CSCI 4022: Stock Market Clustering

Nathaniel Lee

I. Introduction:

The stock market is a complex and stochastic system that is difficult to forecast accurately. Financial planning is one of the most important decisions an individual can make, with many individuals trusting mutual funds to manage their retirement and many others actively managing their funds. According to the New York Times, however: "fewer than 10 percent of active U.S. stock funds managed to beat their benchmarks." (Sommer, 2022). These funds have capital, resources, and connections that most individuals need help accessing and still underperform. As such, it is not advisable for most individual investors to try and beat the market but rather to maintain a diversified portfolio to minimize risk while capturing substantial long-term rewards.

To improve their decision-making, investors can use algorithms and machine learning techniques to gain insights from data. For instance, according to Investopedia, a committee of neural networks (a form of machine learning called deep learning) can evaluate trading strategies and potentially enhance returns by increasing efficiency by up to 10% (Vonko et al., 2022). 10% is certainly significant when a trading strategy can be determined viable or not by a change as small as 1 or 2%.

The corresponding project code (submitted to Canvas) uses the data science techniques learned in class to cluster stocks in the U.S. based on fundamental stock market indicators: market cap, percent change, beta, dividend yield, price-to-earnings ratio, and return on equity. The research question for this project is: How accurate are the stock clusterings generated by K-Means and Gaussian Mixture Models (GMMs) compared to some common-sense metrics? The hypothesis is that the clusterings will be reasonably accurate and consistent with the metrics. The remainder of this paper into sections about the data used (its sourcing and method of collection), the real-world impact of the project, the methods used, and the results.

II. Data:

The data for this project comes from Yahoo Finance and Nasdaq. The data from Nasdaq was downloaded from the nasdaq.com screener webpage. The reader can access it by visiting this site (3) and clicking “Download CSV.” The data from yahoo finance was acquired via the Python libraries yfinance (Aroussi, *yfinance* 2023) and yahoofinancials (Sanders, *yahoofinancials* 2023).

The features of the data used in this project are symbol (ticker), name, yearly percent change, market cap, country, volume, sector, industry, beta, dividend yield, price-to-earnings ratio (P/E ratio), and return on equity (RoE). Concise definitions for clarity are as follows:

- Beta - A measure of a stock's volatility in relation to the overall market. It represents the slope of the linear regression of the stock's returns against the market returns. Different analysis tools calculate this differently, but, for this project the S&P 500 was used as the relative market.

- Dividend Yield - The amount of money a company pays out in dividends each year relative to its stock price. For this project if there was no dividend the stock was assigned a dividend yield of zero instead of undefined.
- Price to Earnings Ratio - A measure of a company's valuation calculated by dividing its current stock price by its earnings per share.
- Return on Equity - A measure of a company's profitability, calculated by dividing its net income by its shareholder's equity.

It is worth noting that downloading data from Nasdaq was straightforward, whereas finding the supplemental information from Yahoo Finance (beta, dividend yield, P/E Ratio, RoE) was more involved. The added complexity is because an external API was called for every stock. According to some rudimentary calculations, this would have taken ~two days on a personal computer and wifi and thus was not feasible. As such, the code was refactored to make requests in parallel and run on a multicore Google Cloud Compute Engine to make the requests. This switch decreased the time needed from greater than two days to roughly thirty minutes (presumably, a large part of this speedup was the significantly better internet connection in a data center in addition to more threads running at once). The Python code for this additional data retrieval is located here (4) for completeness.

III. Real-world impact:

This section discusses the real-world impact of clustering U.S. stocks based on fundamental indicators. It reviews existing literature on this topic and identifies potential benefits and challenges of applying this technique in practice.

Clustering U.S. stocks based on fundamental indicators can help investors create diversified portfolios that reduce risk and maximize return. Fundamental indicators such as earnings, profitability, and growth reflect a company's financial performance and health. By grouping stocks with similar fundamental characteristics, investors can avoid overexposure to individual sectors, industries, or market segments that may be affected by common factors or events.

