Chapter 1

INTRODUCTION

**Chapter 1 Introduction**

Time series analysis is a major branch in statistics that mainly focuses on analyzing data set to study the characteristics of the data and extract meaningful statistics in order to predict future values of the series. There are two methods in time series analysis, namely: frequency-domain and time- domain. The former is based mostly on Fourier Transform while the latter closely investigates the auto correlation of the series and is of great use of Box-Jenkins and ARCH/GARCH methods to perform forecast of the series.

Stock price prediction is an important topic in finance and economies which has spurred the interest researchers over the years to develop better predictive models. The Auto Regressive Integrated Moving Average (ARIMA) models have been explored in literature for Time Series prediction. This paper presents extensive process of building stock price predictive model using the ARlMA model.

Nowadays, the stock markets have attracted the interest of many people for investing in the money and buying shares, due to its high returns. High returns also indicate the presence of high risk in investing and volatility of the stock prices. Recently, people are more interested in building appropriate models for prediction of stock prices to reduce the risk in investing and getting high Stock returns.

Published stock data obtained from Bombay Stock Exchange (BSE) are used with stock price predictive model developed. Results obtained reveal that the ARIMA model has a strong potential or short-term prediction and can compete favorably with existing techniques for stock price prediction. Later, we will use ARCH model to tackle the problem of volatility in the prices.

Before proceeding towards the basic assumptions, data description, data analysis and interpretations, first we will learn the basic terminologies that are involved in this field including Stock Markets, the two most popular Indian Stock Exchanges and their Stock lndices.

**1.1 – Basic Terminologies:**

**Stock Market**: A stock market, equity market or share market is the aggregation of buyers and sellers (a loose network of economic transactions, not a physical facility or discrete entity) of stocks (also called shares), which represent ownership claims on businesses; these may include securities listed on a public stock exchange as well as those only traded privately.

Mark Twain once divided the world into two kinds of people: those who have seen the famous Indian monument, the Taj Mahal, and those who haven't. The same could be said about investors. There are two kinds of investors: those who know about the investment opportunities in lndia and those who don't. India may look like a small dot to someone in the US, but upon closer inspection, you will find the same things you would expect from any promising market. There are several stock exchanges in India. Most of the trading in the Indian Stock Market takes place on its two main stock exchanges the Bombay Stock Exchange (BSE) and the National Stock Exchange (NSE). These are prominent in terms of both volume and popularity.

Bombay Stock Exchange( BSE ): The Bombay Stock Exchange in India, popularly called the BSE is the oldest stock exchange in Asia which came into existence in 1875. It has the third largest number of listed companies in the world with 5500 listed companies and also the worlds l l'h largest stock exchange with an overall market capitalization of about $1.43 Trillion as of March, 2016. It is located at Dalal Street, Mumbai, Maharashtra. National Stock Exchange comes second to BSE in terms of popularity.

National Stock Exchange (NSE): The National Stock Exchange in India, popularly called the NSE is now the leading stock exchange in India which came into existence in 1992 as the first demutualized electronic exchange in India. It was the first exchange in the country to provide a modern, fully automated screen based electronic trading system which offered the easy trading facility to the investors spread across the length and breadth of the country. It has the worlds 12th largest stock exchange with an overall market capitalization of about $l .4] Trillion as of March, 2016 placed just after the Bombay Stock Exchange (BSE). It is also located in Mumbai, Maharashtra.

* **Stock Market Index**: A stock market index is a statistical indicator which gives an idea about how the stock market is performing (daily, weekly, monthly, quarterly, yearly, etc.). In India the main indexes to be tracked are - The BSE SENSEX and The NSE NIFTY. Stock indexes are updated constantly throughout the trading day to provide instant information.
* **Description of BSE INDIA**: BSE INDIA is a official website network. It provides data on the all the BSE Senxes of the Bombay Stock Exchange of India.

The two prominent Indian market indexes are Sensex (BSE 30) and Nifty (Nifty 50).

* **Sensex (BSE 30):** Sensex is the oldest market index for equities; it includes shares of 30 firms listed on the BSE, which represent about 45% of the index's free-float market capitalization. It was created in 1986 and provides time series data from April 1979, onward. The base year of Sensex is 1978-79 and the base value is 100. It is also known as BSE 30.
* **Nifty (Nifty 50):** Another index is the Nifty; it includes 50 shares listed on the NSE from 24 different sectors, which represent about 62% of its free-float market capitalization. It was created in 1996 and provides time series data from July I990, onward. The base year of Sensex is 1995 and the base value is 1000. It is also known as Nifty 50.

Stock price forecasting is a popular and important topic in financial and academic studies. Time series analysis is the most common and fundamental method used to perform this task. This project aims to combine the conventional time series analysis technique (ARIMA, ARCH modelling) with the daily historical data of the National Stock Exchange Index that is the Nifty 50 of the last 5 years to predict the daily changes in stock price.

**1.2 - Objectives :**

Recently, forecasting stock market returns or stock market indices are gaining more and more attention in these modern times because, if somehow, the direction of the stock market is predicted before hand, the investors will have better guidance about investing money and naturally, they will have higher chance to earn more from these stock markets. And this can be well accomplished by building up an appropriate model that will predictor forecast the stock market indices before hand with a certain amount of precision and will also provide a set of confidence limits to the predictions.

That is the primary objective of this project is to set up a model initially and check for the validity of the proposed model using- various diagnostic measures and plots and verify whether it can be used for forecasting' of the future values.

After achieving the first objective successfully, we now proceed towards the fact that as the data is taken on the stock prices on a daily basis, the time series is highly volatile in nature, that is, the residuals plot of the initially fitted show some kind of pattern, or reveal the presence of some irregular variation among the residuals (heteroscedastic residuals), which clearly implies that our initial fitted model is not quite appropriate for forecasting and it should be improved. So, this indicates the need for the second objective where another model is to be used to take care of this Varying- Variance among the residuals of the initially fitted model. Thus after building the appropriate model for the heteroscedastic residuals we move forward as usual to check the validity of the second proposed model as in the previous case.

Now, as our second objective gets fulfilled, we now proceed towards the validation of the combined model and use it for forecasting and compare it the previously fitted ARlMA model. Thus, finally, the third objective will be to combine the two proposed models into one and check for its validity. If all the diagnostic measures prove that the combined model is valid, we can use the final combined for forecasting of the future values.

**Chapter 2**

**Methodology**

**Chapter 2 - Methodology**

The methodologies include the techniques and tools used in the analysis of the stock prices. The main models that are used in this project to fit and forecast the stock prices of the National Stock Exchange (NSE) the Auto – Regressive Integrated Moving Average (ARlMA) model and Generalized Auto Regressive Conditional Heteroscedasticity model.

Due to the presence of high volatility in the stock prices taken on a daily basis and as a result, presence of heteroscedasticity, i.e. changing variances in the residuals (we will show during the analysis), I will make use of the Generalized Auto - Regressive Conditional Heteroscedasticity (GARCH) model to take care of the nuisance. At last, I will finally consider the whole ARlMA + GARCH model and perform the relevant diagnostic checking to judge the proposed model.

The general forms and concepts of the ARlMA model and the GARCH model are given in the next few pages along some basic assumptions of forecasting.

Some of the tests that are used here for various purposes are Augmented Dickey - Fuller test for sting stationarity, Ljung - Box test for testing the presence of independence (autocorrelation) in the residual series, etc. Short descriptions are given on these tests in the next pages.

The different plots that are used here are Time – series plots, Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plots, Residual plots, Forecasting plots, Normal Quantile- Quantile plots (Q Q plots) of the residuals, Conditional Heteroscedasticity plots, etc.

All required calculations, construction of different plots, model buildings, etc. are done using the R Statistical Software. The packages that are used to perform the relevant actions are “tseries”,”lmtest”, “forecast”, “ timeseries”, “fGarch”, etc.

**2.1- ARIMA Model:**

ARIMA stands for Autoregressive Integrated Moving Average. ARIMA is also known as Box-Jenkins approach. Box and Jenkins claimed that non-stationary data can be made stationary by differencing the series, Yt. The general model for Yt is written as,

**Yt =ϕ1Yt−1+ϕ2Yt−2…ϕpYt−p +ϵt + θ1ϵt−1+ θ2ϵt−2 +…θqϵt−q**

Where Yt is the differenced time series value, ϕ and θ are unknown parameters and ϵ are independent identically distributed error terms with zero mean. Here,Yt is expressed in terms of its past values and the current and past values of error terms.

The ARIMA model combines three basic methods:

* **Auto Regression (AR)** – In auto-regression, the values of a given time series data are regressed on their own lagged values, which are indicated by the “p” value in the ARIMA model.
* **Differencing (I-for Integrated)** – This involves differencing the time series data to remove the trend and convert a non-stationary time series to a stationary one. This is indicated by the “d” value in the ARIMA model. If d = 1, it looks at the difference between two-time series entries, if d = 2 it looks at the differences of the differences obtained at d =1, and so forth.
* **Moving Average (MA)** – The moving average nature of the ARIMA model is represented by the “q” value which is the number of lagged values of the error term.

This model is called Autoregressive Integrated Moving Average or ARIMA(p,d,q) of Yt.  We will follow the steps enumerated below to build our model.

**2.1.2 -Step 1: Testing and Ensuring Stationarity**

To model a time series with the Box-Jenkins approach, the series has to be stationary. A stationary time series means a time series without trend, one having a constant mean and variance over time, which makes it easy for predicting values.

**Testing for stationarity –**We test for stationarity using the Augmented Dickey-Fuller unit root test. The p-value resulting from the ADF test has to be less than 0.05 or 5% for a time series to be stationary. If the p-value is greater than 0.05 or 5%, you conclude that the time series has a unit root which means that it is a non-stationary process.

