Final Project – Spring 2023 – JHU AS.440.617

Jarquisha Croom
Sam Dreyfuss
Jeff Gerlica
Sarah Miller
Owen Sullivan

Abstract

In this research, we found a quantitative approach to identifying potential stock investments for an investment fund. To focus on broad market coverage selections, we elected to use the S&P 500 index for our study. By leveraging a K-means clustering process, we identified the most promising stocks based on their proximity to the ten centroids on a rolling 1-year simple daily average return/volatility plane across the entire index. A minimum variance portfolio optimization process was then used to derive the optimal weights for the selected stocks in the portfolio created by the K-means process. We employed an autoregressive integrated moving average model to simulate the return time series of the individual stocks and provide short-term forecasts. Applying the portfolio weights to these forecasts provided a promising estimate of future portfolio values to provide an investment manager. Our approach demonstrated a 0.03% average daily return and a standard deviation of under 1%, highlighting its potential utility for quantitative investment decisions.

I. Introduction

As quantitative analysts, we were tasked to make an educated choice of 5-10 stocks for potential investment. The problem that needed to be solved was how can an analyst take the past pricing data of an index of stocks, extract performance analytics at the daily level, and model the data accurately enough to predict future performance. The layers of the problem involved shrewdly selecting the stocks from the index, obtaining portfolio weights based on the assumption of a risk-averse investor, and fitting and confirming the appropriate model for predictions.

To construct a well-diversified portfolio given only long positions in equity positions, we needed to choose securities that are able to perform well through business cycles. Due to the 5- to 10-stock constraint, we needed to incorporate an intentional, logical way to choose which tickers will be included in the optimal portfolio of a risk-averse investor. To do this, we employed a machine learning unsupervised classification algorithm called K-Means Clustering. The algorithm incorporated a 10-centroid constraint which revealed a set of 10 quantitatively-diverse stocks which were identified as being closed to the K-Means centroids.

The set of stocks were input into a minimum-variance optimization framework to identify the weights of our portfolio. Obtaining the weights that cater to a minimum variance portfolio was suited to our assumption of the risk averse investor and hedging against potential losses during periods of bearish markets.

To ensure our portfolio performed over time, we needed to fit a model to generate reliable future values of returns for the portfolio and individual assets. An auto-regressive integrated moving average (ARIMA) model was adopted to solve this problem. The ARIMA framework allowed predictions and confidence intervals in the short term out to 10 days. The forecasts contributed to investor confidence and provided actionable information to the investment fund manager on investment strategy, which was the main problem we were aiming to solve.

II. Methodology

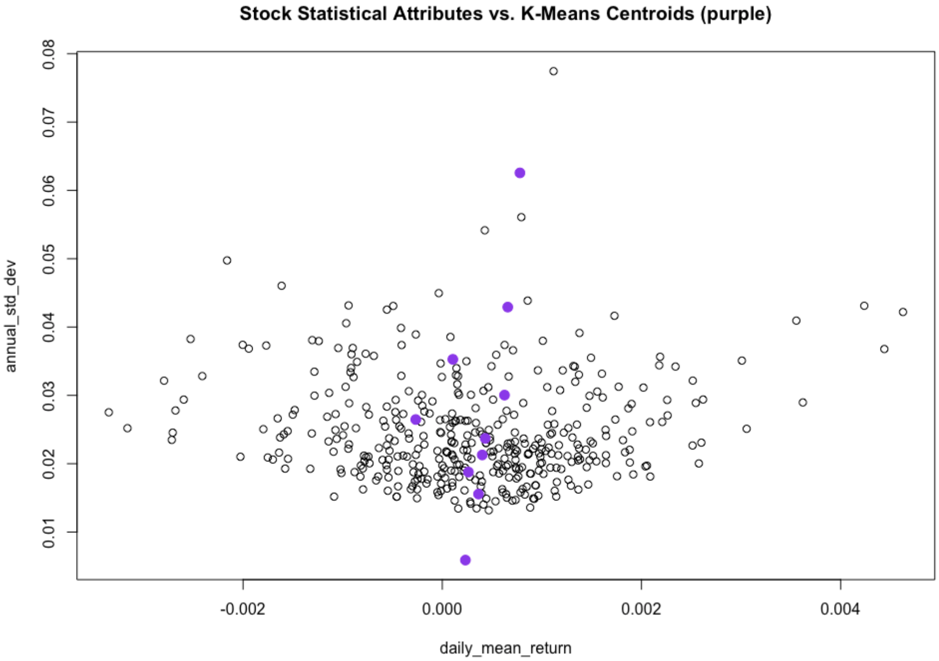
- **Detailed explanation of the stock index selection process**

