

**Rose Wine**

ANAlysis and Prediction

Suhas Pawar | Time Series Forecasting

# Problem Statement:

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

Data set for the Problem:  [Rose.csv](https://olympus.greatlearning.in/courses/24538/files/2989115/download?verifier=rzSSNxcpSW1kb1sb4HaFEtL7P6hTU0LirfpH8Z4N&wrap=1)

Please do perform the following questions on each of these two data sets separately.

1. Read the data as an appropriate Time Series data and plot the data.
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.  
   Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.  
   Note: Stationarity should be checked at alpha = 0.05.
6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.
8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

# Read the data as an appropriate Time Series data and plot the data.

First of all we will import all the necessary libraries like Pandas, Numpy, seaborn, os to set the path and Statsmodels as base library for all the models.

After setting the path we will import Rose dataset

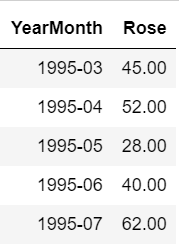
We will read the dataset First by using pandas library.

Then We will Apply head and tail to see how data looks first and last 5 rows.

First 5 Rows:

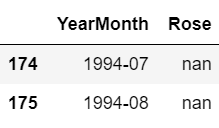


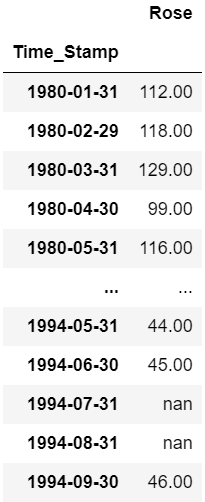
Last 5 Rows:



**Null Value check:**

We will check for null condition, to see whether Dataset has null value or not.



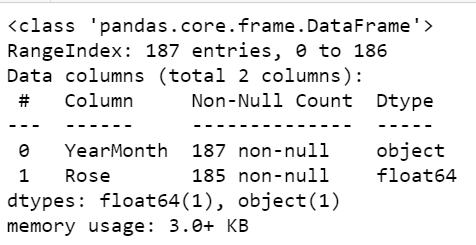
Row no. 174 and 175 has Null values, we will impute this By using  *forward filling method.* 

**Shape of the Data:**

(187, 2)

So Dataset has 187 rows and 2 Columns.

**Data Info:**

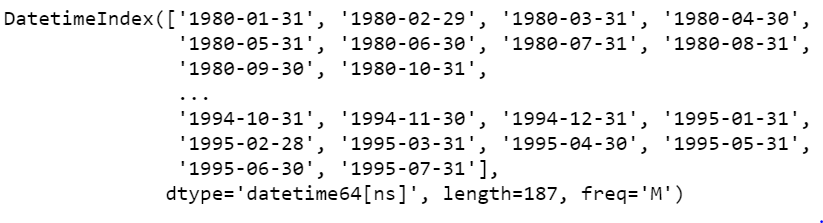


We can see that YearMonth column is of Object Data type while Rose is having float data type.

Rose column has 2 missing values.

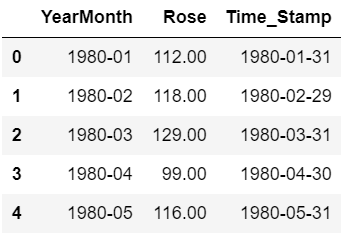
Here we make an assumption that the date starts and ends as mentioned below

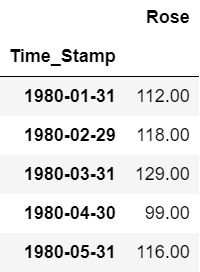
We are assuming the Date Starts from 1980-01-31 to 1995-07-31.



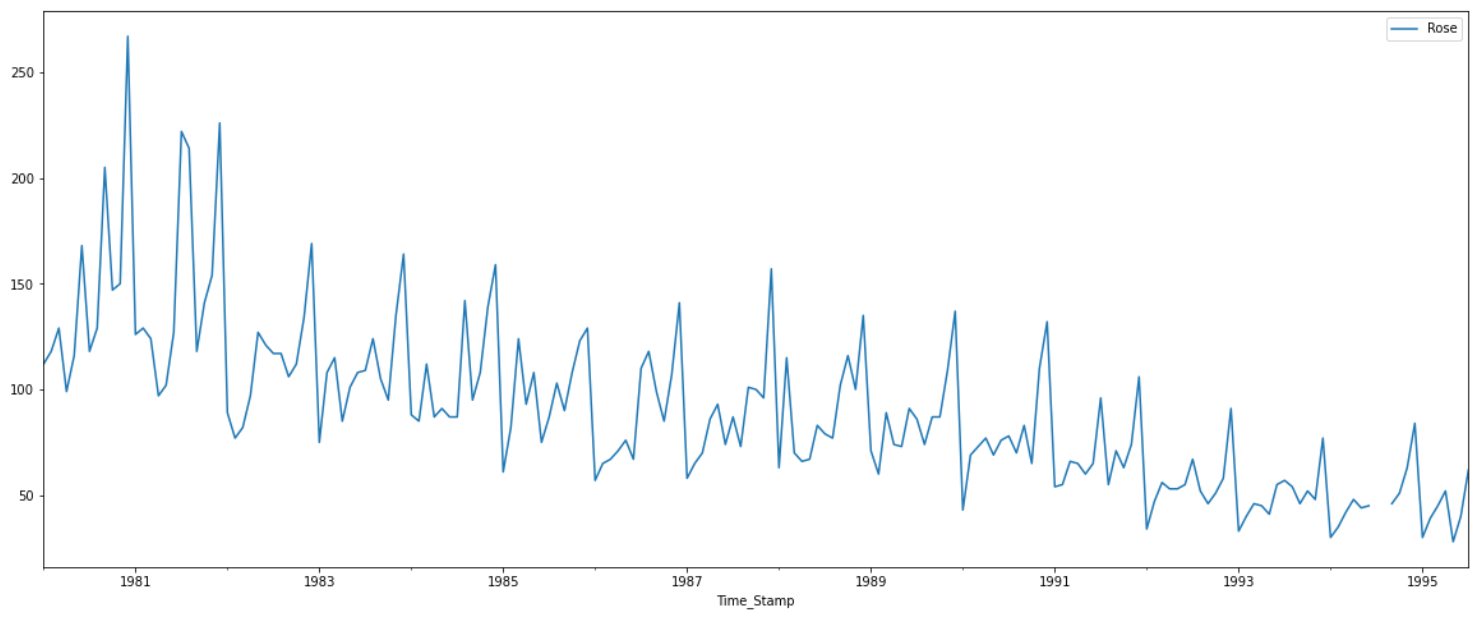
Total 187 entries needs to be added to Original Dataframe while we need to remove YearMonth Column.

And need to make This new Column as Index.





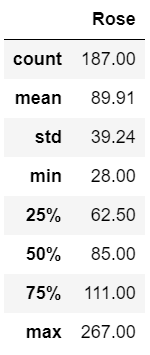
We will plot the original dataset of Rose.



We can see discontinuity in Year 1994-95 as 2 values are missing of Sales-Rose Wines. From Above plot we can see that Data has Negative Trend, Seasonality.

# Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Description:



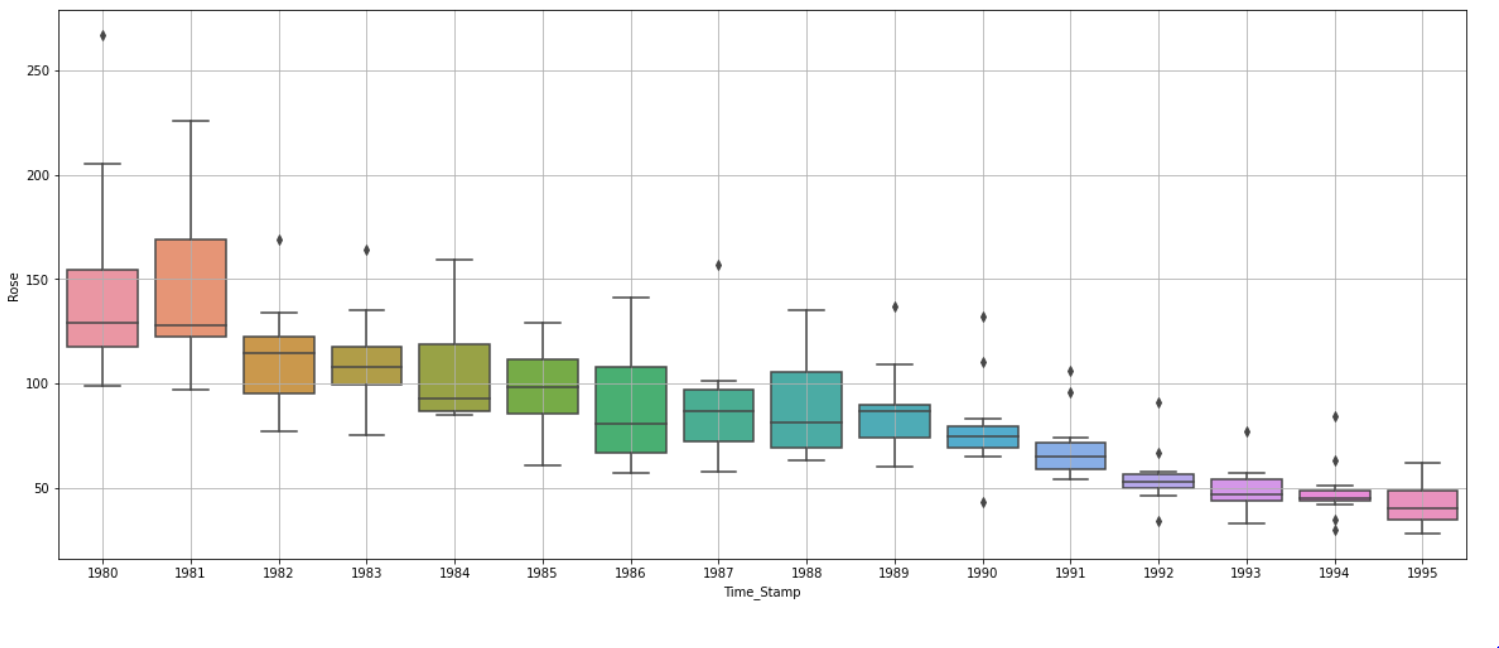
from above describe Function we can see that Highest Sales of Rose wine is 267.Mean and Median are almost same here.

Info check after Imputation of Null values:



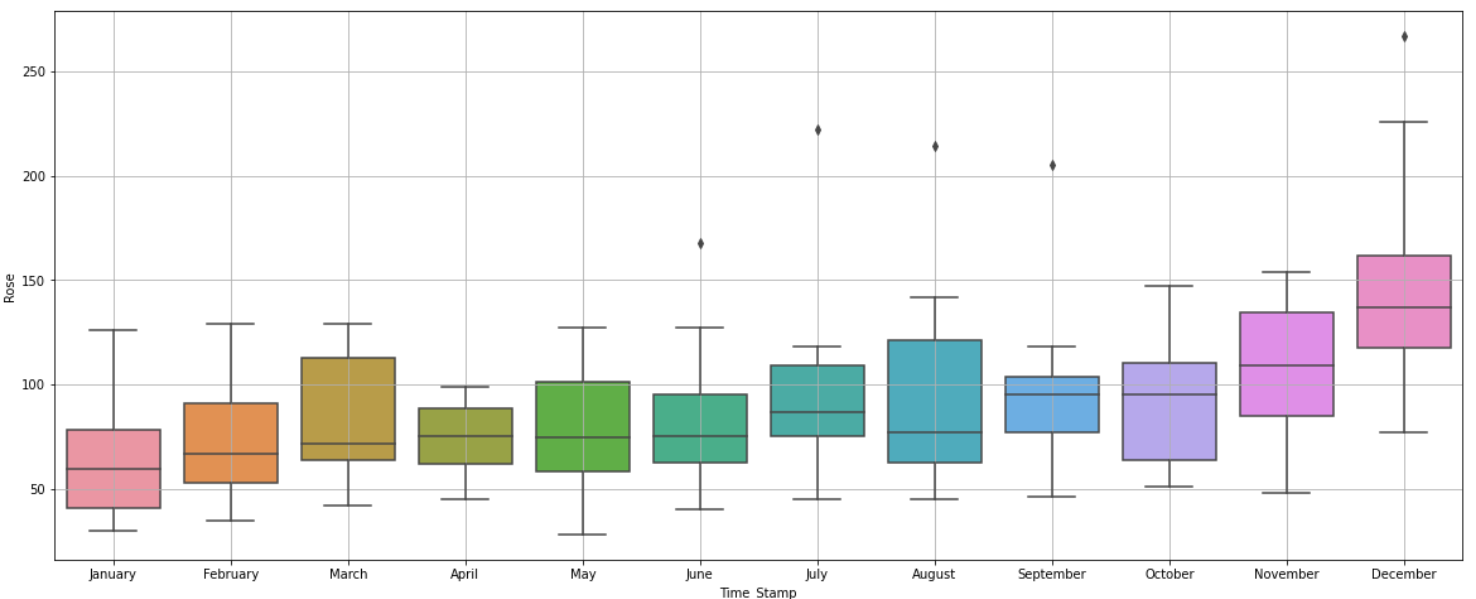
Total 187 rows we have enties from Jan 1980 to July 1995.