**2.2.1 - Differencing –**To convert a non-stationary process to a stationary process, we apply the differencing method. Differencing a time series means finding the differences between consecutive values of a time series data. The differenced values form a new time series dataset which can be tested to uncover new correlations or other interesting statistical properties.

We can apply the differencing method consecutively more than once, giving rise to the “first differences”, “second order differences”, etc.

We apply the appropriate differencing order (d) to make a time series stationary before we can proceed to the next step.

**Step 2: Identification of p and q**

In this step, we identify the appropriate order of Autoregressive (AR) and Moving average (MA) processes by using the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF).  Please refer to our blog, “Starting out with Time Series” for an explanation of ACF and PACF functions.

**Identifying the p order of AR model**

For AR models, the ACF will dampen exponentially and the PACF will be used to identify the order (p) of the AR model. If we have one significant spike at lag 1 on the PACF, then we have an AR model of the order 1, i.e. AR(1). If we have significant spikes at lag 1, 2, and 3 on the PACF, then we have an AR model of the order 3, i.e. AR(3).

**Identifying the q order of MA model**

For MA models, the PACF will dampen exponentially and the ACF plot will be used to identify the order of the MA process. If we have one significant spike at lag 1 on the ACF, then we have an MA model of the order 1, i.e. MA(1). If we have significant spikes at lag 1, 2, and 3 on the ACF, then we have an MA model of the order 3, i.e. MA(3).

**Step 3: Estimation and Forecasting**

Once we have determined the parameters (p,d,q) we estimate the accuracy of the ARIMA model on a training data set and then use the fitted model to forecast the values of the test data set using a forecasting function. In the end we cross-check whether our forecasted values are in line with the actual values.

**2.2.2- Volatility of the Stock Prices:**

Investors in the stock markets are obviously interested in the volatility of the stock prices, for high volatility could mean huge losses or gains and hence results in greater uncertainty. In volatile markets, it is very difficult for the companies to raise capital in the capital markets. The concept of modelling the varying variance due to volatility gave rise to the so-called Auto Regressive Conditional Heteroscedasticity Model which was originally developed by Engle.

**2.2.3 -GARCH model:**

GARCH processes differ from homoskedastic models, which assume constant volatility and are used in basic ordinary least squares (OLS) analysis. OLS aims to minimize the deviations between data points and a regression line to fit those points. With asset returns, volatility seems to vary during certain periods of time and depend on past variance, making a homoskedastic model not optimal.

GARCH processes, being autoregressive, depend on past squared observations and past variances to model for current variance. GARCH processes are widely used in finance due to their effectiveness in modeling asset returns and inflation. GARCH aims to minimize errors in forecasting by accounting for errors in prior forecasting and, thereby, enhancing the accuracy of ongoing predictions.

**2.2.4- When should we choose ARCH / GARCH model?**

Firstly, we should check if residual plot of the chosen ARIMA model displays any cluster of volatility. Next, we should observe the squared returns plot. If yes, ARCH / GARCH should be used to model the volatility to reflect more recent changes and fluctuations in the series. Finally, ACF & PACF plots of the squared residuals will help to confirm if the residuals (noise term) are not independent and can be predicted.

**2.2.6 Ljung-Box Test for residuals:**

The Ljung-Box Test is used to check the independence of the returns extracted from the chosen model. In other words, it is used to reveal the absence or presence of Auto-Correlation in the Returns with the help of the p-value of the test. The hypothesis for this test is,

Ho : Independence (No AC) vs H1 : Dependence (AC)

Here, AC refers to Auto-Correlation. Thus, if p-value for a certain lag value of the test is greater than α= 0.05, we accept Ho and conclude that the Returns are independent and they have no Auto-Correlation. Otherwise, we conclude that the returns are Auto-Correlated.

CHAPTER 3

DATA ANALYSIS

AND

INTERPRETATIONS

**CHAPTER 3 -DATA Analysis and Interpretations**

For the modelling, analysis and forecasting we have considered the S&P BSE SENSEX share price data of the Bombay Stock Exchange extracted from the ‘BSE INDIA’ website. The data contains the columns namely Date, Open, High, Low, Close, Adjusted Close. In our discussion, we are only considering the Adjusted Close price staking into account that the rest of the price scan also be modeled in the similar manner.

The R Statistical Software is used to do the analysis.

**3.1 Data Extraction Layout and Cleaning in R :**

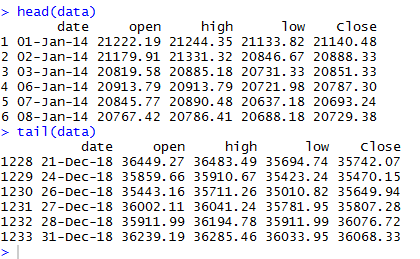
This part deals with the extraction of data on the Adjusted Closing Prices from the original source and revealing the layout and making it suitable for analysis.

**3.1.1 Extracting the data into Microsoft Excel and Reading it in R :**

We extracted the historical data of the S&P BSE SENSEX from the BSE INDIA website by downloading it. It will get saved in an Excel file. Then, we read the data into the R platform to do the rest of the work. The R- code is given below.

|  |
| --- |
| # Extracting the data into ms excel and reading it in R  data= read.csv(file.choose(),header=T)  head(data)  View(data)  tail(data) |

**3.1.2 – Data Description and Layout:**



**Here the BSE SENSEX data layout is given which includes the first oldest 6 observations from 1"January, 2014 and the latest observations up to 31st December, 2018. The date is in the format MM-DD-YYYY and currency of the prices is INR (Indian Rupee). In order to extract**

**The full data, one may open the website BSE INDIA to get access to the data or simply refer to the link or URL given below and extract the desired data.**

[**https://www.bseindia.com/indices/IndexArchiveData.html**](https://www.bseindia.com/indices/IndexArchiveData.html)

The data that is used in this analysis is the Yahoo Share Price Data of the last 5 years. It is actually the daily data of the Bombay Stock Exchange of the BSE SENSEX stock market index from 15January, 2013 to 31StDecember,2017 extracted from the BSE INDIA website. This data set contains the date, open prices, high prices, low prices, close prices and adjusted close prices of the BSE SENSE throughout these five years on a daily basis. To achieve consistency, the adjusted close prices are used as a general measure of stock price of BSE SENSEX over the past five years.

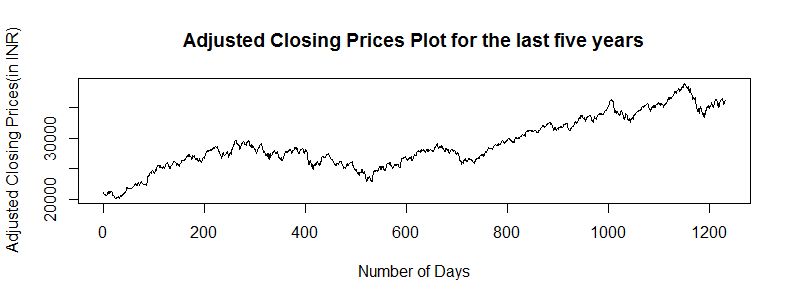
**3.1.3 -Cleaning and Plotting the Adjusted Closing prices:**

After reading the data in R platform, I deleted the null or the unavailable values in the original data and then extract the numerical values of the Adjusted Closing prices from the raw data. Then, the values which are not available (NA values) are removed from the Adjusted Closing prices. Then, the time series plot of the cleaned Adjusted Closing values is constructed.

The R- codes for the above is given below.

|  |
| --- |
| # Extracting, cleaning and plotting the adjusted closing prices  d=na.omit(data)  sp=data.frame(data$Close)  adj\_cl\_values=as.numeric(paste(d$Close))  sp=adj\_cl\_values[!is.na(adj\_cl\_values)]  plot.ts(sp,xlab="Number of Days",ylab="Adjusted Closing Prices(in INR)",  main="Adjusted Closing Prices Plot for the last five years") |

Using the above R code the following time series plot is generated. The time series plot shows the behavior of the Adjusted Closed Prices for the last 5 years from 2014 to 2018 - Diagram 3.1



From the above time series plot, we can say that the data on the adjusted closing prices of the BSE SENSEX are not in the stationarity condition graphically. Further, we will do a stationary test known as the Augmented Dickey-Fuller test to test the stationarity on a theoretical basis.

**3.2 -:** **Stationarity and Transformations:**

3.2.1- Non-Stationarity of the original Adjusted Closing Price data:

We check for the stationarity of the Adjusted Closing Prices using the Augmented Dickey Fuller test using a=5%=0.05 level of significance in R Statistical Software. The hypothesis to be tested is,

Ho: Non -Stationarity vs H1: Stationarity

The R code for the above test is given below.

|  |
| --- |
| ## NON-STATIONARITY OF THE ORIGNAL DATA ##  print(adf.test(sp)) # Augmented Dickey fuller test to test stationary |

We found the following information on running the R code Table 3.2

|  |  |  |
| --- | --- | --- |
| **Dickey Fuller Test Statistic** | **Lag order** | **P-value** |
| -2.1344 | 10 | 0.5214 |

Clearly, the p-value of the test is 0.5214 which is greater than the level of significance, α=0.05. Thus, we accept Ho at 5% level of significance and conclude that the Adjusted Closing Price time series is non-stationary.

3.2.2 - Transformation of the original data: From the Augmented - Dickey Fuller test, it is found that the original Adjusted Closing Prices data is not stationary. Thus, the data is not quite appropriate to use it for analysis or forecasting purpose. The original data can be transformed the into a more reliable data using logarithm transformation to make it linear and also differencing of the data is required to make the data stationary. The transformation of the data is done using R Statistical software.

