Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Systematic Clustering and Prediction of S&P 500 Stocks

By

Lu Zhou, Gonzalo González, Giovanni Battistella, Ángel Echaide and Laxmi Prabhu Kundapur

MSc. Business Analytics

Module Leader: Nicholas Danks

Trinity Business School
TRINITY COLLEGE
UNIVERSITY OF DUBLIN

November 2024

Contents

1 Executive Summary **1**

2 Problem Description **2**
2.1 Business Objective 2
2.2 Data Mining Goal 2

3 Data Source, Description and Preparation **3**

4 Data Mining **3**
4.1 Determining the Optimal Number of Clusters 3
4.2 Stock Classification Model 5
4.3 Prediction 6

5 Conclusions **7**

A1 Appendix **9**
A1.1 AI Declaration 9
A1.2 Code 9
A1.3 Tables and Images 9

1 Executive Summary

Making the right stock choices is a challenge for all market operators, especially with the vast number of stocks to consider and the constant changes in the market. Operators often have different goals and risk preferences, and sorting through all available options to find the right fit can be overwhelming and time-consuming. Without an efficient way to categorize and assess stocks, operators risk making decisions that don't align with their objectives.

To address this challenge, a model was developed using historical data from the S&P 500 to group stocks based on their risk and return characteristics. The goal was to create a simple, flexible framework that helps operators quickly identify stocks that match different investor profiles.

Three groups in the index were stocks with high return and moderate volatility, suitable for aggressive profile investors, stocks with low return and high volatility, and stocks with moderate return and low volatility, recommended for conservative profile investors.

The model classifies stocks into distinct categories based on their risk levels and return potential, making it easier for operators to match stocks with the right profiles. To ensure consistency and reliability, the model follows a clear, structured approach to evaluating and classifying stocks.

The model has also been tested on stocks from the NASDAQ 100, demonstrating that it can be applied across different markets. It is recommended for investors with conservative profile to invest in Coca-Cola Europacific Partners and AstraZeneca, meanwhile, aggressive profile investors are recommended to invest in Arm Holdings.
By simplifying the stock selection process, the model saves time and enables more informed decisions. It ensures that stock choices align with client goals while offering flexibility to adapt to changing market conditions. This approach helps better meet the diverse needs of clients and improves the overall efficiency of investment strategies.

2 Problem Description

Market operators face significant challenges in aligning stock selection with specific risk-return profiles, such as conservative, balanced, and aggressive strategies. The complexity of analyzing and categorizing stocks stems from market noise, systemic co-movements, and the difficulty of identifying stable clusters that remain consistent over time (Zema et al. (2021); Nagy & Ormos (2018)). Without systematic methods, the risk of inefficiencies, such as misaligned benchmarks or poorly optimized portfolios, increases (Nanda et al. (2010) ; Gharanchaei & Panda (2023)). This highlights the opportunity to develop a simple and practical system that can streamline stock classification and enhance decision-making processes.

2.1 Business Objective

The business objective is to develop a data-driven approach that categorizes stocks based on their risk-return characteristics, using patterns observed within the S&P 500 index to identify groups that align with established investor profiles. This will enable the firm to more quickly and accurately classify other stocks, supporting streamlined investment decisions that align with strategic risk-return objectives. Key Business Questions:

- How can we identify groups of stocks that naturally correspond to investor profiles like conservative, balanced, and aggressive?
- Will this categorization method make stock analysis faster and more consistent, helping align with the company's investment strategies?
- What are the benefits of using a classification system for stock assessment?

2.2 Data Mining Goal

To achieve the business objective, we will apply two primary data mining techniques: clustering and classification tree. First, clustering analysis using the K-Means algorithm will be employed to identify natural groupings of stocks within the S&P 500 based on risk-return characteristics. Next, a classification model, utilizing decision trees, will be developed to assign stocks to these clusters based on features such as volatility and mean return. This combined approach ensures a systematic and efficient categorization process, enabling the identification of stock groups aligned with investor profiles while supporting streamlined and scalable investment decision-making. The R code can be found following this **link**.

