Analysis and Prediction/Forecasting of stock prices of   
Vodafone Idea Limited and Bharti Airtel Limited

Gladwin Gracias, Rakhee Singh, Noella Joy

SS and LS Patkar Varde College

# ABSTRACT

This Project/Analysis is mainly confined to the companies Vodafone Idea Limited and Bharti Airtel Limited which trade on the BSE. The goal of this project is to study the behavior of the stock prices of these two companies and create a model using the Autoregressive Integrated Moving Average (ARIMA) method. The model will be trained on historical data from the last 5 years (5th October 2015 to 30th September 2020) and will forecast stock prices on a test set derived from the training set.

Keywords: Time Series, Stock prices, Forecasting

Table of Contents

[ABSTRACT 2](#_Toc52996237)

[CH -1 : INTRODUCTION 5](#_Toc52996238)

[CH – 2 : PROJECT OBJECTIVES 7](#_Toc52996239)

[CH – 3 : PROPOSED APPROACH 8](#_Toc52996240)

[Phase 1: Identification 8](#_Toc52996241)

[Step 1: Data Preparation 8](#_Toc52996242)

[Phase 2: Model Selection 8](#_Toc52996243)

[Phase 3: Estimation and Testing 8](#_Toc52996244)

[Step 1: Estimation 8](#_Toc52996245)

[Step 2: Diagnostics 8](#_Toc52996246)

[Visualizing the series 9](#_Toc52996247)

[Getting an overview of the datasets 9](#_Toc52996248)

[Plotting the series against the time index 11](#_Toc52996249)

[Plotting the series with Moving average (25 and 50 windows) 11](#_Toc52996250)

[Percentage change in the stock prices 12](#_Toc52996251)

[CH – 4 : TESTING FOR STATIONARITY 13](#_Toc52996252)

[ADF (Augmented Dickey-Fuller) Test 13](#_Toc52996253)

[Stationarity and Dickey Fuller test for VI 14](#_Toc52996254)

[Stationarity and Dickey Fuller test for BA 15](#_Toc52996255)

[CH – 5 : MAKING THE TIME SERIES STATIONARY 17](#_Toc52996256)

[Estimating and eliminating the trend in data 17](#_Toc52996257)

[Differencing 18](#_Toc52996258)

[Performing dickey fuller test and stationarity test for VI data 19](#_Toc52996259)

[Performing dickey fuller test and stationarity test for BA data 20](#_Toc52996260)

[Decomposing 21](#_Toc52996261)

[Decomposing the scaled VI data 21](#_Toc52996262)

[Performing ADF test on the residuals of the VI series 22](#_Toc52996263)

[Decomposing the BA scaled data 23](#_Toc52996264)

[Performing ADF test on the BA residuals 24](#_Toc52996265)

[Splitting the VI scaled data into training and testing splits 25](#_Toc52996266)

[Splitting the BA scaled data into training and testing splits 25](#_Toc52996267)

[CH – 6 : FORECASTING A TIME SERIES 26](#_Toc52996268)

[Plotting the ACF/PACF graphs for VI 26](#_Toc52996269)

[Plotting the ACF/PACF graphs for BA 28](#_Toc52996270)

[THE MODEL 31](#_Toc52996271)

[Residual Diagnostics 32](#_Toc52996272)

[FORECAST 34](#_Toc52996273)

[Plotting the forecasts 34](#_Toc52996274)

[CH – 7 : PERFORMANCE EVALUATION OF THE MODEL 37](#_Toc52996275)

[VI Performance Evaluation 37](#_Toc52996276)

[BA Performance Evaluation 37](#_Toc52996277)

# CH -1 : INTRODUCTION

A stock or share (also known as a company’s “equity”) is a financial instrument that represents ownership in a company or corporation and represents a proportionate claim on its assets (what it owns) and earnings (what it generates in profits).

The stock market is a market that enables the seamless exchange of buying and selling of company stocks.

**How does the stock market work?**

The stock market works through a network of exchanges — you may have heard of the New York Stock Exchange, NASDAQ or Sensex. Companies list shares of their stock on an exchange through a process called an initial public offering or IPO. Investors purchase those shares, which allows the company to raise money to grow its business. Investors can then buy and sell these stocks among themselves, and the exchange tracks the supply and demand of each listed stock.

That supply and demand help determine the price for each security or the levels at which stock market participants — investors and traders — are willing to buy or sell.

Stock market analysis is divided into two parts – Fundamental Analysis and Technical Analysis.

* Fundamental Analysis involves analyzing the company’s future profitability on the basis of its current business environment and financial performance.
* Technical Analysis, on the other hand, includes reading the charts and using statistical figures to identify the trends in the stock market.

**Time Series**

TS is a collection of data points collected at constant time intervals. Time series is thought to consist of three systematic components including level, trend, seasonality, and one non-systematic component called noise.

These components are defined as follows:

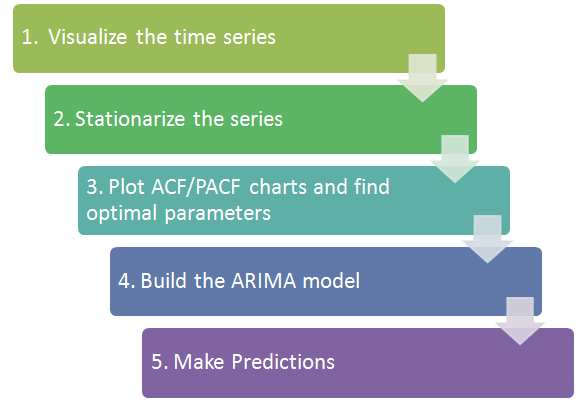
* Level: The average value in the series.
* Trend: The increasing or decreasing value in the series.
* Seasonality: The repeating short-term cycle in the series.
* Noise: The random variation in the series.

**Forecasting Methods**

The time series forecasting methods are generally classified into two broad categories based on statistical concepts and computation intelligence techniques such as Neural Networks and Genetic Algorithms. Statistical time series forecasting methods are as follows:

* Exponential smoothing methods
* Regression methods
* Autoregressive integrated moving average (ARIMA) methods
* Threshold methods
* Generalized autoregressive conditionally heteroskedastic (GARCH) methods

# CH – 2 : PROJECT OBJECTIVES



ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. ARIMA models are fitted to time series data to better understand the data or to forecast future points in the series i.e. forecasting.

ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step can be applied one or more times to eliminate the non-stationarity.

A TS is said to be stationary if its statistical properties such as mean, variance remain constant over time. But why is it important? Most of the TS models work on the assumption that the TS is stationary. Intuitively, we can state that if a TS has a particular behavior over time, there is a very high probability that it will follow the same in the future.

