Group 116: Building Prediction Model For New York Stock Exchange

First Name	Last Name	Monday or Tuesday class
Mallikarjuna	**Sirabadige Nagaraju**	**Monday**
Vishwas	**Reddy Naga**	**Monday**

Table of contents

1.Introduction:

When person plans gain an edge in the stock market he needs to analyze 3 important sectors of the stock market – the trading instrument of a particular sector, the investment sector or the entire market.

Today, Stock Analysis is very crucial for investors and traders to make day-to-day and even long term trading plans. Applying analysis and prediction techniques to data collected from the past and current scenario to calculate future trends.

While doing Stock Analysis there are two types: Fundamental Analysis and Technical Analysis.

a*) Fundamental Analysis*: It focuses on analysis financial records, economic reports, company assets, and market share of the company.

b) *Technical Analysis*: It concentrates on past market trends to predict future market costs. Its focus is mainly on the demand and supply with its effects on the prices and volumes.

Most traders use these methods to get an upper hand in the market. Using these methods, we can predict the future market trends and analyze the stock market.

Tool used: R-Programming

R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing.

R-programming is used in our project for model creation, selection and forecasting the future prediction.

2.Data:

New York stock exchange contains data around 150 company stocks information and it is very huge data. Yahoo company data is chosen and manually Aggregated the data to monthly basis and created the new csv file by name Yahoo.csv .

Data set collected from:

https://www.kaggle.com/dgawlik/nyse

The data set contained:

a)prices.csv

In prices.csv file

Total Rows: 851265

Total Columns: 7

Period:Jan2010-Dec2016


YAHOO.csv


prices.csv

3. Problems to be Solved:

Investors has to manually analyze the market trends to invest and gain profit, but this process is not accurate and needs more effort and time.

Investors can make use of a prediction from time series model for future investment and profit.

If they know the future trends in the market, there is not necessary to watch the prices of stocks frequently ,it reduces a lot of efforts of investors.

Since from our model ,investors can able to predict the future open and high prices. they can

easily make profit by buying the shares at low open price and sell the shares at good high price.

4.Data Processing:

Since there is no null values or bad data in the obtained data set, there is no requirement of preprocessing of data.

Since there is presence of many companies data set and data set is huge. we selected a Yahoo company data for analysis and we aggregated monthly wise Yahoo company data from Jan 2010 to Dec 2016.

5. Methods and Process:

Hypothesis testing

Hypothesis test is performed by selecting 10% of data as sample data

Null Hypothesis: H0: μ>27

Alternative Hypothesis Ha: μ<27.

Confidence level:95%

Since p value .5883 is greater than 0.05 and Z stat < 1.96 ,we accept null hypothesis and reject alternative hypothesis , that is mean value of open column of Yahoo.csv is more than 27

Original mean open value: 27.28

```
> data <-read.csv("YAHOO.csv",header=T,sep=',')
> x=data$open
> opendata=data%>%sample_frac(0.1)
> open=opendata$open
> z.test(open,NULL,alternative="less",mu=27,sigma.x=sd(open),conf.level=.95)

        One-sample z-Test

data:  open
z = 0.22312, p-value = 0.5883
alternative hypothesis: true mean is less than 27
95 percent confidence interval:
       NA 34.24173
sample estimates:
mean of x
   27.865
```

Time series models

A time series data is always equally spaced interval data like daily or monthly or yearly.Time series forecasting is the use of a model to predict future values based on previously observed values. In our project monthly data is used for the time series data analytics task .we created MA,AUTOARIMA and AR model for forecast and predict the future values. AIC metrics is used for the model selection on test data set. RMSE metrics is used for the model selection for future prediction and forecasting.

Time series prediction model built as below steps:

a. Libraries for the time series was loaded.
b. DEC 2016 month data is used as testing data
c. Data for Yahoo company is loaded into R from JAN 2010 to NOV 2016

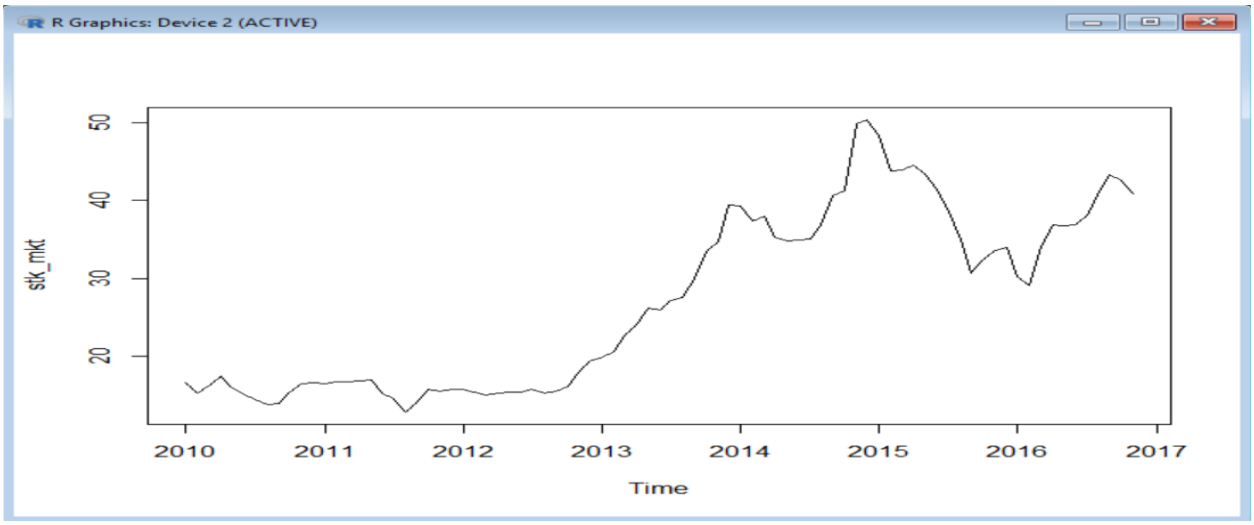
1)Data loaded into R

```
>
> data<-read.csv("YAHOO.csv",header=T,sep=',')
>
>
>
>
>
```

