Portfolio Diversification Based on Clustering Analysis

Young Sim,
Meng Lu,
Kit Qian,
Shanshan Ding,
Jenny Gray

May 17, 2019

1.Introduction

According to Gallup (2017), the average percentage of Americans who own stocks from the year 2009 to 2017 is 54%, which has dropped 8% since the financial crisis. Many people have left the stock market because of non-ideal return or the lack of a resultful investment strategy that generate reasonable profits while withstanding the market's volatility. This is not a problem that only concerns retail investors but also many institutional investors. The modern portfolio theory suggests that investors can achieve this object through portfolio diversification, reducing risk by spreading a portfolio over many different investments (Bodie et al., 2014). Through portfolio diversification, investors can avoid excessive exposure to any one source of risk; an investor with a well-diversified portfolio may be immune to many of the risks faced by a corporation (Hull, 2018). The objective of the project is to use clustering methods to construct well-diversified portfolios that reduce volatility and losses and increase capital preservation.

The first part of the project constructs portfolios using different clustering methods on stocks included in S&P 500 and selecting the stock with the highest Sharpe ratio from each cluster. For k-means clustering, we use daily log returns as features whereas for agglomerative clustering, we use correlations between stocks as our custom distance metric as several previous studies suggest (León et al., 2017). Using Sharpe ratio and silhouette scores, we select our final model to test on the test data. In the second part of the portfolio, the test results are evaluated and compared based on Sharpe ratio with Dow Jones Industrial Average and portfolios of randomly selected stocks as benchmarks. We also compute the annualized volatility of the portfolio and compare it to that of each cluster as well as the S&P 500 to evaluate our model. We conclude our project by providing interpretations of the results from the first and second parts and suggesting further ways of improvements.

2. Data

The data analyzed is comprised of the daily close price of 470 common stocks that are issued by S&P 500, or the 500 large-cap companies and traded on American stock exchanges, from Feb 7, 2013, to Feb 6, 2018. The original data come from the Investor's Exchange API. The source script is available in the references section.

The five-year dataset is split into training period followed by test period of the last one year. The data in the training period is used to fit models to create clusters for portfolio. Training period 1 indicates the first four years of the data. Training period 2 indicates February 2016 to February 2017, which is one year prior to the test data. The data in test period is applied to evaluate the constructed portfolio during the test period to see if the model successfully created a well-diversified portfolio.

3.Methodology

A. Clustering

In this project, we use clustering as a method to find which assets are similar to or different from each other. The goal of clustering methods is to divide datasets into groups of similar data points. In general, clustering methods can be divided into three types: k-means clustering, agglomerative clustering, and DBSCAN.

K-means method is a partitional clustering algorithm that directly classifies objects within a dataset into k clusters based on their similarities. It iteratively groups each data point to its nearest cluster and relocates the cluster centers by their means until the intra-cluster sum of squares is minimized. K-means method works well for simple clusters of regular shape and similar size, and it is sensitive to the range of feature values. In scikit-learn, only Euclidean distances of the features is supported to quantify the similarities between data points, so we cannot use correlation-based distances as our distance metric. Hence, we use daily log returns on each date as features instead.

Agglomerative method belongs to hierarchical clustering algorithm. It classifies objects by collecting small clusters from bottom to top in a tree structure. The clustering process starts with declaring each point as its own cluster, and then merges the two most similar clusters into a single cluster according to their linkage. The stopping criterion in scikit-learn is the input number of clusters so the above is repeated until the specified number of clusters are left. Unlike in k-means, Agglomerative Clustering in scikit-learn allows custom distance metric. Therefore, correlation-based distance metric can be used. The custom distance metric we use follows:

$$distance = (1 - correlation)$$

There are four different linkage criteria to determine how similarities between two clusters are measured: single, complete, average, and ward. Single linkage merges the two that have the smallest minimum distance between their points. On the other hand, Complete linkage merges the two clusters that have the smallest maximum distance between their points. Average linkage merges the two clusters with the smallest average distance between all their points. Ward linkage, which is the default choice in scikit-learn, merges two clusters such that the variance within all clusters increases the least. Since ward only allows for Euclidean distance metric, it is left out for our purpose of using correlation distances.