# CH – 3 : PROPOSED APPROACH

## Phase 1: Identification

### Step 1: Data Preparation

* Transform the data to stabilize the attributes
* Find the difference if the data is not stationary
* Make the series attain stationarity through differencing iterations

## Phase 2: Model Selection

* Examine the data
* Plot ACF and PACF to identify the models to use

## Phase 3: Estimation and Testing

### Step 1: Estimation

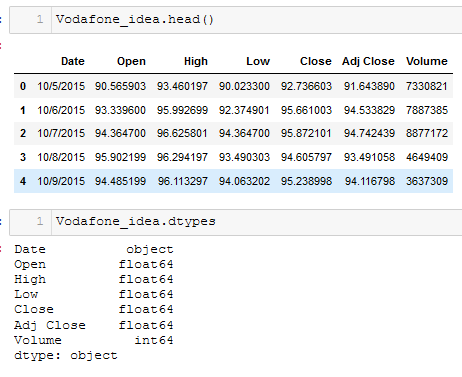
* Estimate the parameters in the potential models
* Select the best model using the AIC BIC criteria

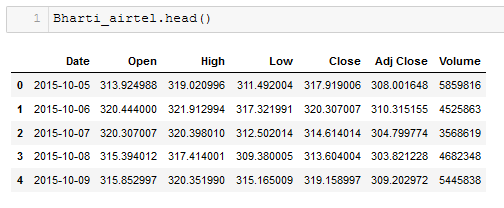
### Step 2: Diagnostics

* Check ACF/ PACF of the residual values
* Test the residuals for White Noise

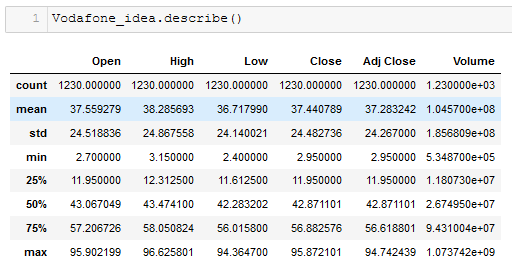
## Visualizing the series

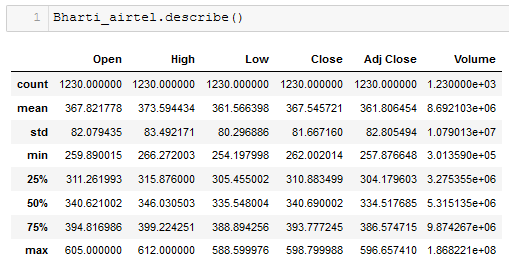
### Getting an overview of the datasets





By using the describe() function we can see that we have a total of 1230 rows in both the datasets respectively. Describe() also gives us the statistical parameters of the dataset such as mean, std deviation, quartiles and the min-max values.

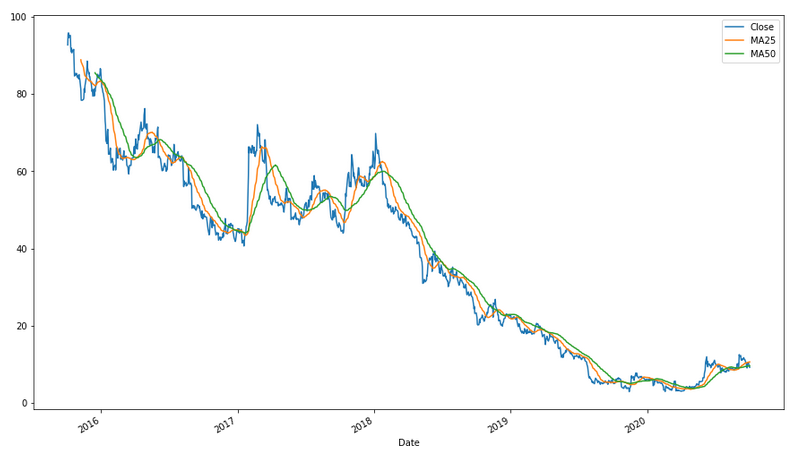




### Plotting the series against the time index

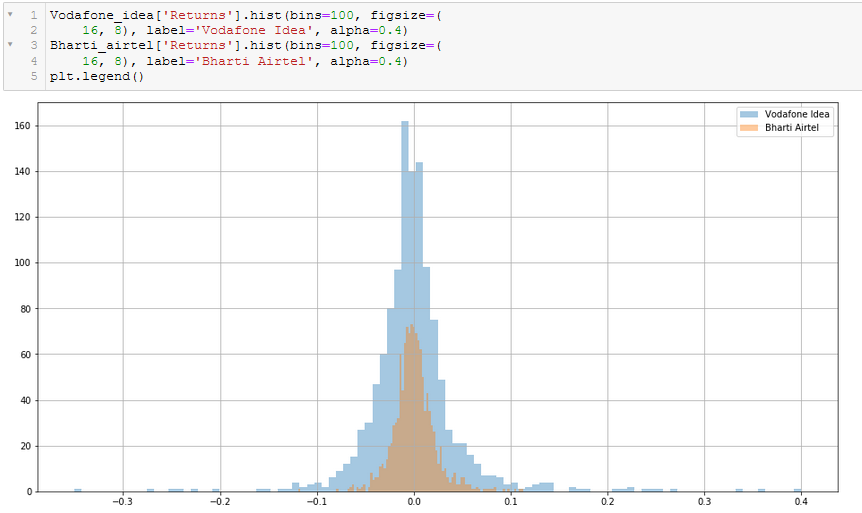


### Plotting the series with Moving average (25 and 50 windows)



### Percentage change in the stock prices

By looking at a daily percentage change, it can be concluded stock's volatility and how risky it is. The formula is simple: more volatility = more risk. Percentage change is simply how much the stock has changed relative to the previous period’s closing price.



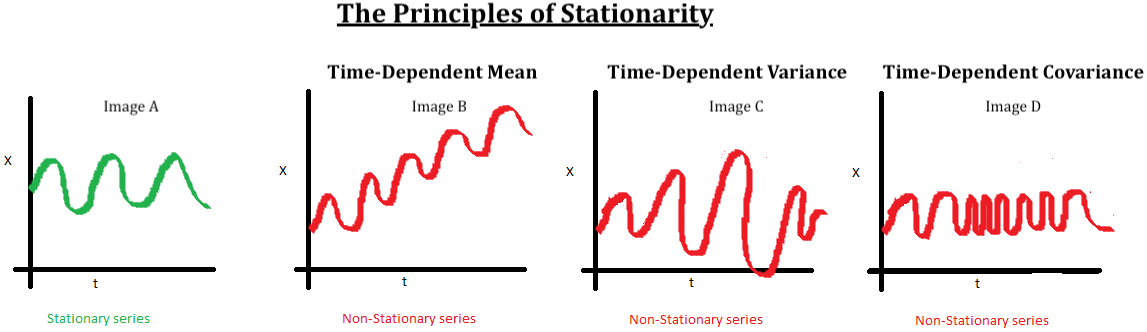
We can see from the plot that the stock prices of both companies have a relatively similar percentage change, both hovering around the mean of 0 up to 0.1 percent.