3 Data Source, Description and Preparation

To set up the analysis, we followed a well-defined data preparation process, which is key to ensuring the accuracy and reliability of any analysis. The process involved the following steps:

- **Data Collection:** We used the "Tidyquant" package in R to obtain the historical stock prices for the S&P 500 index (Iglewicz & Naik (2020)), which includes 500 of the largest companies that are traded on stock exchanges in the United States over the specified time period, from 1 September 2021 to 1 September 2024.
- **Data Description:** The original dataset contains 375,134 entities and 8 columns, representing daily stock data for the S&P 500 companies. Key variables include stock symbol (identifying each stock), date (the trading day), and financial metrics such as open, high, low, close, and adjusted prices, along with volume (the number of shares traded). These features provide insight into daily price movements, volatility, and liquidity. Raw data sample available in Table A1.1.
- **Data Cleaning:** We focused on retaining the most relevant columns: stock symbol (ticker), trading date, trading volume, and adjusted closing price. This process eliminated unnecessary columns (open, high, low, close), simplifying the dataset and facilitating easier analysis and subsequent data processing tasks.
- **Data Transformation:** A logarithmic transformation was applied to the daily prices, and the first difference was computed to capture the daily returns. Results can be found in Table A1.2.
- **Preprocessing:** We compiled the preprocessed dataset, which includes the mean and standard deviation of daily returns for each stock.

4 Data Mining

4.1 Determining the Optimal Number of Clusters

Before applying k-means clustering, it was necessary to determine the optimal number of clusters to best group the stocks based on return and volatility behavior. We used two key methods:

1. **Elbow Method:** We ran k-means clustering for a range of k values (from 1 to 10) and plotted the within-cluster sum of squares (WSS) against k. The "elbow" in the plot occurs at k = 3, where adding more clusters no longer significantly reduces the WSS. This indicates that three clusters offer the best balance between simplicity and clustering performance, as further increases in k do not yield substantial improvements, as shown in Figure A1.1.

2. **Silhouette Score:** We conducted a Silhouette Score analysis (in Figure A1.2) to determine the optimal number of clusters for our dataset, testing values of k from 2 to 10. The results show that k = 3 yields the highest Silhouette Score, indicating that three clusters provide the best separation and cohesion.

The evaluation of clustering metrics indicates that k = 3 is the optimal number of clusters for our project. By using both the Elbow Method and the silhouette score, we can effectively analyze the trade-offs between how compact the clusters are and how well-separated they are from each other. The insights gained from these methods strongly support the choice of "K = 3", as this configuration not only maximizes the silhouette score but also shows a clear inflection point in the Elbow Method, leading to a balanced and effective clustering solution.

After determining the optimal number of clusters, we carried out a k-means analysis and visualized the results using a scatter plot in Figure A1.3.

This plot illustrates the results of the cluster analysis, showing the distribution of stocks based on their mean return and volatility. Each cluster is visually represented by distinct colors, helping to highlight the different groups of stocks with similar characteristics.

The X-axis represents the average return of each stock. Stocks positioned further to the right have higher average returns, while those on the left have lower or negative returns.

The Y-axis indicates the volatility of each stock. Stocks higher up on this axis are more volatile, meaning they carry more risk, while those lower down are more stable.

Here is a detailed analysis of the characteristics of the three identified clusters:

Cluster Number	Cluster 1	Cluster 2	Cluster 3
Number of Stocks	73	119	310
Mean Return	Around 1.25	Negative return, around -1	Mostly around 0
Volatility	Standard Deviation is between -1.5 and 2	Standard Deviation is between 0 and 3	Standard Deviation is close to -1 to 0
Risk Level	Moderate to High Risk, Positive Return	High Risk, Negative Return	Low Risk, Stable Performance

Table 4.1: Cluster Characteristics

The three clusters represent different types of stocks based on their risk and return characteristics. Cluster 1 (Blue) includes stocks with high returns and moderate to high volatility, which are typically aggressive investments, offering the potential for higher rewards but also carrying more risk. Cluster 2 (Red) contains stocks with low or negative returns and higher volatility, making them poor performers that are also high-risk, low-return options. Finally, Cluster 3 (Green) has stocks with moderate to low returns and low volatility, which are usually conservative investments, offering stability with lower risk but also more modest returns.

