2020

K-Means Stock Clustering Analysis Based on Historical Price Movements and Financial Ratios

Shu Bin
Claremont McKenna College

Claremont McKenna College

K-Means Stock Clustering Analysis Based on Historical Price Movements and Financial Ratios

written by

Shu Bin

submitted to

Professor Chiu-Yen Kao

Senior Thesis

2020 Spring

May 11, 2020

Abstract

The 2015 article *Creating Diversified Portfolios Using Cluster Analysis* [18] proposes an algorithm that uses the Sharpe ratio and results from K-means clustering conducted on companies' historical financial ratios to generate stock market portfolios. This project seeks to evaluate the performance of the portfolio-building algorithm during the beginning period of the COVID-19 recession. S&P 500 companies' historical stock price movement and their historical return on assets and asset turnover ratios are used as dissimilarity metrics for K-means clustering. After clustering, stock with the highest Sharpe ratio from each cluster is picked to become a part of the portfolio. The economic and financial implications of the clustering results are also discussed. In the end, portfolios constructed with clustering results of stocks' historical price movements perform poorly, but portfolios constructed with clustering results of companies' financial ratios consistently exceed market average. Using an alternative portfolio construction method that represents each cluster proportionally with regards to their sizes, portfolios constructed with historical stock price movements gain an increase in performance, while the returns of portfolios constructed using companies' financial ratios decrease. Further studies should be done with a different portfolio performance index and a larger dataset.

Contents

1 Introduction

In the field of stock market investing, investors often seek to diversify their portfolio in order to minimize massive losses due to a black swan event. Diversifying a portfolio is the practice of purchasing securities that are exposed to different risks, as opposed to putting all the eggs into one basket. In doing so, the hope is that if a black swan event negatively impacts the price of one stock, the other stocks will be unaffected or even have inverse price movements, due to them not being exposed to the same risks. Although speculators might try to achieve this through naive diversification, a market-tested diversification method that's backed up by reasoning, such as the sector diversification method of ten to thirty stocks suggested by Benjamin Graham [13], is often preferred. Due to the randomness of the market, there is no single "standard" method of portfolio diversification, even during times of bright economic outlook. In 1952, the modern portfolio theory proposed by Harry Markowitz [17] gave foundation to a myriad of mathematical methods of maximizing a portfolio's expected return while maintaining the same level of risk exposure, or minimizing the risk while maintaining the same level of expected returns. One method to determine the similarities of the risks that different stocks are subjected to is using K-means clustering on stocks' historical price movements. This paper seeks to assess the validity of K-means clustering with different dissimilarity metrics on a sample of S&P 500 component companies' stocks. The dissimilarity metrics include stocks' historical daily price movements and companies' historical return on assets and asset turnover ratios. A diversified portfolio is then constructed with the combination of the clustering result and the stocks' Sharpe ratio. Two portfolio-building methods are tested. The first method picks one stock from each cluster of the K-means partitioning. The second method represents each cluster of stocks proportionally with regards to their sizes. Validity of the algorithm is examined in the current economy impacted by the COVID-19 shutdown.

2 Literature Review

The 2015 paper *Creating Diversified Portfolios Using Cluster Analysis* [18] proposes a portfolio diversification method using K-means clustering. Instead of using daily stock movements as the dissimilarity metric, the paper proposes using companies' historical return on assets and asset turnover ratios as the dissimilarity metric. The reasoning is that clustering approaches that are based on stock prices' correlation, such as the ones proposed in *Portfolio Construction Using Clustering Methods* [19] and *Correlation Based Clustering of the Stockholm Stock Exchange* [21], are "not stressful in nature" because these studies were conducted before the great recession of 2008. Global economic crises such as the 2008 recession often wipe out years of growth in a matter of months. To put matters into perspective, in March 2009, the Dow Jones index fell 7721.16 points or 54% from its October 2007 high of 14164.43 [8]. The S&P 500 index experienced a similarly large decline of 38.5% in 2008 [18]. The 2008 recession wiped out more than 10 years of economic growth in under 18 months and saw collapses of industry giants such as Lehman Brothers. According to statistician, trader, and creator of the black swan theory Nassim Nicholas Taleb, such low probability events are usually considered statistical outliers, but are often much more impactful than events that are commonplace and easy to predict [26]. Thus, a good investment portfolio should always take the possibility of high-impact, low-probability events into consideration [27]; in this case, potential impact of unforeseen economic crises need to be considered when creating a portfolio. [18] claims that a company's return on assets and asset turnover ratios are much better indicators of a company's financial health during times of stress. Thus, instead of using historical stock price data, [18] proposes using a weighted average of companies' quarterly reports of those two ratios as the dissimilarity metric for clustering.

To implement the theory proposed in [18], K-means clustering with Lloyd's algorithm is used. Quarterly data of 668 companies listed on NYSE and NASDAQ from 2000 to 2015 is collected to calculate each company's return on assets and asset turnover ratios. The two ratios are combined into one dissimilarity metric by having a 50% weight assigned to each variable. Euclidean distance is used to measure the distance between points. After partitioning, the stock with the highest Sharpe ratio from each cluster is picked to form a portfolio. It is shown that, overall, the portfolio constructed by the algorithm has higher volatility than the S&P 500 index but lower volatility than

the average volatility of all stocks tested. The study also finds that the algorithm portfolio holds up well during the 2008 financial crisis by having a lower volatility than the average. In an ending remark, [18] notes that further test of the algorithm should be conducted over a longer time span and other periods of financial stress. However, it should be noted that the methodology of "stress testing" in [18] is not rigorous. That is because data during the 2008 recession is included in the training of the clustering algorithm. Financial crises are, by definition, unforeseen events. In the real world, it would be impossible to have data on a future financial crisis and use that data to adjust the clustering algorithm and portfolio accordingly. Thus, it is crucial to separate the data before and during a financial crisis as training and testing data. The currently ongoing recession caused by the COVID-19 shutdown provides an excellent testing ground for the performance of a few variations of the algorithm proposed in [18].

3 Methodology

This project seeks to enhance and implement the algorithm proposed by [18] by using a more rigorous way of choosing the number of clusters K and an alternative method of portfolio construction. To determine the effectiveness of companies' financial ratios as a dissimilarity metric, clusterings with financial ratios and companies' daily stock price movements are ran separately. Elkan's algorithm [9] is used to optimize each cluster and K-means++ [6] is used to find initial points. The value of K is chosen by using the Silhouette method [22]. In order to form portfolios from the clustering results, companies' Sharpe ratios are calculated. One stock with the highest Sharpe ratio in each cluster is selected, as is consistent with the methodology in [18]. Alternatively, portfolios that represent each cluster's size proportionally are also constructed. To evaluate the performance of the portfolios, their returns during the beginning period of the COVID-19 shutdown (2-3-2020 to 4-14-2020) is compared to S&P 500 return of the same period. Furthermore, clustering outliers are investigated by checking the corresponding companies' historical financial reports. This section will give a detailed description of the algorithms and financial ratios used.

3.1 Return on assets and asset turnover

Return on assets and asset turnover are two crucial ratios in gauging a company's finance health. They are calculated [7] as

$$\text{Return on assets} = \frac{\text{Net income}}{\text{Total assets}},$$
$$\text{Asset turnover} = \frac{\text{Revenue}}{\text{Total assets}}.$$

Here, revenue is the total amount of sales by a company during a period, and net income is revenue with all operating expenses deducted. These expenses usually include items such as cost of goods sold, salaries payable, depreciation, income tax expense, and others, although the specifics of the items are decided by individual companies while following Generally Accepted Accounting Principles as required by the SEC. Total assets is the total book value of the assets a company possesses. Some

usual items in a company's list of assets include cash and cash equivalents, short and long term investments, accounts receivable, inventory, PP&E (property, plant, and equipment), and goodwill.

Because a company's net income and revenue typically scale with its size, it is important to assess these two numbers with a frame of reference. By dividing net income by total assets, return on assets assesses the profitability of a company and how effective it is at using its assets to generate income. On the other hand, asset turnover measures how effective a company is at using its assets to generate sales. Typically, a higher return on assets ratio indicates a higher profit margin, suggesting that the company has an effective business model. As such, return on assets and asset turnover are important in determining a company's efficiency and profitability [7].

