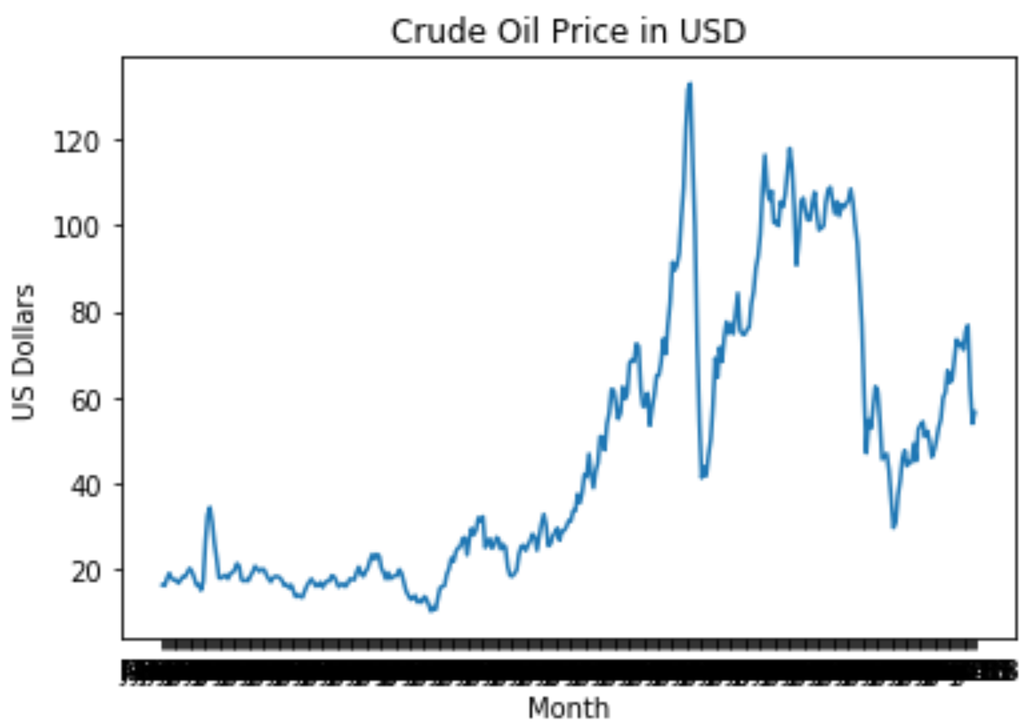
Introduction

In this project, crude oil prices the United States was forecasted. Data was obtained from an index website. The data are monthly from January 1989 to January 2019. The Arima model was used along with the AR and MA models.

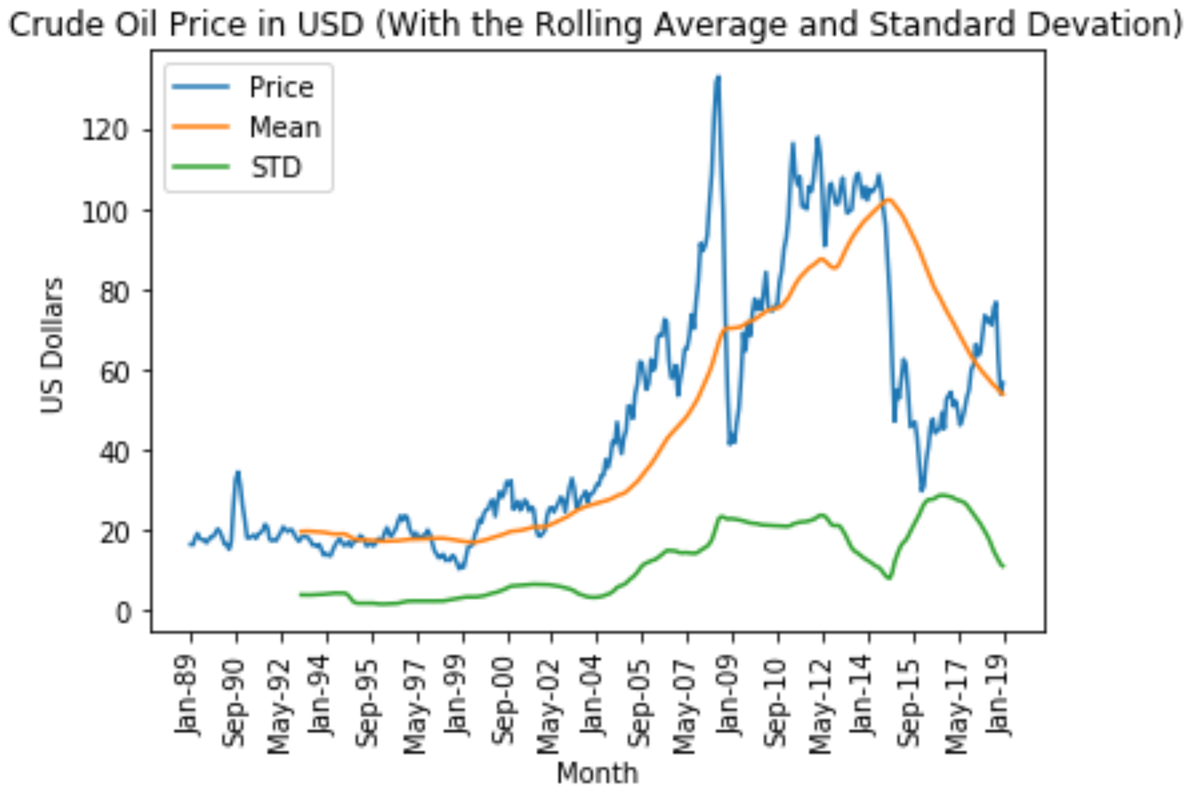
The data were made into a stationary time series and then models were fitted after exploring the auto correlations and effects of the lags of the error terms. The models were evaluated using residual sums of squares. The data were then split into training and testing sets, an ARIMA was built with the optimized parameters, and the average error from the test results was used as a metric to evaluate the model. The average error was 6.3%.

