A Clustering-Based Approach as a Potential Alternative to Industry Classification

Report prepared by:
Zheng En Than (1586814) and Shubham Sahoo (1824661)

Supervisors: Christoph Frei and Rudabeh Meskarian

August 2024

Content

1 Abstract

This project presents an alternative approach to classifying companies using a correlation matrix and features engineered from stock returns as inputs for clustering algorithms. The proposed method aims to provide a more dynamic and data-driven alternative to the traditional Global Industry Classification Standard (GICS). By analyzing 20 years of daily return data of S&P 500 companies, clusters formed using this approach are able to capture the evolving economic relationships between companies across different market conditions. Additionally, the case study on Amazon highlights the limitations of static classification systems like GICS and the need for more flexible methods to reflect the ever-evolving market landscape. Future work may include exploring more data sources and testing the proposed model in real-world scenarios to further validate its practical utility.

2 Introduction

2.1 Project Description

The main goal of this project is to develop a potential alternative to the GICS classification system. The alternative proposed in this project uses clustering to classify companies based on their historical stock performance and aims to be more representative of the dynamic nature of companies' economic relationships. By providing a more data-driven approach, the proposed classification method can be used as a valuable tool in investment management and risk analysis.

2.2 Background Knowledge

The GICS is a widely used framework developed by MSCI and S&P Dow Jones to help categorize companies based on their key business activities. The GICS consists of 11 main

sectors: Energy, Materials, Industrials, Consumer Discretionary or Consumer Cyclical, Consumer Staples or Consumer Defensive, Health Care, Financials, Information Technology, Communication Services, Utilities, and Real Estate. These 11 sectors are further subcategorized into 25 industry groups, 74 industries, and subsequently 163 sub-industries (MSCI, n.d.).

While the GICS is a standard tool, it has limitations in accurately representing companies whose business models have evolved significantly over time. For example, Amazon was initially classified as a Consumer Cyclical company due to its core business operation as an online retailer. However, even though Amazon is now arguably more of an Information Technology company due to its substantial investment in their cloud computing services through Amazon Web Services (AWS), its GICS classification remains to be Consumer Cyclical.

Clustering is an unsupervised machine learning technique that groups similar subjects or data points together based on their features. In this project, clustering is used to group companies based on their historical daily stock return data. As this approach uses dynamic data, the clustering results will be more representative of the companies' behaviors over time, potentially capturing shifts in their economic relationships that a static classification system like GICS might overlook.

2.3 Project Objectives

The objectives of this project are as follows:

- Developing a more data-driven company classification system

        - This project seeks to answer the question of "How can we develop a more meaningful classification system for companies that better captures their dynamic economic relationships and business activities compared to the traditional, static GICS classification system?"
    - Providing insights into how companies' relationships change over time
        - This project intends to examine how companies are grouped differently during different economic conditions. By analyzing the clustering results across different periods, this project will reveal the dynamic relationships among companies and how these relationships shift in response to varying economic conditions.

The proposed classification system offers several potential use cases:

- Long-term investment strategies
    - Users interested in long-term investment strategies may use data spanning a longer period, such as 20 years, to identify overall company groupings and any overarching trends.
- Period-specific analysis
    - Users interested in understanding company behaviors during specific periods, such as the 2008 financial crisis, may use data from those times to gain insights into the different ways that companies react and adapt to economic downturns.
- Current market analysis

- Users interested in making investment decisions based on the current market trends may use data from recent periods, such as the last 1 to 2 years, to identify the latest trends and company relationships.

3 Data Collection and Preprocessing

3.1 Datasets and Collection

The current list of S&P 500 companies was obtained by scraping the SlickCharts website using the BeautifulSoup library. Additionally, the historical daily adjusted closing prices and the GICS sector classifications for these companies were retrieved using the yfinance API. These datasets were accessed and downloaded on August 6, 2024. The historical daily adjusted closing price data spans from January 1, 2004, to January 1, 2024. This 20-year period allows for a comprehensive analysis of market dynamics across different economic cycles which include critical events such as the 2008 financial crisis and the COVID-19 pandemic. The adjusted closing price is used instead of closing price because it is adjusted for dividends, stock splits, and other corporate actions which makes it a more accurate reflection of a company stock's value (Groww, n.d.).

3.2 Data Preprocessing

In this project, we will classify companies based on the characteristics of their stocks' daily return instead of daily adjusted closing price because daily return provides a normalized view of each company's performance over time. This allows us to more accurately compare the volatility and growth between companies regardless of the absolute values of their stock prices.

In order to obtain the daily return of these companies' stocks, the `pct_change` function from the Pandas library was used. The formula of the daily return is

$$\frac{Today's\ Adjusted\ Closing\ Price - Yesterday's\ Adjusted\ Closing\ Price}{Yesterday's\ Adjusted\ Closing\ Price}$$

After all 500 companies' daily returns are calculated, we generated the descriptive statistics, box plots, and line charts to visualize the data.

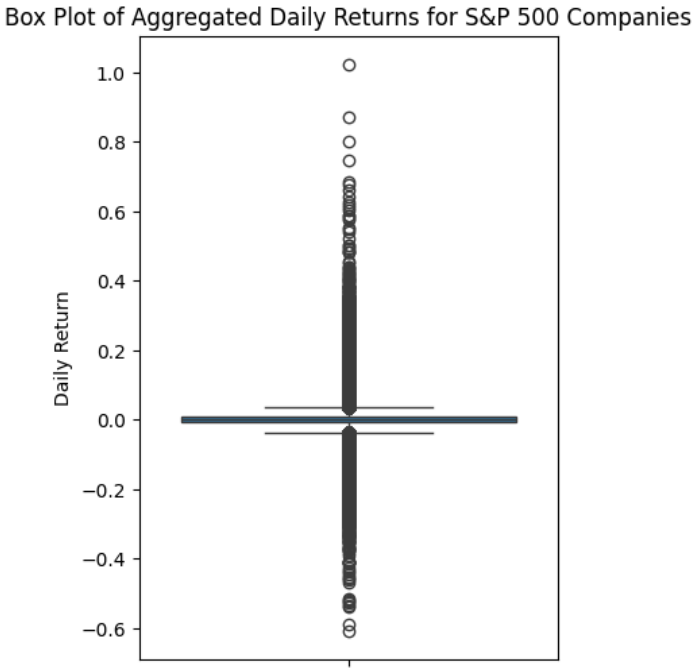


Figure 1. Box plot of the daily returns for all S&P 500 companies over the 20-year period. The y-axis represents daily returns where, for example, a value of 0.2 indicates a 20% increase in adjusted closing price from one day to the next. The narrow box indicates that the majority of daily returns are tightly clustered around 0, suggesting low variability in day-to-day performance. However, the presence of numerous points outside the whiskers may be attributed to significant market events that caused volatility spikes, such as the 2008 financial crisis.

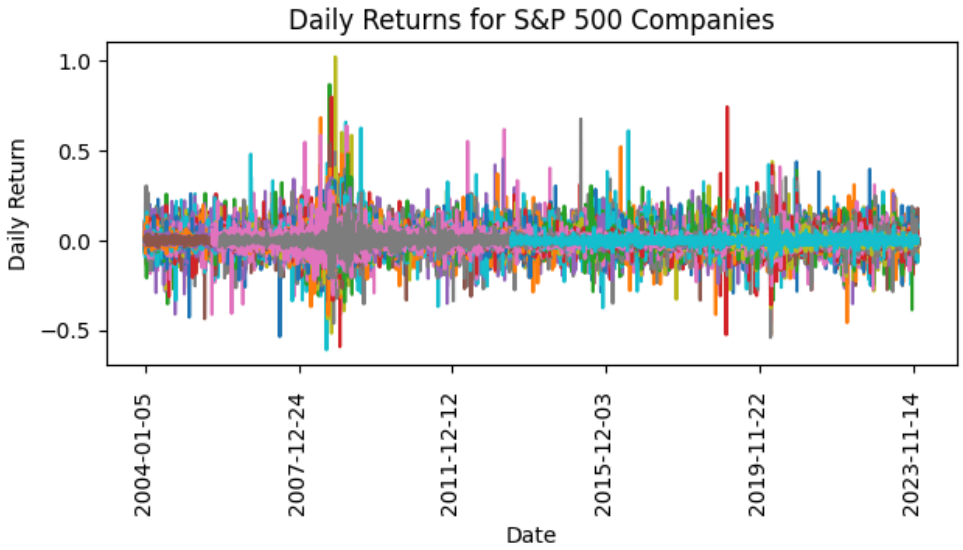


Figure 2. Time series line chart showing the daily returns of all S&P 500 companies. The period between 2007 and 2009 exhibits an increased volatility in daily returns, likely due to the 2008 financial crisis.

As of August 6, 2024, not all S&P 500 companies have complete data over the full 20-year period. Missing data can make it challenging for the clustering algorithm to accurately compute distances between companies, potentially resulting in lower-quality clusters. Additionally, imputing missing values might introduce bias, as the imputed values may not accurately represent a company's true performance. To simplify the analysis and avoid potential issues with missing or imputed data, we excluded companies with incomplete data, leaving us with 385 companies that have complete daily return data for the entire 20-year period. However, this exclusion may introduce survivorship bias into our analysis, which will be further discussed in Section 4.3.

Next, we also segmented the daily return data into five distinct periods that are defined as follows:

- Pre-Financial Crisis: January 1, 2004, to December 31, 2006
- Financial Crisis: January 1, 2007, to June 30, 2009
- Post-Financial Crisis: July 1, 2009, to December 31, 2019
- COVID Period: January 1, 2020, to December 31, 2022
- Post-COVID Period: January 1, 2023, to January 1, 2024

Segmenting the data into these periods allows us to examine how companies' economic relationships evolved in response to significant economic events as shown by their clustering results.

4 Methodology

4.1 Approaches

We used the following three approaches for clustering:

- **Correlation matrix**

  The first approach uses a correlation matrix as input for our clustering algorithms. We computed the correlation matrix by assessing the pairwise correlations between the daily returns of companies using the `corr` function from the Pandas library. This matrix quantifies the linear relationships between the daily returns of different companies.

- **Feature engineering**

  The second approach uses engineered features as input for our clustering algorithms. We created 10 features for each company that capture the average daily returns and

volatilities across different economic phases. Specifically, two features were created for each economic phase, resulting in 10 features as shown below:

1. **Pre-financial Crisis (2004-2006):** A time of economic growth and stability
    - **‘**avg_return_pre_crisis**’**: Average daily return of the company.
    - **‘**volatility_pre_crisis**’**: Standard deviation of the company’s daily return.
2. **Financial Crisis (2007-2009):** Marked by economic downturn and high volatility
    - **‘**avg_return_crisis**’**: Average daily return during the 2008 financial crisis.
    - **‘**volatility_crisis**’**: Standard deviation of daily returns during the crisis.
3. **Post-financial Crisis (2010-2019)**: Recovery phase with varying levels of growth
    - **‘**avg_return_post_crisis**’**: Average daily return during the recovery phase.
    - **‘**volatility_post_crisis**’**: Standard deviation of daily returns post-crisis.
4. **COVID-19 Period (2020-2022):** Global economic disruption and uncertainty
    - **‘**avg_return_covid**’:** Average daily return during the COVID-19 pandemic.
    - **‘**volatility_covid**’**: Standard deviation of daily returns during the pandemic.
5. **Post-COVID-19 (2023 onwards):** Adjustments and new trends after pandemic
    - **‘**avg_return_post_covid**’**: Average daily return during the post-COVID period.
    - **‘**volatility_post_covid**’**: Standard deviation of daily returns post-pandemic.