The following graph presents the probability density distributions of returns for three clusters of S&P 500 stocks, categorized based on mean return and volatility. Each cluster has a distinct distribution shape:

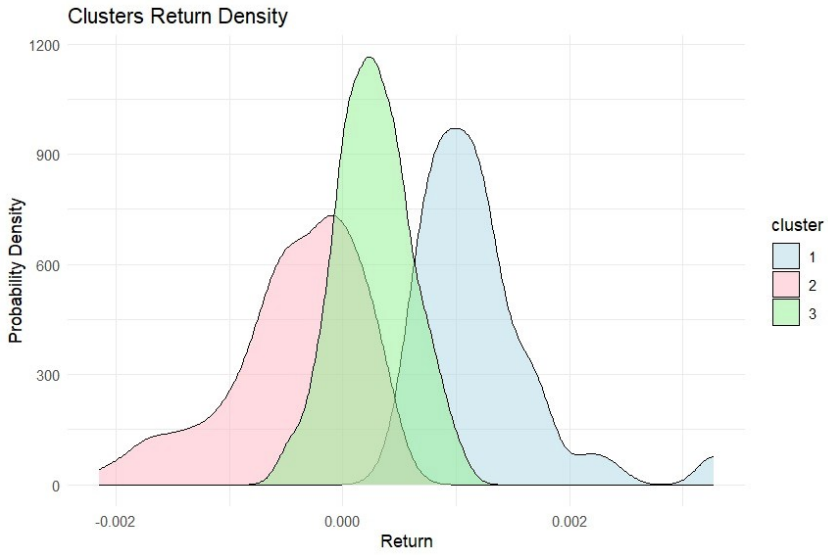

Figure 4.1: Density plot of 3 clusters

The X-axis represents the daily returns of the stocks. Stocks on the right have higher returns, while those on the left have lower or negative returns.

The Y-axis shows the probability density of returns, indicating how often specific returns appear within each cluster. A higher peak means those returns are more common within the cluster.

Volatility, inferred from the spread of the distribution, indicates risk. A wider distribution suggests higher volatility (more risk), while a narrower distribution indicates lower volatility (more stability).

- **Cluster 1:** The distribution is shifted to the right, suggesting higher returns with a wider spread, which implies higher volatility (more fluctuation in returns). These stocks are likely to be aggressive investments with higher potential rewards but come with higher risk.
- **Cluster 2:** This cluster has a narrower distribution skewed to the left, indicating negative returns with moderate to high volatility. Stocks here tend to be high-risk, low-return investments with a higher chance of poor performance.
- **Cluster 3:** The distribution is centered around near-zero returns, suggesting moderate to stable returns with a narrower spread. This indicates low volatility, representing conservative investments with lower risk but more modest returns.

4.2 Stock Classification Model

The next step was to use the established clusters to categorize new stocks, providing investors with a consistent way to evaluate risk and return. We applied two models:

- K-Nearest Neighbors (KNN) model, which tested values of k from 1 to 20 and achieved its highest accuracy of 99% at k = 4 (values at Table A1.3).

- A decision Tree, which classified stocks based on volatility and returns with an accuracy of 95%.

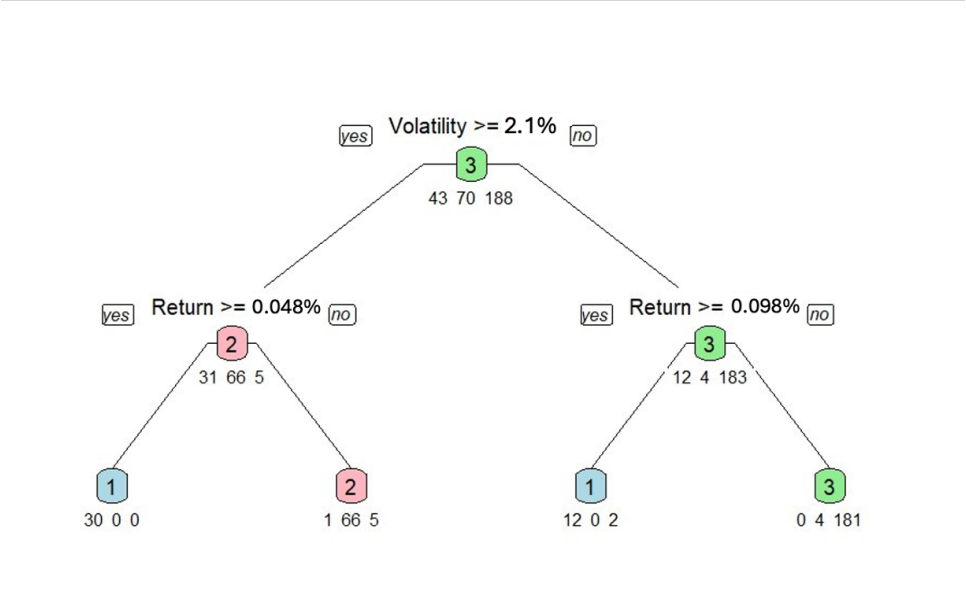

Figure 4.2: Decision Tree

Although the decision tree had slightly lower accuracy, its interpretability and visual clarity make it a more practical option (James et al. 2013, p. 24-26).

