S&P500 and MSFT
Time Series Analysis

By,

Jai Katariya

Abstract

The purpose of this report is to analyze S&P 500 from 2014 to 2018 and MSFT data for a random day. Perform exploratory analysis to analyze the trend, seasonality, auto correlation, distribution, skewness and kurtosis. Perform differencing or log transformation if the time series is not stationary. Perform normality and independence tests. Build forecast models using arima. Select model based on the AIC and coefficient significance. Use the model to predict the future values. We will be using R and required libraries to analyze the datasets.

The Dataset

Download the S&P500 data from 02-26-2014 to 02-23-2018 from this link

https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC. Also download MSFT minutes data for a random day.

S&P500 daily returns: 1007 observations and 7 variables

```
sp500daily=read.csv("https://raw.githubusercontent.com/NikhilKorati/Time-Series/master/GSPCdaily26Feb2018.csv",header=T)

head(sp500daily)
tail(sp500daily)
```

	Date <fctr>	Open <dbl>	High <dbl>	Low <dbl>	Close <dbl>	Adj.Close <dbl>	Volume <dbl>
1002	2018-02-15	2713.46	2731.51	2689.82	2731.20	2731.20	3684910000
1003	2018-02-16	2727.14	2754.42	2725.11	2732.22	2732.22	3637460000
1004	2018-02-20	2722.99	2737.60	2706.76	2716.26	2716.26	3627610000
1005	2018-02-21	2720.53	2747.75	2701.29	2701.33	2701.33	3779400000
1006	2018-02-22	2710.42	2731.26	2697.77	2703.96	2703.96	3701270000
1007	2018-02-23	2715.80	2747.76	2713.74	2747.30	2747.30	3189190000

MSFT minutes data: 391 observations and 10 variables

```
MSFT_day=read.csv("https://raw.githubusercontent.com/NikhilKorati/Time-Series/master/TestMSFT20051111.csv",header=T)

head(MSFT_day)
tail(sp500daily)
```

	X <fctr>	Time <int>	open <dbl>	close <dbl>	low <dbl>	high <dbl>	Vol.trades <int>	Nr.trades <int>	Average <dbl>
1	MSFT	930	27.15	27.17	27.13	27.17	754229	764	27.1438
2	MSFT	931	27.17	27.19	27.13	27.18	328597	471	27.1650
3	MSFT	932	27.18	27.21	27.13	27.19	343711	390	27.1916
4	MSFT	933	27.19	27.20	27.18	27.18	231715	339	27.1957
5	MSFT	934	27.19	27.19	27.16	27.17	151506	269	27.1733
6	MSFT	935	27.17	27.18	27.16	27.17	169217	195	27.1707

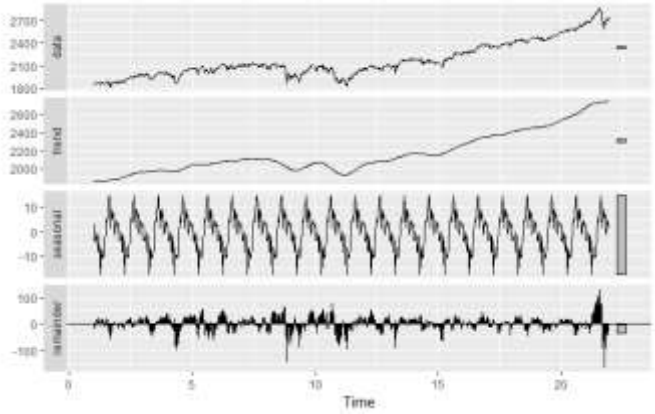
We will be using adjusted close for S&P500 and close for MSFT to build the forecast model and use it as base to perform some prediction to predict future values.

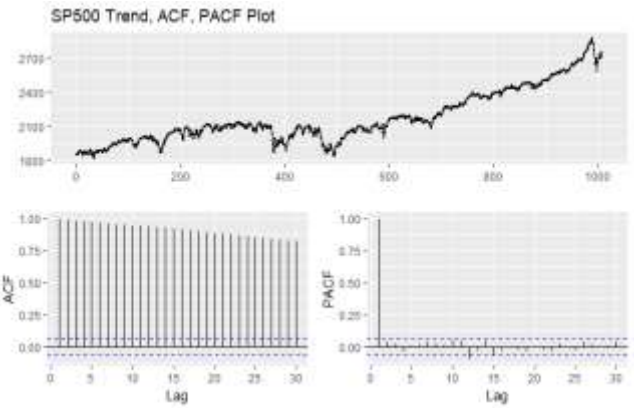
Exploratory Data Analysis of S&P500 daily returns

We will make use of ggplot, ggfortify, stat and forecast libraries to perform the EDA.

```
qplot(sp500daily$Adj.Close, fill = I("blue"), alpha = I(.5), col= I("red"),xlab = "Adj Close",
main="Histogram of SP500 Adj Closing Price", bins = 30)
```

```
sp500close = ts(sp500daily$Adj.Close, frequency = 48)

autoplot(stl(sp500close, s.window = 'periodic'), ts.colour = 'blue')

ggtsdisplay(sp500daily$Adj.Close, main = "SP500 Trend, ACF, PACF Plot")
```