However, clustering U.S. stocks based on fundamental indicators is a complex task. It requires choosing appropriate similarity measures, selecting optimal cluster sizes and numbers, and dealing with dynamic and stochastic market conditions. To illustrate some of the issues and solutions involved in this technique, this section reviews two papers that use clustering algorithms to create diversified portfolios based on financial ratios.

The first paper is "Creating Diversified Portfolios Using Cluster Analysis" (Karina Marvin, 2015). This paper proposes using cluster analysis to create diversified portfolios based on similarity measures of financial ratios, specifically revenues/assets, and net income/assets. First, the K-means algorithm was carried out for historical values, using these financial ratios as the similarity measure. Then, portfolios were picked based on the results of the algorithm (high Sharpe ratios). The study found that the diversified portfolios based on the generated clusters generally outperformed the S&P 500 benchmark in normal years, and was comparable to the market in turbulent years (2007-2009), suggesting that clustering can be a valuable tool for creating diversified portfolios and achieving positive returns in a stochastic market.

The second paper is “Efficient Stock Portfolio Construction by Means of Clustering” (Korzeniewski, 2018). This paper utilized k-means and PAM algorithms on the Warsaw Stock Exchange from 2011-2016 and focused on the elimination of ad-hoc measures to determine the number of clusters. In contrast to the first paper, this paper did not use financial ratios, but instead used the daily return rates of the last 120 trading days. The paper concluded several things from its results. Additionally, rather than choosing the number of clusters arbitrarily, this paper Caliński-Harabasz index. Firstly, it concluded that the economic index of the Sharpe ratio is predictively valid despite simplicity. Additionally, it concluded that short time frames are much more helpful targets for clustering algorithms than more extended time frames. Specifically, it states: “It is absolutely inadvisable to try to extrapolate trends from the most recent six months onto the six months to follow. The most efficient in this respect were short investments, shorter than one month.” Finally, it concluded that using the Caliński-Harabasz index on short time frames does perform better than ad-hoc methods.

Both papers use clustering algorithms to create diversified portfolios based on financial ratios, but they differ in several aspects. The first paper uses revenues/assets and net income/assets as the similarity measures, while the second paper uses the daily return rates of the last 120 trading days. Both papers use the L2 norm for calculating distance. The first paper tests the portfolios on 60 quarters between 2000 and 2015, while the second paper tests them on 20, 50, and 120 day periods between 2011 and 2016. The first paper finds that the clustered portfolios outperform the S&P 500 benchmark, while the second paper confirms that clustered portfolios outperform the market benchmark and finds that short-term investments are more efficient than long-term ones.

The section shows that clustering U.S. stocks based on fundamental indicators can significantly impact portfolio performance and risk management. However, it also highlights some of the challenges and limitations of this technique, such as choosing appropriate similarity measures, selecting optimal cluster sizes and numbers, and dealing with dynamic and stochastic market conditions. Solutions to these challenges were implemented in each paper with varying success.

IV. Methods:

I used two different data science techniques, K-Means clustering and Gaussian Mixture Model (GMM) clustering, to determine clusters for a subset of stock market tickers.

