Stock Market Analysis & Prediction

Problem Statement

The financial industry revolves around the stock market and ultimately to make money for the shareholders. This report analyzes the time series data of the S&P 500 and a portfolio of stocks to see how they perform against each other and also to see how effective ARIMA modeling is for forecasting stock prices.

The information in this report helps financial professionals such as stock brokers, hedge fund managers, and individual investors. By assessing the differences of how an index performs compared to a portfolio, the client can use this information with other resources to best make their decisions in trying to maximize their returns on investments in the stock market.

Dataset

The data was collected using the Alpha Vantage API. The API key is free to obtain and the stock market data, which looks back up to 10 years, is free. They also provide a Python package for easy use with their API.

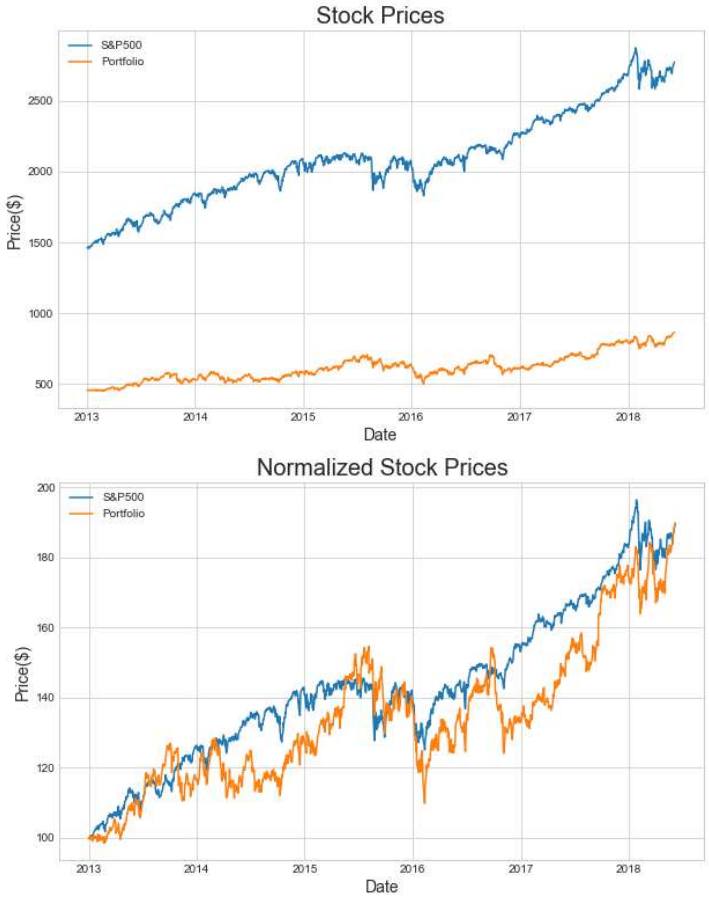
The random stocks for the portfolio were selected by referencing random stock generators online and also filtering for stocks that at least have 5 years of history starting from January 2013. The data collected from the API is for daily closing prices for this report. There are no NaN values in the data set since all the stocks selected and the S&P 500 have records from January 2013.

It should be noted that this introduces a level of survivorship bias to the random selection process as only the companies that survived for this time frame could have been selected for this procedure. The random stock tickers included for the portfolio are as follows:

'SRPT', 'FNLC', 'EVT', 'MDLZ', 'AGI', 'ERIC', 'PERY', 'ELLI', 'FYX', 'CSOD', 'AUMN', 'RPXC', 'OSUR', 'NEOG' ,'MZF', 'VCRA', 'YUM', 'MYGN', 'UVE', 'ALNY'.

The data set used in the main report includes the S&P 500 and the portfolio, which is created by adding all the stock history of the 20 stocks selected. The index is set as DateTime and is a daily frequency. Please refer to the Jupyter notebook Data_Wrangling for more details.

S&P 500 vs Stocks Portfolio



Upon initial look of the plotting for the data, the S&P 500 and Portfolio are not comparable due to their scale difference. The S&P 500 index contains 500 large companies' stocks and thus will be on a higher scale compared to the portfolio of 20 stocks of variable sizes. So for a better comparison, I normalized the stock prices by dividing each set by their respective first stock price value and multiplying by 100 to simulate what the returns of a $100 investment from the start of the data.