**Yearly plot:**

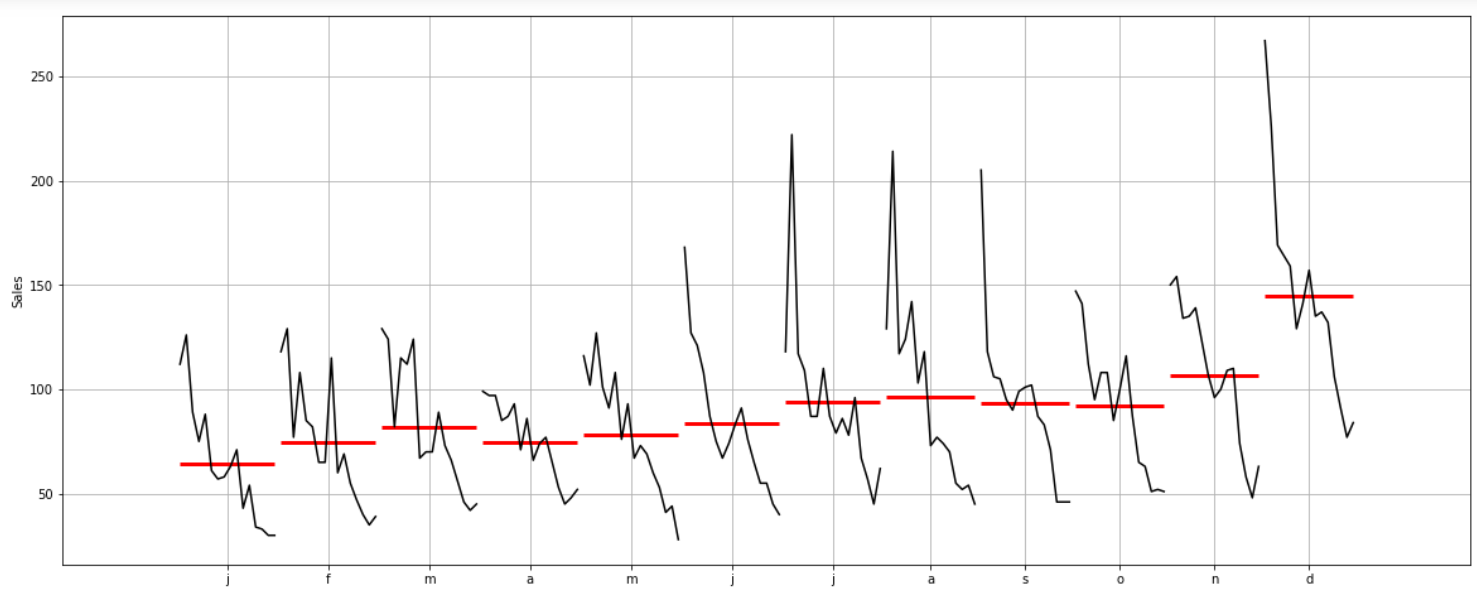


From Yearly plot we can see that Sales of Rose wine are getting down by every year, This continuous decrease in the Sales needs a deep study.

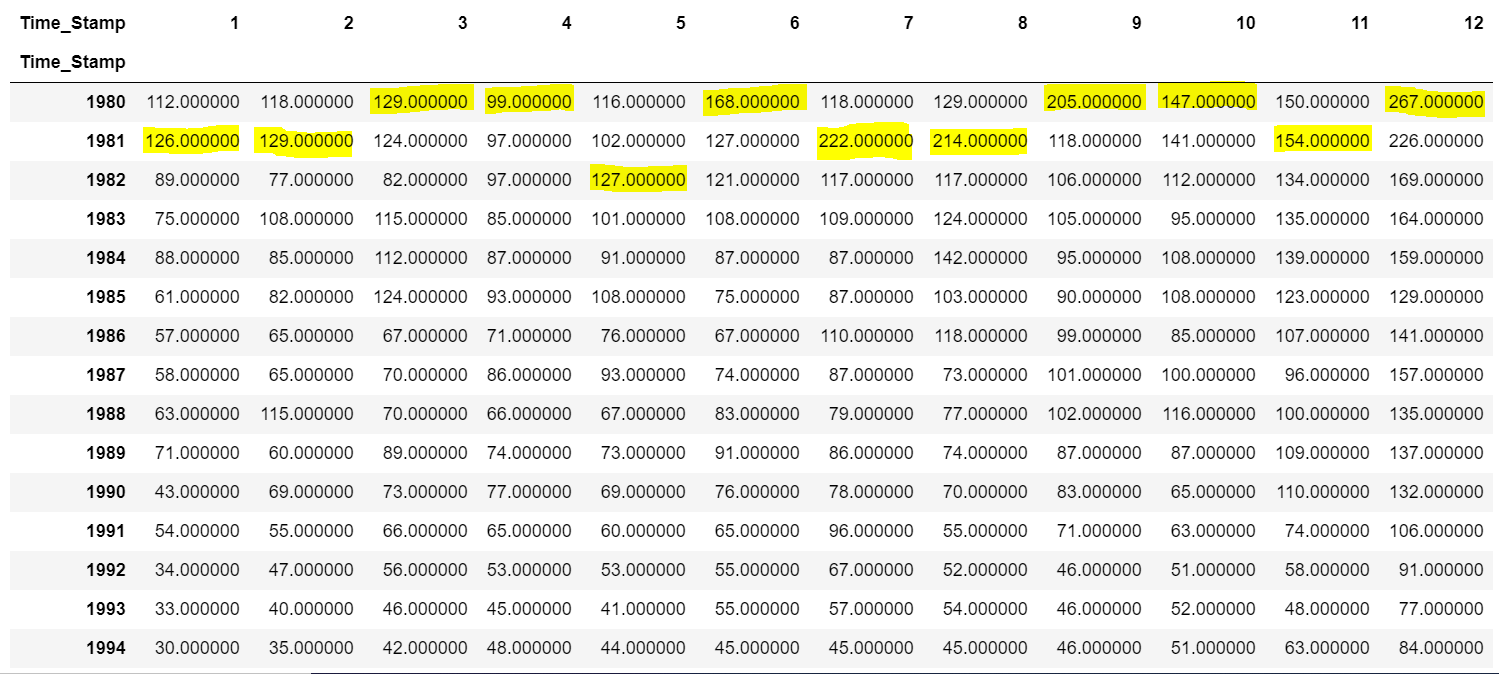
**Monthly plot:**



Here we come with interesting insights from Monthly plot, From January to October Median of Sales is almost Same. Only Sales are increasing in November and December. In This December Only Sales has Crossed figure of 250 wines.

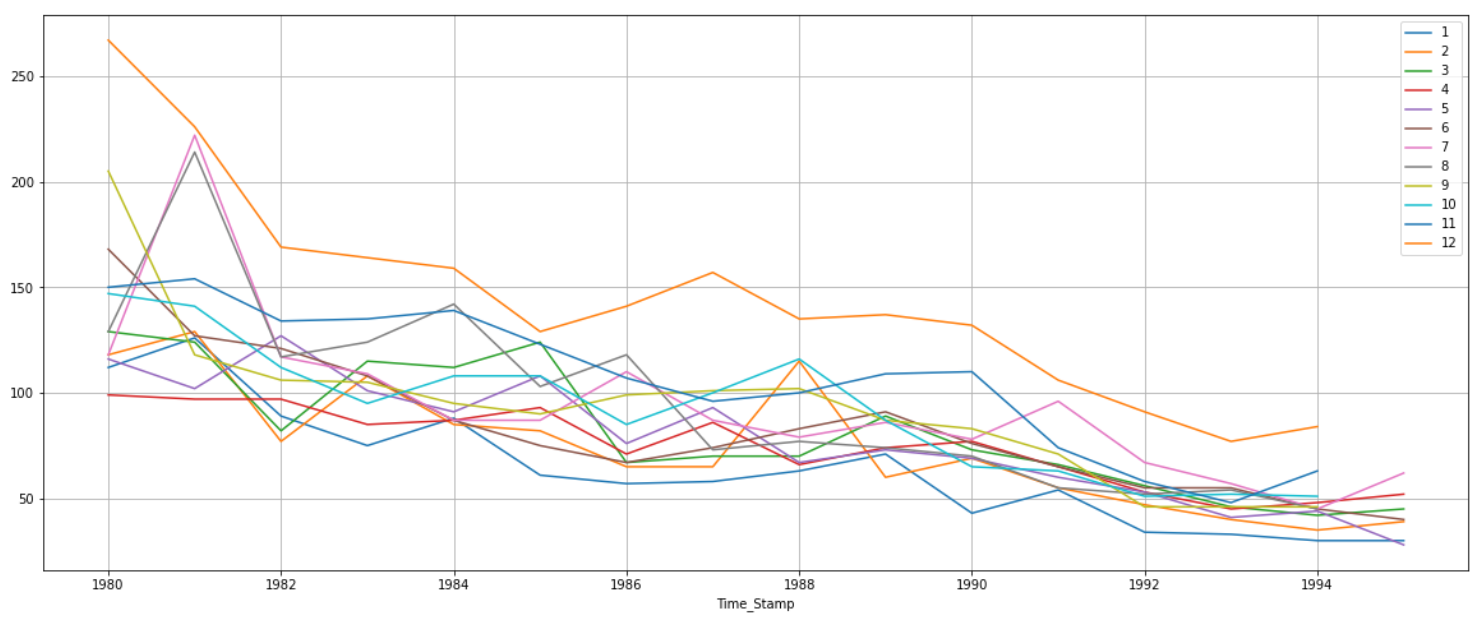


**Highlight The maximum sales:**

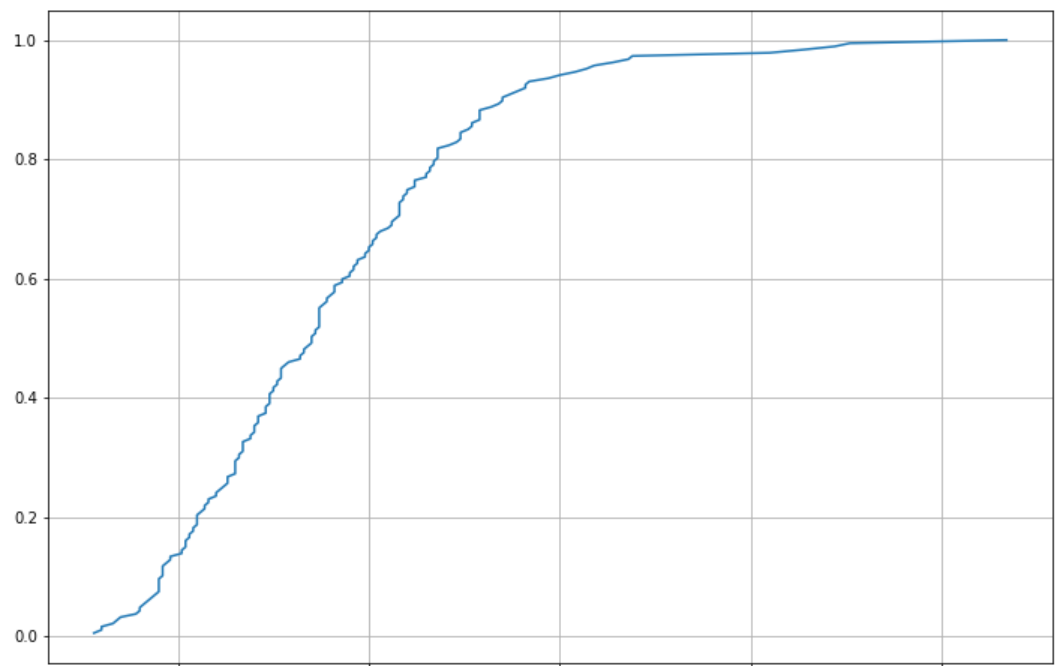


1980 and 1981 are only Two years where Sales are Highest, But from this point the sales has decreased.

The orange line is for December month, It is Alienated from Other lines.

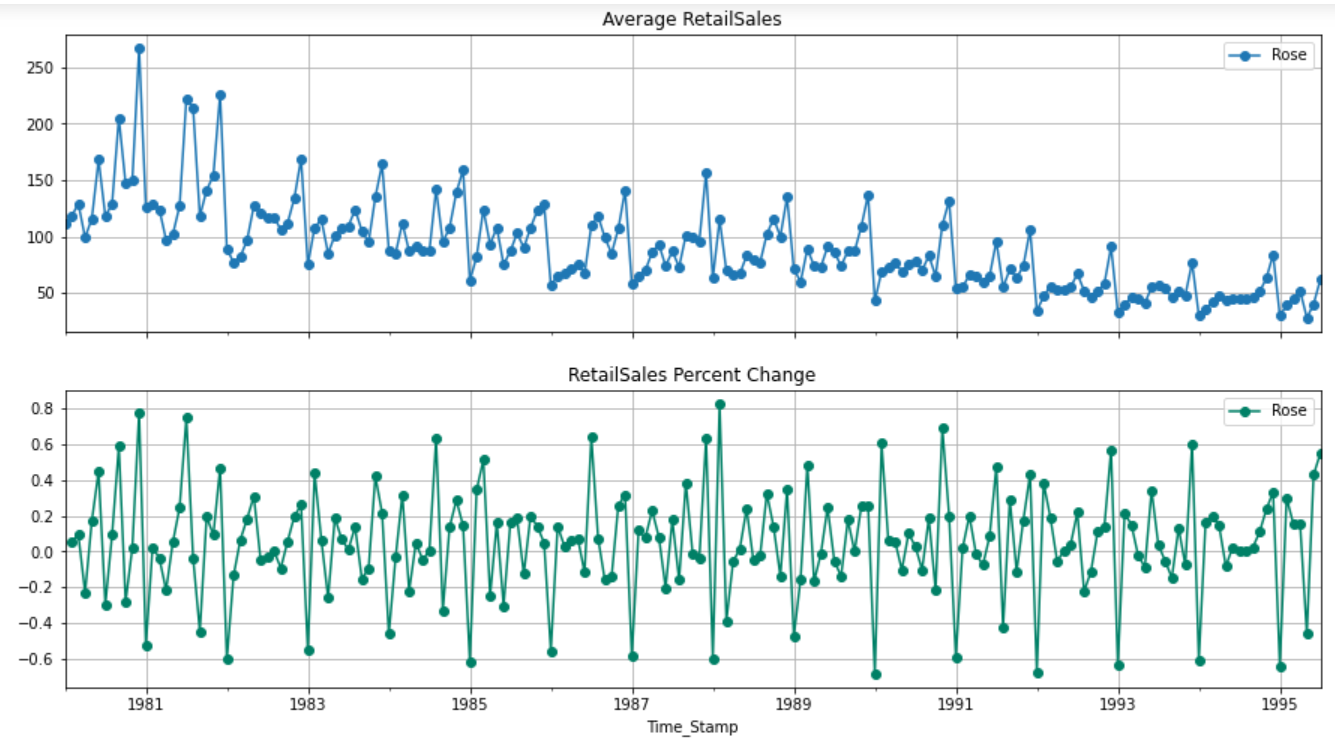


**Empirical Cumulative Distribution.**



This particular graph tells us what percentage of data points refer to what number of Sales. 50% of the sales are below 100. Maximum sales is close 260.