Exploring the Data

As aforementioned, the month data are from January 1989 to January 2019. Below is a plot of the time series data.



It appears that the mean and variance of the data changes significantly over time, with both increasing as time moves forward. To get a better look at this, we can visualize the rolling mean and the rolling standard deviation. Below is the moving average and rolling stand deviation plotted along with the price.



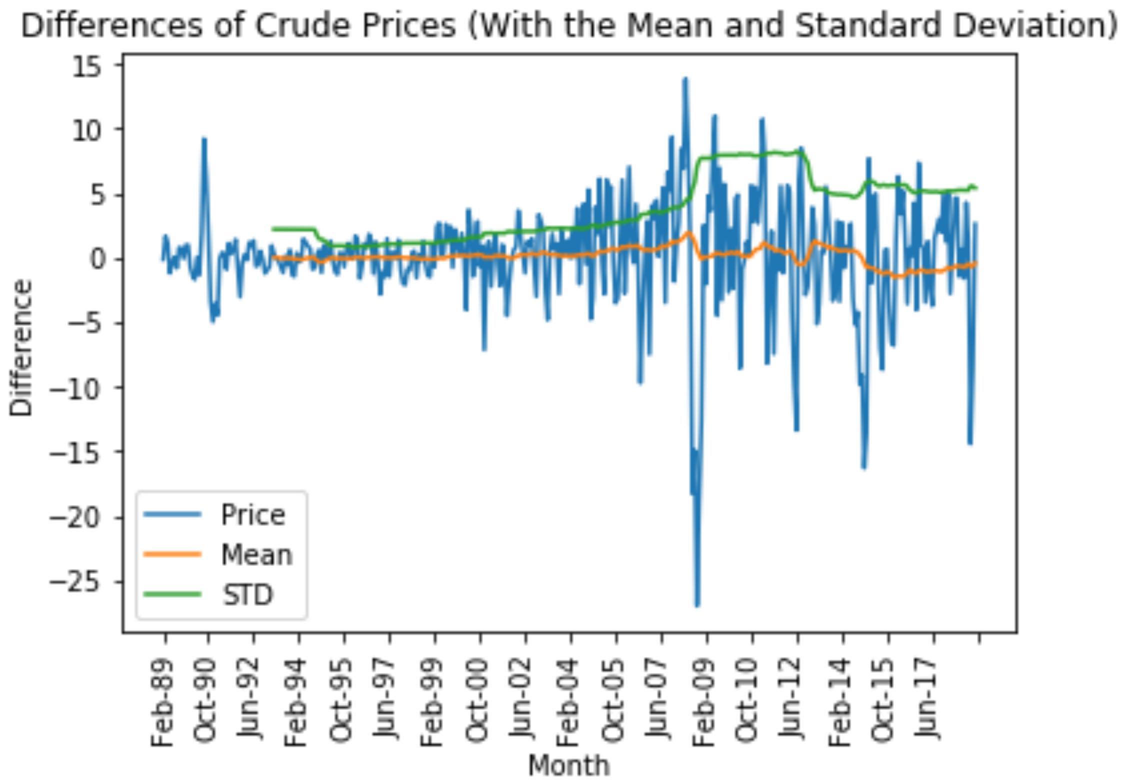
The rolling average and standard deviation do in fact increase with time. The standard deviation does decrease temporarily in later periods but it hovers at noticeably higher levels than is does during the early half of the data. In order to forecast data, it must be stationary, meaning that the mean and variation is not dependent on time. It seems obvious that the data as is is not stationary. But to be certain, we can evoke the AD Fuller test which gives a test statistic to test the null hypothesis that the time series is not stationary and has a unit root.

The AD Fuller technique tests the null hypothesis that a time series data has its coefficients of its lagged terms as one. That means that the lagged terms do not have any forecasting properties concerning subsequent time periods. The price in one period is the price of the previous time period plus an error. The AD Fuller test is used here to examine the state of stationarity in the data.

