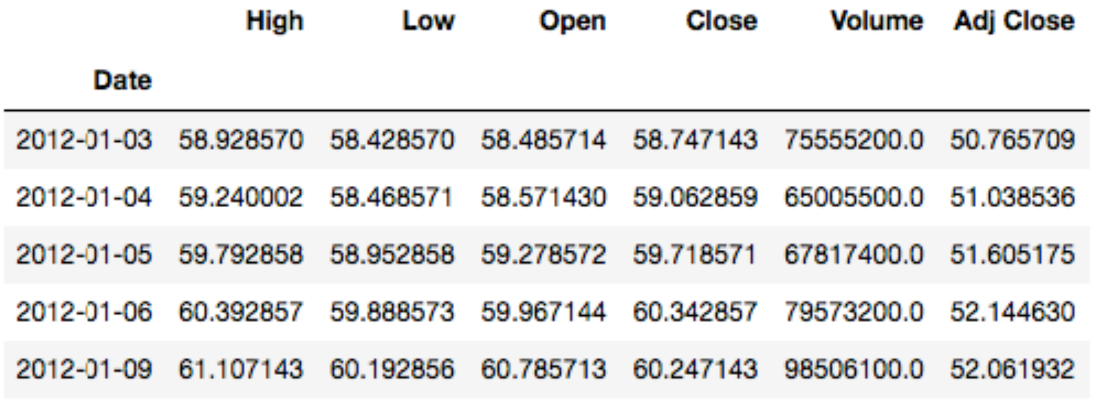
Table of Contents

INTRODUCTION

Analysis of Stock Price:

1. Data Frame

Standard stock price data has five columns the high, low, open, close, adjusted close and date. The high signifies the maximum price for the day while the low indicates the minimum price for the day, the open is the price the stock was at the beginning of the day, the close is the price at the end of the day, the adjusted is the price at the end of the day but factors in anything that might affect the stock price after the market closes.



Date	High	Low	Open	Close	Volume	Adj Close
2012-01-03	58.928570	58.428570	58.485714	58.747143	75555200.0	50.765709
2012-01-04	59.240002	58.468571	58.571430	59.062859	65005500.0	51.038536
2012-01-05	59.792858	58.952858	59.278572	59.718571	67817400.0	51.605175
2012-01-06	60.392857	59.888573	59.967144	60.342857	79573200.0	52.144630
2012-01-09	61.107143	60.192856	60.785713	60.247143	98506100.0	52.061932

figure1: sample apple stock head

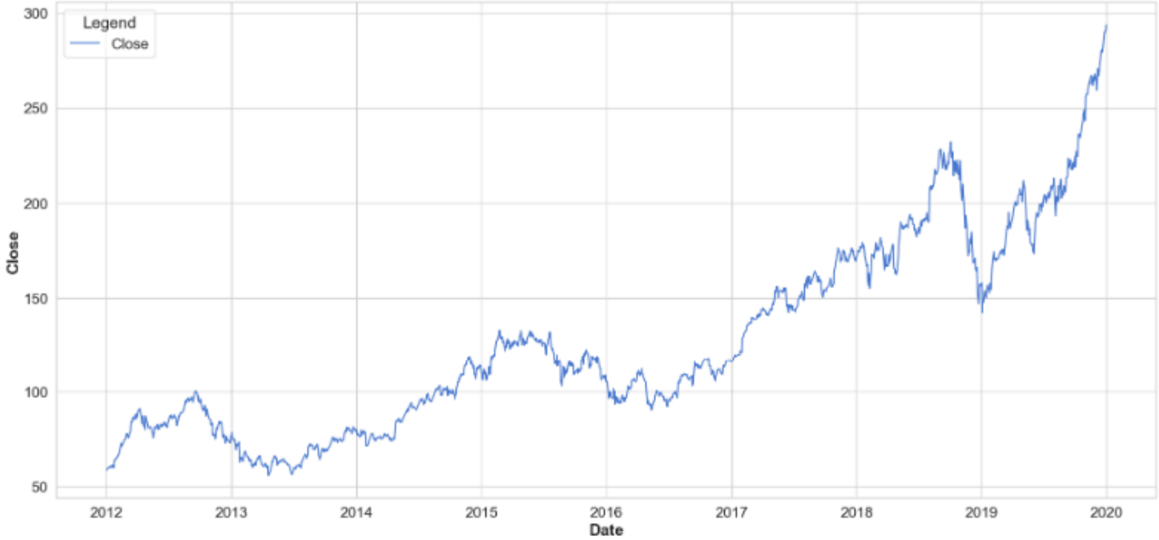
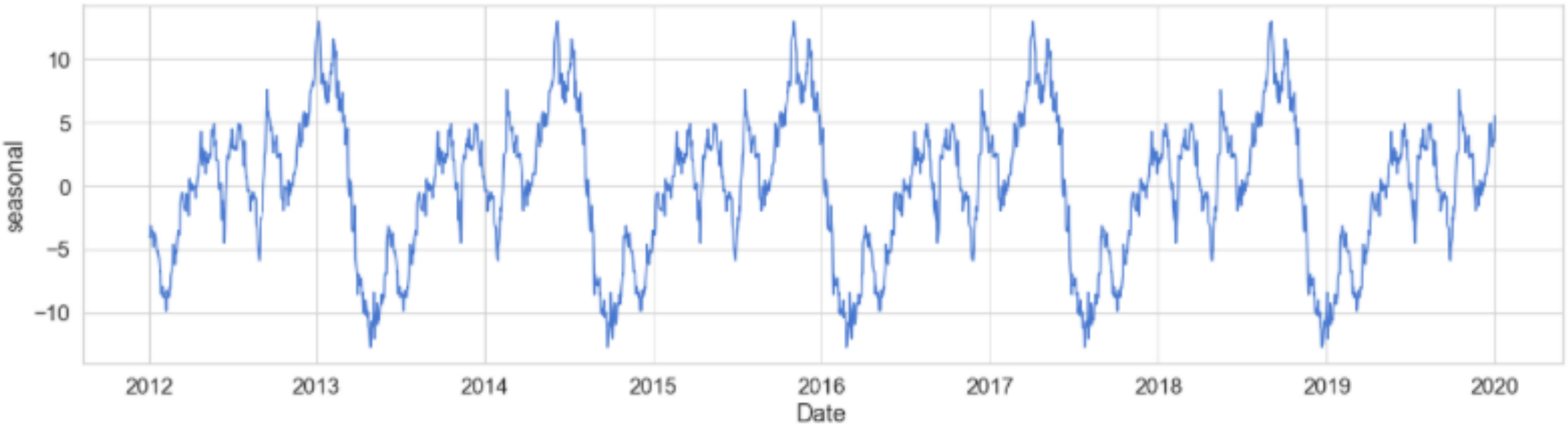


