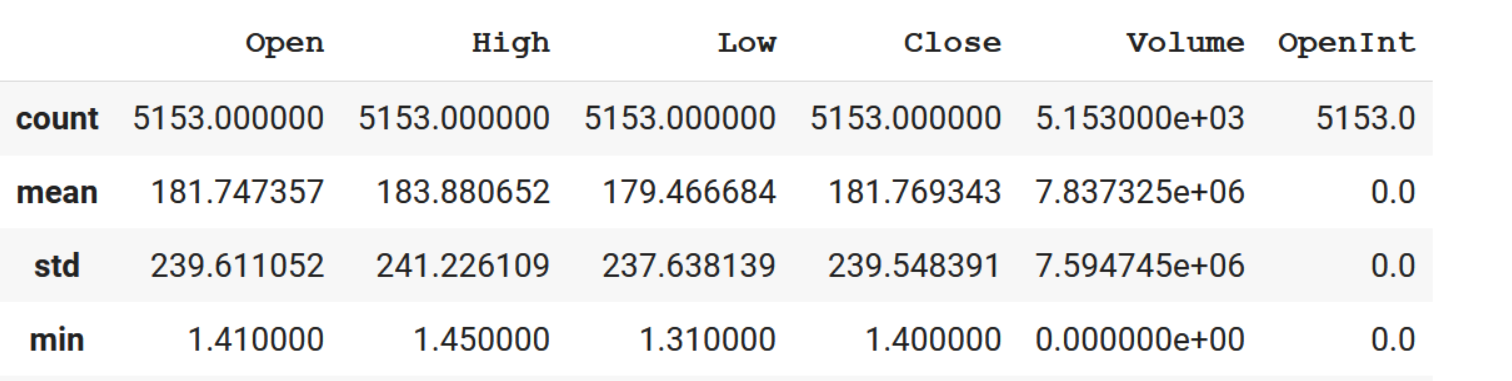
**Amazon Stock Forecasting**

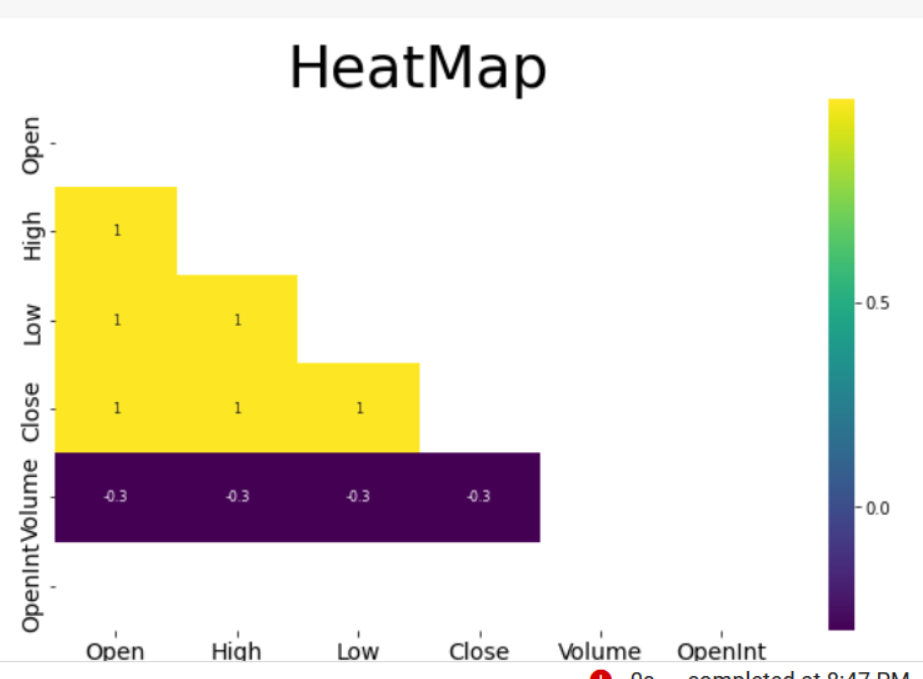
**Using ARIMA and Feed Forward Neural Network**

The Amazon stock prices from the given dataset is used to analyse and forecast. For the current analysis ,the volume of the Amazon Stock are analysed and modelled for forecasting. The data set consists of 5153 rows in columns Date, Open, High, Low, Close , Volume and OpenInt