3.2 Sharpe ratio

The Sharpe ratio was first proposed by William F. Sharpe in 1966 as an index for evaluating mutual fund performance [23, 24]. Let a be a portfolio, R_a the rate of return of portfolio a, and R_f the risk-free rate of return. The Sharpe ratio of a is defined as

$$S_a = \frac{\mathbb{E}(R_a - R_f)}{\sigma(R_a - R_f)}.$$

Here, $(R_a - R_f)$ is the rate of return above the market risk-free rate (a.k.a. excess return), and $\sigma(R_a - R_f)$ is the standard deviation of the excess return, which is used as a proxy for the risk of portfolio a. In industry practice, it is commonplace to use the rate of return of U.S. treasury bills of the same investment length as the risk-free rate of return. The Sharpe ratio follows modern portfolio theory's insight that risk and return should always be assessed together [17], and provides a benchmark to evaluate a portfolio's return relative to its risk [23]. For this project, each stock's excess return is derived by deducting 1-month U.S. treasury bills' rate of return (R_f) from the monthly rate of return for the stock (R_a). Each stock's Sharpe ratio is then calculated.

3.3 K-means clustering

The K-means algorithm is one of the most well-known partitional clustering algorithms. This subsection will give a brief description of the K-means algorithm. The very first versions of the K-means

algorithm were proposed by Edward Forgy in 1965 [11] and James McQueen in 1967 [16], and the most wide-spread version of the algorithm was published by Stuart Floyd in 1982 [15].

3.3.1 Lloyd's algorithm

Lloyd's algorithm is often considered the "standard" K-means algorithm and is the one used in [18]. Below is a description of Lloyd's algorithm, also known as the naive K-means algorithm. This summary closely follows the one in *Clustering* by Xu and Wunsch [29].

Given a set of points $\{x_1, \cdots x_n\}(x_i \in \mathbb{R}^m)$,

1. Initialize K number of clusters $\{C_1, \cdots C_K\}$ with centers $\{m_1, m_2, \cdots m_K\}(m_i \in \mathbb{R}^m)$. The centers can be picked randomly or calculated based on some methods.

2. For all points $x_i(i \in \{1, \cdots n\})$, find the center closest to it based on some distance metric d. Assign x_i to the cluster corresponding to the closest center. In other words:

$$x_i \in C_j \text{ if } d(x_i, m_j) \leq d(x_i, m_l) \quad (\forall l \in \{1, \cdots K\})(j \neq l)(\forall i \in \{1, \cdots n\}).$$

3. Recalculate the center for each cluster $C_l(l \in \{1, \cdots K\})$. The new cluster centers are the mean of the sum of all points in the same cluster. In other words:

$$m_l = \frac{1}{|C_l|} \sum_{x_p \in C_l} x_p \quad (\forall l \in \{1, \cdots K\}).$$

4. Repeat step 2 and 3 until no cluster has any change in point assignment.

3.3.2 The silhouette method

To determine the value of K for the best clustering results, the silhouette method is used [22]. This method evaluates the overall goodness of fit of a partitioning. A summary of the silhouette method can be seen below. This summary closely follows the one presented by Rousseeuw in 1986 [22].

Given n data points $\{x_1, \cdots x_n\}$, a partitioning result of K clusters $\{C_1, \cdots C_K\}$, and distance

metric d, for each point x_i in cluster C_l, define

$$a(x_i) = \frac{1}{|C_l| - 1} \sum_{\forall x_j \in C_l, i \neq j} d(x_i, x_j).$$

Here, $a(x_i)$ is the mean dissimilarity between x_i to all other points within the same cluster.

For each point x_i in cluster C_l, define

$$b(x_i) = \min_{\forall p \in \{1, \cdots K\}, p \neq l} \frac{1}{|C_p|} \sum_{\forall x_j \in C_p} d(x_i, x_j).$$

Here, $b(x_i)$ is the minimum mean dissimilarity between x_i and all points in some cluster C_p which does not contain x_i. In other words, $b(x_i)$ is the mean dissimilarity to the closest cluster that x_i is not assigned to.

For each point x_i in cluster C_l, define their silhouette value as

$$s(x_i) = \begin{cases} \frac{b(x_i) - a(x_i)}{\max(b(x_i), a(x_i))} & \text{if } |C_l| > 1, \\ 0 & \text{if } |C_l| = 1. \end{cases}$$

or

$$s(x_i) = \begin{cases} 1 - \frac{a(x_i)}{b(x_i)} & \text{if } a(x_i) < b(x_i), \\ 0 & \text{if } a(x_i) = b(x_i), \\ \frac{b(x_i)}{a(x_i)} - 1 & \text{if } a(x_i) > b(x_i). \end{cases}$$

It should be noted that $s(x_i) \in [-1, 1]$. For the silhouette value to approach 1, $a(x_i)$ needs to be significantly smaller than $b(x_i)$, meaning that within-cluster mean dissimilarity is much less than the smallest between-cluster mean dissimilarity, and thus the model does a good job clustering similar points together. For the silhouette value to approach 0, $a(x_i)$ needs to be significantly greater than $b(x_i)$, meaning that within-cluster mean dissimilarity is much greater than the smallest between-cluster mean dissimilarity, and thus the model does a poor job clustering similar points together.

The overall goodness-of-fit of the clustering is measured by the average silhouette value of all points

$$\overline{s} = \frac{1}{n} \sum_{i=1}^{n} s(x_i).$$

The average silhouette value is calculated for a number of K values. Although the clustering result of the K values that corresponds to the highest average silhouette scores should, on average, have the lowest within-cluster dissimilarity, the final values of K are chosen with regards to a combination of high average silhouette score and how realistic the number of clusters is in a real-world stock market setting.

4 Data

Historical stock and financial data is collected on S&P 500 component companies as of April 4, 2020. The data and code used for this project can be accessed through the links in Appendix 4. However, it should be noted that the list of S&P 500 component companies is constantly updated to reflex changes in companies' market cap and private or public status. To avoid discontinuity of data in the case that a component company only just became public in the past few years, all companies that weren't consistently a part of the S&P 500 index since 2006 were dropped from the dataset. Because the changes in component companies are not made publicly available by S&P Global, historical changes of Ishares Core S&P 500's holdings are used as a proxy. Ishares Core S&P 500 (IVV) is an ETF that seeks to match the S&P 500 returns by holding the exact same stocks that make up the S&P 500 index. The fund usually updates its portfolio with a one-day delay after S&P 500 updates its list of component companies, and is thus deemed an adequate proxy for changes in S&P 500 components. Data provided by Newport Quantitative Trading and Investments on IVV's historical portfolio changes is used to determine S&P 500 historical component changes. There have been 904 changes to the list of component companies from 9-29-2006 to 4-3-2020. In the end, 268 companies were dropped and only 232 companies remain.

For the remaining 232 companies, their historical daily stock data from 10-31-2006 to 4-14-2020 is collected from Yahoo Finance. The stock data includes the daily opening price, closing price, adjusted closing price, trade volume, high, and low. A total of 3385 days of data is observed for each company. It should be noted that the price level of stocks can vary significantly from company to company. For example, on 10-31-2006, Google's share price opened at \$238.43 and closed at \$233.98, a \$4.45 or 1.9% decrease. On the same day, Gilead Sciences' stock price went from \$17.23 to \$16.89, a \$0.34 or 2% decrease. Although the \$4.45 decrease in Google's stock value is more than ten times that of Gilead Sciences' stock value, the two stocks' daily movements should be treated similarly by the clustering algorithm due to their similarity in percentage changes. A company's stock value is calculated with the following formula:

$$\text{Market stock price} = \frac{\text{Market cap}}{\text{Number of shares outstanding}}.$$

While the market cap should always roughly reflect the true value of the company, and is thus not something easily changed, the number of shares outstanding is determined by the corporation's management. This makes the magnitude of a company's stock price a somewhat arbitrary number. Therefore, daily percentage price movement is a better metric for reflecting a stock's price trends, as opposed to raw price movements. All companies' historical price data is thus normalized with the following formula:

$$\text{Daily percentage price movement} = \frac{\text{Daily closing price} - \text{Daily opening price}}{\text{Daily opening price}}.$$

Companies' quarterly balance sheet and income statement data from 2009 and onward is collected from an API provided by The Financial Modeling Prep Company.[1] Companies' quarterly net income, revenue, and total assets are then extracted from the financial statements. Their quarterly return on assets and asset turnover ratios are calculated from the data. Because many companies have different fiscal year cycles and thus don't report their financial data at the same time, an exact matching of the reporting times of all companies' historical ratios would be impossible. Effort is made to match ratios that were reported no more than 3 months apart. After matching, the data roughly spans from Q3 2009 (August - October, 2009) to Q3 2019 (August - October, 2019), or 41 quarters in total. 109 more companies are dropped from the dataset due to missing values, leaving the financial ratio dataset with 114 companies. Each company's return on assets data and asset turnover data are then concatenated together to form a single row vector of length 82. In other words, let $ROA_i, AT_i \in \mathbb{R}^{41}$ be the return on assets and asset turnover data of company i, respectively, then the data entry for company i in the financial ratio dataset is

$$\begin{bmatrix} ROA_i^\top & AT_i^\top \end{bmatrix} \forall i \in \{1, \cdots, 114\}.$$

The nature of the financial ratios means that they are already scaled to the companies' sizes. The financial ratio dataset thus need not be normalized. In the end, the financial ratio dataset is a 114×82 matrix.