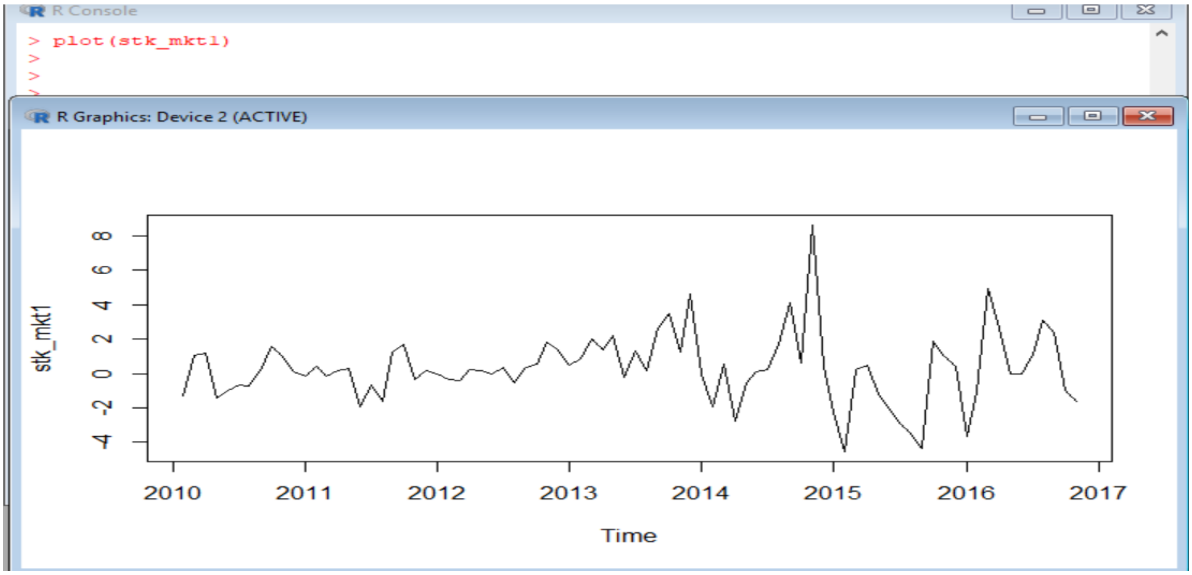
2)Data set contain data up to November 2016 and December 2016 month data is used as testing data

```
>
> stk_mkt=ts(data[,3],start=c(2010,1), end=c(2016,11),frequency = 12)
>
>
>
>
>
```

3)Plot the time series mean and variance is not constant over a period of time in the plot



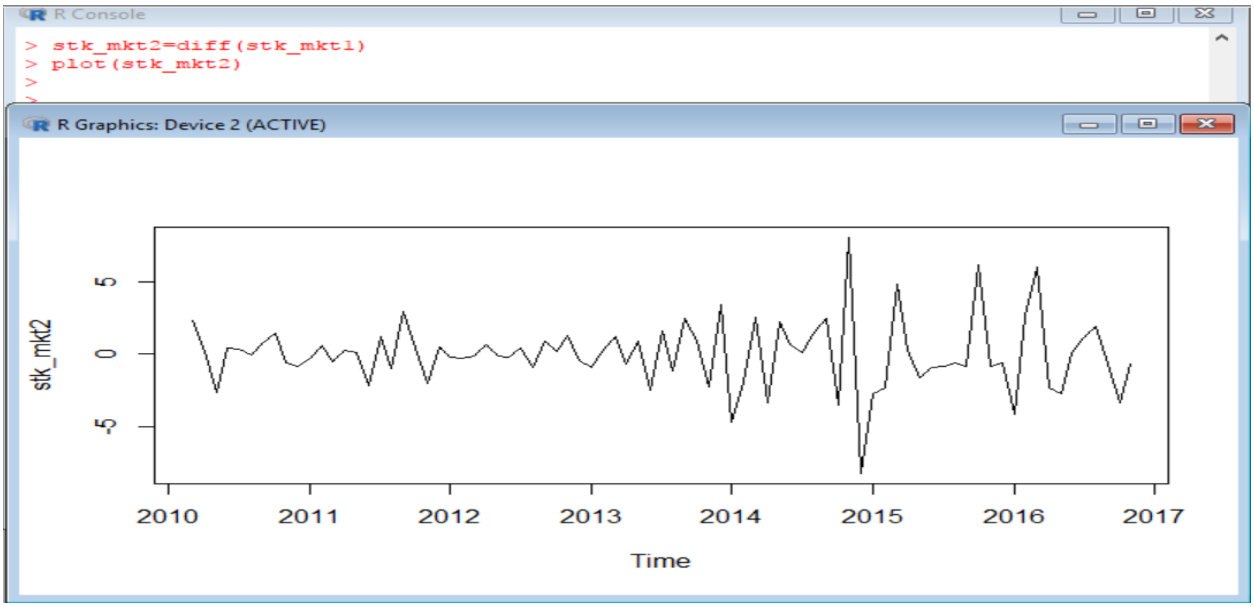
4)Applied first differencing and plot is as below