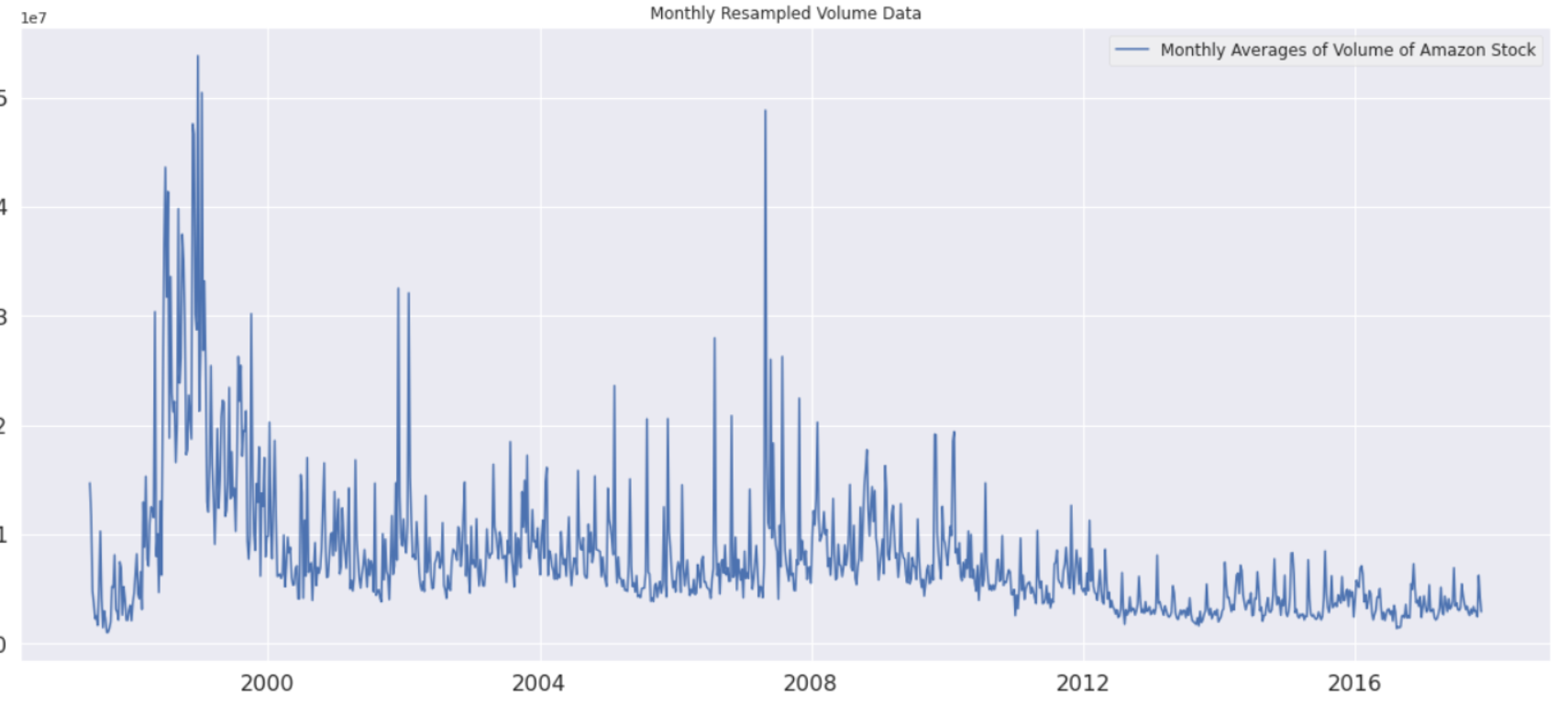
**Correlation**

A heatmap is generated using the columns and it shows Volume have very low correlation with other columns



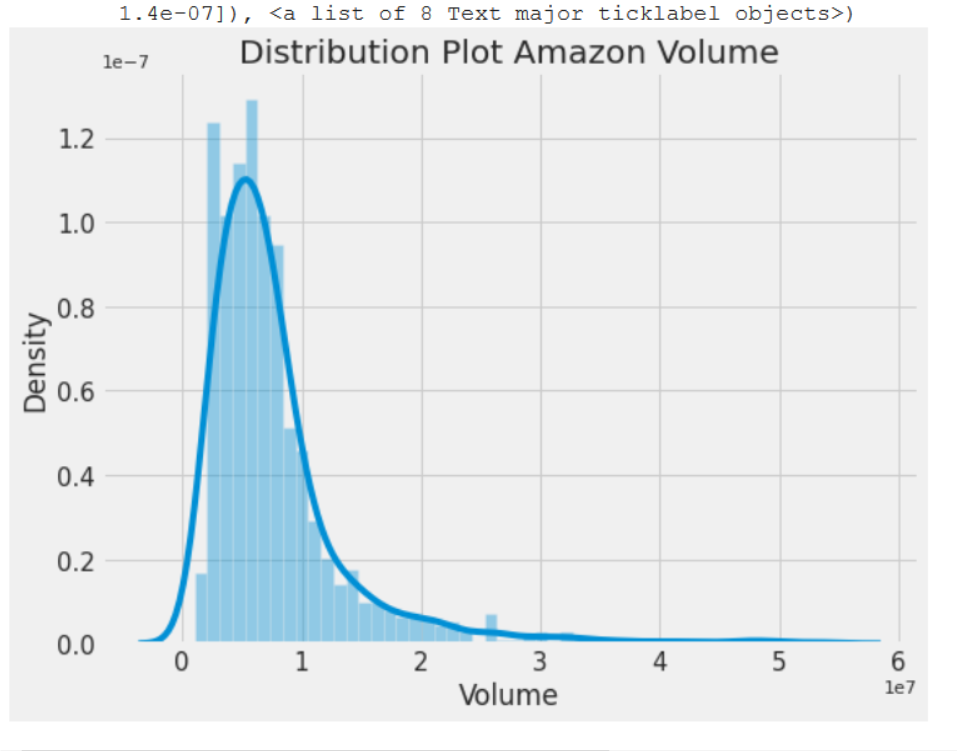
Since the data is daily data with high volatility. To make the analysis easier, the volume is used to compute monthly mean and the monthly mean volume is used for the further analysis. The index of the dataset is set to Date of the stock price.