[1] The API can be accessed at https://financialmodelingprep.com/developer/docs/.

5 Results and Discussion

This section will present and analyze the clustering results of the K-means algorithms. Outlier clusters and companies are investigated, and performance of the constructed portfolios are examined. For the clustering results, see Appendix 4. For illustrations and graphs of the clustering results, see Appendix 1. A description of the constructed portfolios can be found in Appendix 2 and 3.

5.1 Clustering results of stock price movement dataset

Initially, clustering of the entire stock price movement dataset from 9-29-2006 to 4-3-2020 is conducted. However, the clustering results has very high within-cluster sum-of-squares (WCS), and the density histogram of groups is far from uniformly distributed. Instead, over 75% of the observations fall into two groups, while some of the other groups would have very few points in them. To analyze this phenomenon further, the stock data is reduced to $\mathbb{R}^3$ through principle component analysis (PCA), and each observation is plotted and color-coded according to their group. Through visual inspection, there are no apparent borders between each of the groups. Thus, the stock data clustering result is deemed poor. This is likely due to the stock dataset containing data of the 2008 recession and 2020 COVID-19 recession, during which the stock market average return dropped by more than 50% and 34%, respectively. Most stocks saw similarly large declines during those two periods. Those large price shocks are likely what causes most of the data to be grouped together. To improve the clustering results, data before the end of the 2008 recession and data after the start of the COVID-19 recession are dropped. The dropped stock data includes data before 7-1-2009 and data after 2-20-2020.[2] Afterwards, 2764 days of stock price movement data remain.

To determine the optimal value of K for the updated stock dataset, values from 4 to 11 are tested. For each of the K values, the K-means algorithm is ran 1000 times, and the average silhouette value for the clustering results is recorded. The result is shown in Figure 1 in Appendix 1. As is apparent in the figure, $K = 4$ yields the optimal silhouette value and $K = 6$ yields a local optima. Although a natural choice is $K = 11$, as S&P Global divides all companies into 11 major sectors, this is not

[2] According to the National Bureau of Economic Research, the 2008 recession officially ended in June of 2009 [1]. While there is no official date for the start of the COVID-19 recession, many believe that it started on 2-20-2020, the day when the global stock market crashed.

optimal since $K = 11$ has a relatively small average silhouette value.

5.1.1 Clustering result of $K = 4$

Picking $K = 4$, 5000 iterations of the clustering algorithm are ran. The iteration with the lowest WCS at 83.306 is chosen as the best iteration. Figure 2 shows the distribution of the stocks between groups, and Figure 3 is a scatterplot of the companies, color-coded by groups. Over 75% of the observations are clustered into group 2 and 3. Group 4 contains only 11 observations, the smallest of any group. All companies in group 4 in fact belong to the energy sector. As can be seen in Figure 3, these companies' stocks are outliers that are far away from the other groups.

Further investigation reveals that stock price trends in the energy sector had been behaving notably differently from most other sectors. Instead of seeing a steady increase in stock prices as most other sectors have enjoyed during the longest economic expansion in U.S. history, the energy sector's price trend since 2009 has been turbulent. The prices of almost all stocks in group 4 saw sharp declines in 2016 and 2018, and has been steadily decreasing since 2019. These stocks' sharp declines in 2016 and 2018 were caused by a surge in U.S. shale oil production. U.S. shale oil production had been steadily increasing since 2014, driving U.S. to become the largest crude oil producer in the world. However, this surge in production also drove market supply up and caused crude oil prices to decrease significantly. Between 2011 and 2014, West Texas Crude (WTI), price benchmark of U.S. light crude oil, averaged above 90$ per barrel. Due to the shale oil production surge, WTI went down to a 13-year low of 26.55$ per barrel. This is even lower than the price seen during the 2008 recession's energy bear market, when WTI was priced around 30$. In 2018, the surge in shale oil production drove the U.S. oil market into a bear market again, causing a 20% decline in WTI [5]. During these two oil bear markets, many U.S. oil companies filed for bankruptcy, while others needed to save cash to survive through the bear markets and did not have capital to expand their business, driving their stock prices down [12]. In 2019, the energy sector is hit by the gloomy global economic outlook caused by the China-U.S. trade war. The projected low global economic growth caused energy demands to fall, yet again harming the earnings and stock prices of U.S. oil companies [20].

5.1.2 Clustering result of $K = 6$

In order to further break down the groups, another clustering is conducted with $K = 6$. $K = 6$ is chosen due to it being a local optima in terms of average silhouette value, thus the clustering would still have a relatively low WCS value. The clustering result has a WCS value of 78.3192. A histogram of the distribution of the observations can be seen at Figure 4. Similar to the grouping results of $K = 4$, more than 160 of the observations fall into two large groups. Observations in the largest group with more than 100 observations remain largely unchanged from that of the results of $K = 4$. The smaller groups are group 5, 4, 3, and 2. Group 5 has 22 observations, which are all companies in the financial services sector. Group 4 has 21 observations, which all belong to the energy sector. Group 3 has 10 observations, which are all real estate companies. Group 2 has 11 observations and is identical to the outlier group (group 4) of the clustering result of $K = 4$.

Although all observations in group 5 technically belong to the energy sector, they are fundamentally different from the companies in group 2. While group 2's companies are all oil producers, group 5's companies provide electricity and natural gas to residents and businesses. These companies have benefited from the longest economic expansion in U.S. history and their stock values have been slowly but steadily increasing since 2008. For the real estate companies in group 3, their stock prices have been largely stagnant. One possible explanation is the increase in government spending since 2008. Real estate companies are mostly low-risk and slow-growth businesses that usually do not have sudden boosts to their income like technology companies. As U.S. government spending increased since 2008, government borrowing also increased. This caused more treasury bills to be issued, lowering the price and increasing the risk-free yield, effectively decreasing the excess return of real estate stocks in the process. This has drawn many investors away from investing in real estate companies despite the ever-increasing home prices, leading to mostly stagnant stock price trends in the sector [28].

5.2 Clustering results of financial ratio dataset

The average silhouette score plot of the financial ratio dataset (Figure 6) shows that $K = 4$ is the optimal number of clusters. To further divide up the clusters, another clustering is conducted with

$K = 7$.

5.2.1 Clustering result of $K = 4$

The best iteration of clustering with $K = 4$ has a WCS value of 45.0159. The histogram (Figure 7) shows that there is one group with only two observations, Cardinal Health and Costco Wholesale. These two companies stand out due to their unusually high historical asset turnover ratios. While most other companies' asset turnover can barely reach 0.2, Cardinal Health and Costco Wholesale's asset turnover ratios are consistently above 0.7. This indicates that both companies are efficient at generating sales given a fixed amount of assets. However, the two companies' high operating costs mean that their return on asset ratios do not stand out when compared to the rest of the dataset.

5.2.2 Clustering result of $K = 7$

The best clustering result for $K = 7$ has a WCS value of 27.6499. Cluster 3, 4, 6, and 7 are notable as they contain very few observations. Cluster 4 and 6 each only have one observation, which are Cardinal Health and Costco Wholesale. Although these two companies have similarly high asset turnover ratios, Costco Wholesale's return on asset ratios are much higher than that of Cardinal Health, indicating that Costco has a considerably higher profit margin. Cluster 7 includes Gap Inc, Home Depot, and Texas Instruments. These three companies are clustered together because their asset turnover ratios are higher than average, but lower than those of Cardinal Health and Costco Wholesale. Cluster 3 tells a similar story. It includes Amazon.com, Humana Inc, McDonald's, UnitedHealth Group, and Xerox Corp. While all of these companies' asset turnover ratios are higher than average, they are lower than those of companies in cluster 4, 6, and 7.