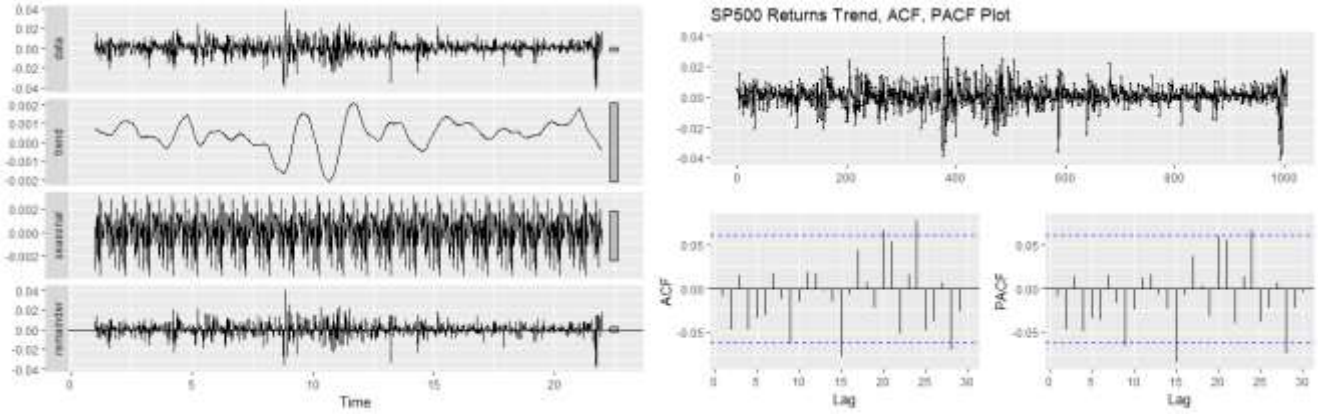




The time series shows an upward trend. There is no sign of stationarity. ACF is exponentially decreasing with a strong autocorrelation.

S&P 500 Daily Return Analysis:

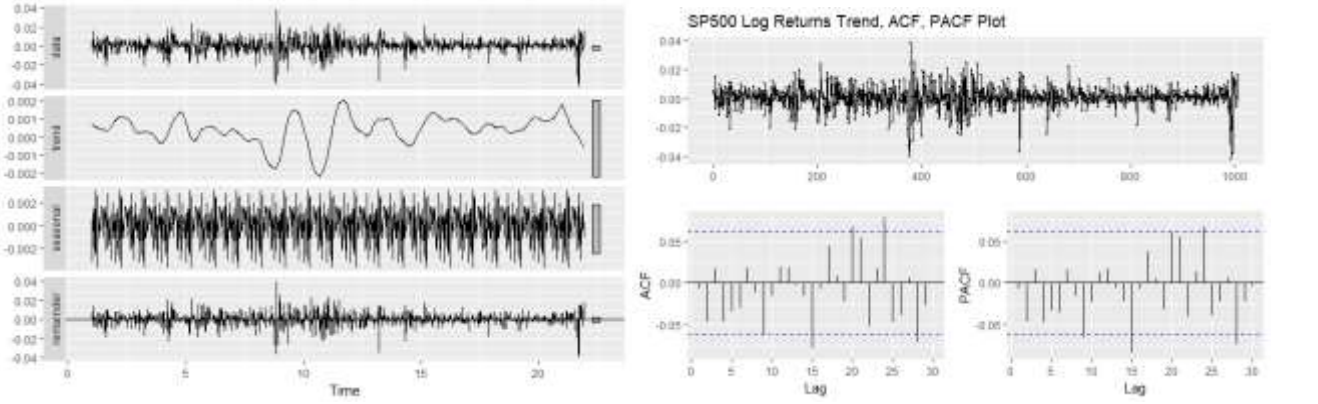
```
sp500_price = sp500daily$Adj.Close
sp500returns =diff(sp500_price)/sp500_price[-length(sp500_price)]


qplot(sp500returns, fill = I("blue"), alpha = I(.5), col= I("red"),xlab = "SP500 Returns",
main="Histogram of SP500 Returns", bins = 30)


sp500_returns = ts(sp500returns, frequency = 48)

autoplot(stl(sp500_returns, s.window = 'periodic'), ts.colour = 'blue')

ggtsdisplay(sp500returns, main = "SP500 Returns Trend, ACF, PACF Plot")
```



The daily returns after differencing is close to stationary. There is no trend but variance seems to increasing and decreasing at some intervals. There is no significant auto-correlation between the lags.

S&P 500 Log Return Analysis:

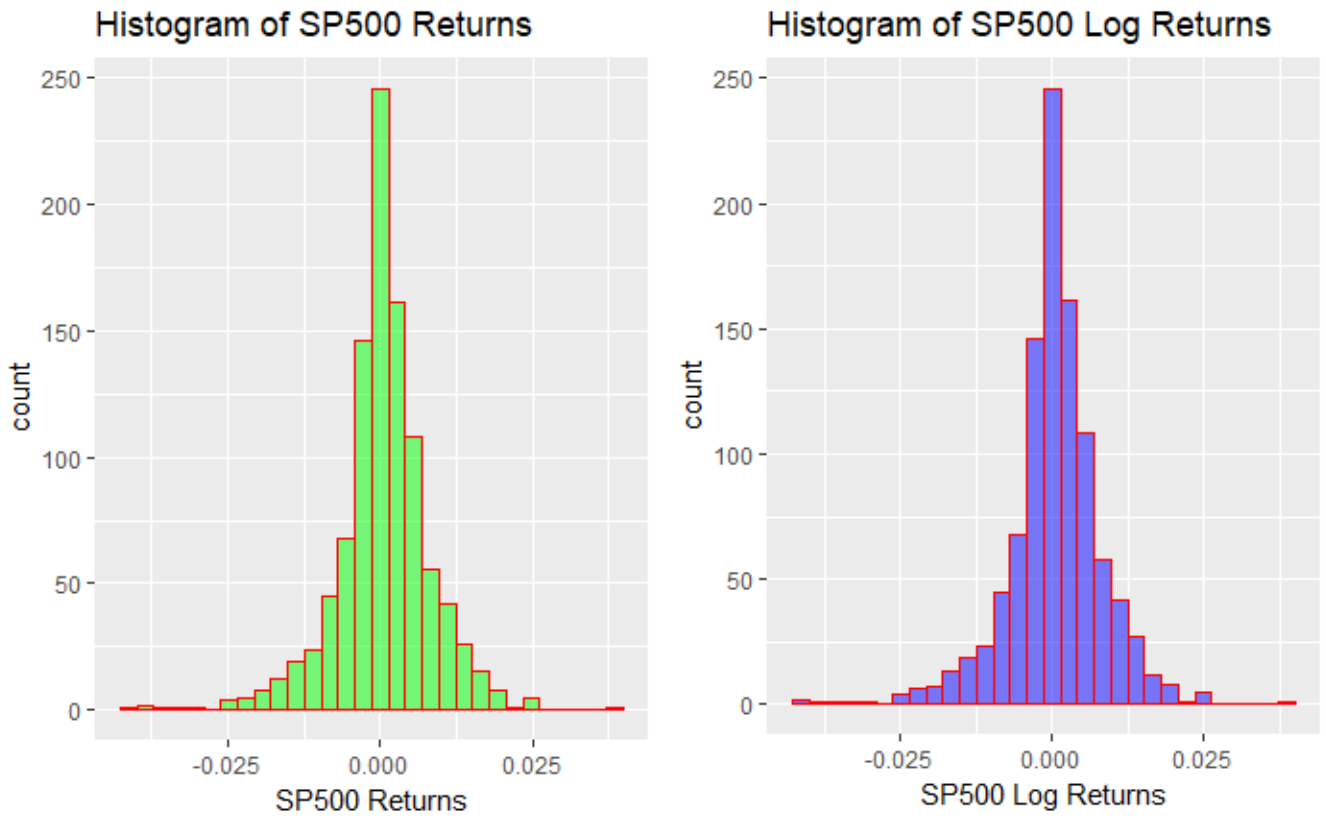
```
sp500_logreturns1=diff(log(sp500_price))

qplot(sp500_logreturns1, fill = I("blue"), alpha = I(.5), col= I("red"),xlab = "SP500 Log Returns"
 main="Histogram of SP500 Log Returns", bins = 30)

sp500_logreturns2 = ts(sp500_logreturns1, frequency = 48)

autoplot(stl(sp500_logreturns2, s.window = 'periodic'), ts.colour = 'blue')

ggtsdisplay(sp500_logreturns1, main = "SP500 Log Returns Trend, ACF, PACF Plot")
```



The log returns after differencing is close to stationary. There is no trend but variance seems to increasing and decreasing at some intervals. There is no significant auto-correlation between the lags.

An interesting observation here is, you can see that there is no significant difference between the daily return and log return. They both look similar.

```
require(gridExtra)
Daily_Returns = qplot(sp500returns, fill = I("green"), alpha = I(.5), col= I("red"),xlab = "SP500
Returns",  main="Histogram of SP500 Returns", bins = 30)

Log_Returns=qplot(sp500_logreturns1, fill = I("blue"), alpha = I(.5), col= I("red"),xlab = "SP500
Log Returns",  main="Histogram of SP500 Log Returns", bins = 30)
grid.arrange(Daily_Returns, Log_Returns, ncol=2)
```



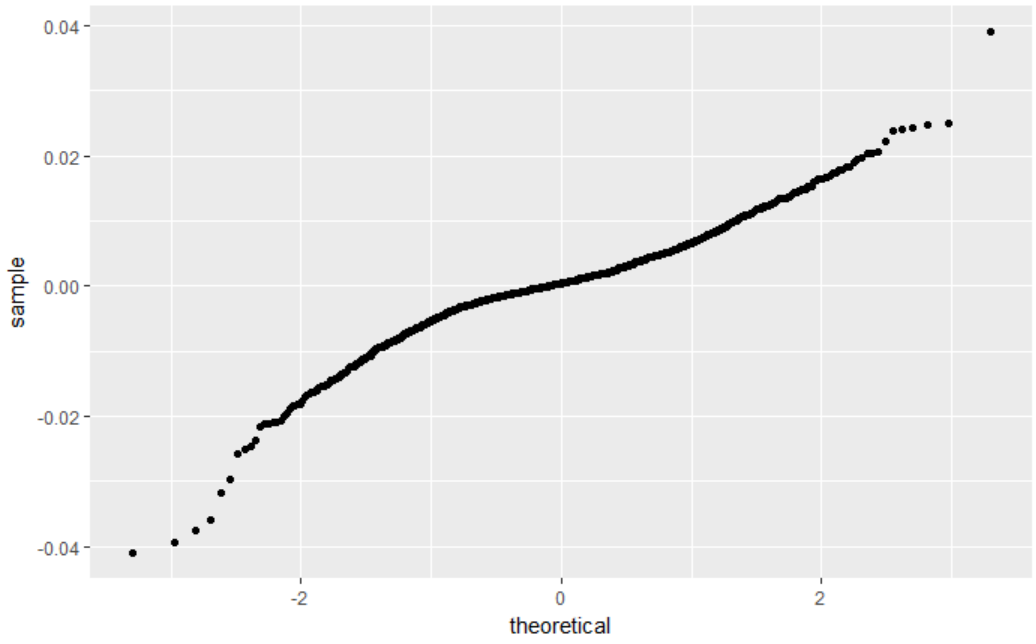
As observed the histograms of both S&P500 daily returns and S&P500 log returns look the same. So we can either use S&P500 daily returns or S&P500 log returns to build our forecast model.

We will be using the S&P500 daily return to build the forecast model.

