# A close up of a sign Description automatically generated

**Forecasting and Time Series Methods**

**BANA 7050**

**Final Project**

**Submitted by:**

**Pratik Gundawade (M13262480)**

**Kaustubh Saraf (M13457296)**

# Problem Statement

Forecast the number of shares sold of Amazon Inc. based on latest 5 year data

# Amazon Share Data

We acquired the data from Yahoo finance. The data has opening and closing share price, along with total number of shares traded of Amazon Inc. on month-to-month basis. The date ranged from May 2015 to April 2020.

# Exploratory Data Analysis

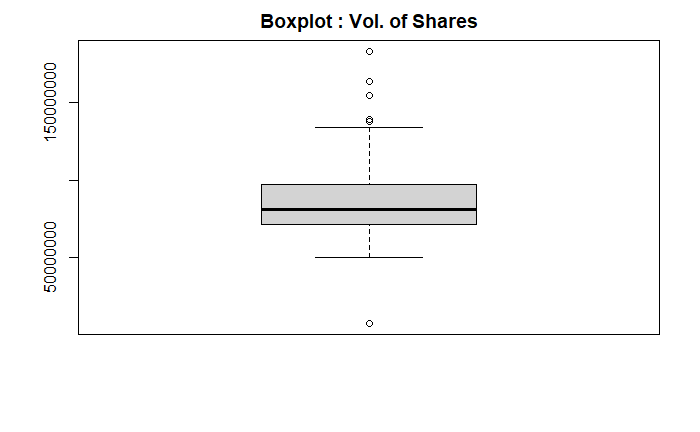
We first plotted the time series data to gauge on the stationarity.

|  |  |
| --- | --- |
| C:\Users\gundawade.p\OneDrive - Procter and Gamble\Documents\Spring 2020\Forecasting\Project\Time-Series\images\4.png | C:\Users\gundawade.p\OneDrive - Procter and Gamble\Documents\Spring 2020\Forecasting\Project\Time-Series\images\3.png |

Observations:

1. The plot suggested that the series is stationary
2. The PACF plot suggests that the series might be generated from MA1
3. The ACF plot suggests that the series might be generated from AR1 or AR2

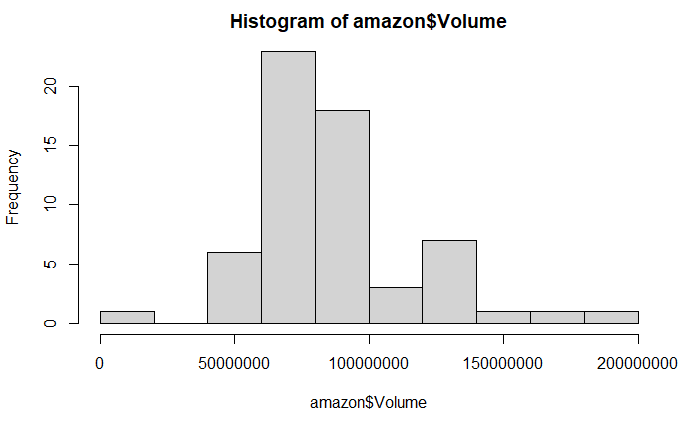
Since we wanted to forecast volume of shares, we wanted to gauge on outliers in the volume of shares traded. Following boxplot depicts range of values:



There were months when the number of shares traded were outside the quantile. However, this data was procured from Yahoo finance, we decided to trust the data at its source.

Distribution of the volume of shares traded

It appears that the total volume of shares of Amazon Inc. traded in a month lies between 60mn to 100mn



Additionally, the timestamp in this data was of number format, and we wanted it converted into names. Hence, we used the factor relevel method to get the data in desired format. Once we had the format, we filtered out data from 2015 and 2020.

# Time Series Modeling

## Methodology

We split the data into training and testing. Then we perform ADF test which suggests that the series is stationary and we do not need to take difference.