The P value of the AD Fuller test on the base data is 0.23, which makes it so that the null hypothesis cannot be rejected. Therefore, the data as is cannot be used for forecasting. The data must put into a stationary state in order to make estimated predictions. This is because the models we are using rely on the correlation between the values of lagged terms and the subsequent prices.

Stationarity and Differencing the Data

While exploring the data, it was discovered that the data is not stationary. The mean and variation increase as time moves forward making them time dependent. Also the price in time period t is the price of time period t-1 plus an error. If the coefficient of time period t-1 is not zero, then there would be implications of prediction capability. Making the data stationary would solve both of these problems. In order to do that, the data is differenced and a new time series of the differences is created. Below are the differences along with the rolling mean and stand deviation plotted.

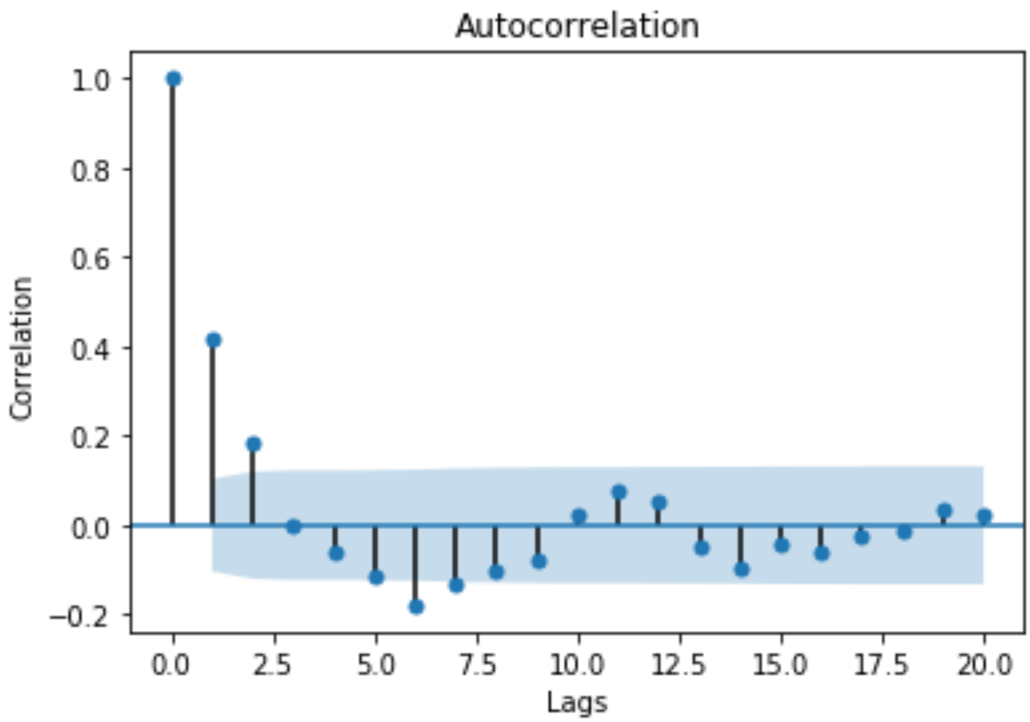


Above it is seen that characteristics of the differenced data differs from the base data. It moves consistently sideways, however, just by eye balling the differences one can notice the increase in variation as time moves forward. The plotted standard deviation illustrates this further with an increase showing between June of 2007 and June of 2012. Though the mean varies where it is obvious (where is the stand deviation increases), it hovers around the same value throughout. We can again turn to AD Fuller to test the null hypothesis that that data is stationary with a unit root to go beyond visual inspection.

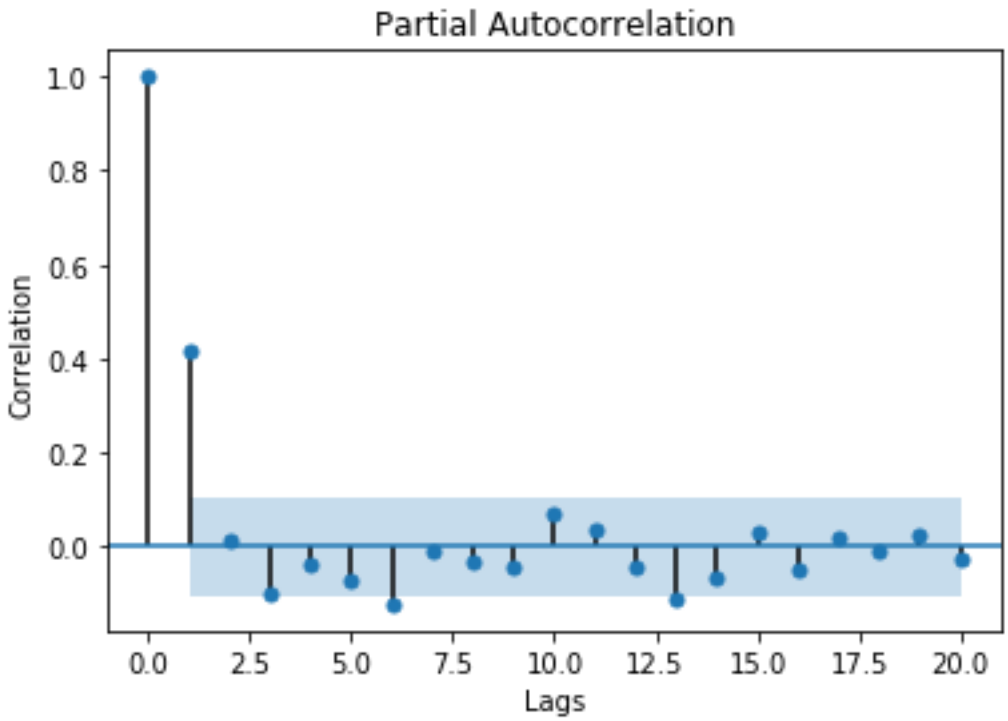
Despite the increase of variation in the differenced data, the P value of the AD Full test statistic is 1.4e-14, which makes is highly unlikely that the differenced data is stationary. This confirms that the differenced time series can be utilized for predictive purposes for some extent.

