Assignment 3

Udit Pant

22/10/2019

```
library("fUnitRoots")
library(lmtest)
library(FitAR)
library("forecast")
library(mlbench)
library(dplyr)
library(fitdistrplus)
library(logspline)
library(tseries)
library(questionr)
```

Question 1

Seasonality: It is the phenomenon of a systematic and calendar-related effect showing up in a time series which can often mask the true underlying movements in a time series. Seasonality can be identified by regularly-spaced crests and troughs having a consistent direction.

Trend: A trend is a long-term movement in a time series. It can be identified by observing increase or decrease of a certain quantity over time.

a) Non-stationary: Trend visible in the graph.

b) Stationary: Except an outlier, the graph shows stationary behaviour as no statistical parameters seem to be changing.

c) Non-stationary: Trends visible.

d) Non-stationary: Seasonality - crests and troughs in the graph indicate periodicity.

e) Non-stationary: Trends visible.

f) Non-stationary: Trends visible

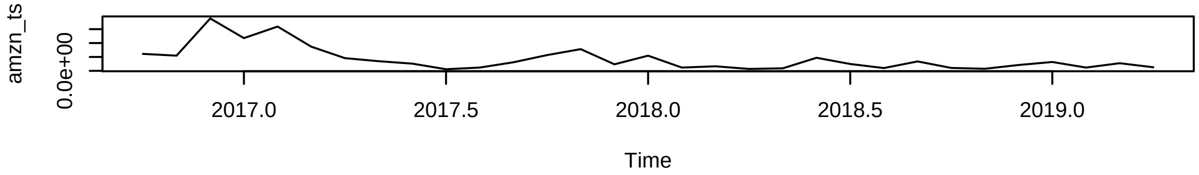
g) Stationary: Statistial parameters constant over time.

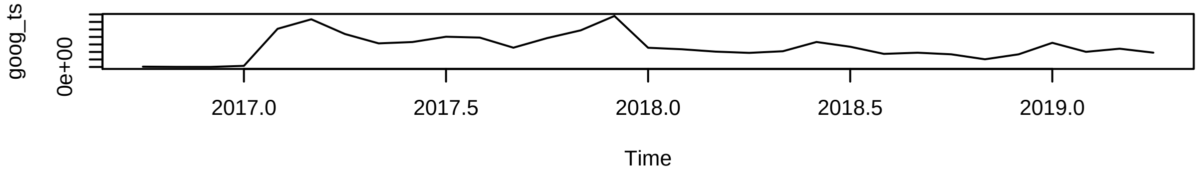
h) Non-stationary: Seasonality - periodicity.

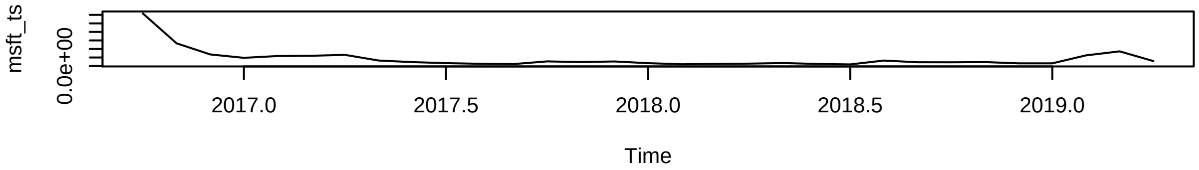
i) Non-stationary: Seasonality and Trends visible in graph.

Question 2

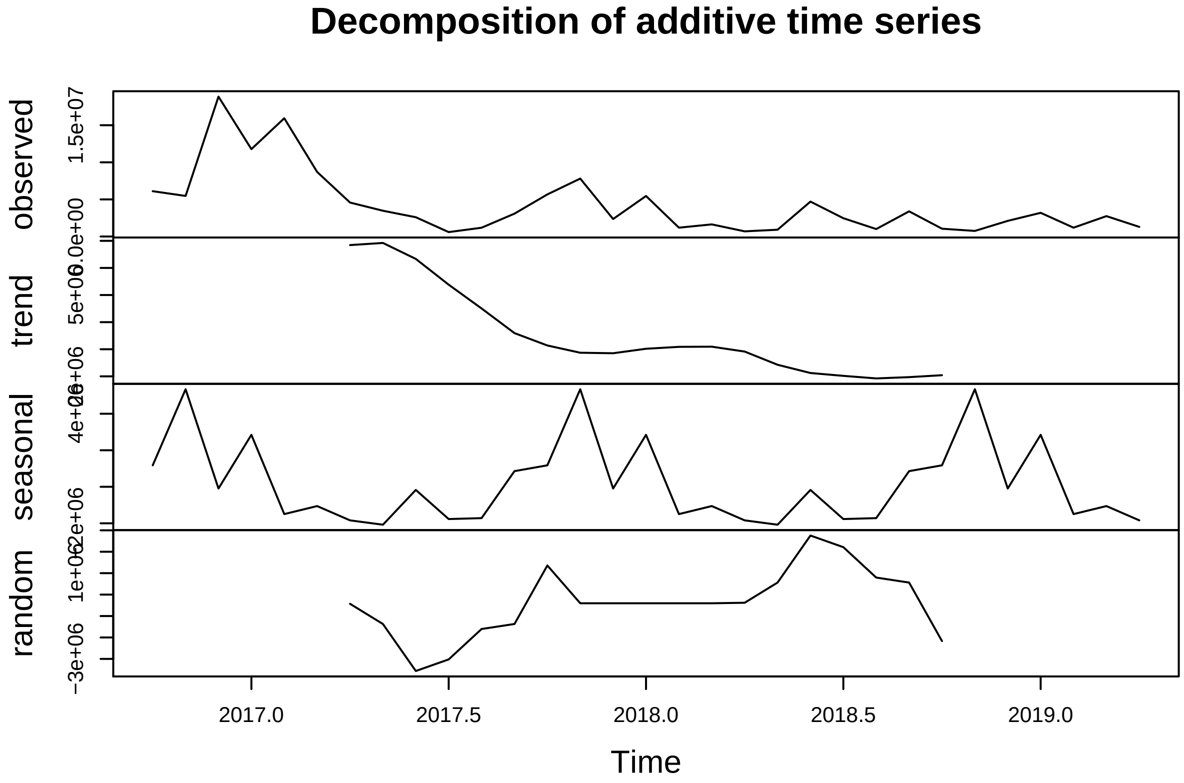
Out of the three years data, a major chunk was taken as training data. The remaining was reserved as testing data.

