1 Introduction

When it comes to investing in stock markets, there are numerous strategies one can employ in the effort to maximize profit and minimize loss. One such strategy involves the diversification of a stock portfolio. In this exploratory paper, I will use cluster analysis with the end goal of building a diversified portfolio.

Portfolio diversification can take on many forms, but it typically includes investing in different asset classes, risk classes, or by geographic location. For the ease of this exploration, I will focus on stocks; specifically, the stocks found within the S&P 500 index. Due to my choice to use the S&P 500, I will not have the luxury of diversifying based on asset class or geographic location, rather, I'll have to use diversification methods typically used for stocks.

Although I would have loved to further explore the many ways one can develop a diversified stock portfolio, I decided on two popular methods. I will begin by clustering, in the attempt to classify different risk classes. I will take this one step further, by clustering the risk classes in each sector individually to see if I can gain any insightful information this way.

I'm still not entirely sure what the best set of parameters to cluster on. While reading academic papers helped me pick some of these parameters, I still think there may be a better set of them out there.

I hypothesize that the diversified portfolio can be built from the best performing stock, or by picking the highest Sharpe Ratio found within each cluster. Determining the best way to score or rank the stocks of a cluster is another major goal of this exploration.

2 Data

I built my code to operate on any stock index if first given an initial dataset with the stock symbols, company names, and sectors of that index (Again, I've chosen the S&P 500 for this exploration). From there, I use an API called 'yfinance' that is built on top of the Yahoo Finance service.[1] This API provides access to a ton of data for a given company, from historical stock prices, general information, dividends, quarterly balance sheets, and financial statements.

I collect historical open-high-low-close-volume for each stock of the index. Although the time period is adjustable, I used 1 year of historical price data in this exploration. This is used to calculate the total and average return, both as a dollar amount and as a percentage. This will be used to measure a given portfolio of stocks against the index they are pulled from. I can also calculate the Sharpe ratio using the historical price data.

---

[1] https://pypi.org/project/yfinance/

The Sharpe ratio is derived by dividing the return of a stock by its risk. This ratio is useful for scoring a portfolio or stock by factoring in the risk. While other similar ratios exist and can be more informative, the Sharpe ratio is relatively simple to compute, and can still be a reasonable scoring metric.

The API also provides additional data for each stock. I collect several items that are given in the general information request. These things are: beta, forward and trailing PE ratios, PEG ratio, and the Current ratio. These ratios can be used to rank the stocks of a given cluster, and even if I don't end up using any of them for this purpose, they are still valuable to have readily available.

Beta is a measure of the volatility or risk of a stock compared to the market as a whole.[2] This may not be the most useful metric to score the stocks of a cluster, but it may be helpful to include in the parameters that I actually cluster on.

The forward and trailing PE ratios, or price-to-earnings ratios, is the current share price over the earnings per share.[3] Trailing PE ratios are calculated using historical share price and EPS values, while forward PE ratios rely on performance predictions. These ratios are useful in determining whether a stock is overvalued or undervalued. Again, not necessarily the best way to score stocks of a cluster, and further, may not be useful to cluster on. However, PE ratios are still valuable to experienced investors (not necessarily myself).

The PEG ratio, or price/earnings-to-growth ratio, provides a more complete picture than a standard PE ratio. The PEG ratio divides the PE ratio by the growth rate of earnings.[4] It tells a similar story to that of PE ratios, but also factors in the potential for growth, so it may be a better indicator of whether a stock is overvalued or undervalued. This could be slightly more useful to score the stocks of a cluster, and *could* be used to cluster on, but once again, I have a feeling it may not be used. Regardless, the PEG ratio another useful piece of information to collect on each stock.

Finally, the Current ratio measures a company's ability to pay its short term debts with its current assets.[5] This ratio can be used to measure the similarity of two companies within the same sector, although this can breakdown if used across different sectors. This could become quite valuable when I cluster each sector individually.

---

[2] https://www.investopedia.com/terms/b/beta.asp

[3] https://www.investopedia.com/terms/p/price-earningsratio.asp

[4] https://www.investopedia.com/terms/p/pegratio.asp

[5] https://www.investopedia.com/terms/c/currentratio.asp

I also collect data from financial statements and balance sheets. I get total revenue, net income, and total assets. These are used to calculate the return on assets and asset turnover ratios. The return on assets and asset turnover are two standard ratios used to measure the financial health of a company. Return on assets can gauge the profitability of a company, while asset turnover can indicate potential for growth and efficiency.[6] These can be very useful to describe the similarity of two given companies, and when I discovered these ratios, I realized this is probably what I should cluster on.

3 Real-world impact:

I'm incredibly confident many companies perform cluster analysis on stocks whether the company is an investment firm trying to optimize their portfolios or a company trying to do research on their competitors. While many companies may benefit from cluster analysis, it was much easier to find academic papers on this topic. These papers provided me with some incredibly helpful insight when encountering a multitude of issues.

