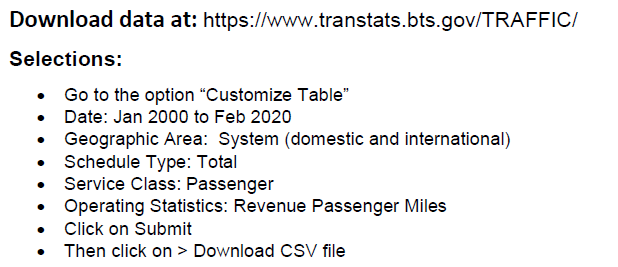
**Problem statement:**

We have the data for the Revenue Passenger Miles for U.S Air Carrier Traffic for the domestic and international Geographic Area. The task here is to predict the revenue for the next 1 year based on the data we have from Jan 2000 to Feb 2020.

Below steps were performed as a part of data collection:

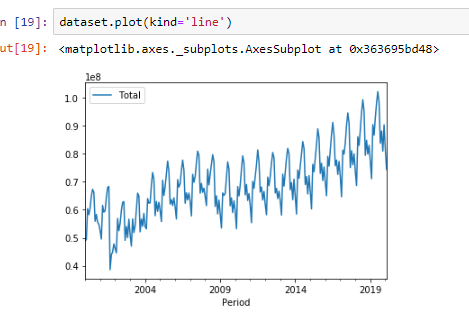
<https://www.transtats.bts.gov/TRAFFIC/>



**Model Preprocessing:**

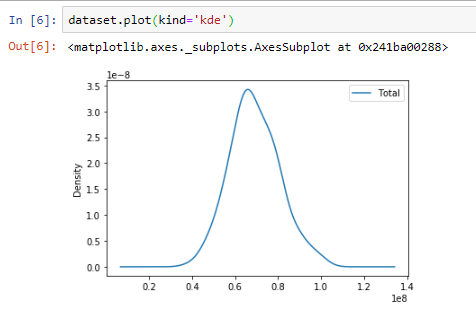


Here we have gathered data from csv file which contains data for 20 years i.e. from Jan 2000 to Feb 2020 which contains revenue on daily basis.



Here we have plotted the line graph for the data to check whether the data contains the trend and the season.

As per the above screenshot we can conclude that data is seasonal and it also contains some upward trend.

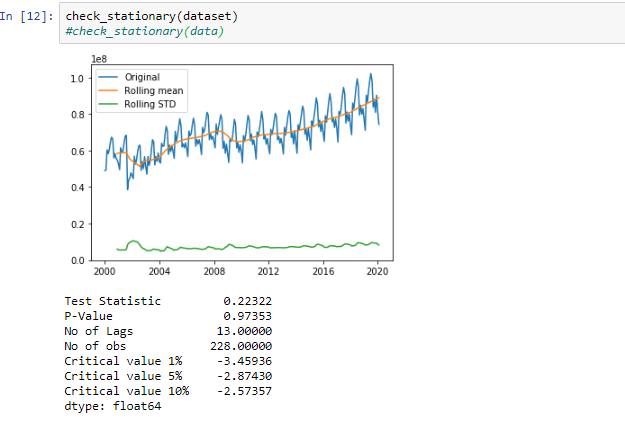


Using above graph we can conclude that data follows the normal distribution.

Now checking if our dataset is stationary or not. There are 2 methods to check whether the stationary or not.

1. Rolling mean
2. Dickey fuller test

We have created a method to check whether data is stationary or not.



In above graph we can see that the mean and std are not constant i.e. the data is not stationary.

Also we can see that p-value is also higher

And test-statistic is greater than the critical value.

Now we will take order of differentiation as 1 to check whether we get the data stationary or not after the transformation.

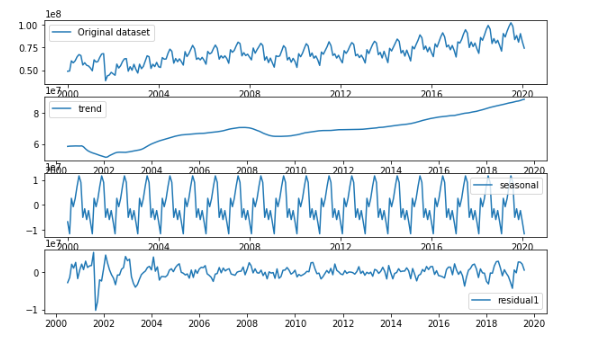


Here we can see that our data is stationary now after taking order of differentiation as 1.

