Project Report

on

Financial Transaction Volume Prediction

Group No: 01

Shinakshi Sankhyan
Viraj Majalkar
Harshad Punghera

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF CYBER SECURITY
SCHOOL OF TELECOMMUNICATION

Problem Statement:

Each public and private sector bank is obliged to report volume details of all financials transactions processed. One of such report is Domestic Payment made in form of NEFT, RTGS, Cards transaction. Future volume prediction can help banks to keep the infrastructure ready to process the transactions.

Overall summary of your solution

Dataset has 5 features namely Month-Year, NEFT Transactions, RTGS Transaction, Debit Card and Credit Card Transactions. As data set pertains to Time Series, 2 different approaches have been used.

Approach 1 : New feature “Total Transaction” is being added which is sum of all four Individual Feature i.e. NEFT , RTGS, Debit Card, Credit Card transactions. New feature “Total Transaction” has been used as a parameter with Month-Year for second approach. For first approach building 5 models have been used.

1. AR
2. MA
3. ARMA
4. ARIMA
5. SARIMA

Approach 2 : Individual Feature i.e. NEFT , RTGS, Debit Card, Credit Card transactions has been used as a parameters with Month-Year for first approach. For second approach building 2 models have been used.

1. AR
2. ARIMA

Detailed description and analysis of solution

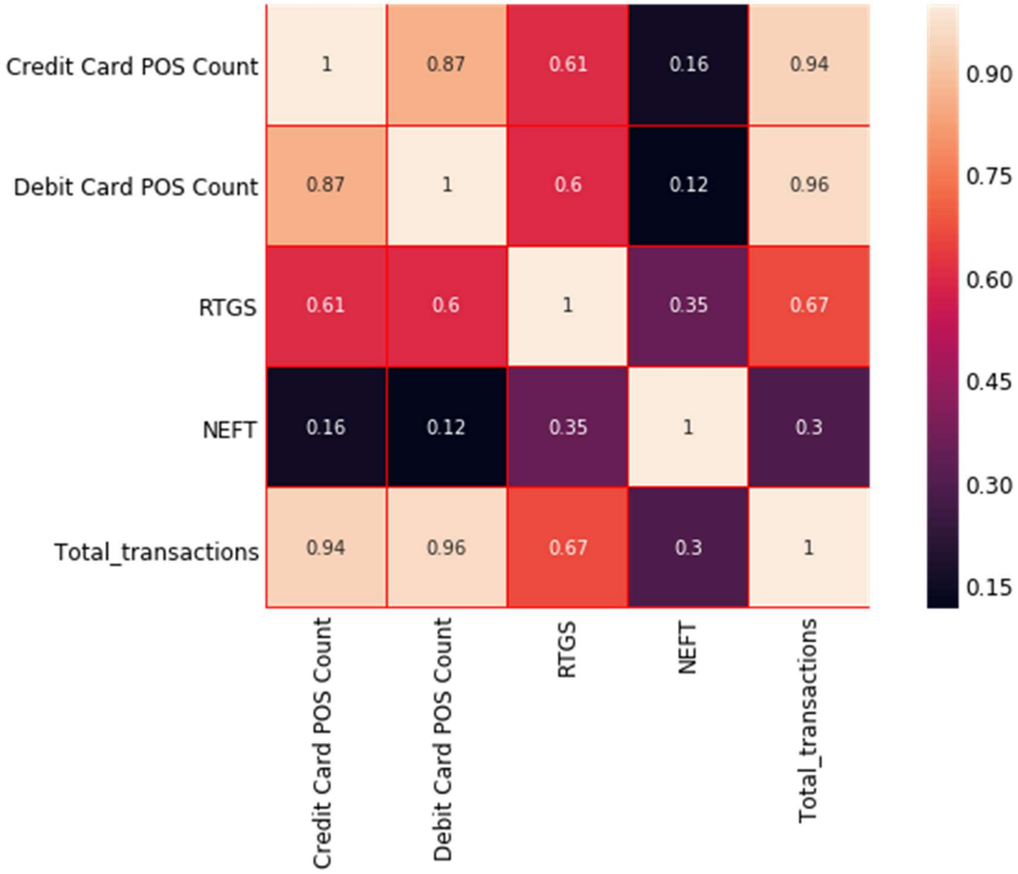
Dataset contains total 60 values starting from Jan 2015 to Dec 2019 and all are numerical hence no need to encode any of the values.

Dataset contains only null values which has been replaced with mean value of the rest of the data.

New feature “Total Transaction” is being added which is sum of all four Individual Feature i.e. NEFT, RTGS, Debit Card, Credit Card transactions.

Heatmap

Heatmap Graph has been used to understand correlation between the variables and it has been found that variable Debit Card POS Count and Credit Card POS Count is highly correlated to Total_transactions.

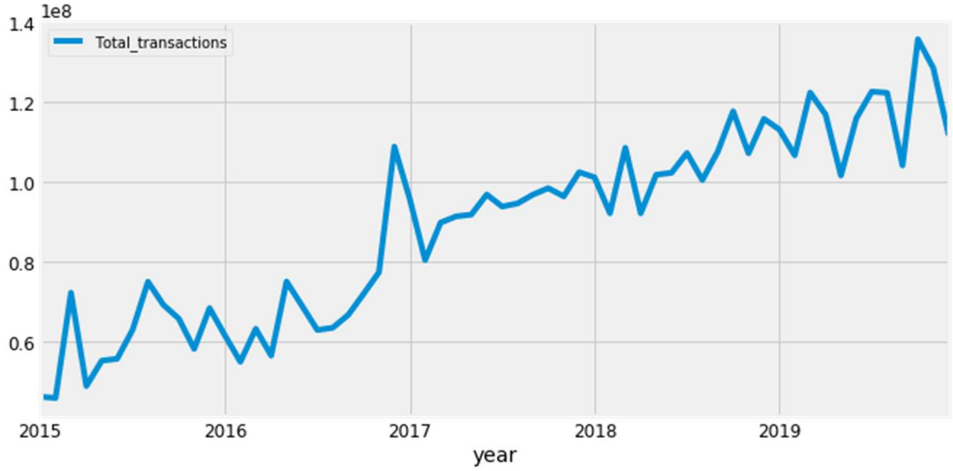


	Credit Card POS Count	Debit Card POS Count	RTGS	NEFT	Total_transactions
Credit Card POS Count	1	0.87	0.61	0.16	0.94
Debit Card POS Count	0.87	1	0.6	0.12	0.96
RTGS	0.61	0.6	1	0.35	0.67
NEFT	0.16	0.12	0.35	1	0.3
Total_transactions	0.94	0.96	0.67	0.3	1

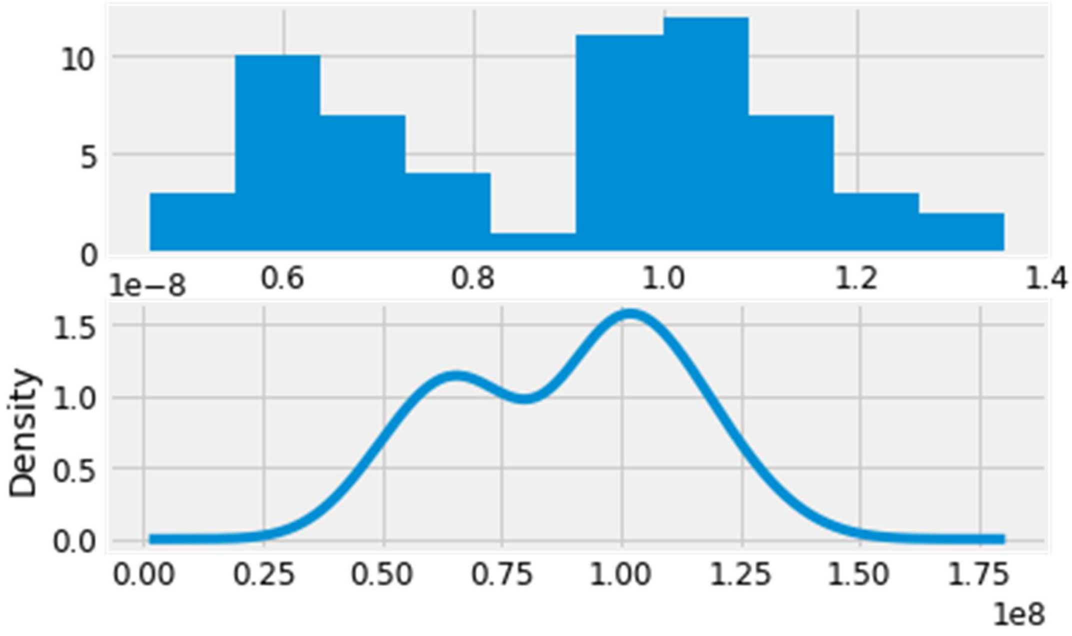
Approach 1

As “Total_transactions” is the only column will be used along with Month-Year rest of the columns have been dropped.

“Month-Year” column is set as index column and below graph is plotted where upward trend of the transaction is clearly seen.

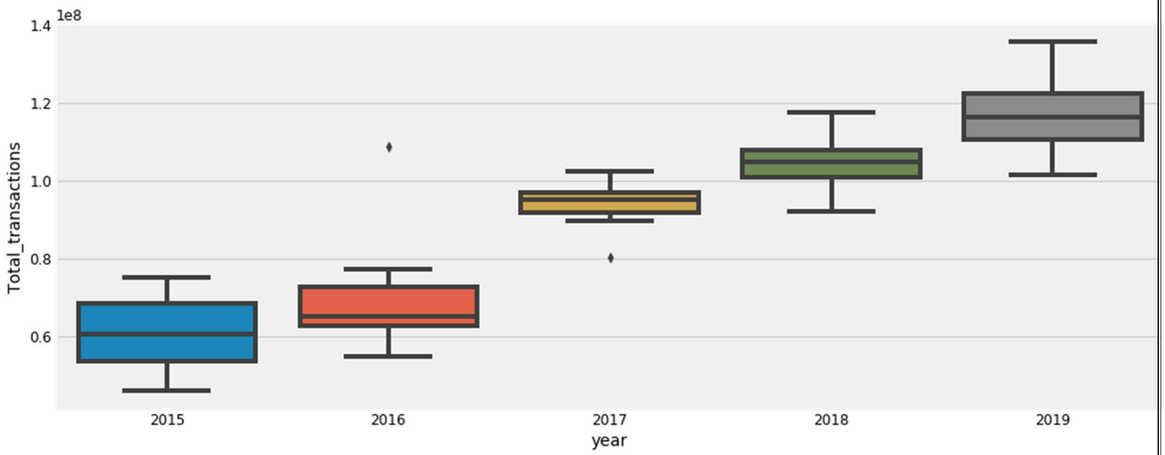


- Below Histograph clearly shows that data is not normally distributed and it is rightly skewed.