The relevant R- codes are given below

|  |
| --- |
| # Transformation of the data applying log and differencing to make the data stationary and linear  d\_sp=diff(sp)  log\_sp=log(sp)  log\_sp2=(log(sp))\*\*2  d\_log\_sp=diff(log\_sp,lag=1) |

Thus, the original Adjusted Closing prices are converted into logarithm of the original prices. Differencing is also done to the original prices and again logarithm is applied on the differenced prices to obtain the Differenced Log Adjusted Closing prices.

**3.2.3 Time Series Plots:**

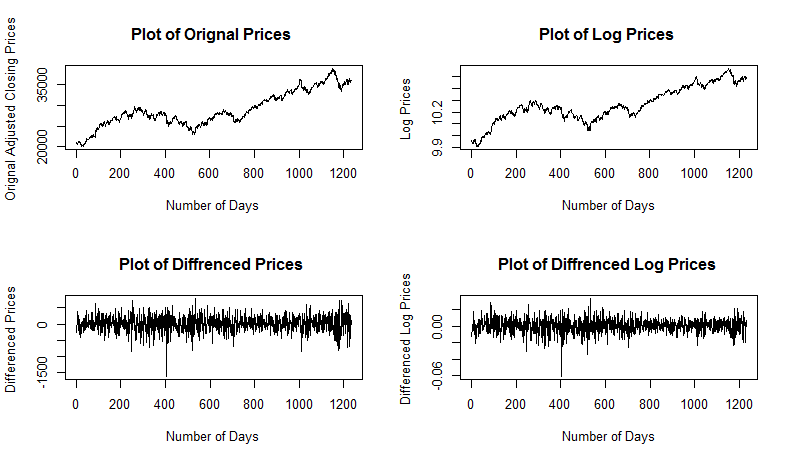
The time series plots of the Original Prices, the Log Prices, the Differenced Prices and the Differenced Log Prices with respect to the number of days generated using R software are constructed to reveal the behavior of each of the transformed prices.

The relevant R – codes are given below.

|  |
| --- |
| # Time series plots    par(mfrow=c(2,2))  plot.ts(sp,main="Plot of Orignal Prices ", xlab="Number of Days",ylab="Orignal Adjusted Closing Prices ")  plot.ts(log\_sp,main="Plot of Log Prices ", xlab="Number of Days",ylab="Log Prices ")  plot.ts(d\_sp,main="Plot of Diffrenced Prices ", xlab="Number of Days",ylab="Differenced Prices ")  plot.ts(d\_log\_sp,main="Plot of Diffrenced Log Prices ", xlab="Number of Days",ylab="Differenced Log Prices ") |

Thus, running the R codes, the time series plots for different forms of the Adjusted Closed prices are obtained as given below,

The time series plots are given below – Diagram (3.2)



* The **upper left graph** is the original time series if Adjusted Closing Stock Prices from 01/01/2014 to 31/12/2018, showing the actual growth.
* The **upper right graph** shows the graph of logarithm of the Adjusted Closing Stock Prices. The series is more linear compared to the original one.
* The **lower left graph** shows the differences of Adjusted Closing Stock Prices. It can be seen that the series is price-dependent in other words, the variance of the series increases as the level of original series increases, and therefore, is not stationary.
* The **lower right graph** shows the differences of the logarithm of the Adjusted Closing Stock Prices. The series seems more and variance is constant and does not significantly change as level of original series changes. Thus, the series is stationary.

**3.2.4- Stationarity of the Differenced Log Adjusted Closing Price data:**

We checked for the stationarity of the Difference of the Logarithm of the Adjusted Closing Prices using the Augmented Dickey Fuller test using a=5%=0.05 level of significance in R Statistical Software. The hypothesis to be tested is,

Ho : Non- Stationarity vs H1 : Stationarity

The R - codes for the above test are given below.

|  |
| --- |
| #Stationarity of the diffrenced log transformed data  print(adf.test(d\_log\_sp)) |

We found the following information on running the R codes Table 3.3

|  |  |  |
| --- | --- | --- |
| Dickey Fuller Test Statistic | Lag order | P-value |
| -10.615 | 10 | 0.01 (<0.05) |

Clearly, the p-value of the test is 0.01 which is lesser than the level of significance, a=0.05. Thus, We reject Ho at 5% level of significance and conclude that the Difference of the Logarithm of the Adjusted Closing Price time series is stationary in nature. Thus, this series is stationary and thus is suitable to use for analysis. Another point that makes the differences of time series data more interesting than the original stock prices series is that many investigators often look at the returns of the stocks rather than its original prices. Differences of the log stock prices represent the returns and are similar to percentage changes of the stock prices.

**3.3 - ARIMA Model Fitting:**

**3.3.1 Model Identification:**

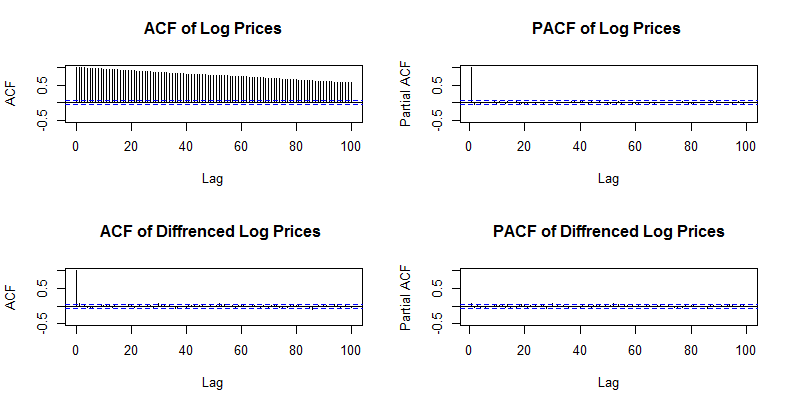
The Auto-Correlation Function (ACF) Plots and the Partial Auto-Correlation Function (PACF)

Plots of the Log Adjusted Closing Prices and the Differenced Log Adjusted Closing Prices are constructed using R Software.

The R - code is given below.

|  |
| --- |
| ##ACF AND PACF PLOTS##  par(mfrow=c(2,2))  acf(log\_sp,main="ACF of Log Prices",lag.max=100,ylim=c(-0.5,1))  pacf(log\_sp,main="PACF of Log Prices ",lag.max=100,ylim=c(-0.5,1))  acf(d\_log\_sp,main="ACF of Diffrenced Log Prices ",lag.max=100,ylim=c(-0.5,1))  pacf(d\_log\_sp,main="PACF of Diffrenced Log Prices",lag.max=100,ylim=c(-0.5,1)) |

The different ACF and PACF plots are thus constructed using the R –codes are given and are displayed in the given plot. – Diagram (3.3)



Clearly, from the ACF and the PACF plots of the Log Adjusted Closed Prices, we observe that the ACF is decaying off gradually which implies that there will be an Auto-Regressive term in the model, and further there is a significant spike in the PACF plot and other spikes are insignificant which reveals the present of a Moving Average term in the model. Now, using the values of the Akaike Information Criteria of different models, we will confirm our claim and find the exact model which will appropriately model the price data.

**MODEL IDENTIFICATION**:

Now we use a R function which automatically selects the best ARIMA model for the data.

The R code is given below

|  |
| --- |
| arima\_out=auto.arima(log\_sp,trace=F)  arima\_out  arima111 = arima\_out |

From the above code of ARIMA model considering the Akaike Information Criteria (AIC), we got an automated model of ARIMA(p=1,d=l, q=l) with a drift gives the least AIC value among all the other models. Thus, we choose the ARIMA(1, l, 1) model with drift to model the logarithm of the Adjusted Closed Prices.

**3.3.2 Parameters Estimation:**

To estimate the parameters of the above selected model, we will implement the R code. The result will provide the estimate of each element of the model containing one Auto-Regressive term (p=1) and one Moving Average term (q=l) with a single differencing (d=l).

The R - codes for building the ARIMA (1, l, 1) model and testing the significance of the parameters are given.

|  |
| --- |
| #parameters estimation  summary(arima111)  coeftest(arima111) |

The coefficient, standard errors, z-values and p-values of the model are provided below in a tabular form- **Table (3.5)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Co-eff(s) | Estimate | Std. Error | Z-value | P-value |
| ar1 | -0.57969632 | 0.15686042 | -3.6956 | 0.0002194\*\*\* |
| ma1 | 0.65876704 | 0.14421009 | 4.5681 | 4.922e-06\*\*\* |
| Drift | 0.00043366 | 0.00025088 | 1.7285 | 0.0838897 |

From the z values and p values of the coefficients of the AR term and the MA term given in the above table, we can infer that the estimates of the parameters of the fitted ARIMA model are significant. But the same imply that the driit is not significant (at 5 % level of significance).

Now, we will write the full fltted model.

Let Yt be the actual log transformed Adjusted Closed Prices.

The ARIMA (1, 1, 1) model is given by,

Yt- Yt-1 = µ + ф(Y t-1 – Yt-2) + θɛt-1 +ɛ

Thus, the parameters of the above model are estimated as,

ф = -0.57969632 ; θ = 0.65876704 ; µ (drift) = 0.00043366

Let Yt be the fitted log transformed Adjusted Closed Prices.