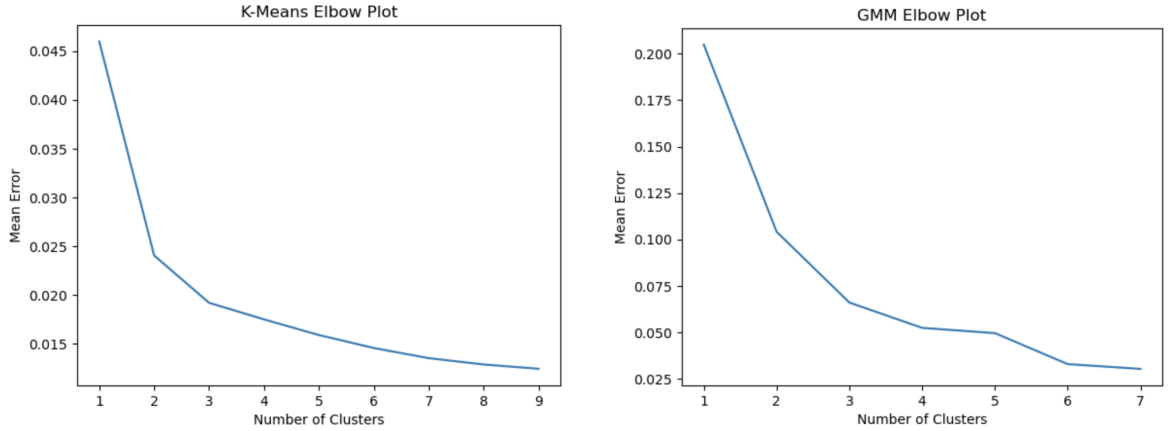
K-Means clustering is an unsupervised learning algorithm that partitions a dataset into K clusters by minimizing the distances between the data points and their cluster’s centroids. This method is useful in identifying distinct groups in a dataset based on similarities or differences in their features. Importantly, K-Means is a hard-clustering model, meaning that every data point is either in a specific cluster or not.

Gaussian Mixture Model (GMM) clustering, on the other hand, is a probabilistic clustering method that assumes that the data points are generated from a mixture of Gaussian distributions. GMM clustering is soft, so every data point has a probability of being in every cluster. For the purposes of analysis, the GMM clustering was modeled as hard-clustering, where every datapoint was assigned to the cluster that it was most likely to be in. The GMM method provided the useful ability to determine non-uniform

clusters. This is in contrast to how K-Means can only generate uniform clusters and thus misses potential linear or other shaped relationships.

The two methods are both appropriate for clustering stock market tickers since they can identify groups of stocks with similar price movements or other financial metrics, but both were used to see which would perform better.

However, there are some ad hoc choices that need to be made when using these algorithms, such as selecting the number of clusters to partition the dataset into. For both algorithms, I ran the algorithm multiple times with a different number of clusters, plotting the mean error for each number of clusters on a line graph and then choosing the point where it appears more clusters has less of an effect (elbow plot method). Seven clusters were chosen for the K-Means experiment, and six clusters were chosen for the GMM experiment.



The probability model f(x) used for the GMM algorithm can be written as:

$$f(x|\Theta) = \sum_{k=1}^{M} (w_k \phi(x \mid \mu_k, (\sigma_k)^2))$$

where $\phi(x \mid \mu_k, (\sigma_k)^2)$ is the probability density function of the kth Gaussian distribution, computed using the mean and variance of that distribution, w_k is the weight of the kth distribution, and M is the number of clusters.

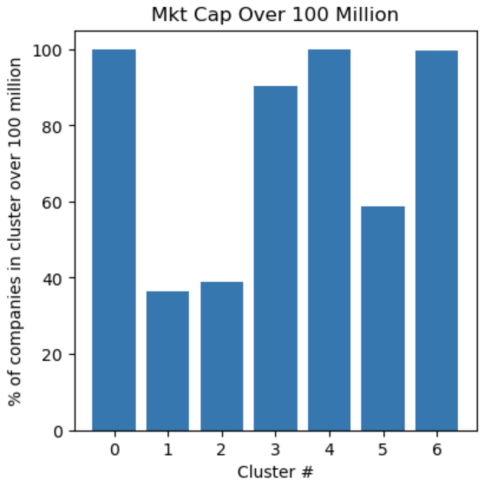
Another important aspect of the project was the preprocessing of the data. Initially, the data used for the clustering was numeric in nature, potentially negative or positive, and ranging in magnitude from zero to several trillion. Additionally, many columns had missing entries for one or more of the features used in the clustering algorithms.