For starters, several papers[7,8] opted for the use of the k-means clustering algorithm. I take a more in-depth look into the issue of which clustering algorithm to choose in the methods section, but it seems k-means is the most popular choice due to its simplicity and how the clusters are determined. The simplicity of k-means in terms of both implementation and rationalization, can aid in the interpretability of the problem and solution. K-means uses a distance function for computing the similarity of two items, while the EM algorithm uses statistical methods. While I'm sure the EM algorithm would do a fine job in the context of this exploration, k-means allows for more flexibility and

Next, I was struggling when it came time to decide how to choose the similarity of two stocks, however, several of the academic papers use return on assets and asset turnover, which ultimately led me to do the same after doing more research on these ratios.

Many of these papers used some form of principal component analysis, which led me to believe they were clustering on more than just return on assets and asset turnover. While this may or may not be the case, I initially planned on performing some form of PCA as well, but I scrapped this idea when my results seemed to be more clear and interpretable without it.

Finally, while I knew I was going to use the Sharpe ratio in some capacity before reading any of the academic papers, these papers helped me confirm that I was on the right track. These papers also inspired me to do

---

[6] https://scholarship.claremont.edu/cgi/viewcontent.cgi?article=3517&context=cmc_theses page 7

[7] https://scholarship.claremont.edu/cgi/viewcontent.cgi?article=3517&context=cmc_theses

[8] https://www.cs.princeton.edu/sites/default/files/uploads/karina_marvin.pdf

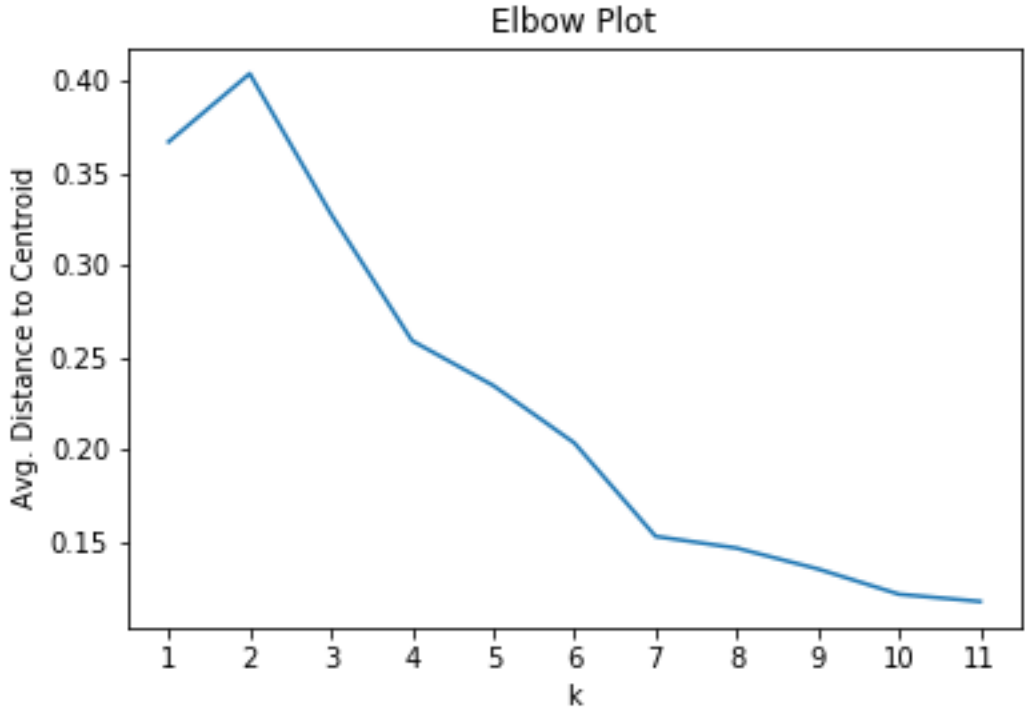
some research on other ratios similar to the Sharpe ratio, which could be used in its place. These other ratios include the Sortino ratio, Calmar ratio, and Max drawdown. These ratios could provide better results depending on the investor, but the Sharpe ratio will be the primary way I score stocks in this exploration.

4 Exploratory results

I wasn't able to pull all of the necessary data for every single stock. I needed to know which stocks from the S&P 500 I wasn't including in my analysis. I ended up not having enough data for the following six stocks: BRK.B, UDR, PWR, ODFL, RJF, and BF.B.

I needed to write several functions to compute total and average changes in dollar amount and percent, as well as functions to compute the return on assets, asset turnover, and the Sharpe ratio. I liked this approach because it will make this project easier to add to in the future. All of the price change functions were tested on the S&P 500 itself, and I used Yahoo Finance to confirm I was getting the correct values. Additionally, because the Sharpe ratio relies on price data as well, I was

I also had to decide which values of k I should use. To do this, I used an elbow plot, with k on the x-axis and the average distance from each cluster's points to their corresponding centroid on the y-axis. I've included this plot on the next page, and found the values 4 and 7 to be optimal k-values. When $k = 4$, our elbow plot makes its first 'bend' and when $k = 7$, the plot's bend flattens out the most. I could have chosen a larger value for k, but I didn't want to overfit the clusters.