I.e. we have the constant mean and standard deviation.

***Note: We have also performed the log transformation on the original data and considered d = 1. After performing above step, we got the data which was stationary.***

Now we will check the components of the time series.



First graph shows the plotting of the original dataset.

Second graph indicate that the data contains a trend in it.

Third graph indicate that the data is seasonal in nature.

Fourth graph shows the noise in the data. I.e. the residuals.

**Model building:**

1. **Holts exponential smoothing:**

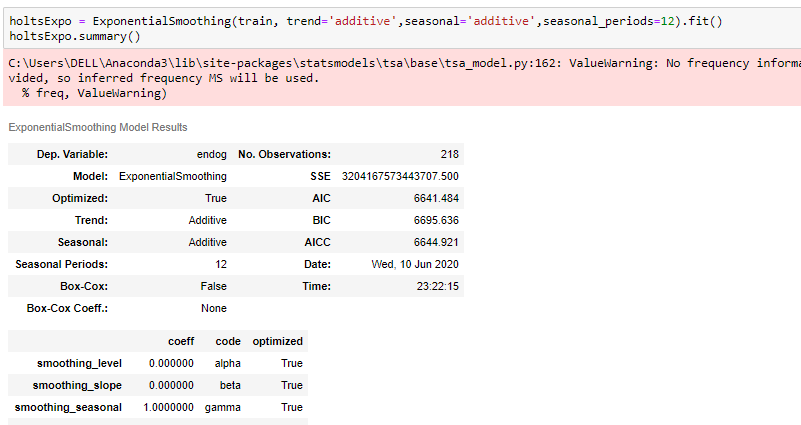
As per the theory we can implement Holts Winters method if the data contains trend and it is seasonal.

So now fitting a model using Holts winters exponential smoothing method.

First converting the data in training and testing set

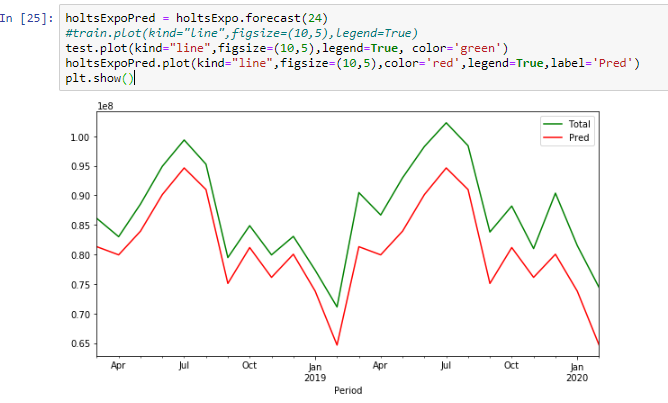


Fitting a model

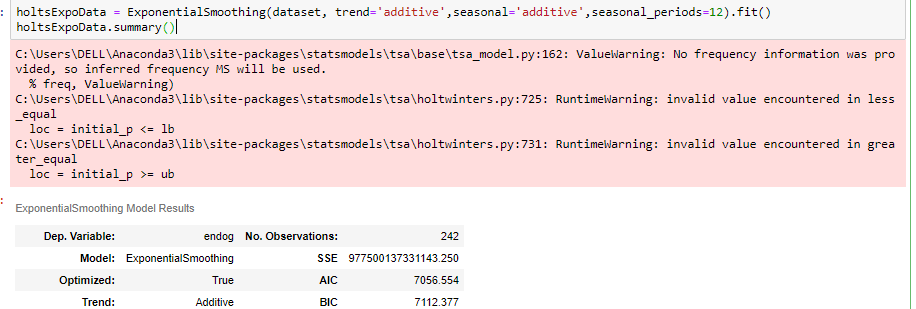


Here we have set the trend and seasonal parameter as additive in nature. Because additive model is useful when the seasonal variation is constant over time.