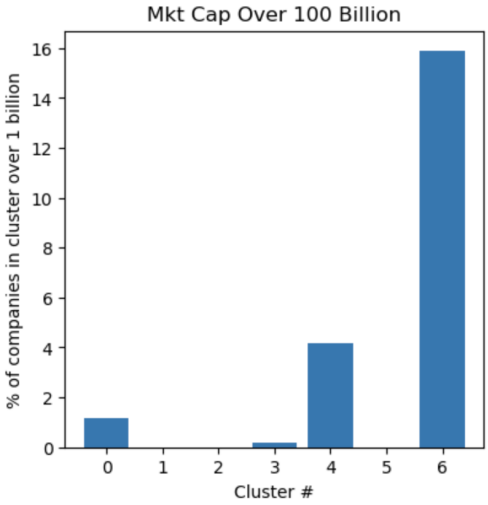
The data was normalized differently for the different algorithms, based on the fact that they had different requirements for their respective data inputs. To normalize the data for K-Means, I scaled the data down so that the maximum data point had a magnitude of one, the minimum data points had a magnitude of 1, and the sign was preserved. For GMM the data was shifted by adding the absolute value of the largest negative value to every value (making it exclusively non-negative), and then scaling the data down so that the maximum data point had a magnitude of one, the minimum data points had a

magnitude of 0. Additionally, for each algorithm the data was preprocessed so that missing entries for the dividend yield feature were interpreted as zero values, since a lack of a dividend yield meant that the stock did not pay a dividend at all. If an entry was missing in any other feature, the stock with the missing entry was removed from the data set. This resulted in significant shrinkage of the data set, from 7635 stocks to 2070 stocks, but this was necessary as further missing entries could cause the clusters to unreliably attract stocks.

V. Results:

K-Means Algorithm:





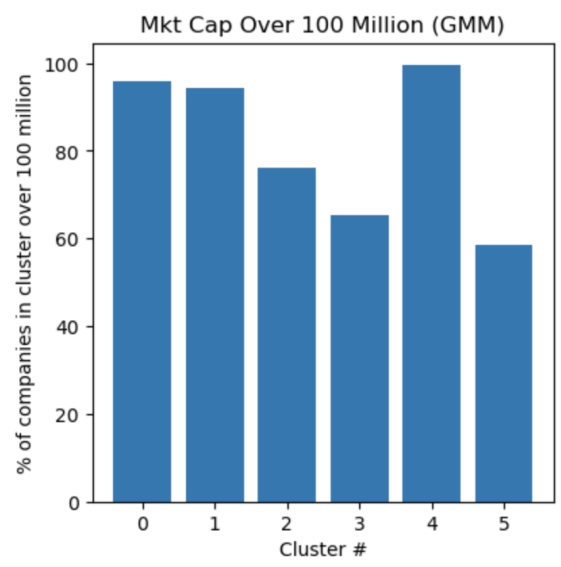
We applied the K-Means algorithm to cluster the stocks in NASDAQ based on their daily returns. The algorithm generated 7 clusters with different characteristics. The following table summarizes some of the features of each cluster, such as the number of stocks, the average market capitalization, the dominant country and sector, and some examples of companies in each cluster.

