0.94	-0.13	-0.12	-1.04	0.08	-0.06	-0.97	-0.43	-0.11	-0.71
	0	0.69	0.19	0.25	0.71	0.17	0.11	0.71	0.16	0.05	0.79	0.16	0.04	0.85	0.13	-0.12
Beta	9	1.62	0.07	-0.42	1.52	0.09	-0.37	1.56	0.09	-0.32	1.43	0.09	-0.32	1.57	0.15	-0.28
	LS	0.94	-0.11	-0.67	0.81	-0.09	-0.47	0.85	-0.07	-0.37	0.64	-0.06	-0.36	0.71	0.02	-0.16
	0	0.76	0.09	-0.02	0.84	0.10	0.06	1.01	0.13	0.07	1.23	0.18	0.12	1.17	0.18	0.18
CompEquIss	9	1.59	0.32	0.61	1.61	0.31	0.57	1.59	0.30	0.57	1.57	0.31	0.56	1.47	0.28	0.52
	LS	0.83	0.23	0.63	0.77	0.22	0.51	0.58	0.18	0.50	0.35	0.13	0.44	0.31	0.10	0.35
	0	1.12	0.39	0.93	1.11	0.34	0.89	1.10	0.32	0.87	0.94	0.24	0.83	0.93	0.20	0.65
DolVol	9	1.11	0.12	-0.78	1.13	0.14	-0.66	1.18	0.15	-0.61	1.24	0.16	-0.58	1.30	0.21	-0.47
	LS	-0.01	-0.27	-1.72	0.02	-0.20	-1.54	0.08	-0.17	-1.48	0.30	-0.08	-1.41	0.37	0.00	-1.13
	0	0.69	0.04	-0.56	0.82	0.08	-0.43	0.82	0.09	-0.39	0.84	0.09	-0.37	0.91	0.11	-0.28
GP	9	1.29	0.23	0.29	1.27	0.22	0.26	1.29	0.21	0.22	1.23	0.20	0.18	1.24	0.20	0.10
	LS	0.59	0.19	0.85	0.45	0.13	0.69	0.47	0.12	0.61	0.39	0.11	0.56	0.34	0.10	0.38
	0	0.74	0.12	-0.13	0.90	0.15	-0.09	0.90	0.15	-0.10	0.96	0.16	-0.11	0.93	0.16	-0.11
IndMom	9	1.13	0.18	0.05	1.10	0.18	-0.01	1.20	0.19	0.01	1.07	0.16	-0.01	1.09	0.16	-0.05
	LS	0.38	0.06	0.18	0.20	0.03	0.08	0.30	0.04	0.11	0.11	0.00	0.10	0.16	-0.01	0.07
	0	1.23	0.21	0.11	1.27	0.22	0.07	1.28	0.22	0.08	1.20	0.20	0.08	1.18	0.20	0.04
InvestPPEInv	9	1.21	0.17	-0.23	1.13	0.15	-0.16	1.03	0.13	-0.13	1.13	0.15	-0.12	1.05	0.16	-0.09
	LS	-0.02	-0.04	-0.34	-0.14	-0.06	-0.23	-0.25	-0.09	-0.21	-0.07	-0.04	-0.21	-0.13	-0.04	-0.12
	0	0.95	0.19	0.35	0.97	0.19	0.34	0.94	0.18	0.33	0.94	0.18	0.33	0.86	0.16	0.32
Investment	9	1.72	0.21	0.08	1.62	0.21	0.09	1.60	0.20	0.09	1.61	0.22	0.09	1.47	0.21	0.11
	LS	0.78	0.03	-0.27	0.64	0.02	-0.24	0.66	0.03	-0.24	0.67	0.04	-0.24	0.61	0.05	-0.21
	0	1.06	0.02	-1.10	1.19	0.06	-1.00	1.16	0.06	-0.98	1.08	0.06	-0.96	1.20	0.11	-0.84
MaxRet	9	0.99	0.40	0.23	1.24	0.26	0.00	1.42	0.27	-0.04	1.40	0.27	-0.10	1.20	0.14	-0.40
	LS	-0.08	0.38	1.32	0.06	0.20	1.00	0.26	0.21	0.94	0.33	0.21	0.86	0.00	0.03	0.44
	(1)	0.50	-0.06	-1.35	0.40	-0.07	-1.15	0.53	-0.06	-1.15	0.80	0.01	-0.95	1.19	0.11	-0.71
Mom12m	(10)	2.21	0.39	0.48	2.30	0.39	0.43	2.14	0.37	0.44	2.16	0.36	0.40	2.35	0.39	0.30
	(LS)	1.71	0.45	1.84	1.89	0.46	1.58	1.61	0.42	1.59	1.36	0.35	1.35	1.15	0.29	1.02
	(1)	0.65	-0.01	1.01	0.66	0.00	-0.93	0.78	0.03	-0.93	0.81	0.03	-0.87	1.04	0.07	-0.81
Mom1m	(10)	1.70	0.23	-0.09	1.71	0.22	-0.16	1.66	0.22	-0.15	1.80	0.24	-0.19	1.57	0.20	-0.24
	(LS)	1.05	0.23	0.92	1.05	0.22	0.77	0.88	0.19	0.78	0.99	0.21	0.69	0.53	0.13	0.57
	(1)	0.65	-0.02	-1.31	0.84	0.02	-1.17	0.84	0.03	-1.18	1.14	0.08	-1.02	1.30	0.11	-0.88
Mom6m	(10)	2.21	0.35	0.36	2.30	0.36	0.19	2.18	0.34	0.20	2.30	0.36	0.14	2.19	0.33	0.04
	(LS)	1.56	0.37	1.68	1.46	0.33	1.36	1.34	0.31	1.38	1.16	0.28	1.16	0.89	0.22	0.93
	(1)	1.09	0.08	-0.59	1.26	0.15	-0.16	1.00	0.11	-0.11	1.39	0.22	-0.09	1.14	0.12	-0.17
OScore	(10)	0.82	0.19	0.71	0.77	0.06	-0.07	1.13	0.07	-0.24	1.04	0.06	-0.62	1.14	0.06	-0.78
	(LS)	-0.28	0.11	1.30	-0.48	-0.08	0.09	0.13	-0.04	-0.13	-0.35	-0.17	-0.53	0.00	-0.07	-0.61
	(1)	1.07	0.17	0.13	1.05	0.16	0.15	1.10	0.17	0.15	1.02	0.16	0.18	0.97	0.16	0.17
OperProf	(10)	1.20	0.34	0.73	1.21	0.27	0.56	1.23	0.25	0.36	1.17	0.23	0.38	1.22	0.23	0.29
	(LS)	0.13	0.16	0.60	0.16	0.11	0.41	0.13	0.08	0.22	0.15	0.08	0.20	0.25	0.06	0.12
	(1)	1.20	0.08	-1.01	1.37	0.12	-1.06	1.22	0.09	-1.06	1.21	0.10	-1.08	1.40	0.13	-1.06
RET	(10)	1.25	0.21	0.32	1.60	0.19	-0.08	1.58	0.19	-0.08	1.43	0.16	-0.07	1.40	0.16	-0.07
	(LS)	0.04	0.13	1.34	0.23	0.07	0.97	0.36	0.09	0.97	0.22	0.06	1.01	0.00	0.03	0.99
	(1)	1.09	0.07	-0.93	1.09	0.07	-0.93	1.07	0.08	-0.89	1.17	0.09	-0.87	1.31	0.12	-0.77
RoE	(10)	1.25	0.20	0.32	1.29	0.20	0.17	1.35	0.19	0.10	1.42	0.19	0.00	1.33	0.16	-0.28
	(LS)	0.04	0.13	1.34	0.20	0.13	1.10	0.28	0.11	0.99	0.25	0.10	0.88	0.02	0.04	0.49
	(1)	1.22	0.27	0.74	1.19	0.25	0.70	1.20	0.25	0.70	1.13	0.23	0.69	1.07	0.22	0.65
ShareIss5Y	(10)	1.20	0.23	0.10	1.10	0.19	0.24	1.19	0.22	0.30	1.19	0.21	0.30	1.13	0.23	0.42
	(LS)	-0.02	-0.04	-0.65	-0.09	-0.06	-0.46	-0.01	-0.04	-0.40	0.06	-0.03	-0.39	0.07	0.01	-0.23
	(1)	1.39	0.18	-0.19	1.46	0.22	-0.07	1.48	0.23	-0.06	1.81	0.33	-0.05	1.39	0.22	-0.02
ShareVol	(10)	0.64	0.14	0.61	1.25	0.17	0.13	1.31	0.21	0.11	1.22	0.17	-0.03	1.18	0.14	-0.10
	(LS)	-0.75	-0.04	0.80	-0.20	-0.05	0.20	-0.17	-0.02	0.16	-0.59	-0.16	0.02	-0.21	-0.08	-0.07
	(1)	1.06	0.15	0.10	1.04	0.15	0.05	0.96	0.13	0.04	1.03	0.14	0.06	0.94	0.13	0.04
dNoa	(10)	1.50	0.19	-0.28	1.23	0.15	-0.20	1.37	0.18	-0.15	1.38	0.18	-0.15	1.19	0.15	-0.09
	(LS)	0.44	0.04	-0.39	0.19	-0.00	-0.25	0.41	0.05	-0.19	0.35	0.04	-0.21	0.25	0.02	-0.12
	(1)	0.95	-0.01	-1.77	0.62	0.01	-0.36	0.94	0.14	0.11	0.98	0.20	0.42	0.90	0.27	0.77
marketcap	(10)	1.18	0.51	1.01	1.19	0.42	0.98	1.19	0.41	1.01	1.19	0.43	1.01	1.21	0.42	1.02
	(LS)	0.22	0.52	2.78	0.57	0.41	1.33	0.25	0.28	0.90	0.21	0.23	0.59	0.31	0.15	0.25
	(1)	1.05	0.03	-0.92	1.20	0.06	-0.84	1.20	0.07	-0.80	1.23	0.09	-0.74	1.32	0.13	-0.66
roaq	(10)	1.22	0.27	0.71	1.32	0.22	0.50	1.33	0.21	0.37	1.25	0.17	0.22	1.31	0.16	-0.11
	(LS)	0.17	0.24	1.63	0.12	0.16	1.34	0.13	0.14	1.17	0.03	0.09	0.96	-0.01	0.03	0.55

Table 15: Long Short Portfolio For Different Configuration of K, for K-Means.