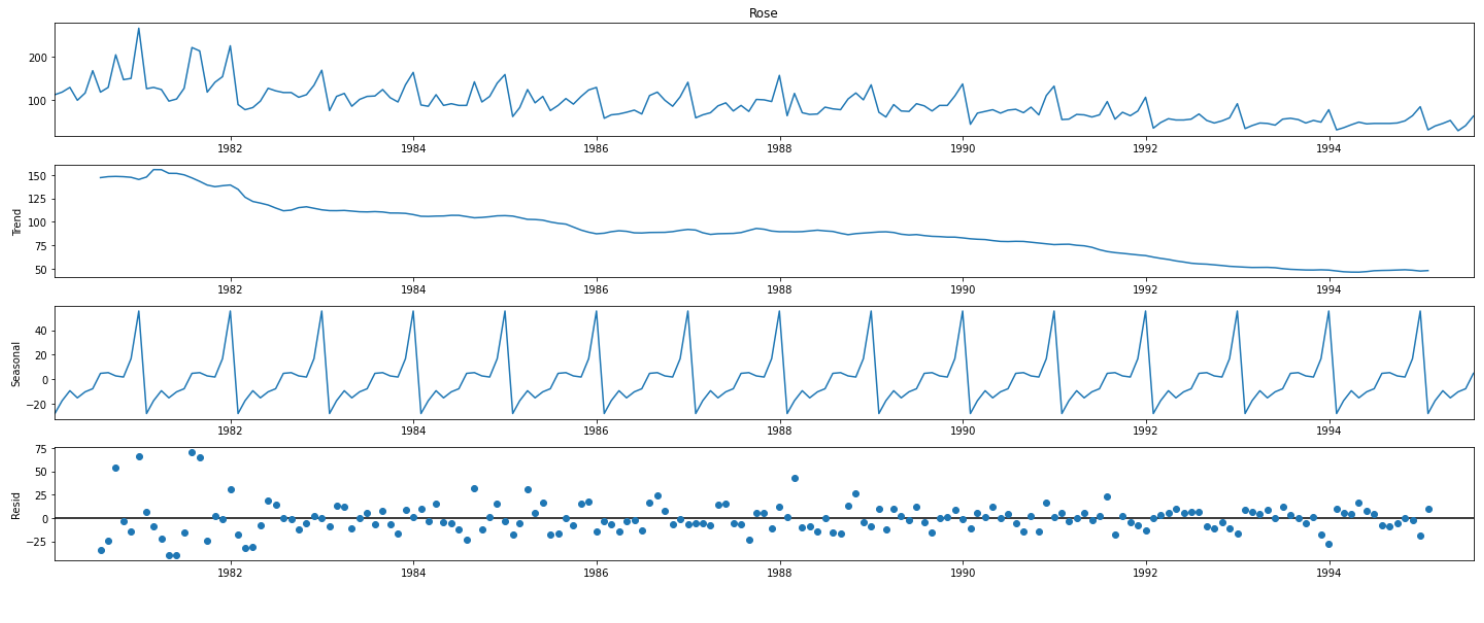
**Plot of average Sales per month and the month on month percentage change of Sales.**



In This plots we can see there is drastic change at the end of every year.

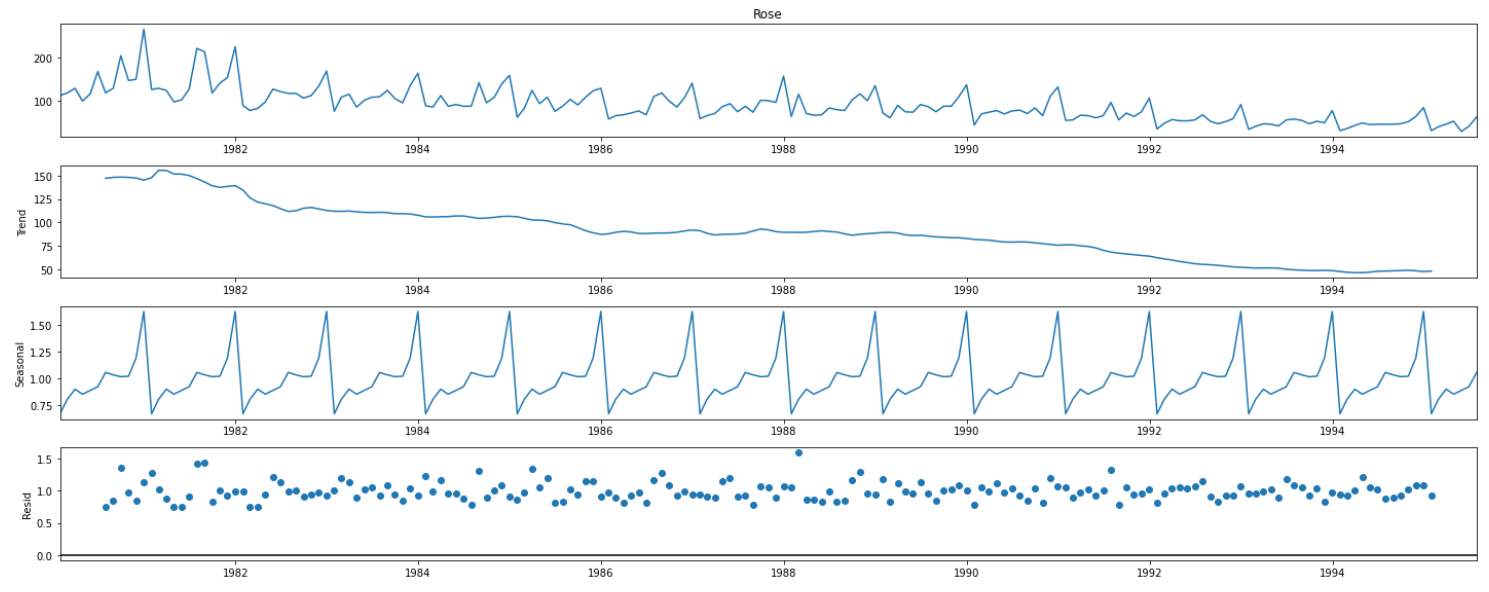
### **Decompose the Time Series and plot the different components.**

Additive Decomposition:



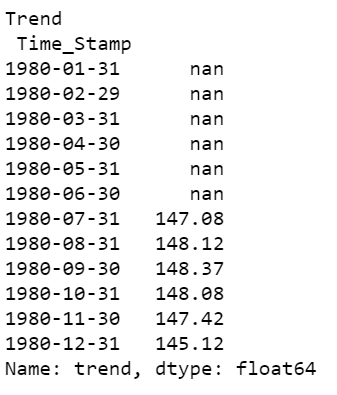
From above Additive Decomposition we can see that There is Negative trend and Seasonality is there.residuals are showing patterns, Might me missing some quality Information.

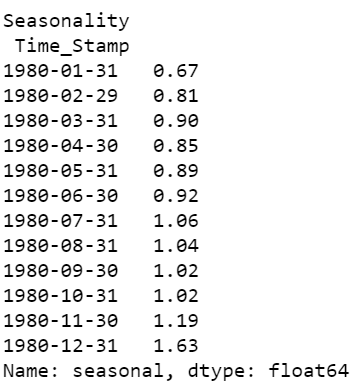
Multiplicative Decomposition:

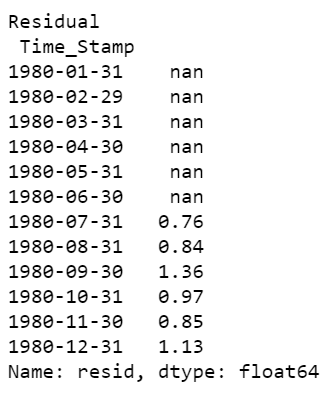


For the multiplicative decomposition series, we see that a lot of residuals are located around 1.

We can see the values of trend and Seasonality for Year 1980. In Case of Seasonality it will mostly repeat and In case of Trend it will keep decreasing overall.

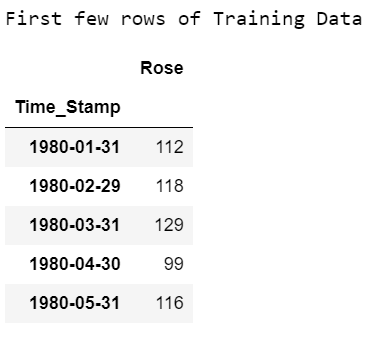


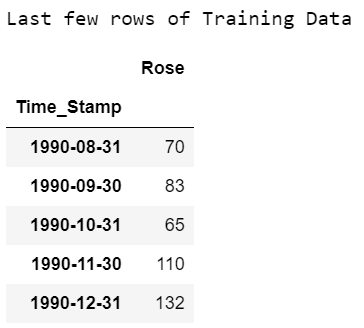


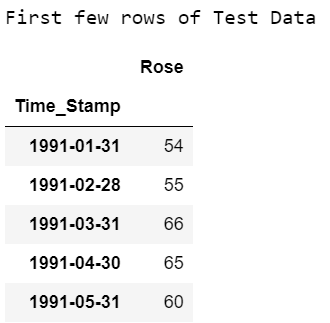


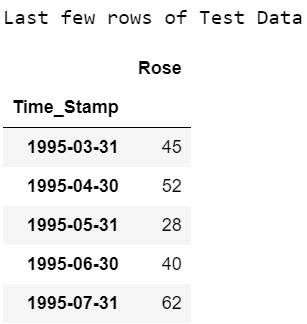
# 3. Split the data into training and test. The test data should start in 1991.

Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.



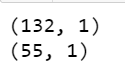




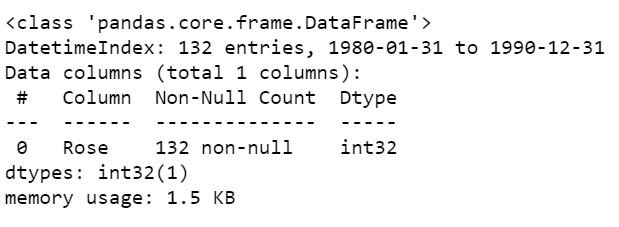


Shape of Train Data: 132 Rows and 1 Column

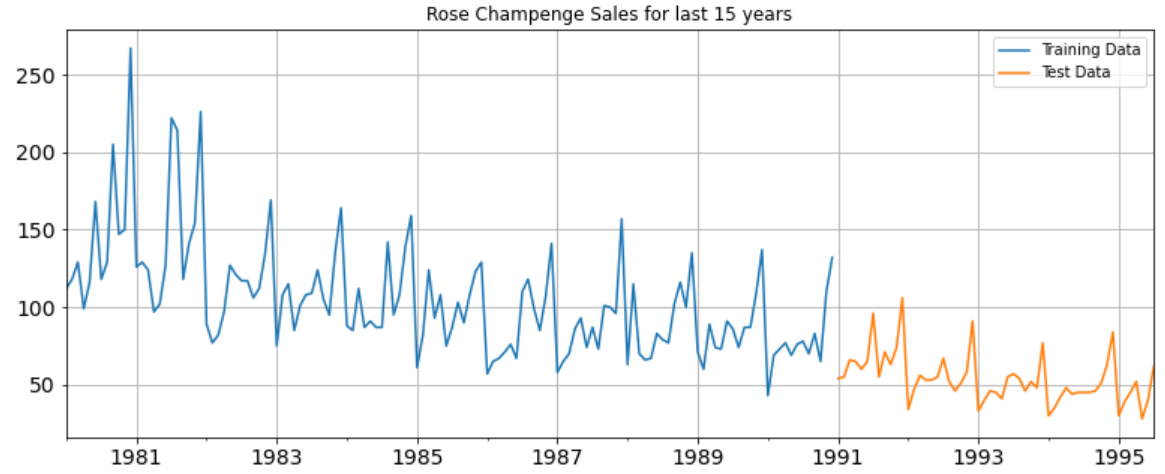
Shape of Test Data : 55 Rows and 1 Column



Train data Information:



**Train and Test data plot:**



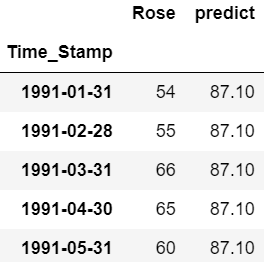
# 4.Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