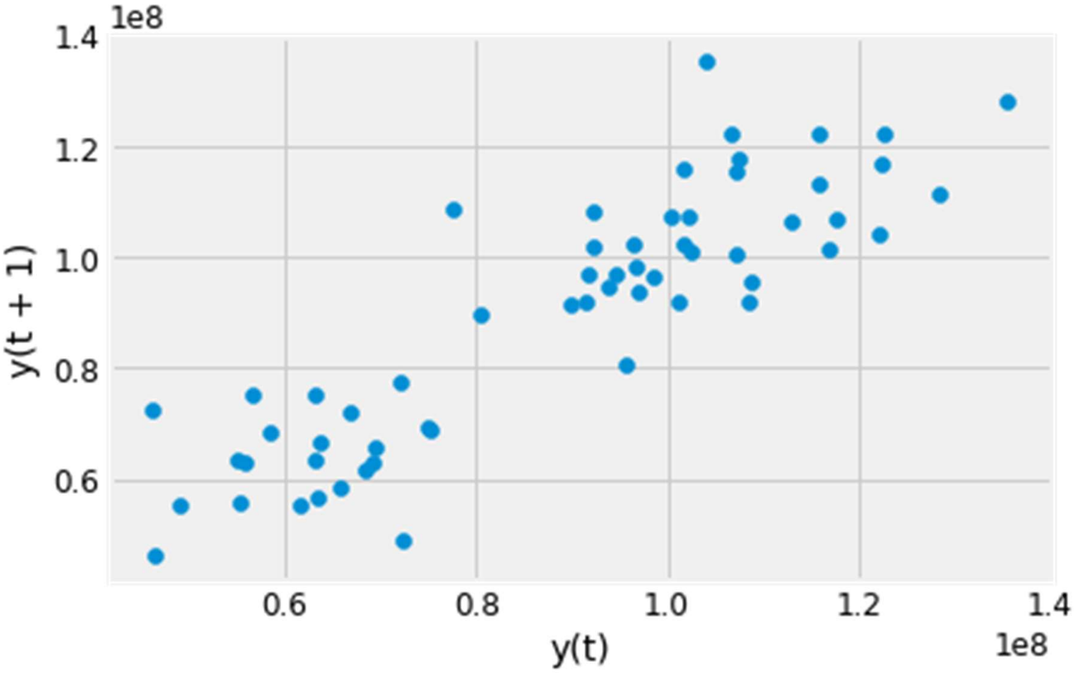


- Box plot for every year :

  Plot shows upward trend for median values.

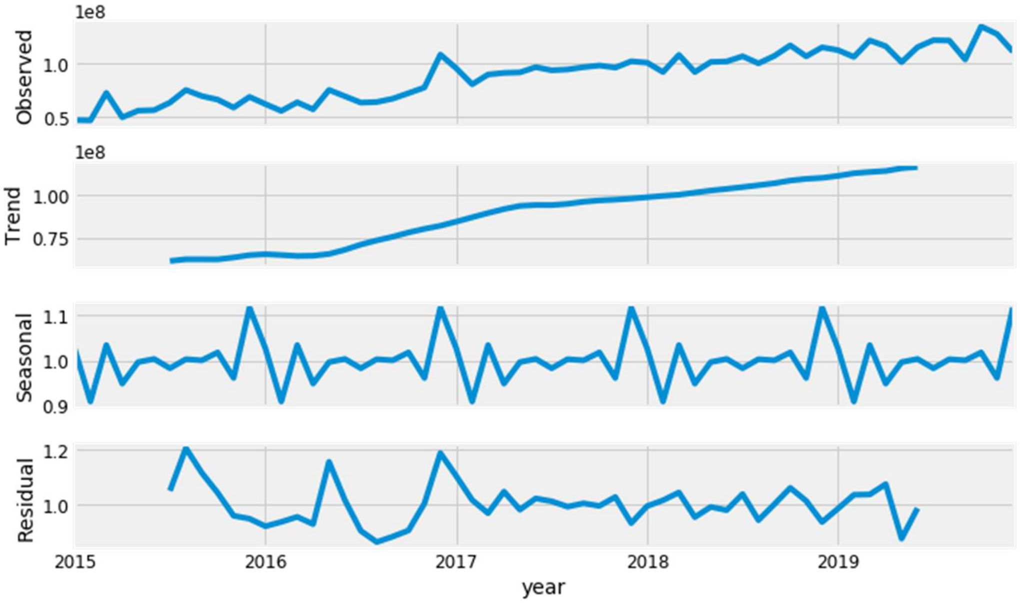


- Lag plot with 1 time lag shows positive correlation



- **Decomposing using statsmodel:**

  Statsmodel is used to perform decomposition as it deconstructs a time series interval component into each one of the underlying categories. Statsmodel help to understand normal trend, seasonal trend and residual components.



Observed /original data plot shows increasing pattern

Trend shows steadiness till mid 2016, post which it shows increasing trend with small downtrends in between.

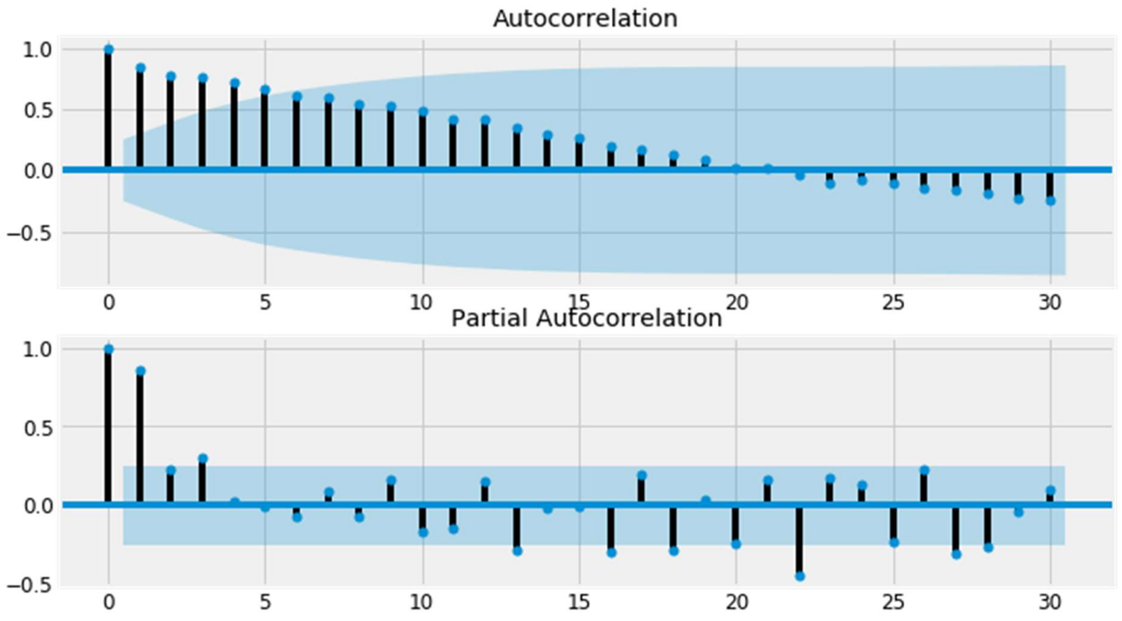
Seasonal plot clearly shows sudden spike in transactions in every year December month.

Residuals show more fluctuations till 2017, post which it shows steady trends.

Stationary means that statistical properties such as mean, variance remain constant over time.

Most of the Time Series models work on the assumption that Time series is stationary. Major reason for this is that there are many ways in which a series can be non-stationary, but only one way for stationarity.

If a Time Series has a particular behaviour over time, there is a very high probability that it will follow the same in the future. Theories suggest stationary series are more mature and easier to implement as compared to non-stationary series.

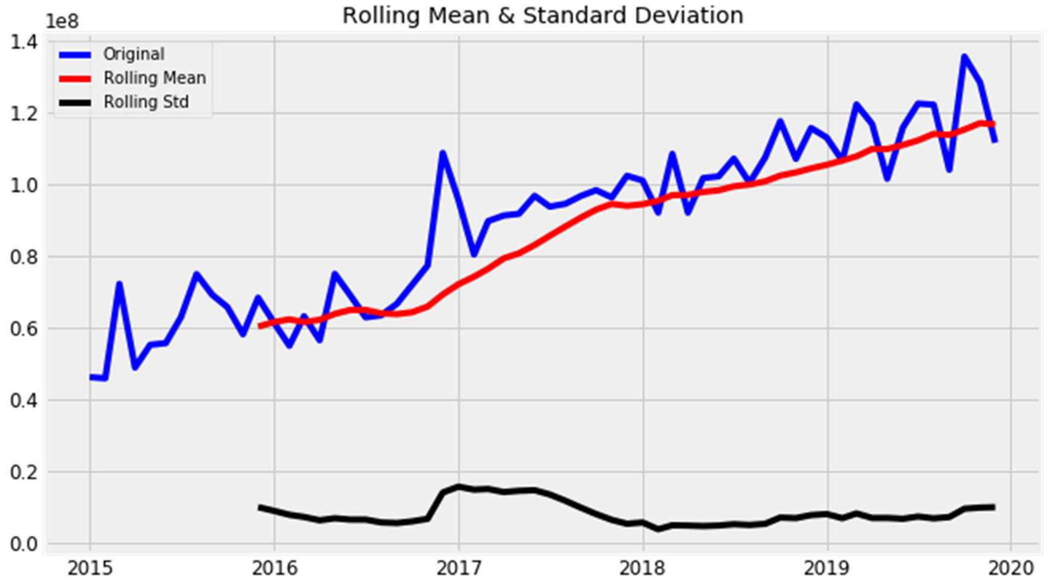


As correlation measures the extent of a linear relationship between two variables, autocorrelation measures the linear relationship between lagged values of a time series.

When data has a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size.
So the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.

When data are seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal frequency) than for other lags.

When data are both trended and seasonal, you see a combination of these effects



We Observe That the Rolling Mean and Standard Deviation are not constant w.r.t Time (Increasing Trend) hence the time series is not stationary

Augmented Dickey-Fuller Test

The intuition behind the test is that if the series is integrated then the lagged level of the series df4(t-1) will provide no relevant information in predicting the change in df4(t).

Null hypothesis: The time series is not stationary. Rejecting the null hypothesis (i.e. a very low p-value) will indicate stationarity and as p value is greater than critical values so reject the H0.