Thus, the full fitted ARIMA(1,1,1) model is given by,

|  |
| --- |
| Yt-Yt-1= 4.3366 \*10-4 – 0.57969632(Yt-1 –Yt-2 ) + 0.65876704ɛt-1 |

**3.4 -Forecasting and Forecasting Plots:**

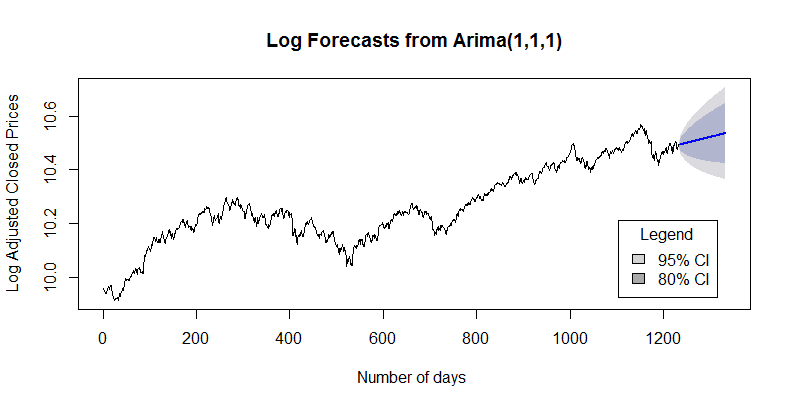
Alter estimating the parameters and writing the full model in the previous section, we now proceed towards the forecasting of future values using the fitted ARIMA (1, 1, 1) model. The forecast values from the chosen model will be in the logarithm form. To get the original prices we will take the exponential of log of the forecasted values.

The R codes for obtaining the forecast values for 100 days and also to obtain the Log -Forecast plot are given below.

|  |
| --- |
| #forecasting and forecasting errors  log\_forecast=forecast(arima111,100)  log\_forecast  log\_fc=as.numeric(log\_forecast$mean)  fc=as.numeric(log\_forecast$mean)  par(mfrow=c(1,1))  plot(log\_forecast,main="Log Forecasts from Arima(1,1,1)",xlab = "Number of days ",ylab = "Log Adjusted Closed Prices")  legend("bottomright",inset=0.05,title="Legend",c("95% CI","80% CI"),fill=c("light Grey","Dark Grey"))  accuracy(log\_forecast) |

Thus, using the above R-code, we will get the log forecast values as well as the original forecast values (by taking exponential of the log values) for the future, say for period of 100 days.

The Log-Forecast plot is given- Diagram (3.4)



The forecast plot of the logarithm of the Adjusted Closed Prices is given above. The part in shades shows the forecasted values.

**The “Darker shades” show the 80 % Confidence Interval of the forecast prices.**

**The “Lighter shades” show the 95 % Confidence Interval of the forecast prices.**

**3.4.1 Forecasting Errors:**

Using the R codes, we got the various forecasting errors.

|  |  |
| --- | --- |
| Mean Error (ME) | 7.884046e-06 |
| Root Mean Squared Error (RMSE) | 0.008333423 |
| Mean Absolute Error (MAE) | 0.006208074 |
| Mean Percentage Error (MPE) | 6.226511e-05 |
| Mean Absolute Percentage Error (MAPE) | 0.06057537 |
| Mean Absolute Scaled Error (MASE) | 0.9915398 |
| Auto-Correlation of Errors at Lag1 (ACF1) | 0.00225678 |

**3.4.2 Comparison of the Original and the Fitted Stock Prices:**

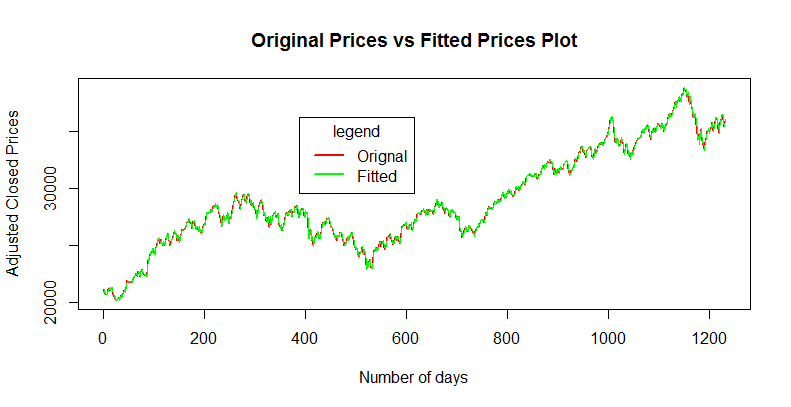
Now, in order to compare the Original Adjusted Closing prices with the fitted Adjusted Closing prices, we will plot both the Original and the Fitted Adjusted Closed Prices obtained from the model in a single graph to check the fit of the chosen model graphically using R.

The R - codes for obtaining the fitted values and the required plot for comparing the fitted and the original prices are given below.

|  |
| --- |
| # Comparision of the original and the fitted stock prices  log\_fit\_values=fitted (arima111)  fit\_Values=exp(log\_fit\_values)  par(mfrow=c(1,1))  plot.ts(sp,col="Red",main="Original Prices vs Fitted Prices Plot",xlab="Number of days",ylab="Adjusted Closed Prices")  lines(fit\_Values,col="green")  legend("bottomright",lty=c(1,1),lwd=2,inset=0.5,title="legend",c("Orignal","Fitted"),col = c("Red","Green"),bty = "o") |

Later, we will proceed towards the diagnostic checking of the residuals of the fitted model using various diagnostic measures, plots and test and verify that whether the fitted ARIMA model is valid can be used for reliable forecasting of the fixture values.

The plot of Original Adjusted Closing prices and Fitted Adjusted Closing prices obtained from the fitted model is given below **Diagram (3.5)**



From the plot of the Original Adjusted Closed Prices vs the Fitted Adjusted Closed Prices obtained from the chosen ARIMA (l, l, 1)model, we can observe that plot of the data points are overlapping on each other. Thus, we can say that the fitted values of the model chosen give a very good fit and less error when plotted simultaneously.

**3.5 Diagnostic Checking- of the Residuals of the ARIMA model:**

The diagnostic checking of the chosen model is done to check whether the residuals of the model are homoscedastic. It also validates whether the fitted ARIMA model can be used for forecasting of the future values.

It includes observing the Residuals Plot and its ACF & PACF plots and interpreting them, checking the Quantile-Quantile Plot for Normality of the residuals and also the squared residuals and finally executing the L-jung Box test of the residuals and the squared residuals to conclude that whetherthe residuals of the fitted model are truly random.

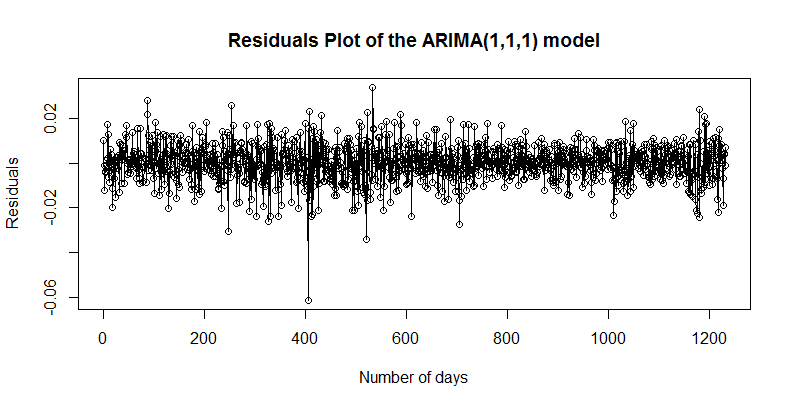
**3.5.1 - Residuals Plot:**

We first extract the residual values from the ARIMA (l, 1, 1) model using relevant R code and then plot the Residuals in the Y-axis against the Number of days in the X-axis known as the Residuals plot.

The R – codes for obtaining the Residual Plot are given below.

|  |
| --- |
| #residual plot  res=resid(arima111)  par(mfrow=c(1,1))  plot(res,type="o",main="Residuals Plot of the ARIMA(1,1,1) model",xlab="Number of days", ylab="Residuals") |

The Residual Plot of the model is given**- Diagram (3.6)**



From the Residual Plot, we can clearly observe that the residuals of the chosen model do not follow any pattern and are more or less random in nature. Thus, observing the Residual Plot we can conclude that the chosen ARIMA model is good and can be used for forecasting of the future values. Mainly, it is to be used for short -term forecasting.

However, 1t 13 to be noted that here in the Residual plot of the residuals of the fitted ARIMA model, there are some traces of difference 1n the variation of the residuals of the fitted ARIMA model, that is, there is a question of randomness among the residuals of the fitted ARIMA model and clearly the randomness of the same is not even all over the plot. Hence, this phenomenon clearly indicates me presence of some kind of volatility or heteroscedasticity in the residuals of the fitted ARIMA model and this nuisance is to be taken care of.

To make our claim stronger, we will further plot the Auto- Correlation Function (ACF) and the

Partial Auto Correlation Function (PACF) of the residuals of the fitted model.

**3.5.2 - ACF & PACF plots of the Residuals:**

The Auto- Correlation Function (ACF) plot and the Partial Auto-Correlation Function (PACF) plot

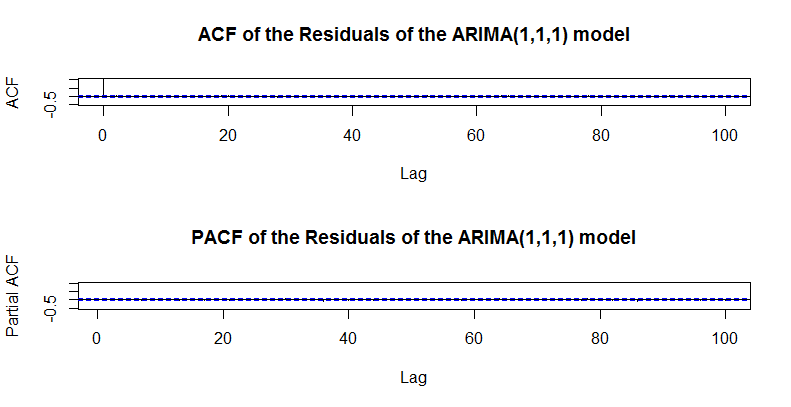
of the extracted residuals from the chosen ARIMA (1, 1, 1) model are built using R.