S&P500 Daily returns Statistical Analysis

```
ggplot(data.frame(sp500returns), aes(sample=sp500returns))+stat_qq()

t.test(sp500returns)

## Normality tests

normalTest(sp500returns,method="jb")

autoplot(acf(sp500returns,lag=15, plot = FALSE)) # Obtain the ACF plot

##Independence tests
Box.test(sp500returns,lag=10)

Box.test(sp500returns,lag=10,type="Ljung")
```

From the QQ plot the data is not normal.



p-value is less than alpha we reject null hypothesis so the true mean is not equal to 0

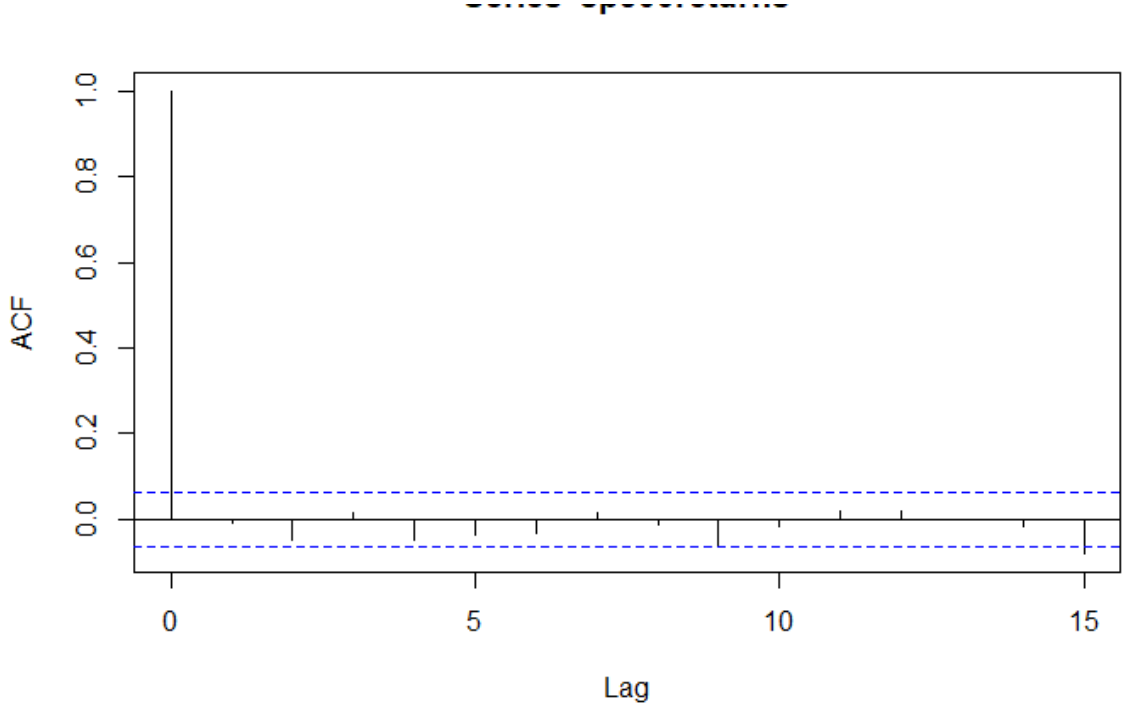
```
        One Sample t-test

data:  sp500returns
t = 1.7285, df = 1005, p-value = 0.0842
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -5.767179e-05  9.105032e-04
sample estimates:
   mean of x
0.0004264157
```

JB test for normality: p value is less than alpha so data does not follow a normal distribution

Box-Pierce test: p value is greater than alpha so we fail to reject null hypothesis so we can say the lags are independent

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 638.1122
  P VALUE:
    Asymptotic p Value: < 2.2e-16

Description:
 Thu Mar 15 22:42:21 2018 by user: nikhil


        Box-Pierce test

data:  sp500returns
X-squared = 11.848, df = 10, p-value = 0.2954


        Box-Ljung test

data:  sp500returns
X-squared = 11.941, df = 10, p-value = 0.289
```

The ACF looks like MA(1) as there is no correlation at lag 1



Model Building S&P500 daily returns

```
model1 = ar(sp500returns, method = 'mle')
model1
ggtsdisplay(model1$resid, main= "Residual Plot, ACF, PACF")
```

```
Call:
ar(x = sp500returns, method = "mle")


Order selected 0  sigma^2 estimated as  6.116e-05
```

The AR model does not show that 1st order is significant so this is not a good model for our data

```
model2=arima(x=sp500returns,order=c(0,0,1))
```

```
arima(x = sp500returns, order = c(0, 0, 1))

Coefficients:
         ma1  intercept
      -0.0098      4e-04
s.e.   0.0332      2e-04

sigma^2 estimated as 6.116e-05:  log likelihood = 3452.7,  aic = -6899.4

        Box-Ljung test

data:  model2$resid
X-squared = 11.958, df = 10, p-value = 0.2878
```

This MA(1) model has a p-value greater than alpha. This is not a good model to do the forecasting.

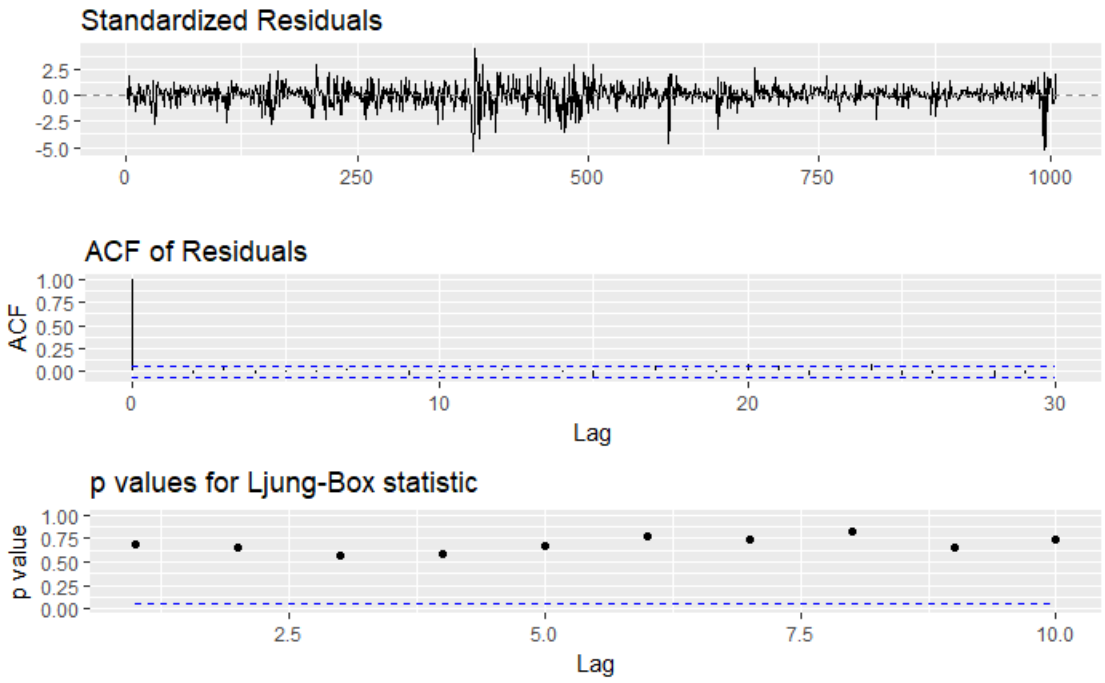
We build 5 other arima models. We select arima(1,0,1) because both the AR and MA coefficients are significant and also the AIC is the best compared to other models.

Model	AIC	Significant coefficients
AR("mle")		
arima(0,0,1)	-6899.4	✗
arima(1,1,1)	-6884.6	✗
arima(1,0,1)	-6905.41	✓
arima(2,1,2)	-6883.9	✓
arima(3,0,2)	-6901.82	✓
arima(3,0,3)	-6900.17	✓

ARIMA(1,0,1) model S&P500 daily returns