5)First differencing not having constant mean and variance and checked for serial correlation through Ljung box test.

```
> Box.test(stk_mkt1,lag=6,type = 'Ljung')

        Box-Ljung test

data:  stk_mkt1
X-squared = 9.7874, df = 6, p-value = 0.1339

> Box.test(stk_mkt1,lag=12,type = 'Ljung')

        Box-Ljung test

data:  stk_mkt1
X-squared = 19.707, df = 12, p-value = 0.07284
```

6)Applied Second differencing and plot is as below

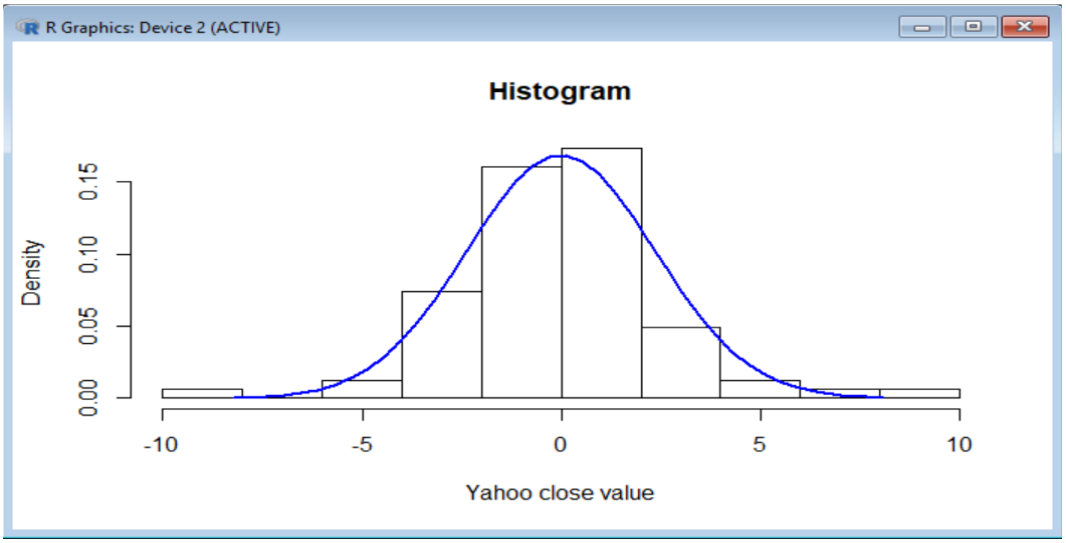


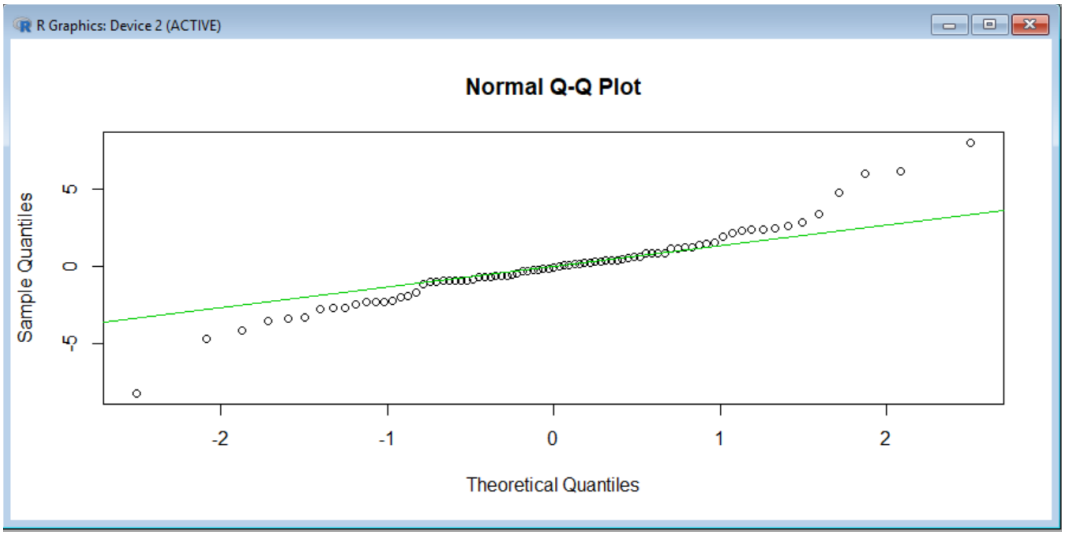
Second differencing is better compared previous plots ,it is almost having constant mean and variance. we are using Second differencing for the model creation and selection.