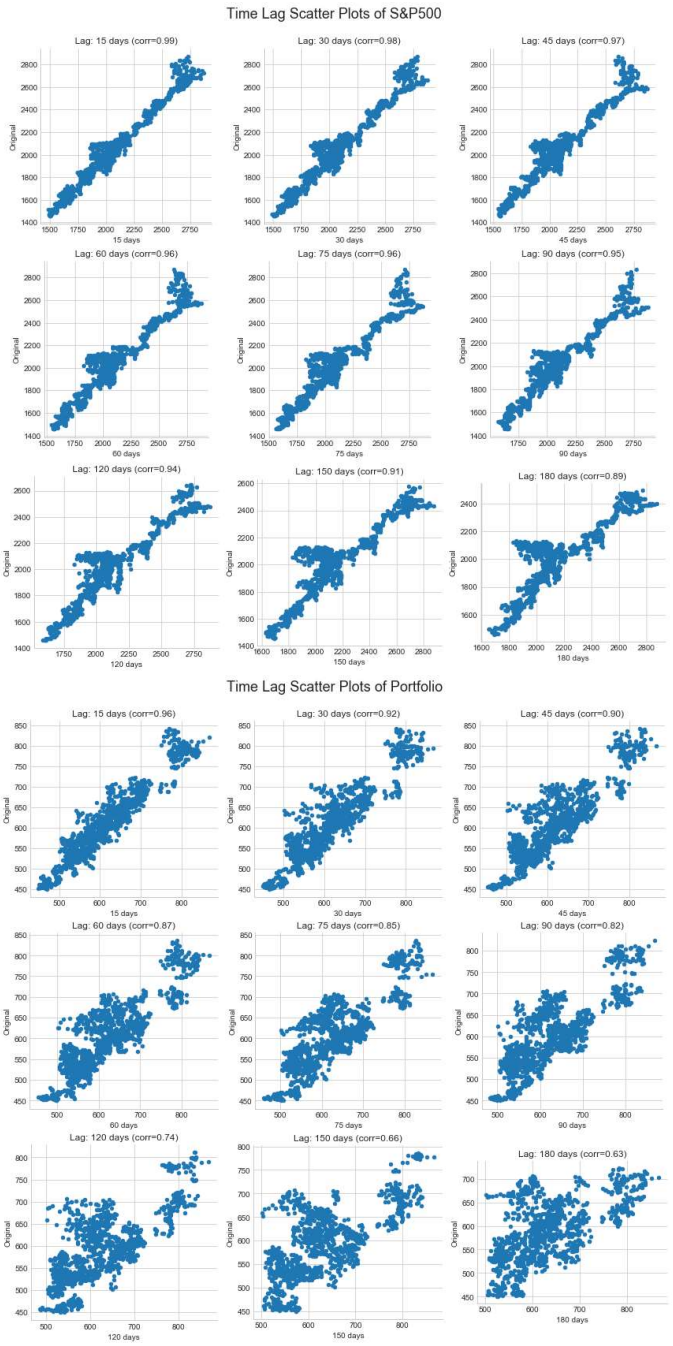
As you can see from the normalized stock prices plot, it is evident that the portfolio underperforms against the S&P 500 index. Although there are peaks where the the portfolio delivers higher returns, in the overall sense, the S&P 500 line is above the portfolio line. Also, the S&P 500 has less volatility as there is a steady growth without erratic vertical movements. On the other hand, the portfolio has more volatility with noticeable vertical movements of the

line. Intuitively, the S&P 500 contains 500 large capitalization stocks, which would buffer against large movements, while the portfolio of 20 stocks is more prone to large movements due to any effect on a couple stocks would make a major effect on the line. It can be argued that with timing the portfolio can outperform the index, but as a whole, it is evident that the S&P 500 index outperforms the portfolio of 20 random stocks.