Autocorrelation and Partial Autocorrelation

After making the data stationary, we can focus on working towards making predictions by looking at the correlations between the price in time period t and the prices in time periods prior, autocorrelation. Below, the autocorrelation of lags one to twenty is plotted.



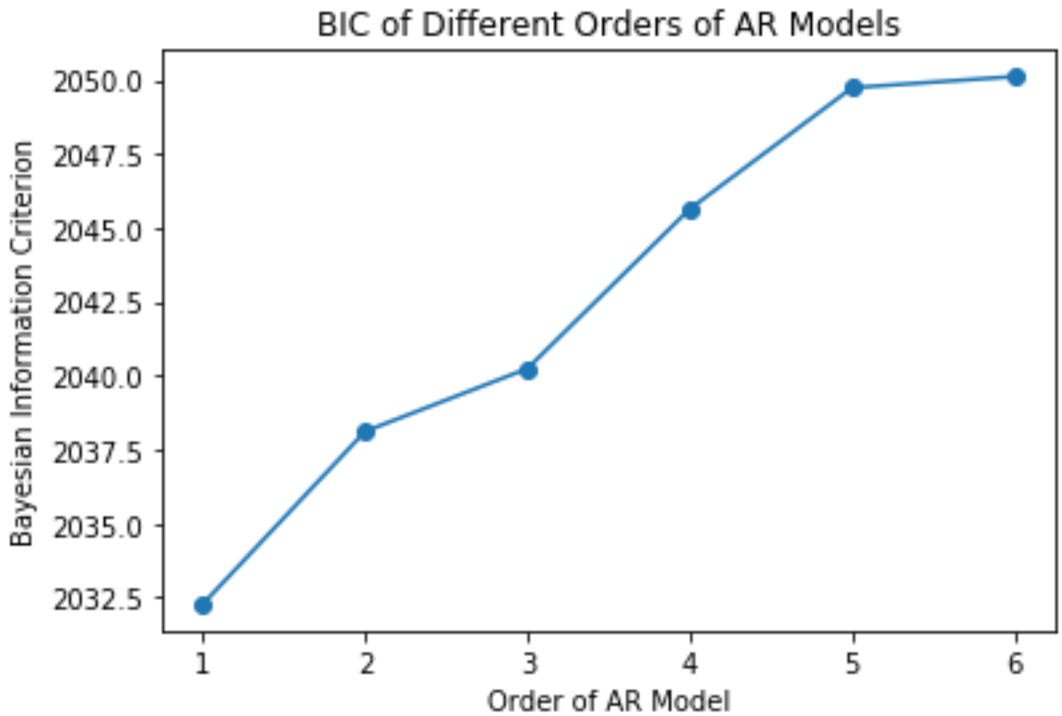
Above we can see that the first and second lags are positively correlated beyond the 95% confidence interval which is illustrated by the blue shade. Assuming the null hypothesis that the correlation of any given lag is actually zero, the probability of a lag having a correlation outside the 95% confident interval is less than 5%. For this reason, any lag with a correlation outside the 95% confident interval is statistically significant, as in the cases of the first and second lags. However, some of the correlation of the second lag is partially due to the correlation of the first lag, in a sort of spill over effect. For this reason, the partial autocorrelation is used to get a truer correlation between each lag and the price difference of time period t.