7)Basic Statics as below

```
> basicStats(stk_mkt2)
               stk_mkt2
nobs          81.000000
NAs            0.000000
Minimum       -8.220000
Maximum        8.060000
1. Quartile   -0.900000
3. Quartile    0.910000
Mean          -0.004198
Median        -0.060000
Sum           -0.340000
SE Mean        0.263459
LCL Mean      -0.528498
UCL Mean       0.520103
Variance       5.622270
Stdev          2.371133
Skewness       0.256677
Kurtosis       2.484720
>
```

8)Plotting Histogram and QQ plot for data normal distribution check





9)Normality test by using Jarque Bera test

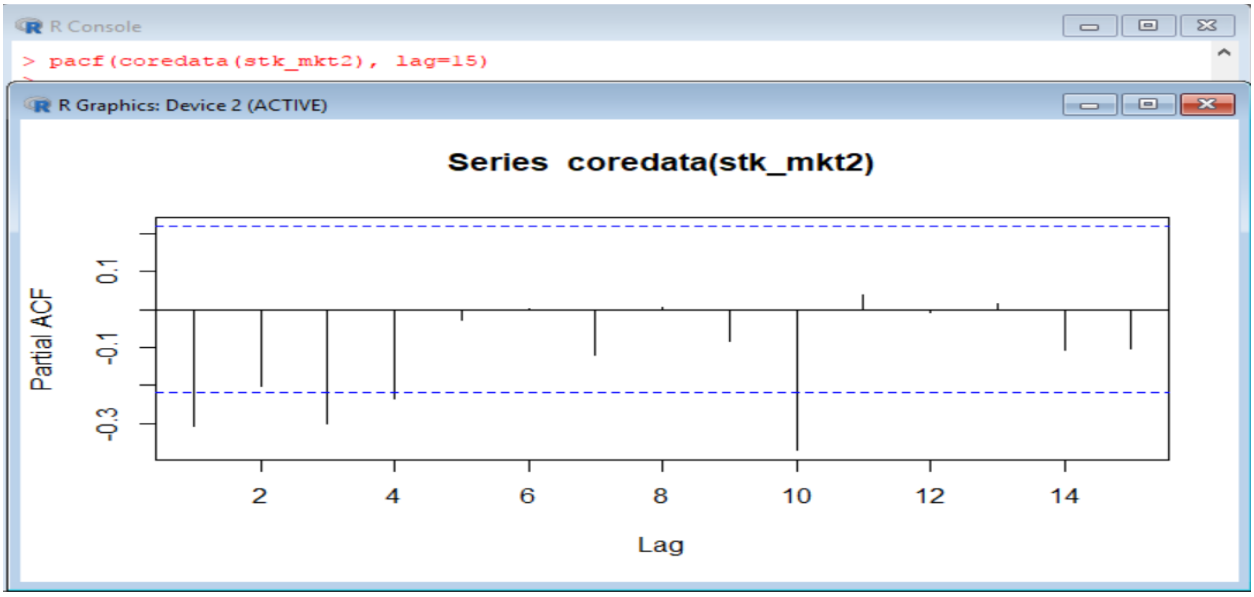
```
> normalTest(stk_mkt2,method=c("jb"))

Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 24.1382
  P VALUE:
    Asymptotic p Value: 5.734e-06

Description:
 Mon Nov 27 15:08:44 2017 by user: arjun
```

10)Serial correlation check By using Ljung box test and data is good to use past to predict the future

Since P value is less than 0.05.

```
> Box.test(stk_mkt2,lag=12,type = 'Ljung')

        Box-Ljung test

data:  stk_mkt2
X-squared = 26.554, df = 12, p-value = 0.008952
```