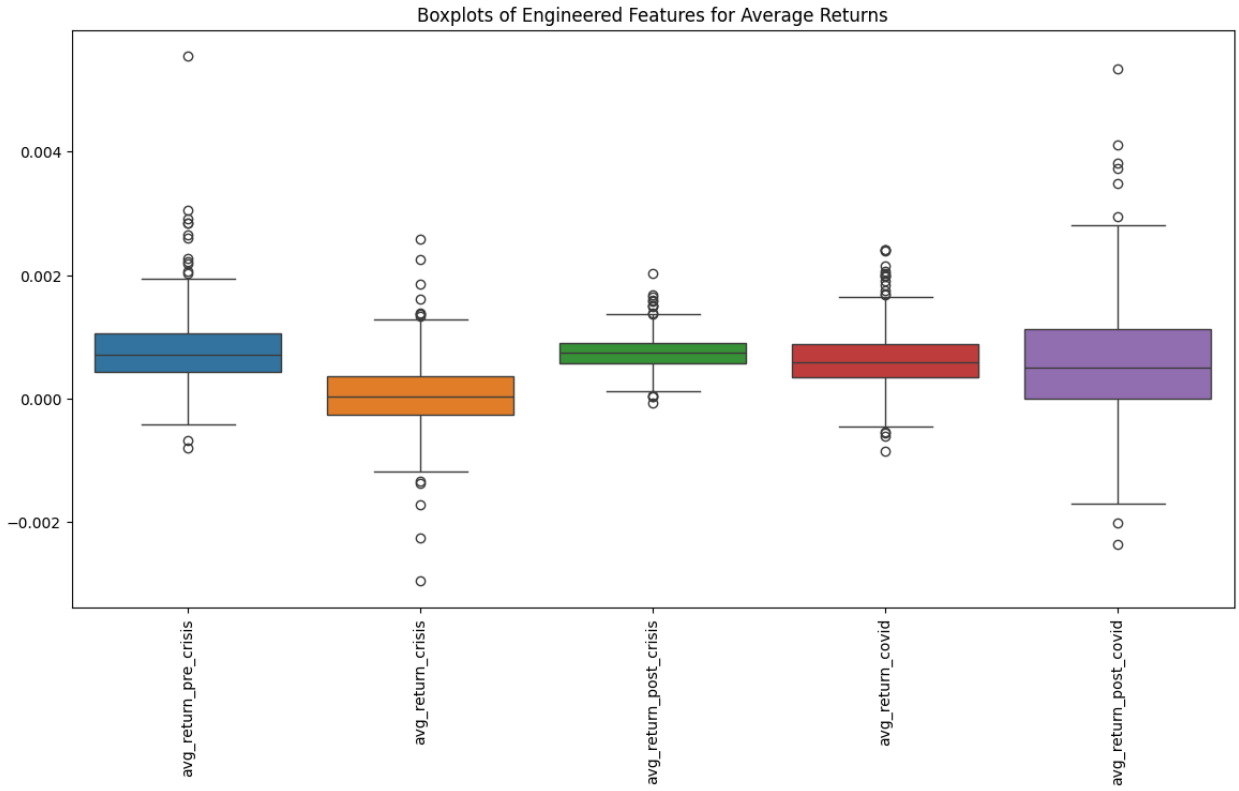


Figure 3. Boxplots of engineered features representing average daily returns of companies across different economic phases.

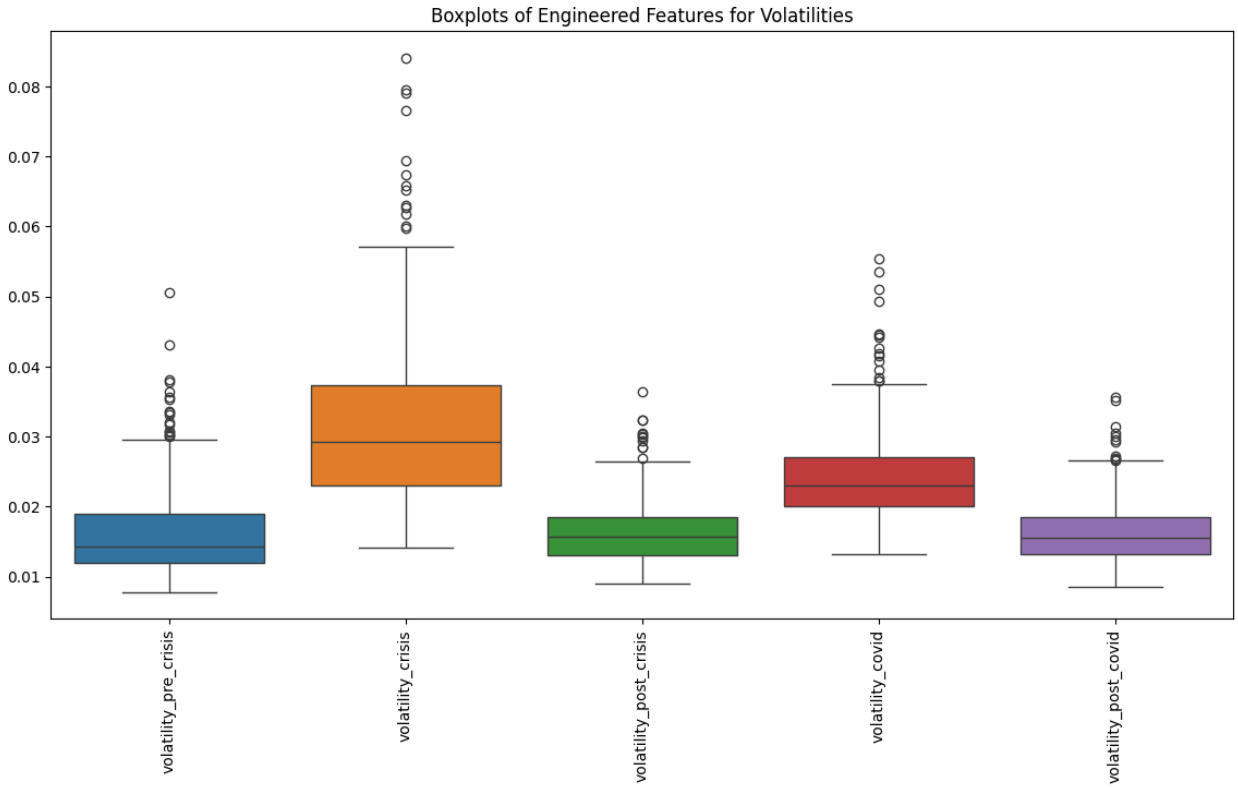


Figure 4. Boxplots of engineered features representing the daily return volatilities of companies across different economic phases.

By including both average daily return and volatility for each period, the engineered features offer a better view of a company's performance and risk across different economic conditions. This enables a more granular analysis of company resilience and growth potential. The spread observed in the boxplots, especially for volatilities, further validates the effectiveness of these features in capturing meaningful distinctions among the companies.

- **The combined approach, using both correlation matrix and feature engineering**

  The third approach uses both the correlation matrix and engineered features as inputs for our clustering algorithms. Initially, the correlation matrix had 385 columns. To balance

the influence between these correlation matrix columns and the engineered features, we applied Principal Component Analysis (PCA) using the `PCA` function from Scikit-learn on the correlation matrix. This process reduced the 385 columns to 10 principal components which were able to capture 89% of the variance in the dataset. Each company ended up with 10 engineered features and 10 PCA-transformed correlation matrix columns, which are then used as inputs for our clustering algorithms.

4.2 Hyperparameter Tuning and Model Selection

For each of the three approaches listed in Section 4.1, we implemented two clustering algorithms — K-means using the `KMeans` function and hierarchical clustering using the `AgglomerativeClustering` function from scikit-learn. Therefore, we ended up with six clustering model combinations.

For each model, we performed hyperparameter tuning using data for the entire 20-year period to identify the set of hyperparameters that gave the highest silhouette score. The silhouette score ranges from -1 to 1, with higher values indicating that the data points — or, for our use case, companies — are well matched within their own cluster and poorly matched to other clusters. The hyperparameters, their meanings, and options that were explored are as follows (Scikit-learn, n.d.):

- **For K-means clustering:**

Hyperparameter	Description	Options explored

n_clusters	Number of clusters to form	Integers from 2 to 20 inclusive
n_init	Number of times the K-means algorithm will run with different centroid seeds.	'auto', 10, 20, 30
max_iter	Maximum number of iterations for a single initialization.	300, 400, 500
init	Method for initializing the centroids.	'k-means++', 'random
algorithm	Algorithm used for K-means computation.	'lloyd', 'elkan'

- **For hierarchical clustering:**

Hyperparameter	Description	Options explored
n_clusters	Number of clusters to form	Integers from 2 to 20 inclusive
linkage	Method used to calculate the distance between clusters.	'ward'
metric	Distance measure used to compute the distance between companies.	'euclidean'

We tested with various linkage methods such as ‘single’, ‘complete’, and ‘average’ with metrics such as ‘l1’, ‘l2’, ‘manhattan’, ‘cosine’, or ‘precomputed’ for hierarchical clustering. However, these methods often resulted in highly imbalanced clusters, with one cluster containing nearly all companies and another containing only one or two companies. Even after potential outliers were removed, these methods would still frequently produce highly imbalanced clusters. This issue may be due to the chaining effect associated with ‘single’ linkage or the sensitivity to outliers observed with ‘complete’ and ‘average’ linkage methods. Therefore, we decided to only focus on the ‘ward’ linkage method which produced more balanced clustering results.

4.3 Model Assumptions

As discussed in Section 3.2, we excluded companies without complete data for the entire 20-year period to avoid the complications associated with imputing missing values. However, this may introduce bias as we are excluding companies that have gone out of business, merged, or are relatively newly established. As a result, the remaining dataset could be skewed towards more stable and long-standing companies. Despite this, we assume that the remaining companies provide a sufficiently representative sample for the objectives of this project.

We also assume that the features used, including the correlation matrix and engineered features, are both relevant and adequate to capture the essential structure of the daily return data. This ensures that the clustering results reflect meaningful and relevant patterns in the performance and economic relationships of these companies.

5 Model Analysis

5.1 First Approach (Correlation Matrix)

In the first approach where we used the correlation matrix of the companies' daily return data as input for the K-means and hierarchical clustering algorithms, the best `n_cluster`'s in terms of silhouette scores are relatively small. Specifically, the best `n_cluster` for K-means clustering is 3 while that of hierarchical clustering is 5. Given that there are 11 GICS sectors that make up these 385 companies, the suggested numbers of clusters appear to be quite low. This may suggest a limitation of the models, indicating that the correlation matrix alone is not adequate to capture more nuanced differences between these companies.