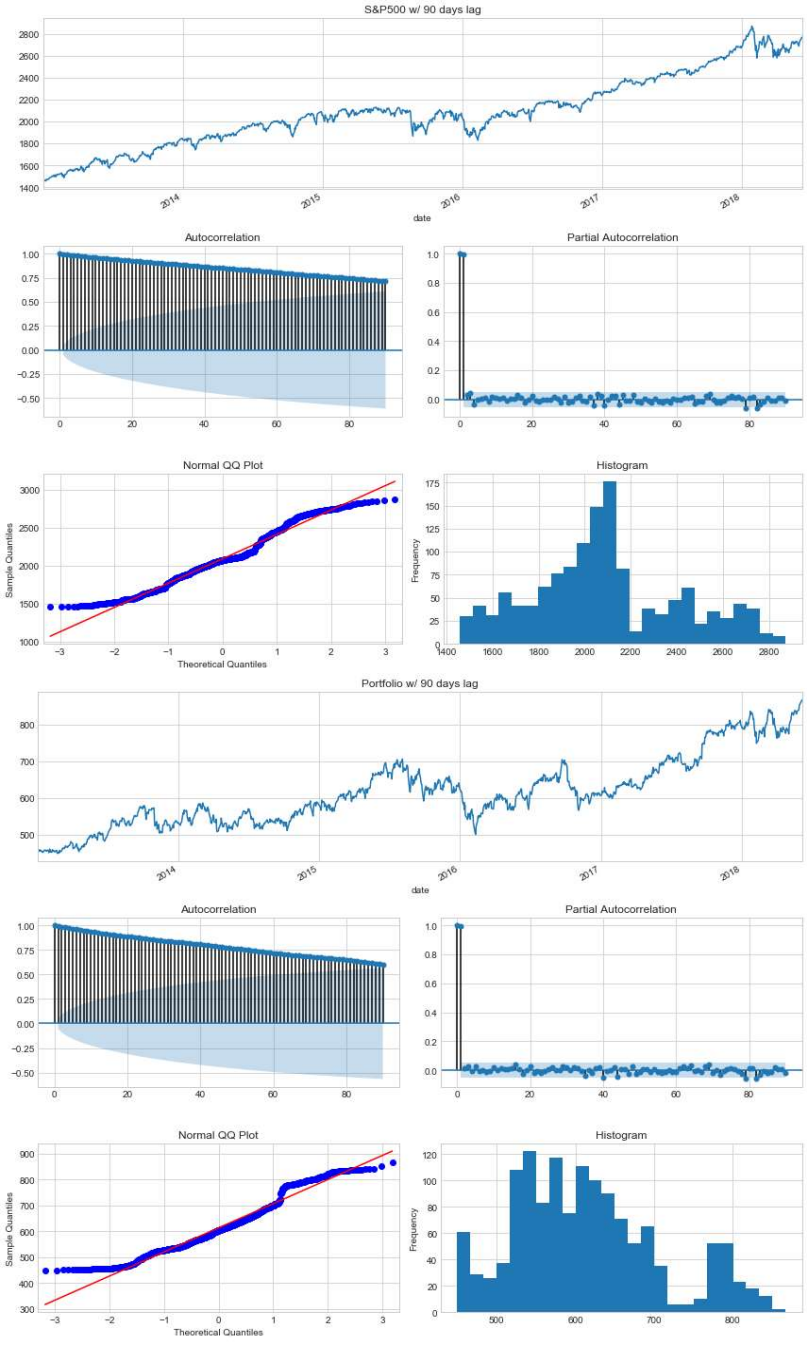
Time Series Analysis

Autocorrelation

The time lag scatter plots on the right reveal the differences in the time series of the S&P 500 and the stocks portfolio. From the top left to the bottom right, the time lags specified for each time series are 15, 30, 45, 60, 75, 90, 120, 150, and 180 days. Judging from the visuals of the scatter plots, it is evident that there is a stronger correlation with the S&P 500 compared to the portfolio. Also the correlation holds stronger over longer time lags compared to the portfolio, which quickly degrades with more time lags. This again highlights the higher volatility and variance of the portfolio compared to the index. The overall linearity for both data

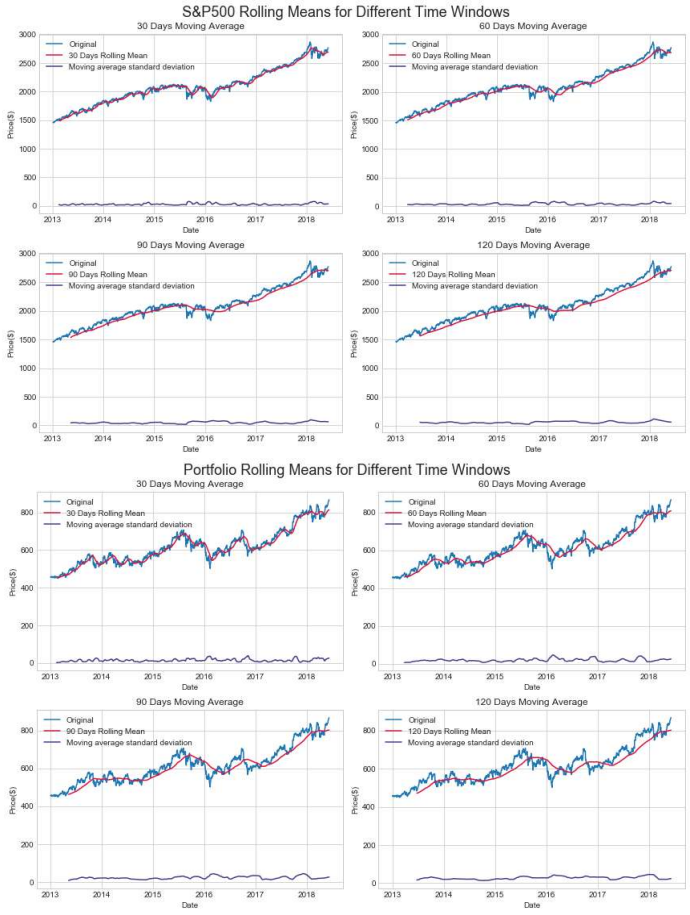


presents strong evidence for autocorrelation in the time series data.



The graphs on the left reveal the S&P 500 and Portfolio data with 90 days lag. The autocorrelation holds strong over gradually decreasing, which was to be expected from the high linear correlation. The partial autocorrelation reveals only one lag value of significance outside the confidence interval, indicating that the autocorrelation is a cascading effect of only the first value after. The normal QQ plot and histogram shows that the S&P 500 is not a normal distribution. All these features are those of a random walk process, which is non-stationary. The same can be seen in the portfolio visual graphs as well with similar effects.

Trend

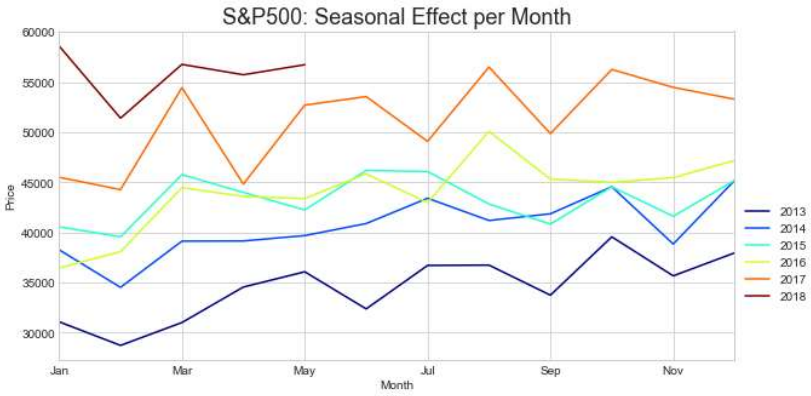


In order to visualize the trend, I applied a rolling mean for different time windows of 30, 60, 90, and 120 days. With increasing day count for the time windows, the rolling mean line is smoother and shifts right against the original line. The higher day count does a better job of giving a general trend line compared to the lower ones, which still have some noise to the lines translated from the original.