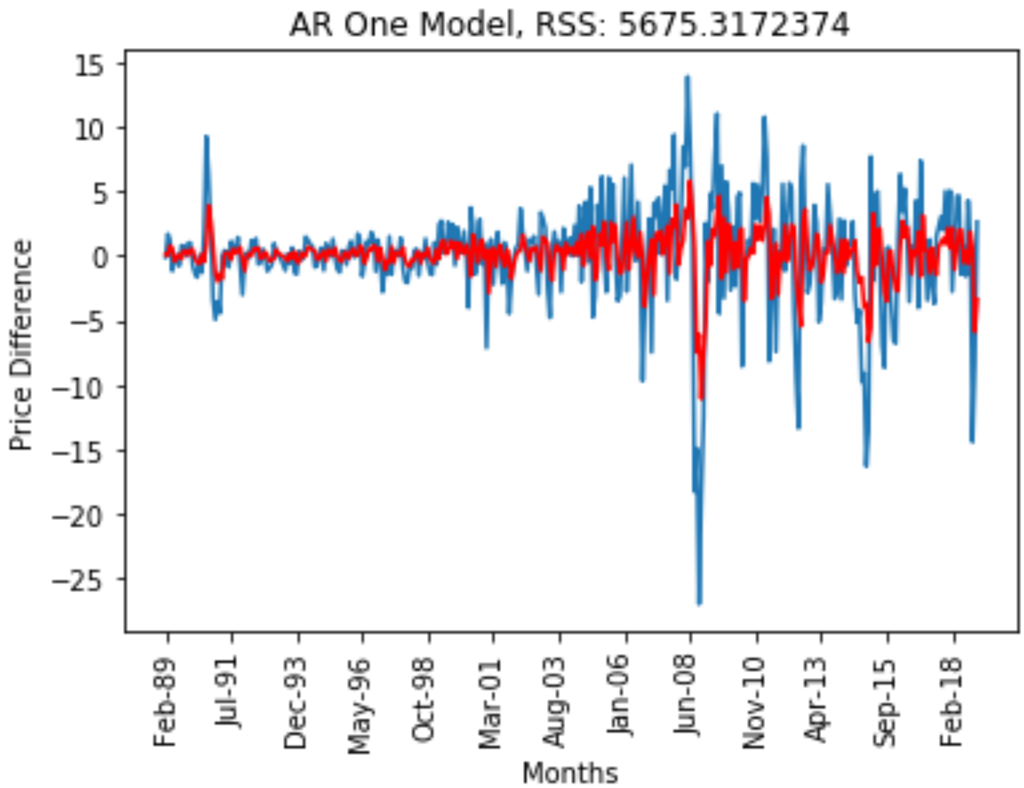
It is seen above that only lag period that has direct correlation with the price difference of time period t that is statistically significant. This indicates that only one lag time period, time period t-1, is useful in making an estimation as to what the price in time period t might be. The first model that is built is an autoregression model, also known as AR model, which uses the price difference and coefficient of time periods prior to the one being predicted. The partial autocorrelation plot shows that the first order AR, the AR of one lag period, should be used. To corroborate the Bayesian Information Criterion is used and the AR model is explored.

Autoregression Model

Before building an autoregression model, the partial correlation and the Bayesian Information Criterion were consulted upon for the order of the AR model. As seen, the partial autocorrelation plot of the differenced data indicates that only the first lag, the price different in time period t-1, and the price different in time period t has a correlation that is statistically significant. The Bayesian Information Criterion of different lags of the price differences can be examined to get a fuller picture in this regard. Below is a plot of the BIC of different lags.



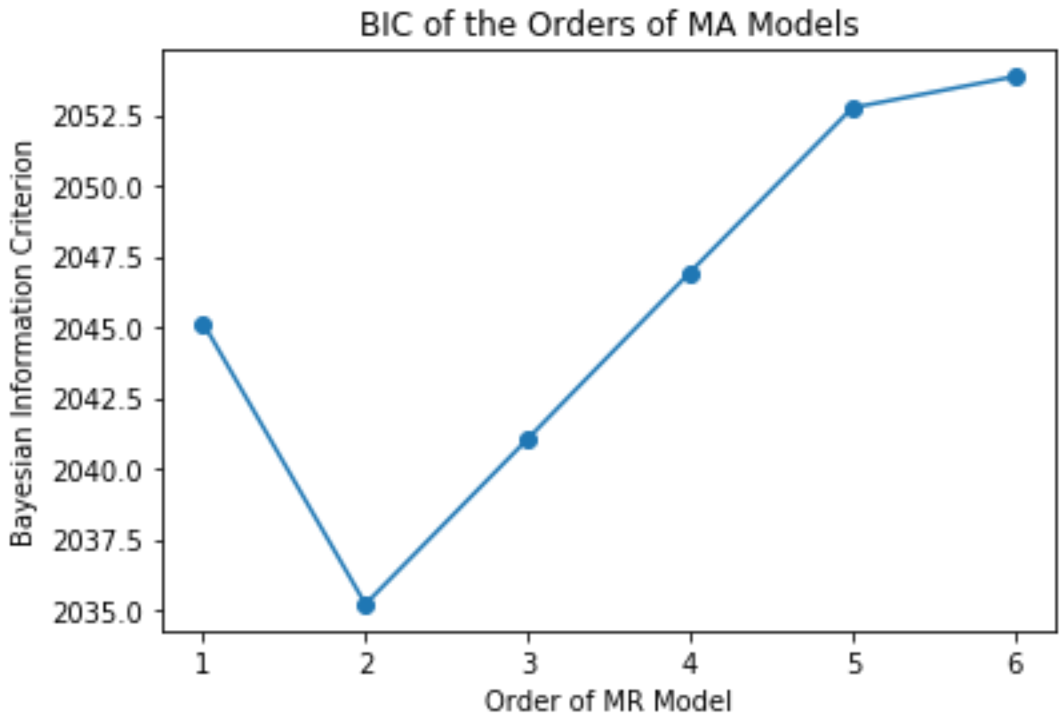
The AR models of order one has the lowest BIC. The lower the BIC, the better. Therefore, according to the BIC, the AR of order one is best, which corroborates the analysis of the partial autocorrelations. This is enough to move forward with the AR of order one.