DBSCAN (Density Based Spatial Clustering of Applications with Noise) method is often used to handle complex cluster shapes that contain noise points. It distinguishes core points from noise points by checking if the number of points within the distance of ε is greater than min_sample. At the end, all core samples within the distance of ε are classified into the same cluster. Unlike the first two methods, it can help determine the number of clusters, but it sometimes produces clusters of very different size. For this project, DBSCAN is left out because i) it often results in clusters of very different size, which may not work with the purpose of portfolio diversification by equally sized investment groups; ii) it lacks the flexibility of choosing desired number of clusters; iii) the noise points of our stock price data is already cleaned so the shape of clusters is not expected to be complex.

B. Sharpe Ratio

Sharpe ratio, as widely used for portfolio performance evaluation, is used to test the performance of the portfolios in this project. It is a means of measuring how much a portfolio outperforms the risk-free rate of return on a risk-adjusted basis. The higher the ratio, the better the performance. Computation of Sharpe ratio follows the below formula:

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

where R_p is the return of the portfolio, R_f is the risk-free return, and σ_p is the standard deviation of the portfolio's excess return. For simplicity, we assume that the risk-free rate is zero.

C. Log Returns

Here we are focusing on log returns rather than closing prices or simple returns. the log return at date i, $\log R_i$, can be written as

$$logR_i = lo\,g\left(1 + \frac{p_i - p_{i-1}}{p_{i-1}}\right) = \ log\ (1 + r_i)$$

where p_i is the close price on date i. The advantage of utilizing returns, versus prices, is normalization. By using returns, we can measure all variables in a comparable metric.

There are also several advantages to using log returns over simple returns. First, we can utilize log normality. Assuming prices are distributed lognormally, (which may or may not be true in practice), we have $log\ (1 + r_i)$ normally distributed, which is convenient since much of classic statistics presumes normality.

Also, we can use the approximate equality between log returns and simple returns. When returns are fairly small (which is often the case for trades with short holding durations), the following approximation is true:

$$log(1 + r) \approx r$$

Finally, log returns are time additive. For simple returns, the compounded return, or the running return, of ordered sequence of trades over time follows:

$$\prod_i (1 + r_i)$$

This formula can be difficult to work with, as the product of normally distributed variables is not normal. On the other hand, the sum of normally distributed variables is normal (given all variables are uncorrelated). Therefore, compounded log returns are normal. Moreover, this gives us an easy formula for compounded log returns, simply taking the difference in log between the beginning and the end periods:

$$\sum_i \log(1 + r_i) = \log(p_n) - \log\ (p_0)$$

D. Portfolio Construction and Evaluation

For the dataset of S&P500 stocks, we cluster stocks according to their log returns. As mentioned earlier, we use daily log returns as features for k-means method and the correlations distance metric for agglomerative method. We train 4 models of 30 clusters based on the training data, each using k-means method and agglomerative methods with single, complete, and average linkages. Once each stock is assigned a group index according to the clustering results, 4 portfolios corresponding to each model are constructed by selecting the stock with the highest Sharpe ratio from each cluster with equal weight. As a result, we get 4 portfolios, each consisting of 30 stocks. We choose 30 as the number of clusters for portfolio construction as this allows us to compare our portfolios directly to Dow Jones Industrial Average, which also has 30 stocks.

In order to evaluate how well the stocks are clustered, we use Sharpe ratio, silhouette scores, and annualized volatility. Higher Sharpe ratio indicates better risk adjusted return of the portfolio, which suggests that the portfolio is well-diversified. We compare Sharpe ratio in the portfolios constructed as well as in benchmarks. Since the purpose here is fore mere comparison of different models over the same period, we used daily Sharpe ratio: the mean of log returns over the period divided by the standard deviation of log returns over the period. If our portfolio has a higher Sharpe ratio than the benchmarks, the result is successful. Silhouette score measures how similar an object is to its own cluster in comparison to other clusters. It ranges from −1 to +1, with higher value indicating that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then there may be too many or too few clusters in the model.