This decision tree offers a structured approach to classifying stocks based on two key financial indicators: volatility and return. If a stock's volatility exceeds this level, it indicates higher risk, so the model then examines the return. If the daily return is above 0.048%, the stock is classified as class 1, suggesting it has a favorable risk-reward balance. If the daily return is lower, it falls into class 2, indicating a riskier investment with potentially lower gains. For stocks with volatility below 2.1%, the model applies a stricter daily return threshold of 0.098%. If the return exceeds this, it is again classified as class 1, indicating stability with reasonable returns. However, if the daily return is below this threshold, the stock is categorized as class 3, representing a low-risk, low-return investment.

The model provides clear and intuitive decision pathways that investors can understand and use effectively. It's a reliable method for classifying companies, providing helpful insights into their risk and return profiles and predicting how they might be categorized in the S&P 500 and other markets.

4.3 Prediction

In the final phase of the project, we will use our trained decision tree model to classify 16 selected companies from the NASDAQ 100 (not included in the S&P 500) based on their volatility and mean return, identifying their investment profile clusters. After gathering historical stock data for these companies and calculating their mean returns and volatility, we will input these metrics into the model to predict their cluster assignments.

The results in Figure A1.4 show that most companies fall into Cluster 2, indicating high

risk with negative returns, including firms like Zscaler, Workday, The Trade Desk, and MercadoLibre. Cluster 1, represented solely by Arm Holdings, shows moderate to high risk but with positive returns, suggesting potential reward despite some volatility. Meanwhile, Cluster 3 consists of stable, low-risk stocks like Coca-Cola Europacific Partners and AstraZeneca, offering steady but modest performance.

5 Conclusions

This model gives us a great way to recommend stocks based on different investor styles. For example, investors who are comfortable taking risks might find Cluster 1 stocks appealing because they offer the potential for higher returns, despite their volatility. On the other hand, Cluster 3 stocks are more suited for conservative investors who prefer stability and lower risk. By clearly defining which stocks fit these profiles, we can help investors make better decisions that align with their financial goals.

The model's key strength lies in its ability to classify companies based on average return and volatility, making it particularly effective for single-stock investments. By grouping stocks into clusters that reflect different risk-return profiles, the model allows investors to make informed decisions tailored to their individual preferences, such as prioritizing high-return, high-volatility stocks or seeking out more stable, lower-risk options. This straightforward approach simplifies stock selection, making it especially useful for investors focused on individual stock performance.

However, the model has some limitations. One major weakness is the relatively short time frame of only three years used in the analysis, which may not fully capture long-term trends. Over a longer period, the average return might be more favorable, especially considering that stock markets tend to recover and grow over time, potentially altering the clusters. Additionally, the model is not designed for portfolio optimization, as it does not account for the correlations between stocks. This makes it less useful for investors who want to optimize a diversified portfolio, where understanding how stocks interact with one another is crucial for managing risk and maximizing returns.

As for recommendations for future developments, the model could be improved by analysing different timespans to capture both short and long-term fluctuations, offering a more comprehensive understanding of market dynamics. Additionally, training it on other indices, such as *Dow Jones*, would help validate its versatility and uncover differences across various markets.

Such improvements would significantly increase the model's practicality and relevance, particularly in complex and interconnected financial markets.

Bibliography

Gharanchaei, M. & Panda, P. (2023), 'Constructing an investment fund through stock clustering and integer programming', *ArXiv* .
URL: *http://arxiv.org/abs/2407.05912*

Iglewicz, B. & Naik, S. (2020), 'Core functions in tidyquant'. Accessed 8 November 2024.
URL: *https://cran.r-project.org/web/packages/tidyquant/vignettes/TQ01-core-functions-in-tidyquant.html*

James, G., Hastie, T., Witten, D. & Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R*, Springer, New York.
URL: *https://www.statlearning.com/*

Nagy, L. & Ormos, M. (2018), 'Friendship of stock market indices: A cluster-based investigation of stock markets', *Journal of Risk and Financial Management* **11**(4), 88.

Nanda, S., Mahanty, B. & Tiwari, M. (2010), 'Clustering indian stock market data for portfolio management', *Expert Systems with Applications* **37**(12), 8793–8798.