We chose the S&P 400 composite index to offer a full range of business sectors for our model to choose. The companies in the S&P 400 are generally considered to be some of the largest and most well-known companies in the US, while the companies in the S&P 400 are considered to be slightly smaller but still well-established companies. Each constituent of the index was individually queried and returns and standard deviations for returns were calculated. In addition to the properties of the index ranging across multiple sectors, financial managers over the long term fail to beat the S&P 400 over time in the aggregate due to the efficient market hypothesis (EMH). The EMH says that information, to include private information in the strong form, is incorporated into asset prices; therefore, investors are not able to obtain extra normal returns over the market portfolio. Our analysis led us to investigate building the portfolio of the investment fund with a subset of the index limited to 10 stocks. The limit on 10 stocks could manifest in the real world (due to client constraints and high transaction costs). With a limit on the stock quantity, this could increase expertise and deeper understanding of the underlying companies invested in.. We hedged against these possibilities and provided an optimal portfolio to the risk-averse investor.

- **Explanation of the K-means clustering algorithm**

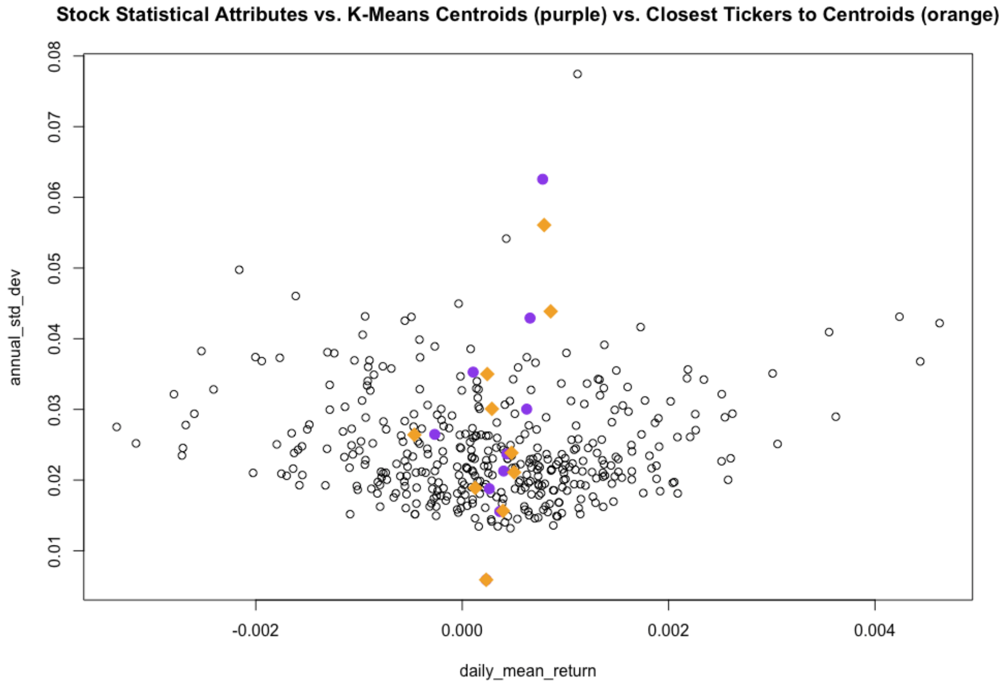
The K-means clustering algorithm is an unsupervised machine learning method primarily used in data classification. A given set of data points, in this case, stock mean returns and standard deviations, is partitioned into k clusters, where k is a user-defined parameter, in this case, up to 10 stocks. It works by iteratively assigning each data point to the cluster whose mean is closest to the data point, and then recalculating the means of each cluster. This process is repeated until the clusters no longer change significantly. The data points are called centroids. The analysts exploited this property of similarity to create diversity among the assets that constitute the final portfolio. The stocks we selected were based on the location of these ten centroids.