C.2.2 *Hierarchical Clustering*

Factor	Decile	N			DMax = 20			DMax = 40			DMax = 60			DMax = 80		
		R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$
	0	1.05	0.14	0.05	1.07	0.13	-0.05	1.09	0.14	-0.02	1.06	0.14	-0.01	1.07	0.14	0.00
Accruals	9	1.48	0.21	-0.23	1.26	0.17	-0.16	1.29	0.18	-0.17	1.30	0.18	-0.18	1.24	0.16	-0.18
	LS	0.43	0.07	-0.28	0.19	0.03	-0.11	0.20	0.04	-0.15	0.24	0.04	-0.17	0.17	0.02	-0.18
	0	1.28	0.21	0.17	1.25	0.20	0.16	1.26	0.19	0.18	1.23	0.19	0.18	1.23	0.18	0.17
AssetGrowth	9	1.30	0.14	-0.52	1.04	0.09	-0.33	0.93	0.09	-0.38	1.11	0.13	-0.39	1.03	0.10	-0.43
	LS	0.02	-0.07	-0.70	-0.21	-0.11	-0.49	-0.33	-0.10	-0.57	-0.12	-0.06	-0.57	-0.20	-0.08	-0.59
	0	1.49	0.25	0.36	1.45	0.24	0.24	1.42	0.24	0.33	1.42	0.25	0.38	1.41	0.25	0.40
BM	9	1.41	0.14	-0.66	1.09	0.14	-0.60	1.20	0.14	-0.66	1.24	0.14	-0.68	1.52	0.18	-0.69
	LS	-0.08	-0.12	-1.02	-0.36	-0.10	-0.84	-0.22	-0.10	-0.98	-0.18	-0.11	-1.06	0.11	-0.07	-1.09
	0	0.69	0.19	0.25	0.80	0.14	-0.07	0.79	0.16	0.01	0.72	0.15	0.07	0.72	0.16	0.10
Beta	9	1.62	0.07	-0.42	1.67	0.15	-0.29	1.68	0.14	-0.31	1.62	0.12	-0.33	1.62	0.11	-0.34
	LS	0.94	-0.11	-0.67	0.87	0.02	-0.22	0.90	-0.02	-0.32	0.91	-0.03	-0.40	0.90	-0.05	-0.44
	0	0.76	0.09	-0.02	1.22	0.20	0.14	1.05	0.16	0.12	0.99	0.15	0.08	1.01	0.14	0.05
CompEquIss	9	1.59	0.32	0.61	1.49	0.27	0.54	1.62	0.31	0.56	1.63	0.32	0.57	1.59	0.31	0.58
	LS	0.83	0.23	0.63	0.27	0.07	0.40	0.57	0.15	0.44	0.64	0.17	0.49	0.57	0.17	0.53
	0	1.12	0.39	0.93	0.91	0.21	0.70	0.93	0.23	0.79	0.94	0.25	0.85	0.95	0.28	0.88
DolVol	9	1.11	0.12	-0.78	1.44	0.24	-0.51	1.37	0.20	-0.58	1.29	0.18	-0.62	1.24	0.15	-0.63
	LS	-0.01	-0.27	-1.72	0.53	0.02	-1.21	0.44	-0.03	-1.38	0.35	-0.06	-1.47	0.29	-0.13	-1.50
	0	0.69	0.04	-0.56	0.92	0.11	0.80	0.09	-0.33	0.79	0.87	0.10	-0.37	0.81	0.08	-0.41
GP	9	1.29	0.23	0.29	1.27	0.21	0.12	1.23	0.20	0.16	1.28	0.21	0.18	1.23	0.20	0.20
	LS	0.59	0.19	0.85	0.35	0.10	0.41	0.44	0.11	0.49	0.41	0.11	0.55	0.42	0.11	0.61
	0	0.74	0.12	-0.13	1.02	0.18	-0.12	0.96	0.17	-0.13	0.90	0.16	-0.13	0.86	0.15	-0.13
IndMom	9	1.13	0.18	0.05	1.10	0.16	-0.02	1.09	0.17	0.01	1.05	0.15	0.03	1.15	0.18	0.04
	LS	0.38	0.06	0.18	0.08	-0.02	0.10	0.13	0.00	0.15	0.14	-0.00	0.15	0.30	0.03	0.17
	0	1.23	0.21	0.11	1.22	0.21	0.06	1.29	0.23	0.08	1.23	0.21	0.09	1.25	0.21	0.10
InvestPPEInv	9	1.21	0.17	-0.23	0.99	0.14	-0.10	1.11	0.15	-0.11	1.11	0.16	-0.12	1.23	0.18	-0.16
	LS	-0.02	-0.04	-0.34	-0.23	-0.07	-0.16	-0.18	-0.07	-0.19	-0.12	-0.05	-0.22	-0.02	-0.03	-0.25
	0	0.95	0.19	0.35	0.86	0.16	0.32	0.97	0.19	0.33	0.91	0.17	0.33	0.89	0.17	0.34
Investment	9	1.72	0.21	0.08	1.54	0.21	0.11	1.47	0.19	0.10	1.44	0.17	0.10	1.58	0.20	0.09
	LS	0.78	0.03	-0.27	0.68	0.05	-0.22	0.50	-0.01	-0.23	0.53	0.00	-0.24	0.69	0.03	-0.24
	0	1.06	0.02	-1.10	1.13	0.09	-0.87	1.09	0.07	-0.95	1.11	0.07	-0.98	0.94	0.05	-1.00
MaxRet	9	0.99	0.40	0.23	1.25	0.16	-0.34	1.20	0.19	-0.18	1.08	0.21	-0.07	1.17	0.26	-0.01
	LS	-0.08	0.38	1.32	0.12	0.08	0.53	0.12	0.12	0.76	-0.04	0.14	0.91	0.23	0.21	0.99
	(1)	0.50	-0.06	-1.35	1.07	0.08	-0.83	0.82	0.02	-1.03	0.70	-0.00	-1.11	0.59	-0.04	-1.22
Mom12m	(10)	2.21	0.39	0.48	2.25	0.38	0.36	2.20	0.37	0.41	2.13	0.35	0.43	2.12	0.36	0.45
	(LS)	1.71	0.45	1.84	1.18	0.30	1.19	1.39	0.35	1.43	1.43	0.35	1.54	1.54	0.40	1.67
	(1)	0.65	-0.01	1.01	1.04	0.06	-0.86	1.02	0.05	-0.90	0.93	0.03	-0.93	0.69	0.01	-0.96
Mom1m	(10)	1.70	0.23	-0.09	1.55	0.19	-0.21	1.50	0.19	-0.18	1.54	0.20	-0.15	1.58	0.21	-0.12
	(LS)	1.05	0.23	0.92	0.51	0.13	0.65	0.48	0.14	0.72	0.61	0.17	0.79	0.89	0.20	0.84
	(1)	0.65	-0.02	-1.31	1.27	0.11	-0.97	1.15	0.08	-1.07	0.93	0.04	-1.12	0.76	0.01	-1.22
Mom6m	(10)	2.21	0.35	0.36	2.24	0.34	0.10	2.31	0.35	0.17	2.34	0.36	0.19	2.34	0.37	0.23
	(LS)	1.56	0.37	1.68	0.97	0.22	1.07	1.15	0.27	1.24	1.41	0.31	1.32	1.58	0.36	1.45
	(1)	1.09	0.08	-0.59	1.04	0.13	-0.14	1.27	0.20	-0.05	1.43	0.23	-0.09	1.55	0.25	-0.14
OScore	(10)	0.82	0.19	0.71	1.25	0.06	-0.73	1.28	0.08	-0.62	1.17	0.06	-0.50	1.17	0.10	-0.25
	(LS)	-0.28	0.11	1.30	0.21	-0.07	-0.59	0.01	-0.12	-0.57	-0.26	-0.17	-0.41	-0.39	-0.14	-0.12
	(1)	1.07	0.17	0.13	1.06	0.18	0.16	1.12	0.18	0.14	1.11	0.16	0.11	1.04	0.15	0.09
OperProf	(10)	1.20	0.34	0.73	1.29	0.25	0.29	1.26	0.25	0.30	1.17	0.24	0.36	1.17	0.25	0.39
	(LS)	0.13	0.16	0.60	0.24	0.07	0.13	0.14	0.07	0.16	0.06	0.08	0.25	0.14	0.10	0.30
	(1)	1.20	0.08	-1.01	1.29	0.12	-0.79	1.23	0.10	-0.83	1.10	0.08	-0.87	1.21	0.09	-0.90
RoE	(10)	1.25	0.21	0.32	1.38	0.18	-0.25	1.31	0.17	-0.09	1.25	0.17	-0.01	1.22	0.18	0.07
	(LS)	0.04	0.13	1.34	0.09	0.06	0.53	0.08	0.06	0.73	0.15	0.09	0.87	0.01	0.08	0.97
	(1)	1.22	0.27	0.74	1.17	0.24	0.66	1.14	0.24	0.68	1.16	0.24	0.70	1.19	0.25	0.71
ShareIss5Y	(10)	1.20	0.23	0.10	1.15	0.22	0.36	1.16	0.23	0.29	1.10	0.21	0.25	1.25	0.24	0.20
	(LS)	-0.02	-0.04	-0.65	-0.03	-0.02	-0.30	0.02	-0.01	-0.39	-0.06	-0.03	-0.45	0.06	-0.01	-0.51
	(1)	1.39	0.18	-0.19	1.39	0.22	-0.07	1.54	0.28	-0.01	1.59	0.28	-0.02	1.73	0.30	-0.04
ShareVol	(10)	0.64	0.14	0.61	1.33	0.18	-0.10	1.32	0.16	-0.14	1.36	0.19	-0.09	1.39	0.20	0.04
	(LS)	-0.75	-0.04	0.80	-0.06	-0.03	-0.03	-0.23	-0.12	-0.13	-0.22	-0.09	-0.08	-0.34	-0.10	0.08
	(1)	1.06	0.15	0.10	0.99	0.15	0.04	1.02	0.14	0.06	1.01	0.14	0.07	1.01	0.14	0.07
dNoa	(10)	1.50	0.19	-0.28	1.24	0.16	-0.10	1.19	0.15	-0.15	1.32	0.17	-0.17	1.20	0.15	-0.19
	(LS)	0.44	0.04	-0.39	0.25	0.02	-0.15	0.18	0.01	-0.21	0.32	0.03	-0.23	0.19	0.01	-0.26
	(1)	0.95	-0.01	-1.77	0.89	0.25	0.71	0.86	0.20	0.52	0.78	0.15	0.27	0.89	0.12	0.03
marketcap	(10)	1.18	0.51	1.01	1.20	0.43	1.02	1.19	0.43	1.02	1.20	0.42	1.02	1.19	0.42	1.03
	(LS)	0.22	0.52	2.78	0.31	0.17	0.30	0.34	0.24	0.50	0.41	0.27	0.75	0.31	0.30	1.00
	(1)	1.05	0.03	-0.92	1.27	0.11	-0.70	1.18	0.09	-0.75	1.08	0.06	-0.78	1.07	0.05	-0.80
roaq	(10)	1.22	0.27	0.71	1.36	0.17	-0.01	1.37	0.18	0.16	1.36	0.19	0.29	1.33	0.20	0.38
	(LS)	0.17	0.24	1.63	0.10	0.06	0.69	0.19	0.08	0.90	0.28	0.13	1.07	0.26	0.15	1.18