Lastly, the annualized portfolio volatility is computed for each cluster to evaluate diversification:

$$vol = \sqrt{w^T \Sigma w}$$

where Σ is the covariance matrix of returns, w and w^{T} are the portfolio weights and its transpose. If our clusters are well constructed and group similar stocks into each cluster, the volatility in each cluster should be higher than the volatility of the portfolio. We also compare the volatility of the portfolio with the volatility of benchmarks: portfolio of 30 randomly selected stocks and equally weighted portfolio consisting of all 470 stocks in the data. Taking into account that the latter portfolio consists of much larger number of stocks and that it roughly represents the market volatility, it would be very difficult for our portfolio to have lower volatility. However, if the volatility of our portfolio is close to the volatility of the newly constructed market portfolio, we can conclude that our portfolio is well diversified.

E. Benchmarks

There are two benchmarks that indicate the market performance used in this project. The first one is Dow Jones Industrial Average (DJIA). DJIA indicates the value of 30 large, publicly owned companies based in the United States that are also included in the S&P 500 Index. While DJIA is a widely used stock market index, it has some limitations as benchmark for the entire market. First, its value is not a weighted mean but rather the sum of the price of one share of stock for each component company price. Weighted average, and thus each company's market capitalization, is not represented. Therefore, DJIA gives more emphasis on higher-priced stocks over the average than their lower-priced counterparts and takes no account of the relative industry size or market capitalization. Moreover, since it consists of 30 companies with highest market capitalization, not only is it arguable that the number of stocks is not sufficiently large to represent the entire market, but also that the index is influenced by economies of scale.

Considering these limitations of DJIA as a benchmark, we also used portfolios of randomly selected stocks as another benchmark, similar to the random selection portfolio constructed by Zhan et al. (2015). To construct this benchmark portfolio, 30 stocks are selected at random using a uniform distribution. But instead of simply drawing 30 stocks at random once, we constructed 30 such portfolios and computed the average Sharpe ratio of the 30 portfolios, in order to control variability. The same benchmark is also used to evaluate volatility.

4. Results

The results for the training period 1 are as follows in Table 1. The portfolio constructed by k-means clustering has Sharpe ratio of roughly 0.122 in the test period. The portfolios constructed by agglomerative methods with single, complete, and average linkages have Sharpe ratio of roughly 0.111, 0.086, and 0.089, respectively. Sharpe ratio for DJIA and the portfolio of randomly selected stocks are 0.15 and 0.102 respectively over the same period.

Figure 1 displays a graph of silhouette scores for each model given different numbers of clusters in training period 1. When number of clusters is 30, single linkage agglomerative clustering has lowest score while the other three clustering methods have similar scores. This indicates that clusters from single linkage model are least well-constructed in terms of similarities within each cluster. These scores are inconsistent with the results above, in which single linkage had the highest performance among the agglomerative models.

Table 2 shows results for training period 2. The portfolios constructed by k-means clustering and agglomerative methods with single, complete, and average linkages have Sharpe ratio of roughly 0.095, 0.075, 0.115, and 0.126, respectively. While all the models performed worse than DJIA, agglomerative methods with complete linkage and average linkage had better performance than portfolio of 30 randomly selected stocks. Figure 2 visualizes silhouette scores in training period 2. The scores indicate that single linkage model perform the worst clustering given 30 clusters, which is consistent with the results from Sharpe ratio. The results from training period 2 are much more consistent with the expectations and silhouette scores, indicating better clusters are constructed. Heatmaps in Figures 3, 4, and 5 also suggest the model from training period 2 do a better job clustering. Therefore, we select the best performing model, which is agglomerative clustering model with average linkage trained with data from training period 2, to evaluate volatility.