The relevant R codes are given below.

|  |
| --- |
| #Acf and Pacf plots  par(mfrow=c(2,1))  acf(res,main="ACF of the Residuals of the ARIMA(1,1,1) model",lag.max=100,ylim=c(-0.5,1))  pacf(res,main="PACF of the Residuals of the ARIMA(1,1,1) model",lag.max = 100,ylim=c(-0.5,1)) |

The ACF and PACF Plots of the residuals obtained from the fitted ARIMA (1,1,1) model are given- Diagram (3.7)

The ACF and the PACF of the residuals obtained from the fitted ARIMA (1,1,1) model are given –**Diagram (3.7)**



The above diagrams are the ACF and the PACF plots of the Residuals of the chosen ARIMA model. It is to be noted that there is only one correlation which is higher than the significant value in the ACF plot of the residuals, and the remaining do not cross the significant limits in both the ACF and the PACF plots of the residuals. Thus, as a whole, we can conclude that there is no significant evidence of the presence of dependence in the residuals of the chosen model.

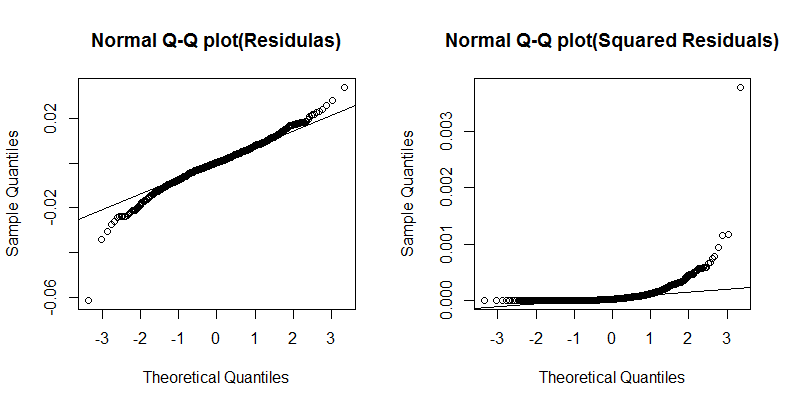
**3.5.3- QQ-Plot of the Residuals and Squared Residuals**

We will see the quantile-quantile plot (Q-Q plot) of the residuals and the square of the residuals to check the presence of auto-correlation.

The R codes for constructing the QQ Plot are given below.

|  |
| --- |
| #QQ PLOT  par(mfrow=c(1,2))  qqnorm(res,main="Normal Q-Q plot(Residulas)")  qqline(res)  qqnorm(res^2,main="Normal Q-Q plot(Squared Residuals)")  qqline(res^2) |

The Normal Q-Q plots are given below- Diagram (3.8)



From the previous two Quantile –Quantile plots, we find that the Residuals of the fitted model look like white noise. But there is a clear evidence of Auto-Correlation in the Squared Residuals. Next, we will use the Ljung-Box test to confirm our findings.

**3.5.4.Ljung Box Test:**

As mentioned earlier, the LJung Box test is used to check the absence or presence of Auto- Correlation in the residuals of particular model.

The hypothesis of interest for the Ljung-Box test is,

H 0 : Independence(No AC) vs H1: Dependence(AC)

Here AC refers to Auto-Correlation.

The R- code for the above Ljung- Box test for both the residuals and squared residuals are given below.

|  |
| --- |
| #Ljung- Box test  Box.test(res,lag=10,type="Ljung-Box")  Box.test(res^2,lag=10,type="Ljung-Box") |

Using R, we get the following information regarding the Residuals using the Ljung-Box test- **Table (3.7)**

|  |  |  |  |
| --- | --- | --- | --- |
| Test | X-squared | Dfs | p-value |
| Ljung –Box | 5.7288 | 10 | 0.8375 |
| Ljung-Box | 63.238 | 10 | 8.795e-10 |

Clearly, at lag 1, the p-value of the testisslightlygreaterthan0.05, but after that as the lags increase, the p-value significantly decreases and gets close to 0, as a result of which we have enough evidence to reject Ho and conclude that, there is a presence of Auto-Correlation among the residuals. In other words, the fitted ARIMA model suffers from the problem of Heteroscedasticity (Unequal variances) which is not desired and should be taken care of.

**3.6 - Presence of Volatility and need for GARCH Modelling:**

Although ACF & PACF of returns of the chosen ARIMA model have no overall significant lag, the time series plot of returns shows some cluster of volatility. It is important to remember that ARIMA is a method to linearly model the data and the forecast width remains constant because the model does not reflect recent changes or incorporate new information. In other words, it provides best linear forecast for the series, and thus plays little role in forecasting model nonlinearly. In order to model volatility, ARCH / GARCH method comes into play.

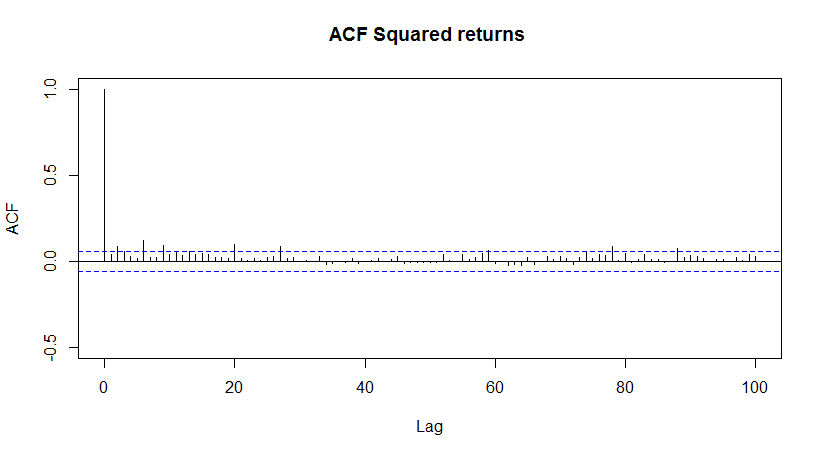
Now, we will study the Auto Correlation Function (ACF) plot and Partial Auto Correlation plot (PACF) plots of the Squared Returns of the previously fitted ARIMA model. These two plots will make us sure about the presence of heteroscedasticity among the returns of the fitted model and hence will ensure the presence of volatility in the time series data of the Adjusted Closing Prices that we are concerned with.

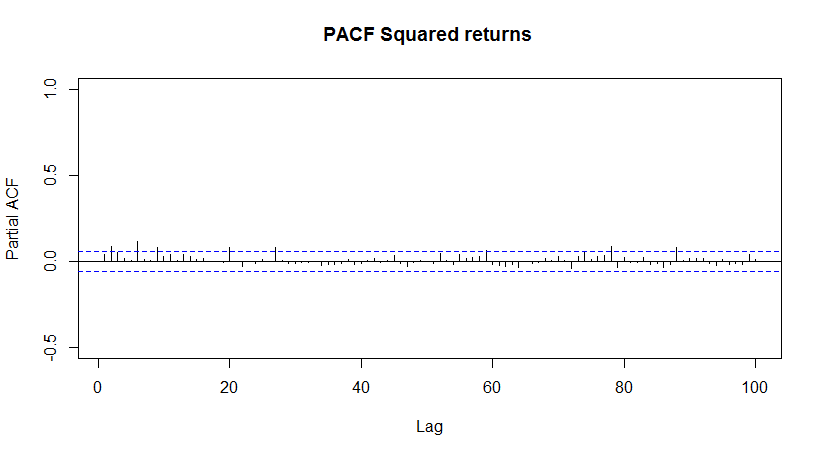
The R - codes for constructing the ACF and PACF plots of the Squared Returns are given below.

|  |
| --- |
| #ACF AND PACF PLOTS FOR SQUARED RETURNS  acf(d\_log\_sp^2,main="ACF Squared returns ",lag.max = 100,ylim=c(-0.5,1))  pacf(d\_log\_sp^2,main="PACF Squared returns ",lag.max = 100,ylim=c(-0.5,1)) |

The ACF and the PACF plots of the Squared Returns are constructed up to a lag of 100. If the Auto Correlation Function (ACF) spikes in the Auto - Correlation Function Plot exceed the significance line (i.e. the blue dotted line) in the plot, then it will reveal the presence of dependence among the residuals of the fitted ARIMA model.

The ACF and PACF plots of the squared Residuals obtained by running the R-codes are given below-Diagram(3.11)





From the above ACF and PACF plots of the Squared Returns, we can clearly see that there are some significant spikes which exceed the significant line (i.e. the blue dotted line in the plot. Thus, we can say that the returns are dependent on each other and they reveal hetroscedastic structure of the model.

**Thus, we conclude that the residuals show some patterns that might be modeled. ARCH /GARCH is necessary to model the volatility of the series. As indicated by their names (ARCH/ GARCH), those models are used to model the Conditional Variance or Conditional Heteroscedasticity present in the series.**

**3.7 - ARCH / GARCH model fitting:**

**3.7.1 - Model Identification:**

As mentioned earlier, that there are traces of volatility in the series conclude during the diagnostic checking of the returns of the fitted ARIMA model, we now proceed towards the identification of the ARCH / GARCH model and build a proper model. The orders and parameters of the ARCH / GARCH model are selected based on the Akaike Information Criteria (AlC) values as done earlier in cases of ARIMA model identification.

Thus, the values of the AIC for different models are calculated using R, and are given below. It is to be remembered that we choose the least value among the rest. The formula for calculating the Akaike Information Criteria values is,

***AIC = -2 \* Log -likehood \*(q + 1) \*(N/(N q 1))***

Here, Log likelihood is the log likelihood value of the fitted model.

q is the number of terms that will be in the model,

N is the total number of observations.