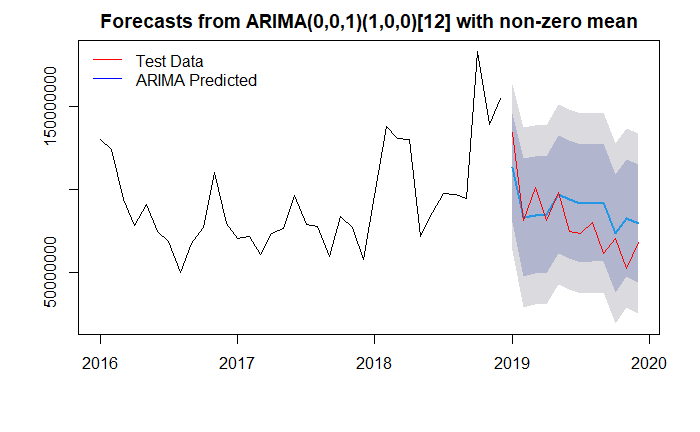
Augmented Dickey-Fuller Test

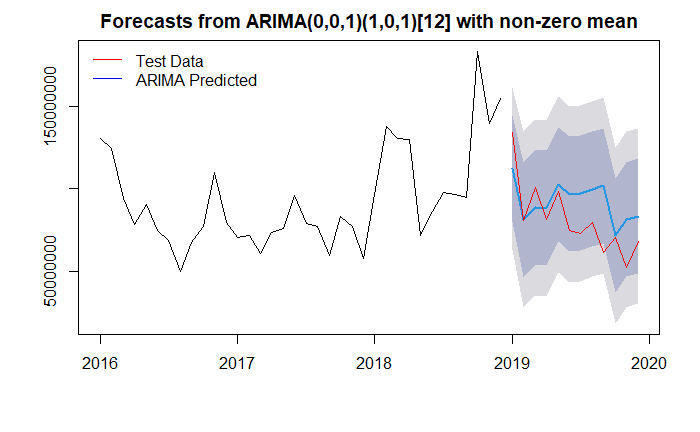
data: amazon.vol.ts

Dickey-Fuller = -3.4874, Lag order = 0, p-value = 0.05379

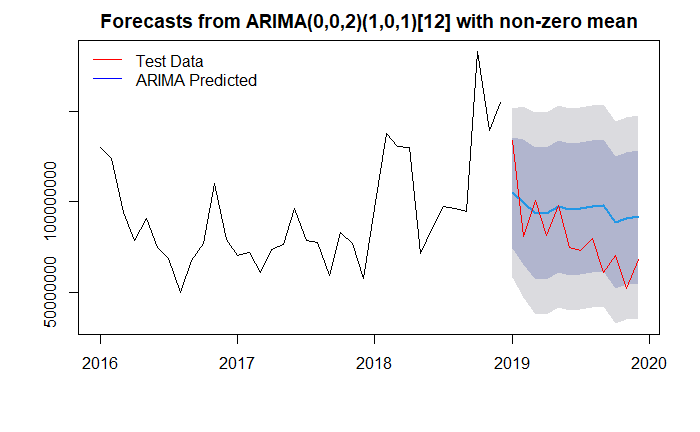
alternative hypothesis: stationary

Model 1 MA 1 Season=1, 0, 0

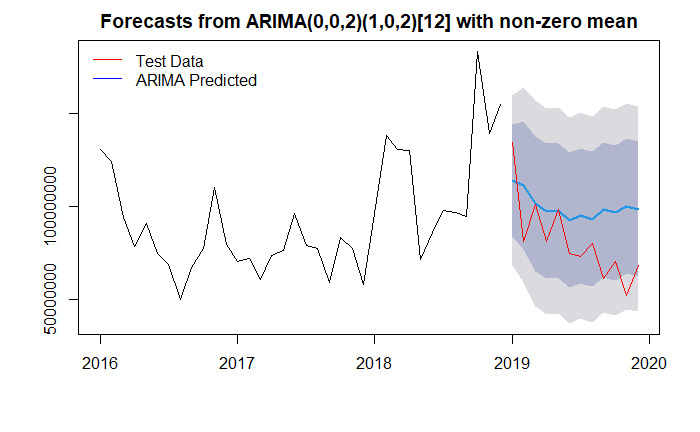


Model 2 MA 1 Season=1, 0, 1 

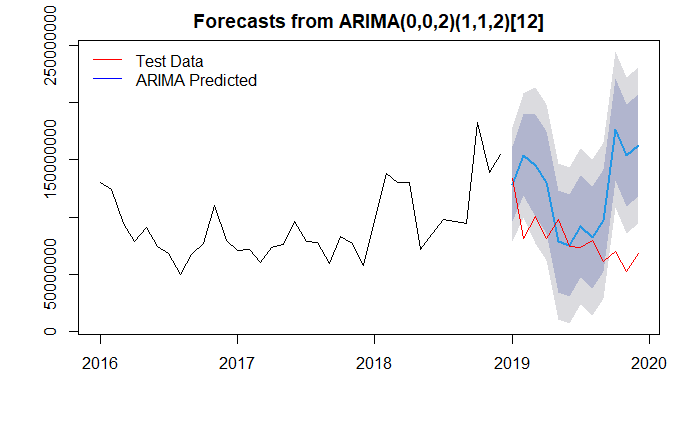
Model 3 MA 2 Season=1, 0, 1



Model 4 MA2 Season= 1, 0, 2

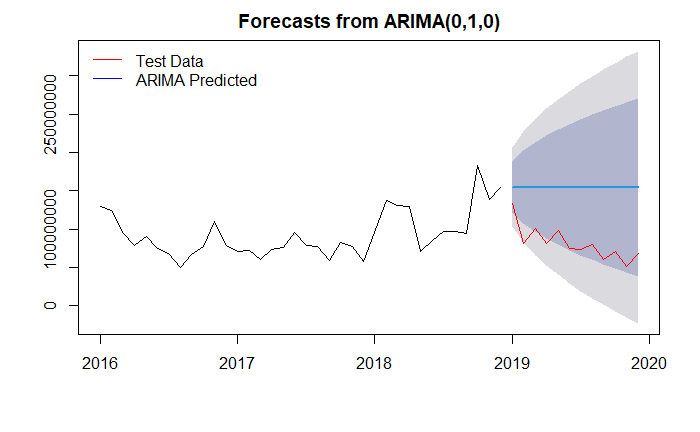


Model 5 MA 2 Season = 1, 1, 2



Auto ARIMA

Auto ARIMA suggests that the model is generated from AR1 series which was our initial understanding