Figure2: close price 2012 - 2019

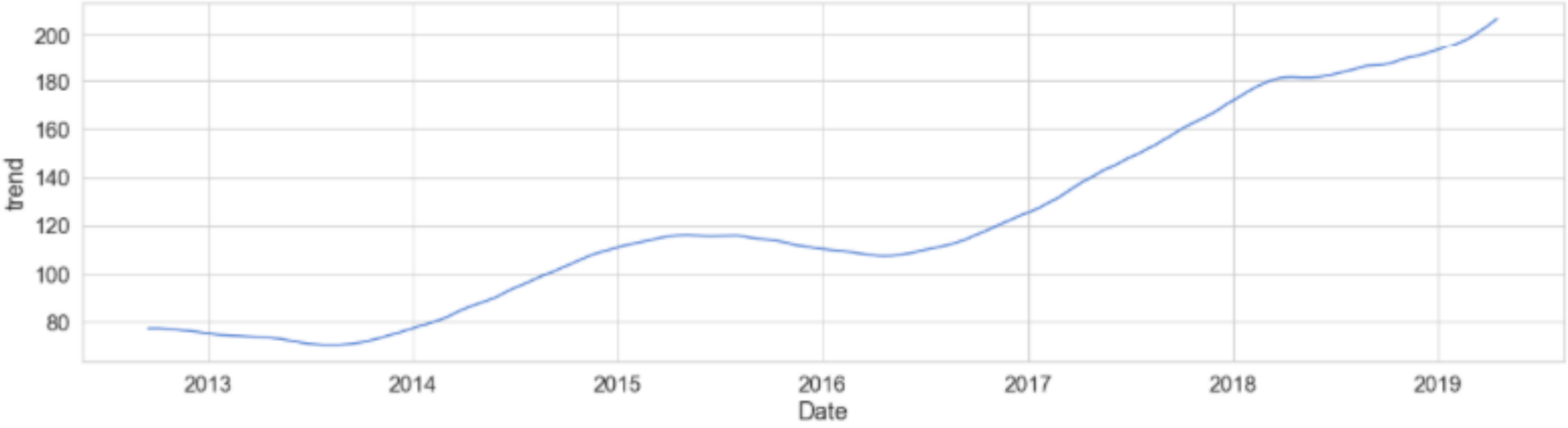
2. Decomposition of Time Series

A series needs to be decomposed in order to ensure that there are some patterns and that the series is not majorly white noise which cannot be used for or casting.