After calculating the daily mean returns and annual standard deviations of the SP400 universe, we calculated our ten centroids (in purple):



In the assignment step, each data point was allocated to its nearest centroid, based on Euclidean distance. This resulted in 10 clusters of data points with unique centroids. In the update step, the centroids' positions were recalculated. These newly calculated centroids represent the new "center" of each cluster. The assignment and update steps were repeated until the centroids produced no significant movement of the clusters, signaling convergence to a solution.

In our research, we employed this method to identify diverse stocks. Each stock was represented by a data point characterized by two attributes: return and volatility. By applying K-means, we identified clusters of stocks with similar performance characteristics, allowing for a systematic selection of diverse stock candidates for our portfolio. The chosen stocks were those closest to each of the 10 centroids, representing a diverse cross-section of performance attributes within the S&P 400.



- **Description of the minimum variance portfolio optimization**

The next step was to determine the weighted composition of the 10-stock portfolio. The goal was to find the weights of each stock that would provide the lowest possible portfolio variance subject to constraints. The first constraint was the portfolio weights must add to 1, therefore the 10 stock weights construct the entire portfolio. The second constraint was that all weights must be greater than or equal to 0. This ensured a “long only” portfolio, in that no shorting of stocks was possible. The portfolio optimization algorithm that was utilized searched 20,000 randomly generated portfolio constructions to identify the composition that yielded the minimum portfolio variance. The stock weights were then used to calculate historical portfolio returns, which were then used to fit an ARIMA model in order to predict future returns of the portfolio.

- **Explanation of the ARMA model for forecasting**

After obtaining the return series using adjusted closing prices, we needed to find the appropriate ARIMA model to fit for accurate predictions of each individual time series. The auto ARIMA function was used to determine the appropriate order for each series, relative to each univariate time series in the selected stocks data frame of daily returns. Various combinations of AR, MA, and ARMA models were selected. The model is dependent on the real time data and when the adjusted prices are pulled; therefore each time the model is run, the ARIMA order will be different. Adjusted closing prices were used to account for events such as dividend issues and stock splits over the past year.

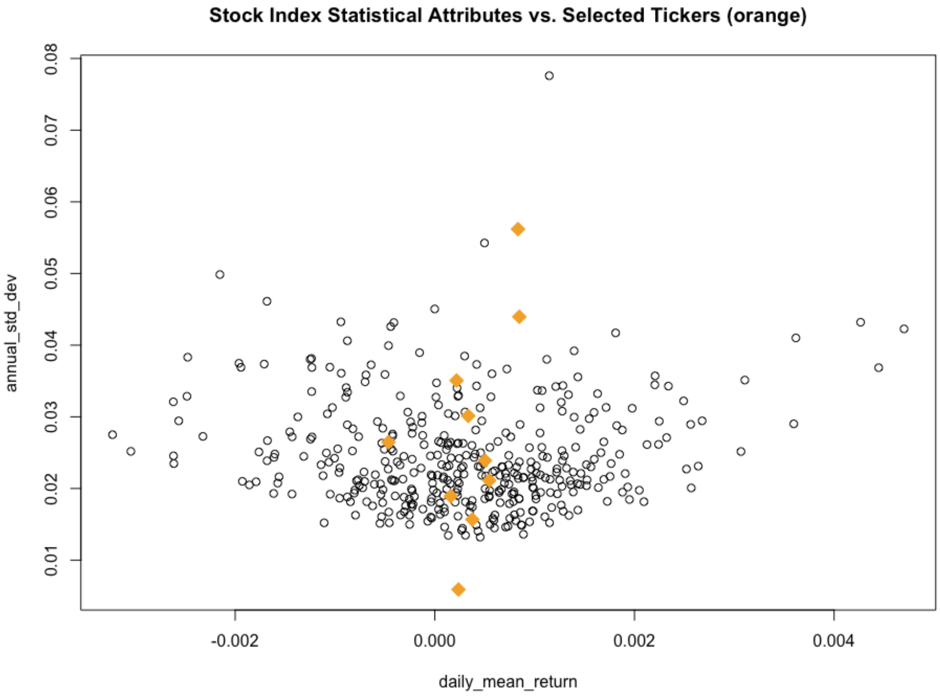
After the order is determined, the researchers checked for statistical significance in the coefficients using a 95% confidence level. Insignificant lags in the autocorrelation function were excluded from the final model. In addition to testing coefficients, we also checked for serial correlation between the lags via the L-Jung Box test defining the null hypothesis as no serial correlation.