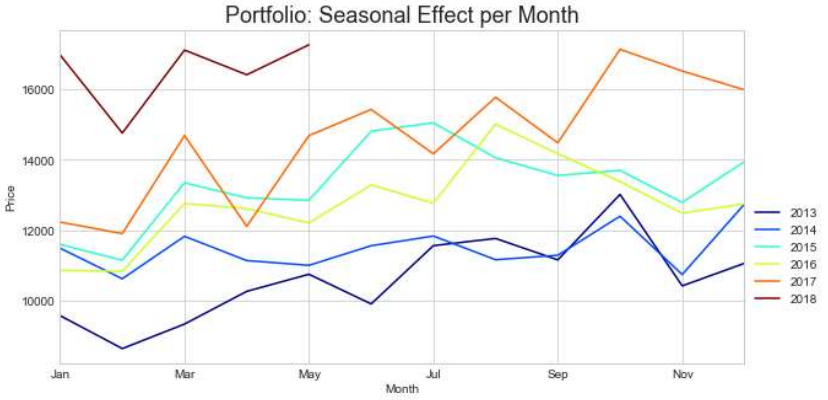
Judging from the trend line, the S&P 500 is a smooth increase with minimal dips while the Portfolio has more erratic movement even visible in the 120 days moving average plot with dips during its gradual ascent. This further adds proof to its volatility compared to the index. Both data is not trend stationary. The standard deviation is difficult to determine from the graphs, but the value seems to be increasing in the long run.

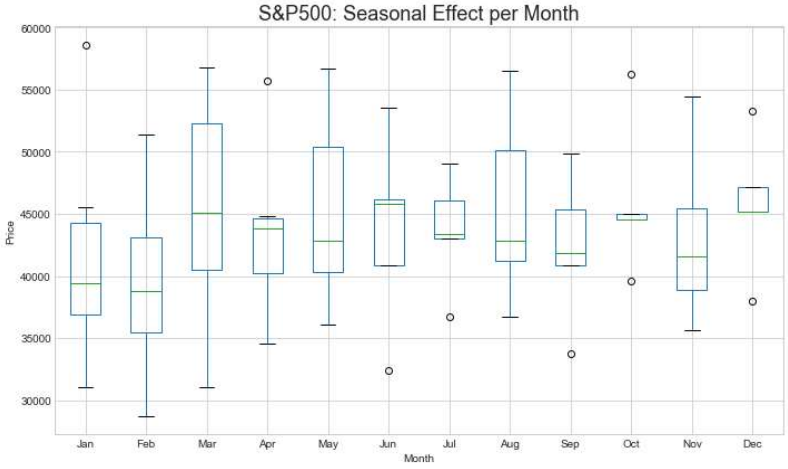
Seasonality

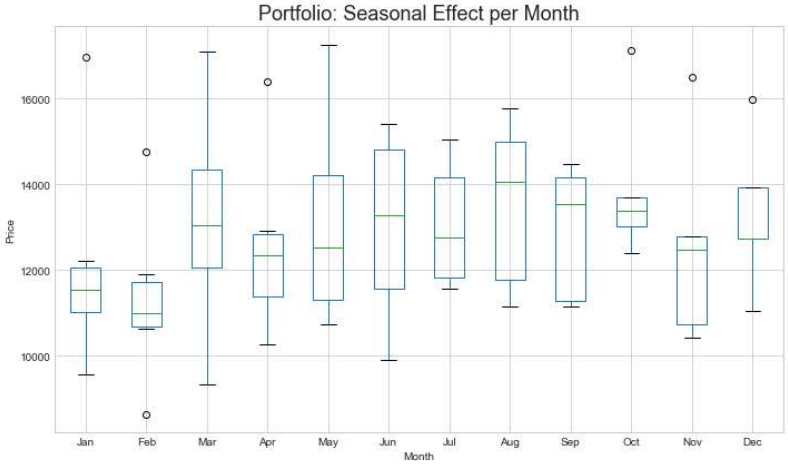
The lines in the plots represent the different years for the S&P 500 data from 2013 to 2018 and reveal seasonal



effects for the months of each year. Intuitively, it would make sense that there are seasonal trends aligning with the earning seasons of publicly traded companies when they release their quarterly earning reports. In general, each earnings season begins one or two weeks after the last month of each quarter (December, March, June, and September). Judging from the line plots for both the S&P 500 and the Portfolio, there seems to be peaks during these time quarter months, but there seems to be overlap in some months. The box plots combine the months to give a better visual representation of the quarter months' seasonal effects with them having higher range of values compared to the others. Looking at the S&P 500 boxplots, it is more evident that there is a pattern of high movement in the months before the quarterly earning reports. This is not as apparent in the Portfolio boxplots, which could be

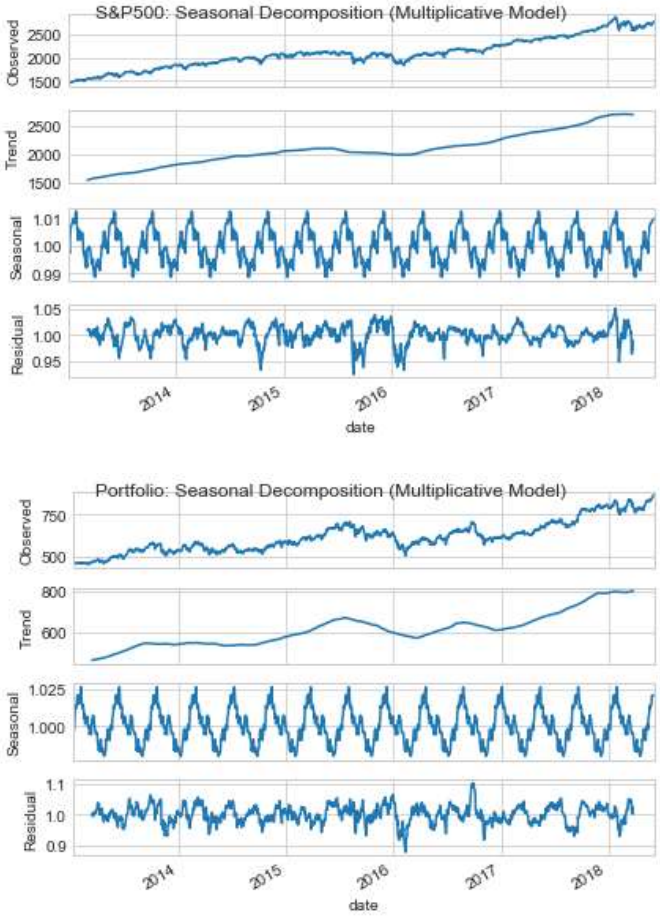






explained by the fact that it encompasses low volume and lower price stocks adding to more volatility.