Figure 6 displays the annualized volatility in the test period. The bars show volatility for an equally weighted portfolio of all 470 stocks, our selected model, equally weighted portfolio of 30 randomly selected stocks, and the average volatility of all 30 clusters in our selected model, respectively. The values are roughly 0.086, 0.09, 0.093, and 0.192, respectively. Also, by comparing the volatility of each cluster to the volatility of our selected model, we find out that only one cluster out of 30 clusters have a volatility lower than the selected portfolio.

5. Discussion

In the results from training period 1, k-means clustering showed the best performance among the models, unlike expected. While none of the models performed better than DJIA in terms of Sharpe ratio, k-means clustering and agglomerative method with single linkage had better performance than portfolio of 30 randomly selected stocks. These results are inconsistent with the silhouette scores shown in figure 1, where single linkage model has the lowest score. This may be because we are adding noise by using training data from too long ago; stock prices from 5 years ago may not necessarily be consistent with the prices today. Suggesting the same, the results from training period 2 look more consistent with silhouette scores displayed in Figure 2. As expected, k-means clustering showed worse performance than complete and average linkage models. The k-means method uses daily log returns as features with Euclidean distances whereas the agglomerative methods use correlation-based distance metric. Thus, the results suggest that correlations work better than returns for clustering.

To conclude that our portfolio is well diversified and generates returns, the portfolio constructed from the model should have higher Sharpe ratio than the benchmarks. The results are not ideal in terms of Sharpe ratio; the selected model had a lower Sharpe ratio than the DJIA unlike in the training period. However, the model performed better than random selection, suggesting that the model is working in the right direction but performing worse than the market index. Note that silhouette scores are higher when number of clusters are smaller in Figure1. This indicates that when number of clusters is smaller, stocks in the same clusters are more “similar”. For unsupervised methods like clustering, however, the selection of parameters is rather qualitative, so this does not necessarily mean smaller number of clusters is better for our purpose of study. This could still suggest, however, that any undesired or unexpected results could be due to too many clusters. Some suggestions on ways to improve our model performance are made in the conclusion section.

The goal of portfolio diversification is to maximize risk adjusted returns but the primary goal is to reduce volatility. In this sense, our model still generates some meaningful results. Taking into consideration that our S&P500 data is composed of 470 companies, we can claim that the volatility of our portfolio is satisfying. Our model also performs better than random selection both in terms of Sharpe ratio and volatility. Since the volatility of the selected portfolio was lower than the average volatility of all 30 clusters as well as the volatility of each of the 30 clusters, our portfolio diversification is satisfactory in terms of volatility reduction.

6. Conclusion

In conclusion, the goal of this project is to construct a well-diversified portfolio using clustering. We trained 4 different models using k-means clustering with daily log returns as features and agglomerative clustering methods with single, complete, and average linkages based on correlation-based distances. By selecting the stock with the highest Sharpe ratio from each cluster, we constructed 4 equally weighted portfolio each consisting of 30 stocks. Based on silhouette scores and Sharpe ratio, we selected agglomerative clustering with average linkage trained on last one year of data as our final model. The performance of our selected portfolio in the test period in terms of Sharpe ratio performed better than random selection but worse than DJIA. Results in terms of volatility showed better performance; our selected portfolio had an annualized volatility lower than random selection as well as the average volatility of all clusters and fairly close to that of equally weighted portfolio consisting of all 470 stocks in the data. Also, only one out of 30 clusters had a volatility lower than that of the selected portfolio.

There are a few ways to possibly improve our model performance. First, we could further adjust the length of training period. We could try adjusting the number of clusters, since we had higher silhouette scores for smaller numbers of clusters. We could also change or extend pool of stocks. Since we are only selecting stocks from S&P500, they are all large companies and may have similarities to some extent. This could be the reason why silhouette scores for smaller number of clusters are higher. Another way to assure well-constructed clusters is to use distance threshold. Distance thresholds define maximum distance within a cluster, so that components of a cluster are “similar” enough. Finally, we could improve the risk adjusted return of the portfolio by assigning weights on each stock of the portfolio based on an optimization problem with maximizing Sharpe ratio as a constraint.