5.3 Portfolio performance

After constructing a portfolio for each clustering result, the portfolios' returns during the COVID-19 period are calculated. The portfolios are comprised of one stock with the highest Sharpe ratio from each cluster. The tables in Appendix 2 give a detailed summary of each portfolio's makeup, return, and return in excess of the market average of -12.40%. The portfolios constructed from clustering results with the stock price movement data perform poorly, each under-performing in

relation to the market average by 13.80% and 17.52%. This is due to the algorithm choosing a few poorly performing stocks. Petroleum company Apache Corporation (APA) alone, with its -70.68% decrease in stock value since the start of the COVID-19 period, has a significant and negative impact on portfolio B's performance.

On the other hand, portfolio C and D constructed from the clustering results with the financial ratio dataset consistently beat market average. Although the returns of both portfolios are still negative, they exceed market average of the same period by 8.72% and 2.63%, respectively. This result is in agreement with the conclusion in [18]. The performance of portfolio C and D are considered satisfactory.

Although [18] intended for the portfolio creation process to be completely left to the algorithm, it should be noted that some human intervention and investigation into why some stocks are clustered together can lead to better portfolios. For example, in the clustering result of the financial ratio dataset with $K = 7$, four groups of ten companies are outliers. A quick investigation finds that these companies are outliers because of their unusually high asset turnover ratios, indicating that they are more effective at generating sales than the other companies. Some further investigation will reveal that, among these ten companies, Cardinal Health, Costco Wholesale, and Amazon Inc. have some of the highest projected yearly revenue growth rates. Yahoo Finance projects the three companies' revenue growth rates in 2020 to be 4.7%, 7.3%, and 22.3%, respectively [3, 4, 2]. A portfolio that consist of these three companies has an average return of 5.98% during the beginning of the COVID-19 period, outperforming the market average by 14.38%. Thus, although the stock-picking algorithm performs admirably well, some human intervention and financial analysis can improve its performance even further.

5.4 Alternative approach to portfolio construction

Because the methodology proposed in [18] is to pick only one stock from each cluster, the clusters aren't represented proportionally according to their sizes. Larger clusters are under-represented in the portfolio, and smaller clusters are over-represented. Theoretically, creating a portfolio that proportionally represents each clusters' sizes should increase diversification and improve portfolio return. To be more precise, this alternative method constructs portfolios in the following way. Let A

be a partitioning that divides n data points into four clusters $\{C_1, C_2, C_3, C_4\}$. Let the clusters' sizes be $\begin{bmatrix} |C_1| & |C_2| & |C_3| & |C_4| \end{bmatrix}$. Thus, the percentage of points in C_1, C_2, C_3, C_4 are $\frac{|C_1|}{n}$, $\frac{|C_2|}{n}$, $\frac{|C_3|}{n}$, $\frac{|C_4|}{n}$, respectively. A total number of ten stocks will be picked from the clusters to form the portfolio, as is consistent with Graham's writing in [13] that a minimum of ten stocks should be in each portfolio for diversification. Hence, the number of stocks picked from clusters C_1, C_2, C_3, C_4 are $\frac{10|C_1|}{n}$, $\frac{10|C_2|}{n}$, $\frac{10|C_3|}{n}$, $\frac{10|C_4|}{n}$, respectively. Stocks with the highest Sharpe ratios in each cluster are chosen. The number of stocks are rounded whenever necessary. Portfolio E and F are constructed from clustering results of the stock dataset and portfolio G and H are constructed from clustering results of the financial ratio dataset. Appendix 3 shows the alternative portfolios' makeups and returns.

Using the alternative method, portfolios constructed from the stock price movement dataset see a significant increase in performance. Portfolio E has an excess return of 3.91% as opposed to portfolio A's −13.80%, and portfolio F has an excess return of −0.87% as opposed to portfolio B's −17.53%. This is likely due to the increased diversification of the portfolio. In the stock price movement dataset, companies are more likely to be clustered together when their stock prices experience similar positive and negative shocks at similar times. This indicates that companies within the same cluster are exposed to comparable business risks. By representing the clusters proportionally, the portfolio is diversified, and thus its performance improves.

On the other hand, portfolios constructed from the financial ratio dataset see a decrease in its average returns. Portfolio G has an excess return of −5.75% as opposed to portfolio C's 8.72%, and portfolio H has an excess return of −10.89% as opposed to portfolio D's 2.63%. For the partitioning results of the financial ratio dataset, stocks with similar asset turnover and return on assets ratios are clustered together. This means that companies within the same cluster have comparable earning power. As analyzed in *Section 5.2*, the outlier groups that contain small numbers of points are often the companies with the higher ratio values. Companies within those clusters are arguably more profitable than those in the larger clusters. Thus, by over-representing the smaller clusters, the method proposed in [18] increases the proportion of high-earning stocks in the portfolio. For the same reason, the alternative portfolios perform poorly because relatively low-earning stocks from the larger clusters are proportionally represented.

6 Conclusion

In this project, the portfolio construction algorithm proposed in *Creating Diversified Portfolios Using Clustering Analysis* [18] is implemented and stressed tested with data from the beginning period of the COVID-19 recession. Although the clustering results by using historical stock price movements as the dissimilarity metric were poor, portfolios constructed with companies' financial ratios as the dissimilarity metric consistently beat the S&P 500 market average. Alternatively, when each cluster is represented proportionally according to their sizes by the portfolios, portfolios constructed using the stock price movements gain a boost in returns due to increased diversification, while portfolios constructed using the financial ratio dataset see a decrease in returns. It should be noted that the algorithm used in this study is limited by a few factors. First, the Sharpe ratio is one of the fundamental building blocks of the algorithm's stock-picking process. A few scholars have pointed out some limitations of the Sharpe ratio, such as that it is directly impacted by the length of the investment horizon [14], its unrealistic assumption that portfolio returns are normally distributed [25], and that it is only a indicator of a portfolio's past performance [10]. The validity of other portfolio performance indicators, such as the one proposed by Stutzer in 2002 [25], should be tested with the Sharpe ratio. This project is also constrained by the limited amount of data available, and only a maximum of 232 companies' data are used for clustering. Because results derived from 232 companies is hardly conclusive for the entire stock market, this project can be expanded to a more realistic scale if more data is available. Further studies should be conducted with these limitations in mind.

Appendix 1: Figures and Illustrations

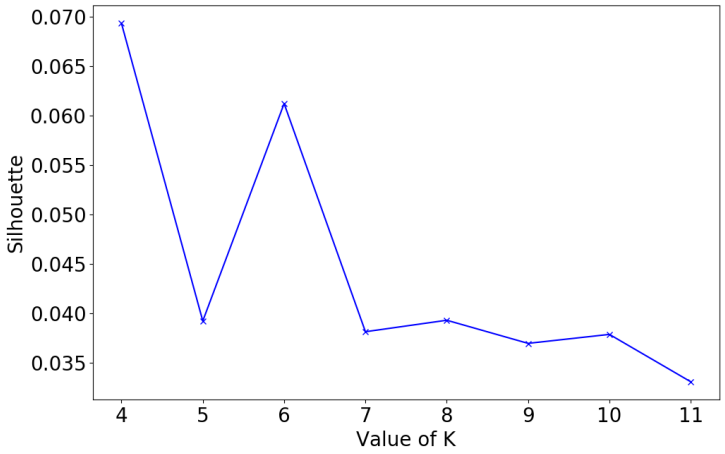

Figure 1: Average silhouette score of 1000 clustering results, stock data

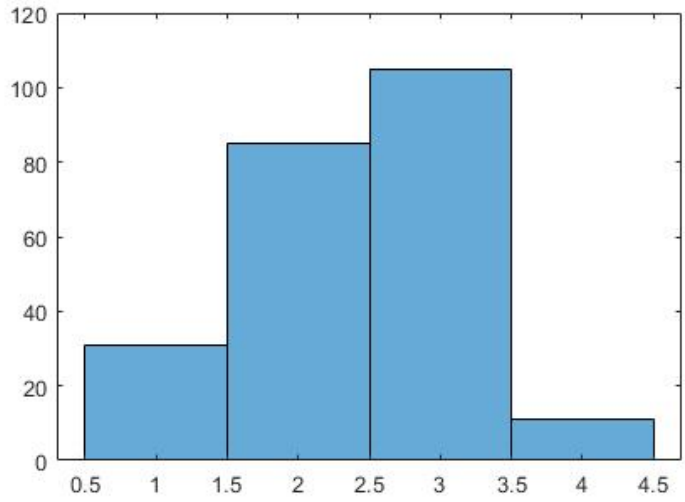

Figure 2: Histogram of the distribution of stocks in groups, stock price movement dataset, $K = 4$