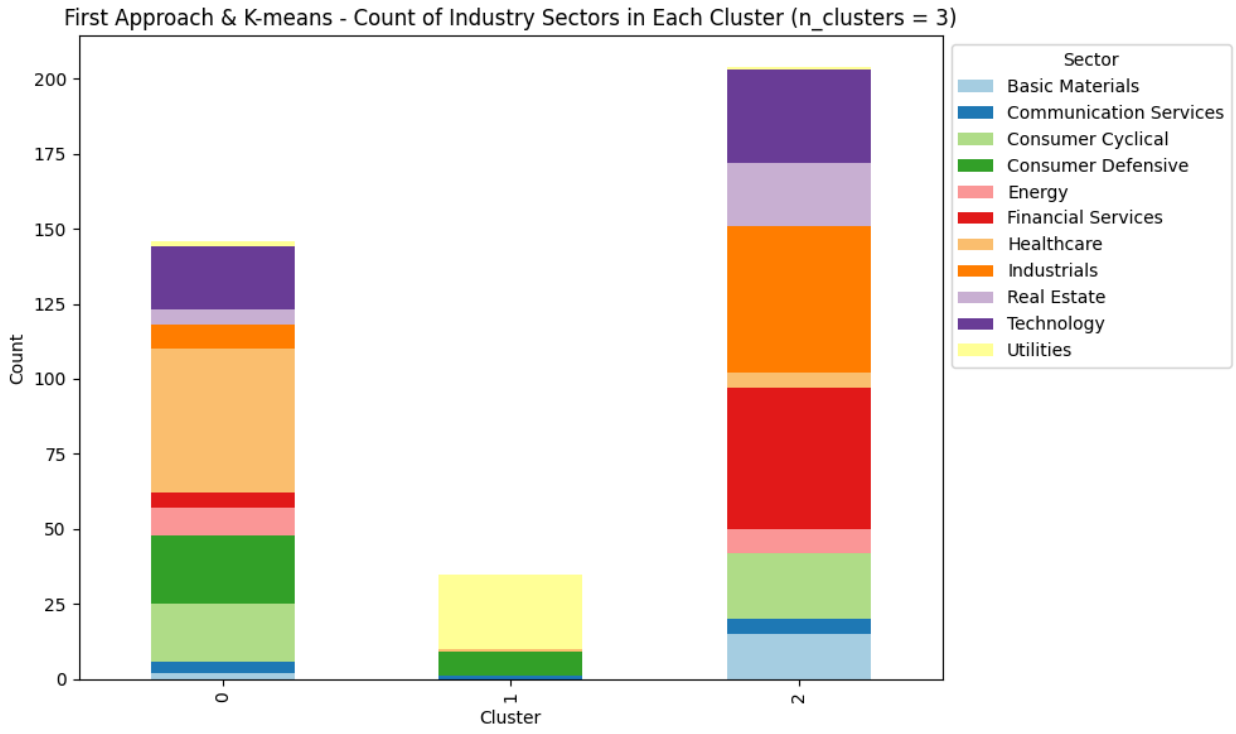


Figure 5. Stacked bar plot showing the count of companies and their GICS sectors in each of the three clusters obtained from K-means clustering using correlation matrix with `n_cluster` = 3.

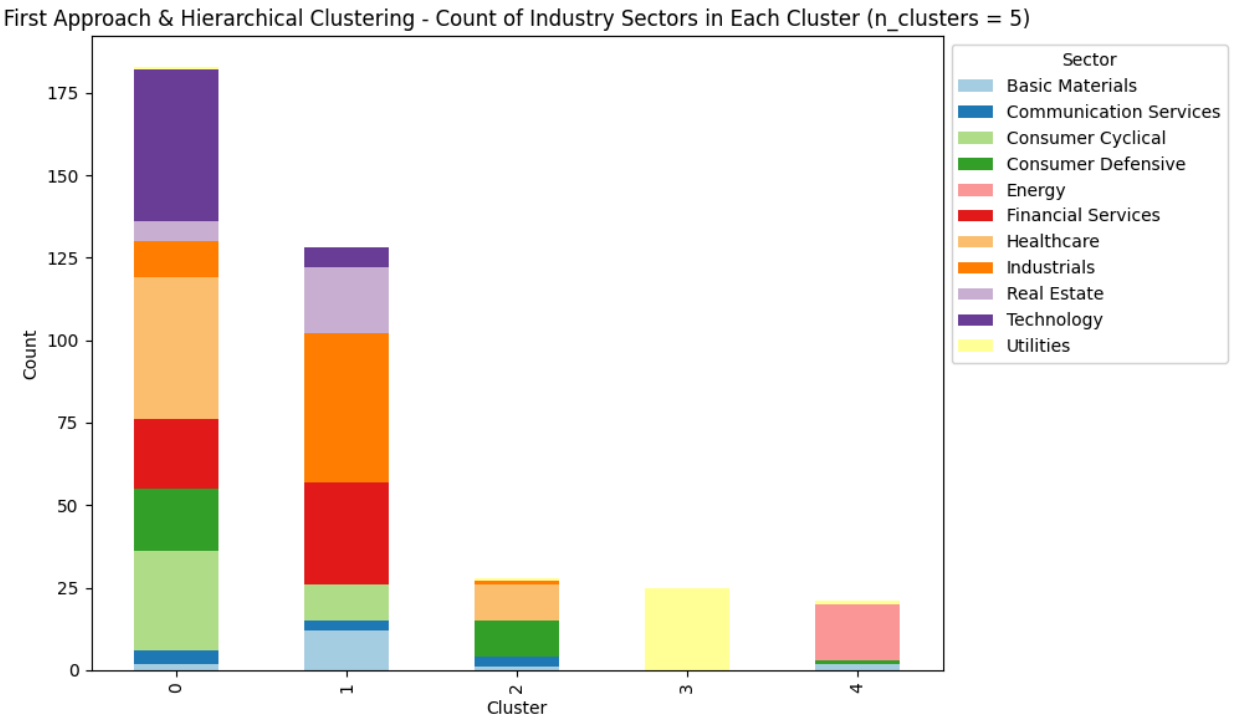


Figure 6. Stacked bar plot showing the count of companies and their GICS sectors in each of the five clusters obtained from hierarchical clustering using correlation matrix with `n_cluster` = 5.

Figures 5 and 6 show the count of companies and their GICS sectors in each cluster obtained from K-means clustering and hierarchical clustering respectively. Both stacked bar plots show a cluster that is primarily composed of Utilities companies, which are Cluster 1 in Figure 5 and Cluster 3 in Figure 6. Cluster 3 in Figure 6 is entirely composed of Utilities companies unlike Cluster 1 in Figure 5, which, while dominated by Utilities, still includes a mix of other sectors. Moreover, both plots also have their largest clusters being composed of a similar sector composition, which include Utilities, Technology, Real Estate, Industrials, Healthcare, Financial Services, Consumer Defensive, Consumer Cyclical, Communication Services, and Basic Materials. However, Figure 5's largest cluster also contains Energy companies, whereas Figure 6

does not. With a higher `n_cluster`, the second model has more flexibility to allocate one smaller cluster predominantly for Energy companies, which is Cluster 4.

5.2 Second Approach (Feature Engineering)

In the second approach where we used features engineered from the companies' daily return data as input for the K-means and hierarchical clustering algorithms, the best `n_cluster`'s in terms of silhouette scores are even smaller. Specifically, the best `n_cluster` for both K-means and hierarchical clustering is 2. Similar to the first approach, this may suggest a limitation of the models, indicating that the 10 engineered features alone is not adequate to capture more nuanced differences between these companies, leading to clustering results that oversimplify the underlying patterns of the daily return data.

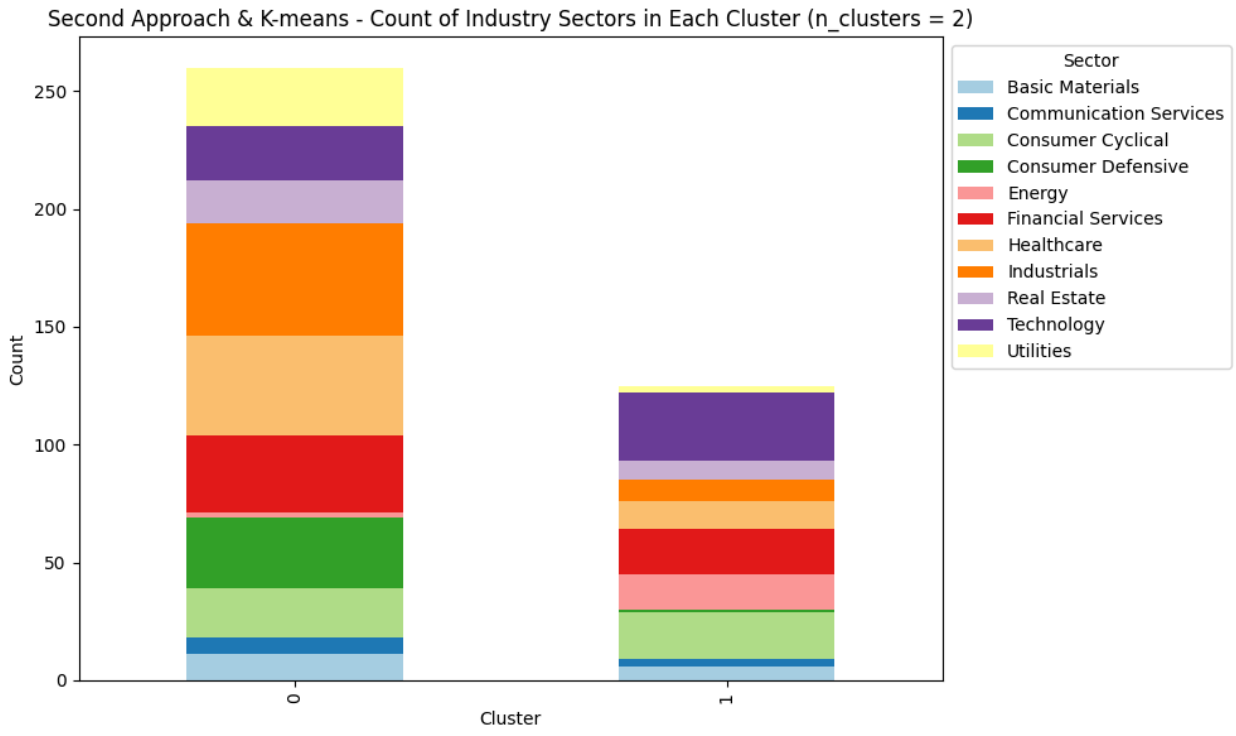


Figure 7. Stacked bar plot showing the count of companies and their GICS sectors in each of the two clusters obtained from K-means clustering using feature engineering with `n_cluster` = 2.