Above is the plot of the fitted values of the AR one model fitted to the differenced time series data along with the data itself. The residual sum of squares is 5675.32. Which in it of itself is meaningless because the residual sum of squares is used in comparison to other models. A moving average model was then built and compared to the AR model.

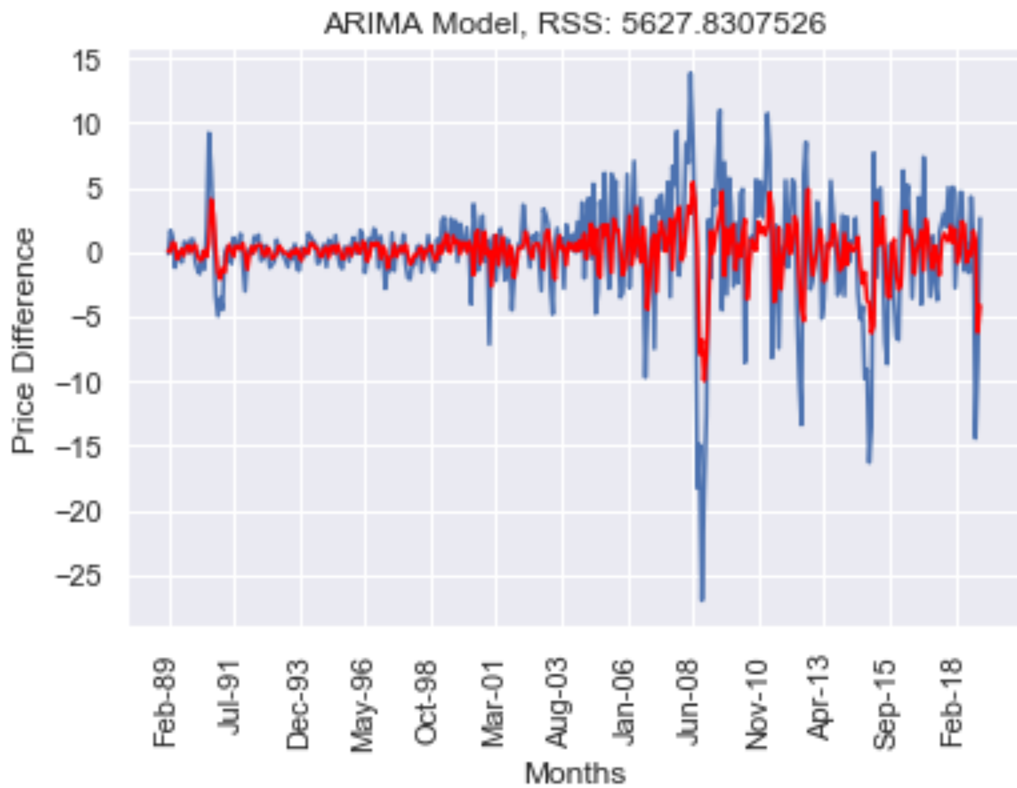
Moving Average Model

The moving average model, also known as the MA model, utilizes the error values of lagged time periods to make estimated predictions. The BIC was again used for diagnosis in regards to the lags of the error terms in relation to making estimation.