Table 16: Long Short Portfolio For Different Configuration of DMax, for Hierarchical Clustering.

C.2.3 *DBSCAN*

Factor	Decile	Normal			ε = 4 , M= 5			ε = 2 , M= 10			ε = 3 , M= 25			ε = 2 , M= 50		
		R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$
	0	1.05	0.14	0.05	1.12	0.15	0.03	1.04	0.13	0.01	1.08	0.14	0.04	1.07	0.14	0.04
Accruals	9	1.48	0.21	-0.23	1.33	0.18	-0.19	1.40	0.20	-0.18	1.34	0.19	-0.21	1.37	0.19	-0.22
	LS	0.43	0.07	-0.28	0.21	0.03	-0.22	0.35	0.07	-0.19	0.26	0.05	-0.25	0.30	0.05	-0.25
	0	1.28	0.21	0.17	1.01	0.14	0.19	1.18	0.19	0.21	1.10	0.17	0.22	1.16	0.19	0.20
AssetGrowth	9	1.30	0.14	-0.52	1.57	0.17	-0.58	1.42	0.15	-0.61	1.48	0.15	-0.71	1.38	0.14	-0.60
	LS	0.02	-0.07	-0.70	0.56	0.03	-0.77	0.24	-0.04	-0.82	0.39	-0.01	-0.93	0.22	-0.04	-0.80
	0	1.49	0.25	0.36	1.52	0.28	0.43	1.43	0.25	0.40	1.49	0.26	0.42	1.45	0.25	0.39
BM	9	1.41	0.14	-0.66	1.18	0.08	-0.74	1.19	0.10	-0.75	1.17	0.09	-0.75	1.23	0.11	-0.74
	LS	-0.08	-0.12	-1.02	-0.35	-0.20	-1.17	-0.23	-0.15	-1.15	-0.31	-0.17	-1.17	-0.22	-0.14	-1.13
	0	0.69	0.19	0.25	0.91	0.21	0.17	0.83	0.18	0.13	0.86	0.18	0.14	0.82	0.18	0.14
Beta	9	1.62	0.07	-0.42	1.46	0.06	-0.35	1.39	0.06	-0.35	1.52	0.08	-0.35	1.42	0.06	-0.35
	LS	0.94	-0.11	-0.67	0.55	-0.14	-0.52	0.56	-0.11	-0.47	0.66	-0.10	-0.49	0.60	-0.12	-0.49
	0	0.76	0.09	-0.02	0.95	0.11	-0.13	0.90	0.11	-0.06	0.91	0.11	-0.11	0.83	0.10	-0.05
CompEquIss	9	1.59	0.32	0.61	1.60	0.32	0.61	1.60	0.32	0.63	1.62	0.33	0.63	1.61	0.33	0.63
	LS	0.83	0.23	0.63	0.65	0.20	0.74	0.70	0.22	0.69	0.71	0.22	0.74	0.77	0.23	0.68
	0	1.12	0.39	0.93	1.13	0.39	0.92	1.12	0.37	0.93	1.12	0.38	0.93	1.13	0.38	0.92
DolVol	9	1.11	0.12	-0.78	1.08	0.11	-0.73	1.05	0.11	-0.75	1.14	0.14	-0.75	1.04	0.12	-0.76
	LS	-0.01	-0.27	-1.72	-0.05	-0.27	-1.66	-0.07	-0.26	-1.68	0.02	-0.24	-1.68	-0.09	-0.26	-1.68
	0	0.69	0.04	-0.56	0.68	0.07	-0.41	0.70	0.07	-0.48	0.71	0.07	-0.41	0.67	0.06	-0.51
GP	9	1.29	0.23	0.29	1.47	0.26	0.25	1.37	0.25	0.27	1.39	0.25	0.27	1.34	0.24	0.28
	LS	0.59	0.19	0.85	0.79	0.19	0.66	0.67	0.18	0.75	0.68	0.18	0.68	0.66	0.18	0.78
	0	0.74	0.12	-0.13	0.78	0.12	-0.14	0.80	0.14	-0.16	0.81	0.14	-0.15	0.76	0.13	-0.17
IndMom	9	1.13	0.18	0.05	1.08	0.16	0.05	1.06	0.17	0.06	1.05	0.16	0.05	1.12	0.18	0.06
	LS	0.38	0.06	0.18	0.30	0.04	0.19	0.26	0.03	0.22	0.23	0.02	0.20	0.36	0.05	0.23
	0	1.23	0.21	0.11	1.01	0.17	0.12	1.22	0.22	0.15	1.13	0.20	0.15	1.25	0.22	0.14
InvestPPEInv	9	1.21	0.17	-0.23	1.34	0.16	-0.36	1.28	0.17	-0.33	1.21	0.15	-0.40	1.28	0.17	-0.31
	LS	-0.02	-0.04	-0.34	0.33	-0.01	-0.48	0.06	-0.05	-0.48	0.08	-0.04	-0.56	0.03	-0.05	-0.44
	0	0.95	0.19	0.35	0.95	0.19	0.34	0.96	0.19	0.35	0.97	0.19	0.35	0.98	0.20	0.35
Investment	9	1.72	0.21	0.08	1.68	0.21	0.08	1.75	0.23	0.07	1.74	0.23	0.07	1.79	0.23	0.07
	LS	0.78	0.03	-0.27	0.72	0.02	-0.26	0.79	0.04	-0.28	0.77	0.04	-0.27	0.81	0.03	-0.27
	0	1.06	0.02	-1.10	0.85	0.00	-1.05	0.91	0.03	-1.06	0.88	0.01	-1.04	0.92	0.02	-1.06
MaxRet	9	0.99	0.40	0.23	1.33	0.29	0.05	1.16	0.26	0.19	1.25	0.25	0.04	1.10	0.27	0.23
	LS	-0.08	0.38	1.32	0.47	0.29	1.10	0.26	0.24	1.25	0.37	0.24	1.07	0.18	0.25	1.29
	(1)	0.50	-0.06	-1.35	0.32	-0.08	-1.33	0.32	-0.08	-1.33	0.47	-0.07	-1.33	0.39	-0.07	-1.33
Mom12m	(10)	2.21	0.39	0.48	2.27	0.39	0.48	2.23	0.39	0.49	2.22	0.38	0.49	2.22	0.38	0.49
	(LS)	1.71	0.45	1.84	1.96	0.47	1.81	1.91	0.46	1.82	1.75	0.45	1.83	1.83	0.45	1.82
	(1)	0.65	-0.01	1.01	0.72	0.01	-1.01	0.68	-0.00	-1.02	0.72	0.01	-1.02	0.66	-0.01	-1.02
Mom1m	(10)	1.70	0.23	-0.09	1.59	0.21	-0.09	1.62	0.22	-0.09	1.58	0.21	-0.08	1.69	0.23	-0.09
	(LS)	1.05	0.23	0.92	0.87	0.20	0.92	0.95	0.22	0.93	0.85	0.19	0.94	1.04	0.23	0.93
	(1)	0.65	-0.02	-1.31	0.63	0.01	-1.31	0.74	0.02	-1.31	0.63	0.00	-1.32	0.76	0.02	-1.31
Mom6m	(10)	2.21	0.35	0.36	2.11	0.34	0.26	2.19	0.35	0.26	2.11	0.33	0.27	2.18	0.34	0.26
	(LS)	1.56	0.37	1.68	1.48	0.33	1.57	1.45	0.33	1.57	1.48	0.33	1.59	1.42	0.33	1.57
	(1)	1.09	0.08	-0.59	1.24	0.10	-0.52	1.29	0.14	-0.38	1.38	0.13	-0.49	1.16	0.11	-0.34
OScore	(10)	0.82	0.19	0.71	1.59	0.24	0.17	0.74	0.03	0.08	1.41	0.16	-0.01	0.78	0.06	0.19
	(LS)	-0.28	0.11	1.30	0.34	0.14	0.69	-0.55	-0.11	0.46	0.03	0.03	0.48	-0.39	-0.06	0.54
	(1)	1.07	0.17	0.13	1.22	0.13	-0.12	1.18	0.19	0.15	1.15	0.16	0.10	1.11	0.17	0.15
OperProf	(10)	1.20	0.34	0.73	1.19	0.28	0.54	1.22	0.35	0.73	1.21	0.34	0.72	1.26	0.36	0.73
	(LS)	0.13	0.16	0.60	-0.02	0.15	0.66	0.05	0.16	0.58	0.06	0.17	0.62	0.15	0.19	0.58
	(1)	1.20	0.08	-1.01	1.33	0.11	-0.89	1.14	0.07	-0.98	1.12	0.08	-0.95	1.19	0.08	-0.99
RoE	(10)	1.25	0.21	0.32	1.18	0.19	0.15	1.24	0.22	0.31	1.18	0.19	0.22	1.24	0.22	0.32
	(LS)	0.04	0.13	1.34	-0.15	0.07	1.04	0.10	0.14	1.30	0.06	0.11	1.17	0.05	0.13	1.31
	(1)	1.22	0.27	0.74	1.16	0.25	0.75	1.19	0.27	0.75	1.19	0.26	0.75	1.18	0.26	0.75
ShareIss5Y	(10)	1.20	0.23	0.10	1.20	0.15	-0.26	1.22	0.19	-0.10	1.15	0.13	-0.43	1.13	0.18	-0.06
	(LS)	-0.02	-0.04	-0.65	0.04	-0.11	-1.01	0.03	-0.07	-0.85	-0.04	-0.13	-1.18	-0.05	-0.08	-0.81
	(1)	1.39	0.18	-0.19	1.35	0.17	-0.23	1.60	0.25	-0.00	1.56	0.24	-0.10	1.58	0.26	0.01
ShareVol	(10)	0.64	0.14	0.61	1.63	0.23	-0.12	0.57	0.03	0.05	1.35	0.17	-0.08	0.67	0.07	0.13
	(LS)	-0.75	-0.04	0.80	0.28	0.06	0.12	-1.03	-0.23	0.05	-0.21	-0.08	0.02	-0.91	-0.19	0.12
	(1)	1.06	0.15	0.10	0.86	0.12	0.10	0.97	0.15	0.14	0.83	0.12	0.13	1.00	0.15	0.14
dNoa	(10)	1.50	0.19	-0.28	1.72	0.20	-0.37	1.60	0.19	-0.38	1.60	0.19	-0.50	1.55	0.19	-0.36
	(LS)	0.44	0.04	-0.39	0.86	0.08	-0.47	0.63	0.04	-0.52	0.77	0.07	-0.63	0.55	0.04	-0.50
	(1)	0.95	-0.01	-1.77	0.99	0.03	-1.44	0.67	0.03	-0.91	0.86	0.01	-1.30	0.62	0.02	-0.90
marketcap	(10)	1.18	0.51	1.01	1.18	0.45	0.96	1.18	0.49	1.01	1.18	0.48	1.00	1.18	0.49	1.01
	(LS)	0.22	0.52	2.78	0.19	0.43	2.40	0.51	0.46	1.92	0.32	0.47	2.30	0.56	0.47	1.91
	(1)	1.05	0.03	-0.92	1.12	0.06	-0.85	1.17	0.05	-0.89	1.04	0.03	-0.86	1.13	0.04	-0.89
roaq	(10)	1.22	0.27	0.71	1.39	0.23	0.38	1.28	0.25	0.63	1.29	0.23	0.45	1.26	0.26	0.66
	(LS)	0.17	0.24	1.63	0.27	0.17	1.23	0.11	0.20	1.52	0.25	0.20	1.32	0.13	0.21	1.55