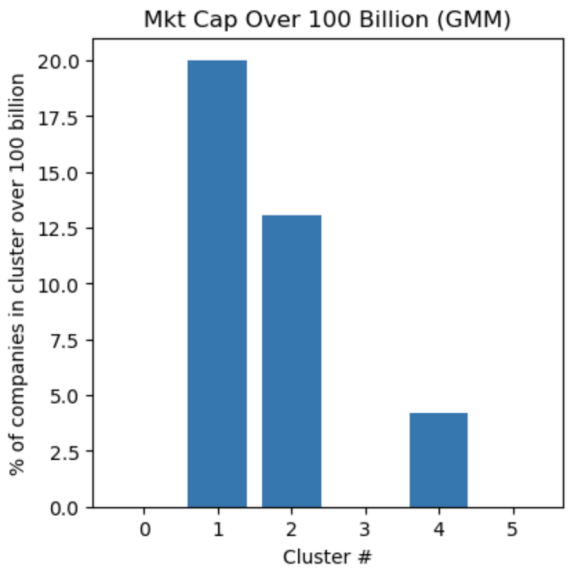
Cluster	# of stocks	Average market cap (in billion USD)	Dominant country	Dominant sector(s)	Examples of companies
0	338	7.08	U.S. (87%)	Finance (27%), Consumer Discretionary (23%)	Match.com, AMD, Delta
1	416	0.46	U.S. (76%)	Healthcare (52%)	BioNTech
2	18	2.20	U.S. (87.5%)	Healthcare (50%)	Pinterest

3	507	3.03	U.S. (83.4%)	Consumer Discretionary (26.8%)	Kroger, Alibaba, Datadog
4	434	19.65	U.S. (89.1%)	Finance (33.9%)	Paypal, Intel, Autodesk, NVIDIA
5	143	1.27	U.S. (81.1%)	Health Care (42.7%)	PDD, Holdings
6	214	90.89	U.S. (88.8%)	Tech (16.4%), Industrial (17.8%), Consumer Discretionary (16.4%), Finance (15.9%)	Comcast, Nasdaq, BlackRock

From the table, we can observe some interesting trends in the data. Cluster 0 seems to represent well-established companies that have a stable performance in the market due to diversity of countries and sectors and lack of low-market cap companies. Cluster 1 seems to represent emerging or risky companies, due to its low average market cap. Cluster 2 seems to represent niche companies due to its lack of high market cap stocks. Cluster 3 seems to represent traditional, mature stocks, due to its high percentage of companies with a market cap over 100 million USD (90%) and dominant industry (consumer discretionary). Cluster 4 seems to represent leading and innovative companies that have a dominant position in the finance industry due to its high average market cap and lack of low market cap stocks. Cluster 5 may represent specialized and stable companies due to its dominant industry (health-care) and moderately high average market cap. Cluster 6 has an extremely high average market cap and as such seems to represent giant and influential companies in all sectors.

GMM Algorithm:





We applied the GMM algorithm to cluster the stocks in NASDAQ based on their daily returns. The algorithm generated 7 clusters with different characteristics. The following table summarizes some of the features of each cluster, such as the number of stocks, the average market capitalization, the dominant country and sector, and some examples of companies in each cluster.

Cluster	# of stocks	Average market cap (in billion USD)	Dominant country	Dominant sector(s)	Examples of companies
0	607	6.50	U.S. (86.6%)	Consumer Discretionary (18.1%), Finance (16.5%)	Hasbro, Allstate, Fidelity
1	105	137.14	U.S. (76.8%)	Health Care (23.8%), Technology (18.1%)	JP Morgan, Microsoft, Blackstone
2	184	36.63	U.S. (84.8%)	Health Care (36.4%)	Amazon, AMD, Amazon
3	859	0.43	U.S. (83.8%)	Health Care (28.3%), Finance (20.7%)	N/A
4	286	22.55	U.S. (84.1%)	Consumer Discretionary (24.8%), Industrials (19.2%)	Dollar Tree, Ball Aerospace, Humana
5	29	14.49	U.S. (74.1%)	Health Care (37.9%)	Motorola, Honda

The table reveals some intriguing patterns in the data. We can see that some clusters consist of massive and influential companies with a global reach and high market caps, such as cluster 1 and cluster 2. These clusters include the most prominent and innovative tech and healthcare organizations like Microsoft, Apple, Alphabet (Google), Amazon, Alibaba, and Comcast. They have moderate to low country and sector diversity, indicating dominance in their respective industries.