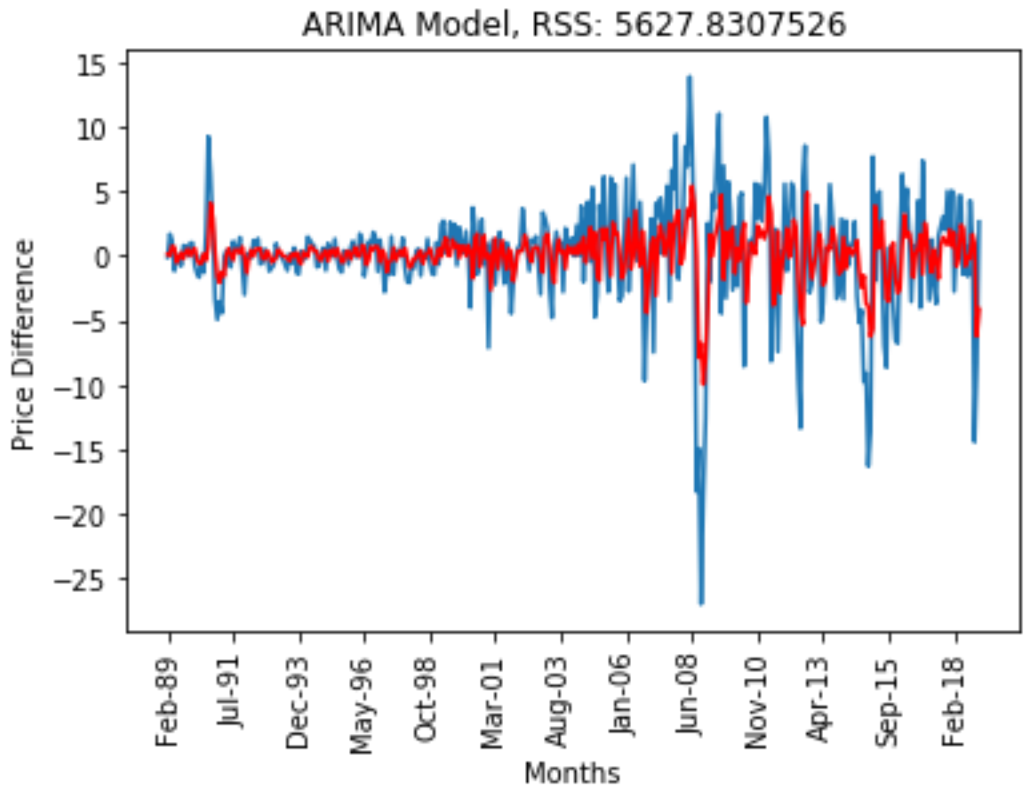
As shown above, the BIC of the different orders MA models is lowest for that order two. Which implies that the second order of the MA models is best. To confirm this, the residual sums of squares of both MA one and MA two models are compared.

The residual sum of squares of the MA model of order one is 5882.77 and that of MA model of order 2 is 5628.54. According to this metric, the MA two model fits the data better. But to check for overfitting, we will test a model with the order of the MA as one and another with the order of the MA as two and compare their average errors on the same test set. Below, the MA models of order two is plotted. It is interesting to note the lower residuals sum of squares of the MA two model in comparison to the AR one model. Next, the AR and MA models are combined and we see if the residuals sum of squares lowers even more.



AR(I)MA Model

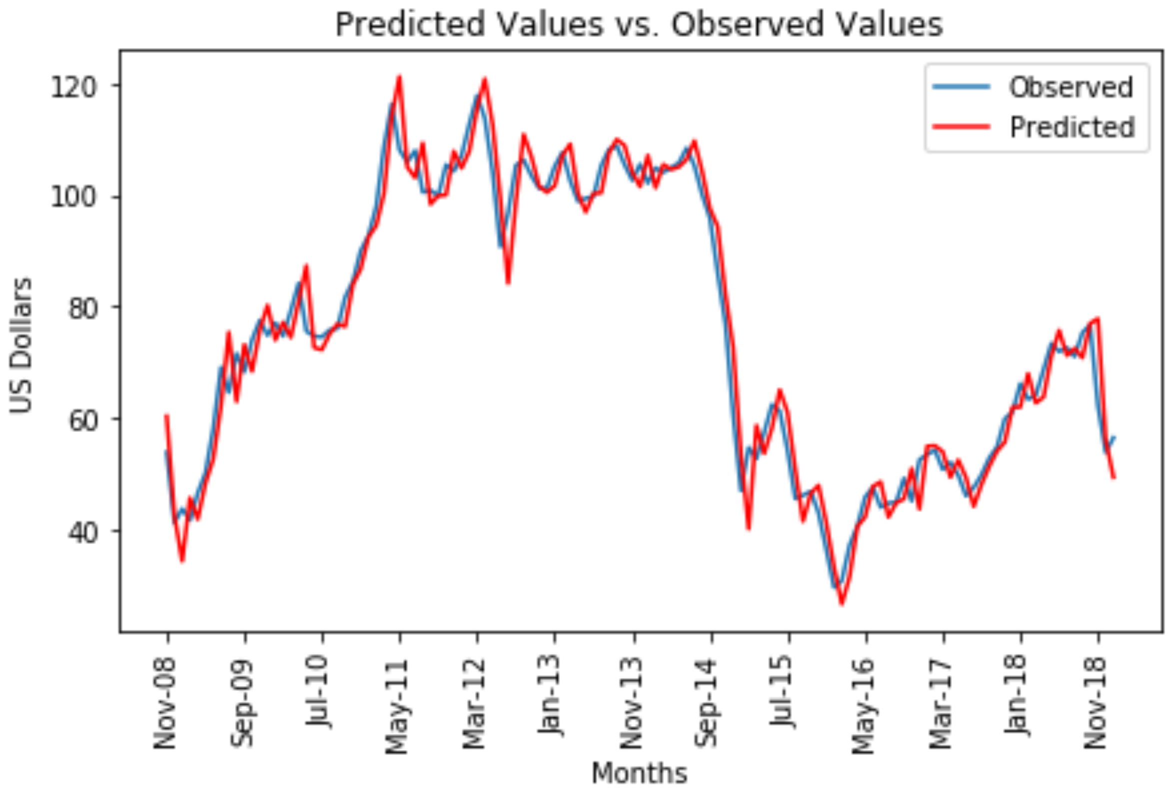
After diagnostics and fitting the AR and MA models using the parameters of best choice, the AR and MA models are combined. Below, the fitted values of the combined models are plotted along with the actual differences. The AR order is one and the MA order is two.