**Model selection criterion**

|  |  |  |
| --- | --- | --- |
| Model | Train MAPE | Test MAPE |
| ma1.100 | 21.174764 | 19.56398 |
| ma1.101 | 21.085901 | 22.98210 |
| ma2.101 | 19.145619 | 28.76997 |
| ma2.102 | 17.979521 | 31.46609 |
| ma2.112 | 9.588964 | 66.07467 |
| Auto Arima (0,1,0) | 20.076593 | 101.55918 |

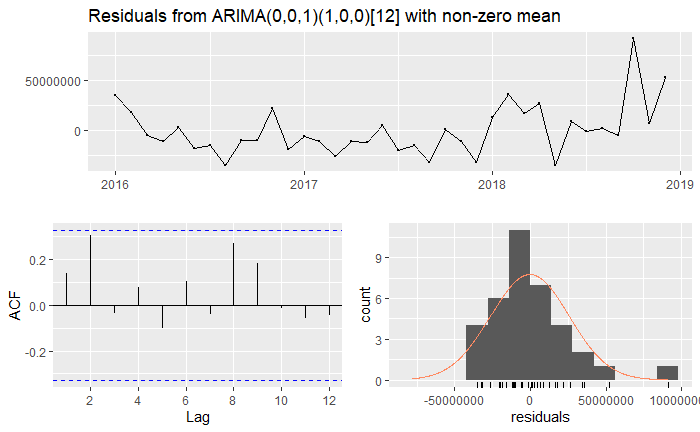
# Residual Diagnostic of the Best Model (Model 1 ma1.100)

Ljung-Box test

data: Residuals from ARIMA(0,0,1)(1,0,0)[12] with non-zero mean

Q\* = 5.8527, df = 4, p-value = 0.2104

Model df: 3. Total lags used: 7

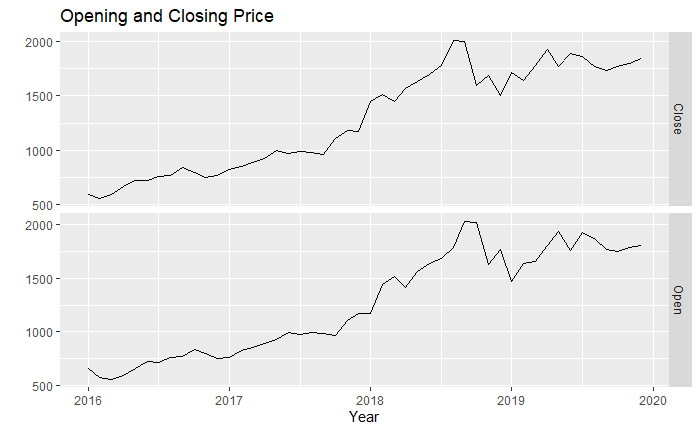


# Time Series Regression

There might be cases that one time series is a combination of the AR and MA components and at the same time depends on one or more Time series. In such cases we can use the ARIMA model along with the xreg input which models the time series as with the predictor variables.

In our case, we have tried to use this concept to forecast the Closing price of the stock given the Opening pric. This kind of model can help us predict the Closing value of the stock at the end of a month. Another underlying fact here is that there is a strong correlation between the Opening and Closing Price of the stock.

We start with checking if the linear regression model fits the data well. Below plot suggests that there might be association in these two variables.



Below is the summary of the linear model:

## Call:  
## lm(formula = Close ~ Open, data = amazon.vol.close.ts)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -418.83 -49.82 -2.97 51.92 252.69   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.42273 47.65039 1.457 0.152   
## Open 0.96312 0.03532 27.272 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 115.6 on 46 degrees of freedom  
## Multiple R-squared: 0.9418, Adjusted R-squared: 0.9405   
## F-statistic: 743.7 on 1 and 46 DF, p-value: < 2.2e-16

The p-value and R-squared suggests that the model holds good and we can use this in the xreg part in ARIMA.

## Fitting ARIMA model:

(fit <- Arima(amazon.vol.close.ts[,"Close"],

order=c(1,1,0), season=list(order=c(2,1,0)),

xreg=amazon.vol.close.ts[,"Open"]))

## Series: amazon.vol.close.ts[, "Close"]   
## Regression with ARIMA(1,1,0)(2,1,0)[12] errors   
##   
## Coefficients:  
## ar1 sar1 sar2 xreg  
## -0.3130 -1.1043 -0.7208 0.0530  
## s.e. 0.2653 0.1361 0.1613 0.2292  
##   
## sigma^2 estimated as 8678: log likelihood=-218.27  
## AIC=446.54 AICc=448.61 BIC=454.32

The AIC value here is lower than the auto.arima function. Hence we go ahead with this model. The below residual plot suggest that there is still one spike in ACF plot and we can investigate further. For the scope of this project, we will use this model.