The mean monthly volume is plotted as shown below



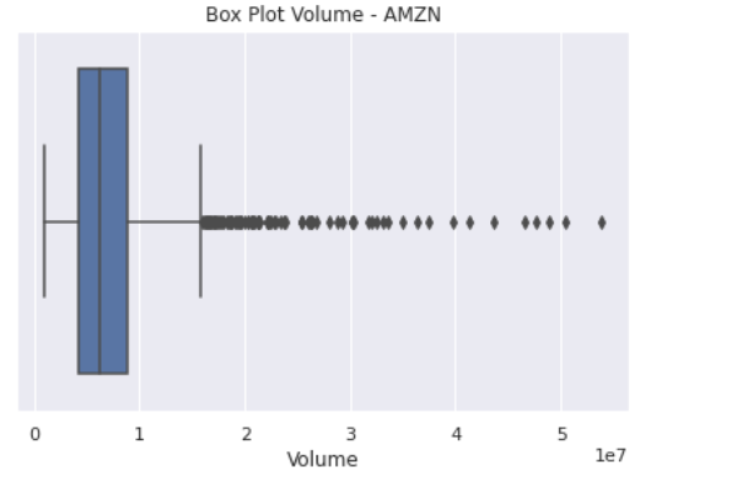
Exploratory Data Analysis

The volume distribution is plotted to view how volume is distributed over period of time.