# CH – 4 : TESTING FOR STATIONARITY

A TS is said to be stationary if its statistical properties such as mean, variance remain constant over time. But why is it important? Most of the TS models work on the assumption that the TS is stationary. Intuitively, we can state that if a TS has a particular behavior over time, there is a very high probability that it will follow the same in the future.

We can assume the series to be stationary if it has constant statistical properties over time, i.e. the following:

* constant mean
* constant variance
* an auto covariance that does not depend on time



## ADF (Augmented Dickey-Fuller) Test

The Dickey-Fuller test is one of the most popular statistical tests. It can be used to determine the presence of unit root in the series, and hence help us understand if the series is stationary or not. The null and alternate hypothesis of this test is:

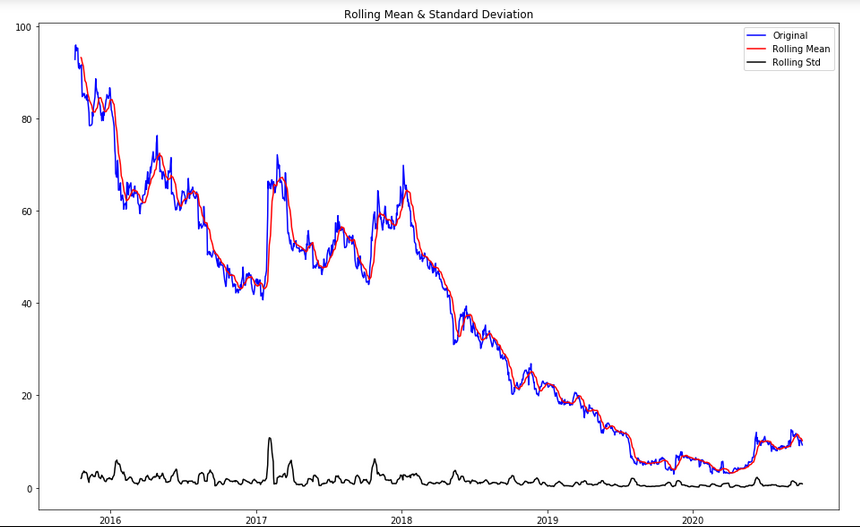
* **Null Hypothesis:** The series has a unit root (value of a =1)
* **Alternate Hypothesis:** The series has no unit root.

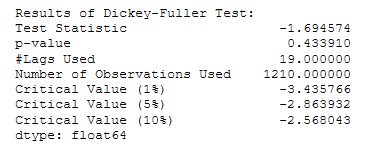
If we fail to reject the null hypothesis, we can say that the series is non-stationary. This means that the series can be linear or difference stationary.

Plotting Rolling Statistics: We can plot the moving average or moving variance and see if it varies with time. i.e. At any instance ‘t’ we’ll take the average/variance of the last year, i.e. last 12 months but this is more of a visual technique.

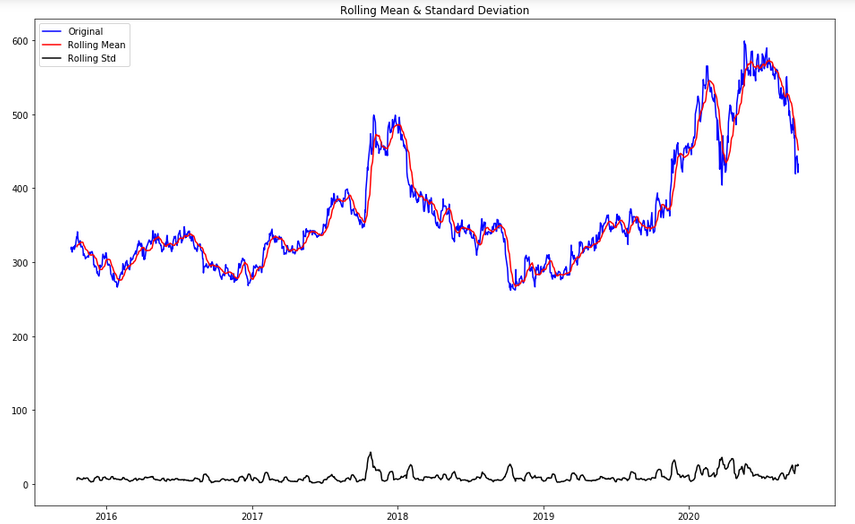


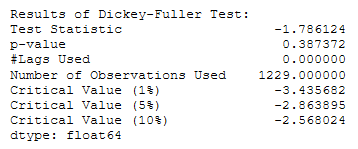
### Stationarity and Dickey Fuller test for VI





### Stationarity and Dickey Fuller test for BA





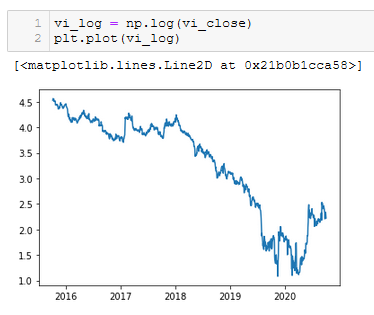
**Test for stationarity:** If the test statistic is less than the critical value, we can reject the null hypothesis (aka the series is stationary). When the test statistic is greater than the critical value, we fail to reject the null hypothesis (which means the series is not stationary).

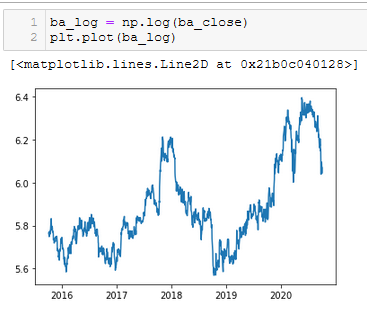
We see that the p-value is greater than 0.05 so we cannot reject the Null hypothesis. Also, the test statistics is greater than the critical values. So the data is non-stationary.

# CH – 5 : MAKING THE TIME SERIES STATIONARY

## Estimating and eliminating the trend in data

Transformations are used to stabilize the non-constant variance of a series. Common transformation methods include power transform, square root, and log transform.