**The AIC values for different ARCH/GARCH models including the number of terms in the model and the log –likelihood values are calculated using R. The R –code for the same are given below.**

|  |
| --- |
| #Model Identification  z1=garch(d\_log\_sp^2,order=c(0,1))  z2=garch(d\_log\_sp^2,order=c(0,2))  z3=garch(d\_log\_sp^2,order=c(0,3))  z4=garch(d\_log\_sp^2,order=c(0,4))  z5=garch(d\_log\_sp^2,order=c(0,5))  z6=garch(d\_log\_sp^2,order=c(0,6))  z7=garch(d\_log\_sp^2,order=c(0,7))  z8=garch(d\_log\_sp^2,order=c(0,8))  z9=garch(d\_log\_sp^2,order=c(0,9))  garch11=garch(d\_log\_sp^2,order=c(1,1))  log\_lik= c(logLik(z1),logLik(z2),logLik(z3),logLik(z4),logLik(z5),logLik(z6),logLik(z7),logLik(z8),logLik(z9),logLik(garch11))  AIC=-2\*log\_lik+2\*(q+1)\*n/(n-q-1) |

The AIC values for the different ARCH/GARCH models including the number of terms in the model and the log- likelihood values the above R-codes are given below in tabular form – **table(3.9)**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | No of terms | Log- Likelihood | AIC |
| ARCH(1) | 1 | 8930.061 | -17856.12 |
| ARCH(2) | 2 | 9076.630 | -18147.25 |
| ARCH(3) | 3 | 9073.299 | -18138.57 |
| ARCH(4) | 4 | 8921.291 | -17832.54 |
| ARCH(5) | 5 | 8911.213 | -17810.37 |
| GARCH(1,1) | 2 | 8930.025 | -17837.85 |

It is to be kept in mind that too many terms should not be included in the ARCH / GARCH model. Thus, from the above table we see that, the z2 model that is the GARCH (0, 2) has the least AIC value among the other models. Thus, we choose the GARCH (0,2) model to explain the volatility in the data.

**Parameters Estimation:**

Using R, we built the GARCH (0,2) model. The R- codes are given.

|  |
| --- |
| # PARAMETER ESTIMATION  z2=garch(d\_log\_sp^2,order=c(0,2))  summary(z2) |

Estimates of z2

α 0 = 1.113e-08 α 1 = 2.506e-02 β 1 = 2.955e+00

The summary of the chosen GARCH (0,2) model including the estimates are significant indicating that they are all statistically independent In addition, the p- value of the Ljung-Box test is greater than 0.05 indicating the acceptance of the null hypothesis that there is no Auto-Correlation among the returns of the fitted GARCH. Thus, full fitted GARCH (0,2) model is given by,

|  |
| --- |
| ht = 1.113e-08 + 2.506e-02\* ɛ2t-1+2.955e+00\*ht-1 |

Thus, we can see that our fitted model consists a constant, a lagged term of the square of the returns of the previously fitted ARIMA (1, 1, 1) model and a lagged term of the variance itself further, we will study the plots of the conditional standard deviations and variances, and also perform the diagnostic checking of the chosen GARCH (0, 2) model.

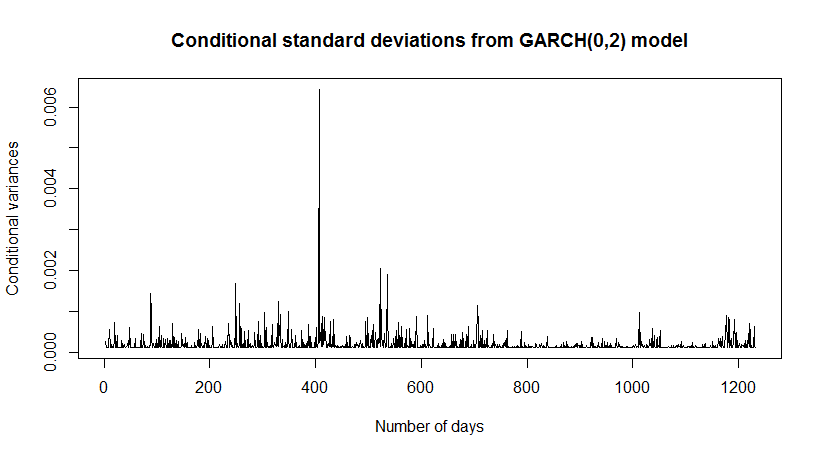
**3.8 - Fitted GARCH Model and Conditional Heteroscedasticity:**

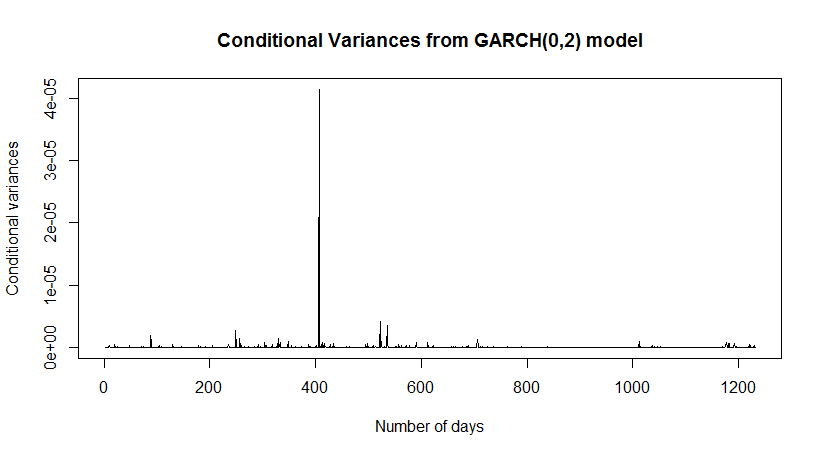
We have estimated the parameters of the fitted GARCH (0, 2) model on the residuals of the fitted ARIMA (1, 1,1) model in the previous part using R software. Now, we will extract the fitted values of the GARCH model using R which are nothing but the Conditional Standard Deviations, and squaring them, we will get the Conditional Variances from the model.

**The R codes for building the Conditional Heteroscedasticity Plots that includes plotting both the fitted Conditional Standard Deviations and Variances obtained from the GARCH model are given below.**

|  |
| --- |
| # CONDITIONAL HETEROSCEDASTICITY  par(mfrow=c(2,1))  sigt.z2=z2$fit[,1]  plot(sigt.z2,main="Conditional standard deviations from GARCH(0,2) model",xlab= "Number of days",ylab="Conditional variances",type='l')  ht.z2=z2$fit[,1]^2  plot(ht.z2,main="Conditional Variances from GARCH(0,2) model",xlab= "Number of days",ylab="Conditional variances",type='l') |

The Conditional Standard Deviation Plot and the Conditional Variances Plot obtained from the GARCH model are given below- **Diagram (3.12)**





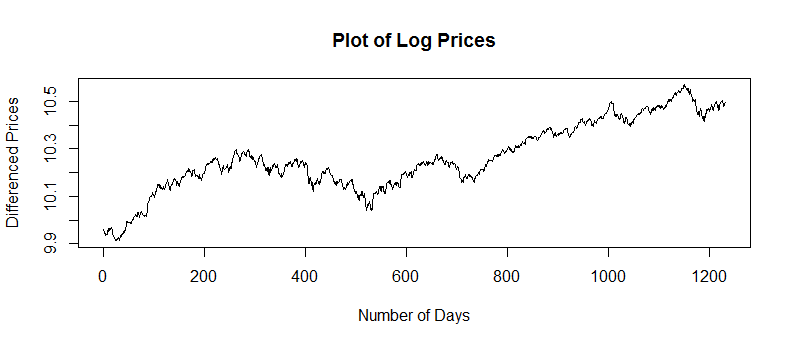
Thus from the given two plots, we can clearly see irregular spikes occurred which are actually the evidence of presence of several clusters of volatility in the daily Adjusted Closing Stock Prices. Thus, clearly there is a presence of Conditional Heteroscedasticity (or Varying Variance) in the time series that is under construction.

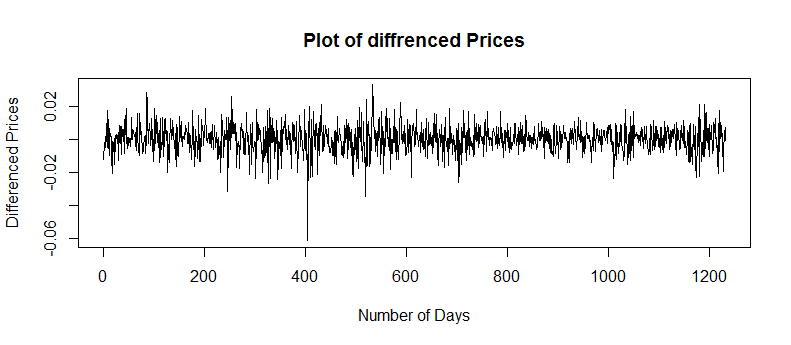
Let us now observe the probable time periods wherever the volatility in the Adjusted Closing Prices had occurred by comparing the original time series plots with the Conditional Heteroscedasticity plots. The time series plot of the log transformed Adjusted Closed Prices (top), the time series plot of the differenced log transformed Adjusted Closed Prices (middle) and the Conditional Variances plot (bottom) are constructed using R.