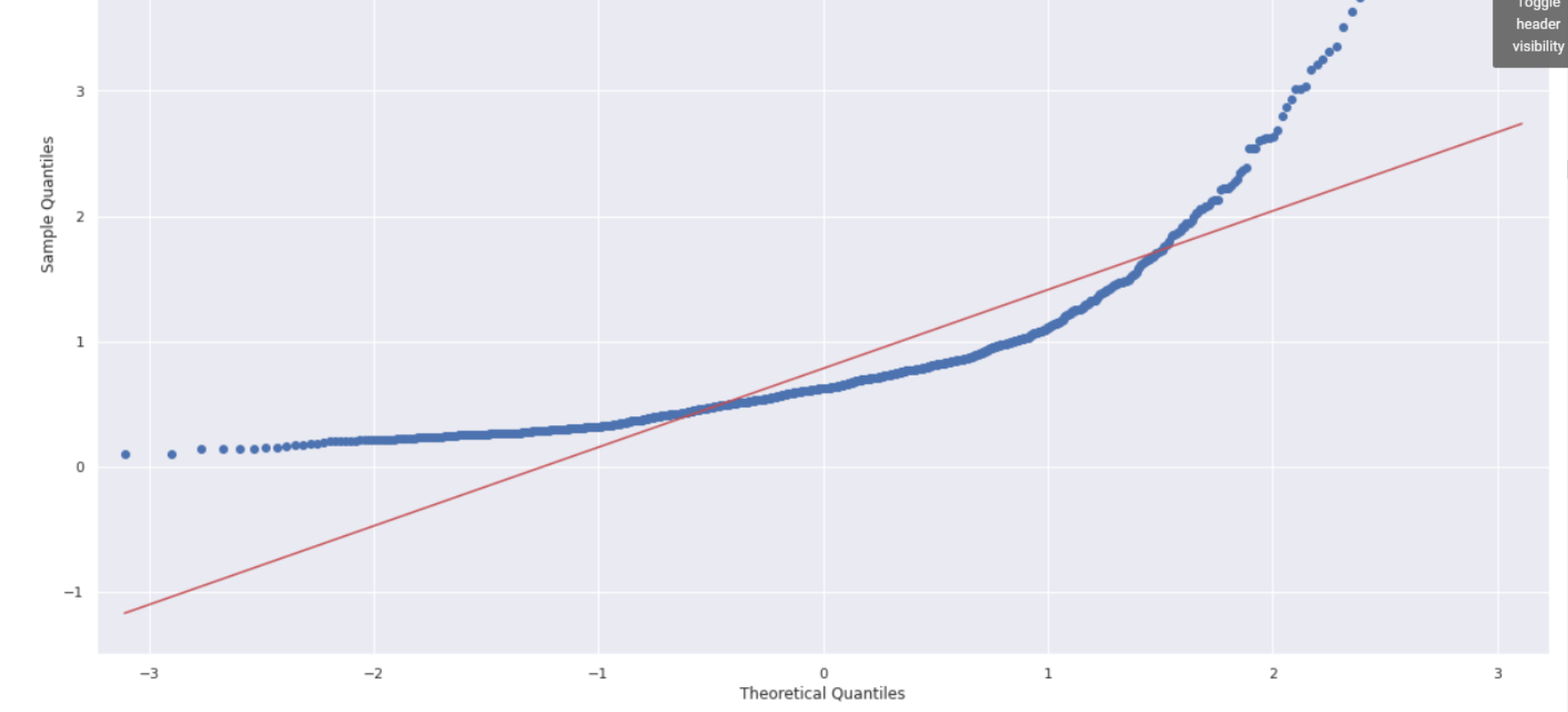


The volume of data is not uniformly distributed but is skewed more towards left indicating larger number of lower values than values in higher ranges.

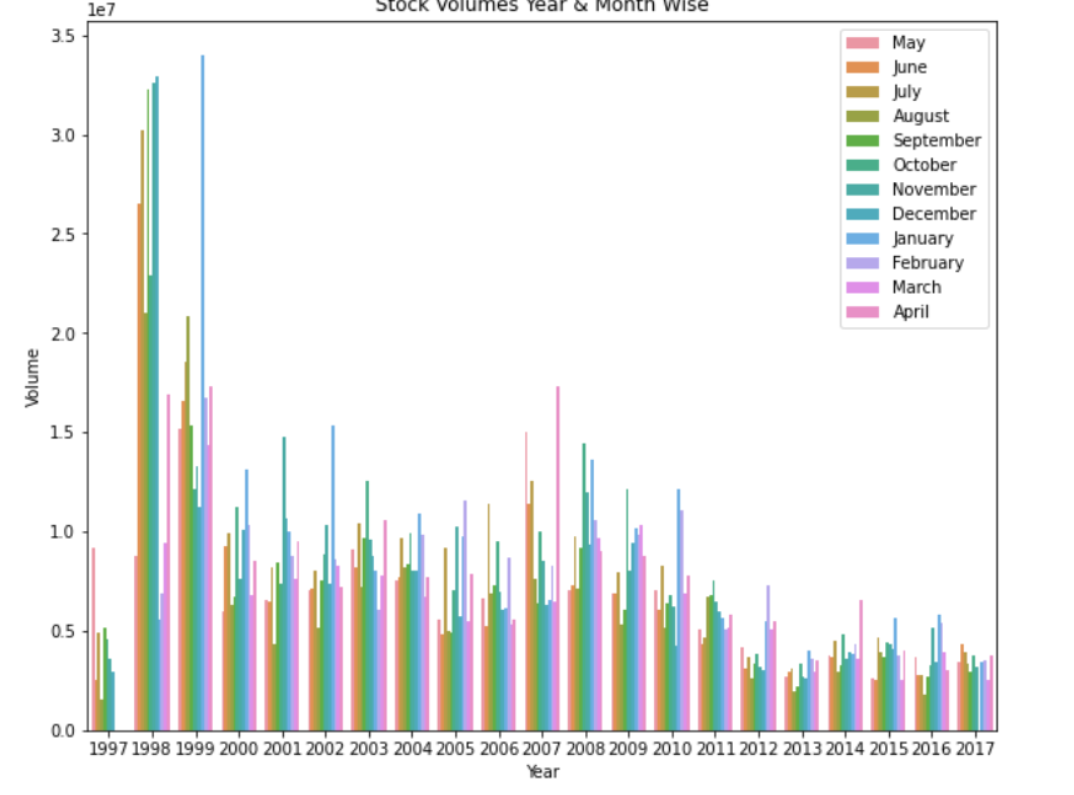
From the box plot, it appears that there are more outliers in the Volume data in the higher ranges.



The QQ plot of volume data shows that both skewness in right and left end of curves.