5 Methods

I use clustering to build a diverse portfolio because it seemed like the most obvious method to group similar items together. While many data science and machine learning tools and techniques are used on financial data, I really enjoyed clustering in this class and was interested when I learned that clustering analysis on stocks was actually a thing.

I was struggling to determine which clustering algorithm from class I should use in this exploration. On one hand, k-means is a simpler algorithm, but the EM algorithm can be more accurate. Ultimately, with the help of several academic papers, I chose k-means.

The parameters I used to cluster were return on assets and asset turnover. I initially tried clustering on all of the parameters I collect, but when I tried running PCA, the visualizations weren't as nice. Further, out of every combination of axes I tried, return on assets and asset turnover proved to look the best.

Once I had the clusters, I had to score the stocks within each cluster so I could choose one. The largest Sharpe ratio value found within each cluster is selected for the portfolio. I used the Sharpe ratio because it was simple to compute, I had a reasonable understanding of it, and because several academic papers used it as well. The Sharpe ratio is given below:

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

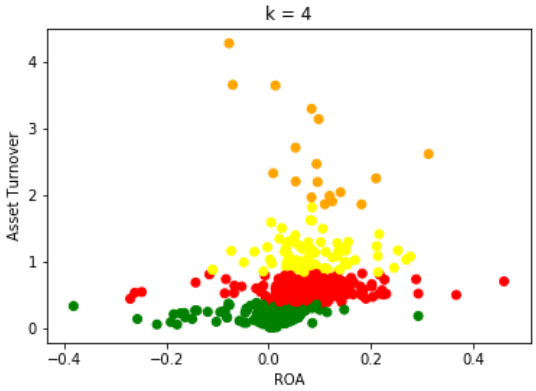
where:

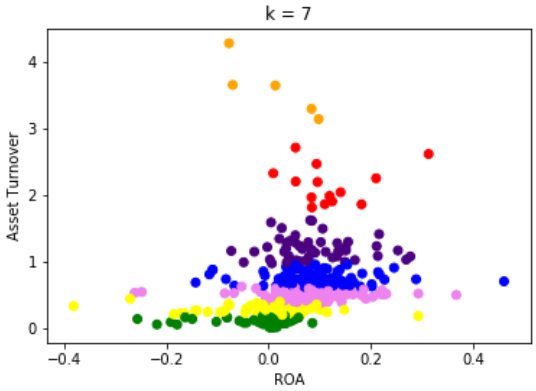
R_p = return of portfolio,
R_f = risk-free rate,
σ_p = standard deviation of portfolio's excess return

From there, I had to calculate the returns of the S&P 500 for a benchmark. Then I calculated the returns for the generated portfolios. Now it was time to see how the portfolios stacked up against the S&P 500 benchmark!

6 Results

I clustered for $k = 4$ and $k = 7$ due to the elbow plot and the clusters look reasonable! Although, they do appear to be clustered in skinny bands, I think the clusters have relatively strong borders, and both values of k seem to do well. Here are those two visualizations of the clusterings:





As it turns out, my generated 'diversified' portfolios outperformed the S&P 500 for both $k = 4$ and $k = 7$. Portfolio 1 with $k = 4$, consisted of the following stocks: AZO, MAA, RHI, FTNT. Portfolio 2 with $k = 7$ consisted of these following stocks: MAA, COST, AZO, RHI, NDAQ, JCI, FTNT. I will provide a table to compare the portfolios to the benchmark for each value of k:

	Portfolio 1	Portfolio 2	Benchmark (S&P 500)
Total Price Change	$1220.62	$1505.00	$958.22
Total Percent Change	83%	75%	26%

As one can tell, these portfolios did quite well in the last year, almost tripling the performance of the S&P 500. It became clear to me that the Sharpe ratio can be a crucial ratio for investors.

7 Conclusion

Although I ended up using return on assets and asset turnover to cluster on and as the axes, I still don't know if this is the best set of parameters to cluster on. I think that this would be an interesting question to explore more. I also think that it would be helpful to perform some form of PCA once an optimal parameter set is found.

Additionally, I'm not too confident in the predictive capabilities of my exploration. I'd love to know: to what extent can this be used to predict returns, and what else I'd need to do for this to be a more robust analysis.

Finally, while I was planning on clustering on each individual sector, I wasn't able to get this finished. It was an interesting project to work on, and I will come back to in the future for sure!

Regardless, I think diversifying is even more important than I did when I started, and the diversification provided by cluster analysis in this

specific project was able to far out-perform the benchmark index, the S&P 500.