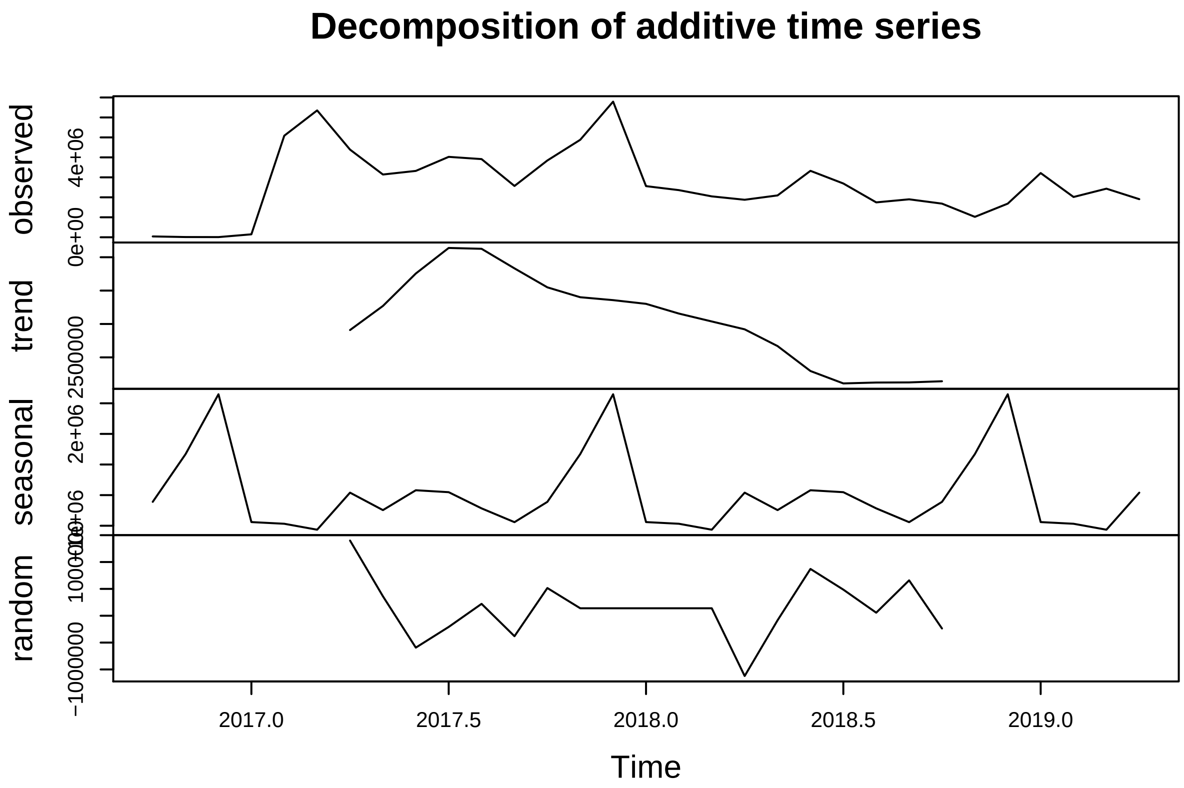


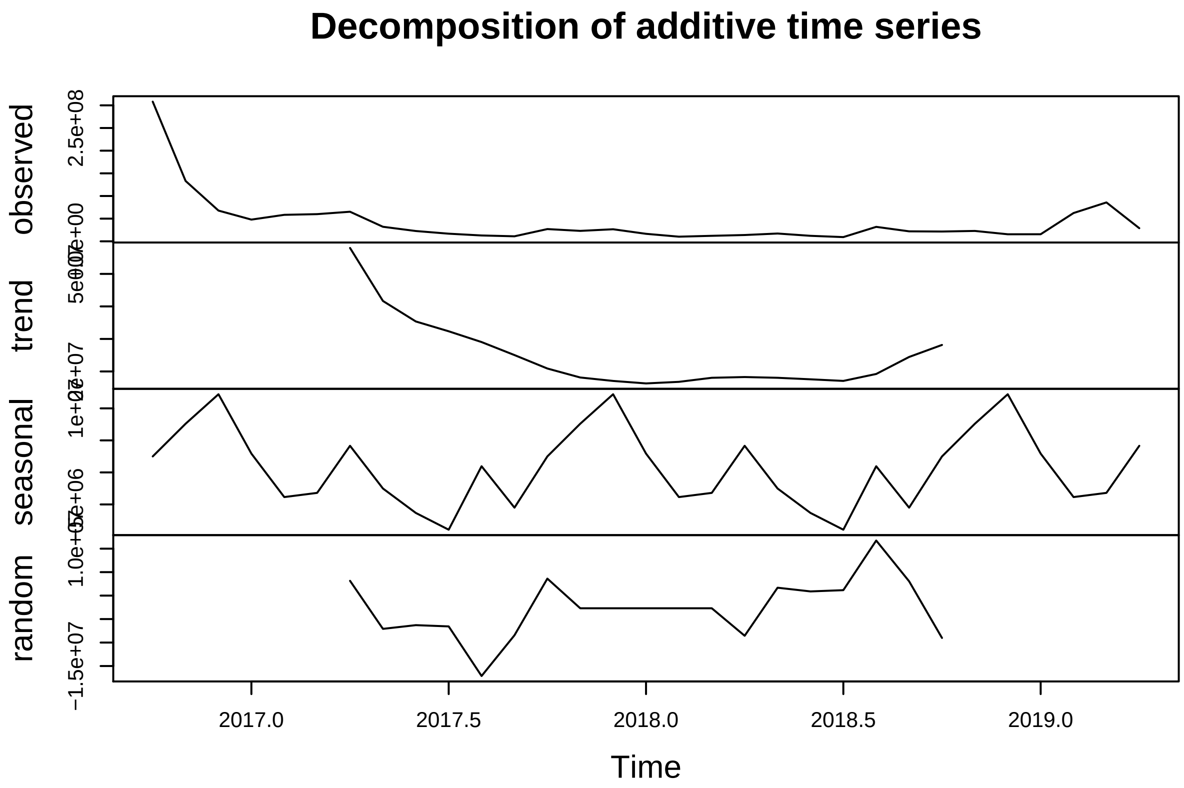




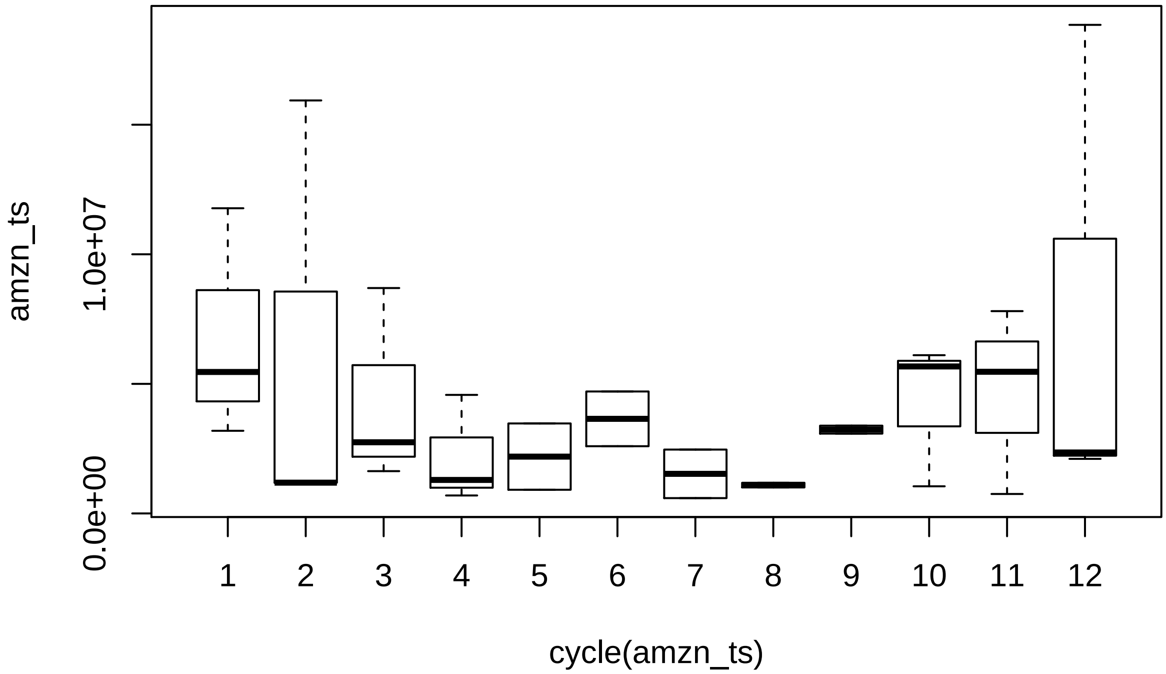
Decomposition of the time-series data yields plots about non-stationarity and seasonality present in the data.