In terms of performance evaluation, we could add additional means of portfolio evaluation like Sortino Ratio, Omega Ratio, VaR, and CvaR to draw more concrete conclusion. We can also construct another benchmark by selecting three stocks with the highest Sharpe ratio in each of 10 industry sectors. We can use GICS number to identify industry sector for each stock.

Furthermore, we could conduct some more qualitative analysis using the industry sector data to understand the commonalities within each cluster, by looking if the stocks in the common cluster are similar in terms of sectors. We can also utilize time series analysis to fit a model that describes the changes in stock prices throughout the duration.

7. List of Tables

A. Table 1: Sharpe Ratio (Training Period 1)

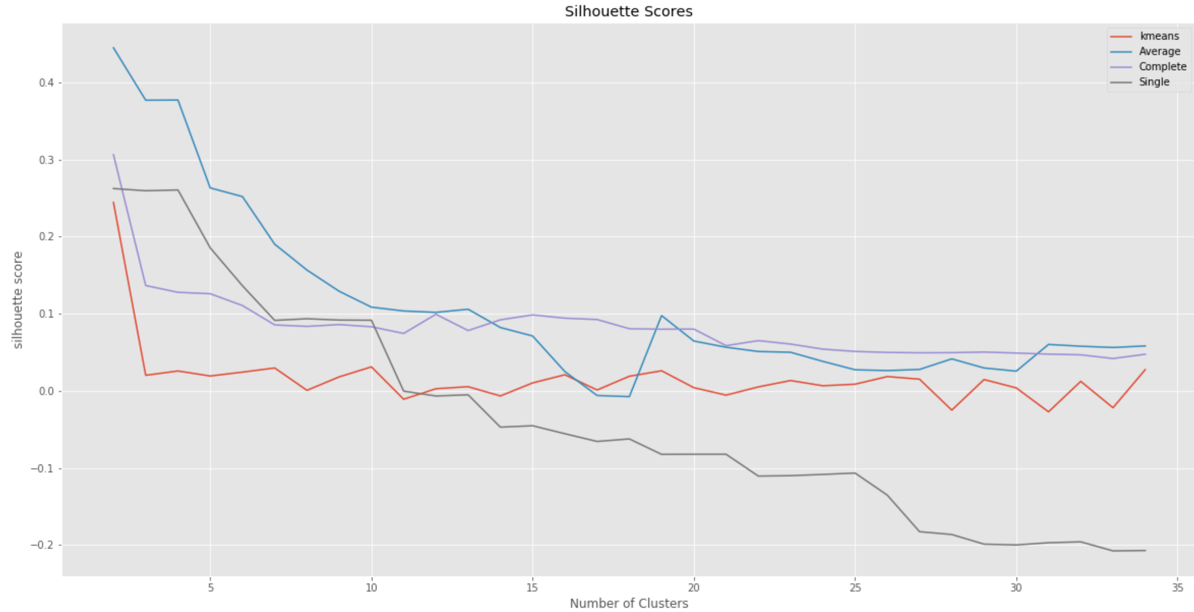
	Models				Benchmarks	
	k-means	Agglomerative (single)	Agglomerative (complete)	Agglomerative (average)	DJIA	Randomly Selected
Sharpe Ratio	0.122447	0.11101	0.085972	0.089122	0.150174	0.102065