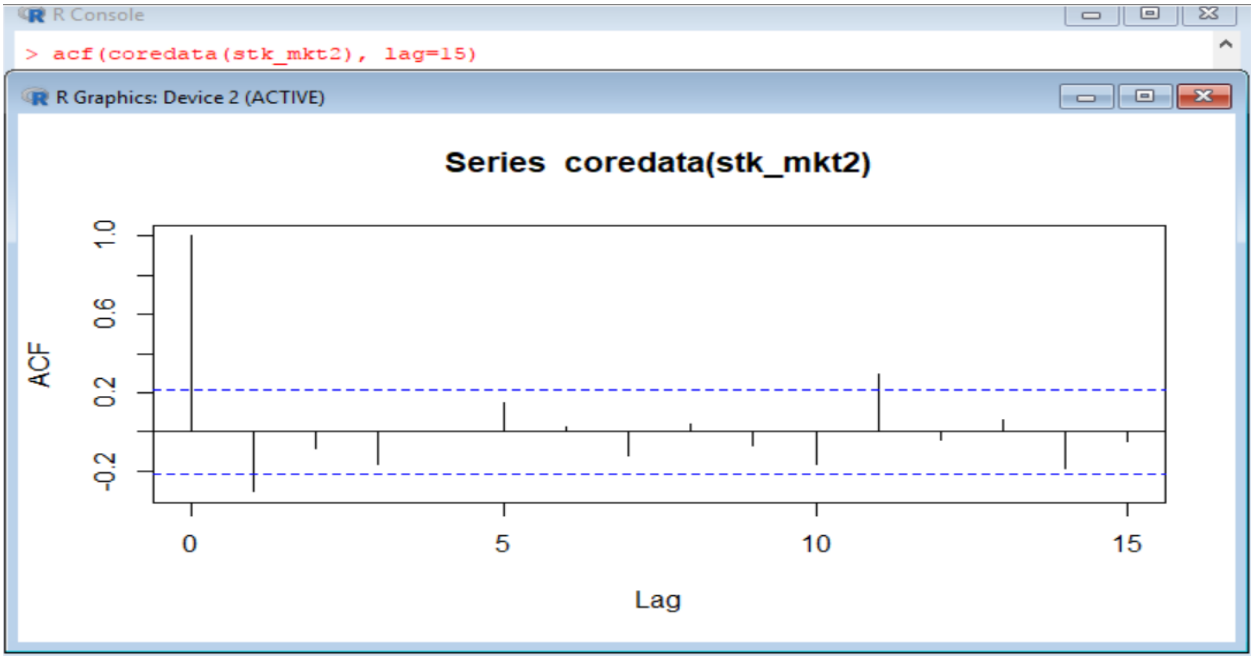
11)Pacf plot for obtaining the p value to build AR model and p value is 10.



12)AR model is built by using the p value obtained in the last step.

```
> AR=arima(stk_mkt2,order=c(10,0,0), method = 'ML',include.mean = T)
> AR

Call:
arima(x = stk_mkt2, order = c(10, 0, 0), include.mean = T, method = "ML")

Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8
      -0.5677  -0.4654  -0.5600  -0.4036  -0.1952  -0.2654  -0.3679  -0.2453
s.e.   0.1009   0.1152   0.1233   0.1326   0.1383   0.1352   0.1292   0.1221
          ar9     ar10  intercept
      -0.3403  -0.4100     0.0064
s.e.   0.1181   0.1035     0.0438

sigma^2 estimated as 3.238:  log likelihood = -163.97,  aic = 351.94
>
```

13)1acf plot for obtaining the p value to build AR model and q value is 11.



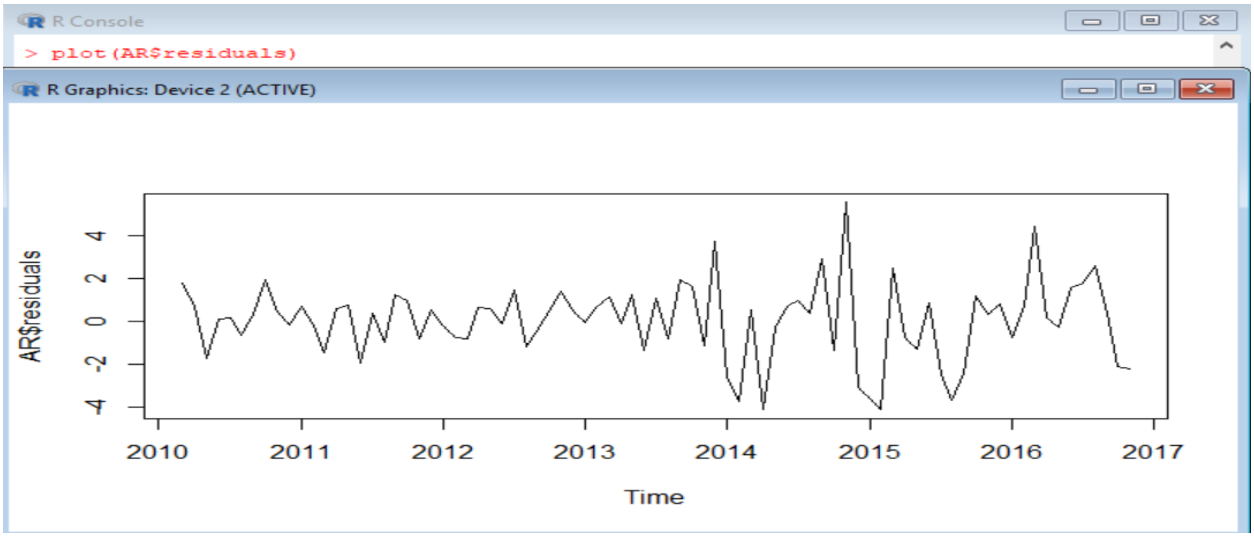
14)MA model is built by using the p value obtained in the last step.

```
>  MA=arima(stk_mkt2,order=c(0,0,11), method = 'ML',include.mean = T)
> MA

Call:
arima(x = stk_mkt2, order = c(0, 0, 11), include.mean = T, method = "ML")

Coefficients:
          ma1      ma2      ma3     ma4     ma5      ma6      ma7     ma8      ma9
      -0.5779  -0.1420  -0.4019  0.0973  0.2533  -0.1772  -0.3371  0.2104  -0.0166
s.e.   0.1262   0.1394   0.1456  0.1672  0.1418   0.1865   0.1686  0.1421   0.1806
         ma10     ma11  intercept
      -0.1184  0.2101     0.0011
s.e.   0.1312  0.1469     0.0103

sigma^2 estimated as 2.802:  log likelihood = -161.43,  aic = 348.85
```

15)AutoArima model is built by using AUTOARIMA function which will automatically take p and q value.