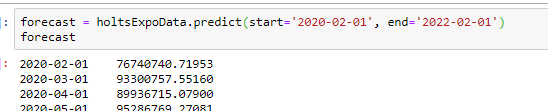
Predictions:



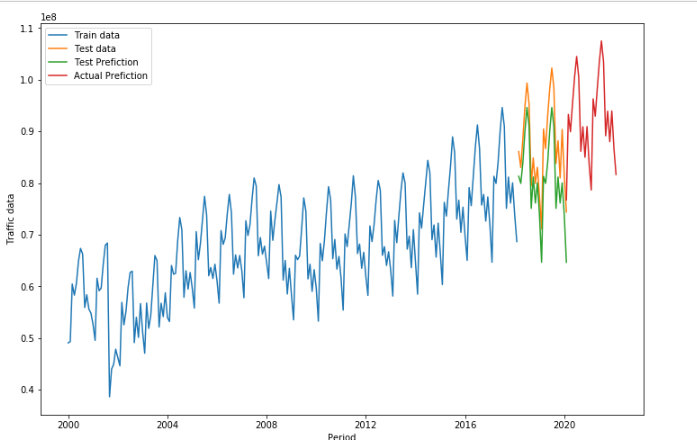
We have also trained the model on the original dataset.

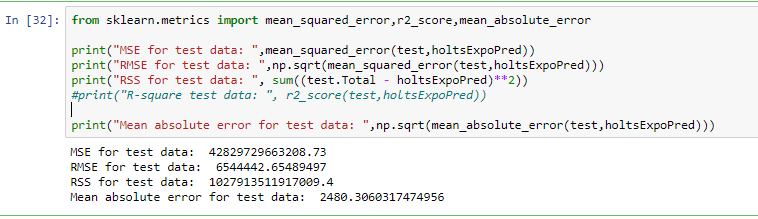


Predictions:



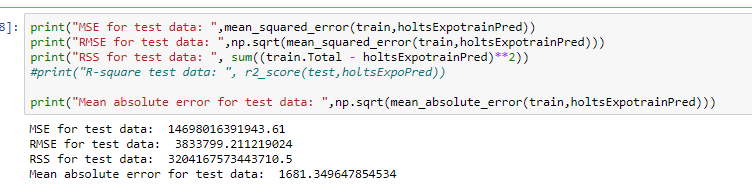
Now plotting the graph for all the predictions





Here we have predicted the error on the test data.

Now predicting error on train data.

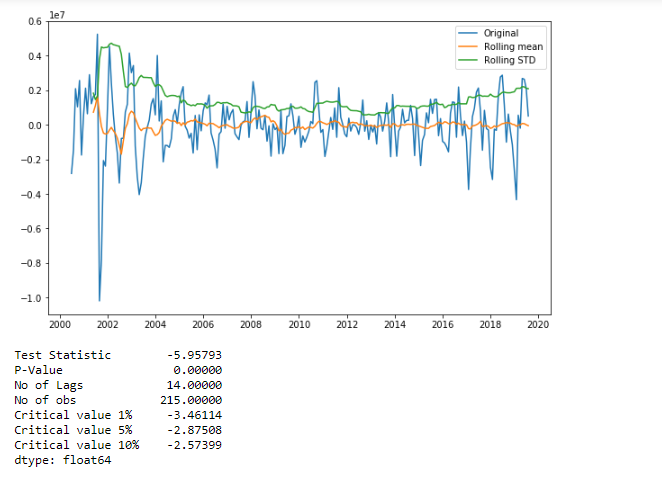


1. SARIMA with normal data:

Data preprocessing will be the same for all the methods.

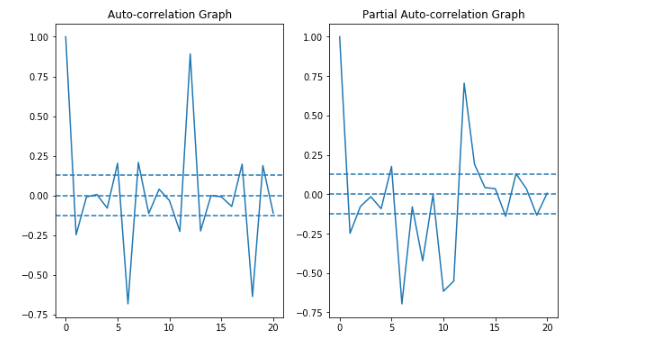
Additional part of data preprocessing:

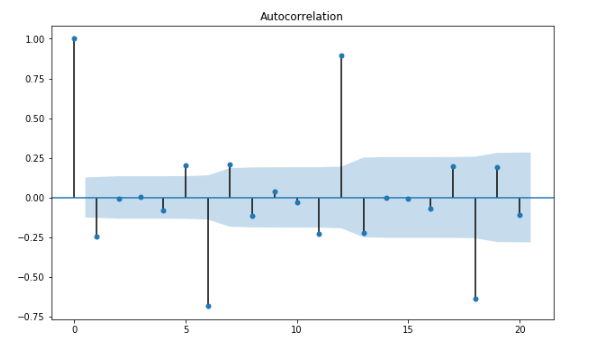
Now we will check if residual follows constant mean and std.

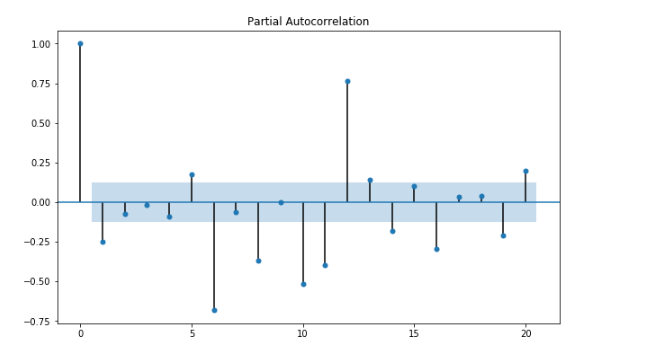


Here we can see that the residuals also constant mean and std.

Test statistic is smaller than critical values and p-values is also relatively small.

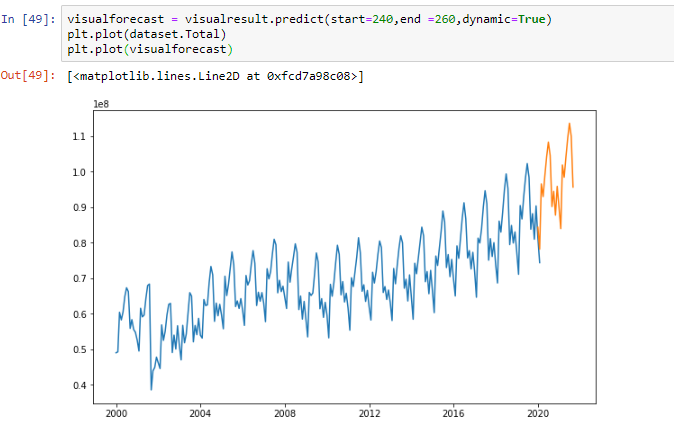






From the above graphs we can consider p and q = 1 at initial stage.

Now build the model using visualized values. i.e. p=q=d=1

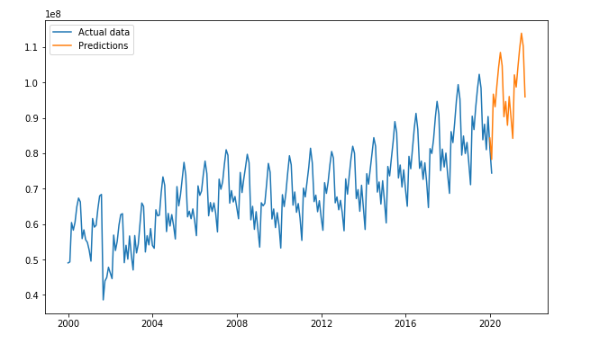


Here we can see that our model has predicted the data for future 20 months.

Now checking with auto\_arima method for the parameters which has less AIC comparatively.