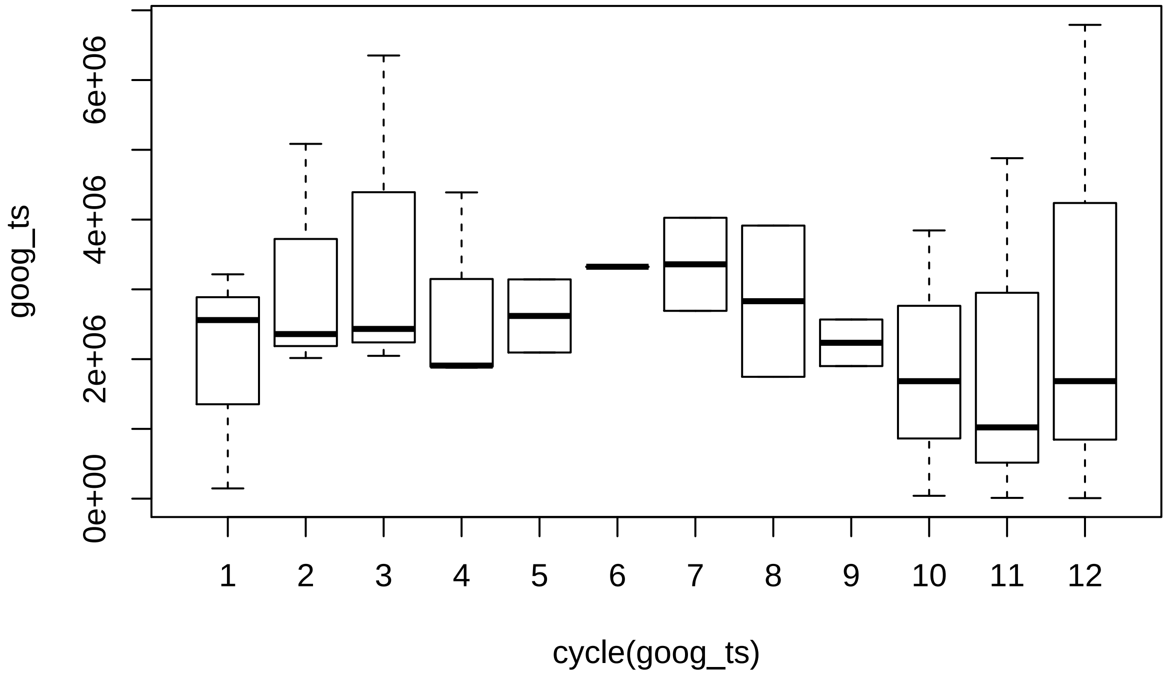

Decomposition of additive time series
observed
1.5e+07
0.0e+00
trend
6e+06
5e+06
2e+06
seasonal
4e+06
2e+06
random
2e+06
1e+06
−3e+06
2017.0
2017.5
2018.0
2018.5
2019.0
Time