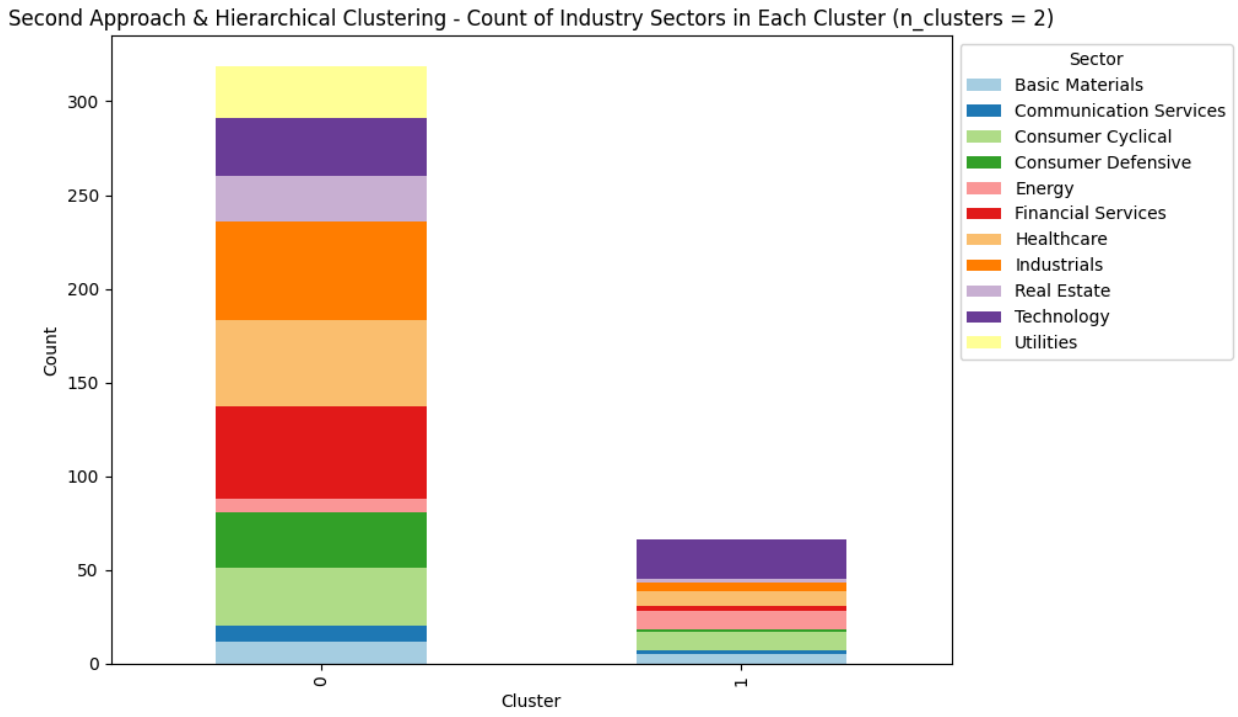


Figure 8. Stacked bar plot showing the count of companies and their GICS sectors in each of the two clusters obtained from hierarchical clustering using feature engineering with `n_cluster` = 2.

Figures 7 and Figure 8 show the count of companies and their GICS sectors in each cluster resulting from K-means clustering and hierarchical clustering respectively. Both stacked bar plots have a large cluster that is at least twice the size of the second cluster. The composition of these clusters is similar across both methods, where both clusters are composed of a mix of various sectors and that there is no cluster that is primarily composed of one sector.

5.3 Third Approach (Correlation Matrix and Feature Engineering)

In the third approach where we combined the correlation matrix with features engineered from the companies' daily return data for K-means and hierarchical clustering algorithms, the best `n_cluster`'s in terms of silhouette scores are relatively larger. Specifically, the best `n_cluster` for K-means is 13, while that of hierarchical clustering is 20. The larger recommended number of clusters may suggest that these models can capture more granular differences between the companies. The combination of the correlation matrix and engineered features likely provides the clustering algorithms with a richer set of information which helps with better differentiation among the companies.

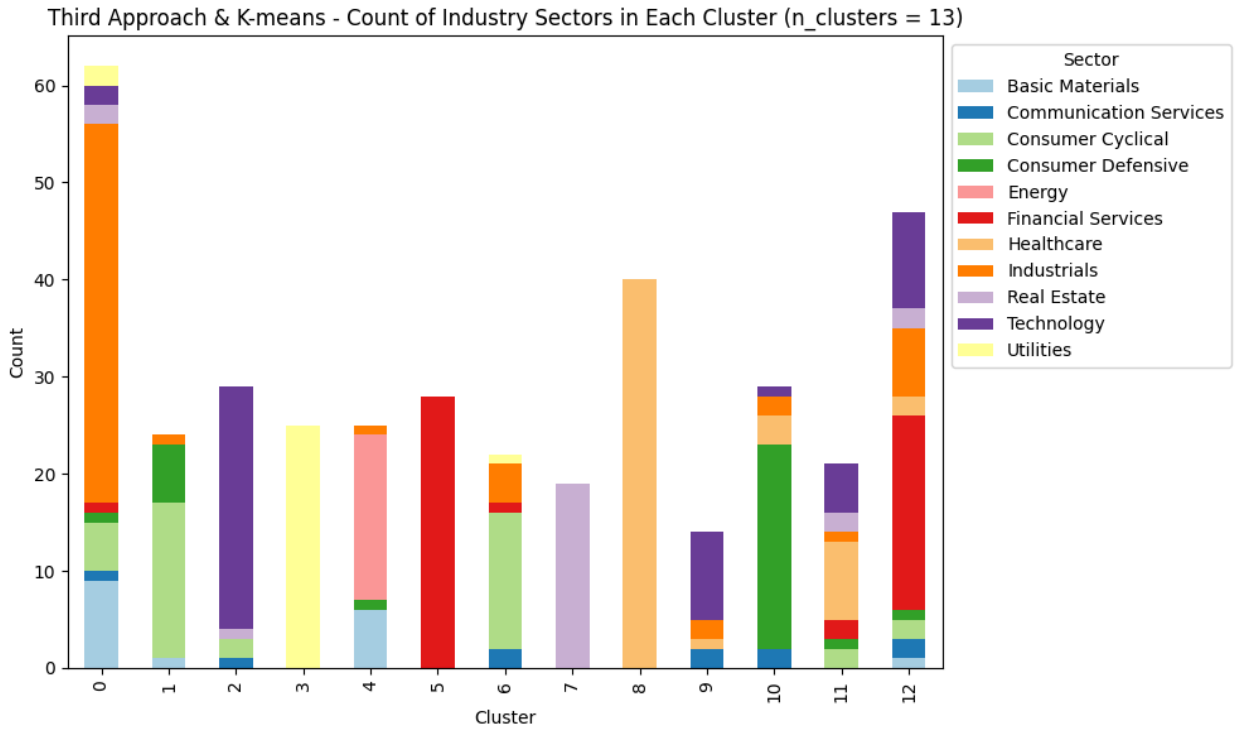


Figure 9. Stacked bar plot showing the count of companies and their GICS sectors in each of the 13 clusters obtained from K-means clustering using correlation matrix and feature engineering with `n_cluster` = 13.

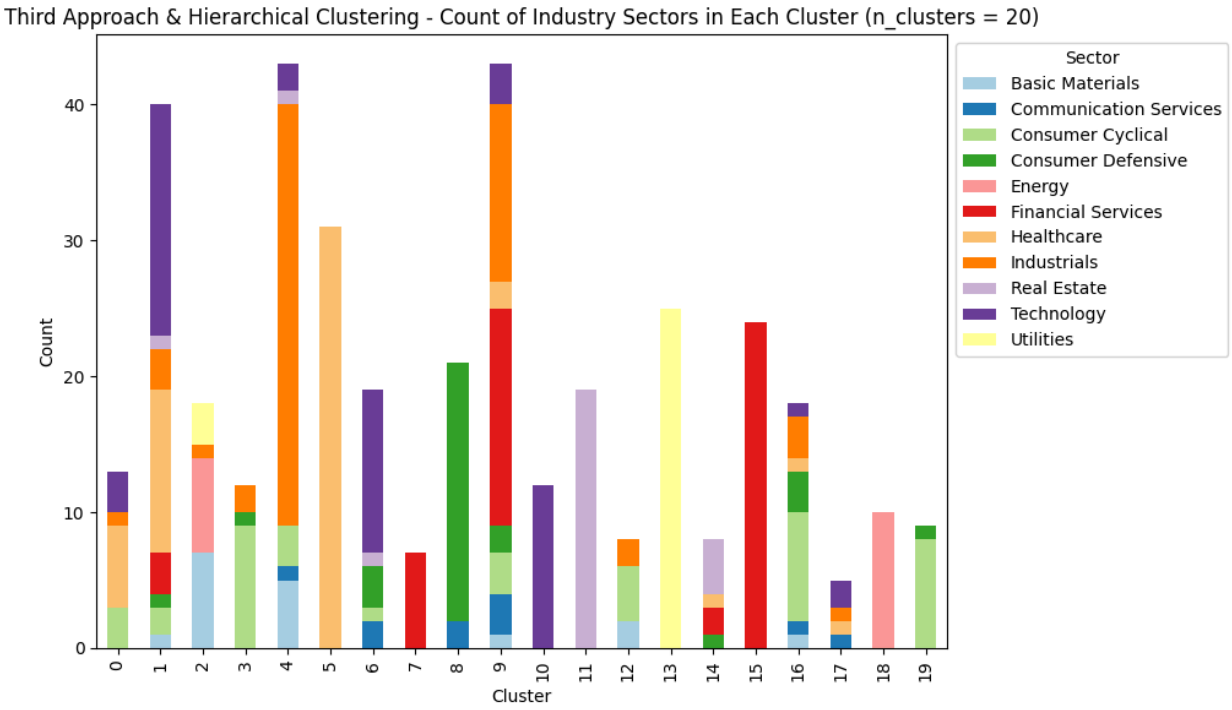


Figure 10. Stacked bar plot showing the count of companies and their GICS sectors in each of the 20 clusters obtained from hierarchical clustering using correlation matrix and feature engineering with `n_cluster` = 20.

Figures 9 and 10 show the count of companies and their GICS sectors in each cluster obtained from K-means clustering and hierarchical clustering respectively. Both stacked bar plots have clusters that are composed entirely of single sectors. For instance, Cluster 3 in Figure 9 and Cluster 13 in Figure 10 are made up of only Utilities companies. Similarly, Cluster 5 in Figure 9 as well as Clusters 7 and 15 in Figure 10 are composed only of Financial Services companies.

Figures 9 and 10 also exhibit some similarities in the composition of their two largest clusters. In particular, Cluster 0 in Figure 9 and Cluster 4 in Figure 10 are both primarily made up of Industrials companies, followed by other sectors such as Technology, Real Estate, Consumer Cyclical, Communication Services, and Basic Materials. Their next largest clusters, which are Cluster 12 in Figure 9 and Cluster 9 in Figure 10, are primarily composed of Industrials and Financial Services companies, followed by other sectors such as Technology, Healthcare, and Consumer Cyclical.

Despite these similarities, the clustering results of the two algorithms differ in several ways. Firstly, hierarchical clustering produced more "pure" clusters, which are clusters that are composed of only a single sector. This may be due to its larger recommended `n_cluster`, which allows for more detailed grouping compared to using a smaller value of `n_cluster`. Although using higher values for `n_cluster` in K-means could also produce many "pure" clusters, we chose to not follow this approach since these higher values did not result in silhouette scores that exceeded those achieved with `n_cluster` set to 13.

Secondly, the larger `n_cluster` used in hierarchical clustering appears to result in a more imbalanced distribution of companies across clusters, with some clusters being significantly smaller than others. This may result in having too many clusters that are overly granular, making it difficult to identify the features that clearly distinguish them.

Approach and clustering algorithm	Best hyperparameter values according to silhouette scores	Best silhoutte score
First approach and K-means	{'algorithm': 'lloyd', 'init': 'k-means++', 'max_iter': 300, 'n_clusters': 3, 'n_init': 10}	0.277778507
First approach and hierarchical clustering	{'linkage': 'ward', 'metric': 'euclidean', 'n clusters': 5}	0.234013117
Second approach and K-means	{'algorithm': 'lloyd', 'init': 'k-means++', 'max_iter': 300, 'n_clusters': 2, 'n_init': 'auto'}	0.275154572
Second approach and hierarchical clustering	{'linkage': 'ward', 'metric': 'euclidean', 'n_clusters': 2}	0.335448831
Third approach and K-means	{'algorithm': 'lloyd', 'init': 'random', 'max iter': 300, 'n clusters': 13, 'n init': 30}	0.204498407
Third approach and hierarchical clustering	{'linkage': 'ward', 'metric': 'euclidean', 'n_clusters': 20}	0.180123061

Table 1. The best hyperparameter values according to silhouette scores for each model.