Zema, S., Fagiolo, G., Squartini, T. & Garlaschelli, D. (2021), 'Mesoscopic structure of the stock market and portfolio optimization', *ArXiv* .
URL: *http://arxiv.org/abs/2112.06544*

A1 Appendix

A1.1 AI Declaration

In this project, we used Generative AI tools, including ChatGPT, to improve different parts of our work. AI helped make the process faster, more accurate, and improved the overall quality. Here's how we used AI:

- Sentence Rephrasing: AI helped us rephrase sentences in our report to make them clearer and more consistent, while keeping the original meaning of our findings.
- R Code Review and Debugging: AI reviewed our R code, pointed out errors, and suggested fixes, making the debugging process easier and more efficient.
- Data Transformation Guidance: While we handled data cleaning and manipulation in R, AI gave us helpful tips and generated code for specific tasks, speeding up parts of the transformation process.
- LaTeX Support: We used LaTeX to write the report for a professional look. AI helped us write and troubleshoot the LaTeX code, making formatting and compiling quicker.
- Tidyquant Integration: AI supported us in using the tidyquant package in R, helping us connect to APIs and get the financial data we needed for the project.

A1.2 Code

R Code for the develpment of the project

A1.3 Tables and Images

Symbol	Date	Open	High	Low	Close	Volume	Adjusted
AAPL	01/09/2021	152.830002	154.979996	152.339996	152.509995	80313700	149.833328
AAPL	02/09/2021	153.869995	154.720001	152.399994	153.649994	71115500	150.953354
AAPL	03/09/2021	153.759995	154.630005	153.089996	154.300003	57808700	151.591949
AAPL	07/09/2021	154.970001	157.259995	154.389999	156.690002	82278300	153.940018
AAPL	08/09/2021	156.979996	157.039993	153.979996	155.110001	74420200	152.387711
AAPL	09/09/2021	155.490005	156.110001	153.949997	154.070007	57305700	151.365982

Table A1.1: Raw Data Sample

Symbol	Mean (Return)	Standard Deviation (Volatility)
A	-0.000246	0.018558
AAPL	0.000562	0.017301
ABBV	0.000899	0.012947
ABNB	-0.000384	0.030660
ABT	-0.000083	0.013862