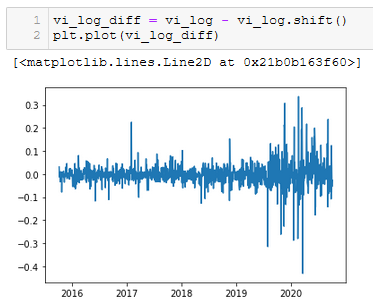


## Differencing

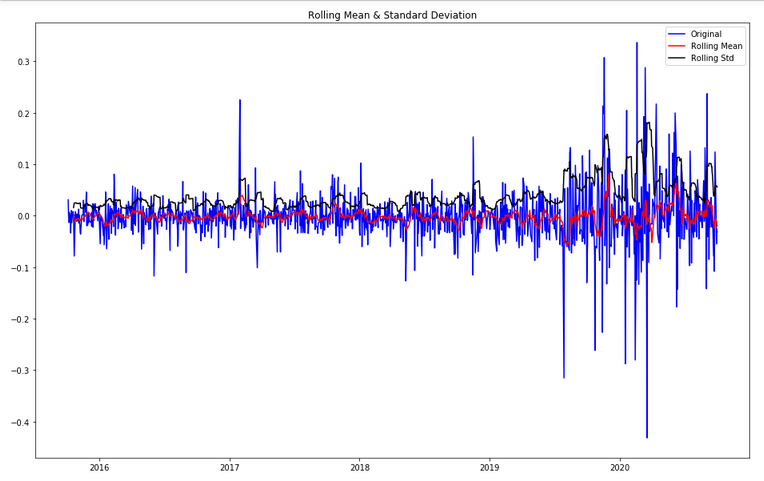
In this method, we compute the difference of consecutive terms in the series. Differencing is typically performed to get rid of the varying mean. Mathematically, differencing can be written as:

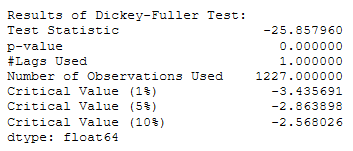
*yt’ = yt – y(t-1)*

Where yt is the value at a time t

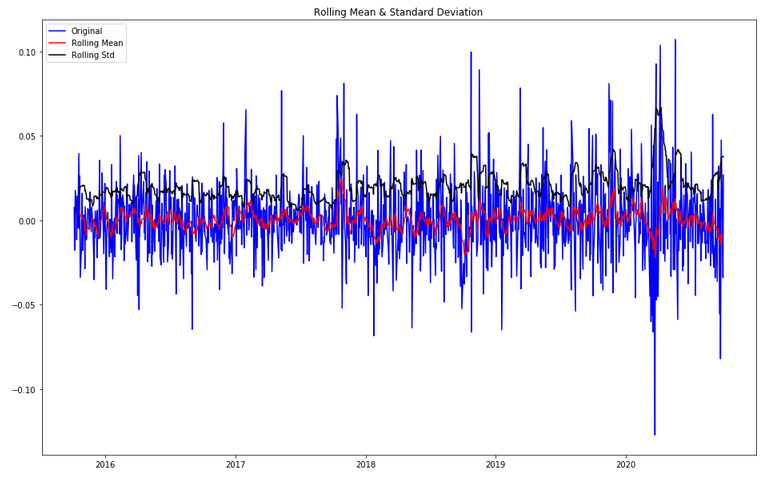


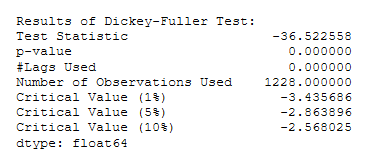
### Performing dickey fuller test and stationarity test for VI data





### Performing dickey fuller test and stationarity test for BA data

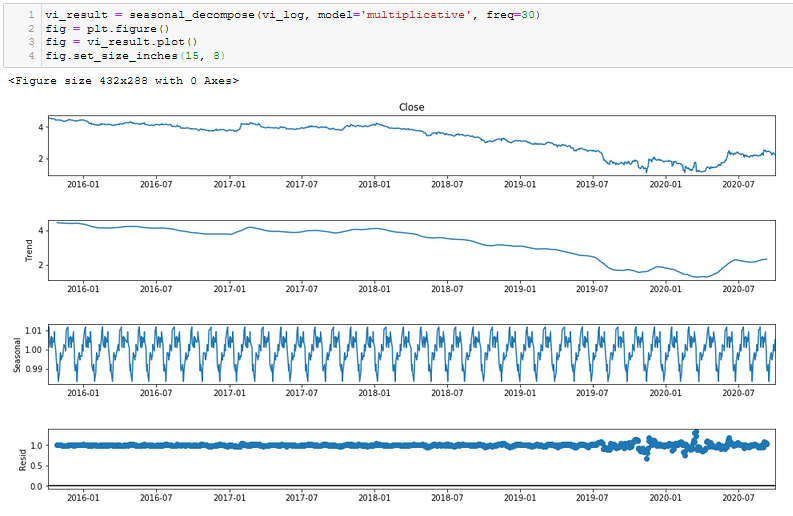




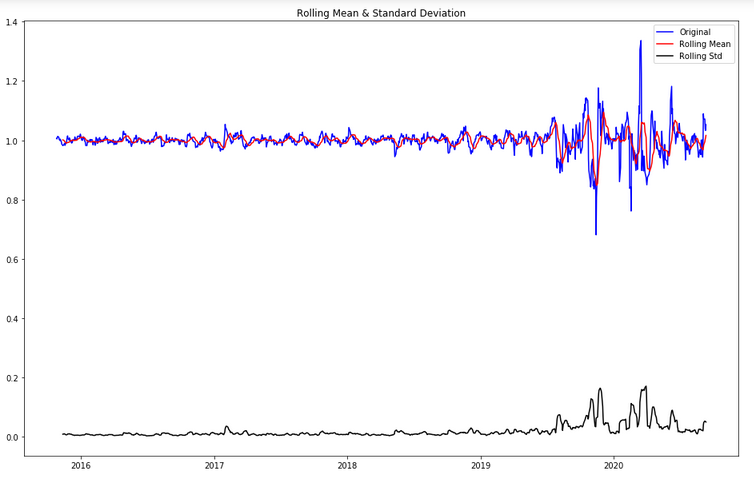
## Decomposing

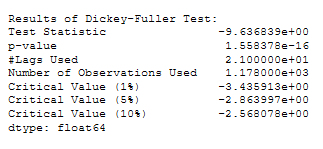
In this approach, both trend and seasonality are modeled separately and the remaining part of the series is returned.