Seasonal Decomposition and Noise



The seasonal decomposition strips away the trend and the seasonal effect to reveal the residual component. The additive model is most appropriate if the magnitude of the seasonal fluctuations or the variation around the trend-cycle does not vary with the level of the time series. When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series, then a multiplicative model is more appropriate. With economic time series, multiplicative models are more common. So, for both data sets, the multiplicative models are applied.

With the seasonal period being set to 90 days to match the quarterly months, the residuals are presented. However, this seasonal decomposition tool is best for a basic overview of the residual and not for predictions, which would be done with more complex models such as ARIMA.

Augmented Dickey-Fuller Test

The previous sections all conclude that the time series for S&P 500 and Portfolio are non-stationary, but for modeling with ARIMA, stationarity needs to be assumed. The Augmented Dickey-Fuller Test presents a form of hypothesis testing. The null hypothesis is that an unit root

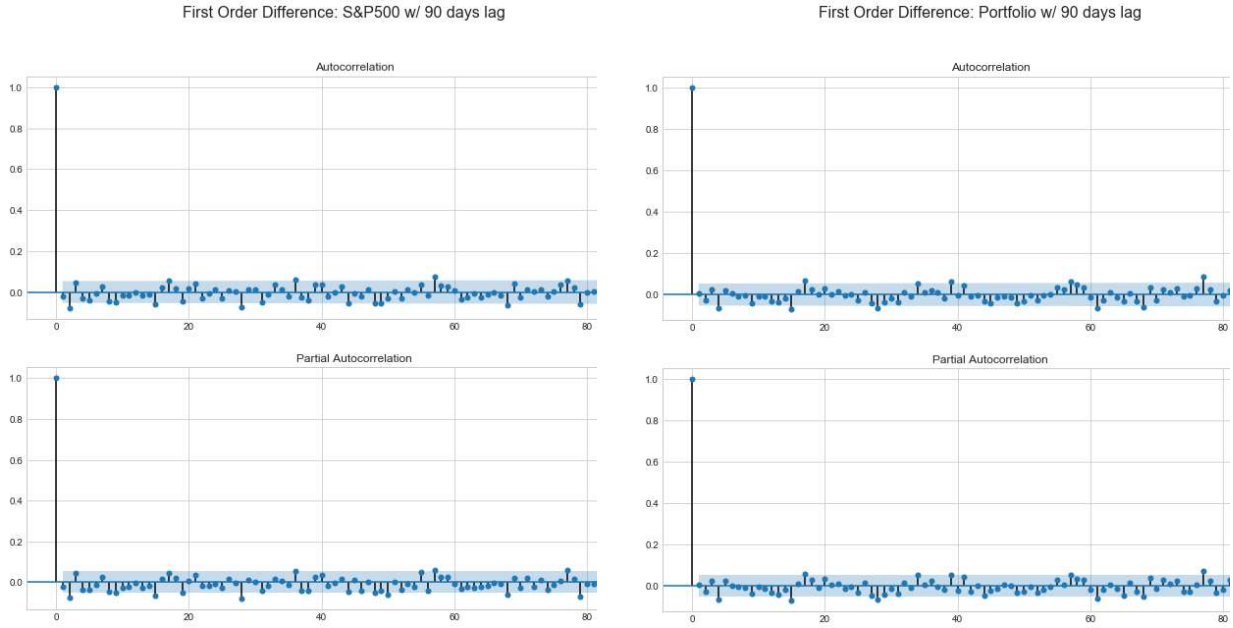
is present in the time series and thus the time series is non-stationary. The alternative hypothesis is that the time series is stationary. For the alpha value or significance level, I selected 0.05. The results of the ADF test for the S&P 500 and the Portfolio reveal a p-value of 0.8437 and 0.8403, respectively, which is way above the significance level. So, the original time series are not stationary.

In order to get stationarity for the time series, I took the differences of the time series values. Looking at the first order differences plots, the graphs more resemble white noise, which is stationary. For further proof of stationarity, the ADF test results' p-values for the S&P 500 and the Portfolio time series are both 0.000. For the test statistics, both values of -21.3979(S&P 500) and -19.4453 (Portfolio) are way below the 1% critical value of -3.4352. So the null hypothesis can be rejected and the first order time series for both the S&P 500 and the Portfolio are stationary.