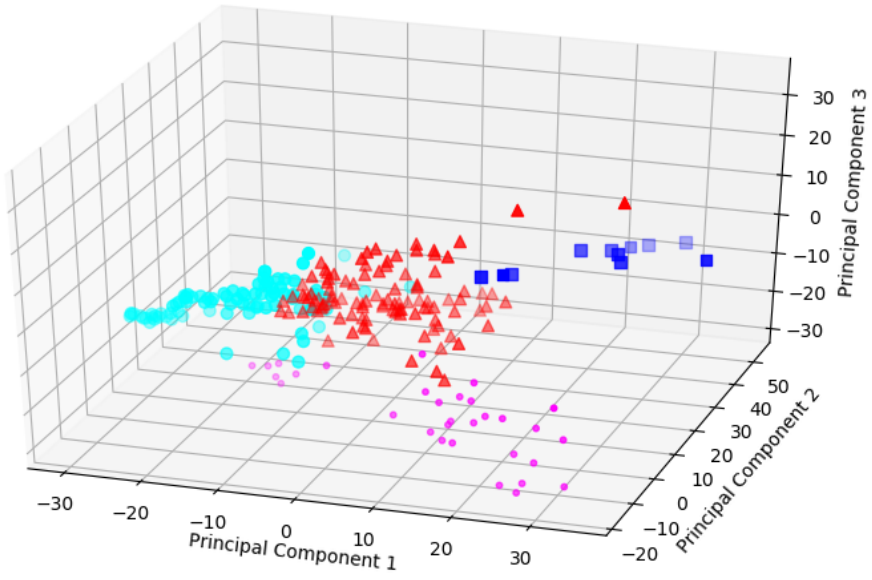

Figure 3: Scatterplot of stock price movement dataset, $K = 4$

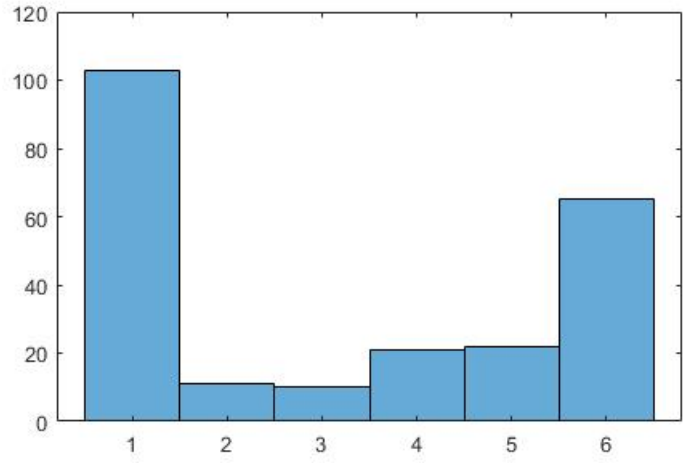

Figure 4: Histogram of the distribution of stocks in groups, stock price movement dataset, $K = 6$

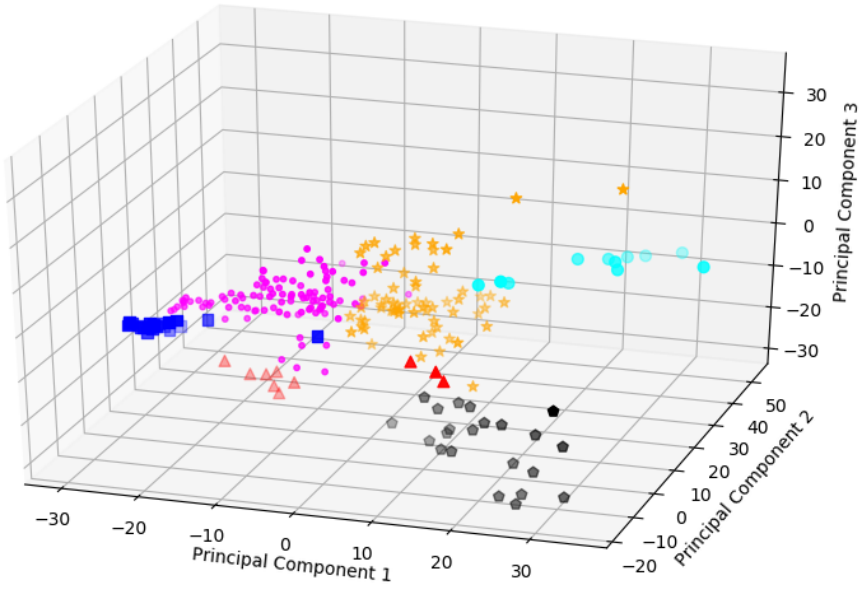

Figure 5: Scatterplot of stock price movement dataset, $K = 6$

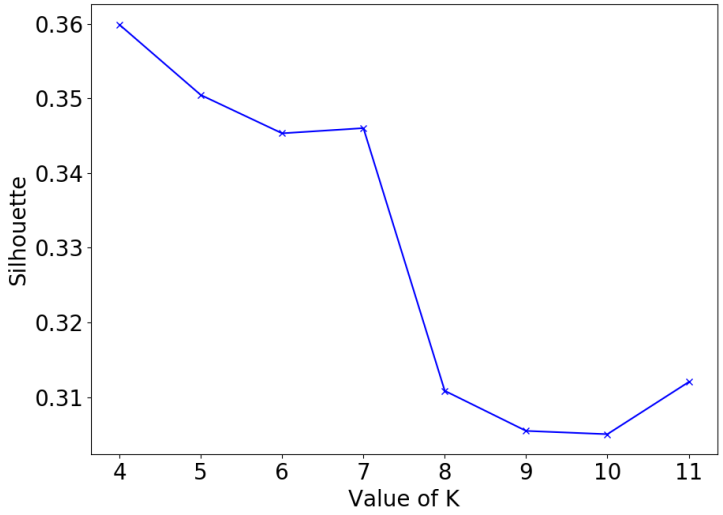

Figure 6: Average silhouette score of 1000 clustering results, financial data

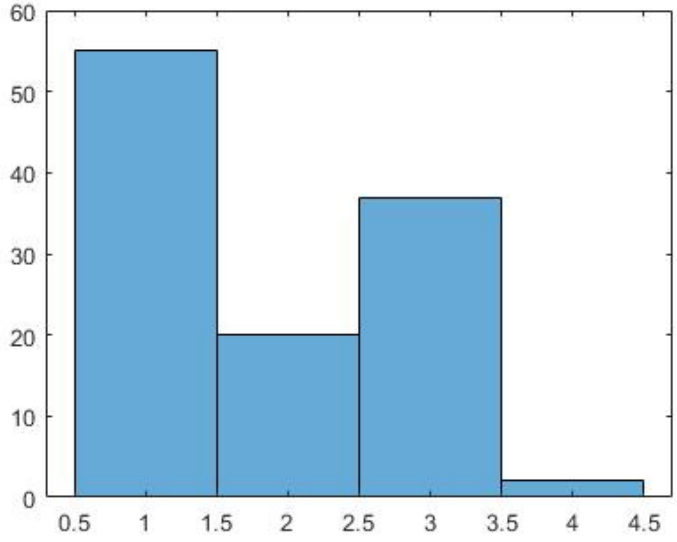

Figure 7: Histogram of the distribution of stocks in groups, financial ratio dataset, $K = 4$