#### **1.Simple Exponential Smoothing**

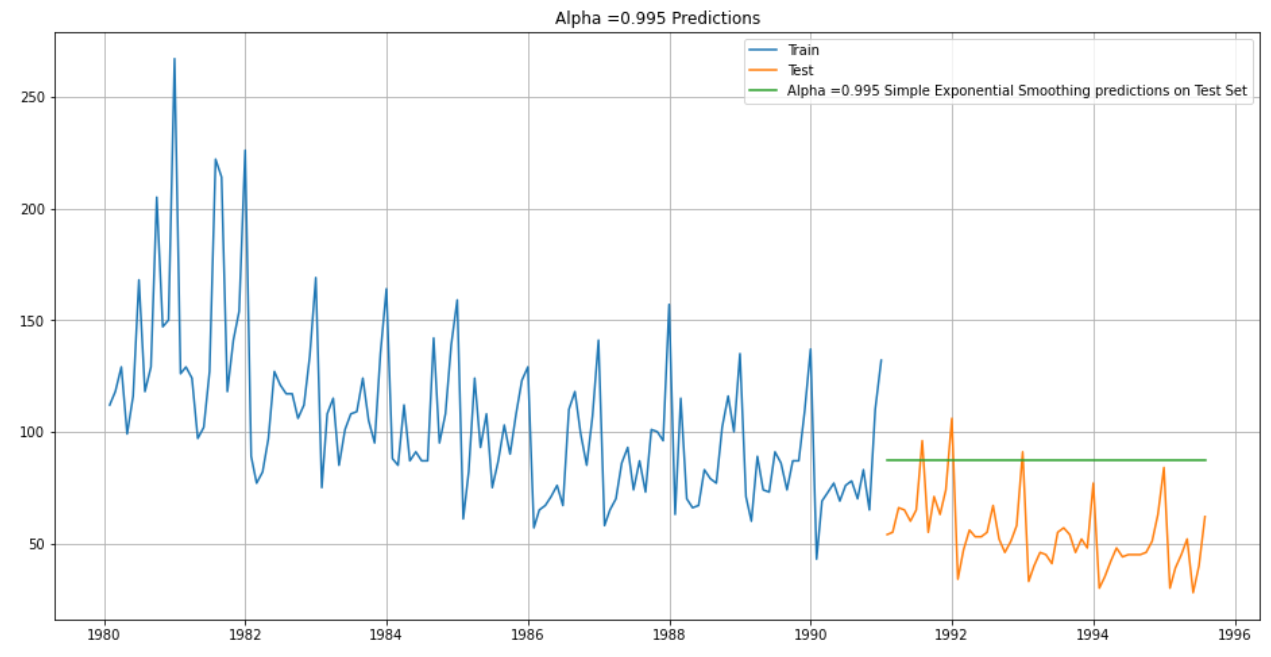
We have created the test and train data for the upcoming Models.

First we will create model on train Data, will test model performance on Test Data.

The Test data predictions for first 5 rows:

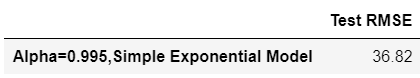


Plotting on both the Training and Test data:



##### ***Model Evaluation for 𝛼 = 0.995 : Simple Exponential Smoothing***

For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.817

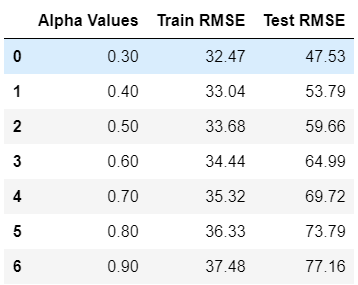


Setting different alpha values. the higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

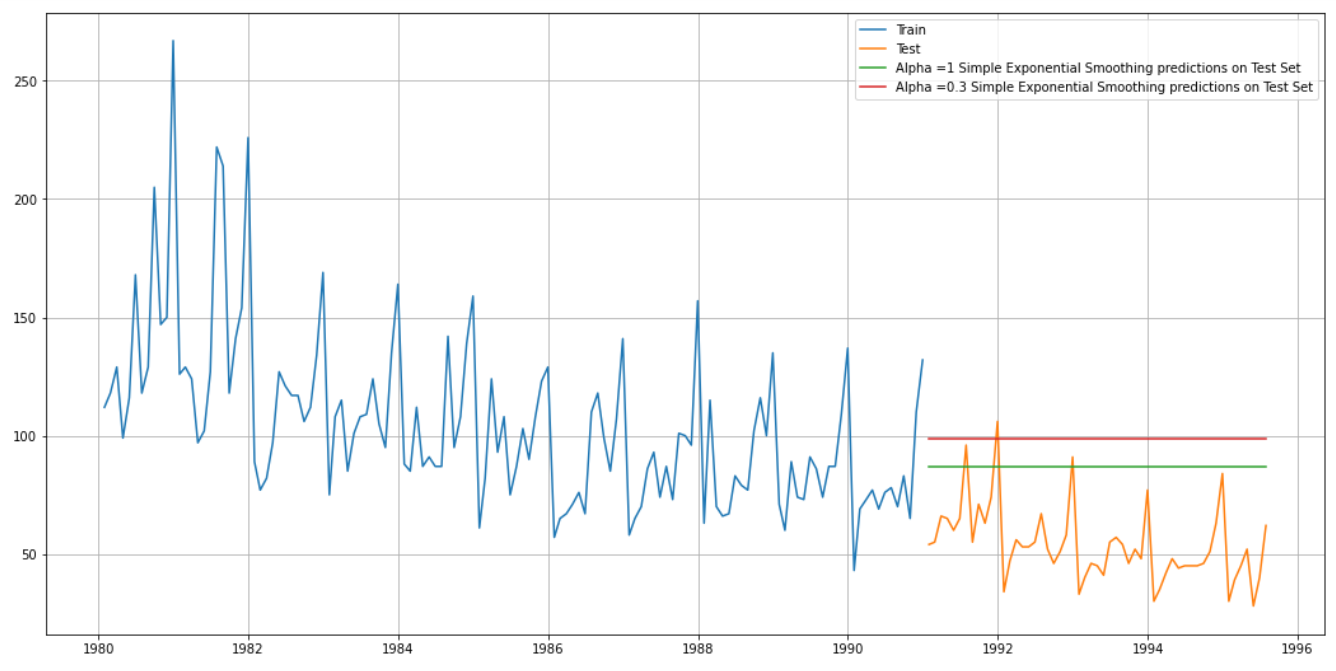
First we will define an empty dataframe to store our values from the loop



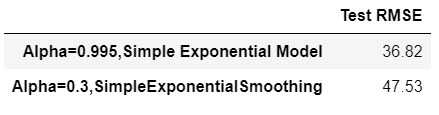
##### ***Model Evaluation:***



Plotting on both the Training and Test data



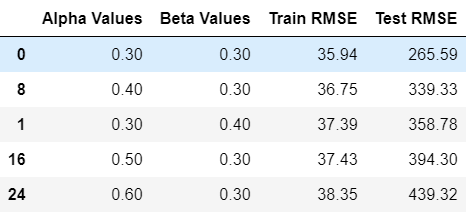
The Final Result of Simple Exponential Smoothing in form of RMSE:



#### **Model 2 : Double Exponential Smoothing (Holt's Model)**

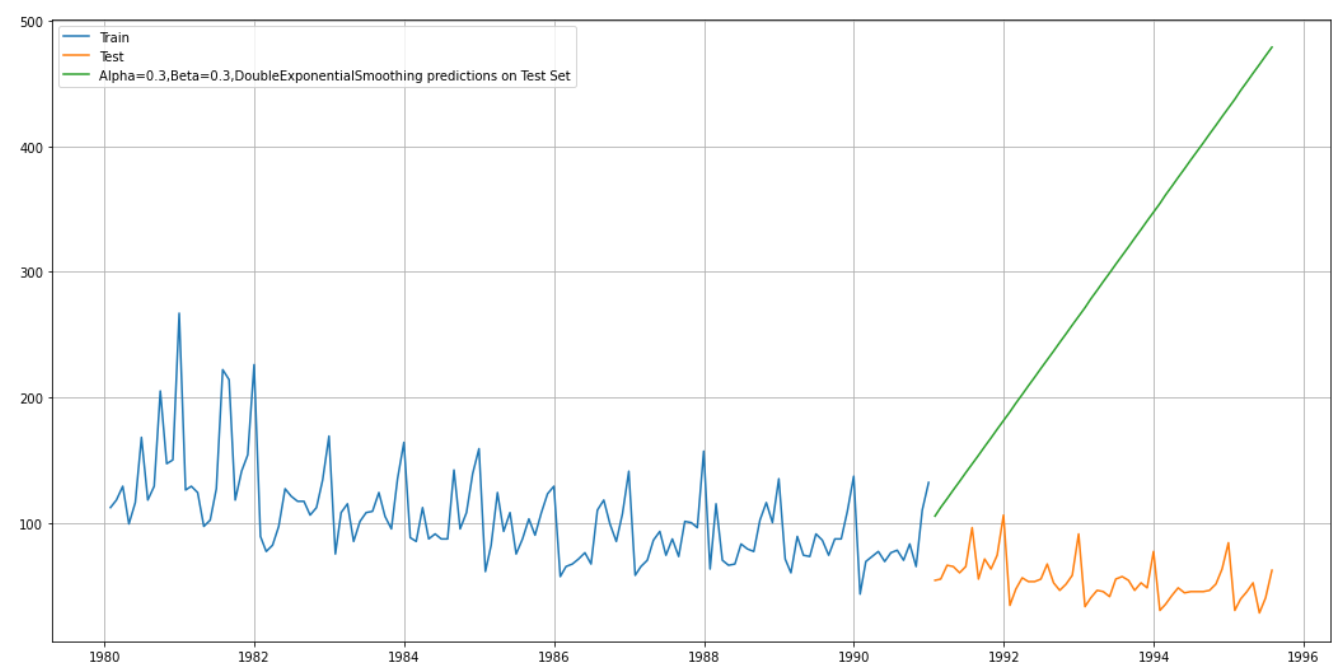
Two parameters 𝛼 and 𝛽 are estimated in this model. Level and Trend are accounted for in this model. First We will create a model and then we will predict on test data.

We will check the results for Different values of Alpha and Beta. Let us sort the data frame in the ascending ordering of the 'Test RMSE' and the 'Test MAPE' values.

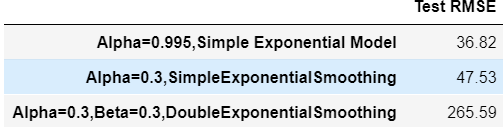


Test RMSE is lowest for Alpha=0.3 and Beta =0.3. hence we will select these values for better performance.

Plotting on both the Training and Test data



**Test RMSE :**

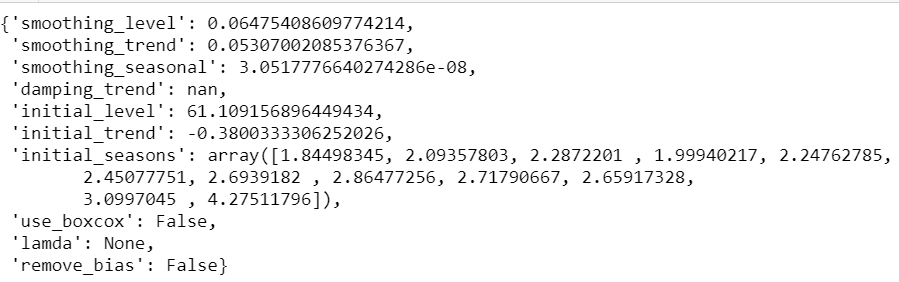


For this the Test RMSE is too high. So we cannot go further with this model. Lets deep dive more, to check about other models.

#### **Method 3: Triple Exponential Smoothing (Holt - Winter's Model)**

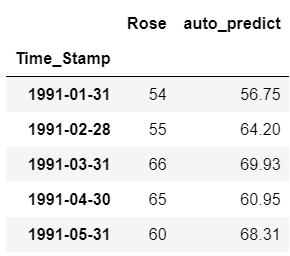
Three parameters 𝛼 , 𝛽 and 𝛾 are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Autofit parameters for the Triple Exponential smoothing:

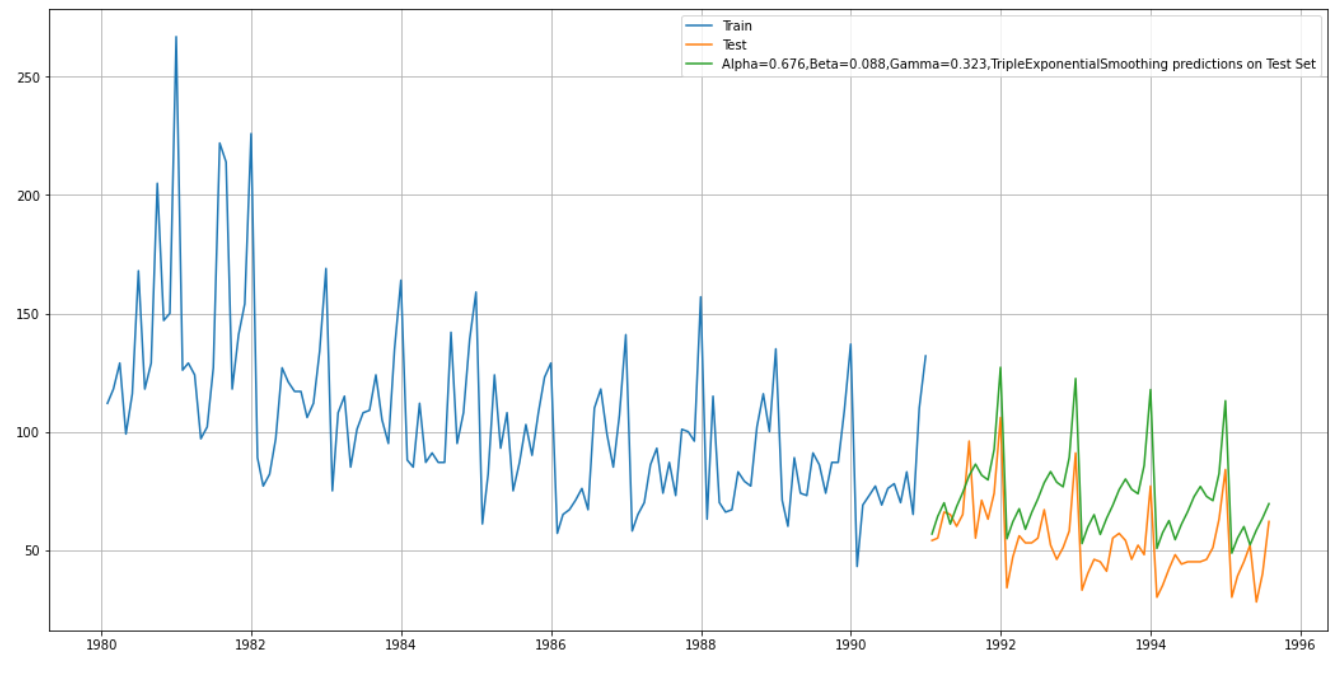


The above fit of the model is by the best parameters that Python thinks for the model. It uses a brute force method to choose the parameters.