ARIMA

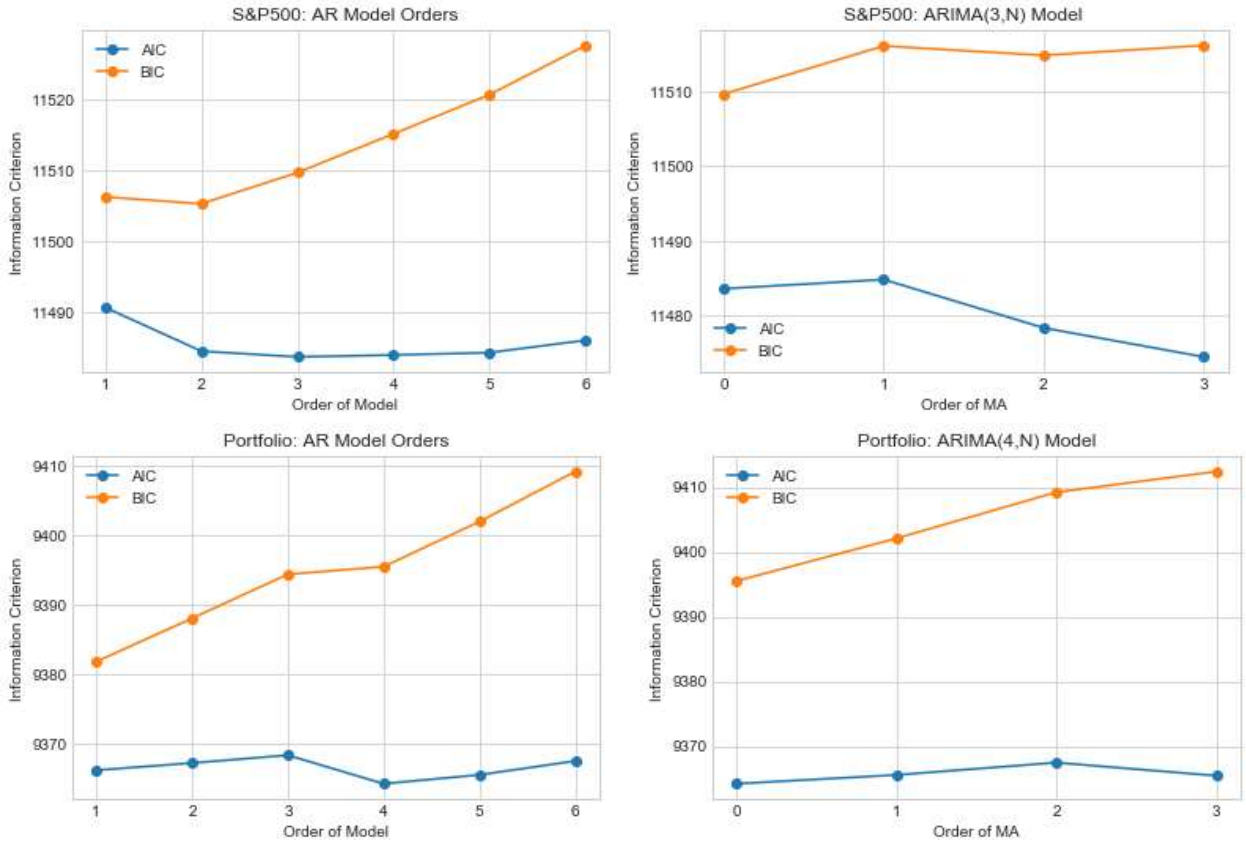
The ARIMA model takes in (p,d,q): p for number of AR (Auto-Regressive) terms, d for number of differences, and q for number of MA (Moving Average) terms. The index and portfolio data requires (at least) one order of non-seasonal differencing to be stationarized, so d value

will be 1 for an ARIMA model. For an ARMA model, using the data with first order difference will function the same as the ARIMA model with d=1.