```
ggtsdiag(model4)
```

```
poly1=c(1,-model4$coef[1:3])
roots=polyroot(poly1)
roots

Mod(roots)
```



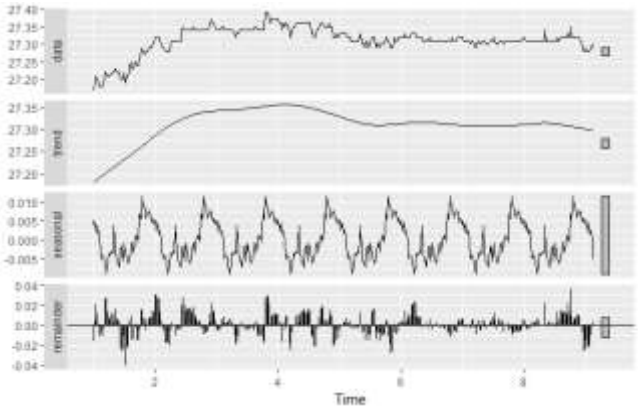

```
1.014283     1.014283 2320.441655
```

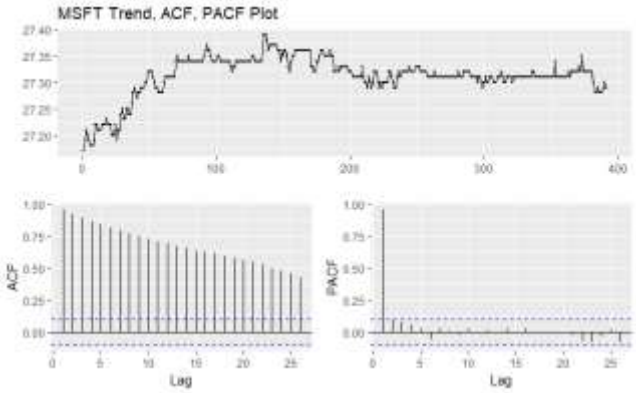
The ACF plot shows a good stationary model. P-values are well above alpha 0.05. The roots are also above 1 so this is stationary.

Exploratory Data Analysis of MSFT minute returns

We will make use of ggplot, ggfortify, stat and forecast libraries to perform the EDA.

```
qplot(MSFT_day$close, fill = I("green"), alpha = I(.5), col= I("red"),xlab = "Close",  main="Histogram of MSFT
Closing Price", bins = 15)


msftminutes = ts(MSFT_day$close, frequency = 48)

autoplot(stl(msftminutes, s.window = "periodic"), ts.colour = "blue")

ggtsdisplay(MSFT_day$close, main= "MSFT Trend, ACF, PACF Plot")
```

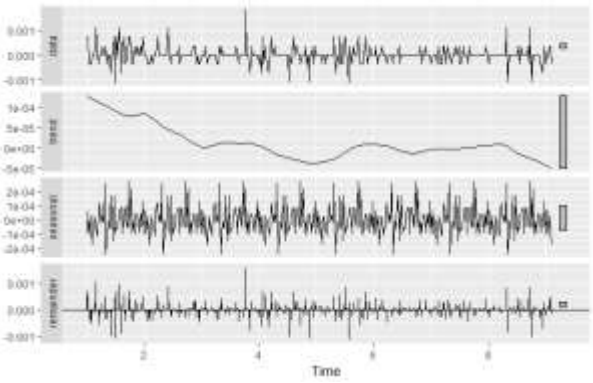


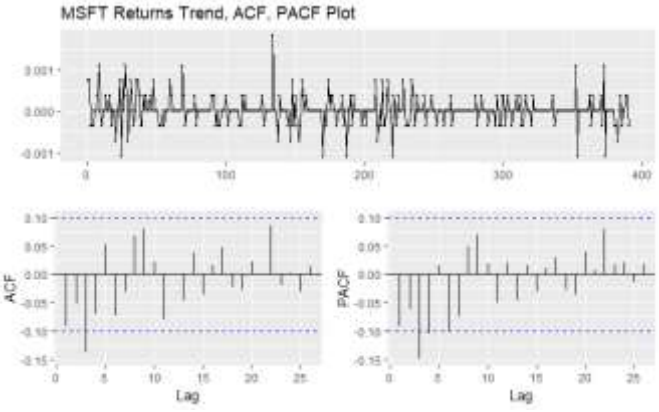


The time series shows an upward trend in the beginning and dips towards the end. There is no sign of stationarity. ACF is exponentially decreasing with a strong autocorrelation.

MSFT Minute Return Analysis:

```
MSFT_price = MSFT_day$close
MSFTreturns =diff(MSFT_price)/MSFT_price[-length(MSFT_price)]


qplot(MSFTreturns, fill = I("blue"), alpha = I(.5), col= I("red"),xlab = "MSFT Returns",  main="Histogram of MSFT
Returns", bins = 30)


MSFT_returns = ts(MSFTreturns, frequency = 48)

autoplot(stl(MSFT_returns, s.window = 'periodic'), ts.colour = 'blue')

ggtsdisplay(MSFTreturns, main = "MSFT Returns Trend, ACF, PACF Plot")
```

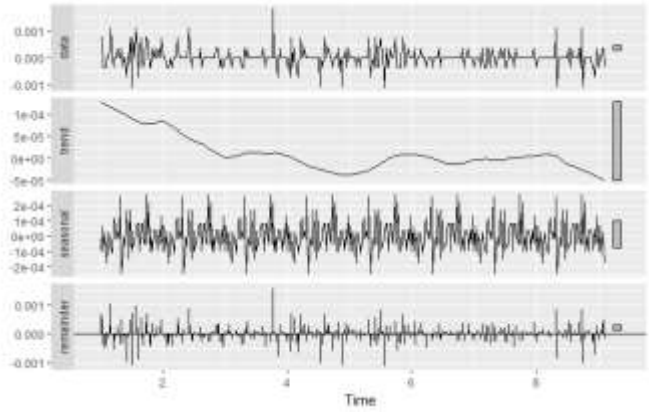


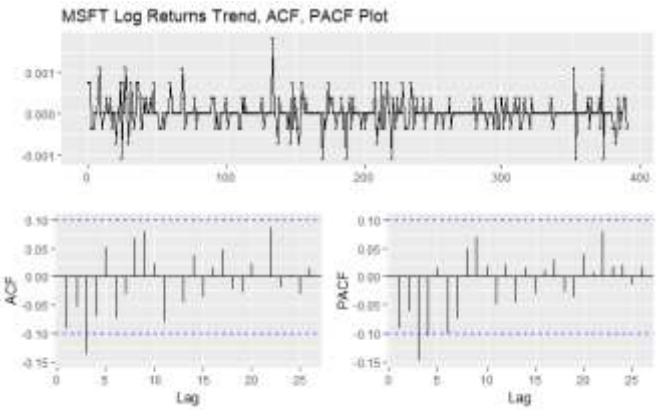


The minute returns after differencing is close to stationary. There is a downward trend but variance seems to increasing and decreasing at some intervals. There is significant auto-correlation between the lags.

MSFT Log Return Analysis:

```
MSFT_logreturns1=diff(log(MSFT_day$close))

qplot(MSFT_logreturns1, fill = I("blue"), alpha = I(.5), col= I("red"),xlab = "MSFT Log Returns",  main="Histogram of
MSFT Log Returns", bins = 30)

MSFT_logreturns2 = ts(MSFT_logreturns1, frequency = 48)

autoplot(stl(MSFT_logreturns2, s.window = 'periodic'), ts.colour = 'blue')

ggtsdisplay(MSFT_logreturns1, main = "MSFT Log Returns Trend, ACF, PACF Plot")
```

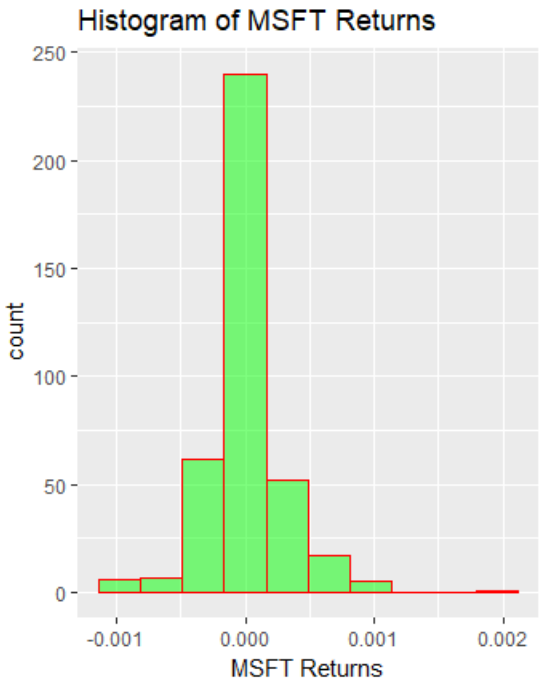


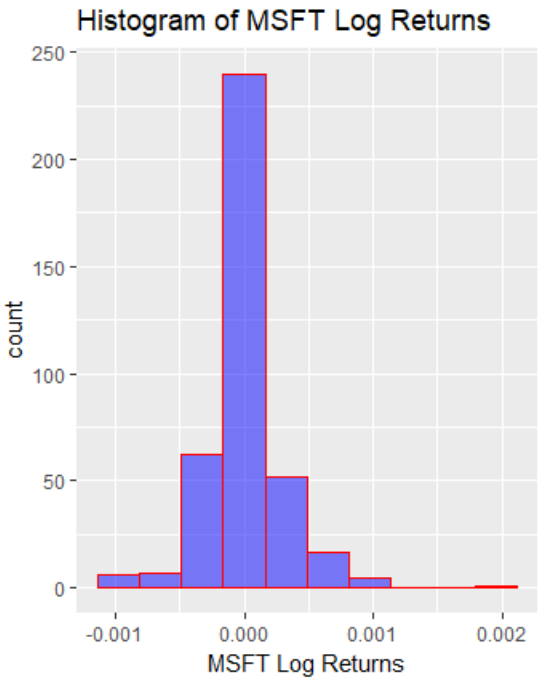


The log returns after differencing is close to stationary. There is a downward trend but variance seems to increasing and decreasing at some intervals. There is significant auto-correlation between the lags.

An interesting observation here is, you can see that there is no significant difference between the daily return and log return. They both look similar.