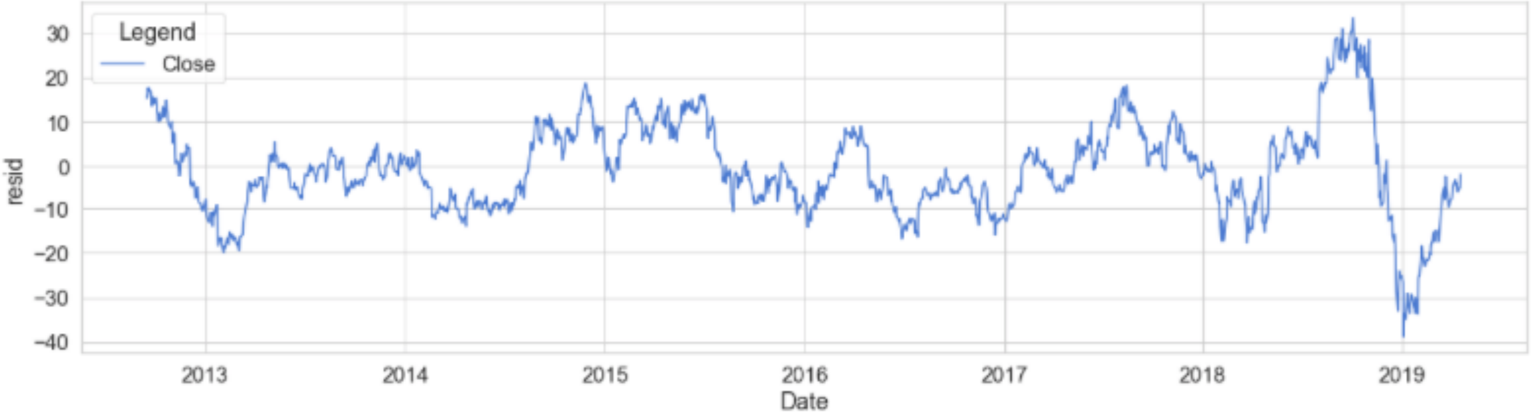
a. Seasonal Chart from this chart we can see a cyclic pattern that occurs but over the a period of a year.



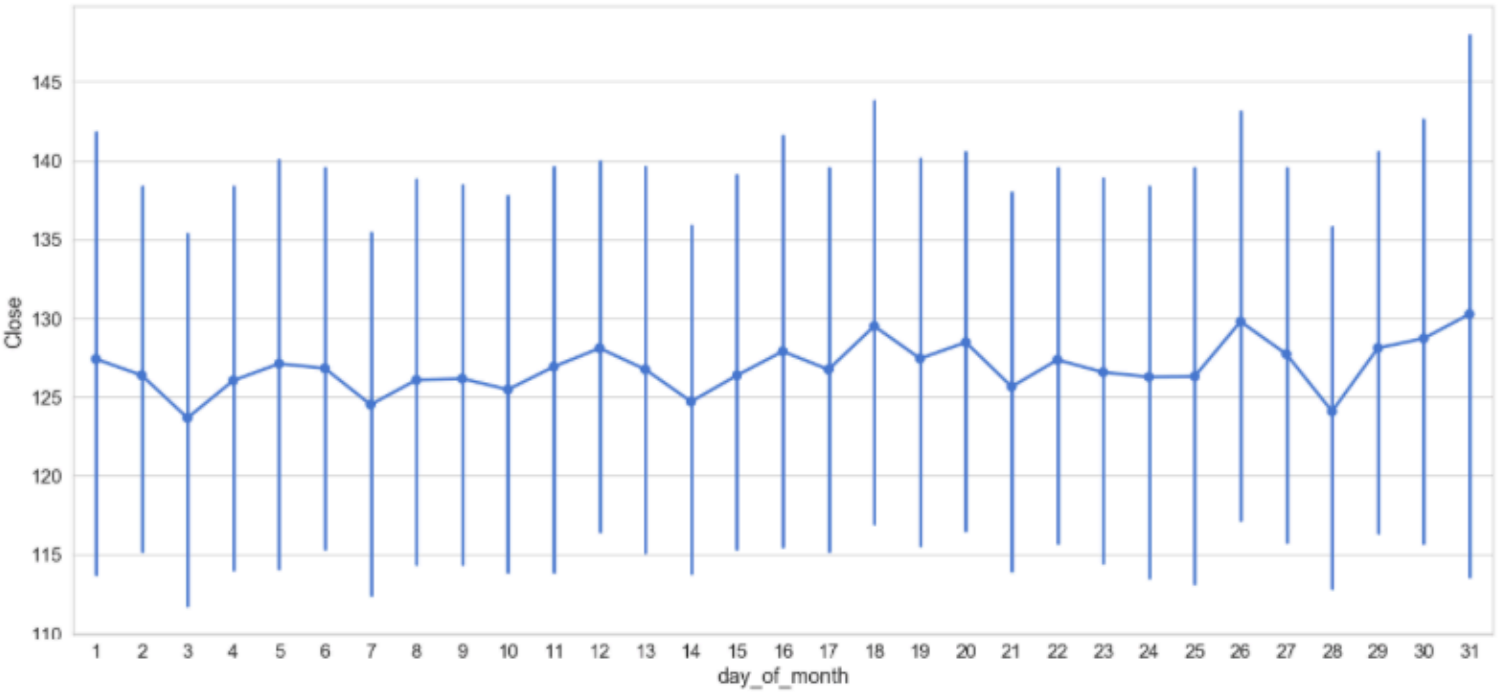
b. Trend: shows the overall direction of the apple stock price over the course of years