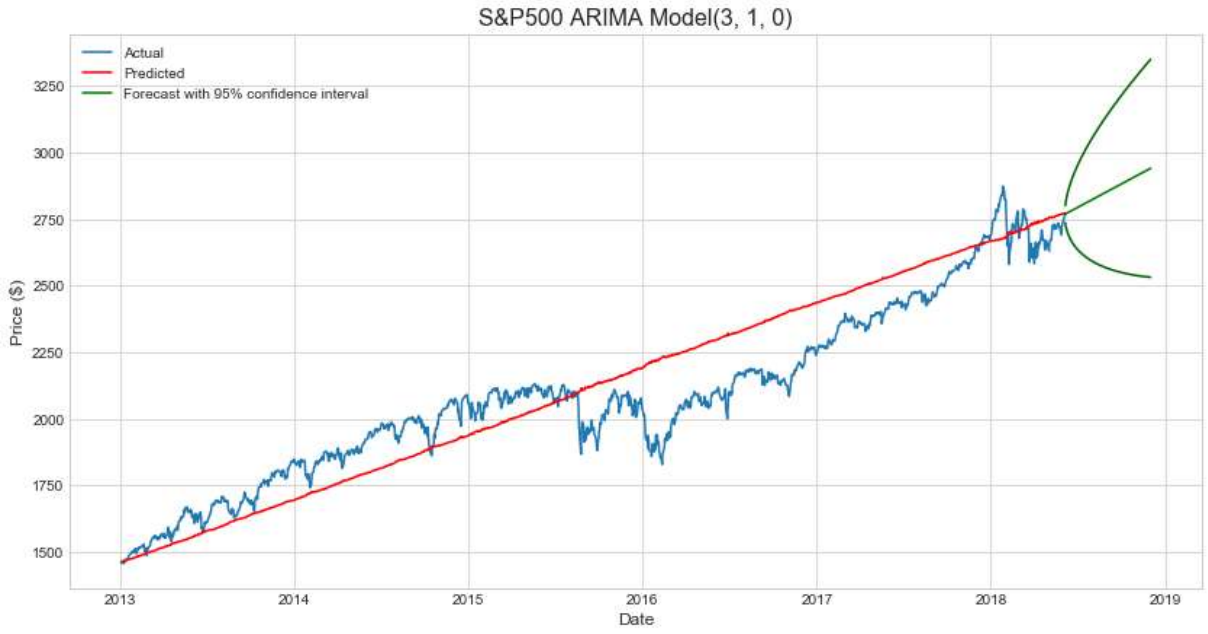


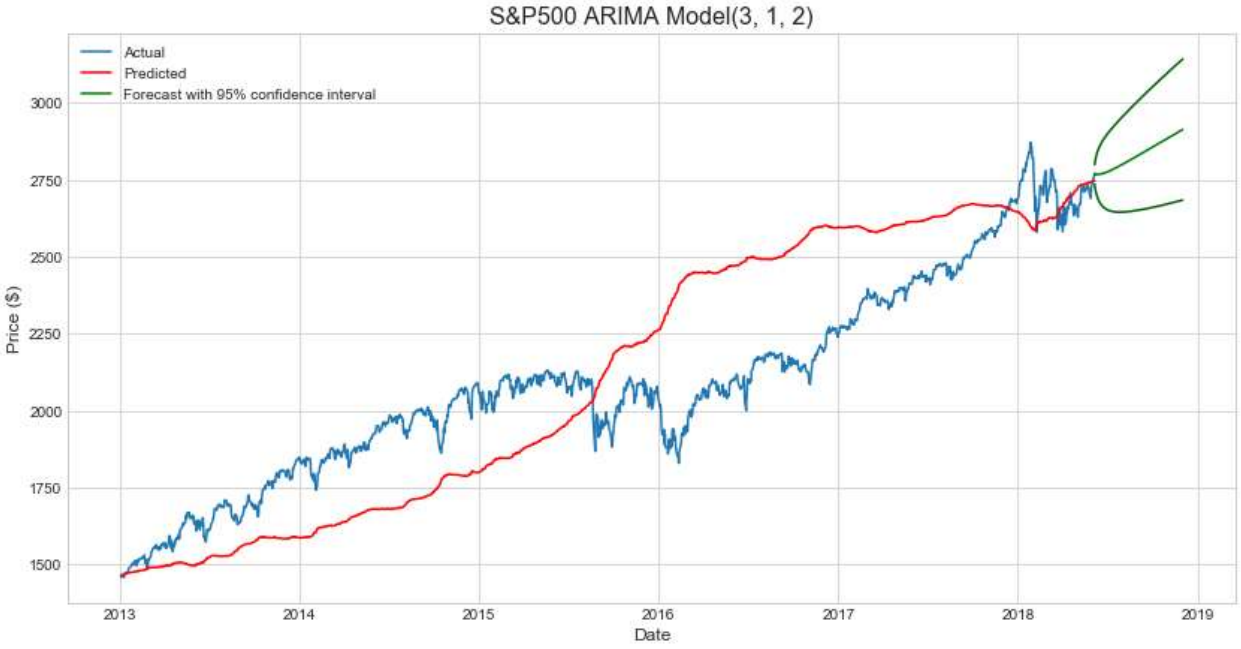
The ACF and PACF graphs don't reveal a clear pattern with no first or second lags peaking beyond the confidence interval of alpha 0.05. As such, I will proceed to test a range of order values measuring the AIC and BIC scores, in which the lower score represent a better model, to see what order values to select.

Though AIC and BIC are both Maximum Likelihood estimate driven and penalize free parameters to combat overfitting, they do so in different manners. AIC accounts for the model degrees of freedom while BIC accounts for the number of observations and model degrees of freedom - according to their equations. Also, AIC is aimed at finding the best approximating model to the unknown data generating process and fails to converge in probability to the true model, if one exists in the group evaluated. On the other hand, BIC does converge as the number of observations tends to infinity.