```
require(gridExtra)
Minute_Returns = qplot(MSFTreturns, fill = I("green"), alpha = I(.5), col= I("red"),xlab = "MSFT Returns",
main="Histogram of MSFT Returns", bins = 10)

Log_Returns=qplot(MSFT_logreturns1, fill = I("blue"), alpha = I(.5), col= I("red"),xlab = "MSFT Log Returns",
main="Histogram of MSFT Log Returns", bins = 10)
grid.arrange(Minute_Returns, Log_Returns, ncol=2)
```





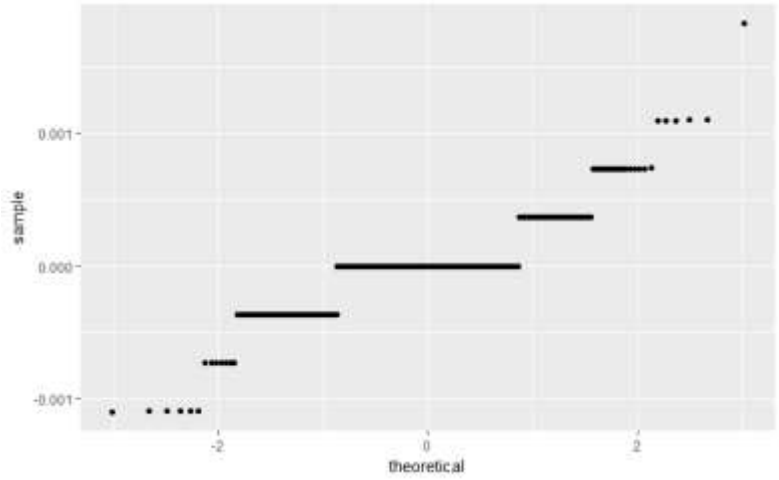
As observed the histograms of both MSFT minute returns and MSFT log returns look the same. So we can either use MSFT minute returns or MSFT log returns to build our forecast model.

We will be using the MSFT minute return to build the forecast model.

MSFT minute returns Statistical Analysis

```
ggplot(data.frame(MSFTreturns), aes(sample=MSFTreturns))+stat_qq()
t.test(MSFTreturns)

## Normality tests

normalTest(MSFTreturns,method="jb")

ggtsdisplay(MSFTreturns, main = "MSFT Returns Trend, ACF, PACF Plot")

#Independence test
Box.test(MSFTreturns,lag=10)

Box.test(MSFTreturns,lag=10,type="Ljung")
```

From the QQ plot the data is not normal because it is not a straight line and data is not continuous as this looks like step function.



p-value is greater than alpha so we fail to reject null hypothesis so the true mean is equal to 0

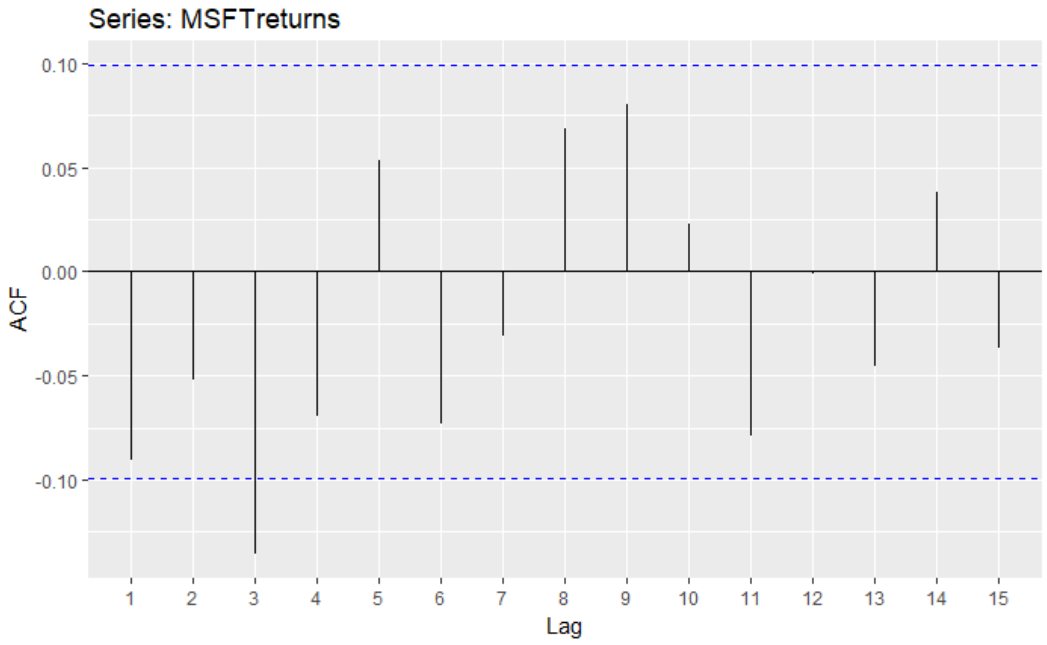
```
        One Sample t-test

data:  MSFTreturns
t = 0.66117, df = 389, p-value = 0.5089
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.241481e-05  4.512925e-05
sample estimates:
   mean of x
1.135722e-05
```

JB test for normality: p value is less than alpha so data does not follow a normal distribution

Box-Pierce test: p value is less than alpha we reject null hypothesis so we can say the lags are not independent

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 270.6437
  P VALUE:
    Asymptotic p Value: < 2.2e-16

Description:
 Sun Mar 18 14:43:57 2018 by user: nikhil


        Box-Pierce test

data:  MSFTreturns
X-squared = 21.415, df = 10, p-value = 0.01838


        Box-Ljung test

data:  MSFTreturns
X-squared = 21.771, df = 10, p-value = 0.01632
```

The ACF shows some correlation



Model Building MSFT minute returns

```
model1=ar(MSFTreturns,method='mle')
model1
model1_residuals = model1$resid[!is.na(model1$resid)]

ggtsdisplay(model1_residuals, main = "Model 1 residuals Trend, ACF, PACF Plot")
```

```
ar(x = MSFTreturns, method = "mle")

Coefficients:
      1       2       3       4       5       6       7
-0.1262  -0.0930  -0.1866  -0.1256  -0.0016  -0.1112  -0.0764

Order selected 7  sigma^2 estimated as  1.078e-07

        Box-Ljung test

data:  model1_residuals
X-squared = 2.923, df = 10, p-value = 0.9832
```

p-value is greater than alpha so the residuals are independent. Not a good model.