Once the appropriate model was fit and checked, the 10-step ahead predictions were made where each step is one day. Utilizing the predicted returns, the weighted predicted returns are found by multiplying the predicted returns of each stock on each day by the portfolio weights of each stock obtained by the portfolio optimizer. The predicted portfolio returns were then found by adding the weighted predicted returns for each stock on corresponding days.

III. Results

- **Evaluation of past performance of selected stocks**

The date range for the stock returns utilized in the portfolio optimization ranged from the past year through the previous business day. The 10 stocks selected by the K-means clustering algorithm as well as the mean daily return and standard deviation and the weights assigned by the portfolio optimization algorithm were:

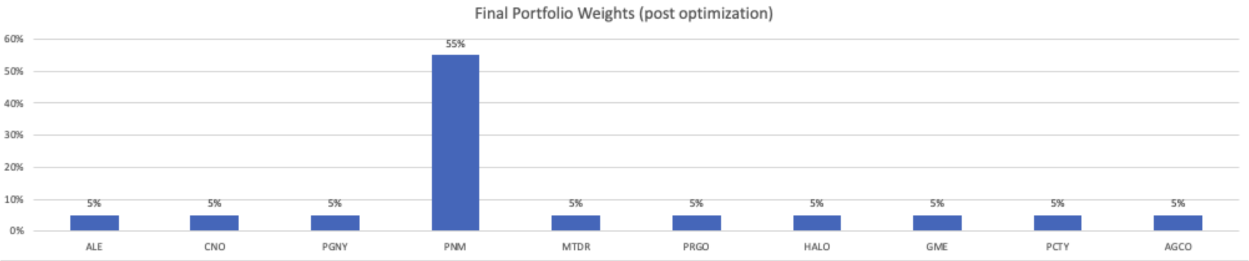


- **Evaluation of past performance of the portfolio**

The weights assigned to each stock respectively were all 5% due to the portfolio optimization constraint with a 55% weighting to the PNM as shown in the “Final Portfolio Weights (post optimization)” chart below. The historical portfolio variance was found to be 0.01%, and the mean daily portfolio return was found to be 0.03%.

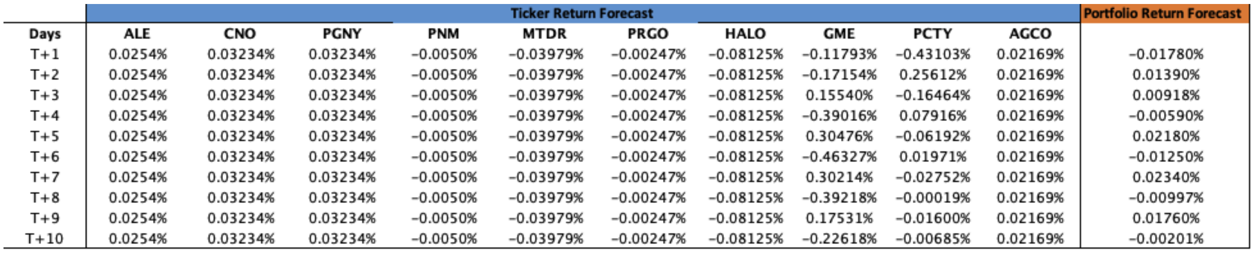
The equation for the fitted autoregressive model fitted to the stock return data for the 10 selected stocks was an autoregressive integrated moving average (ARIMA) model. The 10-day forecasted returns for the 10 individual stocks as well as the 10-day forecasted portfolio returns are provided in the “10-Day Stock and Portfolio Return Forecast” table below.