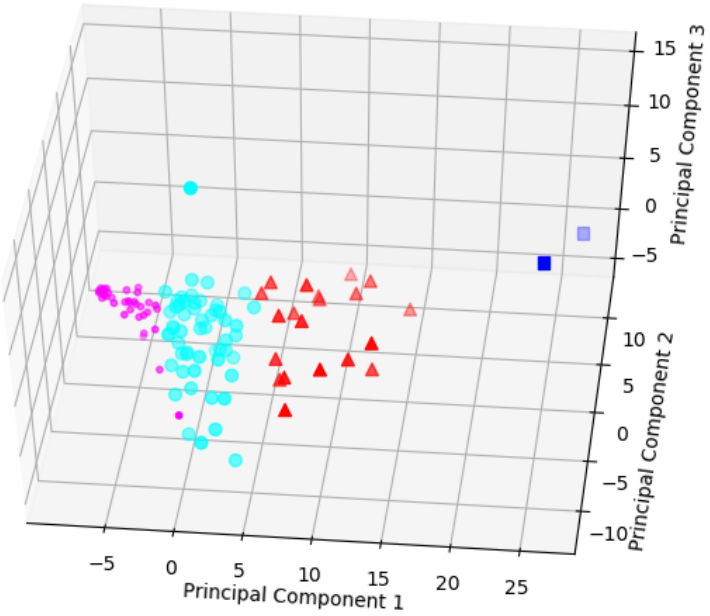

Figure 8: Scatterplot of financial ratio dataset, $K = 4$

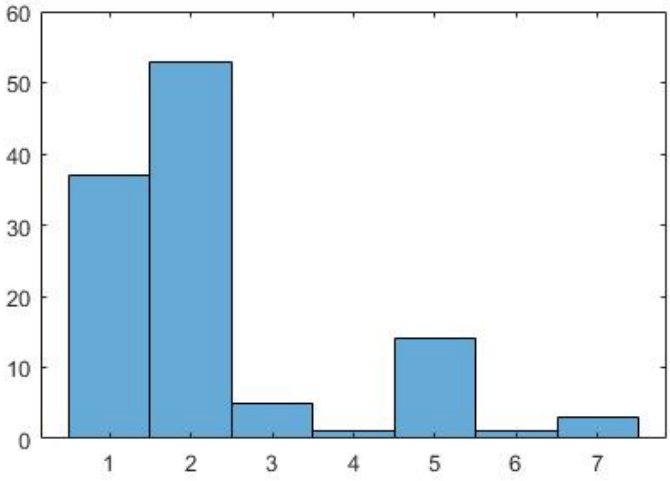

Figure 9: Histogram of the distribution of stocks in groups, financial ratio dataset, $K = 7$

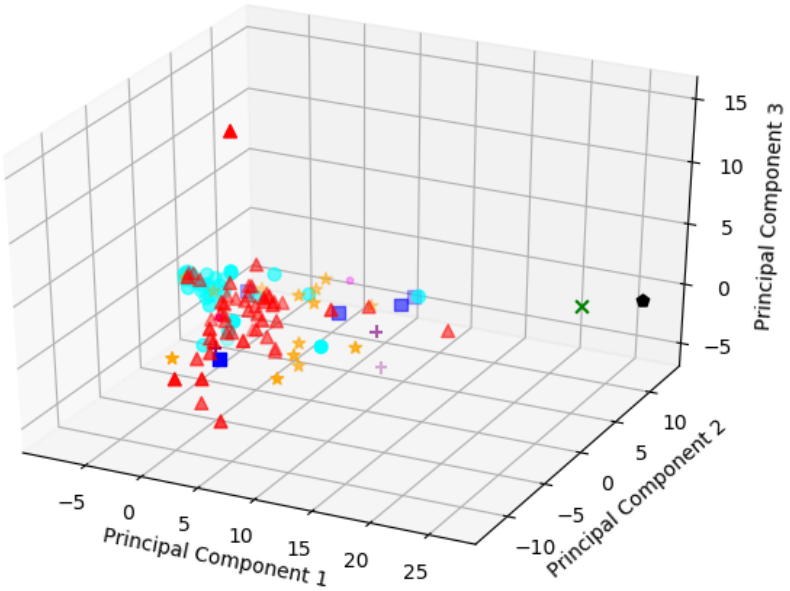

Figure 10: Scatterplot of financial ratio dataset, $K = 7$

Appendix 2: Portfolio Makeup and Return

Portfolio A (Stock price movement data, $K = 4$)

Company Ticker	ADSK	AEE	ALL	APA	**Average**	**Excess Return**
Rate of return	-13.57%	-6.61%	-13.95%	-70.68%	-26.20%	-13.80%

Portfolio B (Stock price movement data, $K = 6$)

Company Ticker	ADSK	AEE	ALL	APA	AEP	AIV	**Average**	**Excess Return**
Rate of return	-13.57%	-6.61%	-13.95%	-70.68%	-17.24%	-25.17%	-29.93%	-17.53%

Portfolio C (Financial ratio data, $K = 4$)

Company Ticker	ADSK	ALL	AMZN	CAH	**Average**	**Excess Return**
Rate of return	-13.57%	-13.95%	13.93%	-0.01%	-3.68%	8.72%

Portfolio D (Financial ratio data, $K = 7$)

Company Ticker	ADSK	ALL	AMZN	TXN	MO	CAH	COST	**Average**	**Excess Return**
Rate of return	-13.57%	-13.95%	13.93%	-26.25%	-8.95%	-0.01%	4.02%	-9.77%	2.63%

Appendix 3: Alternative Portfolio Makeup and Return

Portfolio E (Stock price movement data, $K = 4$)

Company Ticker	ALL	AEE	AEP	AES	AMGN	ADSK	A	AMAT	AMZN	ADBE	**Average**	**Excess Return**
Rate of return	-13.95%	-6.61%	-17.24%	-31.02%	4.57%	-13.57%	-4.04%	-12.14%	13.93%	-4.81%	-8.49%	3.91%

Portfolio F (Stock price movement data, $K = 6$)

Company Ticker	ADSK	A	AIG	AEE	ALL	AMGN	APD	AZO	AMP	BAX	**Average**	**Excess Return**
Rate of return	-13.57%	-4.04%	-50.00%	-6.61%	-13.95%	4.57%	-9.14%	-6.98%	-31.74%	-1.61%	-13.27%	-0.87%

Portfolio G (Financial ratio data, $K = 4$)

Company Ticker	ALL	AMP	AIG	ADSK	A	BA	BEN	AMZN	UPS	BLL	**Average**	**Excess Return**
Rate of return	-13.95%	-31.74%	-49.61%	-13.57%	-4.04%	-55.38%	-35.24%	13.93%	-0.62%	-8.35%	-18.15%	-5.75%

Portfolio H (Financial ratio data, $K = 7$)

Company Ticker	ALL	AMP	AIG	ADSK	A	BA	BEN	CAH	VNO	COST	**Average**	**Excess Return**
Rate of return	-13.95%	-31.74%	-49.61%	-13.57%	-4.04%	-55.38%	-35.24%	-1.11%	-32.26%	4.02%	-23.29%	-10.89%

Appendix 4: Data and Clustering Results

All data and code are available at *https://github.com/Treetion/financial-ratio-clustering*.

The LaTex file can be accessed at *https://www.overleaf.com/read/tqxdkvkxfvwn*.

Clustering results of stock price movement dataset

Group ($K = 4$)	**Group ($K = 6$)**	**Company**	**Ticker**
3	6	Autodesk Inc	ADSK
2	1	Ameren Corp	AEE
1	1	Allstate Corp	ALL
2	1	American Electric Power Company	AEP
3	6	Agilent Technologies	A
1	6	Apartment Investment and Management	AIV
3	1	Applied Materials	AMAT
2	1	The Aes Corp	AES
1	6	Ameriprise Financial Services	AMP
2	4	Amgen Inc	AMGN
3	4	Amazon.com Inc	AMZN
1	4	American International Group	AIG
4	5	Apache Corp	APA
3	3	Adobe Systems Inc	ADBE
3	1	Air Products and Chemicals	APD
3	6	Avery Dennison Corp	AVY
3	1	Boeing Company	BA
2	6	Autozone	AZO
1	6	Bank of America Corp	BAC
3	2	Analog Devices	ADI
2	1	Baxter International Inc	BAX
2	6	Automatic Data Procs	ADP
3	1	Best Buy Company	BBY
2	6	Becton Dickinson and Company	BDX
3	5	Franklin Resources	BEN
2	1	Biogen Inc	BIIB
2	6	Ball Corp	BLL
2	1	Archer Daniels Midland	ADM
1	6	Bank of New York Mellon Corp	BK
3	1	Sherwin-Williams Company	SHW
3	5	Snap-On Inc	SNA
1	1	Simon Property Group	SPG
1	1	State Street Corp	STT
2	1	Sempra Energy	SRE
2	1	Constellation Brands Inc	STZ
2	3	Southern Company	SO
3	1	Stanley Black & Decker Inc	SWK
3	1	Union Pacific Corp	UNP
2	6	Stryker Corp	SYK
2	1	Unitedhealth Group Inc	UNH
2	6	Bristol-Myers Squibb Company	BMY