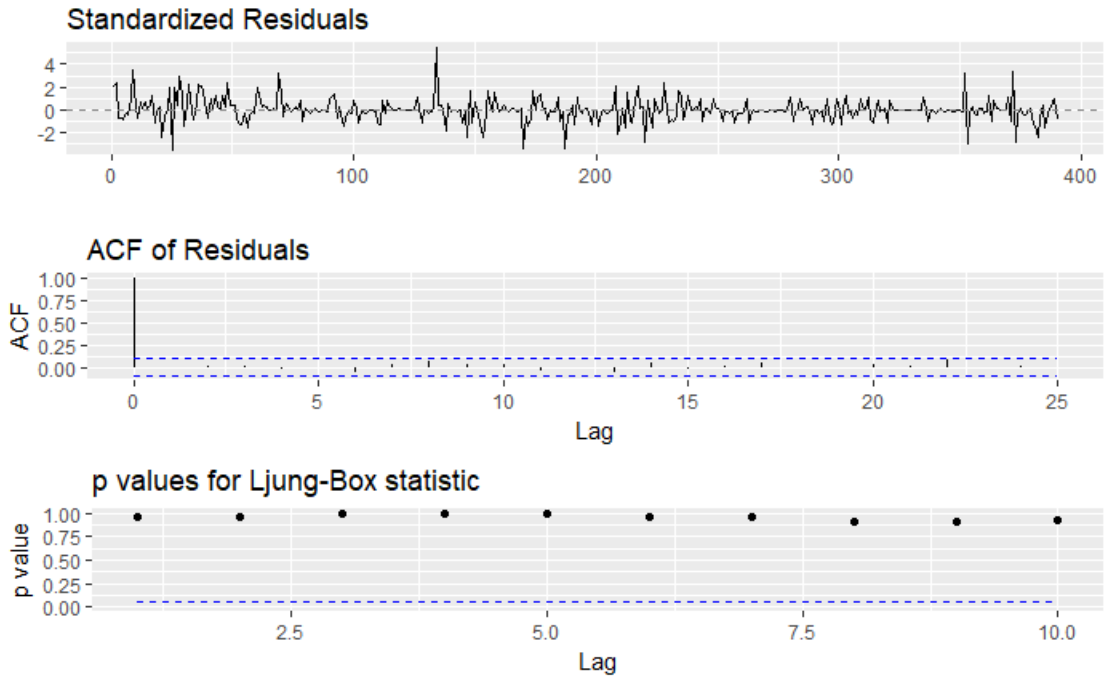
We build 5 other arima models. We select arima(1,0,1) because both the AR and MA coefficients are significant and also the AIC is the best compared to other models.

Model	AIC	Significant coefficients
AR("mle")		✖
arima(4,0,3)	-5131	✓
arima(3,0,4)	-5130.99	✓
arima(3,0,3)	-5132.95	✓
arima(6,0,6)	-5131.08	✓

ARIMA(3,0,3) model MSFT daily returns

```
ggtsdiag(model4)
```

```
poly1=c(1,-model4$coef[1:3])
roots=polyroot(poly1)
roots

Mod(roots)
```




```
[1] -0.082622+1.163864i -0.082622-1.163864i  2.018890-0.000000i
[1] 1.166793 1.166793 2.018890
```

The ACF plot shows a good stationary model. P-values are well above alpha 0.05. The roots are also above 1 so this is stationary.

Prediction

We use ARIMA(1,0,1) to forecast values for next 10 days for the S&P500 daily returns

```
$pred
Time Series:
Start = 1007
End = 1016
Frequency = 1
 [1] 0.0001434447 0.0001593473 0.0001743318 0.0001884512 0.0002017555 0.0002142917 0.0002261041 0.0002372346
 [9] 0.0002477226 0.0002576050

$se
Time Series:
Start = 1007
End = 1016
Frequency = 1
 [1] 0.007788597 0.007792141 0.007795287 0.007798079 0.007800557 0.007802757 0.007804709 0.007806442 0.007807981
[10] 0.007809346
```

We use ARIMA(3,0,3) to forecast values for next 10 minutes for the MSFT minute returns

```
$pred
Time Series:
Start = 391
End = 400
Frequency = 1
 [1] -4.556956e-05  2.140584e-05  7.631031e-05  7.302384e-06 -3.116336e-05  2.096842e-05
 [7]  4.129743e-05 -2.540676e-07 -1.053503e-05  2.103968e-05

$se
Time Series:
Start = 391
End = 400
Frequency = 1
 [1] 0.0003286605 0.0003309622 0.0003319559 0.0003366178 0.0003370852 0.0003377816
 [7] 0.0003377817 0.0003384441 0.0003384442 0.0003387435
```

The error is constant and very low so both the above models are good for forecasting.

Conclusion

- We need to de-difference the prediction and compare it with the actual values to see how well both the models actually predict and how much is the actual forecast error

- We can also try better models like ARCH/GARCH models and see how they perform

- We can also try different transformations like Cox and see if there is any change in the data