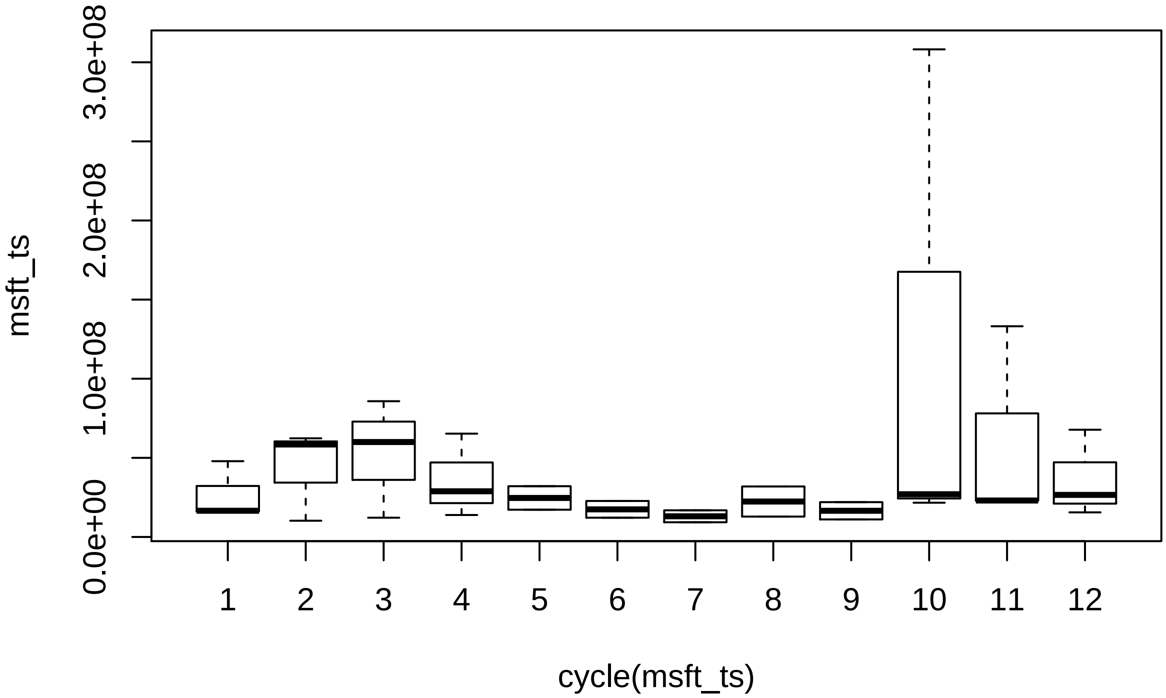

Decomposition of additive time series
observed
0e+00
4e+06
trend
2500000
seasonal
0e+00
2e+06
random
−1000000
1000000
2017.0
2017.5
2018.0
2018.5
2019.0
Time



Box-plots present a neat picture about how the stock volumes are varying across the months.


amzn_ts
0.0e+00
1.0e+07
1
2
3
4
5
6
7
8
9
10
11
12
cycle(amzn_ts)


goog_ts
0e+00
2e+06
4e+06
6e+06
1
2
3
4
5
6
7
8
9
10
11
12
cycle(goog_ts)