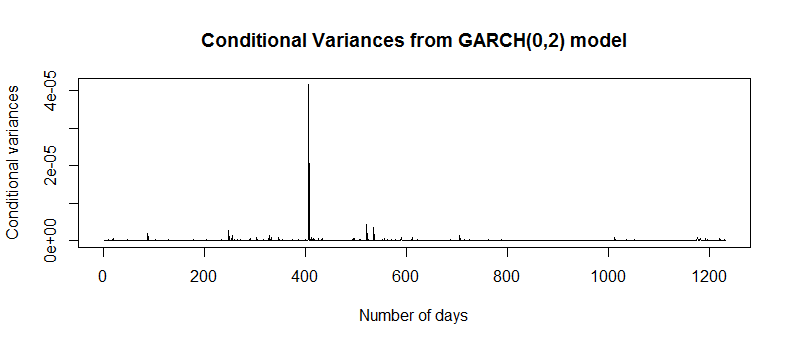
The relevant R code is given below.

|  |
| --- |
| #volatility checking using plots  par(mfrow=c(1,1))  plot.ts(log\_sp,main="Plot of Log Prices",xlab="Number of Days",ylab="Differenced Prices")  plot.ts(d\_log\_sp,main="Plot of diffrenced Prices",xlab="Number of Days",ylab="Differenced Prices")  plot(ht.z2,main="Conditional Variances from GARCH(0,2) model",xlab= "Number of days",ylab="Conditional variances",type='l') |

The plots constructed using the R codes are given**- Diagram (3.13)**







From these plots, we can say that the Conditional Variances plot, obtained after fitting the GARCH model on the residuals of the fitted ARIMA model, successfully reflects the volatility present in the time series under consideration, over the entire period. It is clearly seen that high volatility is related to the time periods where the Adjusted Closed Prices tumbled or had a sudden fall from their usual values.

Thus, our GARCH model is quite successful in modelling the volatility present in the Adjusted Closed Stock Prices. Now, we shall study the two fitted models as a whole.

**3.9-The combined ARIMA (1,1,1) + GARCH (0,2) model:**

We have found the fitted ARIMA (1, 1, 1) model for the log transformed Adjusted Closed Prices and also fitted the GARCH ( 0, 2) model of the returns of the ARIMA model due to the presence of volatility in the series, which was revealed during the diagnostic checking of the model.

Now, we will test the performance of the fitted ARIMA (1, 1, 1) model with that of the fitted ARIMA (1, 1, 1) + GARCH (0, 2) model as a whole.

Thus, the full ARIMA (1, 1, 1) + GARCH (0, 2) model is given as,

|  |
| --- |
| Yt-Yt-1= 4.3366 \*10-4 – 0.57969632(Yt-1 –Yt-2 ) + 0.65876704ɛt-1 1.113e-08 + 2.506e-02\* ɛ2t-1 +2.955e+00\*ht-1 |

Now, we will check the performance of the two models. We have found out the forecasts for the fitted ARIMA model.

The step-1 forecast for the log prices for both the models is given with 95% Confidence Intervals – **Table (3.11)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Point | Forecast | Low 95 | High 95 |
| ARIMA | 1234 | 10.4901 | 10.51168 | 10.47632 |
| ARIMA+GARCH | 1234 | 10.49413 | 10.51252 | 10.47612 |

The step-2 forecast for the actual prices for both the models is given with 95% Confidence Intervals – **Table (3.12)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Point | Forecast | Low 95 | High 95 |
| ARIMA | 1234 | 36098.62 | 35466.36 | 36742.15 |
| ARIMA+GARCH | 1234 | 36102.95510 | 36759.50093 | 3548.56108 |

Thus, the 95% Confidence Interval of ARIMA (1,l, l) is wider than that of the combined model ARIMA (1, 1, 1) + GARCH (0, 2 ). This is because the latter reflects and incorporate recent changes of the stock prices, thus it provides much shorter Confidence Interval. The Combined Model takes into consideration the volatility of stock prices by analyzing the residuals and its conditional variances.

We will now construct the95% Confidence Intervals of the original Adjusted Closing prices using the combined fitted model and plot them along with the original prices.

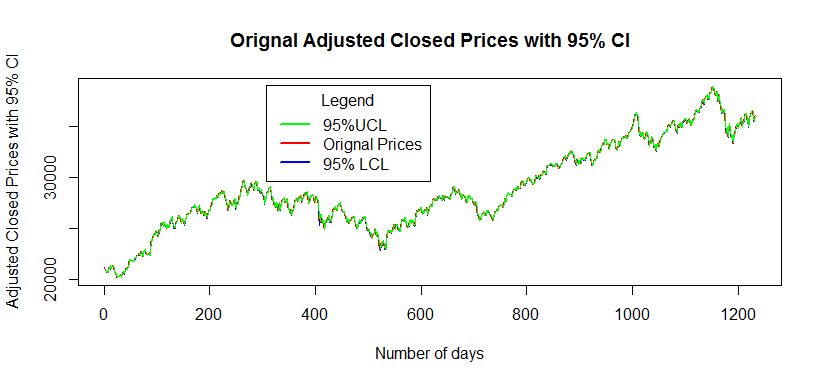
**3.10 - 95% Confidence Interval of the Prices using the combined modelling**

We can plot the95%Confidence Limits of the actual Adjusted Closed Prices using the combined model ARIMA (1,1,1)+GARCH (0,2); we will use the fitted values of the prices of the fitted ARIMA model and the Conditional Variances of the fitted GARCH model.

The relevant R –codes are given below.

|  |
| --- |
| #95% CONFIDENCE INTERVALS OF THE PRICES USING THE COMBINED MODEL  fit111=fitted(arima111)  log\_low= fit111-1.96\*sqrt(ht.z2)  log\_up=fit111+1.96\*sqrt(ht.z2)  low=exp(log\_low)  up=exp(log\_up)  par(mfrow=c(1,1))  plot(sp,type='l',col="Red",main="Orignal Adjusted Closed Prices with 95% CI",xlab="Number of days",ylab="Adjusted Closed Prices with 95% CI")  lines(low,col="Blue")  lines(up,col="Green")  legend("bottomright",lty=c(1,1),lwd=2,insert=0.5,title="Legend",c("95%UCL","Orignal Prices","95% LCL"),col=c("Green","Red","Blue"),bty="o") |

The Time Series plot of the original prices along with the upper and lower 95% Confidence Intervals is given – **Diagram (3.14)**



The above time series plot shows the Original Adjusted Closed Prices in Red, the Upper Confidence Limits in Green and the Lower Confidence Limits in Blue.

**3.11- Diagnostic Checking of Returns of the Combined Model:**

Now, using the diagnostic checking of the returns, we will check the overall performance of the fitted combined model, that is the ARIMA (1,1,1) + GARCH (0,2) model.

**3.11.1- QQ-Plot of Returns:**

The important part of the diagnostic Checking of is to look at the Q-Q Plot of residuals and Returns of combined ARIMA - GARCH model. The formula for the residuals of the overall fitted model is,

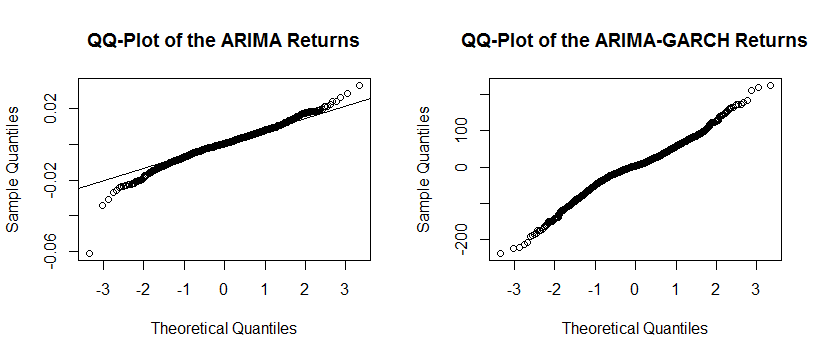
~~e~~t = =

We computed it using the R software, and plotted the Q Q plot to check the normality of the Returns, and also compare it with the Residuals of the previously fitted ARIMA model.

The relevant R- codes are given below.

|  |
| --- |
| #QQ-Plot of the Combined Residuals  comb\_retn= d\_log\_sp/sqrt(ht.z2)  par(mfrow=c(1,2))  qqnorm(d\_log\_sp,main='QQ-Plot of the ARIMA Returns')  qqline(d\_log\_sp)  qqnorm(comb\_retn,main='QQ-Plot of the ARIMA-GARCH Returns')  qqline(comb\_retn) |

The Q-Q plots of the returns of fitted ARIMA model and that of the fitted combined ARIMA- GARCH model are given below-



Clearly, from the above two Q - Q Plots of the Returns of the previously fitted ARIMA model and the combined ARIMA + GARCH model, we can say that, the Returns of the combined model show greater tendency of being Normally distributed and the points are also close to the normal line than the first plot. Thus, graphically, we can infer that our combined ARIMA -GARCH model has performed really well.

**3.11.2 - Ljung- Box Test:**

Using the Ljung- Box test on the Returns of the fitted combined ARIMA/GARCH model, we will confirm our graphical findings.

The R codes for the above test are given below.

|  |
| --- |
| #LJUNG-BOX TEST OF RETURNS,SQUARED RETURNS OF COMBINED MODEL  Box.test(comb\_retn,lag=10,type="Ljung-Box")  Box.test(comb\_retn^2,lag=10,type="Ljung-Box") |

Using R, we got the following information – Table(3.13)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Test | X- squared | Dfs | p-value |
| Returns | Ljung-Box | 18.912 | 10 | 0.04139 |
| Squared Returns | Ljung-Box | 63.063 | 10 | 9.496e-10 |

Thus, according to the Ljung Box test, none of the Returns and the Squared Returns have Significant p value. Hence, we accept the null hypothesis that there is no Auto - Correlation in the returns, implying that the combined model is quite good.

We have, what looks like, a realization of a discrete white noise process, indicating that we have explained the serial correlation present in the squared returns with an appropriate mixture of the ARIMA and the GARCH model of the original Adjusted Closing Stock Prices.