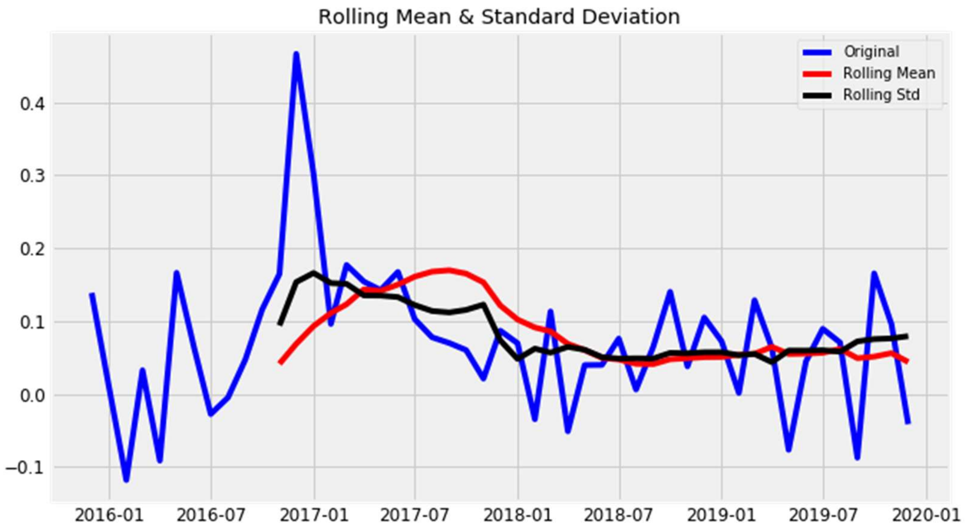
- MAKING TIME SERIES STATIONARY

2 major reasons for time series to be non-stationary.

Trend – varying mean over time. For eg, in this case we saw that on average, the number of passengers was growing over time.

Seasonality – variations at specific time-frames. eg people might have a tendency to buy cars in a particular month because of pay increment or festivals. Transformation which penalize higher values more than smaller values can be done by taking a log, square root, cube root, etc.

- **Log transformation has been taken** whose rolling mean and standard deviation shows following:



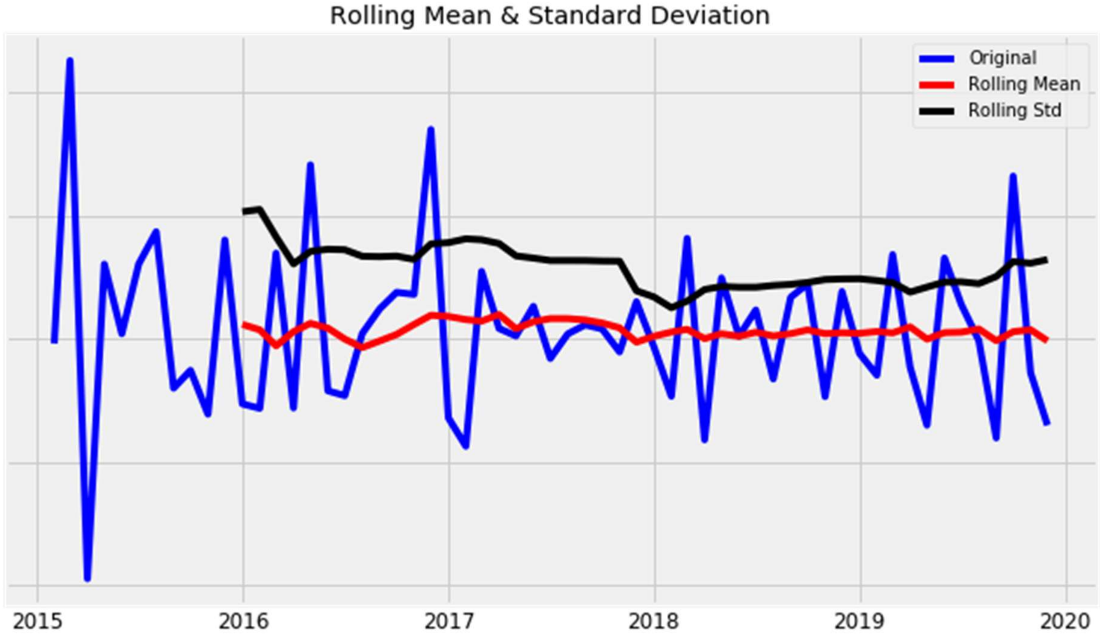
Results of Dickey-Fuller Test:

Test Statistic	:	-4.722569
p-value	:	0.000076

```
Lags Used                          :           0.000000
Number of Observations Used        :           48.000000
Critical Value (1%)                :           -3.574589
Critical Value (5%)                :           -2.923954
Critical Value (10%)               :           -2.600039
dtype: float64
```

As test statistic < 1% critical value. i.e. data is stationary 99%

- other technique to remove seasonality and trend:
  Differencing:
  In this technique, we take the difference of the observation at a particular instant with that at the previous instant.
  First order differencing in Pandas



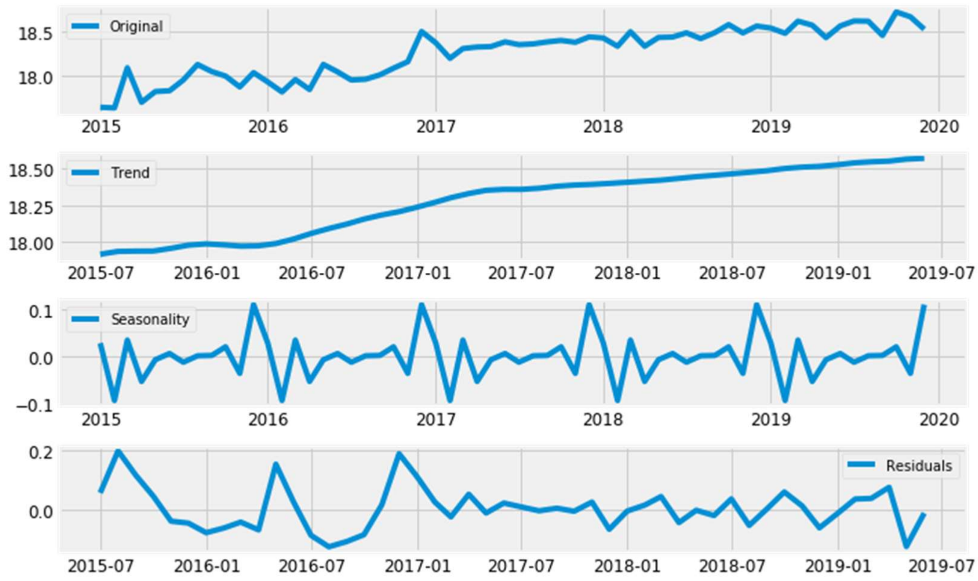
Results of Dickey-Fuller Test:

```
Test Statistic                     :           -9.515190e+00
p-value                            :           3.171034e-16
Lags Used                          :           1.000000e+00
Number of Observations Used        :           5.700000e+01
Critical Value (1%)                :           -3.550670e+00
Critical Value (5%)                :           -2.913766e+00
Critical Value (10%)               :           -2.594624e+00
```

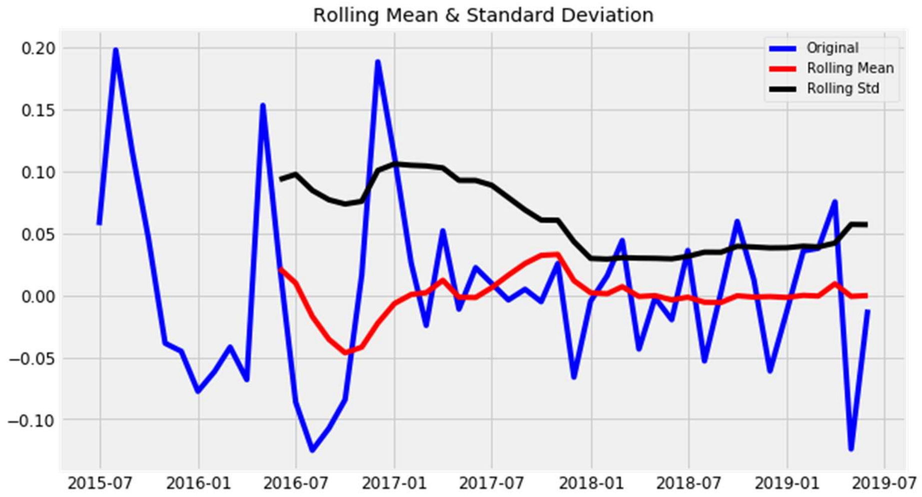
Here also test statistic < 1% critical value thus data is 99% stationary

Decomposition

In this approach, both trend and seasonality are modelled separately and the remaining part of the series is returned.



Rolling Mean and Rolling Standard Deviation:



<u>Results of Dickey-Fuller Test:</u>

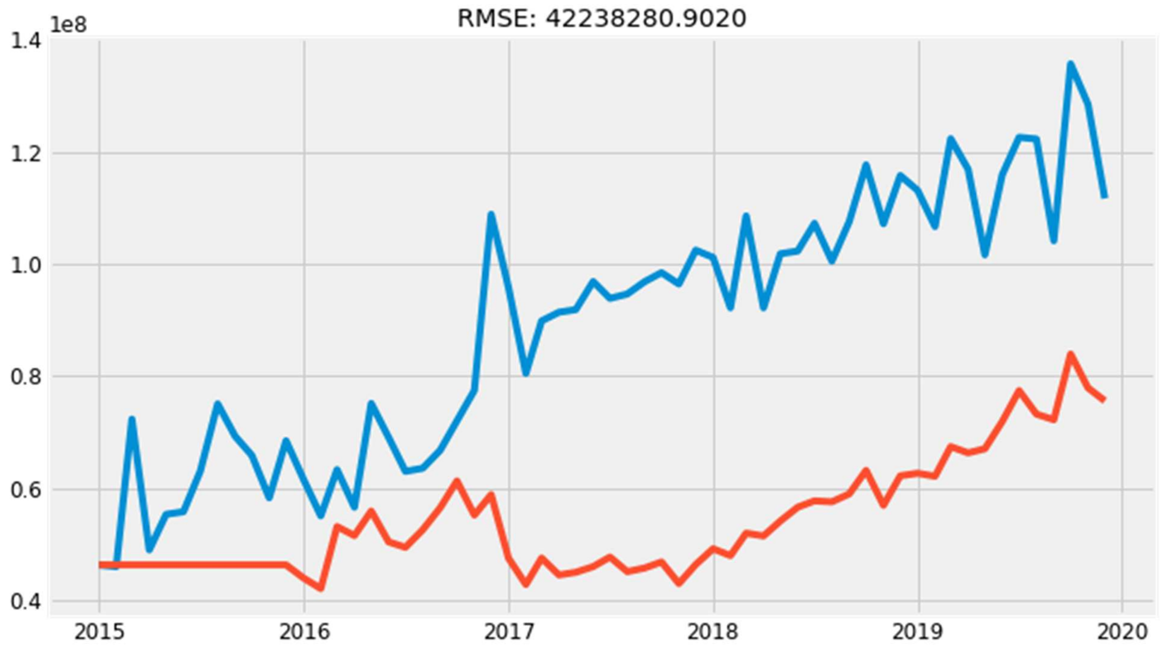
Test Statistic	:	-5.322793
p-value	:	0.000005
Lags Used	:	2.000000
Number of Observations Used	:	45.000000
Critical Value (1%)	:	-3.584829
Critical Value (5%)	:	-2.928299
Critical Value (10%)	:	-2.602344

As test statistics < critical values so its stationary

AUTOREGRESSION (AR)

The autoregression (AR) method models the next step in the sequence as a linear function of the observations at prior time steps.