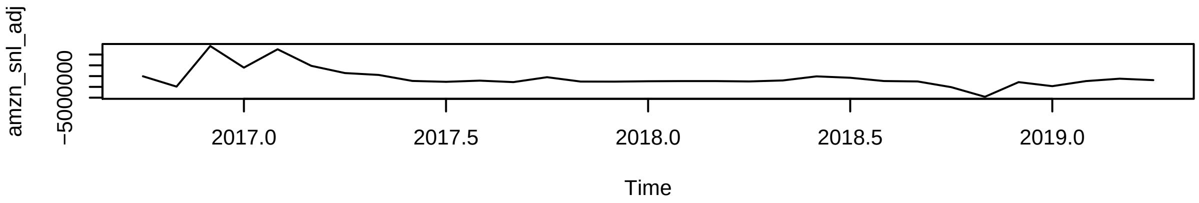


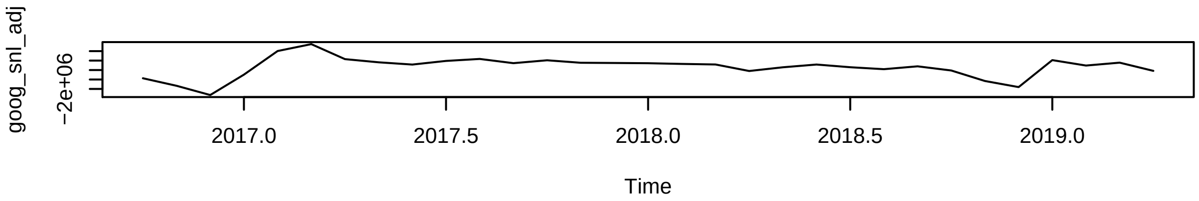
Inferences

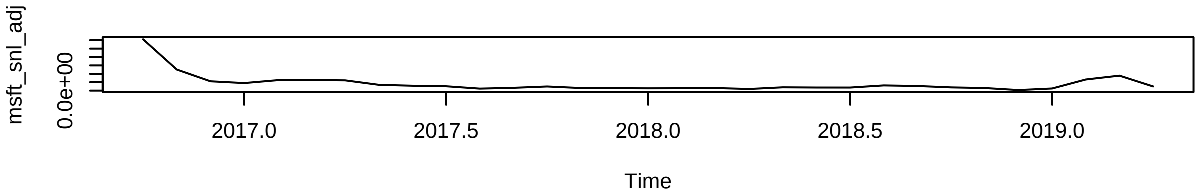
1. We can observe seasonal effects present in the data as during particular months (e.g. October in Microsoft Stock data), the stock volumes tend to drastically change.
2. The increase adnd decrease in stok volumes indicate a trend.
3. These two factors contribute to non-stationarity of data.
4. Comparatively, in order of decreasing health, as per the graphs, Google stands at the top followed by Amazon and then Microsoft.

As we learnt earlier about the seasonality present in data, in order to make it stationary, we must rmove the existing seasonality.

Removing Seasonality

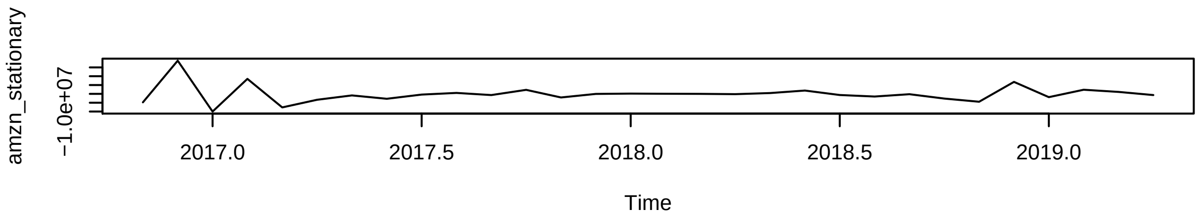


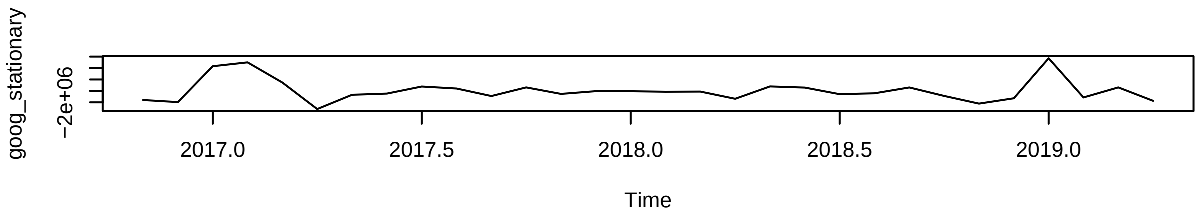


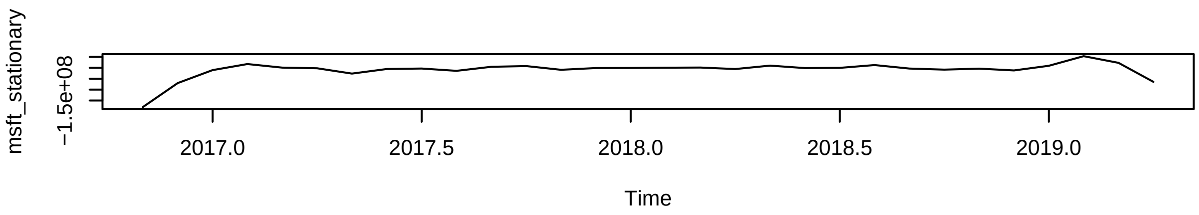


By taking a difference with a lag, we make the data stationary.