Table 17: Long Short Portfolio For Different Configuration of ε (Epsilon) values and MinPts (M), for DBSCAN (NOT taking into account the outlier cap of 10 procent).

Factor	Decile	N			ε = 2, M= 100			ε = 2, M= 500			ε = 3, M= 1000			ε = 4, M= 2000		
		R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$
	0	1.05	0.14	0.05	1.08	0.14	0.04	1.06	0.14	0.04	1.06	0.14	0.04	1.08	0.15	0.04
Accruals	9	1.48	0.21	-0.23	1.38	0.20	-0.22	1.45	0.20	-0.23	1.37	0.19	-0.22	1.33	0.19	-0.22
	LS	0.43	0.07	-0.28	0.30	0.05	-0.25	0.39	0.06	-0.27	0.31	0.05	-0.25	0.25	0.04	-0.26
	0	1.28	0.21	0.17	1.16	0.18	0.20	1.27	0.21	0.18	1.12	0.18	0.21	1.12	0.17	0.21
AssetGrowth	9	1.30	0.14	-0.52	1.29	0.13	-0.59	1.32	0.14	-0.54	1.45	0.16	-0.60	1.45	0.16	-0.66
	LS	0.02	-0.07	-0.70	0.13	-0.05	-0.78	0.06	-0.07	-0.71	0.33	-0.02	-0.80	0.33	-0.01	-0.87
	0	1.49	0.25	0.36	1.45	0.25	0.39	1.47	0.25	0.37	1.43	0.24	0.40	1.51	0.26	0.41
BM	9	1.41	0.14	-0.66	1.26	0.11	-0.73	1.32	0.12	-0.70	1.24	0.10	-0.73	1.15	0.09	-0.73
	LS	-0.08	-0.12	-1.02	-0.19	-0.14	-1.12	-0.15	-0.13	-1.07	-0.19	-0.14	-1.13	-0.36	-0.17	-1.14
	0	0.69	0.19	0.25	0.84	0.18	0.14	0.77	0.18	0.20	0.86	0.20	0.16	0.83	0.19	0.17
Beta	9	1.62	0.07	-0.42	1.43	0.06	-0.36	1.53	0.07	-0.40	1.33	0.05	-0.35	1.47	0.07	-0.36
	LS	0.94	-0.11	-0.67	0.59	-0.12	-0.50	0.76	-0.11	-0.60	0.47	-0.15	-0.51	0.64	-0.12	-0.53
	0	0.76	0.09	-0.02	0.84	0.10	-0.04	0.79	0.09	-0.02	0.91	0.11	-0.05	0.88	0.10	-0.08
CompEquIss	9	1.59	0.32	0.61	1.60	0.33	0.63	1.60	0.32	0.62	1.62	0.33	0.63	1.60	0.32	0.62
	LS	0.83	0.23	0.63	0.76	0.22	0.67	0.80	0.22	0.64	0.71	0.22	0.67	0.72	0.22	0.71
	0	1.12	0.39	0.93	1.13	0.38	0.92	1.12	0.39	0.92	1.12	0.38	0.92	1.12	0.38	0.92
DolVol	9	1.11	0.12	-0.78	1.02	0.11	-0.76	1.13	0.12	-0.81	1.09	0.12	-0.77	1.13	0.13	-0.77
	LS	-0.01	-0.27	-1.72	-0.10	-0.26	-1.68	0.01	-0.26	-1.73	-0.04	-0.26	-1.69	0.02	-0.26	-1.69
	0	0.69	0.04	-0.56	0.66	0.05	-0.52	0.66	0.04	-0.55	0.72	0.07	-0.43	0.72	0.07	-0.42
GP	9	1.29	0.23	0.29	1.34	0.24	0.28	1.29	0.23	0.29	1.40	0.25	0.27	1.40	0.25	0.27
	LS	0.59	0.19	0.85	0.68	0.19	0.80	0.63	0.19	0.84	0.69	0.18	0.71	0.67	0.18	0.70
	0	0.74	0.12	-0.13	0.73	0.12	-0.17	0.73	0.12	-0.15	0.82	0.14	-0.15	0.82	0.14	-0.14
IndMom	9	1.13	0.18	0.05	1.10	0.18	0.06	1.16	0.19	0.05	1.09	0.18	0.06	1.12	0.18	0.05
	LS	0.38	0.06	0.18	0.36	0.06	0.23	0.43	0.07	0.20	0.26	0.03	0.21	0.30	0.04	0.20
	0	1.23	0.21	0.11	1.24	0.22	0.14	1.22	0.21	0.12	1.22	0.21	0.14	1.12	0.19	0.14
InvestPPEInv	9	1.21	0.17	-0.23	1.30	0.18	-0.29	1.22	0.17	-0.24	1.23	0.17	-0.34	1.04	0.13	-0.37
	LS	-0.02	-0.04	-0.34	0.06	-0.04	-0.43	0.00	-0.04	-0.36	0.02	-0.04	-0.48	-0.08	-0.06	-0.51
	0	0.95	0.19	0.35	0.97	0.19	0.35	0.93	0.18	0.35	0.98	0.20	0.35	0.98	0.20	0.35
Investment	9	1.72	0.21	0.08	1.78	0.23	0.08	1.73	0.21	0.07	1.77	0.23	0.08	1.72	0.22	0.08
	LS	0.78	0.03	-0.27	0.82	0.04	-0.27	0.79	0.03	-0.27	0.79	0.03	-0.27	0.74	0.03	-0.27
	0	1.06	0.02	-1.10	0.95	0.03	-1.06	0.95	0.01	-1.09	0.88	0.02	-1.05	0.94	0.02	-1.04
MaxRet	9	0.99	0.40	0.23	1.08	0.28	0.25	0.99	0.35	0.30	1.13	0.26	0.15	1.18	0.26	0.07
	LS	-0.08	0.38	1.32	0.13	0.25	1.31	0.05	0.34	1.39	0.25	0.24	1.20	0.24	0.25	1.12
	(1)	0.50	-0.06	-1.35	0.43	-0.07	-1.34	0.48	-0.07	-1.35	0.39	-0.07	-1.32	0.60	-0.03	-1.34
Mom12m	(10)	2.21	0.39	0.48	2.22	0.38	0.48	2.22	0.39	0.48	2.26	0.39	0.49	2.24	0.39	0.49
	(LS)	1.71	0.45	1.84	1.79	0.45	1.82	1.74	0.45	1.83	1.87	0.46	1.81	1.65	0.42	1.83
	(1)	0.65	-0.01	1.01	0.68	-0.00	-1.01	0.58	-0.02	-1.01	0.55	-0.02	-1.01	0.65	-0.00	-1.02
Mom1m	(10)	1.70	0.23	-0.09	1.71	0.23	-0.09	1.72	0.23	-0.09	1.61	0.22	-0.09	1.57	0.21	-0.08
	(LS)	1.05	0.23	0.92	1.03	0.23	0.93	1.14	0.25	0.92	1.05	0.23	0.93	0.92	0.21	0.93
	(1)	0.65	-0.02	-1.31	0.70	0.01	-1.31	0.59	-0.02	-1.31	0.81	0.03	-1.31	0.69	0.01	-1.32
Mom6m	(10)	2.21	0.35	0.36	2.18	0.34	0.26	2.16	0.34	0.26	2.20	0.34	0.26	2.15	0.34	0.27
	(LS)	1.56	0.37	1.68	1.48	0.33	1.56	1.58	0.36	1.57	1.39	0.32	1.57	1.46	0.33	1.59
	(1)	1.09	0.08	-0.59	0.97	0.09	-0.32	0.89	0.09	-0.32	1.02	0.06	-0.41	1.06	0.06	-0.50
OScore	(10)	0.82	0.19	0.71	0.92	0.12	0.27	0.76	0.16	0.62	1.10	0.09	-0.02	1.47	0.17	0.00
	(LS)	-0.28	0.11	1.30	-0.06	0.03	0.59	-0.13	0.07	0.94	0.08	0.03	0.39	0.42	0.12	0.50
	(1)	1.07	0.17	0.13	1.10	0.17	0.15	1.09	0.17	0.14	1.07	0.15	0.15	1.21	0.18	0.09
OperProf	(10)	1.20	0.34	0.73	1.28	0.37	0.73	1.22	0.34	0.73	1.24	0.35	0.73	1.26	0.36	0.72
	(LS)	0.13	0.16	0.60	0.18	0.19	0.58	0.13	0.17	0.59	0.17	0.20	0.58	0.04	0.18	0.63
	(1)	1.20	0.08	-1.01	1.22	0.08	-1.01	1.22	0.08	-1.01	1.20	0.08	-0.98	1.32	0.10	-0.96
RoE	(10)	1.25	0.21	0.32	1.24	0.22	0.33	1.24	0.22	0.33	1.14	0.20	0.31	1.18	0.20	0.22
	(LS)	0.04	0.13	1.34	0.02	0.14	1.34	0.02	0.14	1.34	-0.05	0.11	1.30	-0.15	0.10	1.19
	(1)	1.22	0.27	0.74	1.20	0.27	0.74	1.20	0.27	0.74	1.18	0.26	0.75	1.19	0.26	0.75
ShareIss5Y	(10)	1.20	0.23	0.10	1.15	0.21	0.05	1.15	0.21	0.05	1.27	0.20	-0.15	1.06	0.13	-0.34
	(LS)	-0.02	-0.04	-0.65	-0.04	-0.06	-0.70	-0.04	-0.06	-0.70	0.09	-0.06	-0.90	-0.14	-0.13	-1.09
	(1)	1.39	0.18	-0.19	1.40	0.19	-0.04	1.40	0.19	-0.04	1.59	0.25	-0.04	1.59	0.25	-0.11
ShareVol	(10)	0.64	0.14	0.61	0.65	0.14	0.42	0.65	0.14	0.42	0.95	0.08	-0.09	1.23	0.14	-0.18
	(LS)	-0.75	-0.04	0.80	-0.74	-0.05	0.46	-0.74	-0.05	0.46	-0.65	-0.17	-0.05	-0.36	-0.11	-0.07
	(1)	1.06	0.15	0.10	1.03	0.15	0.11	1.03	0.15	0.11	0.94	0.14	0.13	0.88	0.12	0.12
dNoa	(10)	1.50	0.19	-0.28	1.50	0.19	-0.30	1.50	0.19	-0.30	1.50	0.19	-0.38	1.41	0.17	-0.44
	(LS)	0.44	0.04	-0.39	0.47	0.04	-0.41	0.47	0.04	-0.41	0.56	0.05	-0.51	0.53	0.04	-0.56
	(1)	0.95	-0.01	-1.77	0.94	-0.01	-1.65	0.94	-0.01	-1.65	0.54	-0.05	-1.35	0.78	-0.02	-1.39
marketcap	(10)	1.18	0.51	1.01	1.18	0.51	1.00	1.18	0.51	1.00	1.18	0.51	1.01	1.18	0.49	1.01
	(LS)	0.22	0.52	2.78	0.24	0.52	2.65	0.24	0.52	2.65	0.64	0.56	2.35	0.40	0.51	2.40
	(1)	1.05	0.03	-0.92	1.09	0.03	-0.92	1.09	0.03	-0.92	1.10	0.04	-0.88	1.11	0.04	-0.87
roaq	(10)	1.22	0.27	0.71	1.21	0.26	0.70	1.21	0.26	0.70	1.22	0.24	0.63	1.27	0.24	0.50
	(LS)	0.17	0.24	1.63	0.12	0.23	1.62	0.12	0.23	1.62	0.12	0.20	1.51	0.17	0.21	1.37

Table 18: Long Short Portfolio For Different Configuration of ε (Epsilon) values and MinPts (M), for DBSCAN (NOT taking into account the outlier cap of 10 procent).