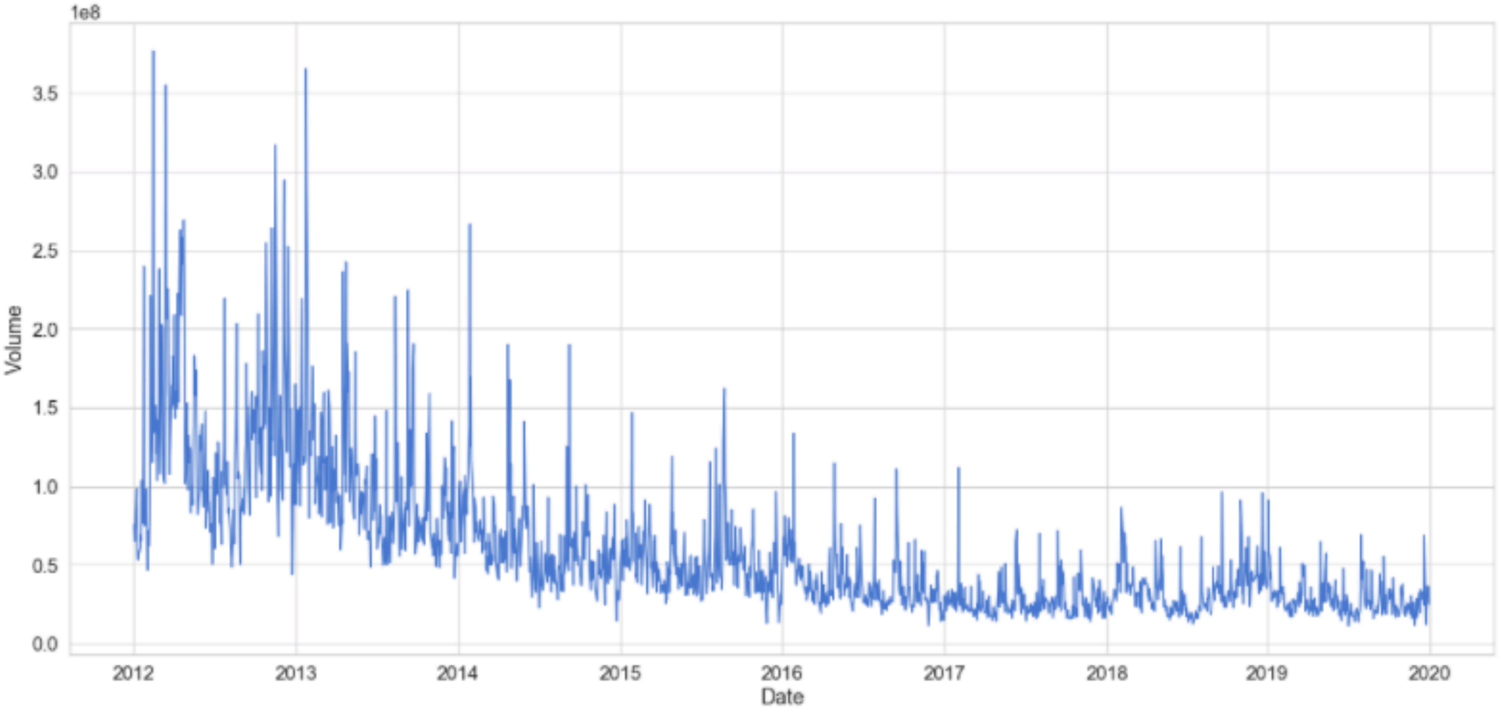
c. Residuals this is the unexplained noise that is found in the data, over those number of years after you remove the effect of the trend and the



3. Check for Patterns over smaller time range
    a. In order to look more closely at data on closer basis we check if we can find any pattern in the time series over the span of a month. From this chart there is no constant pattern that occurs monthly.



    b. This chart tells us high high the buying and selling interests are , here we see volume of shares traded over the years has reduced considerably.