### Decomposing the scaled VI data

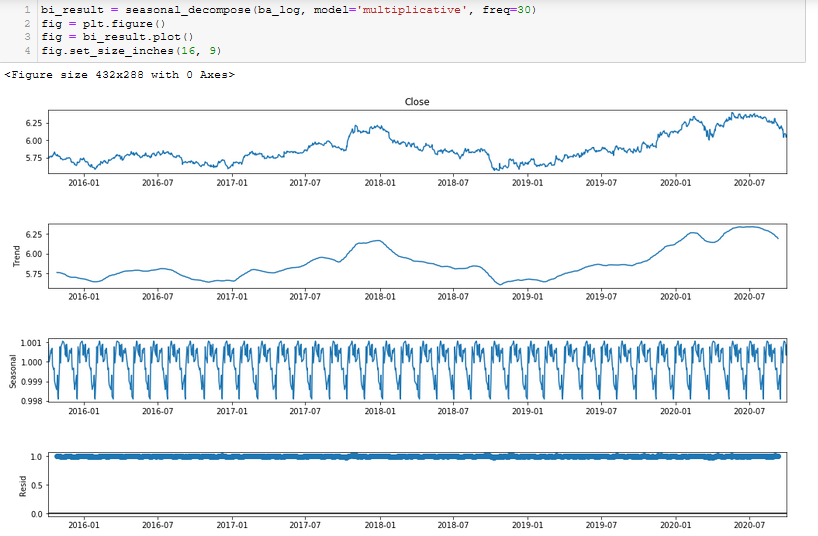


### Performing ADF test on the residuals of the VI series

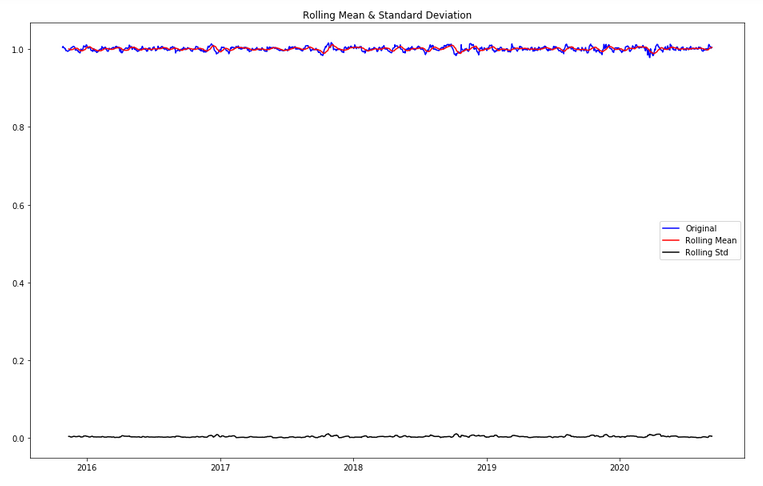


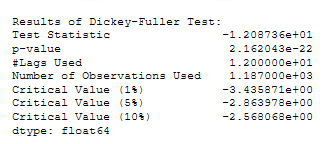


### Decomposing the BA scaled data

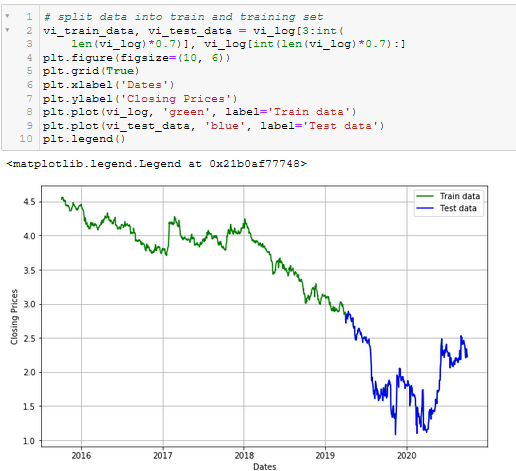


### Performing ADF test on the BA residuals

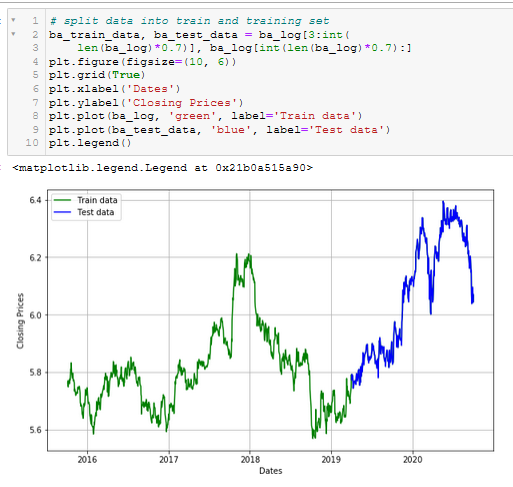




### Splitting the VI scaled data into training and testing splits



### Splitting the BA scaled data into training and testing splits



# CH – 6 : FORECASTING A TIME SERIES

Having performed the trend and seasonality estimation techniques, there can be two situations:

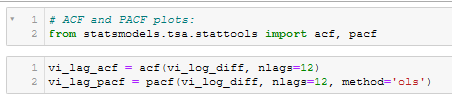
1. A strictly stationary series with no dependence among the values. This is the easy case wherein we can model the residuals as white noise. But this is very rare.
2. A series with significant dependence among values. In this case we need to use some statistical models like ARIMA to forecast the data.

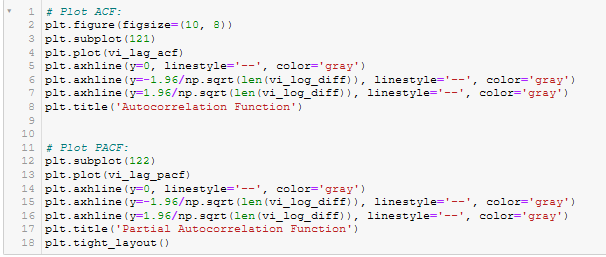
* **Number of AR (Auto-Regressive) terms (p):** AR terms are just lags of dependent variable. For instance if p is 5, the predictors for x(t) will be x(t-1)….x(t-5).
* **Number of MA (Moving Average) terms (q):** MA terms are lagged forecast errors in prediction equation. For instance if q is 5, the predictors for x(t) will be e(t-1)….e(t-5) where e(i) is the difference between the moving average at i th instant and actual value.
* **Number of Differences (d):** These are the number of no seasonal differences, i.e. in this case we took the first order difference. So either we can pass that variable and put d=0 or pass the original variable and put d=1. Both will generate same results.