Factor	Decile	N			ε = 4, M= 1			ε = 3, M= 3			ε = 4, M= 5		
		R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$
	(1)	1.05	0.14	0.05	0.98	0.15	0.11	1.07	0.14	0.03	1.12	0.15	0.03
Accruals	(10)	1.48	0.21	-0.23	1.14	0.16	-0.07	1.32	0.19	-0.20	1.33	0.18	-0.19
	(LS)	0.43	0.07	-0.28	0.16	0.02	-0.18	0.26	0.05	-0.23	0.21	0.03	-0.22
	(1)	1.28	0.21	0.17	0.99	0.16	0.29	1.07	0.16	0.21	1.01	0.14	0.19
AssetGrowth	(10)	1.30	0.14	-0.52	1.26	0.17	-0.28	1.49	0.15	-0.69	1.57	0.17	-0.58
	(LS)	0.02	-0.07	-0.70	0.27	0.01	-0.57	0.41	-0.01	-0.90	0.56	0.03	-0.77
	(1)	1.49	0.25	0.36	1.08	0.20	0.50	1.50	0.27	0.42	1.52	0.28	0.43
BM	(10)	1.41	0.14	-0.66	1.25	0.13	-0.59	1.26	0.10	-0.75	1.18	0.08	-0.74
	(LS)	-0.08	-0.12	-1.02	0.18	-0.08	-1.09	-0.24	-0.16	-1.17	-0.35	-0.20	-1.17
	(1)	0.69	0.19	0.25	0.72	0.21	0.27	0.88	0.19	0.14	0.91	0.21	0.17
Beta	(10)	1.62	0.07	-0.42	1.45	0.09	-0.26	1.56	0.08	-0.35	1.46	0.06	-0.35
	(LS)	0.94	-0.11	-0.67	0.73	-0.12	-0.53	0.68	-0.11	-0.49	0.55	-0.14	-0.52
	(1)	0.76	0.09	-0.02	0.74	0.11	0.10	0.96	0.12	-0.13	0.95	0.11	-0.13
CompEquIss	(10)	1.59	0.32	0.61	1.21	0.26	0.62	1.63	0.33	0.62	1.60	0.32	0.61
	(LS)	0.83	0.23	0.63	0.46	0.15	0.52	0.67	0.22	0.75	0.65	0.20	0.74
	(1)	1.12	0.39	0.93	0.87	0.31	0.88	1.14	0.39	0.93	1.13	0.39	0.92
DolVol	(10)	1.11	0.12	-0.78	1.19	0.16	-0.63	1.17	0.14	-0.74	1.08	0.11	-0.73
	(LS)	-0.01	-0.27	-1.72	0.32	-0.15	-1.50	0.03	-0.25	-1.66	-0.05	-0.27	-1.66
	(1)	0.69	0.04	-0.56	0.77	0.11	-0.28	0.70	0.07	-0.41	0.68	0.07	-0.41
GP	(10)	1.29	0.23	0.29	1.08	0.20	0.25	1.48	0.26	0.25	1.47	0.26	0.25
	(LS)	0.59	0.19	0.85	0.31	0.09	0.53	0.78	0.19	0.66	0.79	0.19	0.66
	(1)	0.74	0.12	-0.13	0.71	0.12	-0.08	0.80	0.13	-0.15	0.78	0.12	-0.14
IndMom	(10)	1.13	0.18	0.05	1.02	0.17	0.09	1.11	0.17	0.05	1.08	0.16	0.05
	(LS)	0.38	0.06	0.18	0.31	0.05	0.17	0.31	0.04	0.20	0.30	0.04	0.19
	(1)	1.23	0.21	0.11	0.94	0.16	0.20	1.12	0.20	0.14	1.01	0.17	0.12
InvestPPEInv	(10)	1.21	0.17	-0.23	1.17	0.20	-0.05	1.29	0.16	-0.42	1.34	0.16	-0.36
	(LS)	-0.02	-0.04	-0.34	0.23	0.04	-0.24	0.17	-0.04	-0.56	0.33	-0.01	-0.48
	(1)	0.95	0.19	0.35	0.85	0.17	0.37	0.94	0.18	0.35	0.95	0.19	0.34
Investment	(10)	1.72	0.21	0.08	1.39	0.21	0.13	1.67	0.21	0.08	1.68	0.21	0.08
	(LS)	0.78	0.03	-0.27	0.53	0.04	-0.23	0.73	0.03	-0.27	0.72	0.02	-0.26
	(1)	1.06	0.02	-1.10	0.88	0.01	-0.96	0.92	0.02	-1.04	0.85	0.00	-1.05
MaxRet	(10)	0.99	0.40	0.23	0.87	0.28	0.18	1.38	0.28	0.02	1.33	0.29	0.05
	(LS)	-0.08	0.38	1.32	-0.01	0.27	1.14	0.47	0.26	1.05	0.47	0.29	1.10
	(1)	0.50	-0.06	-1.35	0.38	-0.06	-1.13	0.57	-0.05	-1.33	0.32	-0.08	-1.33
Mom12m	(10)	2.21	0.39	0.48	1.76	0.33	0.52	2.28	0.39	0.49	2.27	0.39	0.48
	(LS)	1.71	0.45	1.84	1.38	0.39	1.65	1.71	0.44	1.82	1.96	0.47	1.81
	(1)	0.65	-0.01	1.01	0.86	0.04	-0.88	0.68	0.01	-1.01	0.72	0.01	-1.01
Mom1m	(10)	1.70	0.23	-0.09	1.37	0.18	-0.03	1.49	0.19	-0.08	1.59	0.21	-0.09
	(LS)	1.05	0.23	0.92	0.51	0.14	0.85	0.81	0.19	0.93	0.87	0.20	0.92
	(1)	0.65	-0.02	-1.31	0.45	-0.03	-1.14	0.67	0.01	-1.31	0.63	0.01	-1.31
Mom6m	(10)	2.21	0.35	0.36	1.70	0.28	0.31	2.12	0.33	0.27	2.11	0.34	0.26
	(LS)	1.56	0.37	1.68	1.25	0.31	1.46	1.45	0.33	1.58	1.48	0.33	1.57
	(1)	0.65	-0.02	-1.31	0.45	-0.03	-1.14	0.67	0.01	-1.31	0.63	0.01	-1.31
Mom6m	(10)	2.21	0.35	0.36	1.70	0.28	0.31	2.12	0.33	0.27	2.11	0.34	0.26
	(LS)	1.56	0.37	1.68	1.25	0.31	1.46	1.45	0.33	1.58	1.48	0.33	1.57
	(1)	1.09	0.08	-0.59	1.60	0.17	-0.61	1.10	0.11	-0.48	1.24	0.10	-0.52
OScore	(10)	0.82	0.19	0.71	0.91	0.22	0.62	1.57	0.19	-0.13	1.59	0.24	0.17
	(LS)	-0.28	0.11	1.30	-0.68	0.04	1.22	0.46	0.08	0.35	0.34	0.14	0.69
	(1)	1.07	0.17	0.13	0.91	0.15	0.18	1.27	0.17	0.03	1.22	0.13	-0.12
OperProf	(10)	1.20	0.34	0.73	0.95	0.24	0.56	1.21	0.32	0.67	1.19	0.28	0.54
	(LS)	0.13	0.16	0.60	0.04	0.09	0.38	-0.07	0.15	0.64	-0.02	0.15	0.66
	(1)	1.20	0.08	-1.01	1.33	0.13	-0.92	1.30	0.11	-1.05	1.34	0.11	-1.04
RET	(10)	1.25	0.21	0.32	0.85	0.08	-0.00	1.48	0.17	-0.07	1.37	0.16	-0.08
	(LS)	0.04	0.13	1.34	-0.48	-0.04	0.92	0.19	0.06	0.99	0.02	0.05	0.97
	(1)	1.20	0.08	-1.01	1.33	0.13	-0.92	1.30	0.11	-1.05	1.34	0.11	-1.04
RET	(10)	1.25	0.21	0.32	0.85	0.08	-0.00	1.48	0.17	-0.07	1.37	0.16	-0.08
	(LS)	0.04	0.13	1.34	-0.48	-0.04	0.92	0.19	0.06	0.99	0.02	0.05	0.97
	(1)	1.20	0.08	-1.01	1.09	0.09	-0.79	1.27	0.11	-0.93	1.33	0.11	-0.89
RoE	(10)	1.25	0.21	0.32	1.01	0.20	0.31	1.22	0.20	0.17	1.18	0.19	0.15
	(LS)	0.04	0.13	1.34	-0.09	0.11	1.10	-0.05	0.09	1.10	-0.15	0.07	1.04
	(1)	1.22	0.27	0.74	0.89	0.20	0.75	1.18	0.26	0.75	1.16	0.25	0.75
ShareIss5Y	(10)	1.20	0.23	0.10	0.98	0.21	0.22	1.24	0.14	-0.46	1.20	0.15	-0.26
	(LS)	-0.02	-0.04	-0.65	0.09	0.01	-0.53	0.06	-0.12	-1.21	0.04	-0.11	-1.01
	(1)	1.39	0.18	-0.19	1.48	0.17	-0.34	1.21	0.18	-0.16	1.35	0.17	-0.23
ShareVol	(10)	0.64	0.14	0.61	0.86	0.14	0.45	1.64	0.22	-0.19	1.63	0.23	-0.12
	(LS)	-0.75	-0.04	0.80	-0.62	-0.03	0.78	0.43	0.04	0.12	0.28	0.06	0.12
	(1)	1.06	0.15	0.10	0.87	0.14	0.20	0.84	0.12	0.12	0.86	0.12	0.10
dNoa	(10)	1.50	0.19	-0.28	1.23	0.20	-0.05	1.55	0.16	-0.49	1.72	0.20	-0.37
	(LS)	0.44	0.04	-0.39	0.36	0.06	-0.24	0.71	0.05	-0.61	0.86	0.08	-0.47
	(1)	0.95	-0.01	-1.77	1.05	0.04	-1.44	0.95	0.03	-1.31	0.99	0.03	-1.44
marketcap	(10)	1.18	0.51	1.01	1.15	0.42	0.92	1.18	0.45	0.97	1.18	0.45	0.96
	(LS)	0.22	0.52	2.78	0.10	0.38	2.36	0.23	0.42	2.28	0.19	0.43	2.40
	(1)	1.05	0.03	-0.92	1.04	0.07	-0.71	1.20	0.06	-0.85	1.12	0.06	-0.85
roaq	(10)	1.22	0.27	0.71	1.03	0.20	0.53	1.41	0.23	0.39	1.39	0.23	0.38
	(LS)	0.17	0.24	1.63	-0.01	0.13	1.24	0.21	0.18	1.24	0.27	0.17	1.23

Table 19: Long Short Portfolio For Different Configuration of ε (Epsilon) values and MinPts (M), for DBSCAN (taking into account the outlier cap of 10 procent).