- **Optimized min-variance portfolio weights (with 5% diversification constraint)**



- **Forecasted future values of the stocks and the portfolio**

10-Day Stock and Portfolio Return Forecast



	Ticker Return Forecast										Portfolio Return Forecast
Days	ALE	CNO	PGNY	PNM	MTDR	PRGO	HALO	GME	PCTY	AGCO	
T+1	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	−0.11793%	−0.43103%	0.02169%	−0.01780%
T+2	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	−0.17154%	0.25612%	0.02169%	0.01390%
T+3	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	0.15540%	−0.16464%	0.02169%	0.00918%
T+4	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	−0.39016%	0.07916%	0.02169%	−0.00590%
T+5	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	0.30476%	−0.06192%	0.02169%	0.02180%
T+6	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	−0.46327%	0.01971%	0.02169%	−0.01250%
T+7	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	0.30214%	−0.02752%	0.02169%	0.02340%
T+8	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	−0.39218%	−0.00019%	0.02169%	−0.00997%
T+9	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	0.17531%	−0.01600%	0.02169%	0.01760%
T+10	0.0254%	0.03234%	0.03234%	−0.0050%	−0.03979%	−0.00247%	−0.08125%	−0.22618%	−0.00685%	0.02169%	−0.00201%

IV. Discussion

- **Interpretation of the results**

Although sector and industry were not incorporated into the selection process for the stocks directly, it was the desire of the analysts to invest in a diverse set of companies. The goal of the K-means clustering selection process was to ensure diversity in risk and return composition directly, and sector and industry indirectly. The K-means selection process has accomplished this goal, with the Utilities, Financial Services, Healthcare, Energy, Consumer Cyclical, Technology, and Industrial sectors represented among the 10 selected stocks. This diversification among sectors is indicative of a portfolio inherent with minimal risk correlation even before the minimum variance algorithm is applied to determine the appropriate weights of each stock.

	ALE	CNO	PGNY	PNM	MTDR
Name	Allete	CNO Financial Group	Progyny	PNM Resources	Matador Resources Company
Sector	Utilities	Financial Services	Healthcare	Utilities	Energy
Industry	Utilities - Diversified	Insurance - Life	Health IT Services	Utilities - Reg. Electric	Oil & Gas E&P
Market Cap	3.604B	2.433B	3.557B	4.117B	3.557B

	PRGO	HALO	GME	PCTY	AGCO
Name	Perrigo Company	Halozyme Therapeutics	Gamestop	Paylocity Holding Corporation	AGCO Corporation
Sector	Healthcare	Healthcare	Consumer Cyclical	Technology	Industrials
Industry	Drug Manu. - Specialty & Generic	Biotechnology	Specialty Retail	Software - Application	Farm & Heavy Construction Machinery
Market Cap	4.934B	4.669B	6.305B	9.115B	9.51B

One challenge with this method of screening for investments to purchase is that there is a trade off between diversification and assets return predictability. There is no guarantee that the chosen assets for the portfolio will have predictable return patterns, and if they do, will have enough weight as a percentage of the portfolio to influence the portfolio's return even if several assets returns are shown to be predictable at a statistically significant level. As seen below, only two of the assets within our final portfolio amounting to 10% of portfolio weight had an AR or MA model order larger than 0.