Table A1.2: Transformed Data Sample

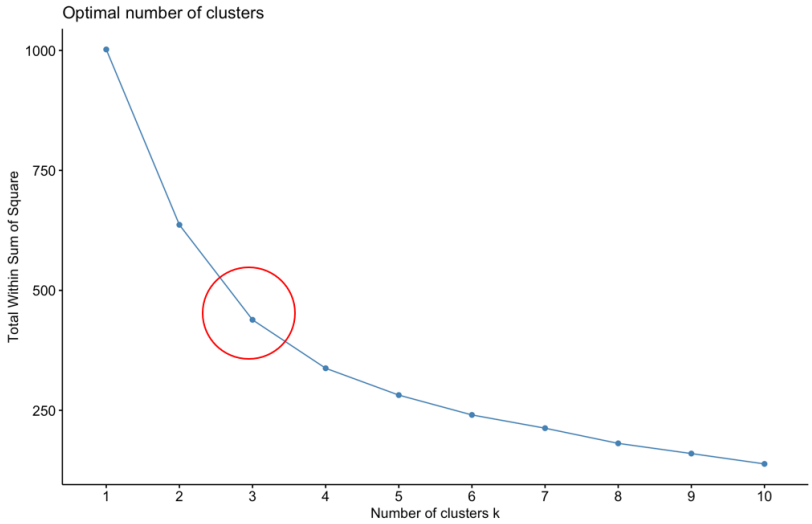

Figure A1.1: Elbow Method for S&P stocks data

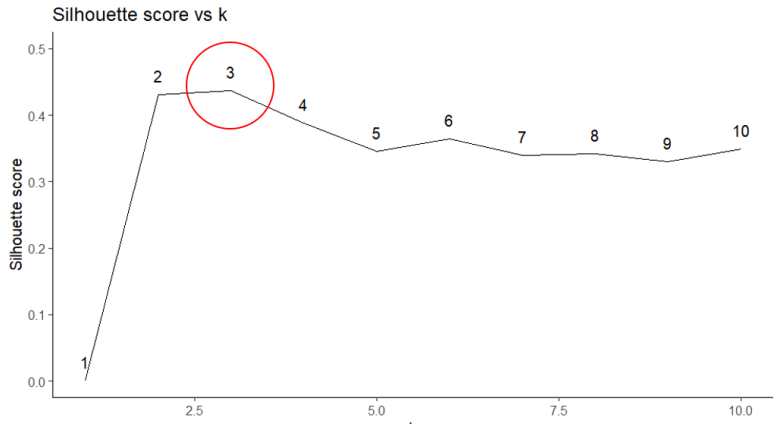

Figure A1.2: Silhouette Score for S&P stocks data

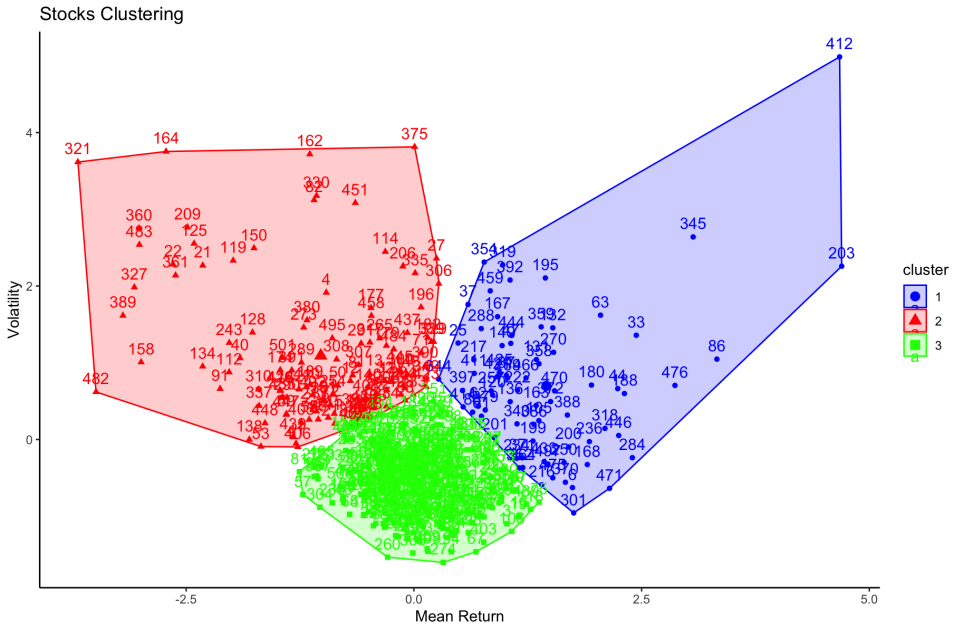

Figure A1.3: Cluster plot

k	Accuracy	k	Accuracy	k	Accuracy	k	Accuracy
1	0.985075	6	0.990050	11	0.965174	16	0.975124
2	0.975124	7	0.985075	12	0.975124	17	0.965174
3	0.985075	8	0.990050	13	0.965174	18	0.965174
4	0.990050	9	0.970149	14	0.970149	19	0.960199
5	0.985075	10	0.985075	15	0.970149	20	0.970149

Table A1.3: Accuracy of the KNN model

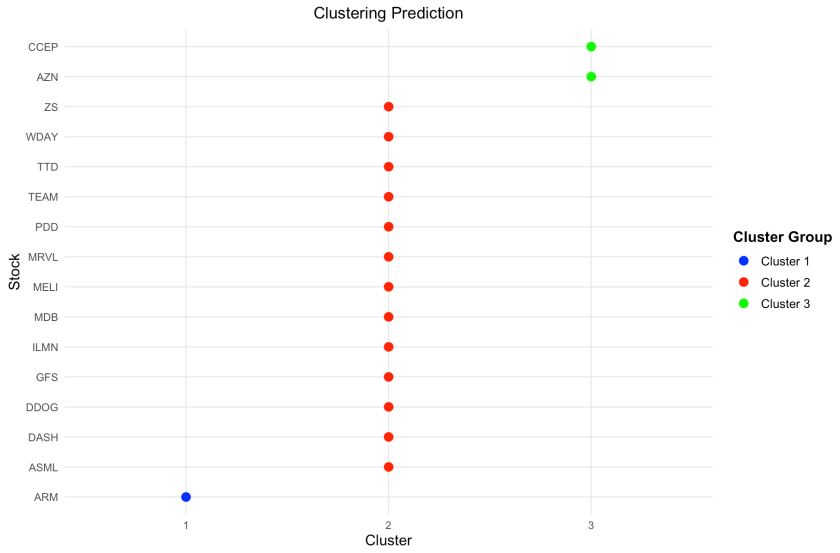

Figure A1.4: Prediction