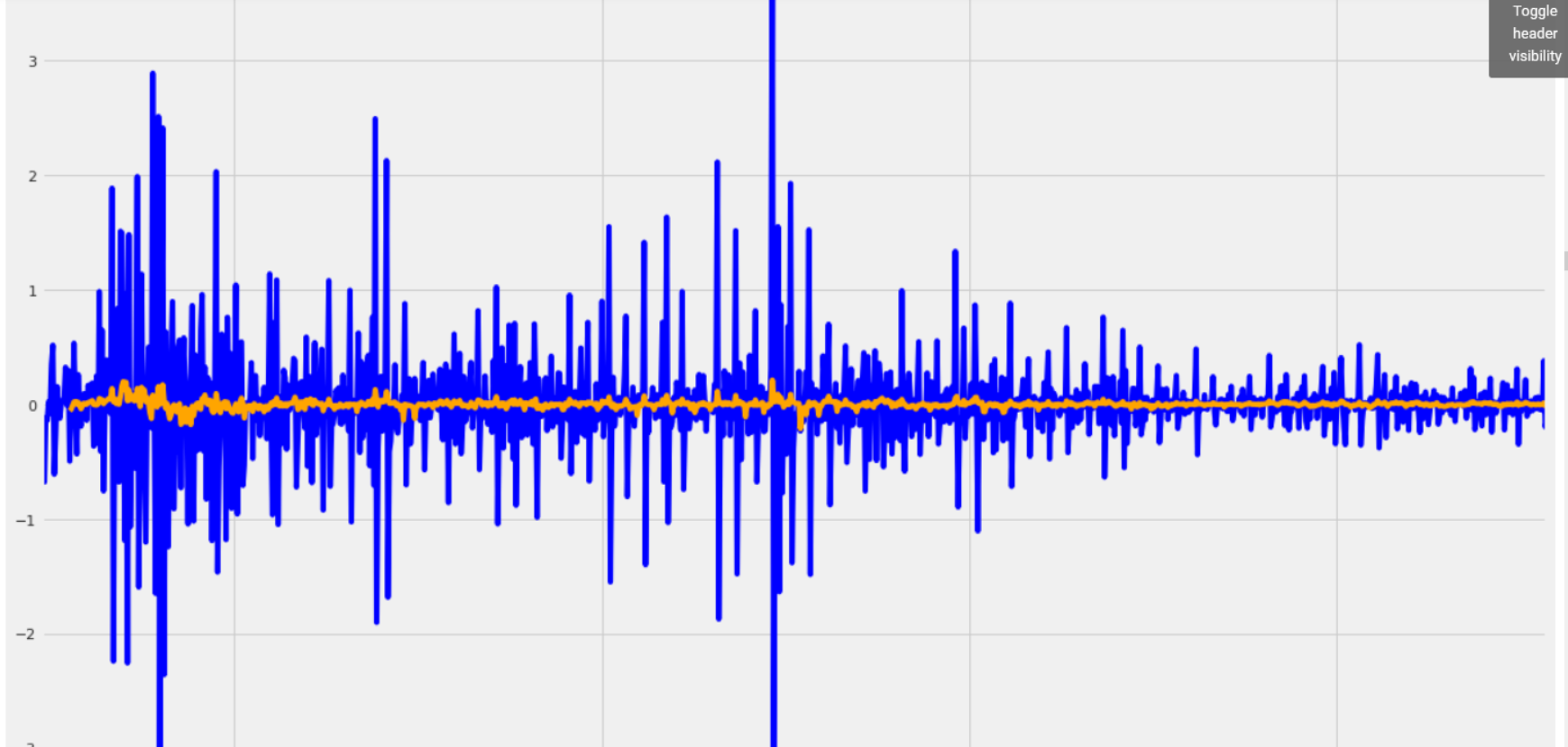


The plot of monthly volume per year spread over the entire dataset. The volumes of the stock were very high during initial period 1998 and 1999 but gradually decreases year over year.