B. Table 2: Sharpe Ratio (Training Period 2)

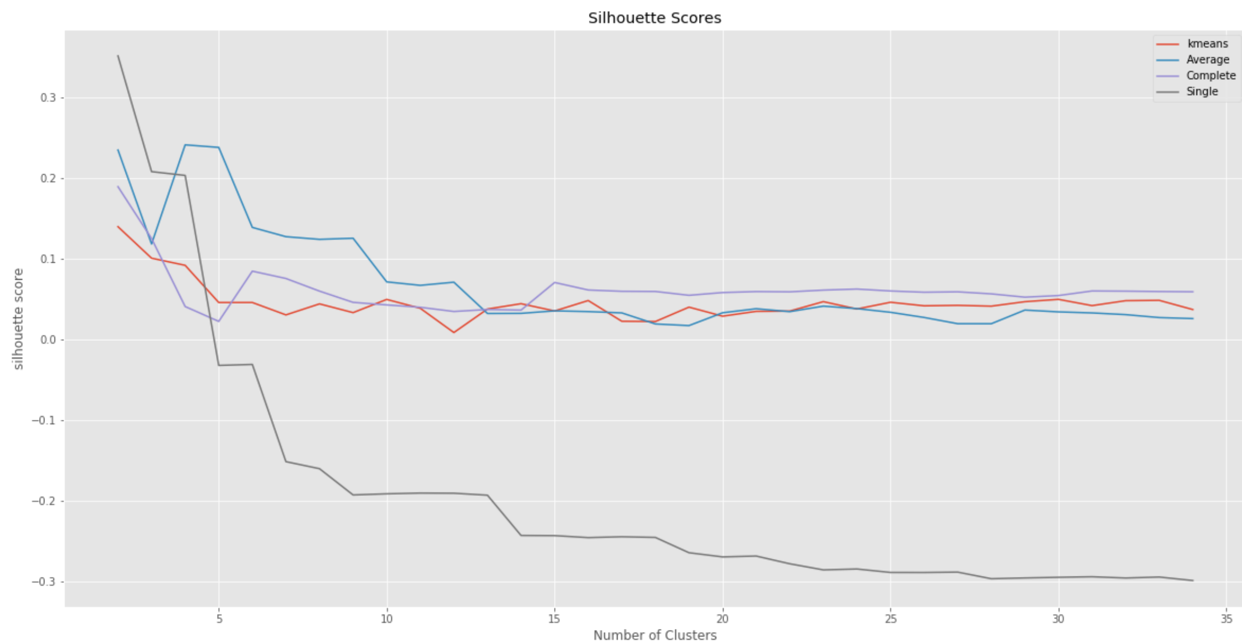
	Models				Benchmarks	
	k-means	Agglomerative (single)	Agglomerative (complete)	Agglomerative (average)	DJIA	Randomly Selected
Sharpe Ratio	0.094728	0.075378	0.115475	0.125963	0.150174	0.102065

8.List of Figures

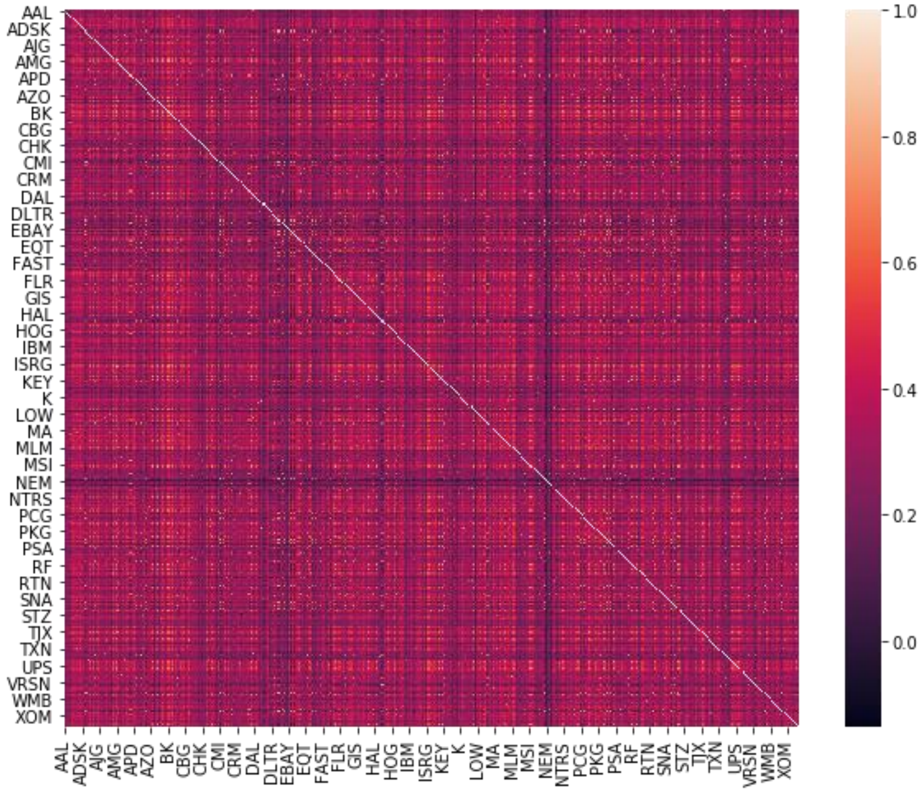
A. Figure 1: Silhouette Scores (Training Period 1)