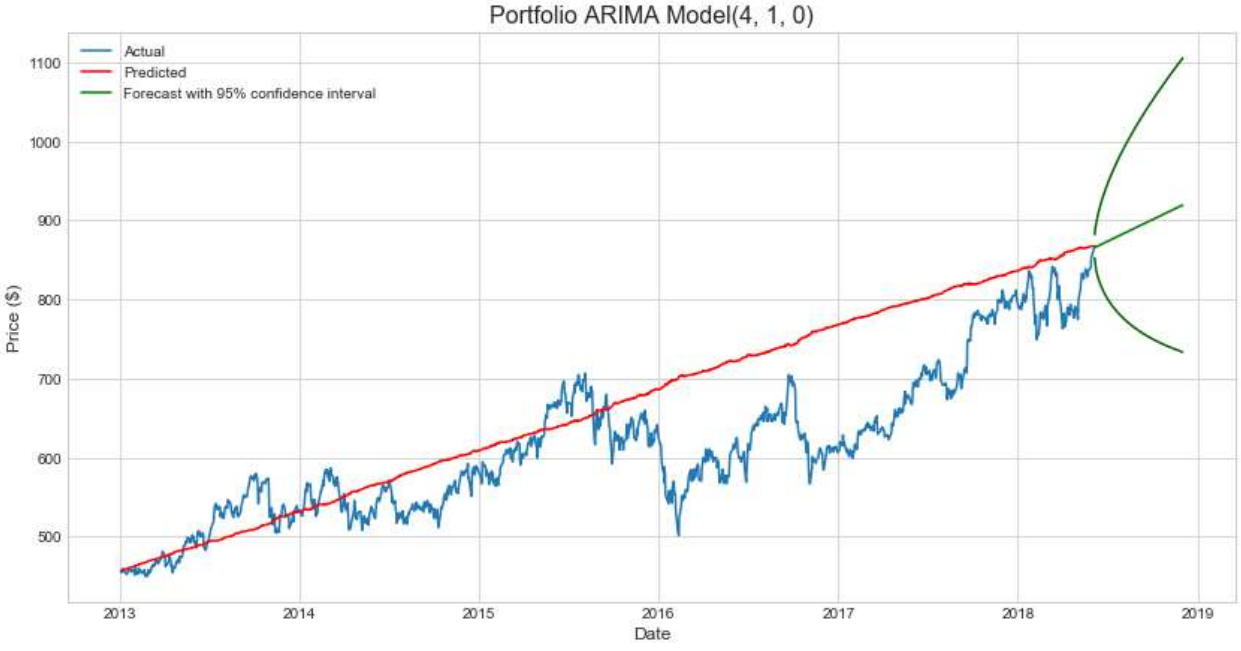


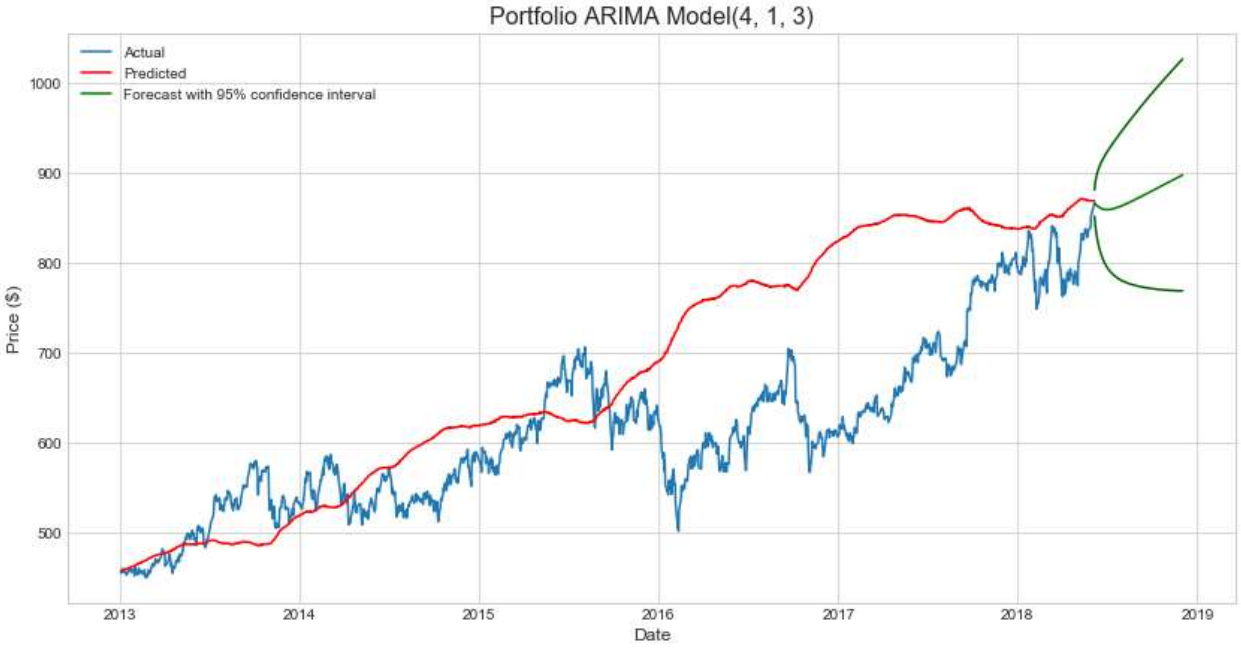
Looking at the visual graphs displaying different order values, I took into account both information criterion in selecting the order values for AR and MA proceeding forward. For the S&P 500 data, the AR value is best at 3. For the Portfolio, the AR value of 4 is best. The different MA values will be tested to see how they affect the model performance.





The S&P 500 ARIMA model with 3 orders for AR, 1 Difference, and no MA orders reveal a linear predicted line with the forecast and 95% confidence range in green. The prediction quality is tested from January 2018 to June 2018 with a 2836.67 mean squared error. The ARIMA model with 2 MA orders on top reveal a more curved line that seems to deviate away from the actual line. However, the prediction quality test reveals a 2547.11 mean squared error, which is lower than the prediction quality of the ARIMA model with no MA order values. Similar observations can be seen in the ARIMA model for the Portfolio data.





In conclusion for both the S&P 500 and Portfolio ARIMA modeling, the pure AR models with first order difference delivered an almost linear prediction. While introducing the MA order values to the AR models changed the shape, at a glance, the model does not seem to be better than the linear, pure AR models.

However, upon further evaluation of the prediction quality from January 2018 to June 2018, the MSE values for the ARIMA models are higher than the AR models. So, if the prediction values are taken from a more current timeframe, the ARIMA model outperforms the pure AR models. In addition, the confidence intervals for future forecasts are narrower for the ARIMA models compared to the AR models, further adding to how it may outperform the linear AR models.

Beyond this Project

Although I did do a thorough time series analysis of the S&P500 and a Portfolio of stocks, the ARIMA model presented in this notebook is a basic starter for time series modeling especially in the financial arena where there are high capital returns to be had. First, instead of the datasets used

in this project, one can consider different indexes or stocks to be chosen for a different portfolio. All the methods used from this project would be easily transferable to those.

In addition, there are many alternatives for modeling these time series beside the ARIMA model such as SARIMAX, perhaps accounting for quarterly seasonal effects, or RNN. Further research into NLP use for finance articles, social media feeds, quarterly earnings reports, and more can help in implementing a more robust and effective ensemble modeling method on top of time series modeling.