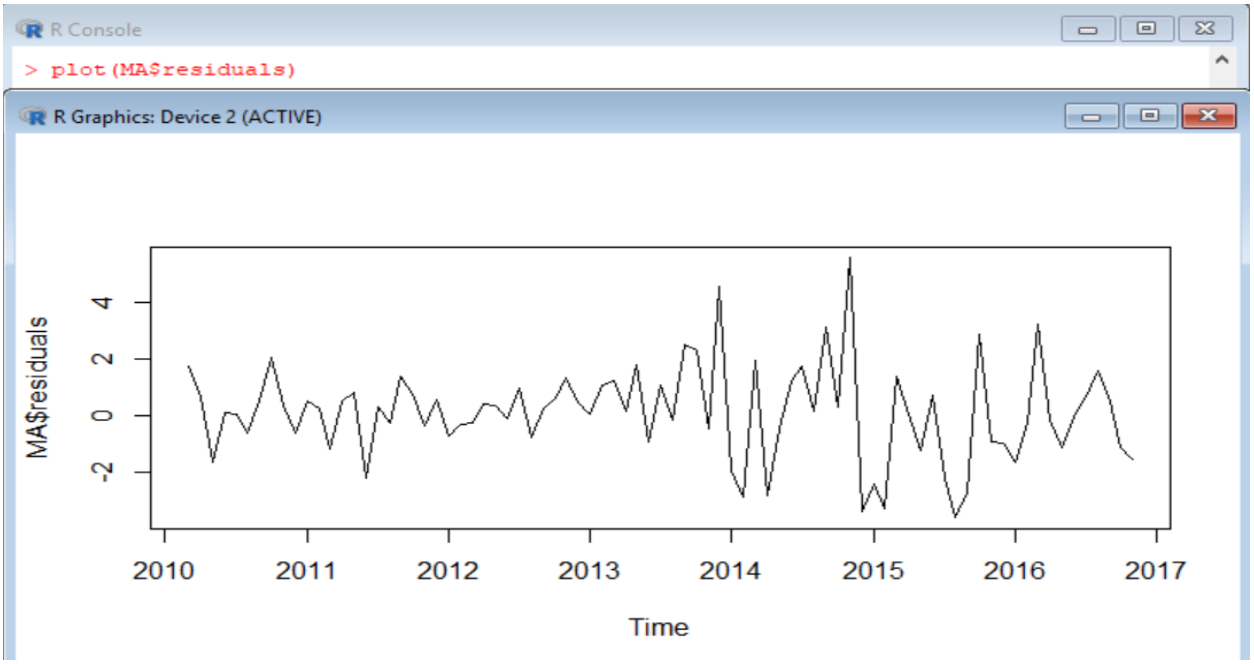
```
> AutoArima=auto.arima(stk_mkt2)
> AutoArima
Series: stk_mkt2
ARIMA(1,0,0)(0,1,1)[12]

Coefficients:
          ar1     sma1
      -0.3983  -0.8405
s.e.   0.1112   0.3126

sigma^2 estimated as 5.125:  log likelihood=-160.1
AIC=326.19   AICc=326.56   BIC=332.89
```

16)Residual plot and Ljung box test for residuals for AR model.



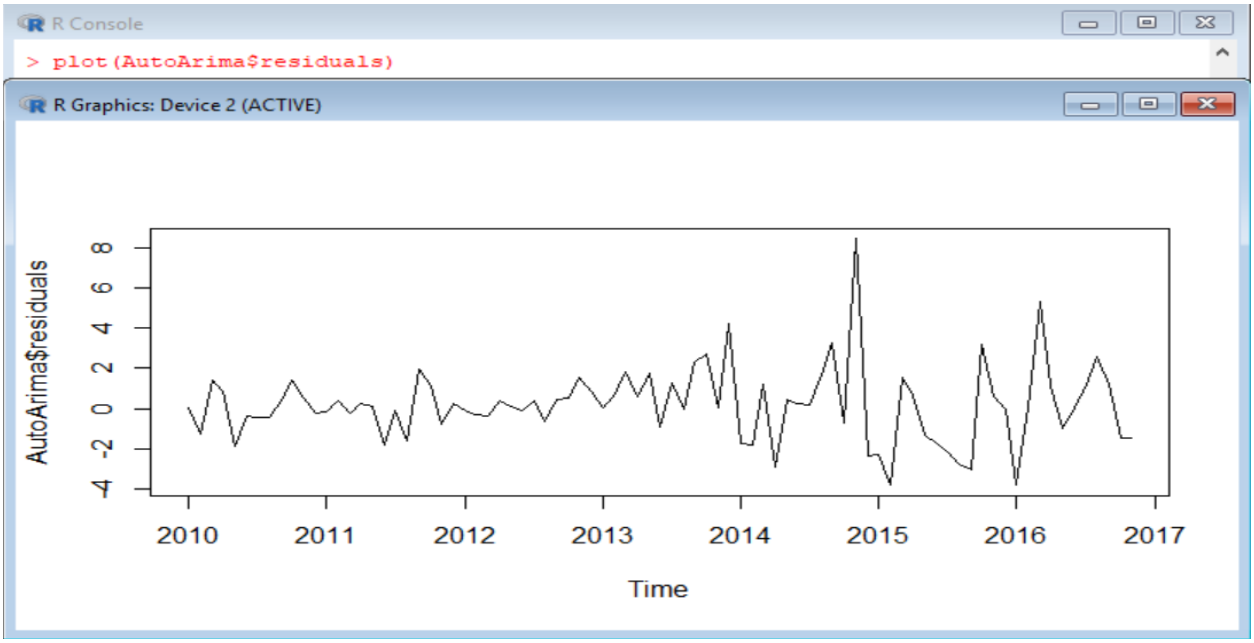

```
> Box.test(AR$residuals, lag=6,type="Ljung")

        Box-Ljung test

data:  AR$residuals
X-squared = 0.93007, df = 6, p-value = 0.9881

> Box.test(AR$residuals, lag=12,type="Ljung")

        Box-Ljung test

data:  AR$residuals
X-squared = 2.2512, df = 12, p-value = 0.9989

> Box.test(AR$residuals, lag=18,type="Ljung")

        Box-Ljung test

data:  AR$residuals
X-squared = 7.2722, df = 18, p-value = 0.9876
```