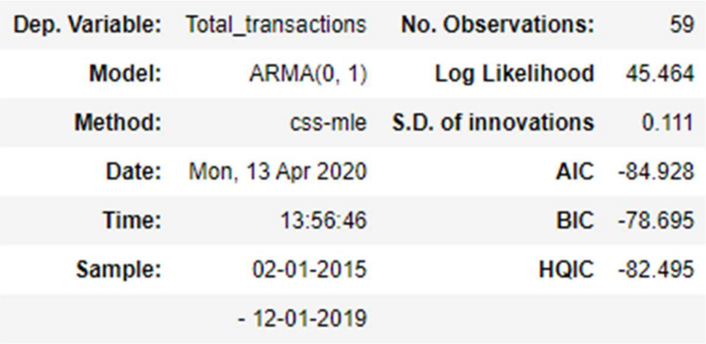
Number of AR (Auto-Regressive) terms (p): p is the parameter associated with the auto-regressive aspect of the model, which incorporates past values i.e. lags of dependent variable. For instance if p is 5, the predictors for x(t) will be x(t-1)….x(t-5).



MAE: 34671616.65035321

MOVING AVERAGE (MA)

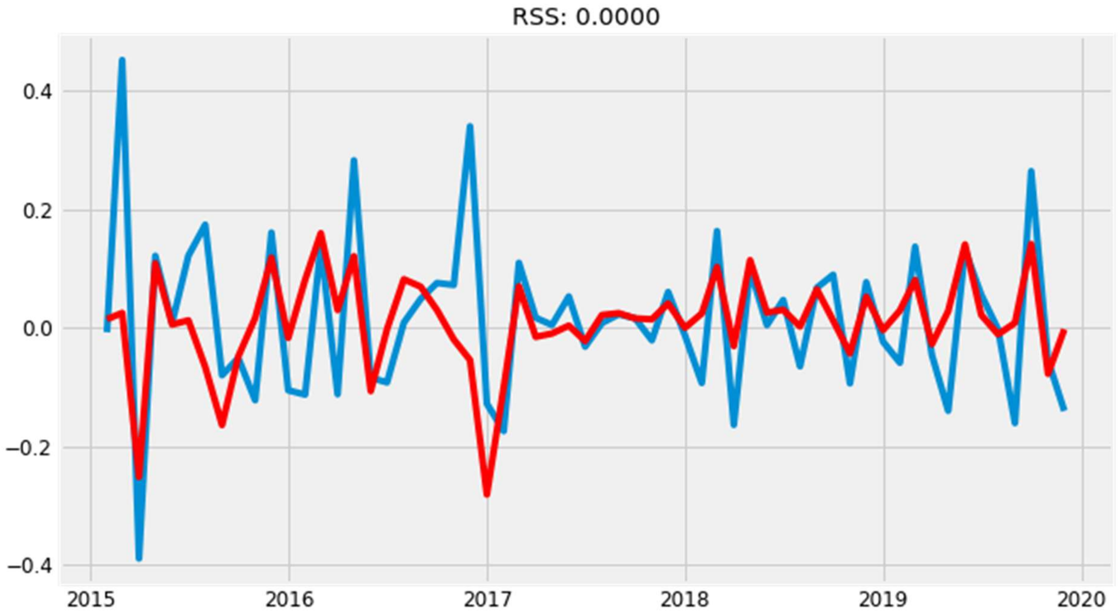
Number of MA (Moving Average) terms (q): q is size of the moving average part window of the model i.e. lagged forecast errors in prediction equation. For instance if q is 5, the predictors for x(t) will be e(t-1)….e(t-5) where e(i) is the difference between the moving average at ith instant and actual value.



Dep. Variable:	Total_transactions	**No. Observations:**	59
Model:	ARMA(0, 1)	**Log Likelihood**	45.464
Method:	css-mle	**S.D. of innovations**	0.111
Date:	Mon, 13 Apr 2020	**AIC**	-84.928
Time:	13:56:46	**BIC**	-78.695
Sample:	02-01-2015	**HQIC**	-82.495
	- 12-01-2019		

MAE : 116749442.63950129

AIC and BIC values are negative.

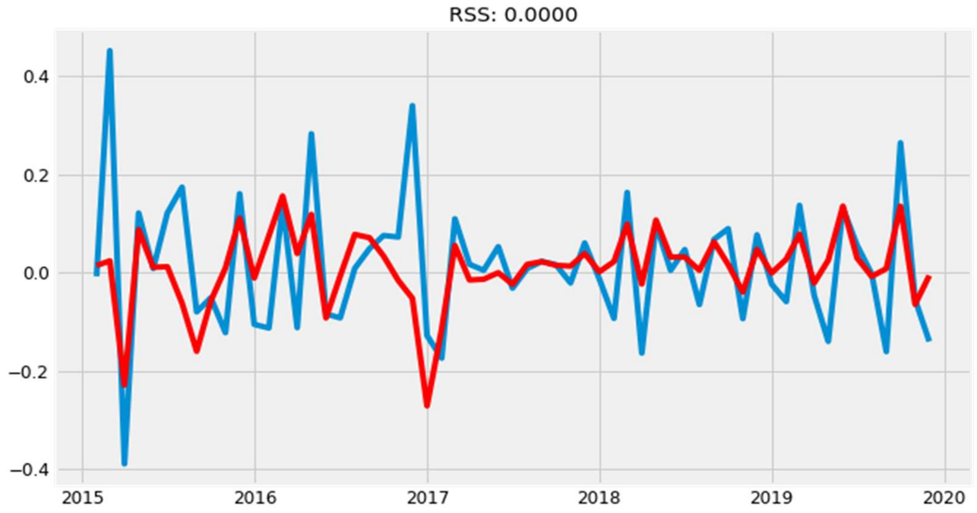


AUTOREGRESSIVE MOVING AVERAGE (ARMA)

Number of MA (Moving Average) terms (q): q is size of the moving average part window of the model i.e. lagged forecast errors in prediction equation. For instance if q is 5, the predictors for x(t) will be e(t-1)….e(t-5) where e(i) is the difference between the moving average at ith instant and actual value.

MAE : 116749442.63950129

AIC and BIC values are negative.



AIC and BIC values are greater than MOVING AVERAGES MODEL

P values does not satisfy for any of the time lag except for no time lag

When p parameter =0 then test statistic value satisfies and this model becomes equal to MOVING AVERAGES MODEL

MAE: 116749442.6384

Here we are taking p and q parameters into the model. But what we observe is that if p Parameter value is equals to 0 then it becomes moving average model and if we take an y time lag lets say p=1 or 2 or higher order then p value i.e. test statistic value doesn’t satisfy the criteria that it should be less than level of significance.
Above graph is for p=0
Thus we can conclude that ARMA becomes MA when we put p=0

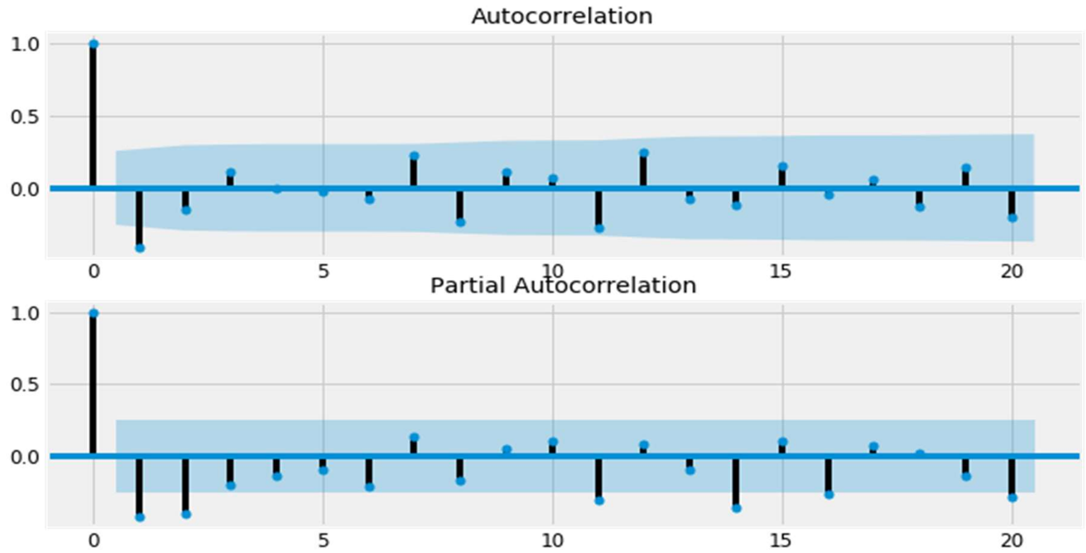
Autoregressive Integrated Moving Average (ARIMA)

In an ARIMA model 3 parameters are used to help model the major aspects of a times se ries: seasonality, trend, and noise. These parameters are labeled p, d, and q

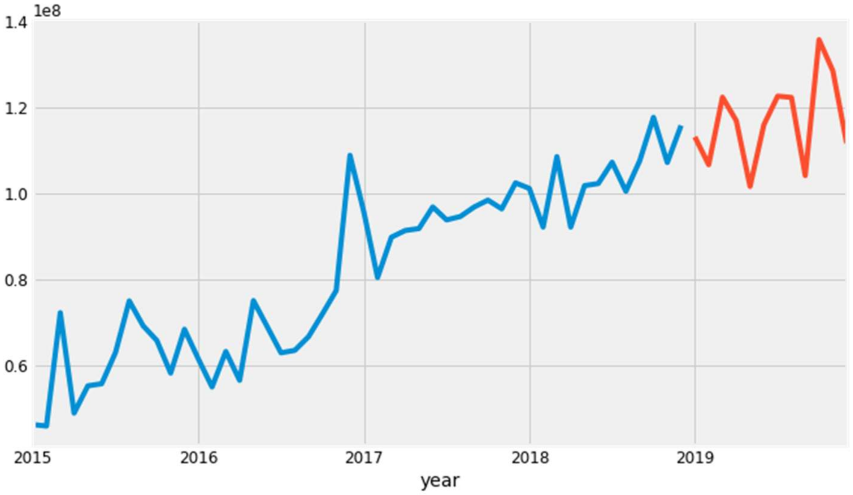
Number of AR (Auto-Regressive) terms (p): p is the parameter associated with the auto-regressive aspect of the model, which incorporates past values i.e. lags of dependent var iable. For instance if p is 5, the predictors for x(t) will be x(t-1)….x(t-5).