CHAPTER-4

CONCLUSION

**Chapter 4 -- Conclusions**

Firstly, ARIMA model focuses on analyzing time series linearly and it does not reflect recent changes as new information is available. Therefore, in order to update the model, users need to incorporate new data and estimate parameters again. The variance in ARIMA model is unconditional variance and remains constant. ARIMA is applied for stationary series and therefore, non-stationary series should be transformed (such as log transformation).

Additionally, ARIMA is often used together with ARCH/GARCH model. ARCH/GARCH is a method to measure volatility of the series, or more specifically, to model the noise term of ARIMA model. ARCH/GARCH incorporates new information and analyzes the series based on conditional variances where users can forecast future values with up-to-date information. The forecast interval for the combined (mixed) model is closer than that of fitted ARIMA model.

It is to be remembered that due to the presences of high volatility in the stock market prices resulting in the Heteroscedastic (Varying Variance) behavior of the residuals, the ARIMA and GARCH models can only be used mainly for Short Term Forecasting of the stock prices. It might not given reliable results in case of long term forecasting of the stock prices due to the presence of high Volatility in the stock markets.

Thus, in short, we transferred the whole Nifty 50 data from Excel to R platform whose layout is given in the Table 3.1. We extracted the Adjusted Closing Prices from the original data, cleaned them and constructed the plot of the extracted Adjusted Closing values given in Diagram 3.1. We checked for the Stationarity of the Adjusted Closing Prices and found that it is non – stationary given in Table 3.2.We did suitable Transformations (For e.g. Log transformation, Differencing. etc.) to the data to make it Stationary and plotted them, given in Diagram 3.2 and interpreted them.

After this, we proceeded towards the ARIMA Model Fitting of the Prices. We identified the model using the Auto - Correlation F unction (ACF) and Partial Auto - Correlation Function (PACF) plots, given in Diagram 3.3 and Akaike Information Criteria (AIC) in Table 3.4.We found that the ARIMA (1, 1, 1) model with drift has the least AIC and thus we choose this model. Next, we estimated the parameters and also found there corresponding standard errors, 2 values and p -values to test their individual significance given in Table 3.5.Then, we tried to forecast the future prices for the upcoming 100 days and the log forecast plot of the log prices and we also provided the different forecasting errors provided in Table 3.6. We also compared the fitted prices with the original prices in Diagram 3.5. To check the validity of the model, we proceeded towards the diagnostic checking of the residuals of the fitted ARIMA model. We provided the Residuals plot in Diagram 3.6 to check the pattern of the residuals, ACF and PACF plots to study the Auto Correlation structure of the residuals in Diagram 3.7and also constructed the Q - Q plots of the residuals and the squared residuals to test their normality in Diagram 3.8.Finally, we performed the Lj ung - Box test of the squared residuals in Table 3.8 to conclude that there is a trace of conditional heteroscedasticity in the residuals of the fitted model.

Thus, the presence of volatility in the concerned time series became inevitable and to be sure, we further studied the Returns, the Squared Returns plots given in Diagram 3.10and also the ACF and the PACF plots of the Squared Returns in Diagram 3.11.We, then, fitted the GARCH model on the returns of the fitted ARIMA model using the values of the MOS for different types of model, given in Table 3.9. We saw that GARCH (1, 1) model had the least AIC. Thus, we estimated the parameters of the GARCH model and also provided their corresponding standard errors, t- values and p- values to check their significance Table 3.10.We plotted the Conditional Standard deviations and the Conditional Variances in Diagram 3.12.

We combined the two fitted models, that is the previously fitted ARIMA (1, 1, 1) model and the fitted GACRH (0, 2) model. The model is given below.

The combined ARIMA (1, 1, 1) + GARCH (0, 2) model is given as,

|  |
| --- |
| Yt-Yt-1= 4.3366 \*10-4 – 0.57969632(Yt-1 –Yt-2 ) + 0.65876704ɛt-1 1.113e-08 + 2.506e-02\* ɛ2t-1 +2.955e+00\*ht-1 |

After this, we proceeded to compare the two models. Using the previously fitted ARIMA model and the above combined model for predicting the stock prices, we calculated the 1 - step forecast values along with their 95% CIs given in Table 3.11 and Table 3.12. We clearly found that the 95% CI in case of the former model was much wider than that of the latter combined model. Thus, the combined model gives a good forecast model as here. The volatility is also taken care of apart from modelling the original prices only. We also plotted the fitted stock prices along with their 95% using the combined fitted model Diagram 3.14. At last, we did the diagnostic checking of the returns of the combined model. We compared the Q Q plot of the returns of the combined model with that of the fitted ARIMA model Diagram. 3.15.After performing the Ljung 'Box test on the residuals of the combined model, provided in Table 3.13, we claimed that as we have successfully modelled the heteroscedastic nature of the returns of the ARIMA model and thus, the combined model is has Performed better than the previously fitted ARIMA model. Thus, it can be used for short –term forecasting of the Adjusted Closing Stock Prices.

CHAPTER-5

APPENDIX

**Chapter 5- Appendix**

**5.1 - Description of the R packages and R – codes used:**

**5.1.1 - The description of R Packages used in the above time series analysis are given below!**

* fGarch - Provides a collection of functions to analyze and model heteroskedastic behavior in financial time series models.
* forecast - Methods and tools for displaying and analyzing univariate time series forecasts including exponential smoothing via state space models and automatic ARIMA modelling.
* lmtest - A collection of tests, data sets, and examples for diagnostic checking in linear regression models. Furthermore, some generic tools for inference in parametric models are provided.
* timeSeries - Environment for teaching "Financial Engineering and Computational Finance". Managing financial time series objects.
* tseries -Time series analysis and computational finance.

**5.1.2 - The description of the R codes used in the above time series analysis are given below!**

* **acf()** - Computes the sample autocorrelation (covariance) function of x up to certain lag.
* **adf.test()** – Performs the Augmented Dickey-Fuller test for the null hypothesis of a unit root of a univariate time series x (equivalently, x is a non-stationary time series).
* **AIC()** – Fits a Generic function calculating Akaike's ‘ An Information Criterion’ for one or several fitted model objects.
* **arima(**) - Fits an ARIMA model to a univariate time series.
* **Arima()** – Largely a wrapper for the arima() function in the stats package.
* **as.numeric()** - Converts into a numeric value.
* **auto.arima()** - Returns best ARIMA model according to either AIC, AlCC or BIC value. The function conducts a search over possible model with in the order constraints provided.
* **Box.test()-** Computes the Box- Pierce or Ljung – Box test statistic for examining the null hypothesis of independence in a given time series.
* **c()** - A quick utility for concatenating strings or vectors together.
* **coeftest()-** Generic function for performing z and(quasi)t Wald tests of estimated coefficients.
* **diff()-** Returns suitably lagged and iterated differences.
* **exp()** -Computes exponential of the specified float value.
* **fitted()** - Generic function which extracts fitted values from objects returned by modelling functions.
* **forecast()** - Generic function for forecasting from time series or time series models.
* **garch()**- Fits a Generalized Auto regressive Conditional Heteroscedastic GARCH(p, q) time series model to the data by computing the maximum-likelihood estimates of the conditionally normal model.
* **head()** - Returns the first 6 observations of a vector, matrix, table, data frame or function.
* **is.na()** - Returns TRUE if any value is NA i.e., Not Available.
* **legend()** - Can be used to add legends to plots with specific location.
* **length()-** Get or set the length of vectors(including lists)and factors, and of any other R object for which a method has been defined.
* **library()** - Load and attach add-on packages.
* **lines()** - Generic function taking coordinates given in various ways and joining the corresponding points with line segments.
* **log()** - Computes logarithms, by default natural logarithms.
* **logLik()** - Used to extract the Log – Likelihood of fitted model, most commonly used for a model fitted by maximum likelihood.
* **na.omit()** - Functions for handling missing values in 'time Series' objects
* **pacf()** – Computes the sample partial auto correlation function of x up to certain lag is that this function allows a drift term.
* **par()**—Can be used to set or query graphical parameters.
* **paste()-** Concatenate vectors after converting to character.
* **plot()** - Generic function for plotting of R objects.
* **plot.ts()** - Plotting method for objects of a time series.
* **qqline()** - Adds a line to a theoretical, by default normal, quantile-quantile plot which passes through the quantiles, by default the first and third quartiles.
* **qqnorm()** -Generic function the default method of which produces a normal QQ plot of the values
* **read.csv()-** Reads a tile in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.
* **resid()** - Generic function which extracts model residuals from objects returned by modelling functions.
* **sqrt()-** Computes the square root of the specified float value.
* **summary()**- Generic function used to produce result summaries of the results ofvarious model fitting functions.
* **tail()**- Returns the last 6observations of a vector, matrix, table, data frame or function.
* **View()**- Invoke a spreadsheet-style data viewer on a matrix-like R objects.

**References**

* R Documentation
* BSE INDIA- <https://www.bseindia.com/indices/IndexArchiveData.html>
* Montgomery D. C. and Johnson L. A. (1977): Forecasting and Time Series Analysis, McGraw Hill.
* Montgomery D.C., C.L. Jennings and M. Kulachi (2006): Introduction to Time Series Analysis and Forecasting, Wiley InterScience.
* Box G. E. P., Jenkins G. and Reinsel G. C. (2004): Time Series Analysis Forecasting and Control, Pearson Education.
* Shumway R. H. and Stoffer.D. S. (2011): Time Series Analysis and its Applications with R examples, Springer.
* Gujrati, D. (2011): Econometrics by Example, Palgrave Macmillan.
* Ruey S. Tsay (2012): An Introduction to Analysis of Financial Data with R, Wiley.
* Pham Ly: Time Series Analysis with ARIMA ARCH/GARCH model in R, L Stem Group.