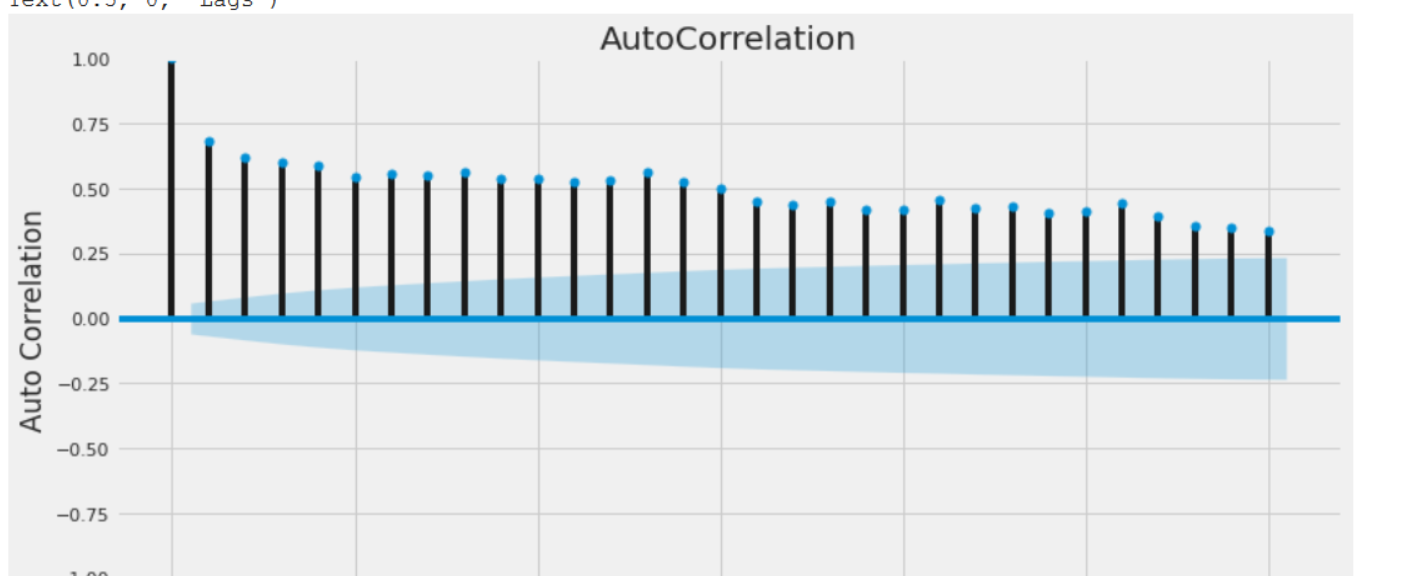
**Stationarity**

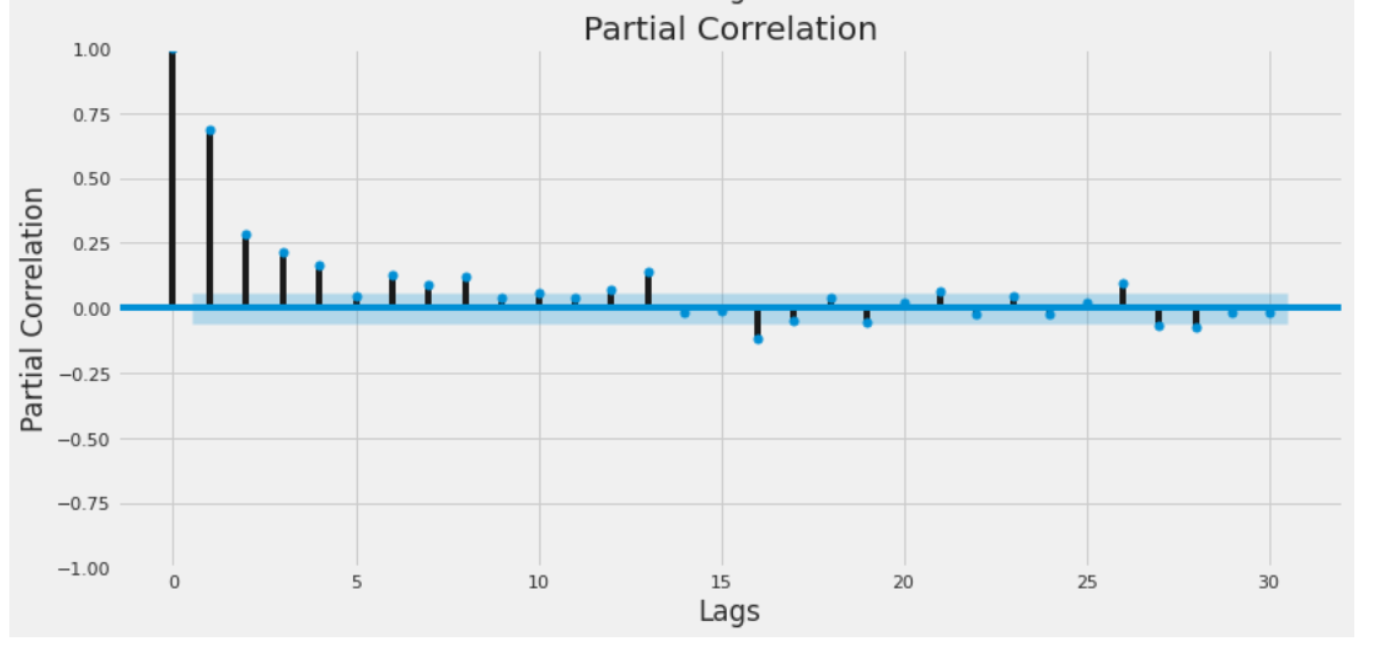
The Volume data is checked for stationarity using a adfuller() and also plotting a diff of the dataset. Though adfuller() has pointed that the series is not startionary ( p-value not < 0.05), the plot of the diff data shows that series is centered around zero and hence it is staionary

****

**ACF and PACF plotting**

The plot of ACF and PACF shows decaying bars over the lags. ACF has many bars over the significant area and the p, q values are not clear from the plot. These values can also be computed using auto\_arima().