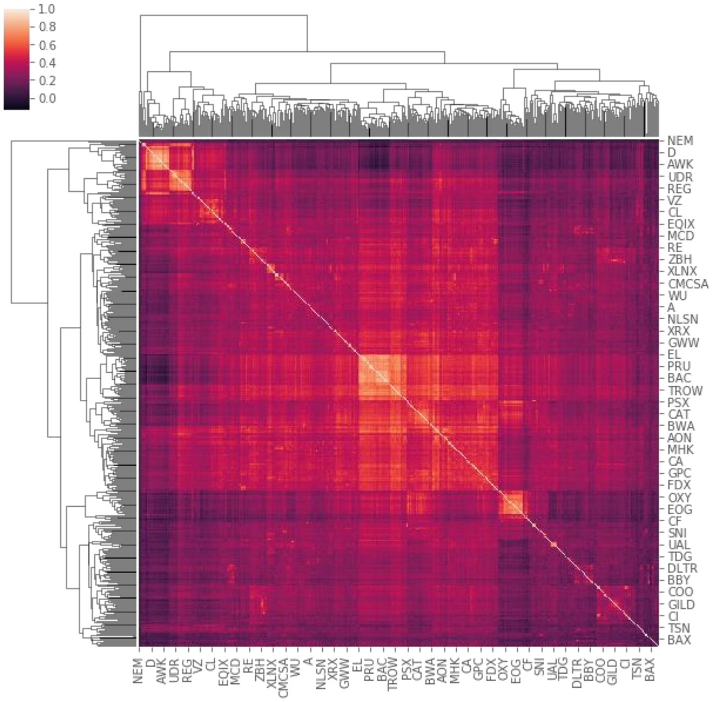
B. Figure 2: Silhouette Scores (Training Period 2)



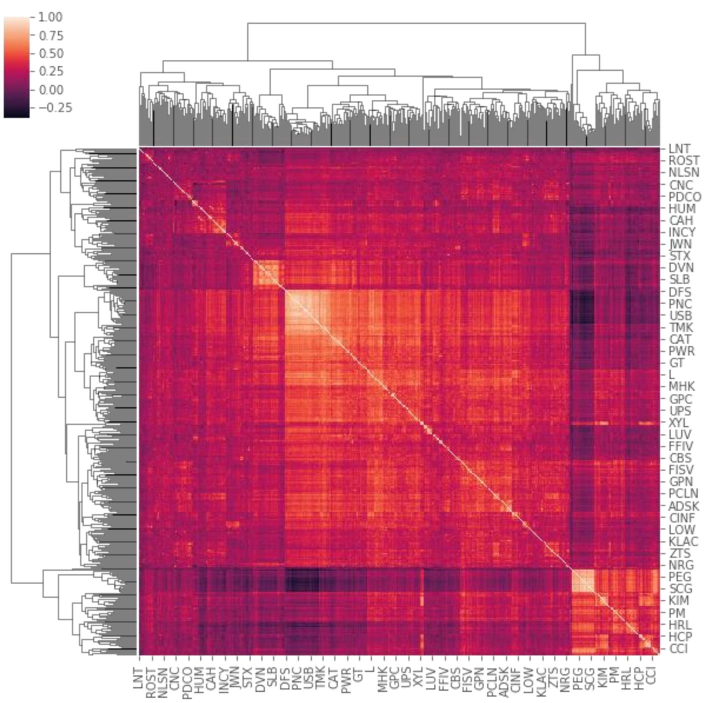
C. Figure 3: Heat Map (Correlation Matrix)