Stock Prediction Models

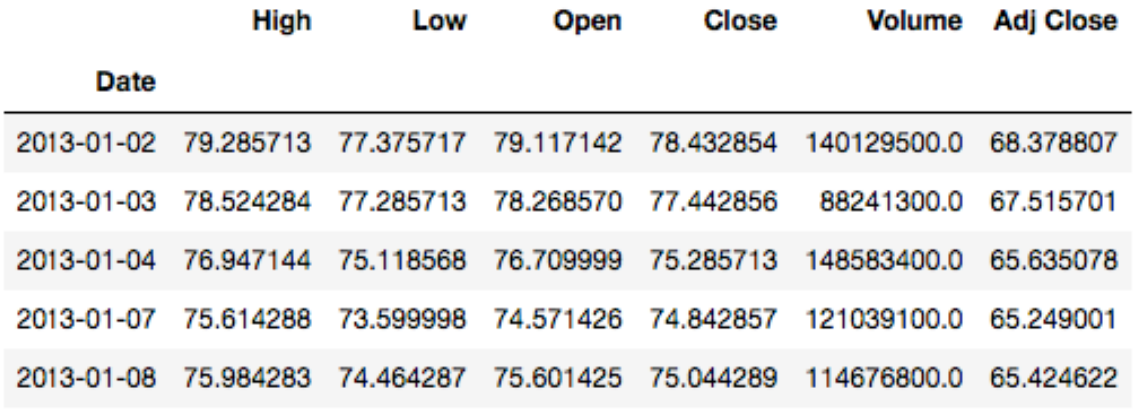
1. ARIMA

ARIMA stands for ‘AutoRegressive Integrated Moving Average’, is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values. ARIMA models must be non- seasonal and exhibits a patterns and is not white noise. An ARIMA model uses it own lag as predictors therefore they will work best when they are not correlated and are independent of each other so we much make the series stationary to achieve this. An Auto regressive model (AR) depends on its own lag $Y_{t.}$ While the moving average (MA) depends on the lagged error forecasts. Hence the model is as follows

Predicted Yt = Constant + Linear combination Lags of Y (up to p lags) + Linear Combination of Lagged forecast errors (up to q lags). Hence the p, d and q variables need to be identified. p is the order of the AR term, q is the order of the MA term and d is the number of differencing required to make the time series stationary.

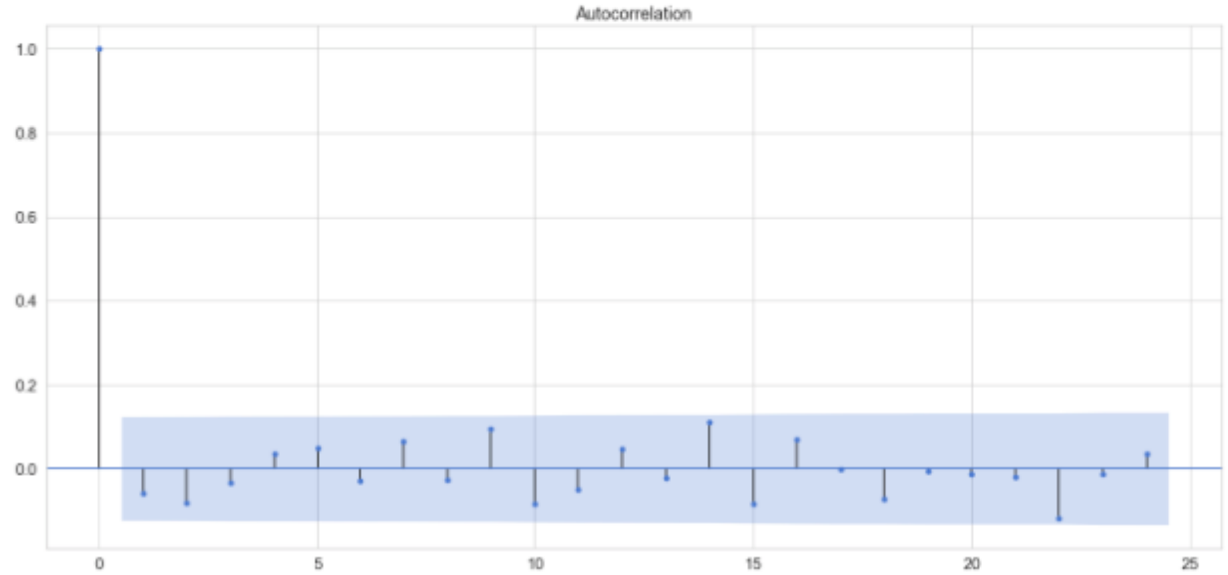
Steps in carrying out in ARIMA forecast :

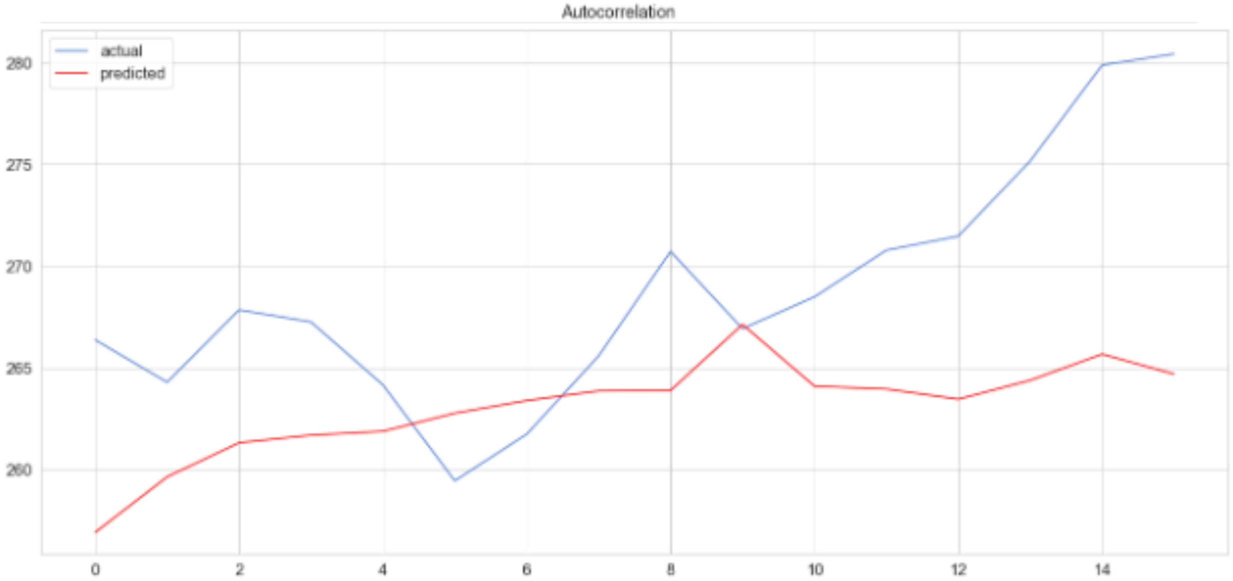
1. Import the dataset.
2. Subset data set to target variable only
3. Build a train and test set.
4. Check stationarity in data.
5. Make data stationery.
6. Build models.
7. Validate the model with the test set