Number of Differences (d): d is the parameter associated with the integrated part of the model, which effects the amount of differencing to apply to a time series.

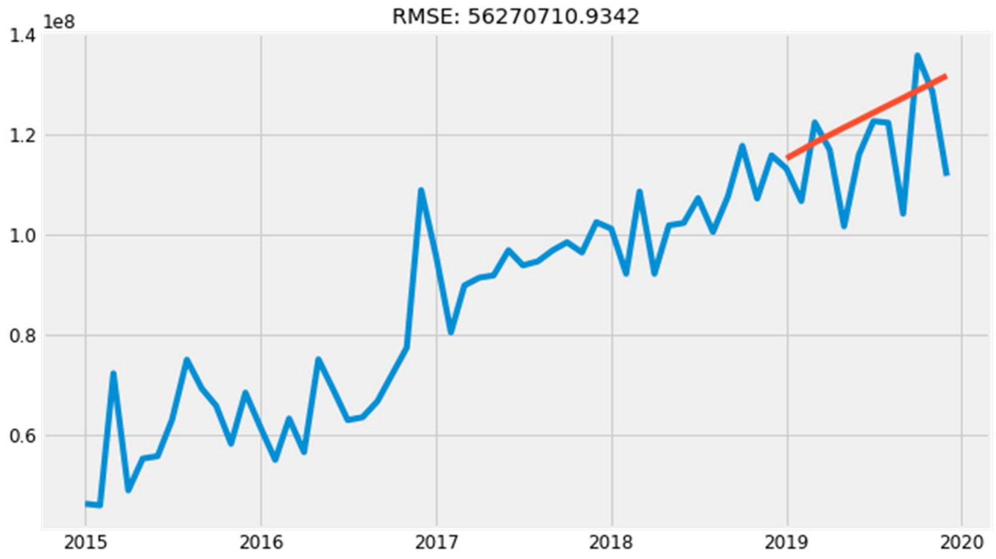
Number of MA (Moving Average) terms (q): q is size of the moving average part window of the model i.e. lagged forecast errors in prediction equation. For instance if q is 5, the predictors for x(t) will be e(t-1)….e(t-5) where e(i) is the difference between the moving average at ith instant and actual value.



As acf plot have 2nd lag negative i.e. data is stationary now splitting data into train test



Red line shows test part and blue one train part
Prediction:



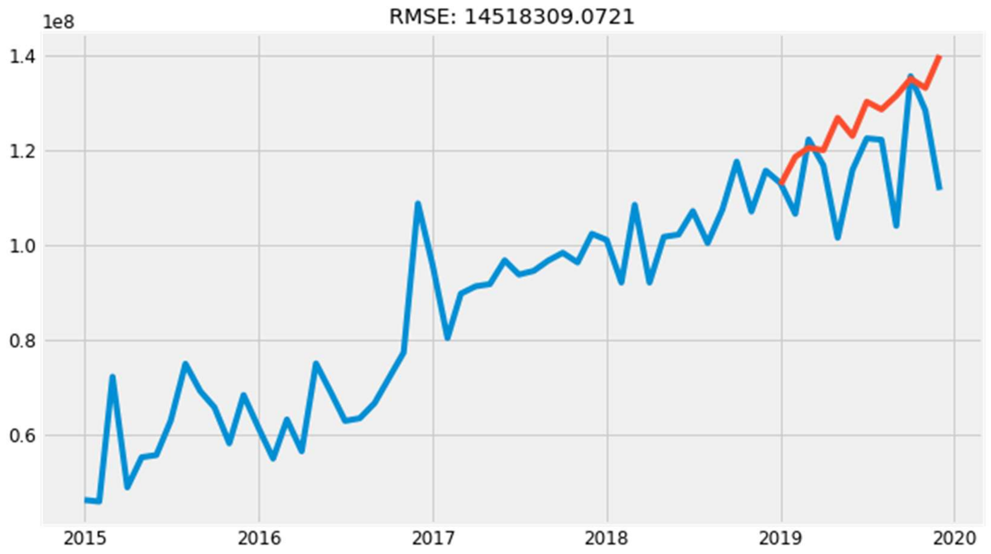
MAE : 8545596.045

As it is clearly seen from the graph that for test values, predicting values which are in red colour, are quite good. Moreover MAE value is way far less than other models. So this is good prediction.

SARIMA

multiplicative seasonal ARIMA model

A seasonal autoregressive integrated moving average (SARIMA) model is one step different from an ARIMA model based on the concept of seasonal trends

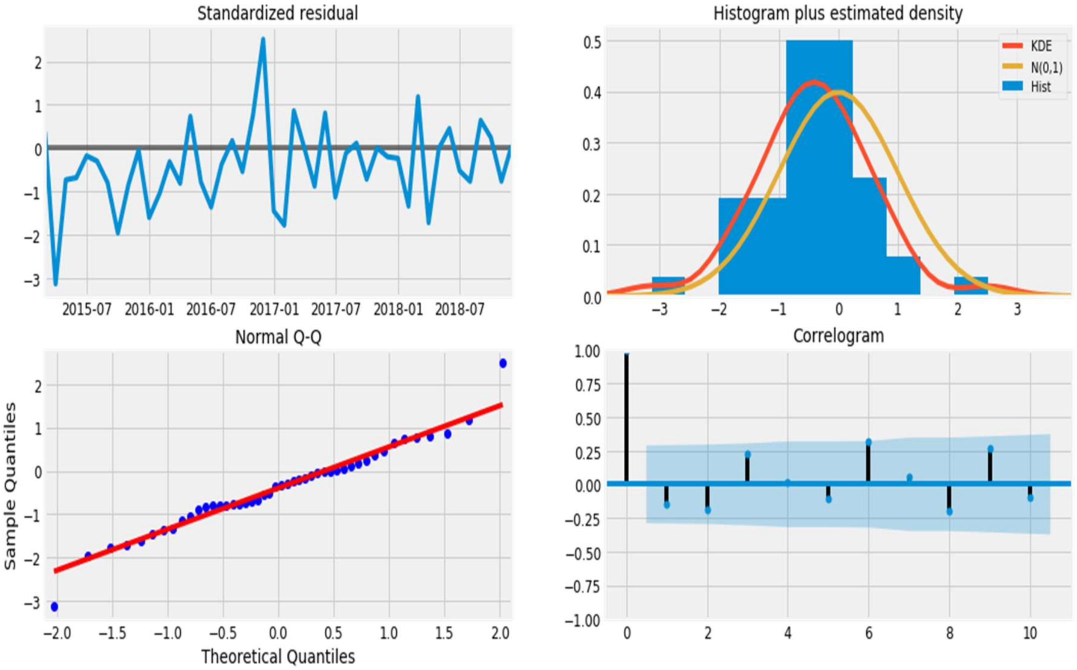


Prediction is not as per expectation as seen from the graph

MAE : 10412364.56

This MAE value is far far greater than that of ARIMA model.

MODEL_DIAGNOSTICS OF SARIMA



Our primary concern is to ensure that the residuals of our model are uncorrelated and normally distributed with zero-mean.

If the seasonal ARIMA model does not satisfy properties, it is a good indication that it can be further improved.

The model diagnostic suggests that the model residual is normally distributed based on the following:

1. In the top right plot, the red KDE line follows closely with the N(0,1) line. Where, N(0,1) is the standard notation for a normal distribution with mean 0 and standard deviation of 1. This is a good indication that the residuals are normally distributed.

2. The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution. Again, this is a strong indication that the residuals are normally distributed.

3. The residuals over time (top left plot) display seasonality

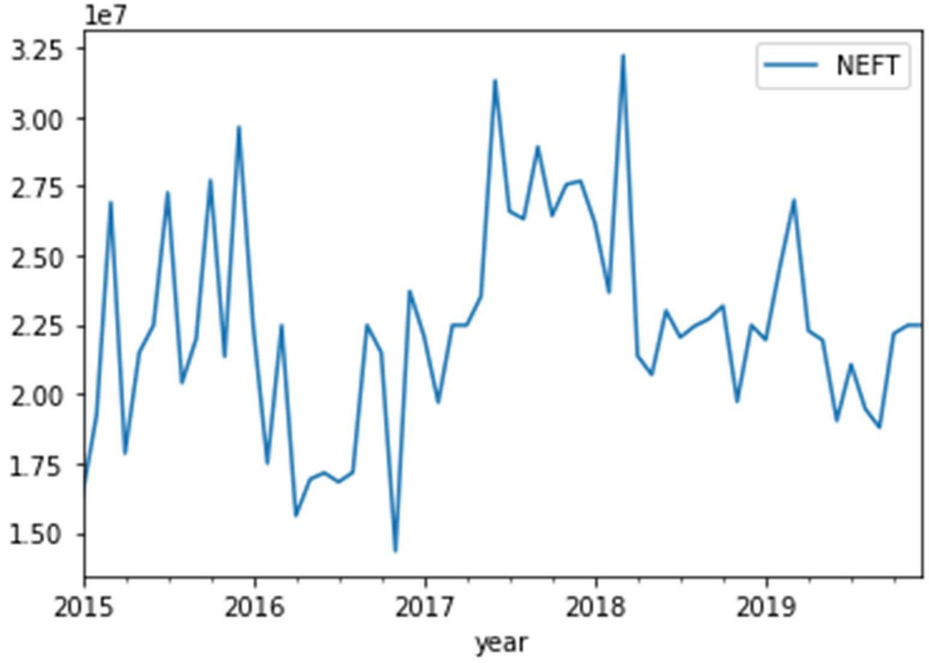
This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

APPROACH2:

Here we have fitted the models on each of the parameter separately. Whatever modelling for AR and ARIMA models have done above , same has been done here.