Factor	Decile	N			ε = 4, M= 10			ε = 4, M= 25			ε = 4, M= 50		
		R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$
	(1)	1.05	0.14	0.05	1.19	0.16	0.03	1.13	0.15	0.03	1.14	0.15	0.03
Accruals	(10)	1.48	0.21	-0.23	1.34	0.18	-0.20	1.33	0.18	-0.20	1.36	0.19	-0.20
	(LS)	0.43	0.07	-0.28	0.15	0.02	-0.23	0.20	0.03	-0.23	0.22	0.03	-0.24
	(1)	1.28	0.21	0.17	1.02	0.15	0.19	1.00	0.14	0.19	1.04	0.15	0.21
AssetGrowth	(10)	1.30	0.14	-0.52	1.54	0.17	-0.59	1.46	0.16	-0.61	1.48	0.16	-0.65
	(LS)	0.02	-0.07	-0.70	0.52	0.02	-0.78	0.46	0.02	-0.80	0.44	0.01	-0.86
	(1)	1.49	0.25	0.36	1.51	0.27	0.43	1.52	0.27	0.42	1.50	0.27	0.43
BM	(10)	1.41	0.14	-0.66	1.09	0.07	-0.74	1.04	0.07	-0.74	1.04	0.07	-0.74
	(LS)	-0.08	-0.12	-1.02	-0.42	-0.20	-1.17	-0.48	-0.21	-1.16	-0.46	-0.20	-1.16
	(1)	0.69	0.19	0.25	0.88	0.20	0.17	0.88	0.20	0.17	0.91	0.20	0.17
Beta	(10)	1.62	0.07	-0.42	1.50	0.07	-0.35	1.50	0.07	-0.35	1.49	0.07	-0.36
	(LS)	0.94	-0.11	-0.67	0.62	-0.13	-0.52	0.62	-0.13	-0.53	0.58	-0.13	-0.53
	(1)	0.76	0.09	-0.02	0.98	0.12	-0.15	0.95	0.11	-0.16	0.77	0.08	-0.14
CompEquIss	(10)	1.59	0.32	0.61	1.61	0.32	0.61	1.58	0.32	0.62	1.59	0.32	0.62
	(LS)	0.83	0.23	0.63	0.63	0.20	0.77	0.62	0.20	0.78	0.82	0.23	0.76
	(1)	1.12	0.39	0.93	1.12	0.38	0.92	1.12	0.38	0.92	1.12	0.39	0.93
DolVol	(10)	1.11	0.12	-0.78	1.16	0.14	-0.74	1.17	0.14	-0.74	1.18	0.14	-0.74
	(LS)	-0.01	-0.27	-1.72	0.04	-0.25	-1.66	0.06	-0.25	-1.67	0.06	-0.25	-1.67
	(1)	0.69	0.04	-0.56	0.70	0.07	-0.41	0.72	0.07	-0.41	0.71	0.07	-0.41
GP	(10)	1.29	0.23	0.29	1.43	0.25	0.25	1.45	0.25	0.25	1.42	0.25	0.27
	(LS)	0.59	0.19	0.85	0.73	0.18	0.66	0.73	0.18	0.67	0.71	0.18	0.68
	(1)	0.74	0.12	-0.13	0.80	0.13	-0.14	0.81	0.13	-0.14	0.80	0.13	-0.14
IndMom	(10)	1.13	0.18	0.05	1.07	0.17	0.05	1.12	0.17	0.05	1.09	0.17	0.05
	(LS)	0.38	0.06	0.18	0.27	0.04	0.19	0.30	0.04	0.20	0.29	0.04	0.19
	(1)	1.23	0.21	0.11	1.02	0.17	0.12	1.07	0.18	0.13	1.08	0.18	0.14
InvestPPEInv	(10)	1.21	0.17	-0.23	1.29	0.15	-0.37	1.33	0.16	-0.40	1.32	0.16	-0.43
	(LS)	-0.02	-0.04	-0.34	0.26	-0.02	-0.50	0.27	-0.02	-0.53	0.24	-0.02	-0.57
	(1)	0.95	0.19	0.35	0.97	0.19	0.35	0.95	0.19	0.35	0.94	0.18	0.35
Investment	(10)	1.72	0.21	0.08	1.70	0.22	0.08	1.69	0.21	0.08	1.67	0.21	0.08
	(LS)	0.78	0.03	-0.27	0.73	0.02	-0.27	0.74	0.03	-0.27	0.72	0.03	-0.27
	(1)	1.06	0.02	-1.10	0.87	0.01	-1.05	0.88	0.01	-1.05	0.90	0.01	-1.05
MaxRet	(10)	0.99	0.40	0.23	1.29	0.29	0.06	1.31	0.30	0.06	1.27	0.29	0.05
	(LS)	-0.08	0.38	1.32	0.43	0.29	1.11	0.43	0.29	1.11	0.38	0.27	1.10
	(1)	0.50	-0.06	-1.35	0.38	-0.07	-1.33	0.42	-0.07	-1.33	0.47	-0.06	-1.34
Mom12m	(10)	2.21	0.39	0.48	2.28	0.39	0.48	2.27	0.39	0.49	2.28	0.39	0.49
	(LS)	1.71	0.45	1.84	1.90	0.46	1.81	1.85	0.45	1.82	1.80	0.45	1.83
	(1)	0.65	-.01	1.01	0.78	0.02	-1.01	0.75	0.02	-1.01	0.79	0.02	-1.01
Mom1m	(10)	1.70	0.23	-0.09	1.61	0.22	-0.09	1.60	0.21	-0.09	1.62	0.22	-0.08
	(LS)	1.05	0.23	0.92	0.84	0.19	0.92	0.85	0.19	0.92	0.84	0.19	0.93
	(1)	0.65	-0.02	-1.31	0.66	0.01	-1.31	0.63	0.01	-1.32	0.62	0.00	-1.33
Mom6m	(10)	2.21	0.35	0.36	2.14	0.34	0.26	2.13	0.34	0.26	2.08	0.33	0.27
	(LS)	1.56	0.37	1.68	1.48	0.33	1.58	1.50	0.33	1.58	1.46	0.33	1.59
	(1)	1.09	0.08	-0.59	1.33	0.11	-0.52	1.31	0.11	-0.52	1.27	0.11	-0.52
OScore	(10)	0.82	0.19	0.71	1.50	0.23	0.14	1.49	0.22	0.09	1.46	0.21	0.05
	(LS)	-0.28	0.11	1.30	0.17	0.11	0.66	0.17	0.11	0.62	0.19	0.11	0.57
	(1)	1.07	0.17	0.13	1.25	0.14	-0.12	1.27	0.15	-0.12	1.29	0.16	-0.09
OperProf	(10)	1.20	0.34	0.73	1.22	0.30	0.55	1.21	0.31	0.61	1.18	0.31	0.66
	(LS)	0.13	0.16	0.60	-0.03	0.16	0.68	-0.06	0.16	0.72	-0.11	0.15	0.75
	(1)	1.20	0.08	-1.01	1.40	0.12	-0.89	1.35	0.13	-0.89	1.36	0.12	-0.91
RoE	(10)	1.25	0.21	0.32	1.17	0.19	0.15	1.19	0.19	0.17	1.16	0.19	0.18
	(LS)	0.04	0.13	1.34	-0.24	0.06	1.04	-0.15	0.07	1.06	-0.20	0.07	1.09
	(1)	1.22	0.27	0.74	1.18	0.26	0.75	1.17	0.26	0.75	1.16	0.25	0.75
ShareIss5Y	(10)	1.20	0.23	0.10	1.18	0.15	-0.31	1.22	0.16	-0.37	1.17	0.15	-0.42
	(LS)	-0.02	-0.04	-0.65	0.00	-0.11	-1.06	0.06	-0.10	-1.12	0.01	-0.11	-1.17
	(1)	1.39	0.18	-0.19	1.45	0.19	-0.23	1.48	0.21	-0.20	1.54	0.23	-0.16
ShareVol	(10)	0.64	0.14	0.61	1.59	0.23	-0.13	1.53	0.21	-0.19	1.42	0.19	-0.27
	(LS)	-0.75	-0.04	0.80	0.15	0.04	0.10	0.05	0.01	0.01	-0.11	-0.04	-0.10
	(1)	1.06	0.15	0.10	0.82	0.11	0.10	0.84	0.12	0.11	0.86	0.12	0.12
dNoa	(10)	1.50	0.19	-0.28	1.64	0.20	-0.38	1.64	0.19	-0.41	1.54	0.18	-0.47
	(LS)	0.44	0.04	-0.39	0.82	0.09	-0.48	0.80	0.08	-0.52	0.68	0.06	-0.59
	(1)	0.95	-0.01	-1.77	0.83	-0.00	-1.45	0.81	-0.01	-1.43	0.77	-0.02	-1.41
marketcap	(10)	1.18	0.51	1.01	1.18	0.45	0.96	1.18	0.46	0.96	1.18	0.48	1.00
	(LS)	0.22	0.52	2.78	0.35	0.45	2.41	0.37	0.46	2.39	0.41	0.50	2.40
	(1)	1.05	0.03	-0.92	1.12	0.05	-0.85	1.09	0.04	-0.85	1.15	0.05	-0.86
roaq	(10)	1.22	0.27	0.71	1.35	0.22	0.38	1.31	0.21	0.39	1.33	0.22	0.38
	(LS)	0.17	0.24	1.63	0.23	0.17	1.23	0.22	0.17	1.24	0.19	0.17	1.24

Table 20: Long Short Portfolio For Different Configuration of ε (Epsilon) values and MinPts (M), for DBSCAN (taking into account the outlier cap of 10 procent).