Making Time series stationary



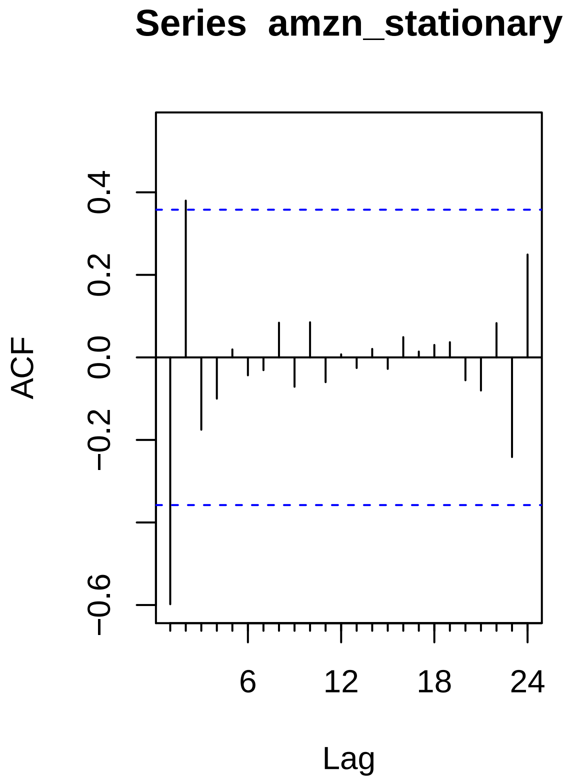


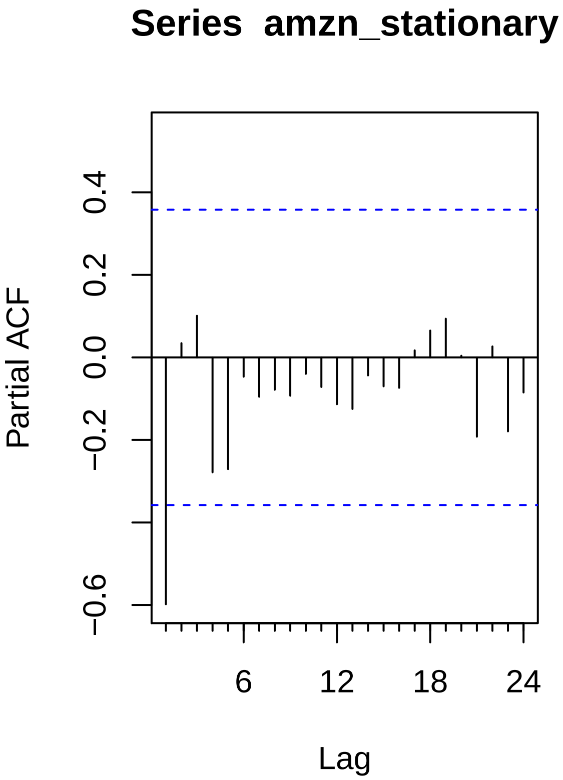


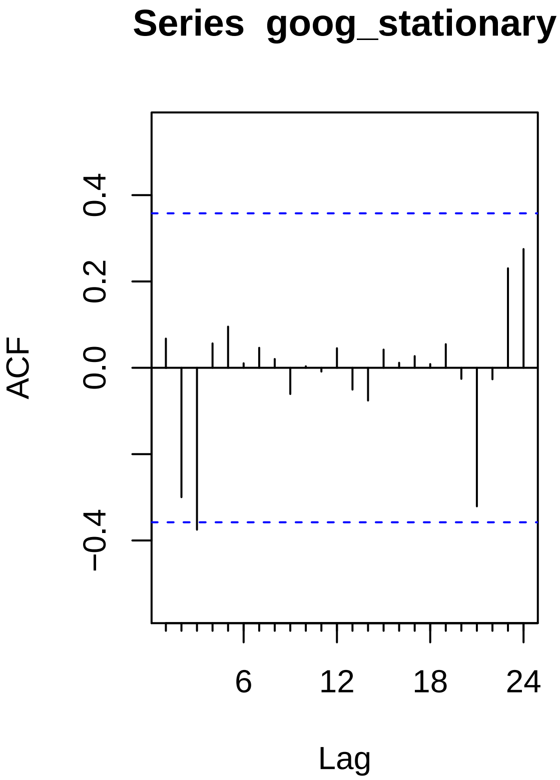
The auto-correlation function (Acf) plot helps determine the Q value which corresponds to moving average.

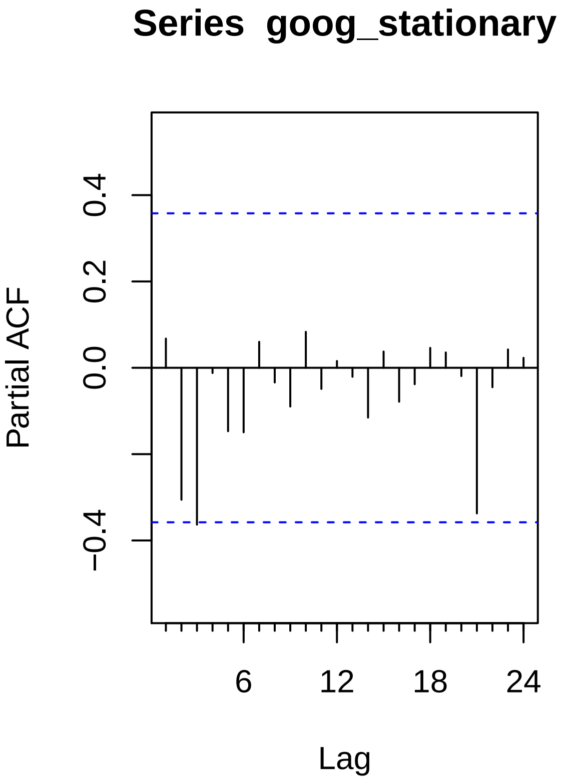
Similarly, the partial auto-correlation function (Pacf) plot helps identify the P value which correpsonds to auto-correlation.

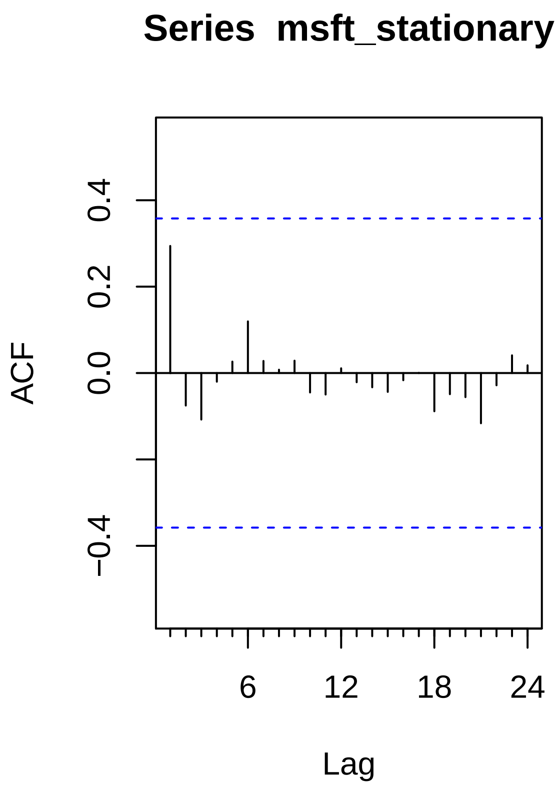
Since, we're observing two years data, we take max lag to be 24.