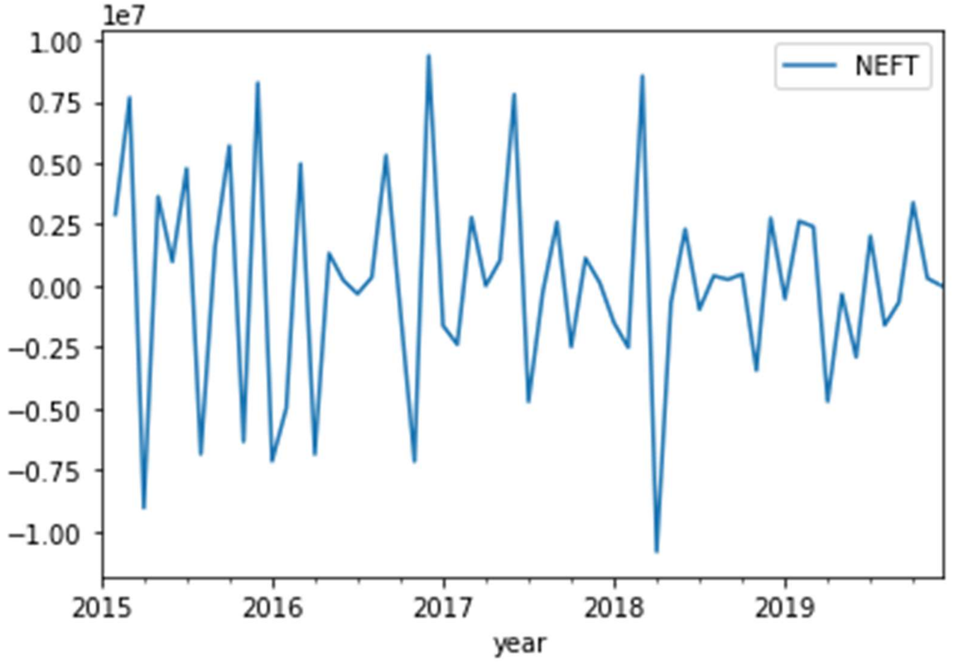
What’s different is that we have taken range in between 0 to 5 and calculated all the combinations for (p,d,q) parameter and corresponding aic values. And whichever was minimum aic value in that, corresponding combination has taken and applied in ARIMA model for all the parameters.

For NEFT :

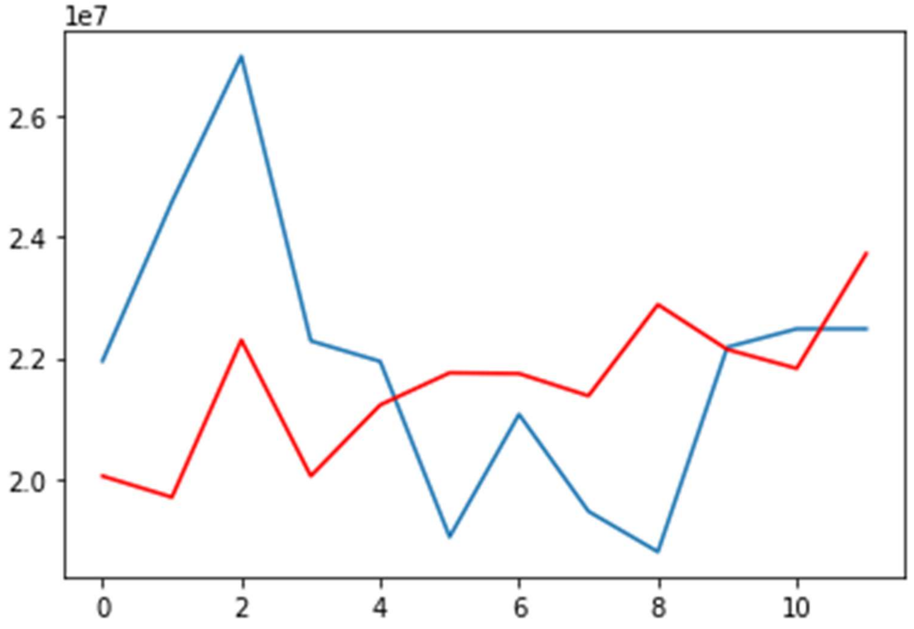


It’s clearly seen that data is not stationary

We have made the data stationary by taking the difference series



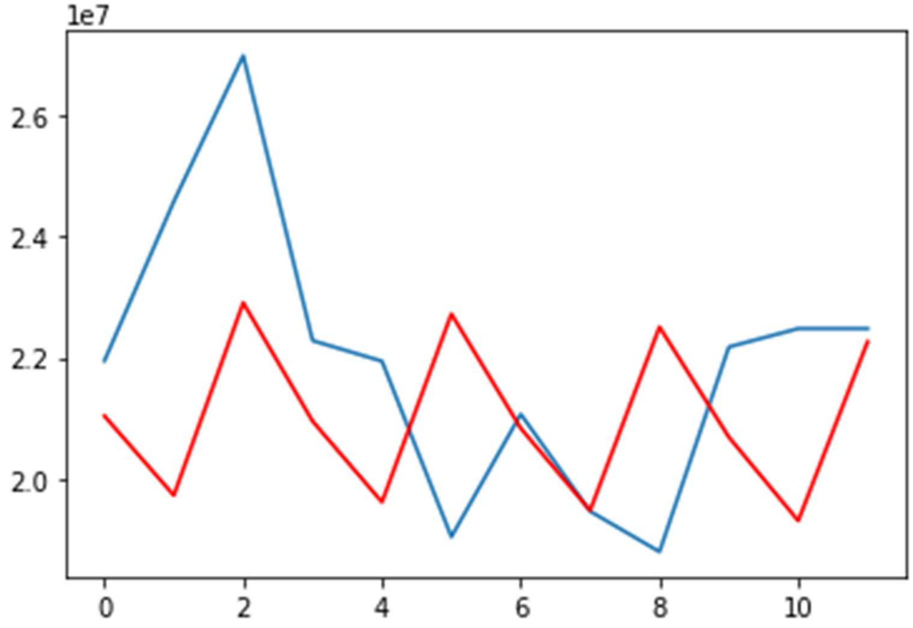
AR MODEL:



As we can see that predicting values are not as per the testing values

And MAE score for this model is :2147112.4291

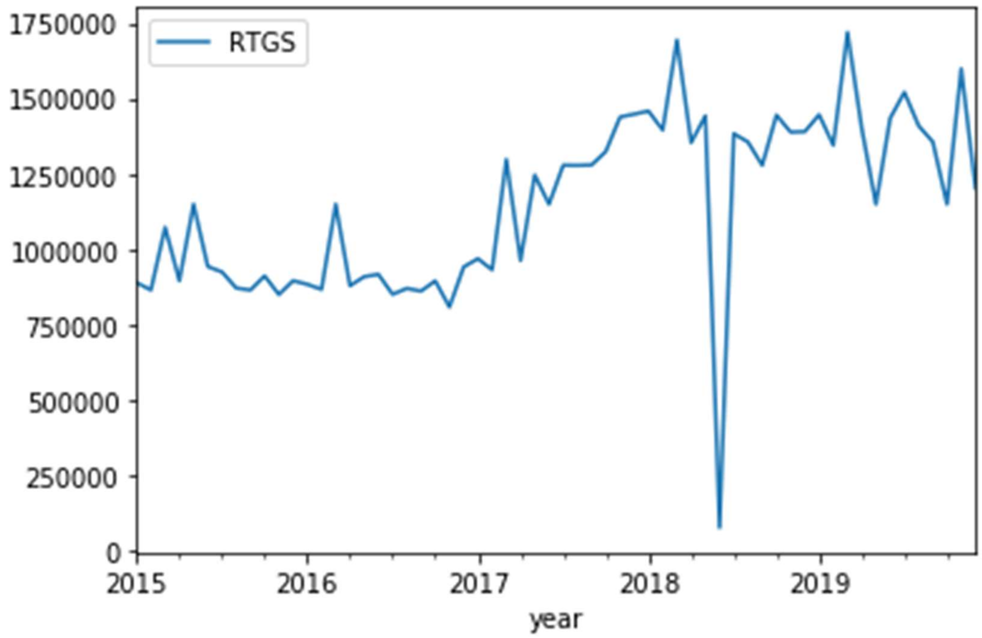
ARIMA MODEL:



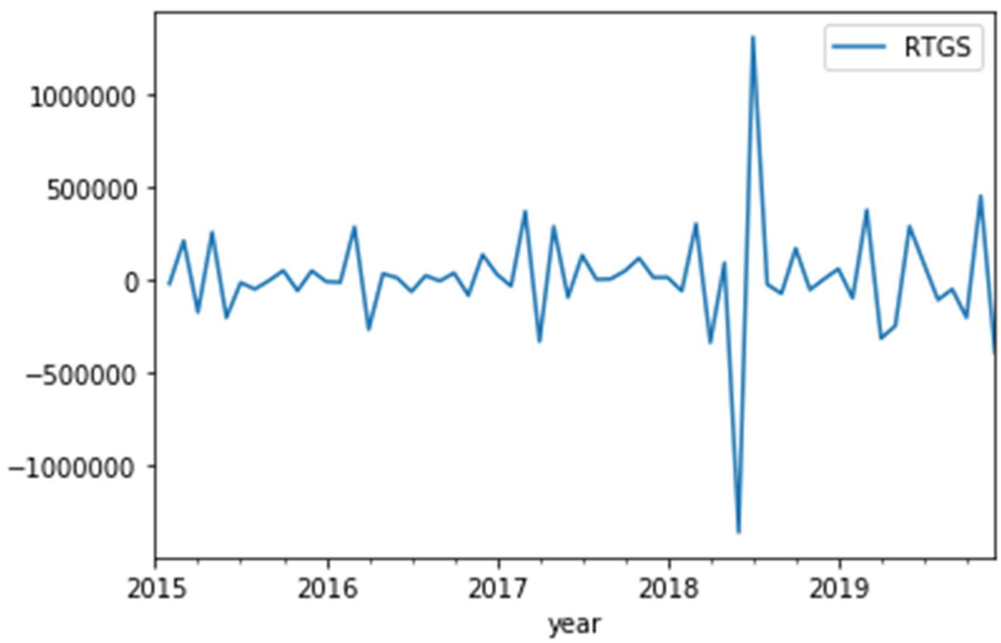
It can be seen that few values are close to actual values and some are not. This is model corresponding to aic value 1534.6851 which is corresponding to (2,2,4) combination.