C.2.4 *Comparing Across Different Cluster Techniques*

Factor	Decile	N			K = 10			DMax = 80)			ε = 4, M= 50			OLS		
		R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$	R	SR	$FF5^{\alpha}$
	(1)	1.05	0.14	0.05	1.05	0.13	-0.01	1.07	0.14	0.00	1.14	0.15	0.03	1.24	0.16	-0.01
Accruals	(10)	1.48	0.21	-0.23	1.33	0.18	-0.17	1.24	0.16	-0.18	1.36	0.19	-0.20	1.06	0.14	-0.18
	(LS)	0.43	0.07	-0.28	0.28	0.05	-0.16	0.17	0.02	-0.18	0.22	0.03	-0.24	-0.18	-0.02	-0.17
	(1)	1.28	0.21	0.17	1.21	0.18	0.17	1.23	0.18	0.17	1.04	0.15	0.21	1.51	0.24	0.14
AssetGrowth	(10)	1.30	0.14	-0.52	1.15	0.12	-0.38	1.03	0.10	-0.43	1.48	0.16	-0.65	1.13	0.12	-0.22
	(LS)	0.02	-0.07	-0.70	-0.06	-0.06	-0.55	-0.20	-0.08	-0.59	0.44	0.01	-0.86	-0.39	-0.12	-0.36
	(1)	1.49	0.25	0.36	1.40	0.25	0.33	1.41	0.25	0.40	1.50	0.27	0.43	1.31	0.20	-0.05
BM	(10)	1.41	0.14	-0.66	1.48	0.18	-0.64	1.52	0.18	-0.69	1.04	0.07	-0.74	1.13	0.14	-0.40
	(LS)	-0.08	-0.12	-1.02	0.08	-0.06	-0.97	0.11	-0.07	-1.09	-0.46	-0.20	-1.16	-0.19	-0.06	-0.34
	(1)	0.69	0.19	0.25	0.79	0.16	0.04	0.72	0.16	0.10	0.91	0.20	0.17	0.85	0.13	-0.14
Beta	(10)	1.62	0.07	-0.42	1.43	0.09	-0.32	1.62	0.11	-0.34	1.49	0.07	-0.36	1.75	0.17	-0.31
	(LS)	0.94	-0.11	-0.67	0.64	-0.06	-0.36	0.90	-0.05	-0.44	0.58	-0.13	-0.53	0.90	0.04	-0.17
	(1)	0.76	0.09	-0.02	1.23	0.18	0.12	1.01	0.14	0.05	0.77	0.08	-0.14	1.26	0.20	0.23
CompEquIss	(10)	1.59	0.32	0.61	1.57	0.31	0.56	1.59	0.31	0.58	1.59	0.32	0.62	1.43	0.22	0.28
	(LS)	0.83	0.23	0.63	0.35	0.13	0.44	0.57	0.17	0.53	0.82	0.23	0.76	0.17	0.02	0.04
	(1)	1.12	0.39	0.93	0.94	0.24	0.83	0.95	0.28	0.88	1.12	0.39	0.93	0.93	0.18	0.53
DolVol	(10)	1.11	0.12	-0.78	1.24	0.16	-0.58	1.24	0.15	-0.63	1.18	0.14	-0.74	1.69	0.28	-0.42
	(LS)	-0.01	-0.27	-1.72	0.30	-0.08	-1.41	0.29	-0.13	-1.50	0.06	-0.25	-1.67	0.76	0.10	-0.95
	(1)	0.69	0.04	-0.56	0.84	0.09	-0.37	0.81	0.08	-0.41	0.71	0.07	-0.41	1.00	0.10	-0.26
GP	(10)	1.29	0.23	0.29	1.23	0.20	0.18	1.23	0.20	0.20	1.42	0.25	0.27	1.19	0.18	0.08
	(LS)	0.59	0.19	0.85	0.39	0.11	0.56	0.42	0.11	0.61	0.71	0.18	0.68	0.19	0.08	0.34
	(1)	0.74	0.12	-0.13	0.96	0.16	-0.11	0.86	0.15	-0.13	0.80	0.13	-0.14	1.03	0.18	-0.11
IndMom	(10)	1.13	0.18	0.05	1.07	0.16	-0.01	1.15	0.18	0.04	1.09	0.17	0.05	1.01	0.15	-0.09
	(LS)	0.38	0.06	0.18	0.11	0.00	0.10	0.30	0.03	0.17	0.29	0.04	0.19	-0.01	-0.02	0.02
	(1)	1.23	0.21	0.11	1.20	0.20	0.08	1.25	0.21	0.10	1.08	0.18	0.14	1.56	0.25	-0.05
InvestPPEInv	(10)	1.21	0.17	-0.23	1.13	0.15	-0.12	1.23	0.18	-0.16	1.32	0.16	-0.43	0.99	0.14	0.00
	(LS)	-0.02	-0.04	-0.34	-0.07	-0.04	-0.21	-0.02	-0.03	-0.25	0.24	-0.02	-0.57	-0.57	-0.11	0.05
	(1)	0.95	0.19	0.35	0.94	0.18	0.33	0.89	0.17	0.34	0.94	0.18	0.35	0.99	0.19	0.31
Investment	(10)	1.72	0.21	0.08	1.61	0.22	0.09	1.58	0.20	0.09	1.67	0.21	0.08	1.71	0.23	0.09
	(LS)	0.78	0.03	-0.27	0.67	0.04	-0.24	0.69	0.03	-0.24	0.72	0.03	-0.27	0.72	0.05	-0.22
	(1)	1.06	0.02	-1.10	1.08	0.06	-0.96	0.94	0.05	-1.00	0.90	0.01	-1.05	1.37	0.12	-0.78
MaxRet	(10)	0.99	0.40	0.23	1.40	0.27	-0.10	1.17	0.26	-0.01	1.27	0.29	0.05	1.47	0.18	-0.43