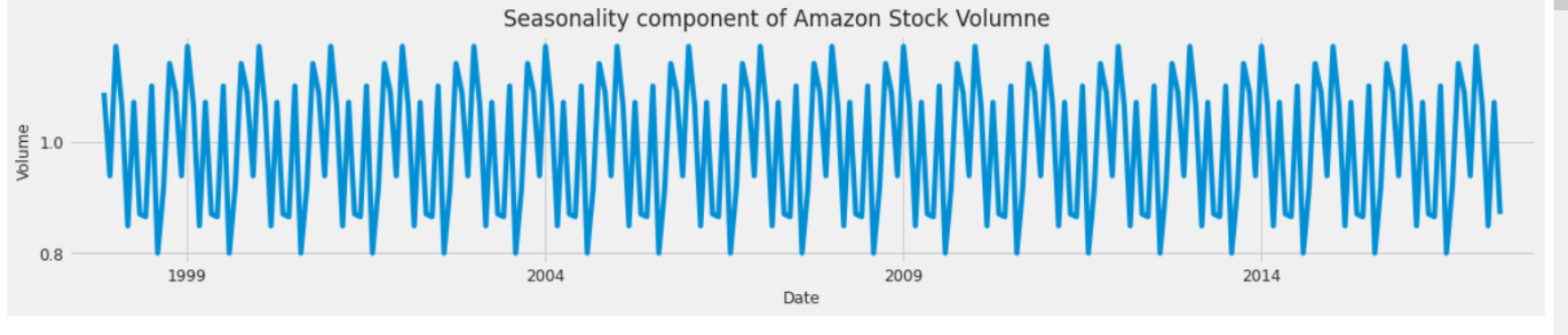
****

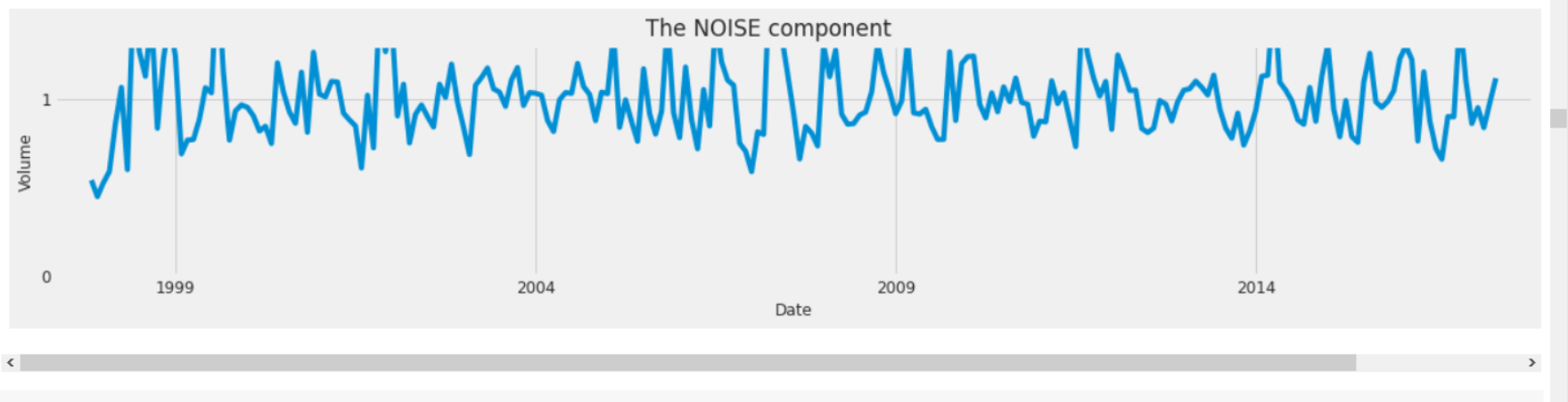
****

**Decomposition of Volume**

The volume data is decomposed manually into trend, seasonality and residue or noise. The manual computed trend, seasonality and residue are plotted as follows.