MAE score: 2170000.6145

For RTGS:

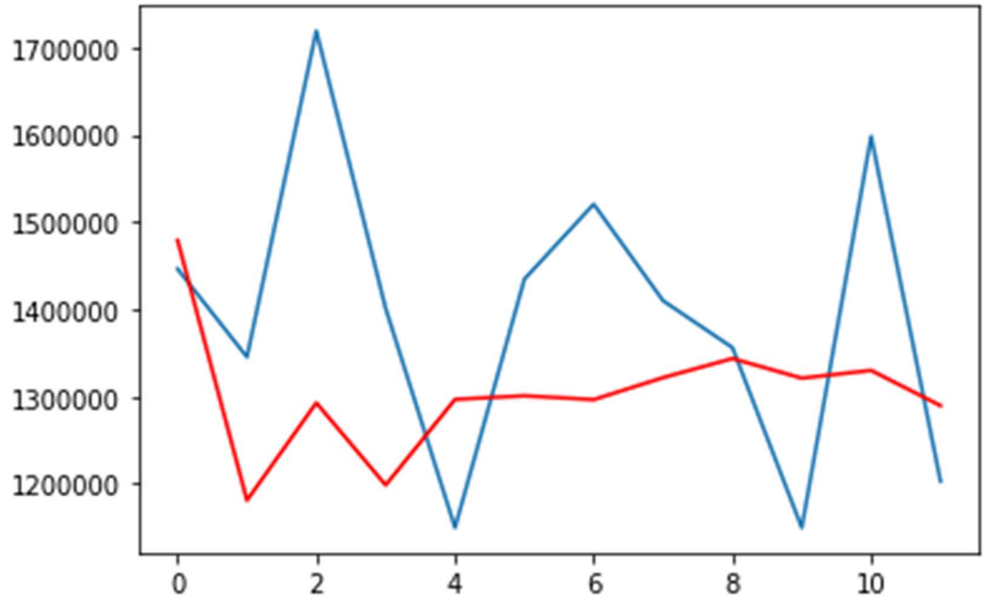


Data shows that it is not stationary.



Its stationary now.

AR MODEL :



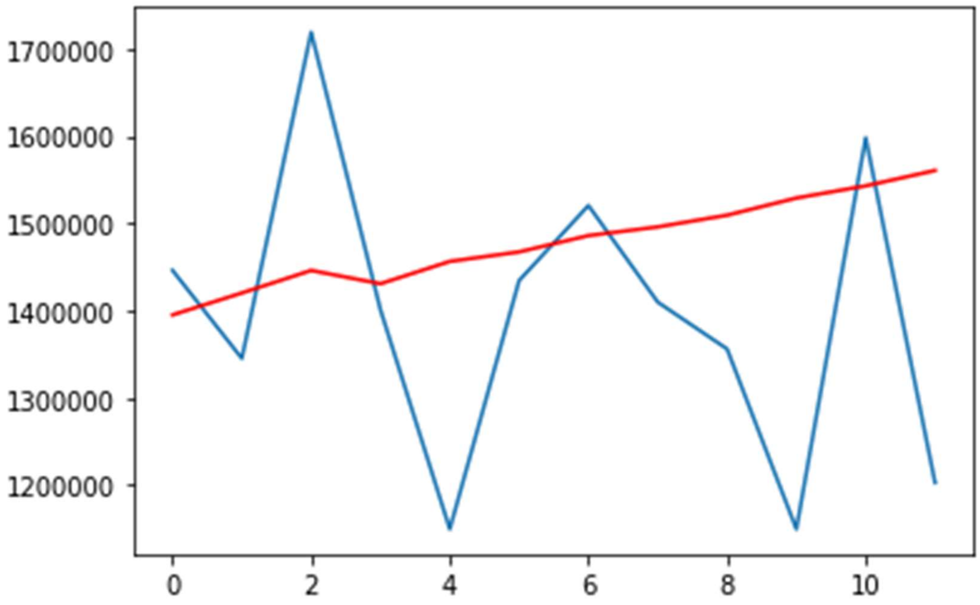
Here also predicting values are not that accurate to test values.

MAE VALUE: 163241.7568

```
2147112.429110894
```

In [26]:

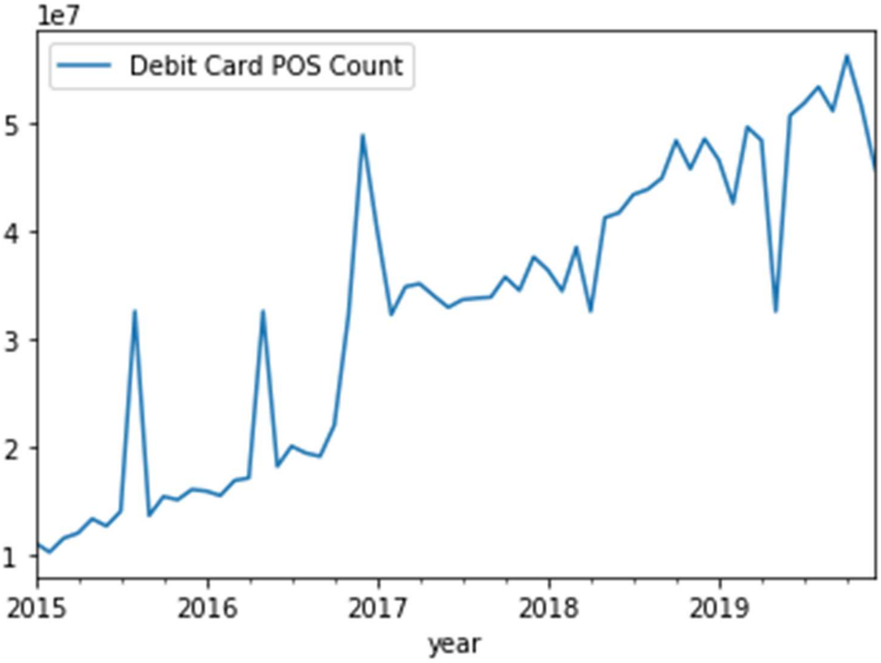
ARIMA MODEL:



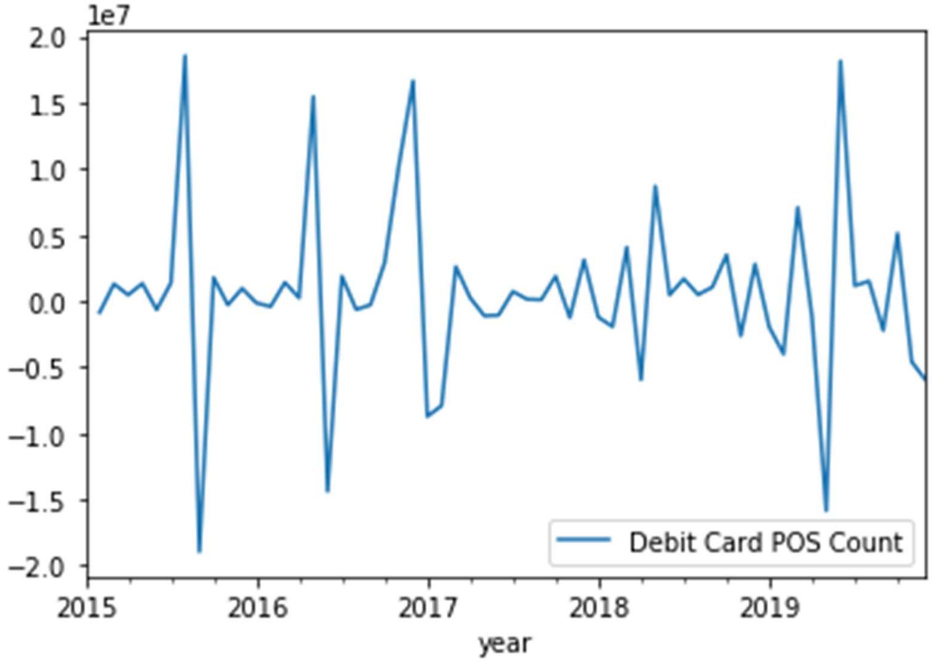
This is the prediction corresponding to aic value 1302.68 corresponding to (4,2,0) combination.

MAE VALUE: 152995.14

For debit card:

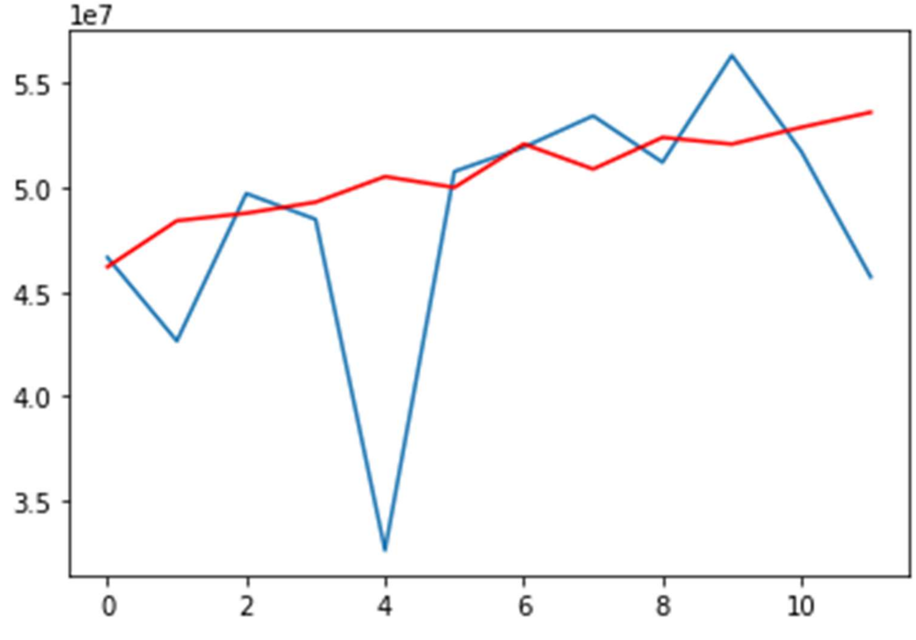


Data is not stationary as it is not constant.



By taking difference series it’s made stationary.