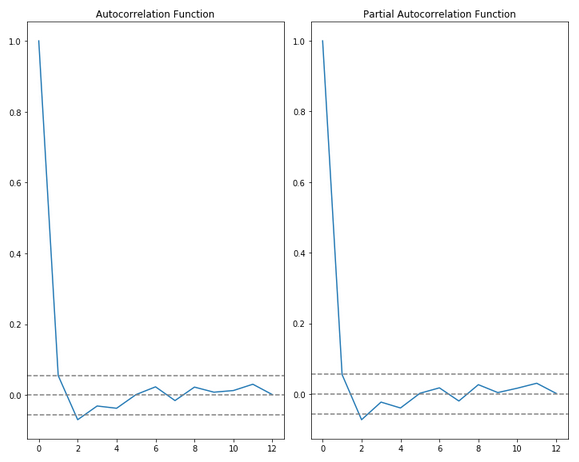
An importance concern here is how to determine the value of ‘p’ and ‘q’. We use two plots to determine these numbers. Let’s discuss them first.

* **Autocorrelation Function (ACF):** It is a measure of the correlation between the the TS with a lagged version of itself. For instance at lag 5, ACF would compare series at time instant ‘t1’…’t2’ with series at instant ‘t1-5’…’t2-5’ (t1-5 and t2 being end points).
* **Partial Autocorrelation Function (PACF):** This measures the correlation between the TS with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons. E.g. at lag 5, it will check the correlation but remove the effects already explained by lags 1 to 4.

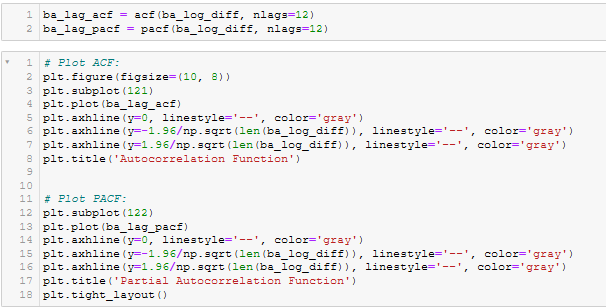
## Plotting the ACF/PACF graphs for VI

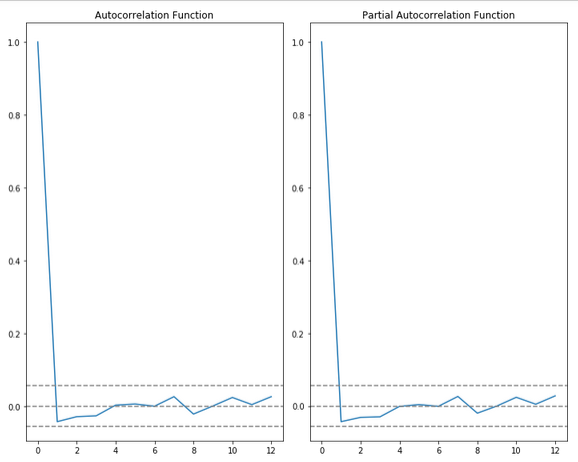






## Plotting the ACF/PACF graphs for BA

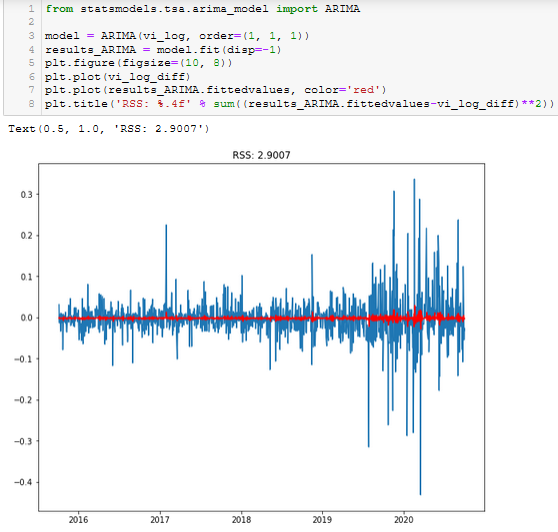


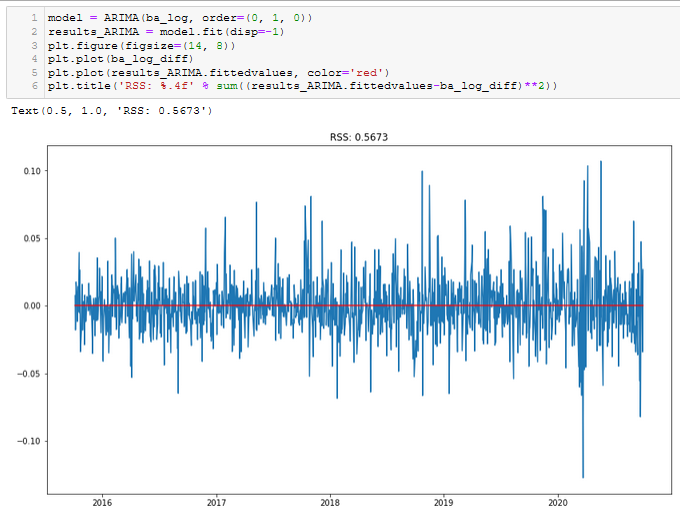


In this plot, the two dotted lines on either sides of 0 are the confidence intervals. These can be used to determine the ‘p’ and ‘q’ values as:

* p – The lag value where the PACF chart crosses the upper confidence interval for the first time. If you notice closely, in this case p=0
* q – The lag value where the ACF chart crosses the upper confidence interval for the first time. If you notice closely, in this case q=0

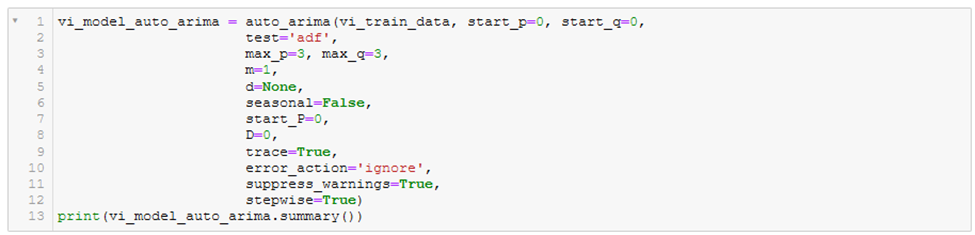
Using the p, d, q values from the ACF/PACF plots we will run an ARIMA model on the scaled data.