	High	Low	Open	Close	Volume	Adj Close
Date						
2013-01-02	79.285713	77.375717	79.117142	78.432854	140129500.0	68.378807
2013-01-03	78.524284	77.285713	78.268570	77.442856	88241300.0	67.515701
2013-01-04	76.947144	75.118568	76.709999	75.285713	148583400.0	65.635078
2013-01-07	75.614288	73.599998	74.571426	74.842857	121039100.0	65.249001
2013-01-08	75.984283	74.464287	75.601425	75.044289	114676800.0	65.424622

The Closing prices was subset out as the target variable with the past values of the close to be used for forecasting. We then look at the data. The series doesn’t look stationary as it is showing an increasing trend. ADF fuller test confirms the null hypothesis. Stationary means mean, variance, and covariance are constant over periods., First we look at autocorrection plot.

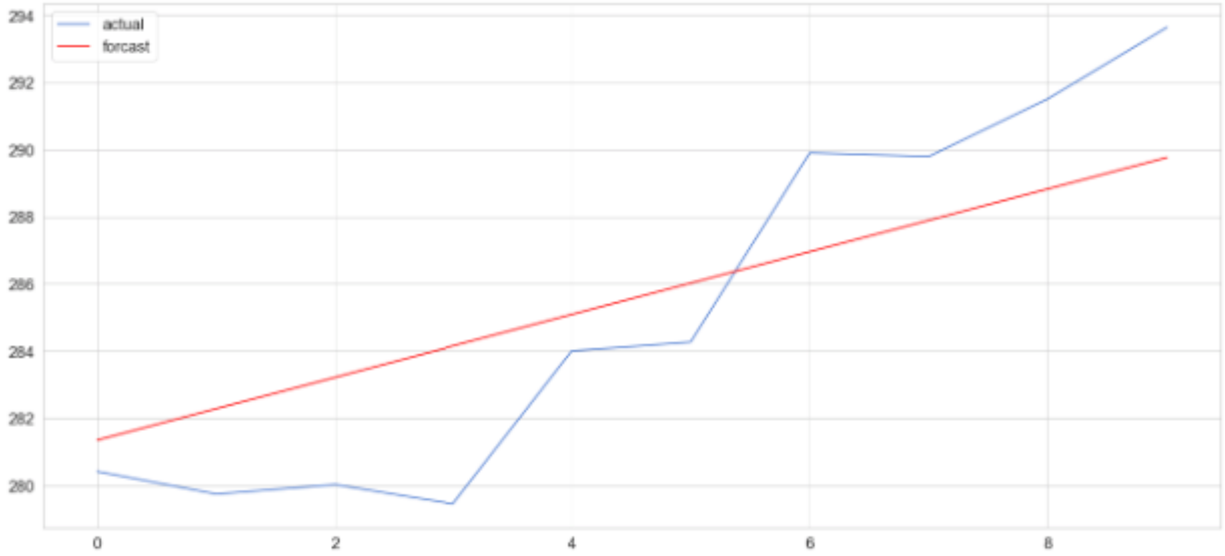




Which shows there is correlation evidence for series being non-stationary Now we will use the augmented dickey fuller test to check if the series are stationary. The results of the ads test is a p values of

0.99 which means we cannot cannot reject the null hypothesis of a unit root which suggests the series is not stationary.

The order of differencing, d = 0 if the time series is stationary if not we have to carry out differencing and check the act plot until we confirm the series is now stationary. For the above series, the time series reaches stationarity with one order of differencing, we confirm this further by looking at the differencing plot, where we confirm that the differencing mean is constant at 0. Also we can check the Residuals and the density plot



here we confirm that one order of differencing is enough as the residual errors seem fine with near zero mean and uniform variance, and the density plot shows and uniform distribution. The next steps are to train the model and use it to predict the test set.

These values though on some occasions coincide with the test values most times the predictions are off, this algorithm may need to be tweaked to further increase accuracy.