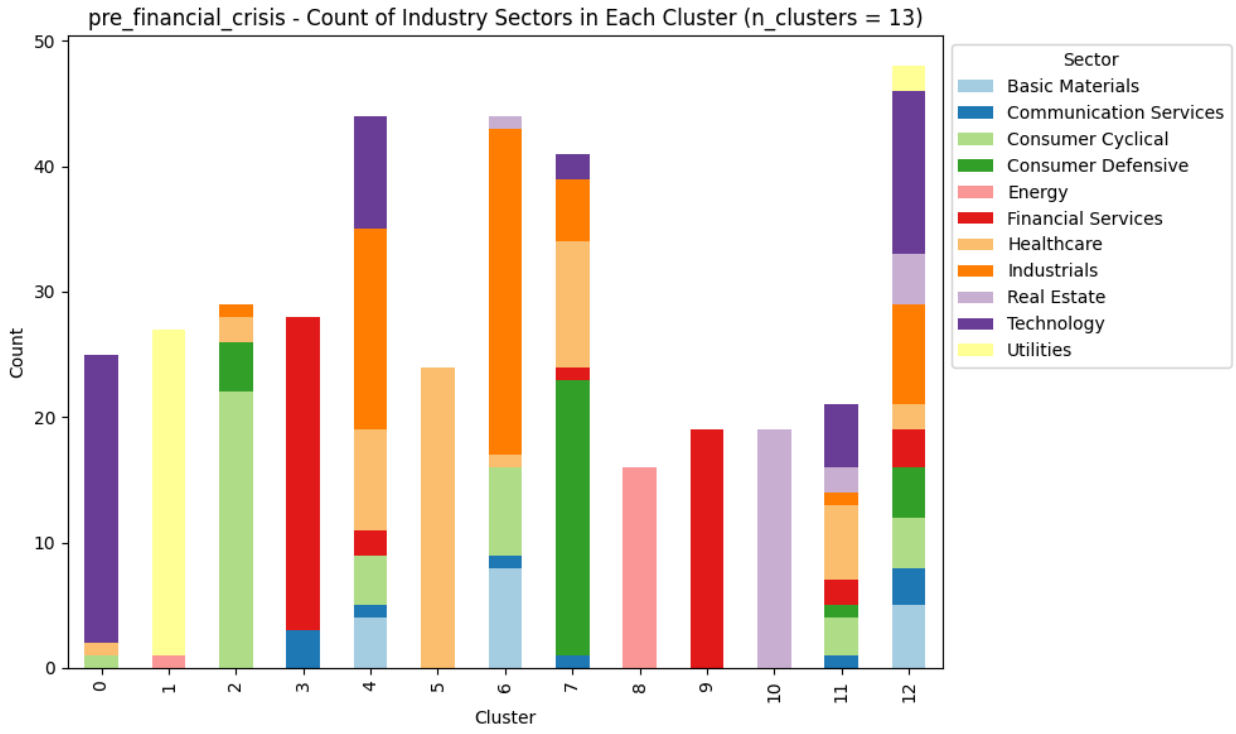
5.4 Choosing the Third Approach and K-Means over Other Models

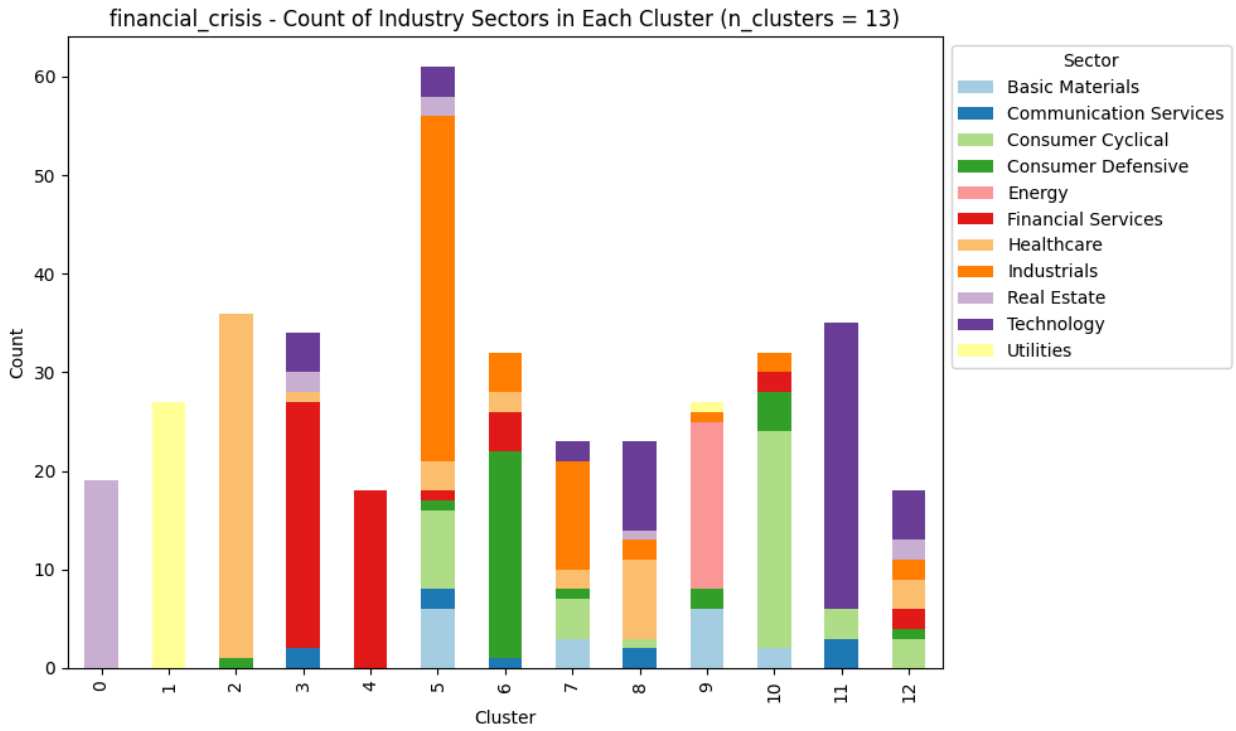
We decided to focus our further analysis on the third approach over the first two. This is because we believe that the combination of correlation matrix and engineered features can ultimately give us more insights about the daily return data. The higher values of `n_cluster` suggested by silhouette scores for the third approach likely indicate that these models can better capture more granular details about the companies and differentiate them more effectively.

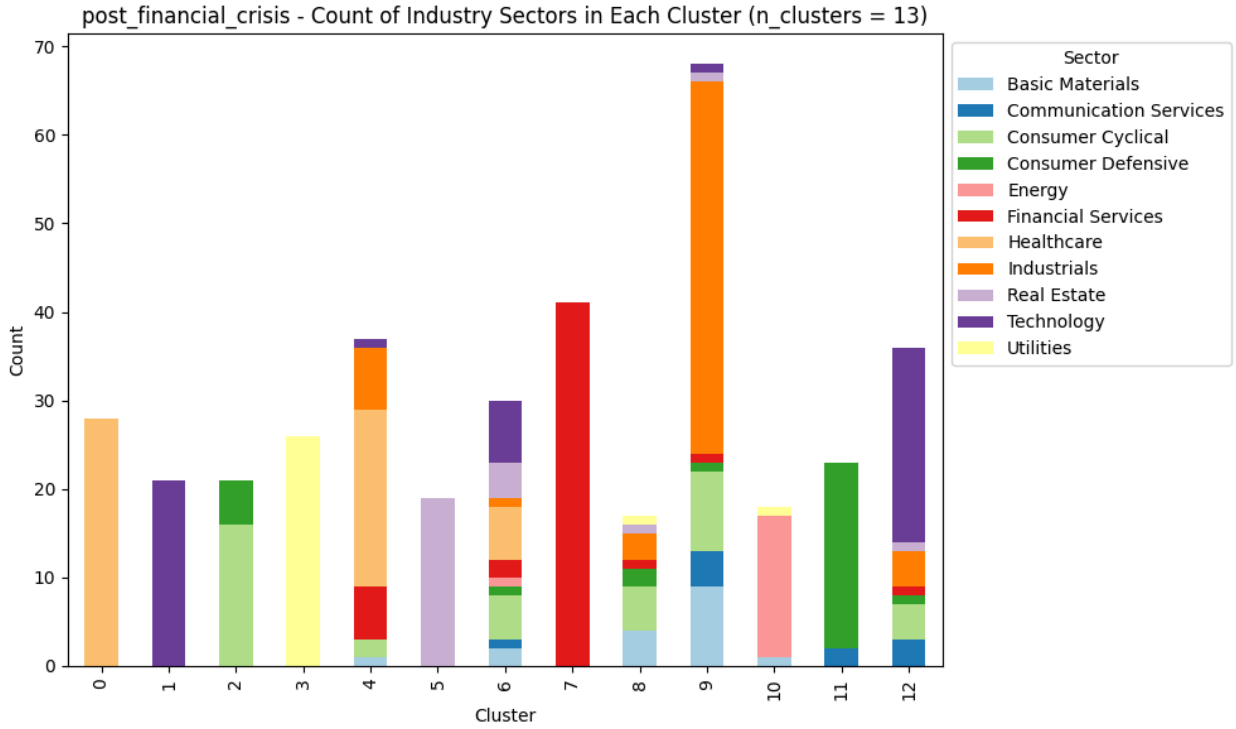
Furthermore, we decided to concentrate our further analysis on K-means over hierarchical clustering with the third approach. The value of `n_cluster` recommended by silhouette score is 13 for K-means clustering. Compared to that of hierarchical clustering which has 20 as the recommended value, K-means clustering appears to strike a better balance between detail and interpretability. K-means also generally has a better time complexity than hierarchical clustering, making it a more suitable choice for scalability if the proposed approach is to be applied to larger datasets.