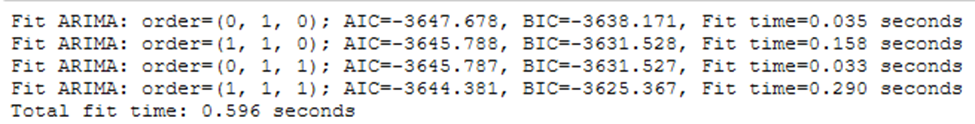




## THE MODEL

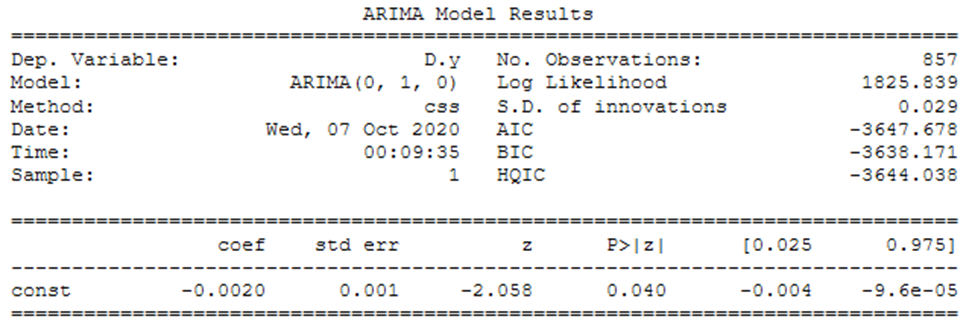
auto\_arima() uses a stepwise approach to choose the best combination of the p, d, q values for the model which has the lowest AIC.

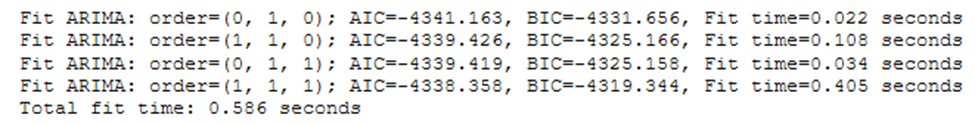




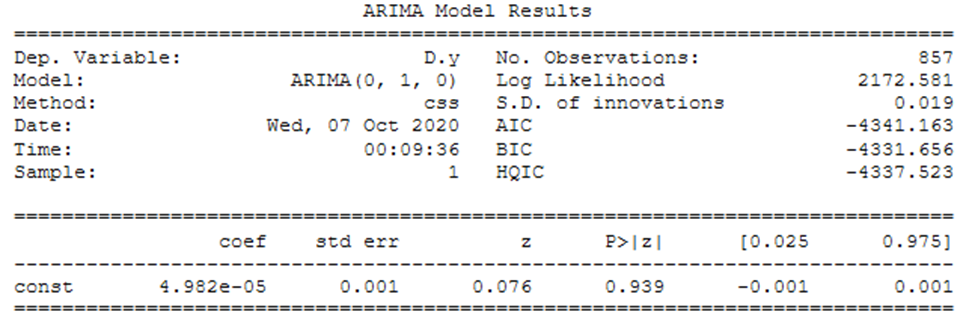
Here, we can see that auto\_arima creates multiple models with difference combinations of the hyperparameters and gives us the parameters which gives us the lowest AIC.

This autoarima gives us the best hyperparameters for the VI model.

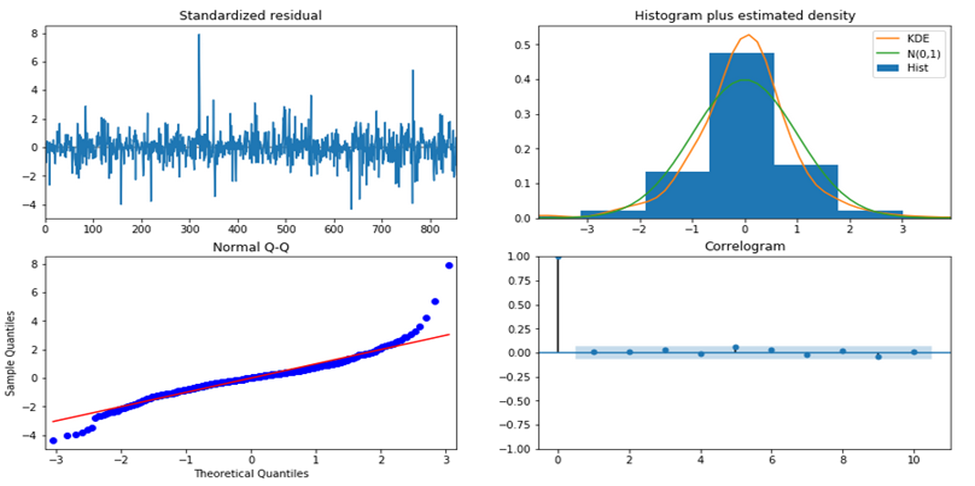


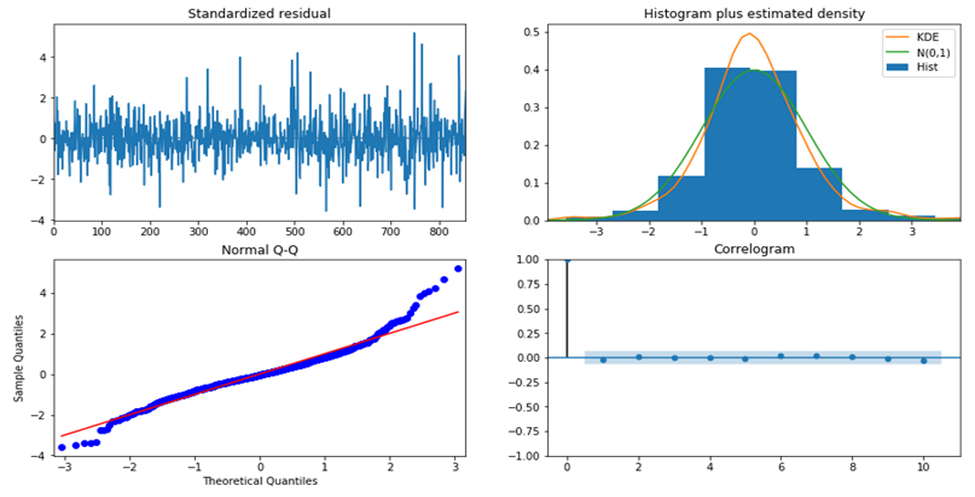


And these are the optimal hyperparameters for the BA model.



### Residual Diagnostics





**Top left:** The residual errors seem to fluctuate around a mean of zero and have a uniform variance.

**Top Right:** The density plot suggest normal distribution with mean zero.

**Bottom left:** All the dots should fall perfectly in line with the red line. Any significant deviations would imply the distribution is skewed.

**Bottom Right:** The Correlogram, aka, ACF plot shows the residual errors are not autocorrelated. Any autocorrelation would imply that there is some pattern in the residual errors which are not explained in the model. So you will need to look for more X’s (predictors) to the model.

## FORECAST

Now, let’s forecast the data with the model created in the step above.

The same forecast function is used for both the VI and BA datasets.

The .forecast() function of ARIMA accepts 3 arguments.