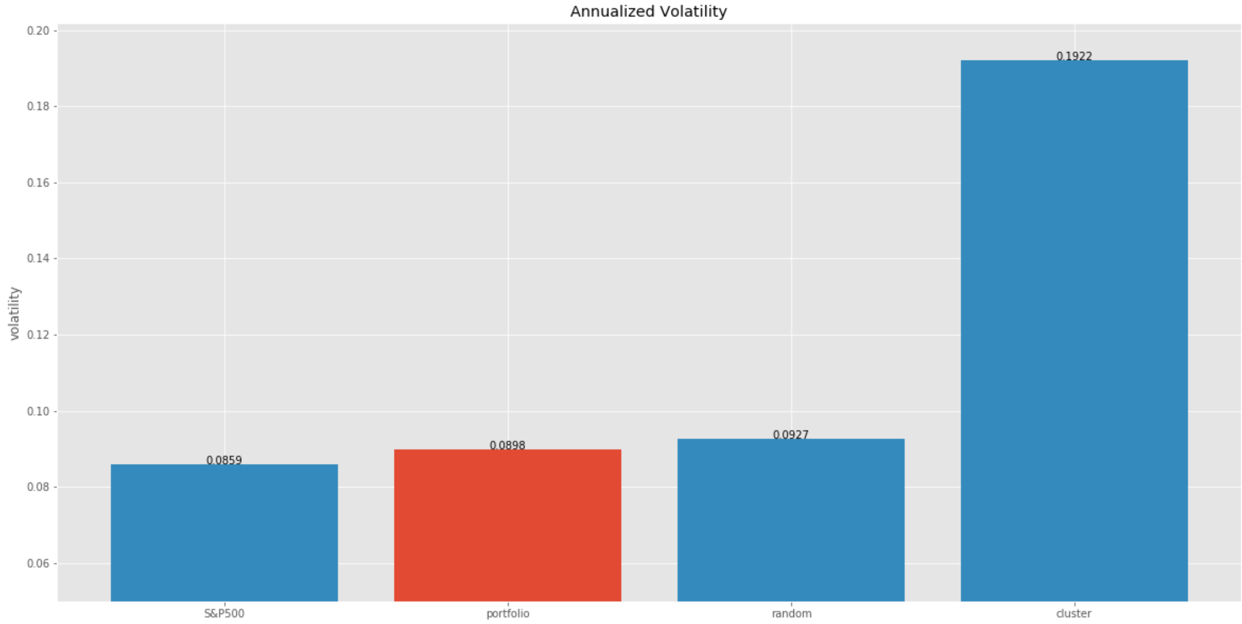
D. Figure 4: Heat Map (Stocks Clustered by Correlation Distances, Average Linkage, Training Period 1)



E. Figure 5: Heat Map (Stocks Clustered by Correlation Distances, Average Linkage, Training Period 2)



F. Figure 6: Volatility



9. References

Bodie, Z., Kane, A., & Marcus, A. (2014). Investments (10th ed.). New York: McGraw-Hill Education.

Cheng, H., Zhan, J., Rea, W., and Rea, A. (2015). An Application of Correlation Clustering to Portfolio Diversification. arXiv preprint arXiv:1511.07945

Gallup. “U.S. Stock Ownership Down Among All but Older, Higher-Income.” 24 May 2017. <https://news.gallup.com/poll/211052/stock-ownership-down-among-older-higher-income.aspx>

Hull, J. (2018). Options, Futures, and Other Derivatives (10th ed.). New York: Pearson Education, Inc.

Kaggle. “S&P 500 Stock Data.” 2018. <https://www.kaggle.com/camnugent/sandp500>

León, D., Aragón, A., Sandoval, J., Hernández, G., Arévalo, A., and Niño, J. (2017). Clustering Algorithms for Risk-Adjusted Portfolio Construction. *Procedia Computer Science, 108*, 1334-1343.