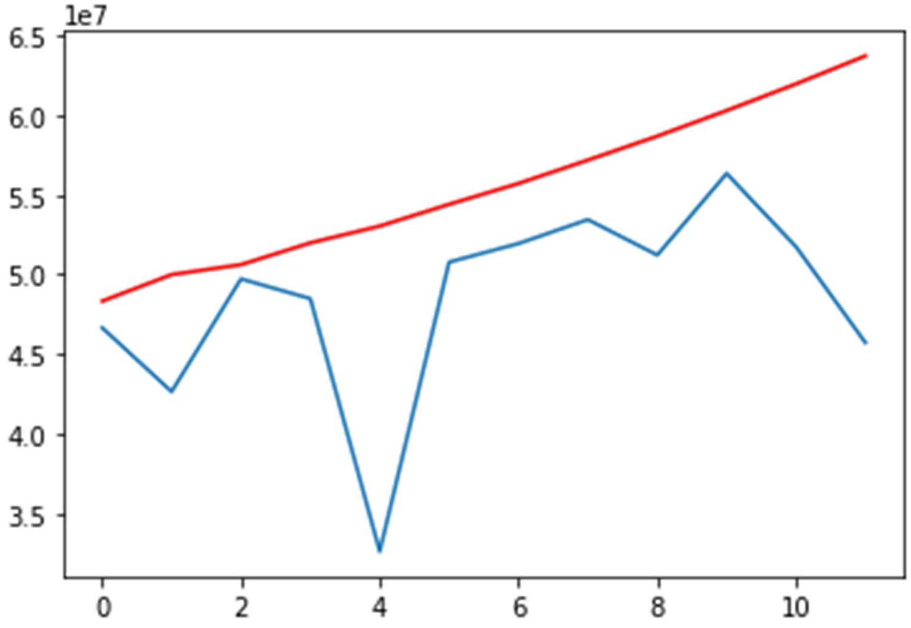
AR model:



Its matching few values correctly to that actual values.

MAE SCORE: 3652323.740

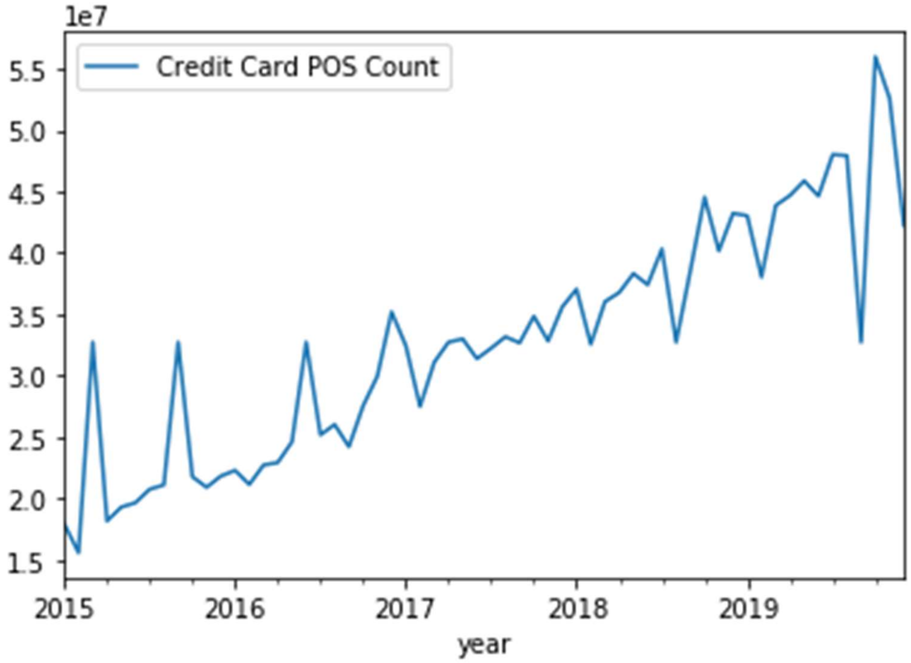
ARIMA MODEL:



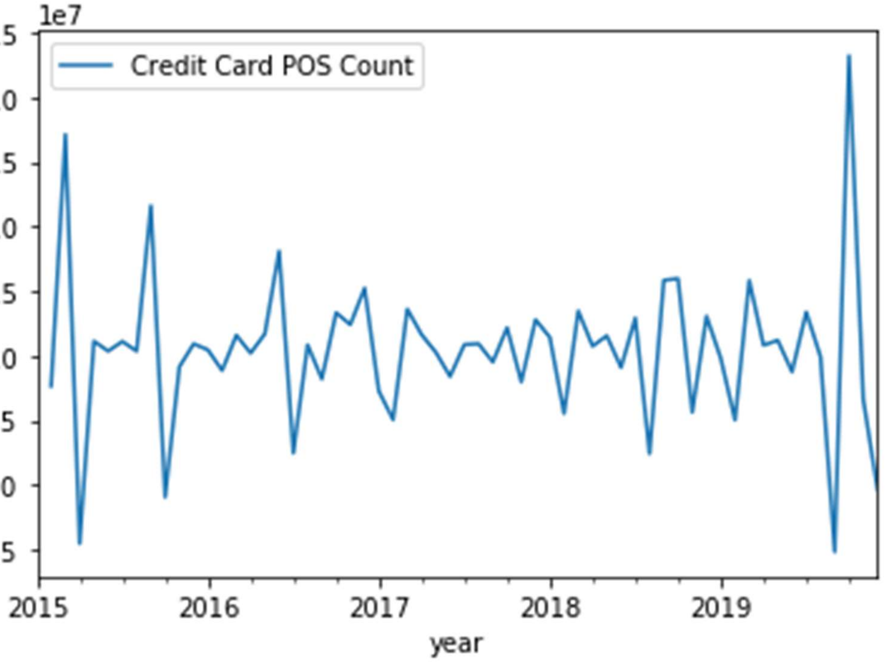
This is the model corresponding to (1,2,0) which has aic value 1602.472

MAE SCORE:7046382.4136

CREDIT CARD:

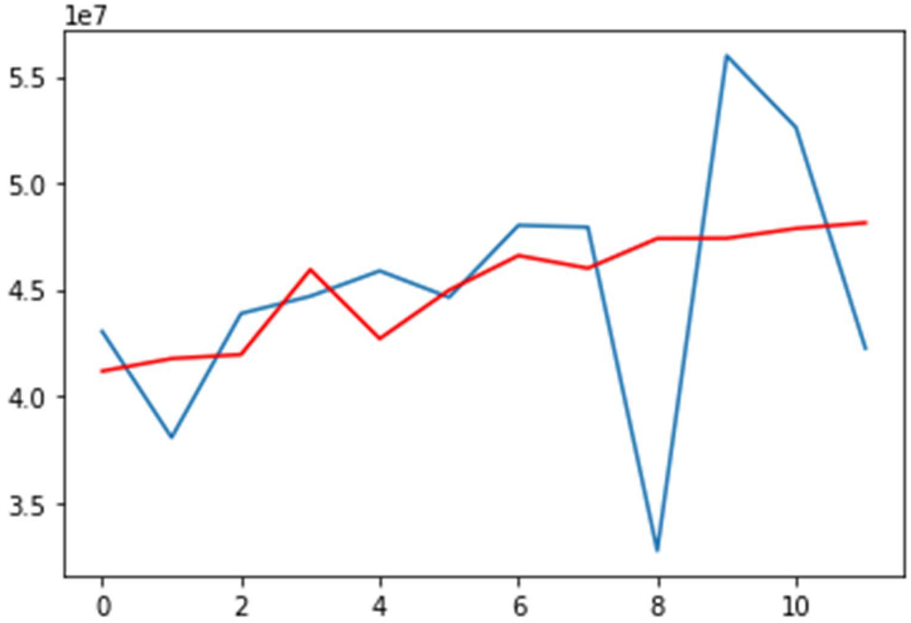


It’s in increasing trend. And this is not stationary.



It’s made stationary by taking difference series.

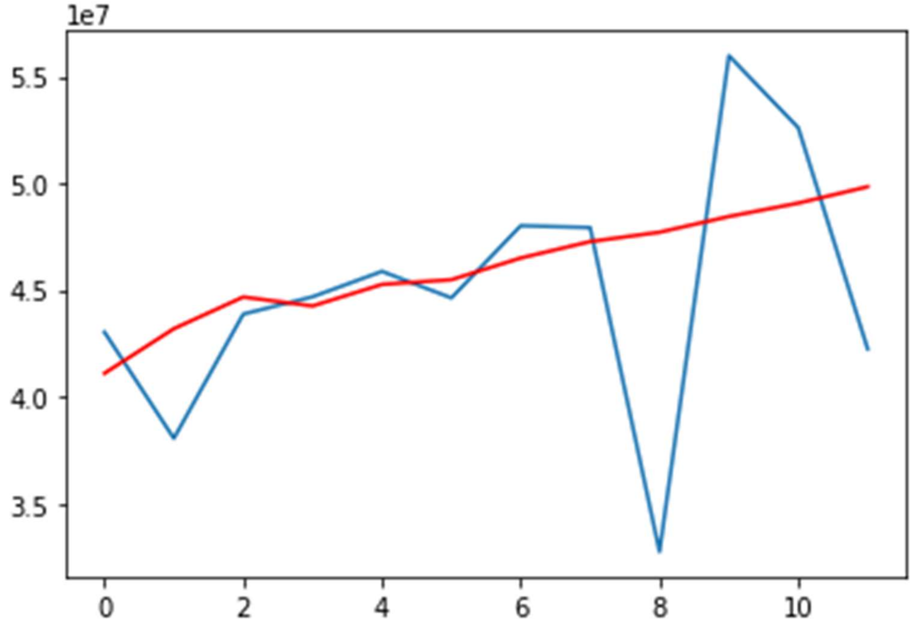
AR model:



Few values are close to actual values and few are not.

MAE VALUE: 4122353.830

ARIMA MODEL:



Can see that few values are close to actual values and few are not. This is the model corresponding to (4,2,1) combination which has minimum aic in the range 0 to 5 i.e. 1547.028

MAE value: 3797136.5481

- Identification of best model backed by logic for selecting it

Mean Absolute Error (MAE) is the best criteria to judge the model.

In case of Approach 1

ARIMA model has the lowest MAE of 8545596.045 hence ARIMA model is chosen as the best model.

In case of Approach 2

AR model has less MAE value of 2147112.4291 hence AR model is the best model for NEFT Transaction.

ARIMA model has lesser MAE of 152995.14 hence ARIMA model is the best model for RTGS transaction.

AR model has less MAE value of 7046382.4136 hence AR model is the best model for Debit Card Transaction.

ARIMA model has low MAE value of 3797136.5481 hence ARIMA model is best model for Credit Card Transactions.

- Lessons learnt from project

Understanding of data from both the perspective i.e. technical and functional (Domain) is very important to proceed ahead with data cleaning and exploration. It is the base of constructing a good classification model.

Data analysis and cleaning is an important task as at this stage lot of important decision are to be made i.e. whether to replace blank data with NaN value or Mean value of rest of the data. Data analysis helps to decide kind of encoding to be used for conversion of categorical data to numerical.

Data exploration helps to visualize the relationship between predictor and response variable. It helps to identify highly correlated variable which can be attributed to train / test data.

More the number of techniques, better the chances to find anomalies or error. Merely applying only technique will not give best model hence it is advisable to use all the techniques and compare the results to find out best of model for a given data set.