* **Steps:** These are the out of sample forecasts from the end of the sample. (In our model, after dropping NA values, we are left with 857 rows in the training split and 369 rows in the test split. These 369 rows are at the end of the sample AKA the last 369 rows.)
* **Exog:** An extraneous array of samples for ARMAX process. We won’t be using this.
* **Alpha:** This is used to attain the confidence interval of the forecast

.forecast() returns the following results

* **Vi\_fc\_series:** This is the array of the forecasted values
* **Vi\_se:** Array of the Standard error of the forecasts
* **Vi\_conf:** A 2D array of the confidence intervals of the forecast values. It includes the lower limit and the upper limit of the forecast.

## Plotting the forecasts

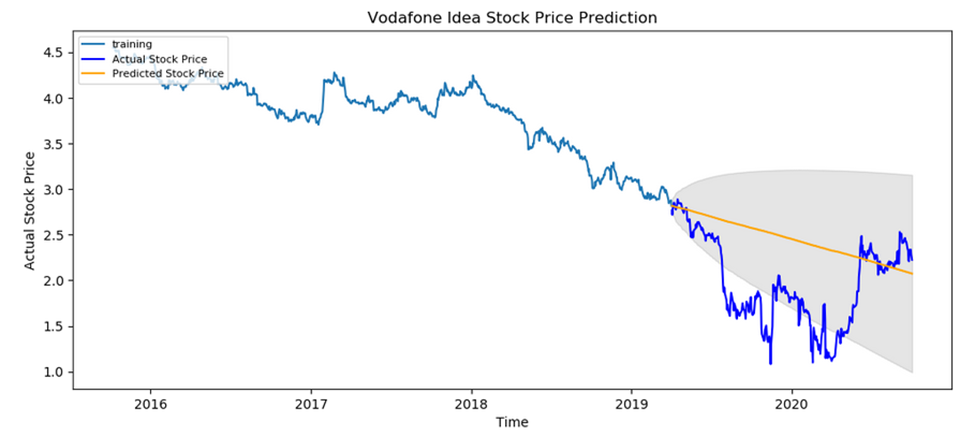
Now that we have forecast values and the range of those values, we can plot it.

Here’s what we need to see in the forecast.

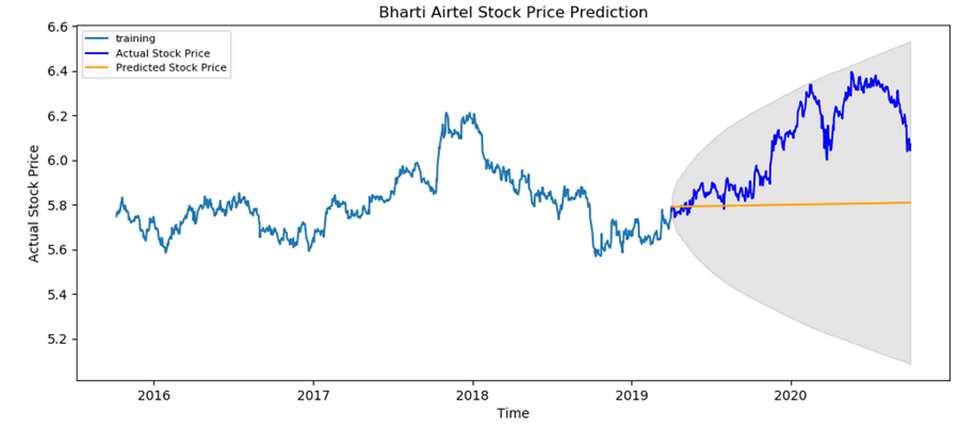
* The training data
* The test data
* The forecast
* Limits of the forecast (Confidence Interval)

To do this we need to convert the outputs of the .forecast function into a plottable structure i.e. a time series.

* Firstly, since all the forecast and conf values are based on the test split, we will set the index of all the series to the same date index as the test index.
* Next we will convert the vi\_forecast array to a time series with the pd.series() function.
* Vi\_lower\_series will contain the lower limit of the confidence intervals i.e. the 0th column of the vi\_conf array.
* Vi\_upper\_series will contain the upper limit of the confidence intervals i.e. the 1st column of the vi\_conf array.
* Finally we need to fill the area between the lower and upper series. We can do this with the help of the fill\_between function of matplotlib



From this plot, we can see that the forecasts follow the general trend of the data for VI i.e. a downward trend.



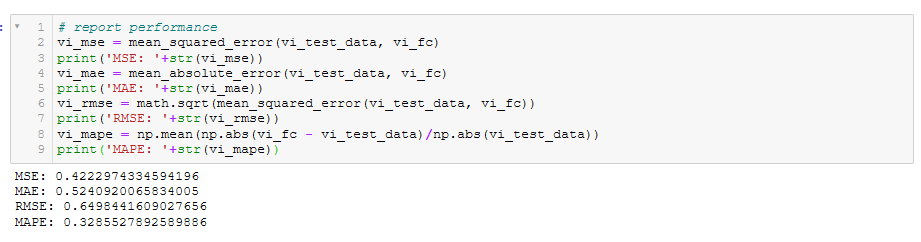
From this plot, we can see that the forecasts follow the general trend of the data for BA i.e. a slight upward trend.

# CH – 7 : PERFORMANCE EVALUATION OF THE MODEL

One of the final steps of any analysis is checking and evaluating the performance of the model.

The main performance indicator we will be looking out for is the MAPE value.

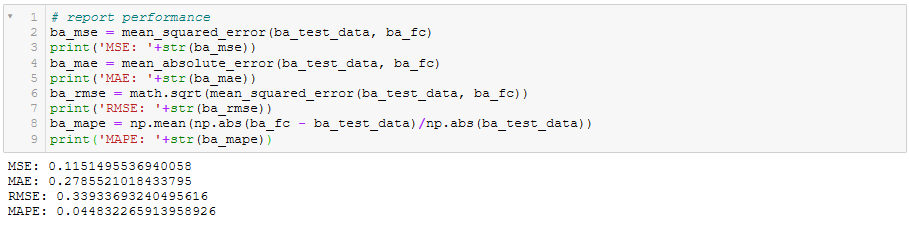
## VI Performance Evaluation



As we can see, the MAPE value for the VI model is very high.

## BA Performance Evaluation

The MAPE value for the BA model is very good.



A MAPE of 4.4% means that the model is upto 95.4% accurate in its forecasts.

# CH – 8 : CONCLUSION AND INSIGHTS

A good way to improve our predictions would be to use Deep Learning methods like LSTM which is able to store past information that is important and forget information that is not needed.

There are multiple factors involved in stock price predictions. Physical factors, Psychological factors, rational vs irrational behaviors etc. All of these factors combine and make the stock prices exceptionally volatile and very difficult to predict with any accuracy.