The test and training sets were created with the train to be 25 days before and the train set to predict 10 days out. Although the When using predict command the in-sample lagged values are used for the prediction which may not give a true independent future forecast. The forecast command in the statsmodel tools will be used to predict values into the future with an out of sample prediction that is not being influenced by in-sample values.

We will now manually choose p, d and q values to see if we can come up with a better predicting model. After looking through most combinations for p, d, q combinations using the combination with the lowest AIC as the deciding parameter values were chosen for p, d, q with we choose the model with AIC value of 80.15, and plot for the next 10 steps

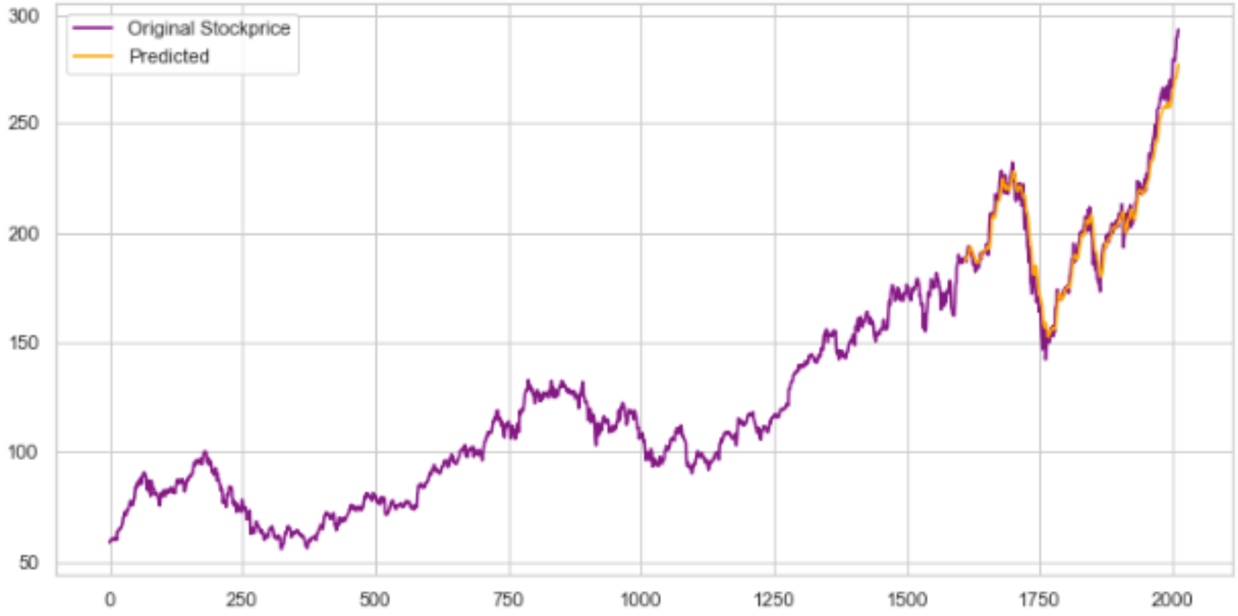
2. Long Short-Term Memory Network (LSTM)

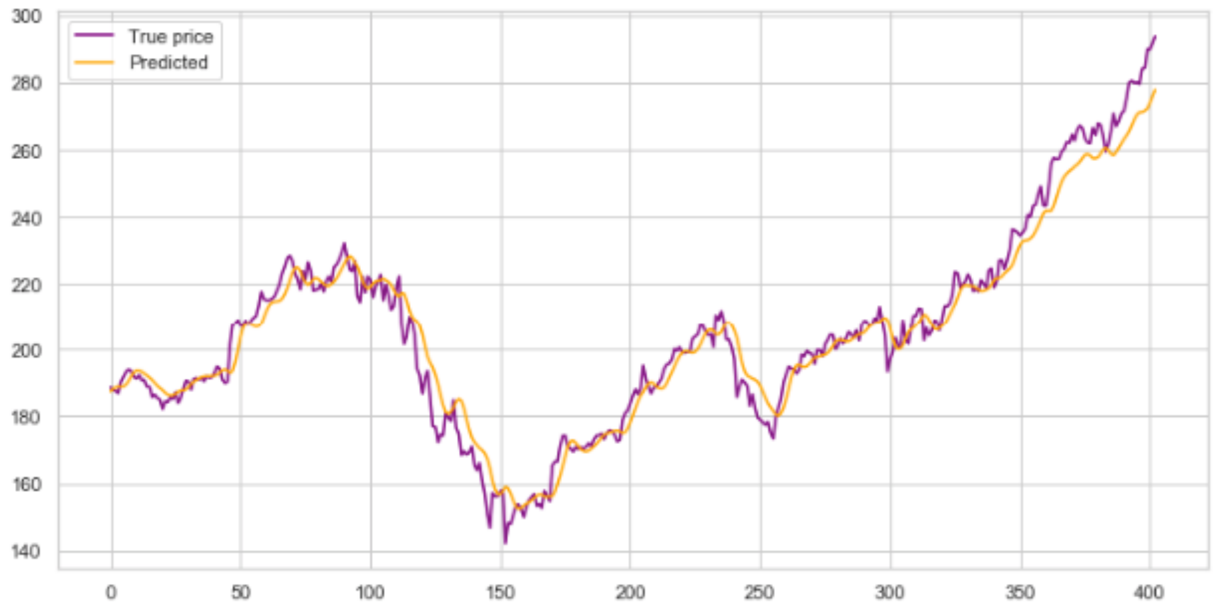
The Long Short-Term Memory network, or LSTM network, is a recurrent neural network (RNN) that is trained using sequential observations learned from the earlier stages to forecast future trends. LSTM can capture long term influences, with a component it has called memory block which replaces hidden layer by managing the blocks state using sate structure. There are 3 gates the forget gate that conditionally decides what information to keep or discard, the input gate which conditionally decides which input to update the state of the memory block and the output gate which decides conditionally what to output based on the input and memory of the block.