Residual plot and Ljung box test for residuals for MA model.




```
> Box.test(MA$residuals, lag=6,type="Ljung")

        Box-Ljung test

data:  MA$residuals
X-squared = 0.33668, df = 6, p-value = 0.9993

> Box.test(MA$residuals, lag=12,type="Ljung")

        Box-Ljung test

data:  MA$residuals
X-squared = 3.5801, df = 12, p-value = 0.9899

> Box.test(MA$residuals, lag=18,type="Ljung")

        Box-Ljung test

data:  MA$residuals
X-squared = 7.2253, df = 18, p-value = 0.9881
```

Residual plot and Ljung box test for residuals for AutoArima model.




```
>
> Box.test(AutoArima$residuals, lag=6,type="Ljung")

        Box-Ljung test

data:  AutoArima$residuals
X-squared = 9.036, df = 6, p-value = 0.1716

> Box.test(AutoArima$residuals, lag=12,type="Ljung")

        Box-Ljung test

data:  AutoArima$residuals
X-squared = 21.137, df = 12, p-value = 0.04842

> Box.test(AutoArima$residuals, lag=18,type="Ljung")

        Box-Ljung test

data:  AutoArima$residuals
X-squared = 31.749, df = 18, p-value = 0.02354
```

Since p-value is larger than 0.05 at 95% confidence level in all the models, we can say residuals is white noise.

Note:

By taking AIC as metrics AutoArima having less AIC value 332.89 compared to AR(351.94) and MA(348.85) model. We are using the model for test the testing data.

We use AIC metrics for testing test data(DEC 2016)

17)Applied reverse differencing on AutoArima model and testing the test data (Decemeber 2016).

We considered AIC as a metrics for testing data set.

```
> ARMAFIT=auto.arima(stk_mkt,d=2,approximation=FALSE,trace=FALSE)
> ZOpen=predict(ARMAFIT,n.ahead = 1, se.fit = T)
> ZOpen
$pred
          Dec
2016 40.81723

$se
          Dec
2016 2.027719
```

Original value of Open price in December 2016 : 39.74

Predicted value of Open price in December 2016 : 40.81

Predicted Open price value is almost matching .

6.Evaluations and Results

6.1 Evaluation Method:

RMSE metrics is used for the model selection for future prediction and forecasting.

1)For feature prediction we choose the RMSE metrics and for AR,MA and AutoArima model is as follows.

```
> accuracy(AR)
                     ME      RMSE       MAE       MPE     MAPE      MASE        ACF1
Training set 0.02105547 1.799567 1.354785 52.62942 215.3864 0.5344055 0.005478185
>

> accuracy(MA)
                    ME      RMSE       MAE        MPE     MAPE      MASE        ACF1
Training set 0.02996579 1.986571 1.370862 0.3127326 4.776296 0.1900443 -0.01422295
>
`

> accuracy(AutoArima)
                   ME      RMSE       MAE      MPE     MAPE      MASE        ACF1
Training set 0.1066964 1.645352 1.204389 54.24666 182.2739 0.4750805 -0.03006268
>
```

AutoArima is having less RMSE value compared to other models, we are using AutoArima model for prediction