Group ($K = 4$)	**Group ($K = 6$)**	**Company**	**Ticker**
3	1	Whirlpool Corp	WHR
3	1	Waters Corp	WAT
3	1	Verisign Inc	VRSN
1	1	Wells Fargo & Company	WFC
1	5	U.S. Bancorp	USB
3	1	United Parcel Service	UPS
3	6	Textron Inc	TXT
3	4	V.F. Corp	VFC
2	4	Waste Management	WM
2	5	Sysco Corp	SYY
4	2	Williams Companies	WMB
2	1	Verizon Communications Inc	VZ
3	1	Texas Instruments	TXN
3	1	Valero Energy Corp	VLO
2	6	Wal-Mart Stores	WMT
2	1	AT&T Inc	T
3	1	Xilinx Inc	XLNX
3	6	Vulcan Materials Company	VMC
3	1	Weyerhaeuser Company	WY
3	1	Starbucks Corp	SBUX
2	4	XCEL Energy Inc	XEL
1	6	Vornado Realty Trust	VNO
2	1	Exxon Mobil Corp	XOM
3	3	TJX Companies	TJX
2	1	Molson Coors Brewing Company	TAP
3	1	Netapp Inc	NTAP
3	6	Thermo Fisher Scientific Inc	TMO
2	1	Tyson Foods	TSN
3	4	Tiffany & Company	TIF
1	4	T Rowe Price Group	TROW
3	2	Sealed Air Corp	SEE
1	6	Northern Trust Corp	NTRS
3	6	Xerox Corp	XRX
3	1	Yum! Brands	YUM
3	4	Target Corp	TGT
1	1	Zions Bancorp	ZION
3	4	Nucor Corp	NUE
2	1	Marsh & Mclennan Companies	MMC
3	6	3M Company	MMM
3	6	Rockwell Automation Inc	ROK
3	2	Parker-Hannifin Corp	PH

Group ($K = 4$)	**Group ($K = 6$)**	**Company**	**Ticker**
3	3	Nvidia Corp	NVDA
2	5	Altria Group	MO
2	4	Public Storage	PSA
3	4	Newell Rubbermaid Inc	NWL
3	6	PPG Industries	PPG
3	2	Oracle Corp	ORCL
3	6	Berkshire Hathaway Cl B	BRK-B
1	4	Prudential Financial Inc	PRU
3	1	Robert Half International Inc	RHI
1	5	PNC Bank	PNC
3	1	Omnicom Group Inc	OMC
3	1	Perkinelmer	PKI
2	1	Progressive Corp	PGR
2	6	Pinnacle West Capital Corp	PNW
2	1	Merck & Company	MRK
3	1	Pultegroup	PHM
2	6	Mccormick & Company Inc	MKC
3	5	Moody's Corp	MCO
1	6	Principal Financial Group Inc	PFG
2	2	PPL Corp	PPL
3	1	Paccar Inc	PCAR
4	5	Occidental Petroleum Corp	OXY
2	1	Procter & Gamble Company	PG
2	2	Paychex Inc	PAYX
2	5	Medtronic Inc	MDT
2	6	Public Service Enterprise Group Inc	PEG
4	6	Marathon Oil Corp	MRO
2	1	Pepsico Inc	PEP
2	1	Pfizer Inc	PFE
3	1	Nordstrom	JWN
1	1	Morgan Stanley	MS
3	1	Microsoft Corp	MSFT
1	1	Boston Properties	BXP
2	6	Kellogg Company	K
3	6	Boston Scientific Corp	BSX
3	6	Micron Technology	MU
3	1	Nike Inc	NKE
3	1	Masco Corp	MAS
3	6	Norfolk Southern Corp	NSC
3	5	Marriot Int Cl A	MAR
3	6	K L A-Tencor Corp	KLAC

Group ($K = 4$)	**Group ($K = 6$)**	**Company**	**Ticker**
1	1	Keycorp	KEY
2	5	Cigna Corp	CI
2	3	Conagra Brands Inc	CAG
2	6	NiSource Inc	NI
2	1	Kimberly-Clark Corp	KMB
2	1	Nextera Energy	NEE
2	1	McDonald's Corp	MCD
4	1	National-Oilwell	NOV
3	6	Mylan NV Ord Shs	MYL
2	3	Coca-Cola Company	KO
1	1	Kimco Realty Corp	KIM
2	1	Mckesson Corp	MCK
2	5	Kroger Company	KR
2	1	Colgate-Palmolive Company	CL
2	6	Newmont Mining Corp	NEM
3	6	Juniper Networks	JNPR
3	6	Kohl's Corp	KSS
2	1	Cardinal Health	CAH
1	1	Comerica Inc	CMA
3	6	Leggett & Platt Inc	LEG
2	1	Cincinnati Financial	CINF
2	1	Clorox Company	CLX
1	1	JP Morgan Chase & Company	JPM
1	1	Capital One Financial Corp	COF
3	1	Johnson Controls Intl	JCI
3	1	Comcast Corp A	CMCSA
3	2	Intuit Inc	INTU
3	5	Lennar Corp	LEN
2	1	Chubb Ltd	CB
2	6	Laboratory Corp of America Holdings	LH
3	1	Cummins Inc	CMI
2	4	Cms Energy Corp	CMS
3	1	Illinois Tool Works Inc	ITW
3	4	International Paper Company	IP
3	1	Lowe's Companies	LOW
2	2	Johnson & Johnson	JNJ
2	6	Eli Lilly and Company	LLY
3	6	Interpublic Group of Companies	IPG
2	5	Campbell Soup Company	CPB
1	6	Lincoln National Corp	LNC
2	6	Costco Wholesale	COST

Group ($K = 4$)	**Group ($K = 6$)**	**Company**	**Ticker**
3	6	Southwest Airlines Company	LUV
3	1	Caterpillar Inc	CAT
3	1	Cintas Corp	CTAS
4	2	Conocophillips	COP
2	1	Centerpoint Energy Inc	CNP
3	6	Cisco Systems Inc	CSCO
3	4	CSX Corp	CSX
3	1	Carnival Corp	CCL
3	1	Citrix Systems Inc	CTXS
2	5	Centurylink	CTL
2	1	CVS Corp	CVS
2	1	Chevron Corp	CVX
3	6	Eastman Chemical Company	EMN
3	3	Deere & Company	DE
2	1	Quest Diagnostics Inc	DGX
2	5	Dominion Resources	D
3	4	Estee Lauder Companies	EL
3	6	Emerson Electric Company	EMR
1	4	Equity Residential	EQR
3	5	D.R. Horton	DHI
3	3	International Flavors & Fragrances	IFF
4	6	Eog Resources	EOG
2	6	International Business Machines	IBM
3	1	Danaher Corp	DHR
3	6	Walt Disney Company	DIS
3	1	Dover Corp	DOV
2	6	Edison International	EIX
2	4	Dte Energy Company	DTE
3	3	Equifax Inc	EFX
2	4	Humana Inc	HUM
3	5	Darden Restaurants	DRI
1	1	E*Trade Finl Corp	ETFC
2	6	Duke Energy Corp	DUK
2	1	Fiserv Inc	FISV
2	1	Entergy Corp	ETR
2	1	Ecolab Inc	ECL
2	1	Consolidated Edison Company of New York	ED
1	1	Fifth Third Bncp	FITB
3	6	General Dynamics Corp	GD
3	1	Electronic Arts Inc	EA
3	1	Ebay Inc	EBAY

Group ($K = 4$)	**Group ($K = 6$)**	**Company**	**Ticker**
2	6	Exelon Corp	EXC
4	1	Devon Energy Corp	DVN
2	6	Hershey Foods Corp	HSY
2	6	Gilead Sciences Inc	GILD
2	1	General Mills	GIS
3	6	Ford Motor Company	F
3	1	Corning Inc	GLW
3	5	HP Inc	HPQ
2	1	H&R Block	HRB
3	6	Alphabet Cl A	GOOGL
4	6	Freeport-Mcmoran Inc	FCX
3	3	W.W. Grainger	GWW
3	1	Gap Inc	GPS
3	1	Genuine Parts Company	GPC
3	1	Fedex Corp	FDX
4	5	Halliburton Company	HAL
1	6	Goldman Sachs Group	GS
2	1	Firstenergy Corp	FE
3	2	Harley-Davidson Inc	HOG
3	1	Hasbro Inc	HAS
3	6	Home Depot	HD
1	4	Huntington Bcshs	HBAN
3	6	Apple Inc	AAPL
4	1	Hess Corp	HES
1	6	Hartford Financial Services Group	HIG
2	1	Amerisourcebergen Corp	ABC
2	5	Abbott Laboratories	ABT