	(LS)	-0.08	0.38	1.32	0.33	0.21	0.86	0.23	0.21	0.99	0.38	0.27	1.10	0.10	0.06	0.35
	(1)	0.50	-0.06	-1.35	0.80	0.01	-0.95	0.59	-0.04	-1.22	0.47	-0.06	-1.34	0.99	0.08	-0.46
Mom12m	(10)	2.21	0.39	0.48	2.16	0.36	0.40	2.12	0.36	0.45	2.28	0.39	0.49	2.45	0.39	0.03
	(LS)	1.71	0.45	1.84	1.36	0.35	1.35	1.54	0.40	1.67	1.80	0.45	1.83	1.45	0.31	0.49
	(1)	0.65	-.01	1.01	0.81	0.03	-0.87	0.69	0.01	-0.96	0.79	0.02	-1.01	1.44	0.13	-0.75
Mom1m	(10)	1.70	0.23	-0.09	1.80	0.24	-0.19	1.58	0.21	-0.12	1.62	0.22	-0.08	1.26	0.13	-0.30
	(LS)	1.05	0.23	0.92	0.99	0.21	0.69	0.89	0.20	0.84	0.84	0.19	0.93	-0.19	0.00	0.46
	(1)	0.65	-0.02	-1.31	1.14	0.08	-1.02	0.76	0.01	-1.22	0.62	0.00	-1.33	2.01	0.22	-0.68
Mom6m	(10)	2.21	0.35	0.36	2.30	0.36	0.14	2.34	0.37	0.23	2.08	0.33	0.27	1.71	0.22	-0.15
	(LS)	1.56	0.37	1.68	1.16	0.28	1.16	1.58	0.36	1.45	1.46	0.33	1.59	-0.30	-0.01	0.53
	(1)	1.09	0.08	-0.59	1.39	0.22	-0.09	1.55	0.25	-0.14	1.27	0.11	-0.52	1.03	0.14	-0.24
OScore	(10)	0.82	0.19	0.71	1.04	0.06	-0.62	1.17	0.10	-0.25	1.46	0.21	0.05	1.48	0.10	-0.86
	(LS)	-0.28	0.11	1.30	-0.35	-0.17	-0.53	-0.39	-0.14	-0.12	0.19	0.11	0.57	0.45	-0.04	-0.62
	(1)	1.07	0.17	0.13	1.02	0.16	0.18	1.04	0.15	0.09	1.29	0.16	-0.09	1.51	0.22	0.01
OperProf	(10)	1.20	0.34	0.73	1.17	0.23	0.38	1.17	0.25	0.39	1.18	0.31	0.66	1.16	0.15	-0.34
	(LS)	0.13	0.16	0.60	0.15	0.08	0.20	0.14	0.10	0.30	-0.11	0.15	0.75	-0.35	-0.07	-0.35
	(1)	1.20	0.08	-1.01	1.17	0.09	-0.87	1.21	0.09	-0.90	1.36	0.12	-0.91	1.24	0.12	-0.72
RoE	(10)	1.25	0.21	0.32	1.42	0.19	0.00	1.22	0.18	0.07	1.16	0.19	0.18	1.35	0.18	-0.11
	(LS)	0.04	0.13	1.34	0.25	0.10	0.88	0.01	0.08	0.97	-0.20	0.07	1.09	0.11	0.06	0.61
	(1)	1.22	0.27	0.74	1.13	0.23	0.69	1.19	0.25	0.71	1.16	0.25	0.75	1.13	0.20	0.46
ShareIss5Y	(10)	1.20	0.23	0.10	1.19	0.21	0.30	1.25	0.24	0.20	1.17	0.15	-0.42	1.09	0.19	0.50
	(LS)	-0.02	-0.04	-0.65	0.06	-0.03	-0.39	0.06	-0.01	-0.51	0.01	-0.11	-1.17	-0.04	-0.01	0.04
	(1)	1.39	0.18	-0.19	1.81	0.33	-0.05	1.73	0.30	-0.04	1.54	0.23	-0.16	1.69	0.27	-0.22
ShareVol	(10)	0.64	0.14	0.61	1.22	0.17	-0.03	1.26	0.16	0.04	1.42	0.19	-0.27	1.26	0.16	0.08
	(LS)	-0.75	-0.04	0.80	-0.59	-0.16	0.02	-0.34	-0.10	0.08	-0.11	-0.04	-0.10	-0.42	-0.11	0.30
	(1)	1.06	0.15	0.10	1.03	0.14	0.06	1.01	0.14	0.07	0.86	0.12	0.12	0.83	0.10	-0.15
dNoa	(10)	1.50	0.19	-0.28	1.38	0.18	-0.15	1.20	0.15	-0.19	1.54	0.18	-0.47	1.36	0.19	0.12
	(LS)	0.44	0.04	-0.39	0.35	0.04	-0.21	0.19	0.01	-0.26	0.68	0.06	-0.59	0.53	0.09	0.27
	(1)	0.95	-0.01	-1.77	0.98	0.20	0.42	0.89	0.12	0.03	0.77	-0.02	-1.41	0.72	0.16	0.74
marketcap	(10)	1.18	0.51	1.01	1.19	0.43	1.01	1.19	0.42	1.03	1.18	0.48	1.00	1.25	0.19	-0.43
	(LS)	0.22	0.52	2.78	0.21	0.23	0.59	0.31	0.30	1.00	0.41	0.50	2.40	0.53	0.02	-1.17
	(1)	1.05	0.03	-0.92	1.23	0.09	-0.74	1.07	0.05	-0.80	1.15	0.05	-0.86	1.18	0.12	-0.57
roaq	(10)	1.22	0.27	0.71	1.25	0.17	0.22	1.33	0.20	0.38	1.33	0.22	0.38	1.50	0.17	-0.17
	(LS)	0.17	0.24	1.63	0.03	0.09	0.96	0.26	0.15	1.18	0.19	0.17	1.24	0.31	0.05	0.40

Table 21: Comparing Across Different Cluster Techniques For The Long Short Portfolio, and OLS regression.