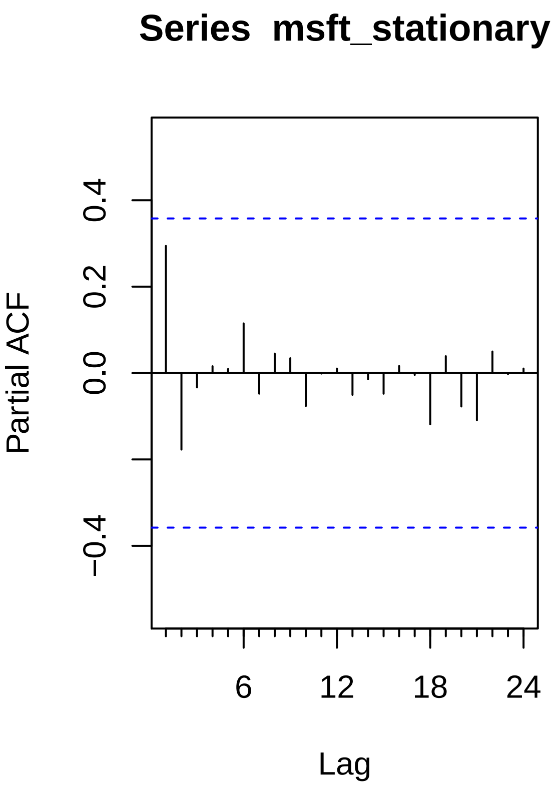

Series amzn_stationary
ACF
0.4
0.2
0.0
−0.2
−0.6
6
12
18
24
Lag


Series amzn_stationary
Partial ACF
0.4
0.2
0.0
−0.2
−0.6
6
12
18
24
Lag


Series goog_stationary
ACF
0.4
0.2
0.0
−0.4
6
12
18
24
Lag


Series goog_stationary
Partial ACF
0.4
0.2
0.0
−0.4
6
12
18
24
Lag

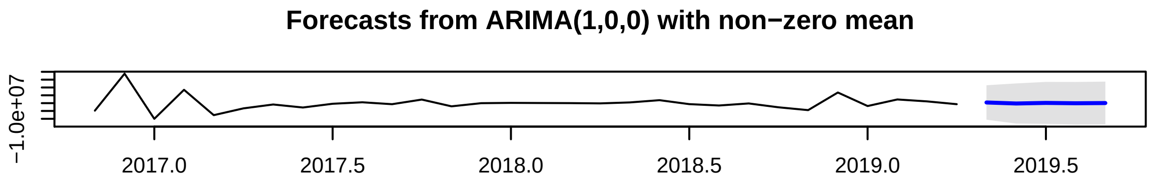




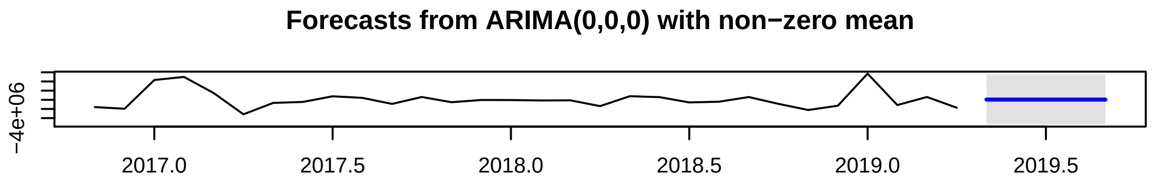
The order or (p,q) values were determined from the Acf and Pacf plots. d was taken to be zero.

Future predictions of stock price:

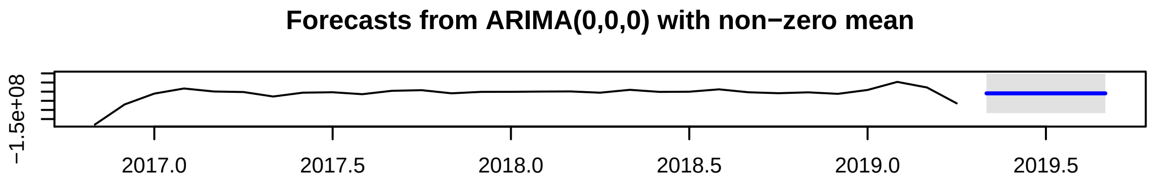
Forecast of the stocks:



Forecasts from ARIMA(1,0,0) with non−zero mean

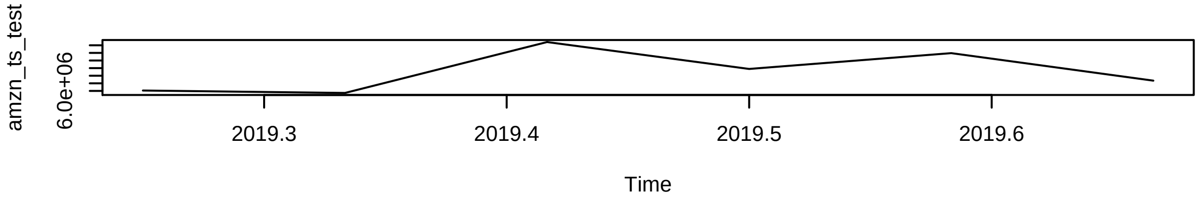


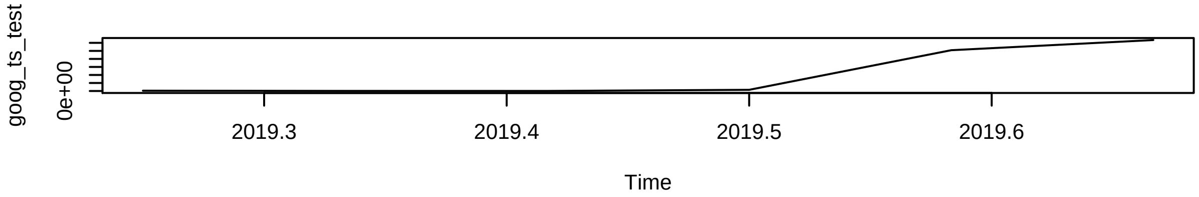
Forecasts from ARIMA(0,0,0) with non−zero mean