K-means clustering with the third approach also shows a rather consistent clustering pattern across different periods. When applied to data for the five different periods as defined in Section

3.2, the resulting clusters show similar compositions over time. Specifically, the Industrials companies tend to be grouped together, forming one of the largest clusters. Moreover, clusters that are primarily composed of Financial Services, Real Estate, Utilities, Healthcare, and Technology companies respectively also tend to appear across these different periods. The number of clusters that are dominated by a single cluster, where at least 80% of the cluster comes from one sector, remains stable across periods. This number ranges from 5 to 8 clusters in each period, with no period showing a sudden absence or a significant increase in these clusters.




financial_crisis - Count of Industry Sectors in Each Cluster (n_clusters = 13)
Count
Cluster
Sector
Basic Materials
Communication Services
Consumer Cyclical
Consumer Defensive
Energy
Financial Services
Healthcare
Industrials
Real Estate
Technology
Utilities


post_financial_crisis - Count of Industry Sectors in Each Cluster (n_clusters = 13)
Count
Cluster
Sector
Basic Materials
Communication Services
Consumer Cyclical
Consumer Defensive
Energy
Financial Services
Healthcare
Industrials
Real Estate
Technology
Utilities

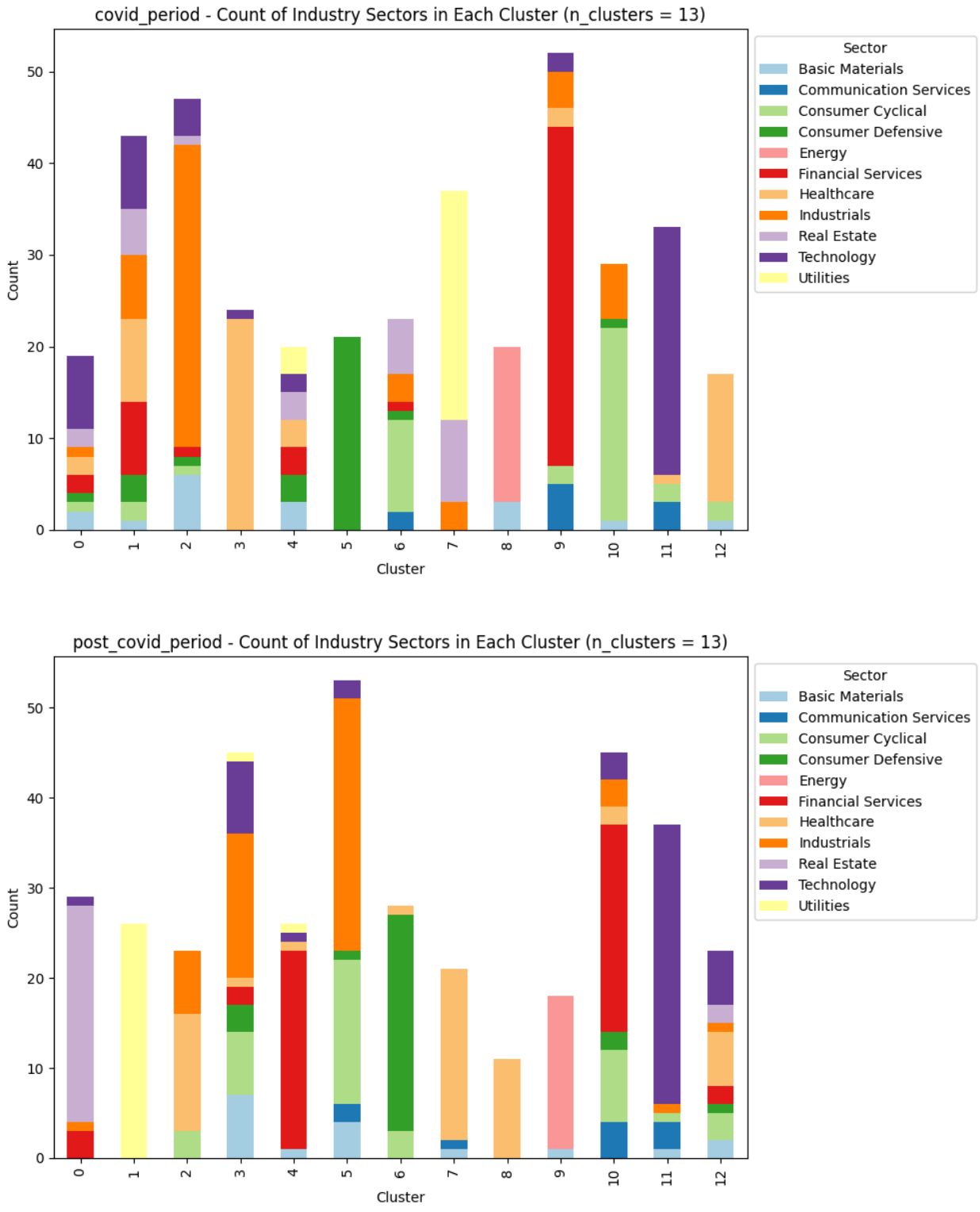


Figure 11. Stacked bar plots showing the count of companies and their GICS sectors in each of the 13 clusters obtained from K-means clustering using correlation matrix and feature

engineering with data for the pre-financial crisis, financial crisis, post-financial crisis, COVID-19, and post-COVID-19 periods respectively.

6 Clustering Results of the Third Approach with K-Means

6.1 Mean Daily Return, Volatility of Each Cluster, and Inter-Cluster Correlation

We will now look at the characteristics of the 13 clusters formed by K-means with correlation matrix and engineered features when trained with the entire 20-year data. Specifically, we will first examine the mean daily return and mean volatility (standard deviation of daily return) for each cluster.

Cluster	R_pre_crisis	V_pre_crisis	R_crisis	V_crisis	R_post_crisis	V_post_crisis	R_covid	V_covid	R_post_covid	V_post_covid
0	0.1080%	1.2525%	-0.0045%	4.3571%	0.0676%	1.4517%	0.0300%	2.5623%	0.0483%	1.6509%
1	0.0509%	1.3486%	0.0014%	2.5077%	0.0823%	1.3061%	0.0586%	2.1582%	0.0394%	1.4482%
2	0.1507%	3.2882%	0.0525%	3.8800%	0.1275%	2.4790%	0.0765%	2.6671%	0.1232%	1.8502%
3	0.0575%	1.1548%	0.0004%	5.4042%	0.0659%	1.7864%	0.0593%	2.7363%	0.0459%	1.9054%
4	0.0460%	1.0588%	0.0060%	1.8105%	0.0560%	1.1701%	0.0413%	1.7236%	-0.0233%	1.2561%
5	0.1376%	1.9084%	0.0326%	3.5238%	0.0537%	1.8719%	0.1171%	2.8474%	0.0179%	1.8421%
6	0.1292%	1.9130%	0.0768%	3.7709%	0.0201%	2.2202%	0.1899%	4.1841%	0.0020%	2.1606%
7	0.0776%	1.7759%	-0.0370%	3.8757%	0.0805%	2.0935%	0.0547%	3.5445%	0.1369%	1.9928%
8	0.0574%	1.5324%	-0.0016%	2.3764%	0.0743%	1.4992%	0.0574%	2.1404%	0.0222%	1.5765%
9	0.1134%	1.9693%	-0.0001%	3.0585%	0.0918%	1.6273%	0.0645%	2.2925%	0.1053%	1.6043%
10	0.0664%	1.4950%	0.0024%	2.9838%	0.0705%	1.6071%	0.0601%	2.2454%	0.0831%	1.5457%
11	0.0715%	1.0014%	-0.0086%	2.0292%	0.0597%	1.0749%	0.0345%	1.9672%	-0.0172%	1.3042%
12	0.0409%	2.3850%	-0.0054%	3.4650%	0.1032%	2.1897%	0.0737%	2.9782%	0.2091%	2.2348%

Table 2. Mean daily return (R) and mean volatility of the daily return (V) for each cluster.

Cluster	R_pre_crisis	R_crisis	R_post_crisis	R_covid	R_post_covid
0	0.1080%	-0.0045%	0.0676%	0.0300%	0.0483%
1	0.0509%	0.0014%	0.0823%	0.0586%	0.0394%
2	0.1507%	0.0525%	0.1275%	0.0765%	0.1232%
3	0.0575%	0.0004%	0.0659%	0.0593%	0.0459%
4	0.0460%	0.0060%	0.0560%	0.0413%	-0.0233%
5	0.1376%	0.0326%	0.0537%	0.1171%	0.0179%
6	0.1292%	0.0768%	0.0201%	0.1899%	0.0020%
7	0.0776%	-0.0370%	0.0805%	0.0547%	0.1369%
8	0.0574%	-0.0016%	0.0743%	0.0574%	0.0222%
9	0.1134%	-0.0001%	0.0918%	0.0645%	0.1053%
10	0.0664%	0.0024%	0.0705%	0.0601%	0.0831%
11	0.0715%	-0.0086%	0.0597%	0.0345%	-0.0172%
12	0.0409%	-0.0054%	0.1032%	0.0737%	0.2091%

Table 3. Mean daily return (R) for each cluster with conditional formatting applied. The data presented here is identical to that in Table 2. Column headers corresponding to periods of economic uncertainty, which include the 2008 financial crisis and the COVID-19 pandemic, are highlighted in bright yellow. Conditional formatting is applied to the cells ranging from green to red, where green represents the lowest value in each column while red indicates the highest.

Cluster	V_pre_crisis	V_crisis	V_post_crisis	V_covid	V_post_covid
0	1.2525%	4.3571%	1.4517%	2.5623%	1.6509%
1	1.3486%	2.5077%	1.3061%	2.1582%	1.4482%
2	3.2882%	3.8800%	2.4790%	2.6671%	1.8502%
3	1.1548%	5.4042%	1.7864%	2.7363%	1.9054%
4	1.0588%	1.8105%	1.1701%	1.7236%	1.2561%
5	1.9084%	3.5238%	1.8719%	2.8474%	1.8421%
6	1.9130%	3.7709%	2.2202%	4.1841%	2.1606%
7	1.7759%	3.8757%	2.0935%	3.5445%	1.9928%
8	1.5324%	2.3764%	1.4992%	2.1404%	1.5765%
9	1.9693%	3.0585%	1.6273%	2.2925%	1.6043%
10	1.4950%	2.9838%	1.6071%	2.2454%	1.5457%
11	1.0014%	2.0292%	1.0749%	1.9672%	1.3042%
12	2.3850%	3.4650%	2.1897%	2.9782%	2.2348%

Table 4. Mean volatility of daily return (V) for each cluster with conditional formatting applied. The data presented here is identical to that in Table 2. Column headers corresponding to periods of economic uncertainty, which include the 2008 financial crisis and the COVID-19 pandemic, are highlighted in bright yellow. Conditional formatting is applied to the cells ranging from green to red, where green represents the lowest value in each column while red indicates the highest.

Table 2 shows the mean daily return as well as the mean volatility of daily return for each cluster formed by K-means with correlation matrix and engineered features. The data in Table 2 is split into Table 3 and Table 4 where the mean daily return and mean volatility of daily return are shown separately with conditional formatting to facilitate our analysis.