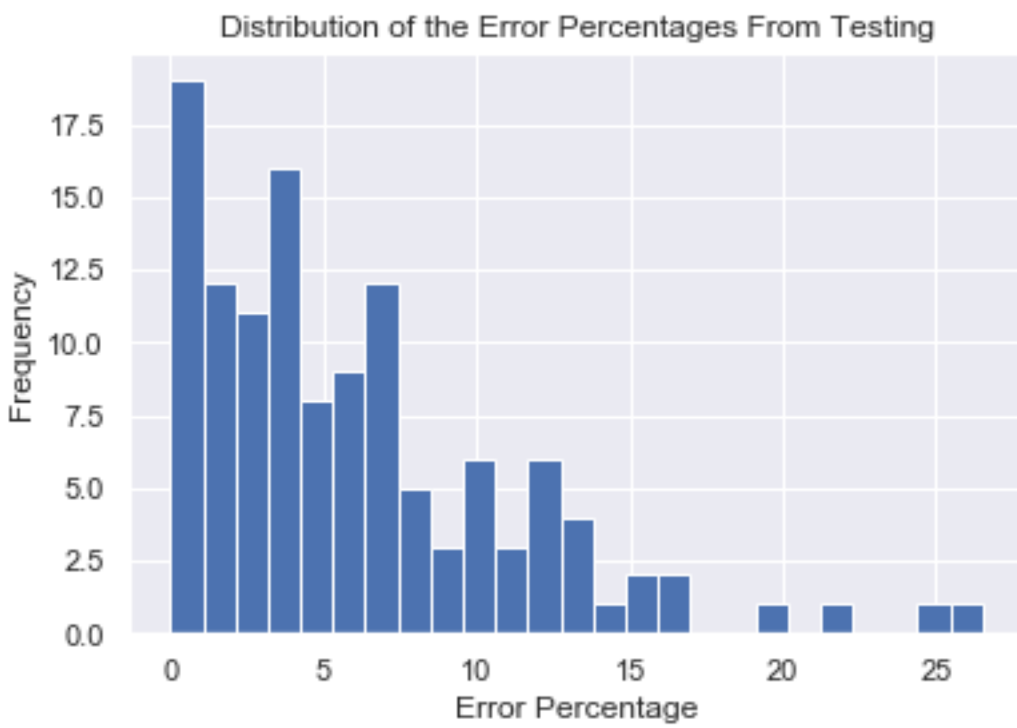
Interestingly, the residuals sum of squares for the AR(I)MA model lower than that of the MA two model by less than 1. It is 5628.54 for the MA two model and 5627.83 for the AR(I)MA model. Given the residuals sum of squares of the models, the AR(I)MA model was used to forecast and test.

Evaluation

The next step is to test the model under the parameters determined as the most optimal. This is done by first splitting the data into a training set and test set (33.3% testing), making a list of predicted values which are undifferenced, plotting the predicted values along with the observed values, calculating the differences between the predicted values and the observes values, the errors, and averaging those errors to obtain the final metric used for evaluation. Below is the plot of the predicted values and observed values.



The average error of the test set is 6.09%, with the max of 28.52%. Below is the distribution of the error percentages, which provides some insight as to model performance.



As seen, the model error percentages are heavily distributed towards the lower end. Overall, the model performed well in regards to evaluation with 6.09% average error.

Conclusion

A model to forecast crude oil prices in the United States was developed and evaluated. In preparation, the monthly time series data of crude oil prices in the United States was modified in order to make into one that could be used for forecasting. The data was made stationary by differencing. This stationarity was confirmed using the AD Fuller test. Autocorrelation was then analyzed as part of diagnosis.

When examining the autocorrelation is was found that the first two lag periods were correlated with the price of lag zero. Upon further investigation, it was found that only the first lag had direct correlation. The second lag was only correlated due to the spill over effect of the first lag. An AR model of order one was then fitted after corroboration from the Bayesian Information Criteria.

The BIC was again used in regards to diagnosis concerning the moving average model. It was determined by the BIC and the residuals sum of squares that the MA two model was optimal. The AR and MA models were then combined, and the combined model was compared to the AR and MA models to see which one would be used for testing and forecasting.

The AR(I)MA model had the lowest residuals sum of squares and was tested. The data was split into training and testing sets. From the test, the model was calculated to have a 6.09% error. The distribution of the error percentages was plotted to get a better understanding in that regard. Overall, the model performed well and could be used to forecast the price of crude oil in the United States one time period ahead of the last month for which the price is obtained with 6.09% average error.