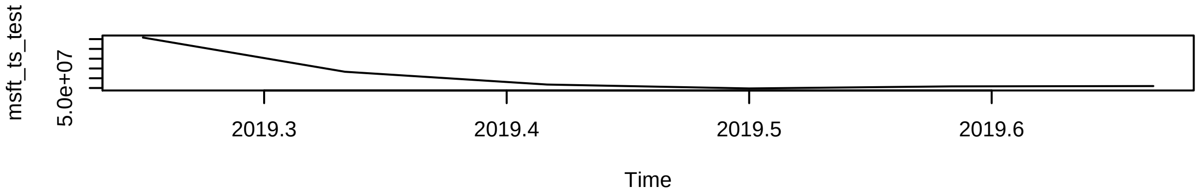


Forecasts from ARIMA(0,0,0) with non−zero mean

Plots of the test data:





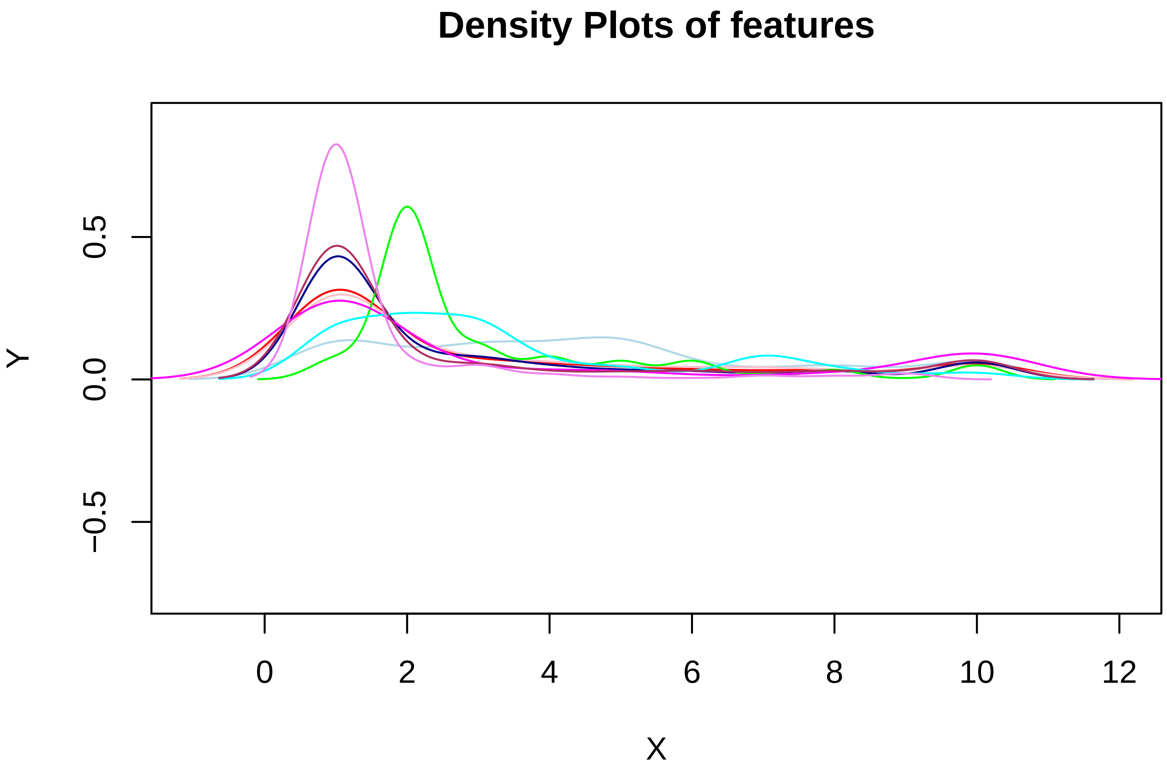


Question 3

Handling Missing Values

Since only 16 data points of “Bare.nuclei” had NA, those were dropped.

To understand the distributon our data might belong to, density plots of the features were used. We observe that the values of the features are greater than zero. The plots show a gaussian curve with mean greater than zero for every feature. Also, we must not ignore the fact that data across all features is discrete and not continuous. Therefore, we must use the poisson distribution which allows discrete values along with a log function as our link function which will limit the distribution to account only for positive values.



GLM model: Using a poisson distribution with log as the link function.

Summary Statistics:

Estimate values: Values of the unknown parameters Standard Error: Standard deviation of the parameters Adjusted R squared value: Goodness of fit AIC: Akaike Information criterion is an estimator of the quality of the model.

CI and odd ratios listed in the summary below.

Call: glm(formula = cancer_data$Class\ cancer_{d}ata$Cl.thickness * cancer_data$Cell.size * cancer_{d}ata$Cell.shape * cancer_data$Marg.adhesion * cancer_{d}ata$Epith.c.size * cancer_data$Bare.nuclei * cancer_{d}ata$Bl.cromatin * cancer_data$Normal.nucleoli * cancer_{d}ata$Mitoses, family = poisson(link = “log”), data = cancer_data)

Deviance Residuals: Min 1Q Median 3Q Max
0.000e+00 0.000e+00 0.000e+00 0.000e+00 2.107e-08

Null deviance: 1.0936e+02 on 682 degrees of freedom Residual deviance: 5.7287e-14 on 203 degrees of freedom AIC: 2472.7

Number of Fisher Scoring iterations: 4

Estimate (Intercept) -1.533e+01 cancer_data$Cl.thickness 5.670e + 00 cancer_{d}ata$Cell.size -3.714e+01 cancer_data$Cell.shape - 2.618e + 01 cancer_{d}ata$Marg.adhesion 3.985e+01 cancer_data$Epith.c.size 2.208e + 01 cancer_{d}ata$Bare.nuclei 2.799e-01 cancer_data$Bl.cromatin - 1.622e + 01 cancer_{d}ata$Normal.nucleoli 5.731e+00 cancer_data$Mitoses 1.468e+01

References

1. Amazon Stock Data

2. Google Stock Data
3. Microsoft Stock Data
4. Time Series Data Analysis