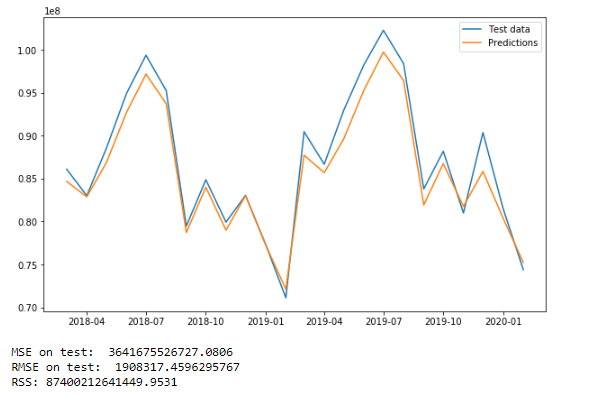
Build the model using same.

Predictions:

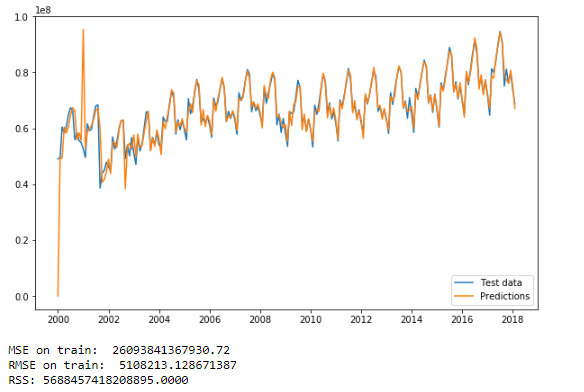


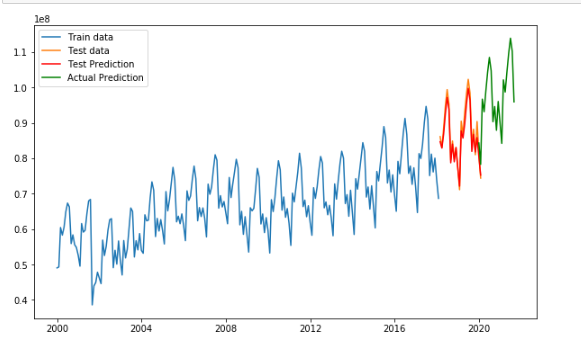
Now create training and testing data.

Fir the model on training data. Predict the test data and calculate the error terms.



Now Predicting for training set





1. SARIMA with Log transformed data:

Same steps are performed in this model prediction. Only as a reprocessing part we have taken the log transformation of the original data for prediction.

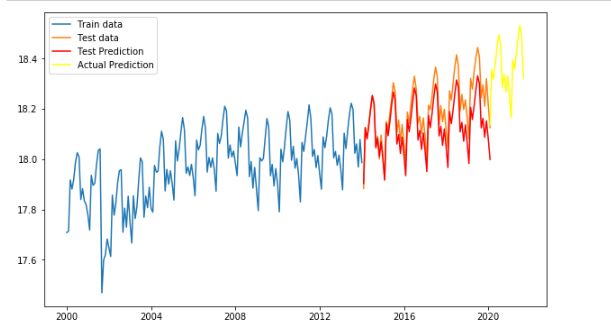
Now errors on test and train data set

Test data:



Train data:

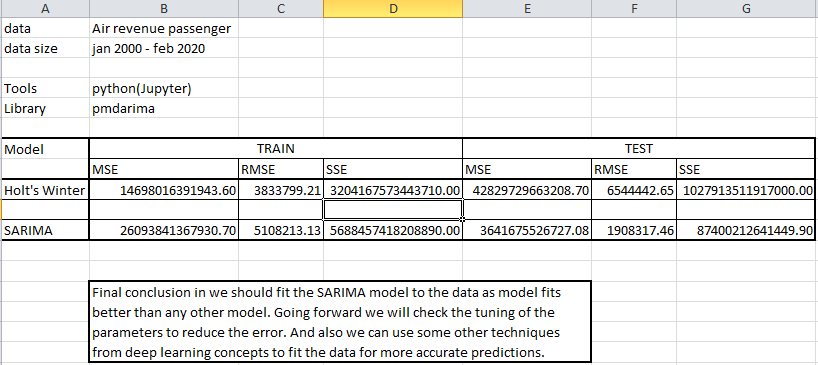




Conclusion:

Holts method has higher RMSE/MSE value than the SARIMA model.

So here we conclude that SARIMA fits well to the data and for future predictions.



Future work:

We can try to fit the other models from the deep learning concepts and compare the RMSE scores. And then based on that we can choose the best model for prediction.