Steps in carrying out in LSTM forecast :

1. Import the dataset.
2. Subset data set to target variable only
3. Convert to numpy array and reshape
4. Build a train and test set.
5. Feature scale the data.
6. Create data step to predict next day price.
7. Build model.
8. Validate the model with the test set.

The full dataset with all columns is shown in fig.XX.X and subset target variable of close prices is shown in fig.XX.X. The data is then feature scaled to scale all data between 0 to 1. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. The test train dataset is divided with 20% test and 80% training. The time steps are then created in this project we use 50 steps to predict 1 step. LSTMs usually accepts a 3d array array format so after the time steps are created the test and train set are reshaped into a 3d array. The LSTM model is created with keras and consist of 4 types of layers; a sequential layer for initialization, the LSTM layer, the dropout layer to prevent overfitting and the dense layer for adding a densely connected neural network layer is normally at the end of the model. The LSTM layer comprises of 3 arguments the units which is the dimensionality of the output space, the return sequences which determines whether to return the last output in the output sequence, or the full sequence so when the last layer is input the sequence should not be returned, and the input shape which is the shape of the training data set. The drop out layers were specified at 20% hence 20% of the layers will be dropped. The dense layer is then specified. This model uses ADAM optimizer and sets the loss as the mean square error.





Conclusion

The loss function ranged from 0.0172 to 0.001 these are fairly good values. However when looking at the plot for the actual versus the predicted the predicted prices were close, always slightly above or below the actual prices. The exact predicted values would not be good for trading as they are mostly not perfectly accurate but indicated the overall trend and direction so could be somewhat effective for stock predictions. This LSTM model could be altered and improved upon to generate more accurate results and from this study it seems LSTM models could be useful in stock predictions.

3. Linear Regression

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the dependency between one dependent variable (usually denoted by y) and one or a series of other variables (known as independent variables and usually denoted by x). Linear regression is used for modeling qualitative dependency, practically it fits a straight line to some data in order to capture the linear relationship between that data. Linear regression helps answer the question of if and how a change in y influences a change in y. The regression line is constructed by optimizing the parameters of the straight line function such that the line best fits a sample of (x, y) observations where y is a variable dependent on the value of x. In this study we will be using the ordinary least squares method. The least squares method minimizes the sum of the errors squared, where the errors are the residuals between the fitted curve and the set of data points. The residual can be calculated using perpendicular distances or vertical distances. The errors are squared so that the residuals form a continuous differentiable quantity.

Steps in carrying out in OLS forecast :

1. Import the dataset.
2. Subset data set to target variable only.
3. Difference the model to make stationary.
4. Shift data to get average for last 5 days.
5. Remove the target variable
6. Build a train and test set.
7. Build model.
8. Validate the model with the test set.

The full dataset with all columns is shown in fig.XX.X and subset target variable of close prices is shown in fig.XX.X. Here the target variable will be the adjusted close. Which is subsets to a new data frame. The data is then differenced in order to make the time series stationary. The differenced values are added to the data frame. Then the data frame is shifted to sequentially five days back to the day before the predicted day's adjusted close, these values are added as five separate columns (explanatory variables) to the data frame and the last variable of the data frame is the average of the past 5 days adjusted close. This allows us to use the previous 5 days individual prices, the average of the past 5 day prices and the differencing between the previous days prices to predict the next day prices. The target variable is then removed for the data set. The model is then trained with data from Apple stock prices from 2015-01-01 to 2018-10-01 which was approximately 917 values and the test data was from 2018-10-01 to 2018-12-15 and approximately 53 values, making the test train ratio about 95% train data to 5% test data.

Conclusion

In this experiment we carried out the experiment we carried out predictions for the close of the stocks using 5 days prior data to predict one day in the future. The complete data set consists of about 8 years from January 2015 to December 2019. The data train to test set was about 95% to 5%. The predictions were good with a RMSE of 5.52e-15, the R^2 score is 1.0 which is the best possible score. When looking at the predicted to the actual values they seem to be very near to accurately predicting the next day values. So we can say this system nay be helpful in stock analysis.