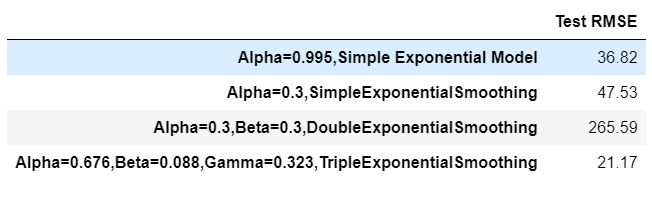
Lets predict on the test data:



Plotting on both the Training and Test using autofit



For Alpha=0.676, Beta=0.088, Gamma=0.323, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 21.170

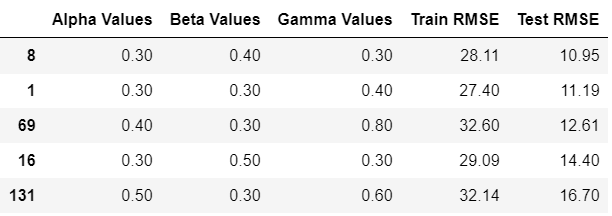


Lets define an empty dataframe to store our values from the loop



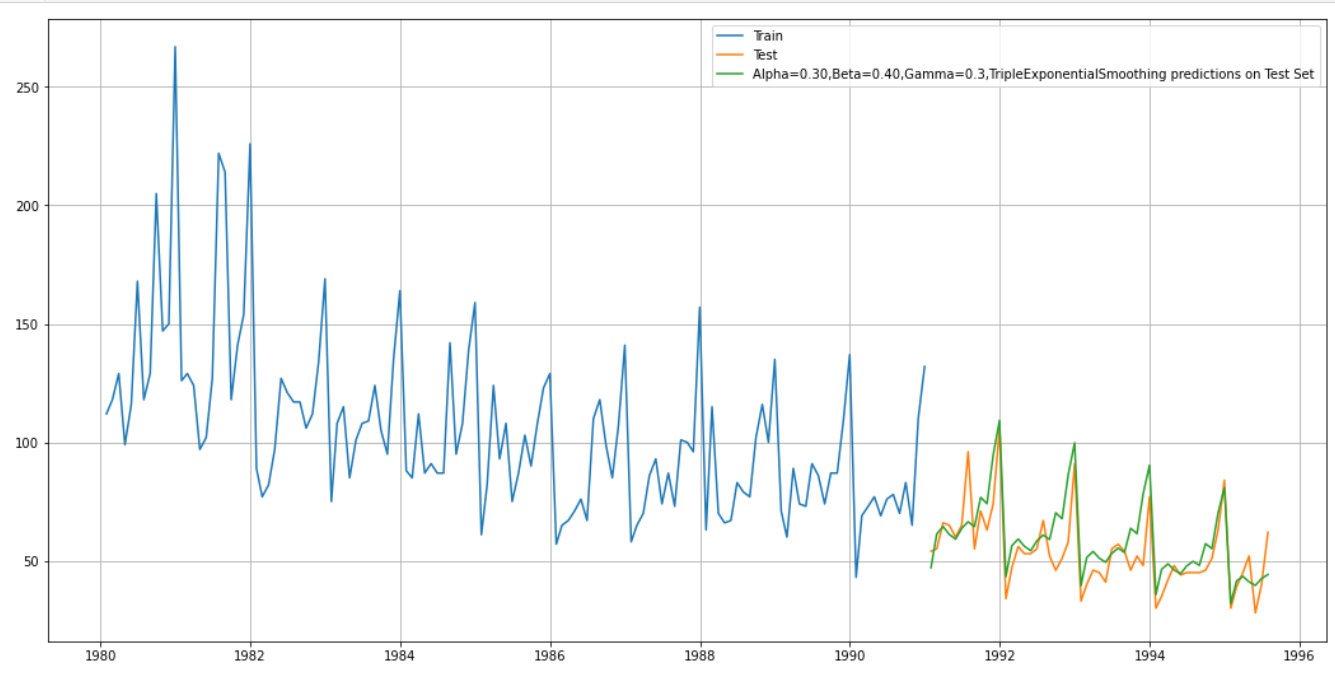
We will append all the results to it, by changing values of alpha, beta , gamma.

After Sorting the first five rows are

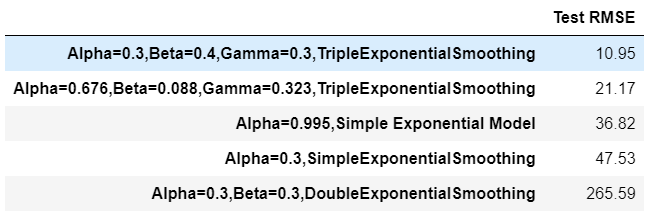


So this model is giving RMSE 10.95 lowest so far, and Ideal values are 0.3,0.4,0.3 for Alpha, Beta , Gamma.

Plotting on both the Training and Test data using brute force alpha, beta and gamma determination



**The sorted results of all the models so far:**



For this data, we had both trend and seasonality so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double Exponential Smoothing. However, since this was a model building exercise we had gone on to build different models on the data and have compared these model with the best RMSE value on the test data.

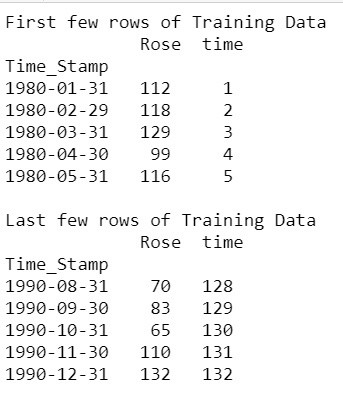
#### **Model 4: Linear Regression**

For this particular linear regression, we are going to regress the 'Rose-Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

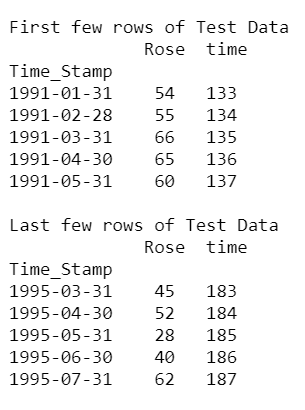


We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

**Training Data:**

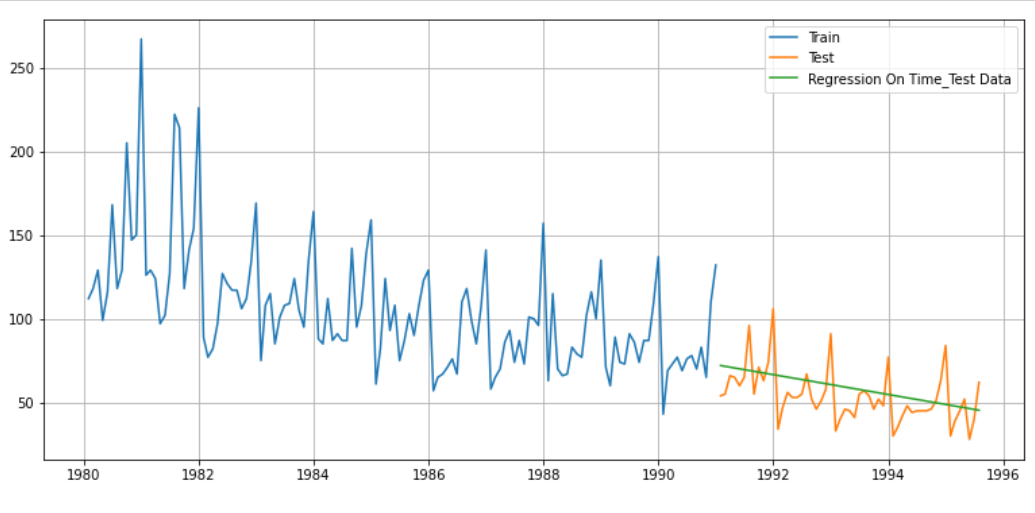


**Test Data:**



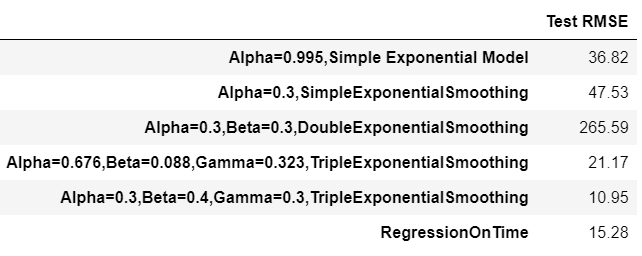
Now our training and test data has been modified, let us go ahead use Linear regression to build the model on the training data and test the model on the test data.

Prediction plot:



#### **Model Evaluation**

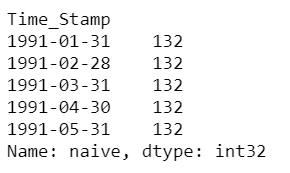
For RegressionOnTime forecast on the Test Data, RMSE is 15.276



#### **Model 5: Naive Approach:**

#### For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today,therefore the prediction for day after tomorrow is also today.

Prediction : First 5 rows:

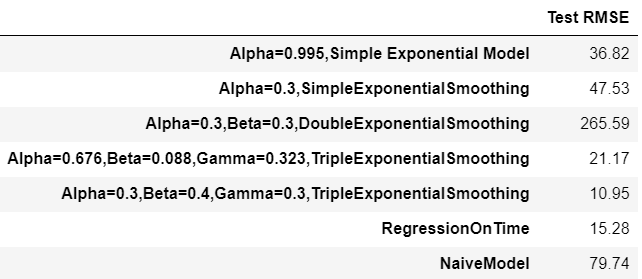


#### **Prediction Plot:**

#### 

#### **Model Evaluation**

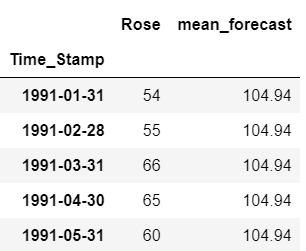
#### For RegressionOnTime forecast on the Test Data, RMSE is 79.739



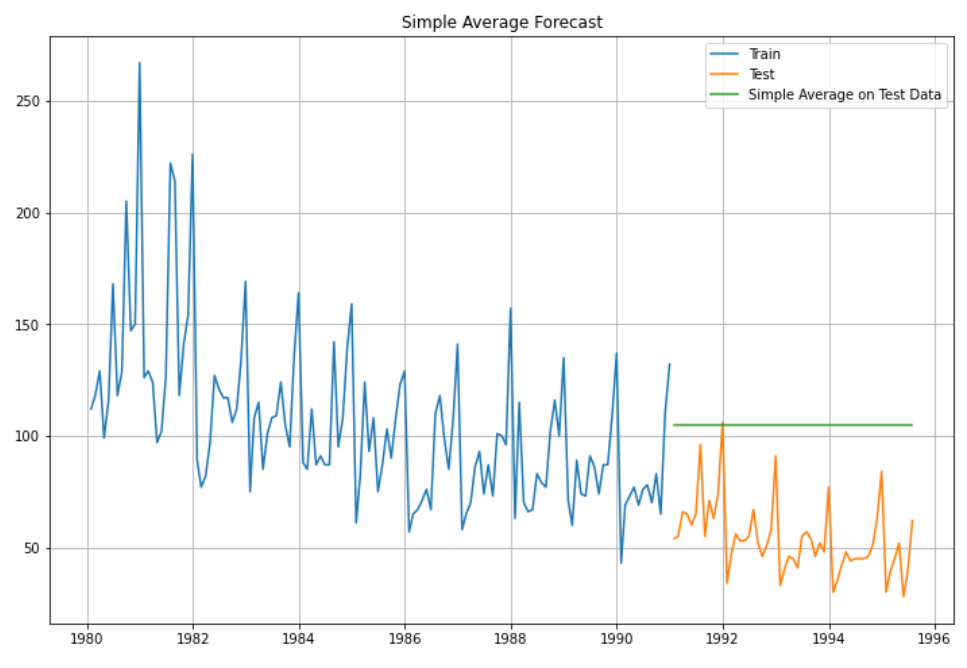
#### **Model 6: Simple Average**

For this particular simple average method, we will forecast by using the average of the training values.