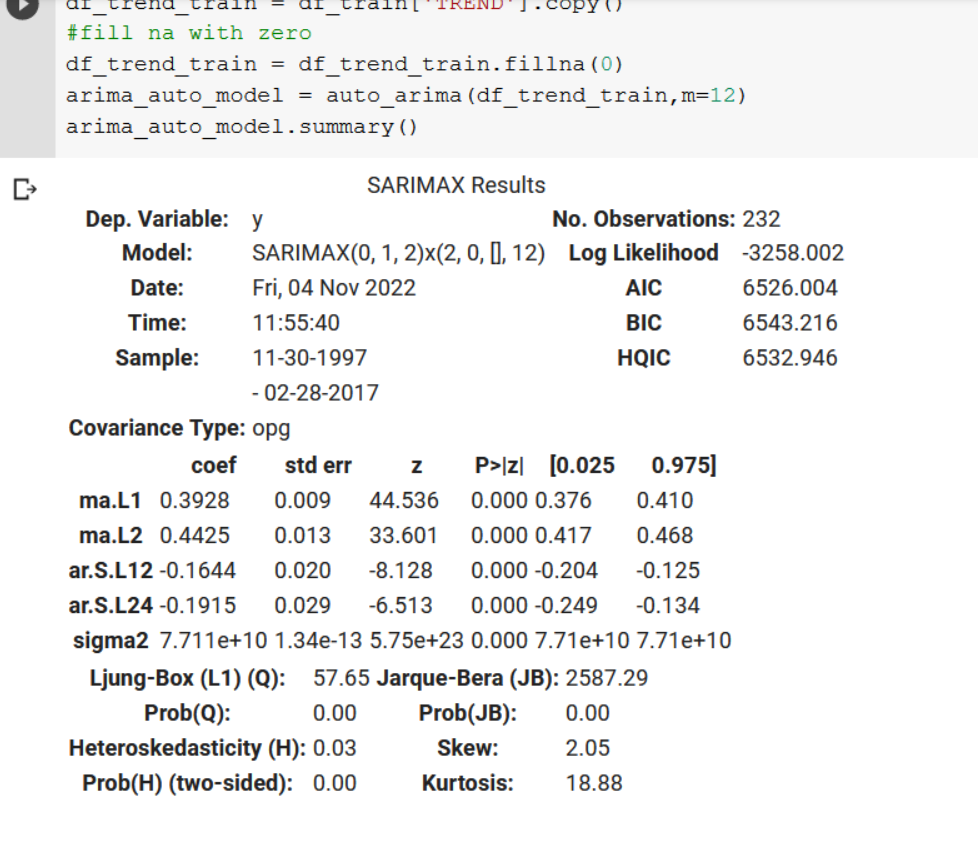




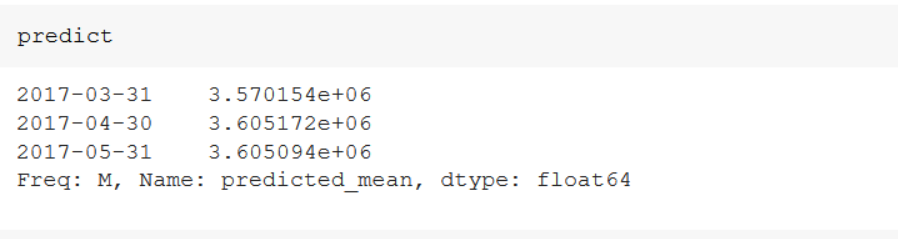
The computed decomposed data has rows with null (NaN) at the top and bottom of the dataset. These rows with Null values are dropped from the dataset as these data can skew the forecasting results

**ARIMA Modelling**

The trend component of the volume data is used in the current analysis. Using Auto ARIMA, the trend time series is analysed to get best ARIMA model with seasonality = 12 as the dataset is monthly and assuming there is a yearly seasonality of volume trend. Auto Arima has suggested to use SARIMAX as the best suitable model shown below



The SARIMAX model is trained with train and test data. The last 3 months of data is used for testing and the rest of the rows are used for training. The model fitted with training data is used to predict the test data. The Sarimax model have predicted the trend component of the volume as follows:



The combined predicted values is computed by adding predicted trend, seasonality of test and residue of test data. The combined data represent the predicted volume using Arima model. The model performance is measured using the predicted and actual volume. The model performance is a measured for the errors for the monthly mean of volume data.

RMSE: 608194.269583

R2: 454188.950589

MAPE: 454188.950589

MAE: 454188.950589

**Feed Forward Neural Network**

The trend component of the Volume data is used to create a Feed Forward Neural Network by using LSTM algorithm. The input data is just a array of trend values. To add more features to the input data, each row of the input data is appended with past 12 months of data as input. This is expected to help to create a better prediction model rather than a model dependent on single feature

Past values of the data are added as follows:

for i in range (12 , len(df\_trend\_train)) :

  X.append(df\_trend\_train[i-12:i])

  Y.append(df\_trend\_train[i])

The X array represents the set of features, Y represent the target value, i.e the trend of the Volume.

A feedforward neural network is created that accepts 1 input layer , 3 hidden layers and one output layer. The input layer accepts array size 12 inputs and outputs one value as the forecasted value

lstm\_model = Sequential()

lstm\_model.add(LSTM(60 , return\_sequences = True , input\_shape = (X\_train.shape[1] , 1)))

lstm\_model.add(LSTM(60 , return\_sequences = False))

lstm\_model.add(Dense(12))

lstm\_model.add(Dense(1))

The model is trained with adam optimizer, mean\_square\_error as loss function and compiled with batch size of 12 for 50 epochs. The compiled model is used to predict the test data that has same array size as input array. The prediction based on test data represents the trend of the volume. The predicted trend is combined with seasonality and residues of test data in the same manner as above.

Predicted Trend component for last 3 months:

array([[50343.44],

[50343.44],

[50343.44]], dtype=float32)

The combined predicted volume and actual volume are used to compute the error and performance of the LSTM model. The error output from the LSTM is as follows

RMSE: 3258276.480828

R2: 3215191.525128

MAPE: 3215191.525128

MAE: 3215191.525128

When compared to ARIMA model, the error in LSTM are far higher which suggests that LSTM has not performed well in the forecast. The ARIMA model has performed better with lower errors.