On the other hand, some clusters comprise smaller and more specialized companies with moderate to low market caps, such as cluster 3 and cluster 5. These clusters contain emerging or stable companies that offer essential or diversified products or services in various sectors. In addition, they have relatively high country diversity, suggesting their regional or niche markets. The other two clusters, cluster 0 and cluster 4, are somewhere in between, with stable and mature companies with moderate market caps and sector diversity.

Overall, it seems that the GMM models did a better job clustering the companies. The GMM model has a more sharp distinction between high-market cap and low-market cap companies, with almost all emerging (and therefore risky) companies grouped together in cluster 3. While cluster 2 from the K-Means model seems to contain many emerging companies as well, it is significantly smaller meaning that many of the emerging companies are also spread across other groups. One advantage of the K-Means model seems to be that it better concentrates companies into clusters based on industry. Specifically, 52% of cluster 1 and 50% of cluster 2 are healthcare in the K-Means model, while the GMM model has a maximum of 37.9% of any one sector in any cluster.

VI. Conclusion:

In conclusion, both the K-Means and GMM algorithms were successful in clustering the stocks in NASDAQ based on their daily returns, generating 7 and 6 clusters, respectively, each with distinct characteristics. In addition, the K-Means algorithm effectively concentrated companies into clusters based on industry. At the same time, the GMM model was able to better distinguish between the high-market cap and low-market cap companies.

Overall, the clustering results and the trends observed in the data have important implications for investors, portfolio managers, and financial analysts. The clusters created are diversified across stock type, industry, and size, despite not taking these measures into account in the calculation. As a result, the K-Means and GMM algorithms could be used to guide investment decisions and identify opportunities in the stock market.

Future research can explore the use of other clustering algorithms and the inclusion of additional features to improve the accuracy and effectiveness of the clustering results. Additionally, measures such as the Sharpe index could be used to pick out a portfolio of stocks and measure returns over time as the prior studies reviewed in the "Real World Impact" section do. One weak area of this project is how many clusters were chosen, as this was essentially arbitrary and could have implemented an objective measure such as the Caliński-Harabasz index. Nonetheless, our study provides valuable insights into clustering stocks based on their daily returns, highlighting the importance of algorithm selection and data interpretation in the stock market analysis.

Citations:

(1) Sommer, J. (2022, December 2). Mutual funds that consistently beat the market? not one of 2,132. The New York Times. Retrieved April 30, 2023, from https://www.nytimes.com/2022/12/02/business/stock-market-index-funds.html#:~:text=And%20over%20a%20full%2020,that%20fail%20to%20do%20so.

(2) Vonko, D., Brown, J. R., & Logan, M. (2022, July 8). Neural networks: Forecasting profits. Investopedia. Retrieved April 30, 2023, from https://www.investopedia.com/articles/trading/06/neuralnetworks.asp

(3) https://www.nasdaq.com/market-activity/stocks/screener

(4)https://docs.google.com/document/d/1K5WiWMH-9ykK757gt1r_dlPJITlaHp6o4Jc4pGusGeU/edit?usp=sharing

(5) Aroussi, R. (2023, April 16). Version (0.2.18). *yfinance*. pypi.org. Retrieved April 25, 2023, from https://pypi.org/project/yfinance/.

(6) Sanders, C. (2023, February 21). Version (1.14). *yahoofinancials*. pypi.org. Retrieved April 26, 2023, from https://pypi.org/project/yahoofinancials/ .

(7) Karina Marvin, K. (2015). *Creating Diversified Portfolios Using Cluster Analysis* (thesis). *princeton.edu*. Princeton. Retrieved April 30, 2023, from https://www.cs.princeton.edu/sites/default/files/uploads/karina_marvin.pdf .

(8) Korzeniewski, J. (2018). Efficient stock portfolio construction by means of clustering. *Acta Universitatis Lodziensis. Folia Oeconomica*, *1*(333). https://doi.org/10.18778/0208-6018.333.06