Forecast:

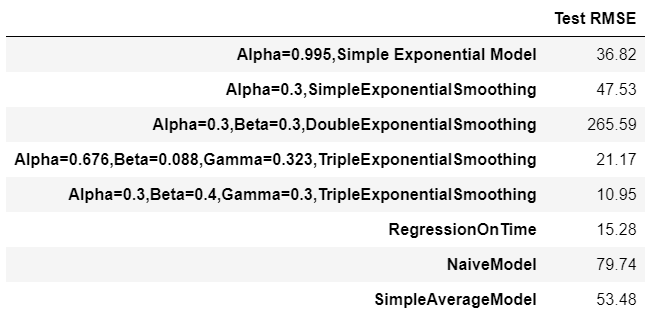


**Prediction Plot vs Test Data**:



#### **Model Evaluation**

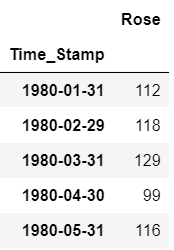
For Simple Average forecast on the Test Data, RMSE is 53.481



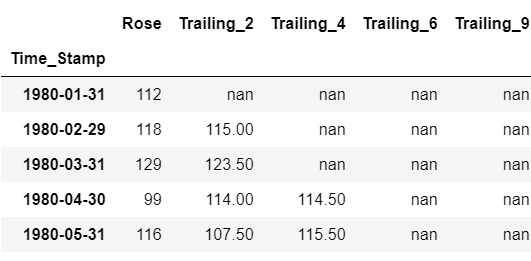
#### **Model 7: Moving Average(MA)**

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

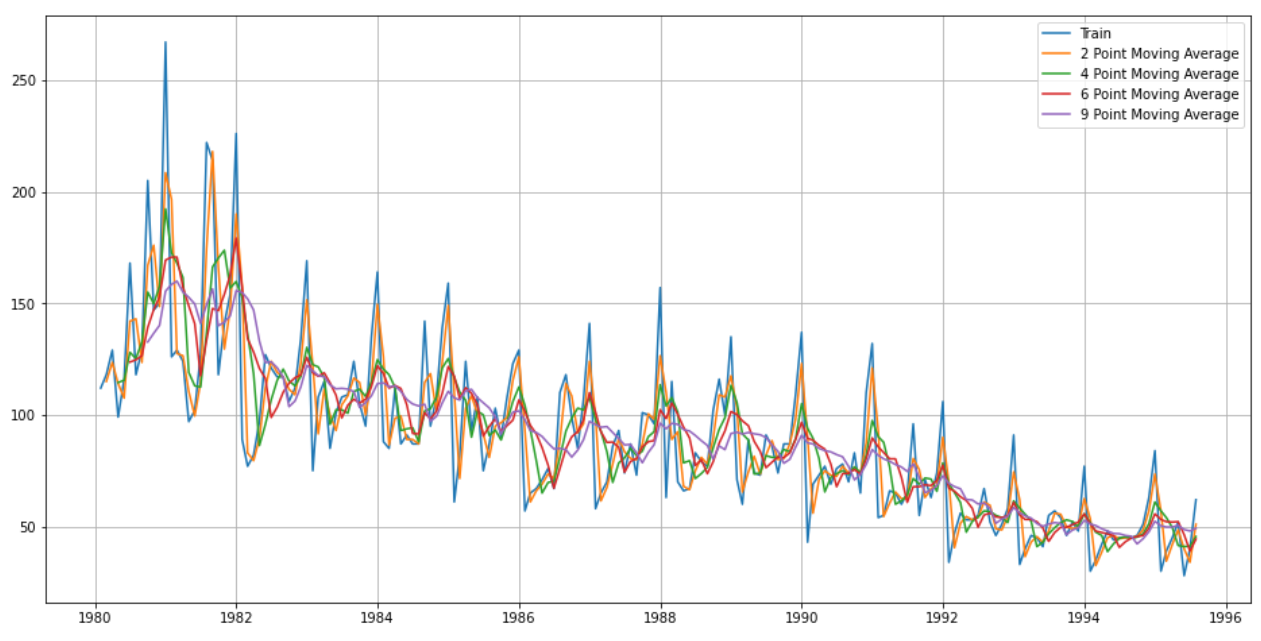
For Moving Average, we are going to average over the entire data.



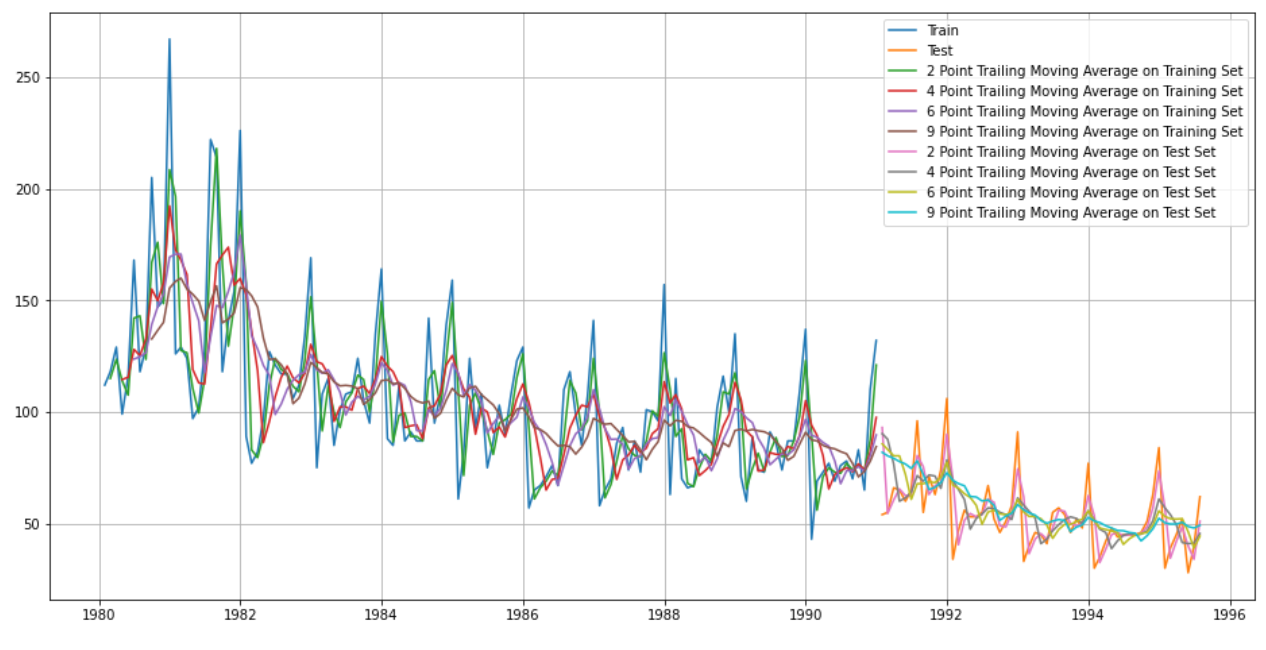
For Trailing Average,



**Plot for 2,4,6,9 Point Moving Average**



Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.



#### **Model Evaluation**

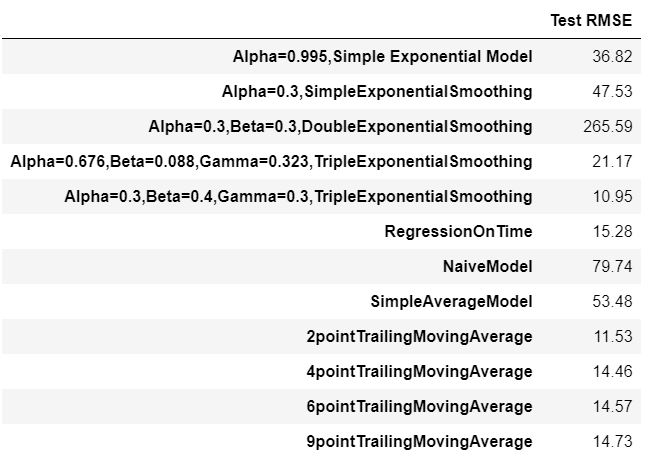
Done only on Test data

For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.529

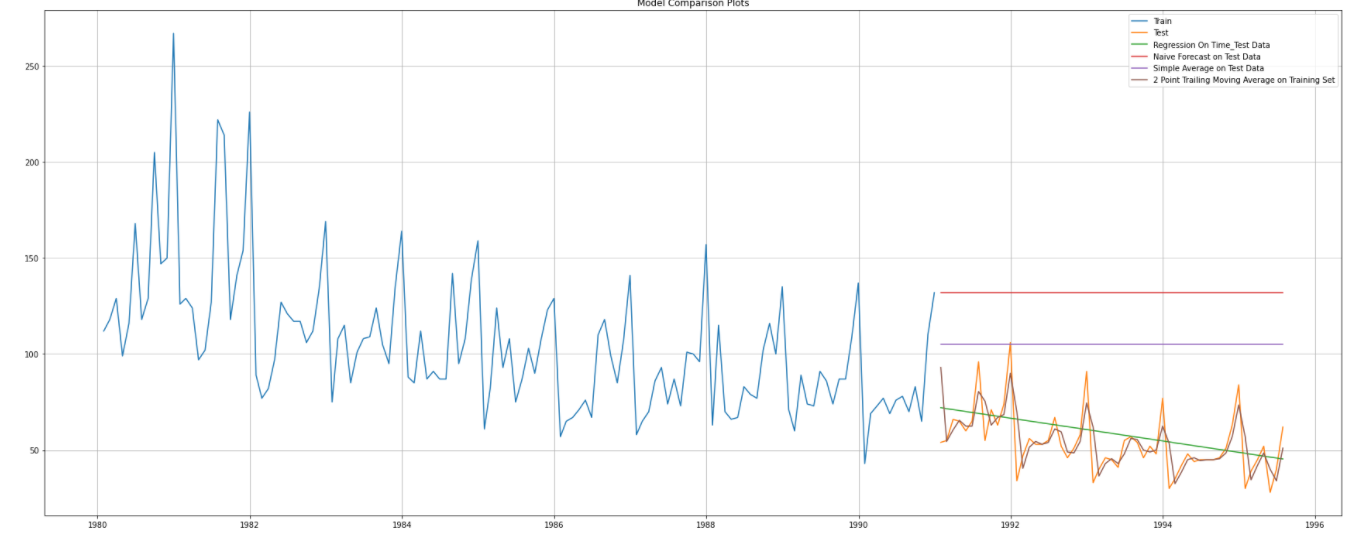
For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.455

For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.572

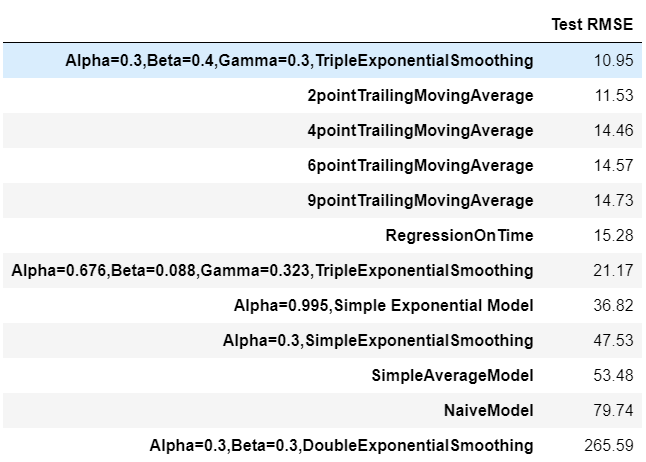
For 9 point Moving Average Model forecast on the Training Data, RMSE is 14.731



**Plotting on both Training and Test data**



**Sorted by RMSE values on the Test Data:**



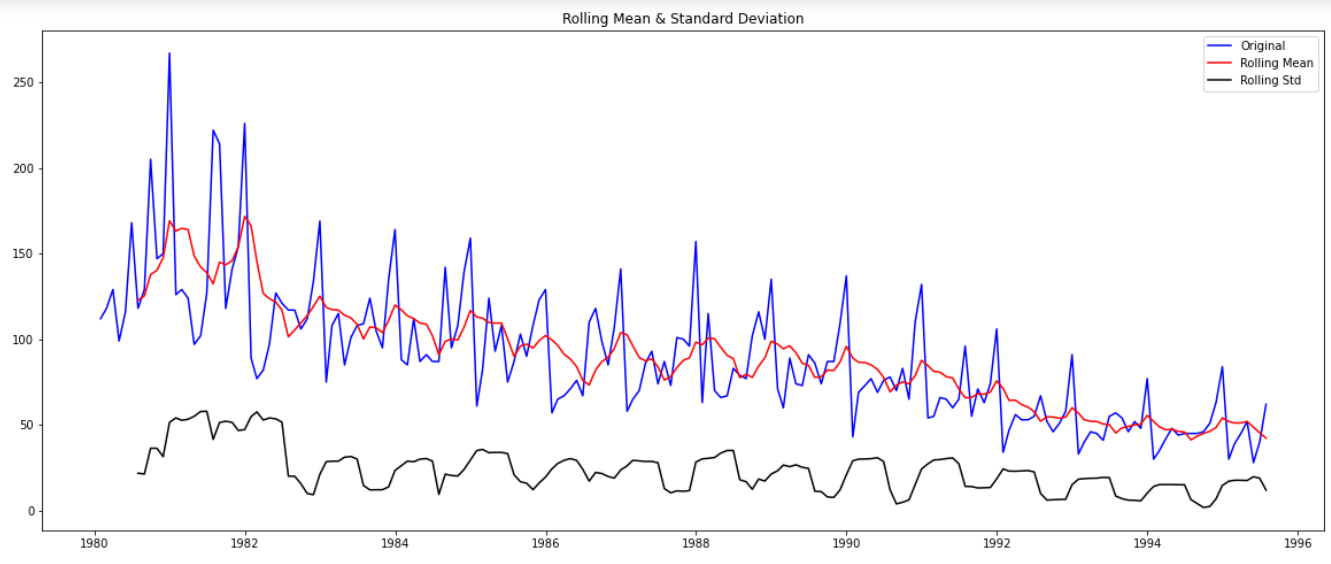
# 5.Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