6.2 Results and Findings

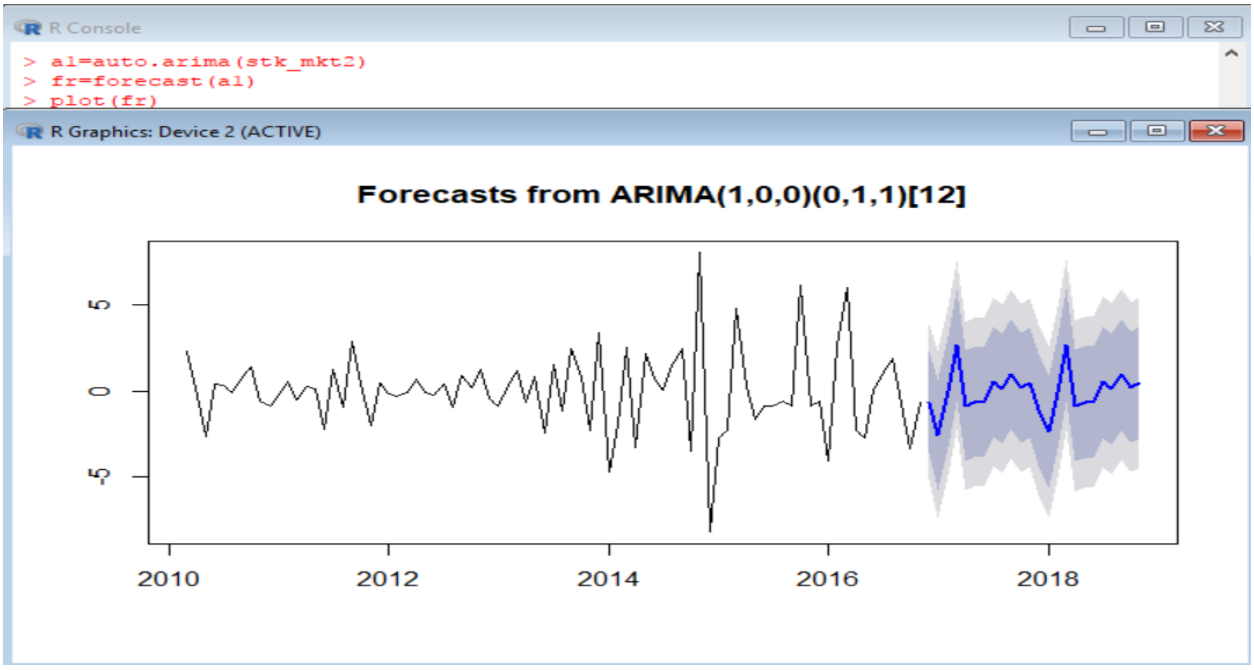
```
> ARMAFIT=auto.arima(stk_mkt,d=2,approximation=FALSE,trace=FALSE)
> ZOpen=predict(ARMAFIT,n.ahead = 21, se.fit = T)
> ZOpen
$pred
         Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2016                                                                                                 40.81723
2017 41.80516 42.74218 42.80370 42.45446 42.43637 42.81691 43.25657 43.45867 43.49868 43.59423 43.81820 44.08446
2018 44.29215 44.44173 44.59449 44.78487 44.99497 45.19258 45.37105 45.54636

$se
          Jan       Feb       Mar       Apr       May       Jun       Jul       Aug       Sep       Oct       Nov       Dec
2016                                                                                                            2.027719
2017  3.642221  5.179116  6.589462  8.076436  9.878018 11.883366 13.985838 16.125767 18.340523 20.681358 23.142782 25.696755
2018 28.321890 31.021577 33.806989 36.679253 39.630824 42.654130 45.747646 48.913064
```

2) New AutoArima model for High price is created and we followed all the previous steps to build the new model and its prediction as below.

```
> ZHigh=predict(ARMAFIT1,n.ahead = 21, se.fit = T)
> ZHigh
$pred
         Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2016                                                                                                 41.17145
2017 42.17766 43.10248 43.11941 42.78560 42.77238 43.14987 43.57134 43.75816 43.79943 43.89625 44.11462 44.36866
2018 44.56650 44.71235 44.86252 45.04724 45.24879 45.43850 45.61140 45.78199

$se
         Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2016                                                                                                 2.062518
2017  3.687292  5.257427  6.699120  8.224336 10.062154 12.101459 14.243184 16.426791 18.689395 21.078119 23.587364 26.191118
2018 28.869009 31.624125 34.466196 37.395975 40.406455 43.490603 46.646857 49.876436
```

From the above 2 model ,investor can able to check the less open price and high price on the same month, if he purchased and sell the stocks in the same month, definitely he can able to make profit .

Forecast for AutoArima (Open price) model



7) Conclusions and Future Work:

7.1 Conclusions:

- Hypothesis testing on Yahoo company stock data is conducted and obtained the desired results.
- By using AIC as metrics and AutoArima model, testing conducted on the test data is passed and obtained nearly matching value as of predicted value.
- RMSE as a metrics, we choose the AutoArima model for future prediction and forecasting.

7.2 Limitations:

- Even we predict future prices of Yahoo stocks, we cannot have guaranteed the Investors that they will make a profit, because company stocks price also depends on many factors like company annual profit, company external or internal affairs, economical and geological factors etc.

7.3 Future work:

- In this project, prediction restricted to only yahoo company stock price. In future this project can be implement in full pledge to predict future market trends of all company listed in New York Stock exchange.
- Any other metrics can be used apart from AIC and RMSE metrics, which is used for model selection to forecast and predict the future stock price of Yahoo.
- Stock market prices may vary in every bit of second, better model to be created that can predict the prices for every second or a minute, that can help the investors to check the daily improvement in the stocks, so they can easily estimate the profit and they can invest money on stocks.