Clustering results of financial ratio dataset

Group ($K = 4$)	**Group ($K = 7$)**	**Company**	**Ticker**
1	2	Agilent Technologies	A
1	2	Apple Inc	AAPL
1	2	Abbott Laboratories	ABT
3	1	Automatic Data Procs	ADP
1	2	Autodesk Inc	ADSK
3	1	American International Group	AIG
3	1	Allstate Corp	ALL
3	1	Ameriprise Financial Services	AMP
2	3	Amazon.com Inc	AMZN
1	2	Boeing Company	BA
1	2	Franklin Resources	BEN
3	1	Bank of New York Mellon Corp	BK
1	2	Ball Corp	BLL
1	2	Bristol-Myers Squibb Company	BMY
4	4	Cardinal Health	CAH
1	2	Caterpillar Inc	CAT
1	2	Cigna Corp	CI
3	1	Cincinnati Financial	CINF
2	5	Colgate-Palmolive Company	CL
3	1	Comerica Inc	CMA
3	1	Comcast Corp A	CMCSA
2	5	Cummins Inc	CMI
3	1	Centerpoint Energy Inc	CNP
3	1	Capital One Financial Corp	COF
4	6	Costco Wholesale	COST
3	1	CSX Corp	CSX
1	2	Cintas Corp	CTAS
3	1	Centurylink	CTL
1	2	Citrix Systems Inc	CTXS
2	5	CVS Corp	CVS
3	1	Dominion Resources	D
1	2	Deere & Company	DE
1	2	Danaher Corp	DHR
1	2	Walt Disney Company	DIS
2	5	Darden Restaurants	DRI
3	1	Duke Energy Corp	DUK
3	1	Devon Energy Corp	DVN
1	2	Electronic Arts Inc	EA
3	1	Ebay Inc	EBAY
1	2	Ecolab Inc	ECL
3	1	Consolidated Edison Company of New York	ED

Group ($K = 4$)	**Group ($K = 7$)**	**Company**	**Ticker**
3	1	Edison International	EIX
2	5	Estee Lauder Companies	EL
1	2	Eastman Chemical Company	EMN
1	2	Emerson Electric Company	EMR
1	2	Fiserv Inc	FISV
3	1	Fifth Third Bncp	FITB
1	2	General Dynamics Corp	GD
1	2	Gilead Sciences Inc	GILD
1	2	Alphabet Cl A	GOOGL
2	7	Gap Inc	GPS
2	5	W.W. Grainger	GWW
1	2	Halliburton Company	HAL
1	2	Hasbro Inc	HAS
2	7	Home Depot	HD
1	2	Harley-Davidson Inc	HOG
2	5	HP Inc	HPQ
2	5	Hershey Foods Corp	HSY
2	3	Humana Inc	HUM
1	2	International Paper Company	IP
1	2	Interpublic Group of Companies	IPG
1	2	Johnson Controls Intl	JCI
1	2	Kellogg Company	K
1	2	K L A-Tencor Corp	KLAC
1	2	Coca-Cola Company	KO
1	5	Leggett & Platt Inc	LEG
1	2	Laboratory Corp of America Holdings	LH
3	1	Lincoln National Corp	LNC
1	1	Marriot Int Cl A	MAR
2	3	McDonald's Corp	MCD
1	2	Moody's Corp	MCO
1	2	3M Company	MMM
0	1	Altria Group	MO
1	2	Morgan Stanley	MS
1	2	Microsoft Corp	MSFT
3	1	Micron Technology	MU
1	2	Mylan NV Ord Shs	MYL
1	2	Nextera Energy	NEE
3	1	Nike Inc	NKE
3	1	Northern Trust Corp	NTRS
2	5	Nucor Corp	NUE
3	1	Nvidia Corp	NVDA

Group ($K = 4$)	**Group ($K = 7$)**	**Company**	**Ticker**
2	5	Newell Rubbermaid Inc	NWL
1	2	Occidental Petroleum Corp	OXY
1	2	Paccar Inc	PCAR
3	1	Public Service Enterprise Group Inc	PEG
1	2	Pepsico Inc	PEP
3	1	Principal Financial Group Inc	PFG
1	2	Procter & Gamble Company	PG
3	1	Progressive Corp	PGR
1	2	Parker-Hannifin Corp	PH
1	2	Pultegroup	PHM
1	2	PNC Bank	PNC
1	2	PPG Industries	PPG
3	1	PPL Corp	PPL
1	2	Prudential Financial Inc	PRU
3	1	Sealed Air Corp	SEE
3	1	Simon Property Group	SPG
1	2	State Street Corp	STT
3	1	Stryker Corp	SYK
3	1	Molson Coors Brewing Company	TAP
1	2	Target Corp	TGT
3	1	Tyson Foods	TSN
2	7	Texas Instruments	TXN
2	3	Unitedhealth Group Inc	UNH
1	2	United Parcel Service	UPS
2	5	Vornado Realty Trust	VNO
2	5	Verisign Inc	VRSN
3	1	Verizon Communications Inc	VZ
3	1	Whirlpool Corp	WHR
1	2	Wal-Mart Stores	WMT
1	5	XCEL Energy Inc	XEL
2	3	Xerox Corp	XRX
3	1	Zions Bancorp	ZION

References

[1] US business cycle expansions and contractions. *The National Bureau of Economic Research*, Sep 2010. Accessed on 5-8-2020. https://www.nber.org/cycles.html.

[2] Amazon.com, inc. (AMZN) analyst ratings, estimates, forecasts. *Yahoo! Finance*, May 2020. Accessed on 5-8-2020. https://tinyurl.com/y7aq28ya.

[3] Cardinal health, inc. (CAH) analyst ratings, estimates, forecasts. *Yahoo! Finance*, May 2020. Accessed on 5-8-2020. https://tinyurl.com/y8453zm4.

[4] Costco wholesale corporation (COST) analyst ratings, estimates, forecasts. *Yahoo! Finance*, May 2020. Accessed on 5-8-2020. https://tinyurl.com/y8rtge2w.

[5] Kimberly Amadeo. Behind the us shale oil boom and bust. *The Balance*, May 2020. Accessed on 5-8-2020. https://tinyurl.com/yacrn9uh.

[6] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[7] Jerilyn J Castillo and Peter J McAniff. *The Practitioner's Guide to Investment Banking, Mergers & Acquisitions, Corporate Finance*. Circinus Business Press, 2007.

[8] Tara Clarke. 2008 stock market crash causes and aftermath. *Money Morning*, Nov 2016. Accessed on 5-8-2020. https://tinyurl.com/hus94yu.

[9] Charles Elkan. Using the triangle inequality to accelerate k-means. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 147–153, 2003.

[10] Simone Farinelli, Manuel Ferreira, Damiano Rossello, Markus Thoeny, and Luisa Tibiletti. Beyond sharpe ratio: Optimal asset allocation using different performance ratios. *Journal of Banking & Finance*, 32(10):2057–2063, 2008.

[11] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

[12] Swetha Gopinath. 2016 brings more pain to U.S. shale companies as crude sinks. *Thomson Reuters*, Jan 2016. Accessed on 5-8-2020. https://tinyurl.com/ya3l6tfc.

[13] Benjamin Graham and Jason Zweig. *The Intelligent Investor: The Definitive Book on Value Investing (Fourth Edition)*. Harper, 2006.

[14] Charles W Hodges, Walton RL Taylor, and James A Yoder. Stocks, bonds, the sharpe ratio, and the investment horizon. *Financial Analysts Journal*, 53(6):74–80, 1997.

[15] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[16] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[17] Harry Markowitz. Portfolio selection. *The Journal of Finance, Vol. 7, Issue 1.*, 1952.

[18] Karina Marvin. Creating diversified portfolios using cluster analysis. *Princeton University*, 2015.

[19] Zhiwei Ren. Portfolio construction using clustering methods. 2005.

[20] Jessica Resnick-Ault. Oil plummets, on track for biggest weekly drop in 2019. *Thomson Reuters*, May 2019. Accessed on 5-8-2020. https://tinyurl.com/y7f9pykj.

[21] Fredrik Rosen. Correlation based clustering of the stockholm stock exchange, 2006.

[22] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[23] William F Sharpe. Mutual fund performance. *The Journal of business*, 39(1):119–138, 1966.

[24] William F Sharpe. The sharpe ratio. *Journal of portfolio management*, 21(1):49–58, 1994.

[25] Michael Stutzer. A portfolio performance index. *Financial Analysts Journal*, 56(3):52–61, 2000.

[26] Nassim Nicholas Taleb. *The Black Swan*. Random House, 2009.

[27] Nassim Nicholas Taleb. *Antifragile: Things that gain from disorder*. Random House, 2014.

[28] Jordan Wathen. Real estate stocks drop after donald trump wins the presidency. *The Motley Fool*, Nov 2016. Accessed on 5-8-2020. https://tinyurl.com/yaj7qjfx.

[29] Rui Xu and Donald Wunsch. *Clustering*. IEEE Press Series on Computational Intelligence, 2009.