#### **Check for Stationarity:**

Dicky Fuller Test

Null Hypothesis H0- Series is not Stationary

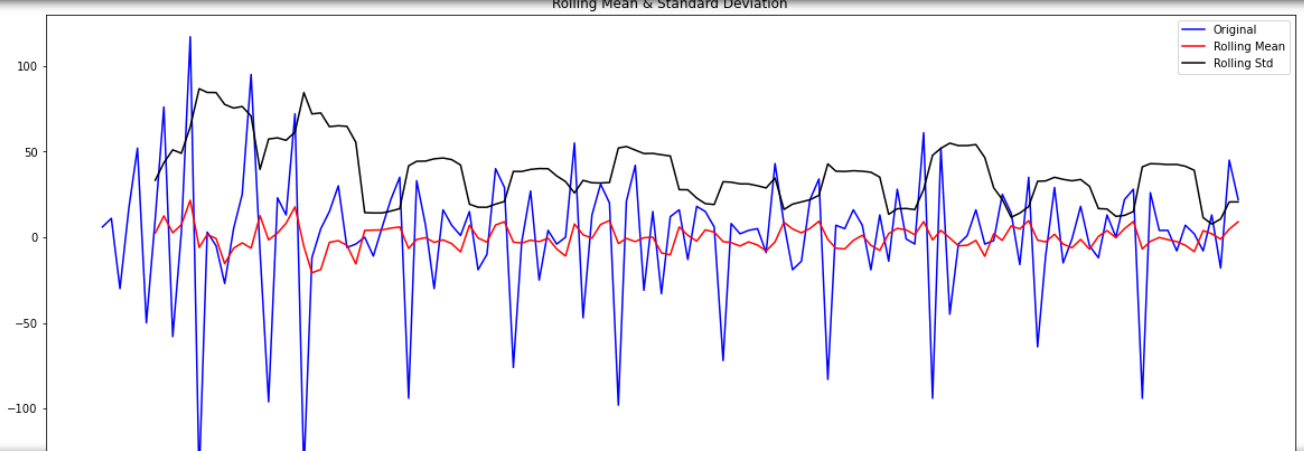
Alternative Hypothesis H1- Series is Stationary

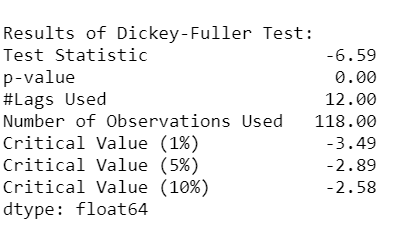


P value is higher that 0.05 Hence Null Hypothesis is True, we will take Order 1 differencing to make series stationary



We check stationarity at initial level, but series is not stationary as P value is higher than 0.05 difference of order 1.



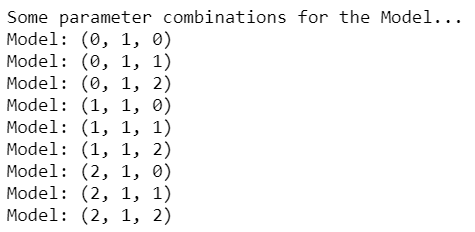


We see that at P value is less than 0.05 the Time Series is indeed stationary.

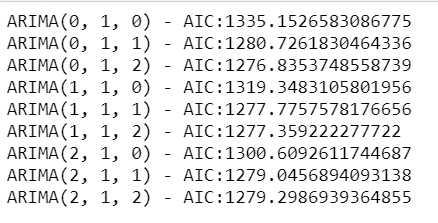
### 6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE. 8

#### **Automated Verion of ARIMA**

The following loop helps us in getting a combination of different parameters of p and q in the range of 0 and 2. We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

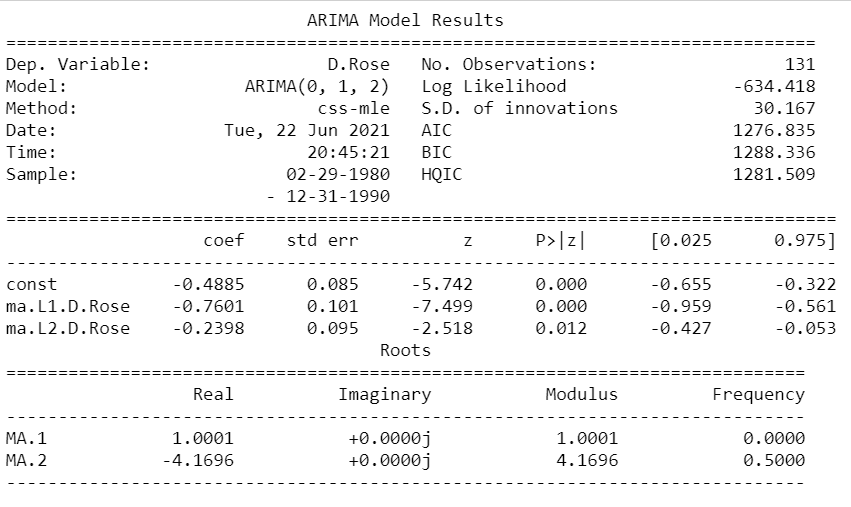


We will Apply all values of p and q. check which combination is giving us the low AIC score.

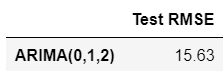


Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value

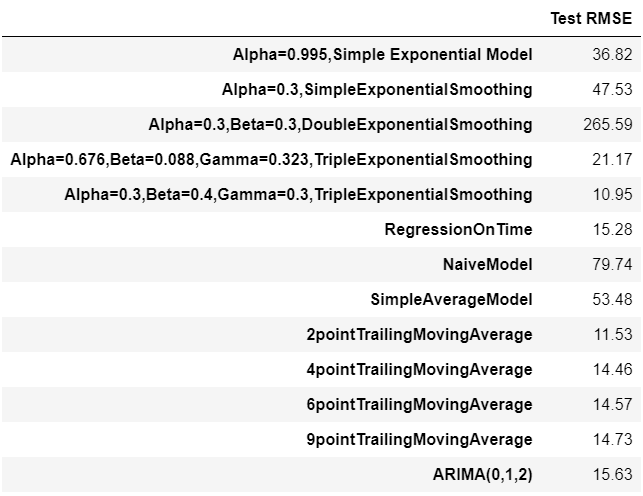




#### **Predict on the Test Set using this model and evaluate the model.**



Performance of Models so far:

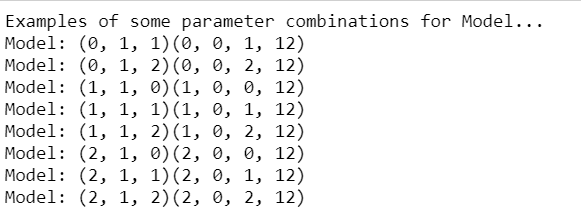


Still this point Triple exponential has performed the best.

#### **Automated Version of SARIMA**

In Model SARIMA, we are considering Seasonal P,D,Q,S.

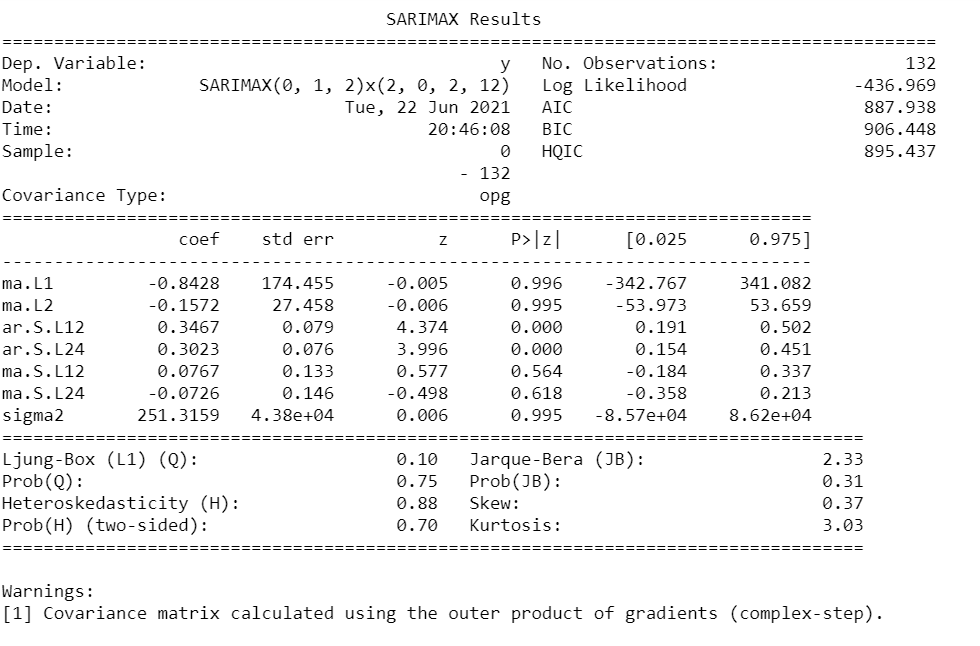
It is extension of ARIMA model by considering Seasonality factor.



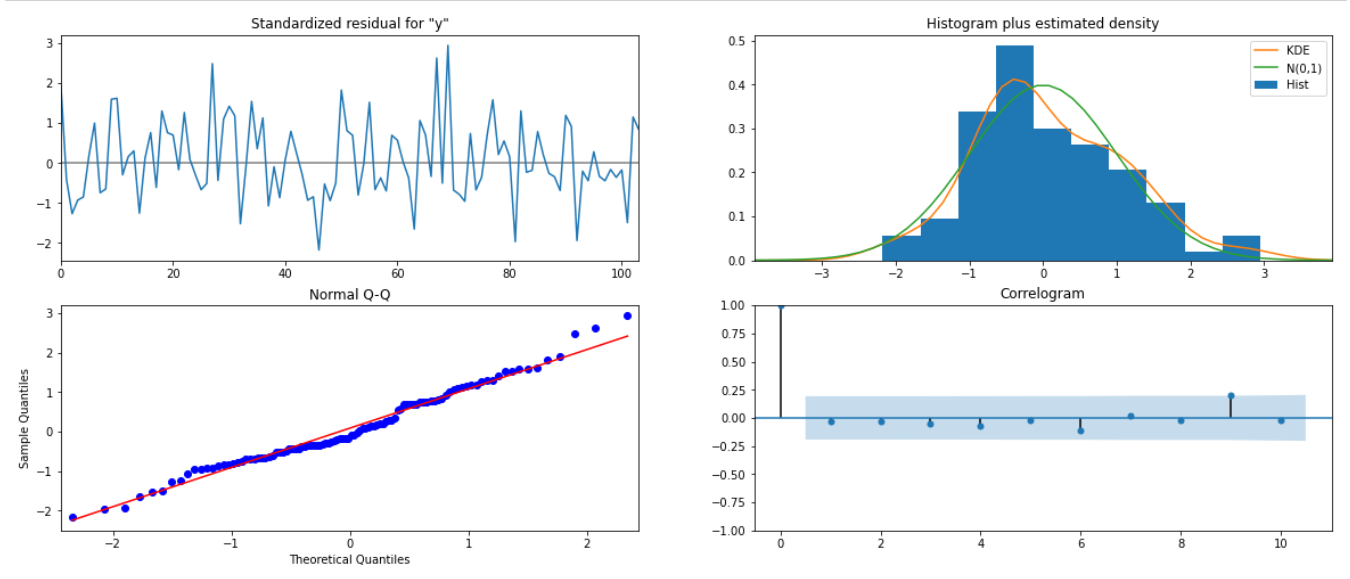
Top 5 Parameters AIC score:



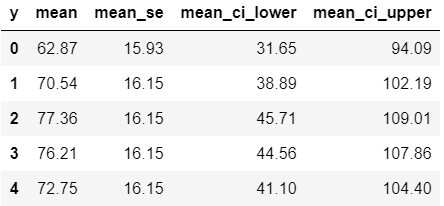
From the above result we will choose (0,1,2)(2,0,2,12).