Overall, we can observe how the clusters may be formed based on the companies' daily return and volatility of daily return. For example, Cluster 4 and Cluster 11 both exhibit a generally lower volatility compared to other clusters. However, Cluster 4 shows lower mean daily returns across most periods, while Cluster 11 has only started to exhibit lower mean daily returns since

the COVID-19 pandemic. On the other hand, Cluster 2 is characterized by relatively higher mean daily returns across most periods but tends to be more volatile than other clusters during certain periods, such as the pre-financial crisis and post-financial crisis periods. Lastly, Cluster 7 is characterized by companies that have lower mean daily returns during the 2008 financial crisis while Cluster 3 is characterized by companies that have higher mean volatility during the same period. Understanding these cluster characteristics can help in tailoring investment strategies. For instance, investors seeking stability might prefer clusters with an overall lower volatility while those seeking higher returns might be willing to accept higher volatility in exchange for a higher return.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.9302	0.2356	0.1409	0.6609	0.2317	0.2680	0.3441	0.3749	0.1381	0.1422	0.5257	0.5265	0.1877
1	0.2356	0.1728	0.0572	0.2366	0.1391	0.0405	0.0452	0.1239	0.1253	0.1292	0.3163	0.1892	0.0960
2	0.1409	0.0572	0.7571	0.1752	0.5262	0.2921	0.5160	0.4297	0.5957	0.2808	0.2422	0.3552	0.5430
3	0.6609	0.2366	0.1752	0.7509	0.0935	0.3624	0.4506	0.4271	0.0932	0.1429	0.5628	0.2333	0.3054
4	0.2317	0.1391	0.5262	0.0935	0.6874	0.2309	0.3652	0.1795	0.5526	0.1978	0.1397	0.6319	0.2669
5	0.2680	0.0405	0.2921	0.3624	0.2309	0.4474	0.5861	0.2629	0.2792	0.1339	0.4458	0.3215	0.3239
6	0.3441	0.0452	0.5160	0.4506	0.3652	0.5861	0.8879	0.3902	0.4181	0.1811	0.4807	0.4158	0.4852
7	0.3749	0.1239	0.4297	0.4271	0.1795	0.2629	0.3902	0.5366	0.2817	0.1970	0.4131	0.1374	0.4340
8	0.1381	0.1253	0.5957	0.0932	0.5526	0.2792	0.4181	0.2817	0.6463	0.2925	0.1249	0.4244	0.4082
9	0.1422	0.1292	0.2808	0.1429	0.1978	0.1339	0.1811	0.1970	0.2925	0.2998	0.2605	0.1612	0.3520
10	0.5257	0.3163	0.2422	0.5628	0.1397	0.4458	0.4807	0.4131	0.1249	0.2605	0.6502	0.2897	0.4195
11	0.5265	0.1892	0.3552	0.2333	0.6319	0.3215	0.4158	0.1374	0.4244	0.1612	0.2897	0.9488	0.1817
12	0.1877	0.0960	0.5430	0.3054	0.2669	0.3239	0.4852	0.4340	0.4082	0.3520	0.4195	0.1817	0.7206

Table 5. Mean correlations between clusters. The mean correlation between Cluster A and Cluster B is obtained by computing the pairwise correlations between each company in Cluster A and each company in Cluster B, then averaging these values. Conditional formatting is applied to the cells ranging from green to red, where green represents the lowest value while red indicates the highest in the whole table.

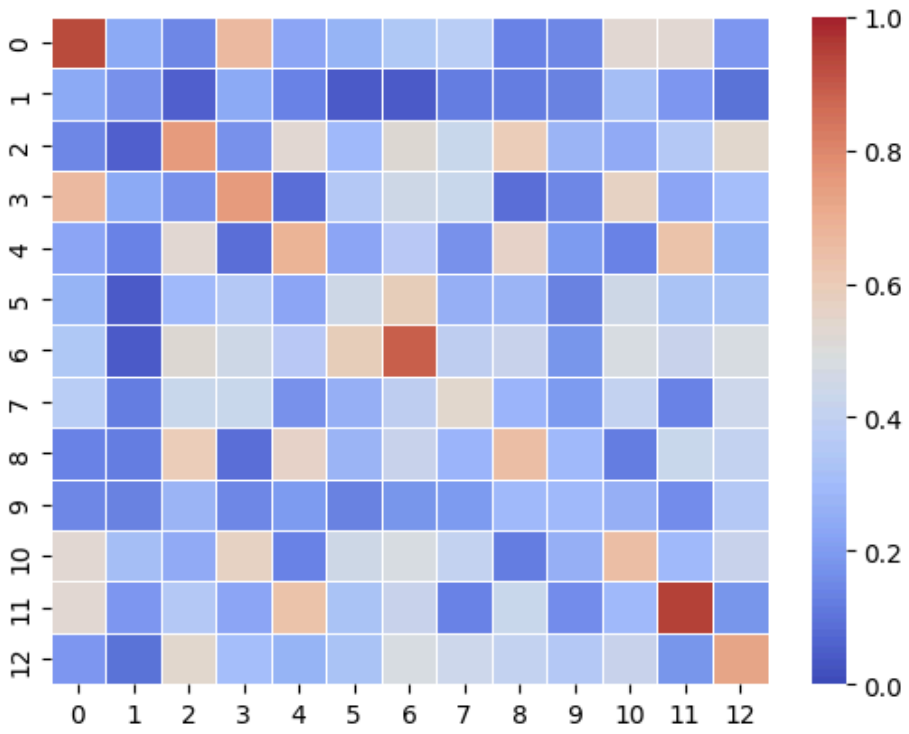


Figure 12. Mean correlations between clusters displayed as a heat map.

The inter-cluster correlations provide valuable insights into how different sectors interact and move together. For example, clusters like Cluster 1 and Cluster 9 have low correlations with most other clusters, making them excellent options for diversification. Including stocks from these clusters in our portfolio along with stocks from other clusters can help reduce overall risk, especially during economic downturns. This is because when stocks from other clusters are underperforming, stocks from Cluster 1 and Cluster 9 may not be affected in the same way, or could even perform well.

On the other hand, intra-cluster correlations are displayed on the diagonal of the correlation matrix in Table 5. Upon closer inspection, we can see that some clusters exhibit low internal cohesion. For example, Cluster 1 has an intra-cluster correlation of 0.1728, which is lower than

its correlations with other clusters such as Cluster 0 (0.2356) and Cluster 10 (0.3163). Similarly, Cluster 9 has a rather low intra-cluster correlation of 0.2998, suggesting that the companies in this cluster do not strongly follow similar patterns in their daily return movements. This lack of strong internal cohesion in Cluster 1 and Cluster 9 suggests that companies within these clusters operate rather independently, making their collective behavior less predictable. While these clusters can still provide opportunities for more diversified investments, investors should consider their internal variability as it may result in more unpredictable outcomes.

In contrast, clusters like Cluster 11 and Cluster 6 have higher intra-cluster correlations which indicate stronger internal cohesion. Cluster 11 has an intra-cluster correlation of 0.9488, indicating that its companies tend to move together more consistently than those in other clusters. Similarly, Cluster 6, with an intra-cluster correlation of 0.8879, shows a high level of internal cohesion. These clusters, characterized by their stronger internal cohesion, may be appealing to investors who are looking for a group of companies with a more uniform behavior. The uniformity in these companies' daily return movement can provide investors with better growth during favorable conditions, since most companies in this group may grow together at the same time. However, such uniformity might also increase exposure to risk in market downturns as companies in this group may all perform badly simultaneously.

In summary, the overall low correlation values between clusters, most of which are below 0.5, indicate that the clusters are well-separated. For more risk-averse investors, investments may be spread across these clusters to minimize risk as downturns in one cluster can be offset by better performance in other clusters. On the other hand, clusters with higher intra-cluster correlations

can be targeted by more risk-tolerant investors for amplified growth during favorable conditions, although at the risk of experiencing bigger losses if the market turns unfavorable. By analyzing both inter- and intra-cluster correlations, investors can balance growth opportunities with risk management.

6.2 Amazon's Cluster Membership Over Time

We will now look at how Amazon's cluster membership changes over the five periods defined in Section 3.2 based on the results given by K-means with correlation matrix and engineered features. For each of the five periods, we calculated the Euclidean distances between Amazon and all 13 clusters.

Period	Top 3 closest clusters and their primary sectors	Top 3 farthest clusters and their primary sectors
Pre-financial crisis	Cluster 0: Technology	Cluster 8: Energy
	Cluster 4: Industrials	Cluster 10: Real Estate
	Cluster 12: Technology	Cluster 1: Utilities
Financial crisis	Cluster 11: Technology	Cluster 0: Real Estate
	Cluster 8: Technology	Cluster 1: Utilities
	Cluster 7: Industrials	Cluster 4: Financial Services
Post-financial crisis	Cluster 12: Technology	Cluster 5: Real Estate
	Cluster 4: Healthcare	Cluster 11: Consumer Defensive
	Cluster 2: Consumer Cyclical	Cluster 3: Utilities
COVID-19 period	Cluster 11: Technology	Cluster 8: Energy
	Cluster 1: Healthcare	Cluster 7: Utilities
	Cluster 6: Consumer Cyclical	Cluster 3: Healthcare
Post-COVID-19 period	Cluster 11: Technology	Cluster 8: Healthcare
	Cluster 5: Industrials	Cluster 9: Energy
	Cluster 10: Financial Services	Cluster 1: Utilities

Table 6. Top 3 clusters closest to Amazon and top 3 farthest clusters, with the farthest cluster listed first for each period.

Period	Cluster where most of the Cyclical Companies are in	Distance between this cluster and Amazon
Pre-financial crisis	Cluster 2	3.308728337
Financial crisis	Cluster 10	4.385433501
Post-financial crisis	Cluster 2	4.070317368
COVID-19 period	Cluster 10	4.913531125
Post-COVID-19 period	Cluster 5	4.594265467

Table 7. Cluster where most of the Cyclical Companies are in and the Euclidean distance between this cluster and Amazon for each period.

Table 6 summarizes the three closest clusters to Amazon as well as the three farthest clusters. For each of these clusters, we also identify the sector that constitutes its majority. Table 7 shows the clusters where most of the Cyclical Companies are in and the Euclidean distances between these clusters and Amazon for all five periods. The data shows an increasing trend in the distances which start from 3.3, followed by lower 4s and eventually going above 4.5.

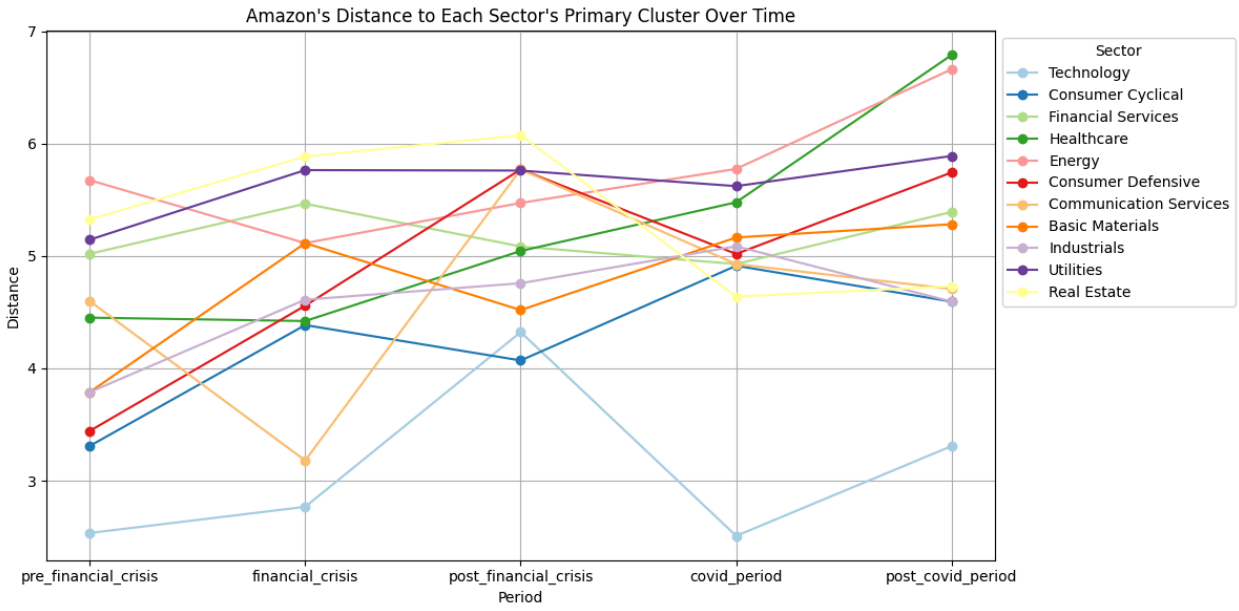


Figure 13. Line chart showing the distances between Amazon and each sector's primary cluster over the five periods. For instance, in the pre-financial crisis period, the cluster for Healthcare is defined as the one where most of the Healthcare companies are clustered in.