- **Areas for additional improvement**

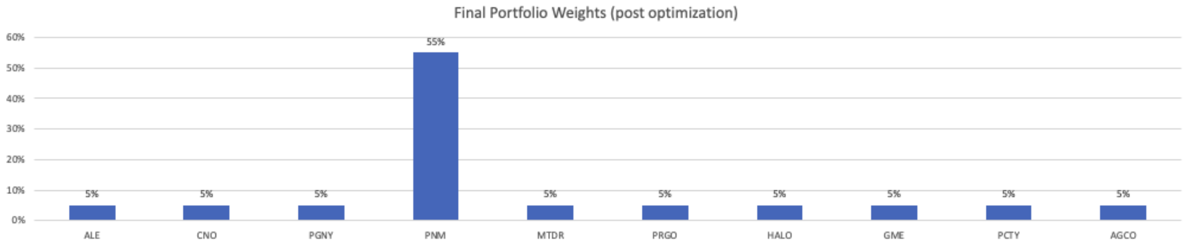
To improve upon the current framework, screening the S&P index for tickers which are shown to have ARIMA model orders larger than zero would make for a more appropriate input ticker universe to be passed to the K-Means algorithm in the case the portfolio managers of the fund wanted to improve their understanding of likelihood of future returns.

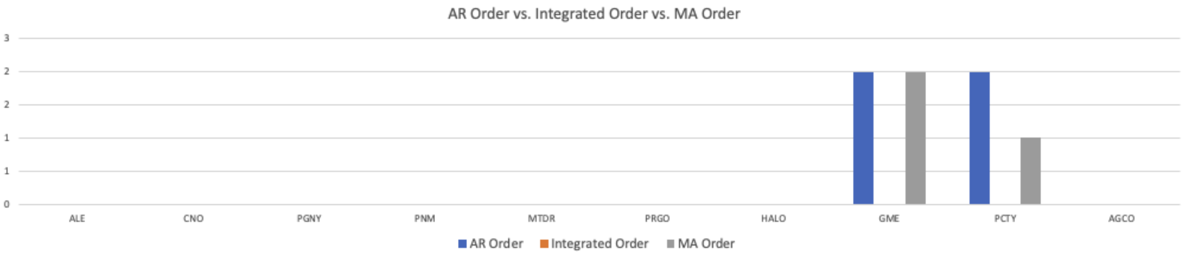
A further expansion area would be to incorporate determining optimal portfolio weights for a risk neutral investor and subsequently a risk seeking investor by modeling an entire efficient frontier and determining a capital allocation line by incorporating a risk free asset in addition to the minimum variance portfolio. Additionally, allowing the optimizer to short stocks could also be beneficial as this could allow the investors to take advantage of potential economic downturns or express explicit negative views towards assets.

V. Conclusion

- **Summary of the findings**

The K-means algorithm selected 10 stocks with various risk and return profiles from the set of S&P MidCap 400 Index. The selected stocks represented 7 different sectors of the economy, ensuring a diverse and minimal risk correlation. After completing our analysis, we successfully constructed a diversified portfolio which had some predictability due to the presence of order 1 or 2 lags depending upon the stock within the AR and MA components of our ARIMA capable model. The minimum portfolio variance optimizer algorithm assigned weights to the 10 stocks; 9 of the stocks were given weights of 5%, and one stock (PNM) was given a 55% weighting. The diversified portfolio demonstrated a historical daily average return of 0.03% and standard deviation of less than 1%.





- **Potential applications**

The analysis conducted has a vast array of further applications. The three portions of the problem solving methodology can each individually be explored further. The K-Means clustering classification process is a widely studied algorithm and has the capacity to take more input parameters. Investing is an individualized discipline, and depending on individual or firm preferences (min variance or not) the firm can adjust the parameters to fit the needs of the firm. In the minimum variance portfolio methodology, we are assuming a risk averse investor. Expansion of the optimization to different labels of risk tolerance will be useful to make the process more robust to individuals.

VI. References

- **Citations to the references used in the research**

- Baganzi, R. , Kim, B. and Shin, G. (2017) Portfolio Optimization Modeling with R for Enhancing Decision Making and Prediction in Case of Uganda Securities Exchange. *Journal of Financial Risk Management*, **6**, 325-351. doi: 10.4236/jfrm.2017.64024. https://www.scirp.org/journal/paperinformation.aspx?paperid=80120
- https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=65f1232434c5eeddd9e658db7ae0dd5c47b6e20d
- https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp
- Various materials from our course https://jhu.instructure.com/courses/36320/modules
- Tsay, R (2005) Analysis of Financial Time Series. Wiley