Figure 13 shows the distances between Amazon and each sector's primary cluster over the five periods. We can see that the distance between Amazon and the cluster where most of the Healthcare companies are in gets larger as time goes by. The same is true for the cluster where most of the Energy companies are in. The distance between Amazon and the cluster where most of the Consumer Cyclical companies are in dropped temporarily in the post-financial crisis period, but is overall following an increasing trend. On the other hand, the distance between Amazon and the cluster where most of the technology companies are in remain relatively low over all periods.

7 Interpretation of results

7.1 Comparison with GICS

The GICS system categorizes companies into predefined sectors, such as "Retail" and "Technology", based on their primary business activities. However, this static classification does not capture the evolving nature of companies, such as a retailer expanding into the technology sector through the development of advanced e-commerce platforms. Unlike GICS, a clustering approach groups companies based on real-time market behavior and relationships, uncovering relationships that GICS might miss. For example, a company like Amazon might traditionally be classified under Consumer Cyclical in GICS, but clustering might reveal its closer ties to the technology sector due to its significant investment in cloud computing. These discrepancies

show how clustering can uncover relationships and patterns that GICS might overlook, thus offering deeper insights for investors.

Furthermore, the clustering results reveal how certain clusters exhibit higher daily returns or increased volatility during certain periods compared to other clusters. This insight allows investors to identify clusters that are likely to perform better or carry higher risks under particular market conditions. By focusing solely on GICS classifications, investors might miss these crucial insights, leading to less informed portfolio diversification. Therefore, integrating dynamic clustering with traditional classifications can offer a more comprehensive understanding of market dynamics and company behavior.

7.2 Amazon Case Study

Amazon's changing cluster membership over time highlights its evolving role in the market. Founded in 1995 as an online book retailer, Amazon has since diversified significantly, expanding into areas such as smart speaker and cloud computing (Zippia, 2023). This shift is reflected in the clustering results where Amazon's distance from the cluster where most of the Consumer Cyclical companies are in have increased as shown in Table 7.

Throughout all five periods spanning from 2004 to 2024, Amazon consistently belongs to clusters that are predominantly composed of technology companies. Amazon has been heavily involved in the technology sector since launching AWS in 2002 and expanding it with Elastic Compute Cloud (EC2) in 2006. Amazon continued with innovations such as the Amazon Deals app in 2010 and the introduction of the smart speaker in 2014 (Zippia, 2023). Unlike GICS

which is still classifying Amazon as a Consumer Cyclical company, our clustering model is able to capture Amazon's significant involvement in the technology sector.

During the pre-financial crisis and financial crisis periods, Amazon also has a relatively closer distance to clusters that are predominantly composed of Industrials companies. This may be due to Amazon's investment in building a robust logistics and fulfillment infrastructure in the 2000s. For example, Amazon introduced Amazon Prime in 2005, offering students free express shipping, and Fulfillment by Amazon (FBA) in 2006, which allows third-party sellers to store their products in Amazon's fulfillment centers and automate order fulfillment and shipping services (Zippia, 2003; BigCommerce, 2024).

Later on, Amazon has a relatively closer distance to clusters that are predominantly composed of Healthcare companies during the post-financial crisis and COVID-19 periods. This may be due to Amazon's involvement in healthcare since the 2010s. For instance, Amazon partnered with Berkshire Hathaway and JPMorgan Chase in 2018 to launch Haven, a non-profit healthcare organization that aims to improve access to healthcare services to employees and their families. Afterwards, Amazon also acquired the online pharmacy PillPack and Health Navigator, a startup specializing in developing APIs for digital healthcare services. Then, Amazon launched Amazon Care to provide urgent and primary care services for their employees as well as Amazon Halo, a wearable device that tracks temperature and sleep in 2020. Amazon Pharmacy, which offers online pharmacy services, was also launched in 2020 (Hawkins, 2021).

Amazon then went back to having a relatively closer distance to clusters that are predominantly composed of Industrials companies in the post-COVID-19 period. This may be due to Amazon's continued initiatives to improve their logistics and fulfillment infrastructure. For example, in 2023, Amazon relaunched Amazon Shipping, a service offering two to five-day delivery for shipments within the contiguous U.S. Additionally, they introduced Supply Chain by Amazon, an end-to-end supply chain solution available to third-party sellers to use across all sales channels, including those outside of Amazon (Garland, 2023).

In summary, we can see how Amazon's evolving involvement across various sectors is reflected in the clustering results. This dynamic clustering approach provides investors with a more nuanced understanding of Amazon's broad market influence, surpassing the insights offered by the static GICS classification system.

8 Conclusion

This project demonstrated that combining the correlation matrix with engineered features and applying K-Means clustering provides a more representative and dynamic classification of companies compared to traditional methods like GICS. The K-means clustering approach revealed deeper insights into how companies are clustered together based on their market behavior across different economic periods.

By understanding how different clusters behave during various economic conditions, investors can better diversify their portfolios, reducing risk and enhancing returns. For example, clusters

that show low correlations with others can offer stability, while those with high correlations might be used for strategic growth opportunities.

The analysis of Amazon's evolving cluster memberships further highlights the limitations of traditional classification systems and underscores the need for more flexible, data-driven methods. This approach proves especially useful during periods of economic uncertainty, where traditional classifications might overlook crucial shifts in the market or relationships between different companies.

9 Discussion

9.1 Limitations

One of the main limitations of this study is that the clustering was performed only on companies with available stock price data. As a result, the findings in this report are limited to publicly traded companies. Privately held and government-owned companies are not represented in this study, even though industry classifications and market dynamics also exist for these entities. By focusing solely on publicly traded companies, the model used in this study may not fully capture the broader economic landscape, especially in sectors where private companies or government-owned companies play a significant role.

Additionally, this study is limited to companies with complete data over the 20-year period, which may introduce potential bias. While this was necessary to ensure the accuracy of the clustering process, this means that companies that went out of business, were merged, or emerged more recently were excluded. As a result, the resulting clusters mostly reflect the

behavior of stable, long-established companies, potentially underrepresenting more volatile or newly established companies.

Another limitation comes from the static nature of the features used in the clustering process. Although we combined correlation matrices with engineered features, these features rely on historical data, assuming that past performance will indicate future behavior. This assumption might not hold in rapidly changing markets or during unanticipated events like the COVID-19 pandemic, potentially causing the model to miss rapidly emerging trends or shifts in market dynamics and thereby limiting its usefulness as a guide for future investment decisions.

Lastly, the model assumes that linear relationships, as captured by the correlation matrix, are sufficient to describe the complex interactions between companies. However, market dynamics are often non-linear and influenced by various external factors that a correlation matrix might not be able to fully capture. This could lead to overlooking important relationships, resulting in less accurate clustering results.

9.2 Future Work

To ensure the clustering approach is practical and effective, it should be further tested in real-world scenarios such as portfolio management and risk analysis. This includes backtesting the model against traditional GICS-based strategies to evaluate its performance and reliability. Comparing the outcomes of investments performed based on the clustering approach with those based on the GICS approach will help determine whether the clustering method offers positive, tangible benefits over the GICS approach.

Feature engineering can also be further improved. By employing advanced techniques like deep learning or time-series analysis, we can extract more nuanced features that capture non-linear relationships and temporal dependencies between the companies' daily returns. This would allow the clustering model to better reflect the complex and evolving nature of companies' market behavior and potentially produce more accurate and meaningful clusters.

We can also explore more data sources to enhance the model's accuracy and utility. Integrating additional data, such as social media sentiment, could provide insights into public perception and investor sentiment towards certain companies that may impact their stock prices and performance. This may lead to more precise and meaningful classifications, thereby improving the model's practical utility in making investment decisions.

10 References

AlphaLayer. (2024, May 1). *Evolution of Amazon's industry membership*. https://alphalayerai.substack.com/p/amazons-industry-membership-evolution

BigCommerce. (2024, May 26). *Amazon FBA: How it works + cost and maximizing sales.* https://www.bigcommerce.com/articles/omnichannel-retail/amazon-fba/#h2_how_does_amazon_fba_work_

Cai, F., Le-Khac, N., & Kechadi, T. (2016, September 4). *Clustering Approaches for Financial Data Analysis: a Survey*. arXiv.org. https://arxiv.org/abs/1609.08520

Garland, M. (2023, December 20). *Amazon's logistics, delivery services expanded beyond its website in 2023*. Supply Chain Dive. https://www.supplychaindive.com/news/amazon-logistics-delivery-services-expanded-2023-roundup/701643/

Groww. (n.d.). *Adjusted Closing Price.* https://groww.in/p/adjusted-closing-price

Hawkins, L. (2021, July 7). *Amazon's move to healthcare - a timeline*. Healthcare Digital. https://healthcare-digital.com/technology-and-ai/amazons-move-healthcare-timeline

Liu, X., Zhu, X., Qiu, P., & Chen, W. (2012). A correlation-matrix-based hierarchical clustering method for functional connectivity analysis. *Journal of Neuroscience Methods*, *211*(1),

94–102. https://doi.org/10.1016/j.jneumeth.2012.08.016

MSCI. (n.d.). *GICS® - Global Industry Classification Standard*. https://www.msci.com/our-solutions/indexes/gics

Scikit-learn. (n.d.). *KMeans*. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

yfinance. (2024, August 24). *yfinance 0.2.43*. PyPI. https://pypi.org/project/yfinance/

Zippia. (2023, July 21). *Amazon company history timeline*. https://www.zippia.com/amazon-careers-487/history/

11 Appendix

The complete code with its output can be found at https://github.com/Hyshubham2504/Clustering-Project/blob/main/Final_Clustering_Stocks_3_Approaches.ipynb

Individual contributions

Shubham Sahoo

1. Used yfinance to obtain the S&P 500 companies' adjusted closing price data and their corresponding GICS labels.
2. Split the data into 5 different periods to represent different economic conditions as discussed in Section 3.2 above.
3. Performed feature engineering on the daily return data to create 10 new features.
4. Performed hyperparameter tuning on the K-means and hierarchical clustering models that use the engineered features as input.
5. Discussed and wrote down the assumptions of the models (Section 4.3).
6. Computed and analyzed the mean daily return, volatility of each cluster, intra- and inter-cluster correlation (Section 6.1).
7. Interpreted the results and compared them with GICS (Section 7.1).
8. Discussed and wrote down the conclusion and limitations of this project (Section 8 - 9).
9. Wrote down the possible future work for this project (Section 10).
10. Recorded the references used (Section 11).

Zheng En Than

1. Used BeautifulSoup to scrape the SlickCharts website to obtain the latest list of S&P 500 companies.
2. Performed exploratory data analysis on the daily return data and produced plots such as those in Section 3.2 above.
3. Computed the correlation matrix of the daily return data.
4. Performed hyperparameter tuning on the K-means and hierarchical clustering models that use the correlation matrix as input.
5. Discussed and wrote down the assumptions of the models (Section 4.3).
6. Analyzed the three different approaches and wrote down the reasoning for choosing the third approach (Section 5.1 - 5.4)
7. Analyzed and plot Amazon's cluster membership over time (Section 6.2).
8. Researched on Amazon's past business ventures and related them to the clustering results (Section 7.2).
9. Discussed and wrote down the conclusion and limitations of this project (Section 8 - 9).
10. Recorded the references used (Section 11).