Results:



#### **Predict on the Test Set using this model and evaluate the model.**



RMSE for SARIMA model is:

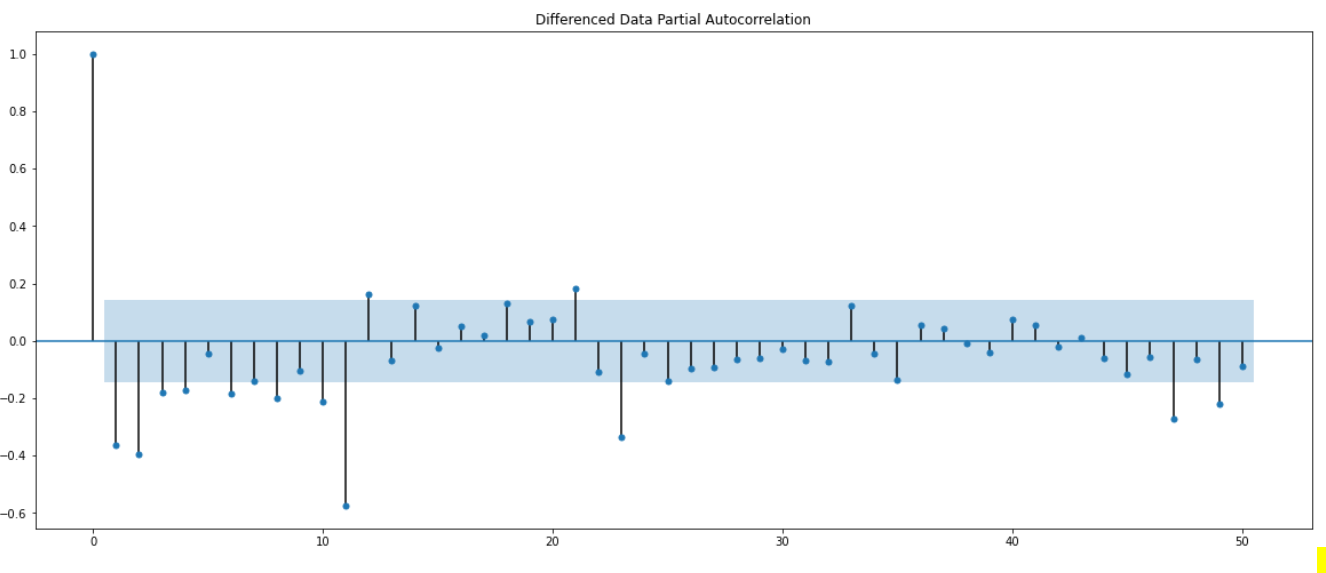


### 7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

#### **Manual ARIMA**

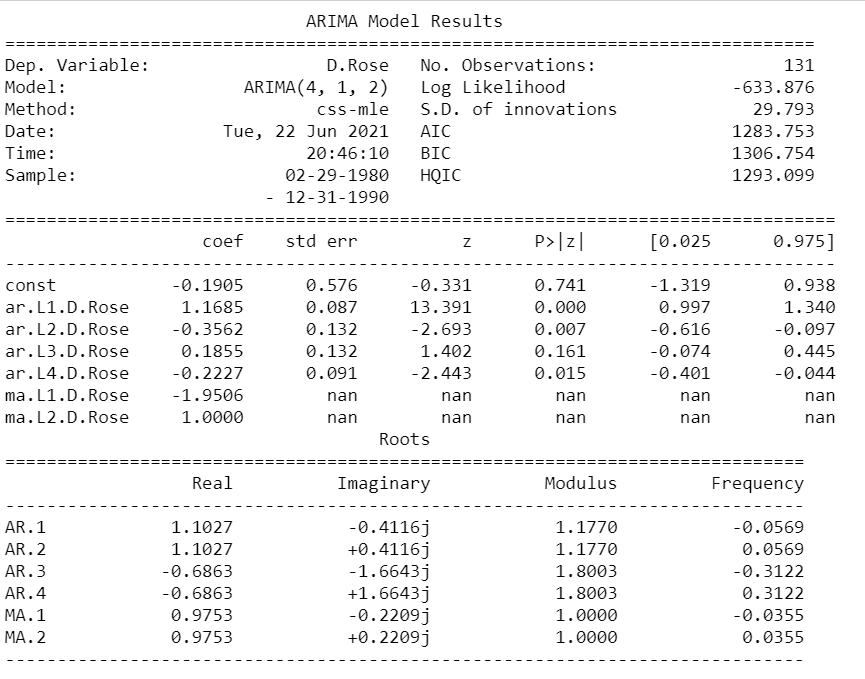
Let us look at the ACF and the PACF plots





Here, we have taken alpha=0.05.

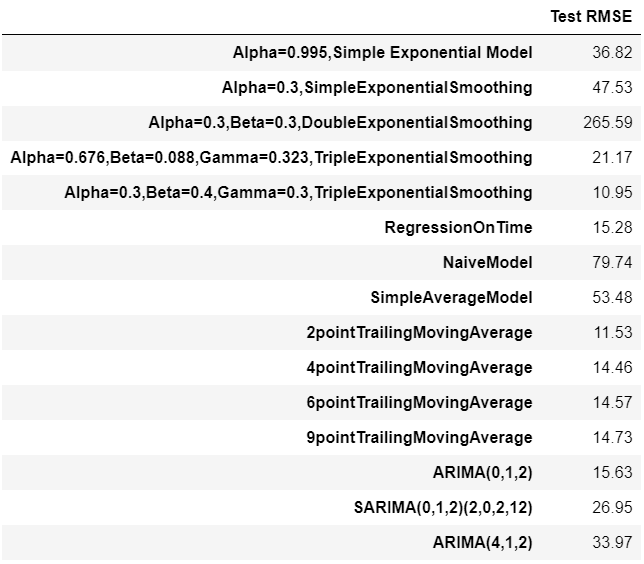
The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0. By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 4 and 2.



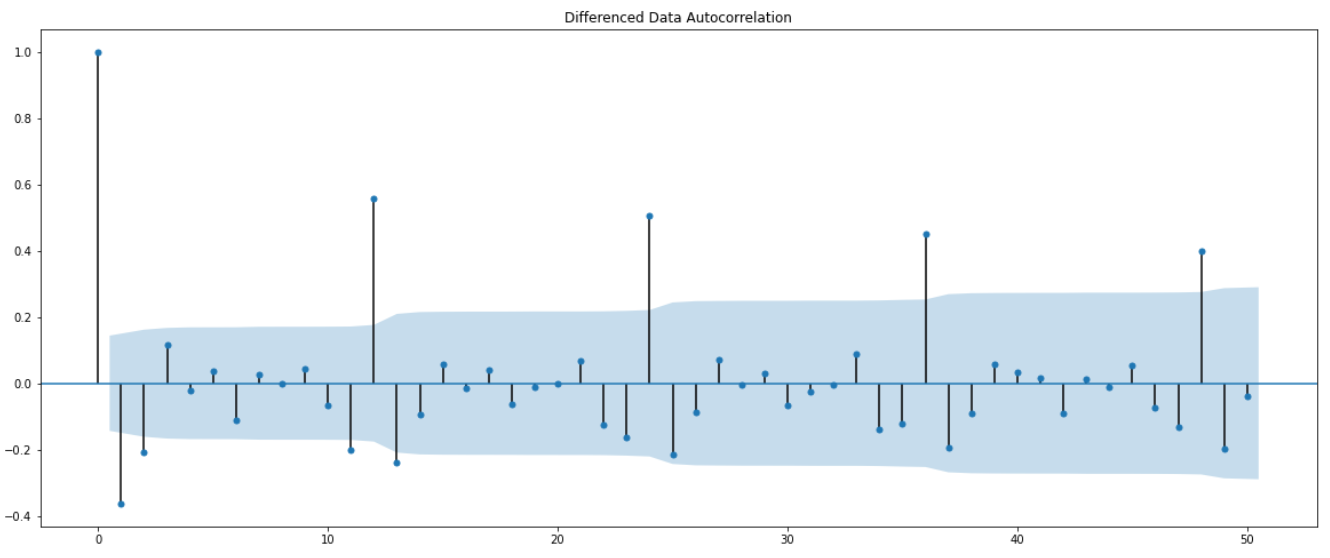
Predict on the Test Set using this model and evaluate the model.

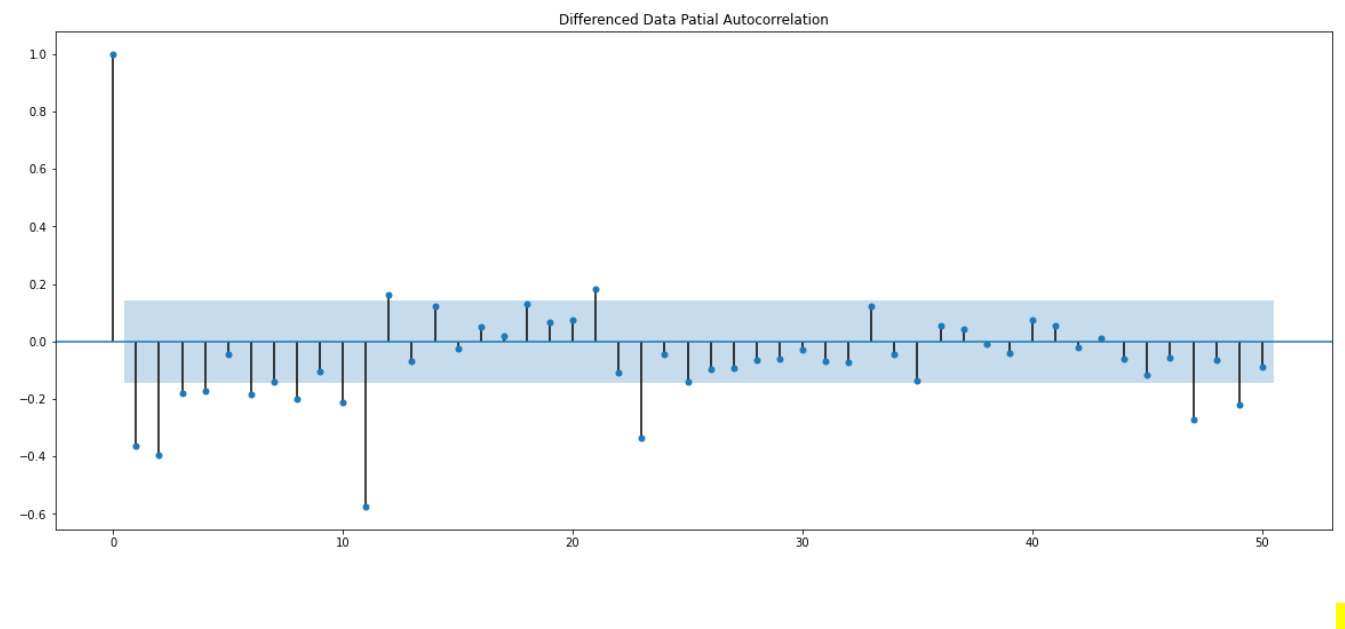
RMSE for Manual ARIMA model is 33.96920076790553.

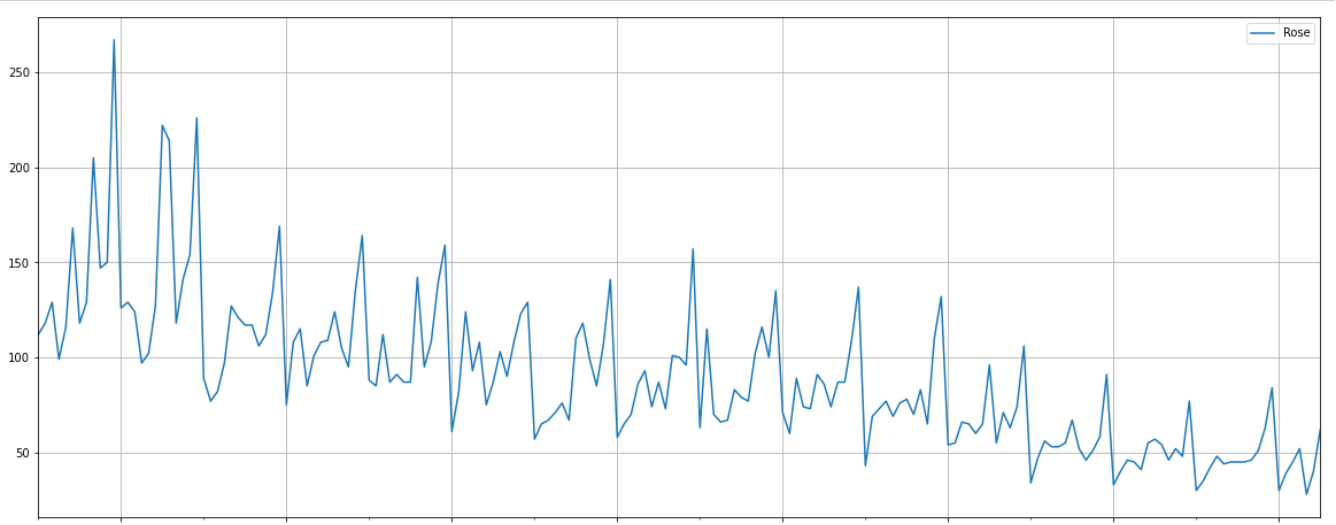
The Overall Results till now,



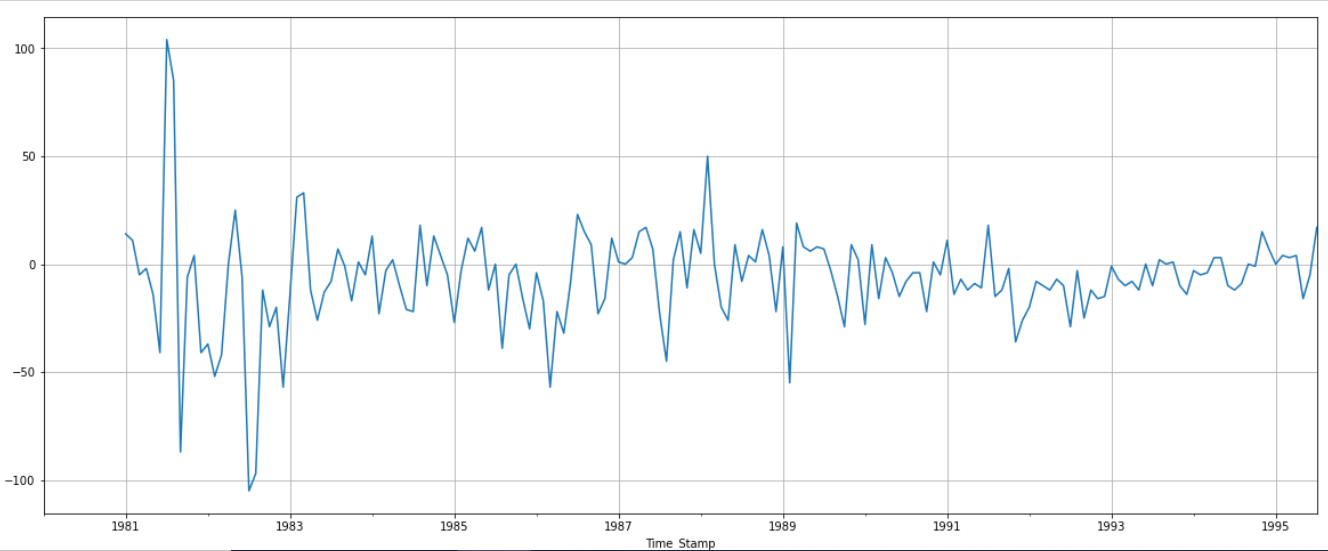
#### **SARIMA Model : Manually looking at ACF and PACF**



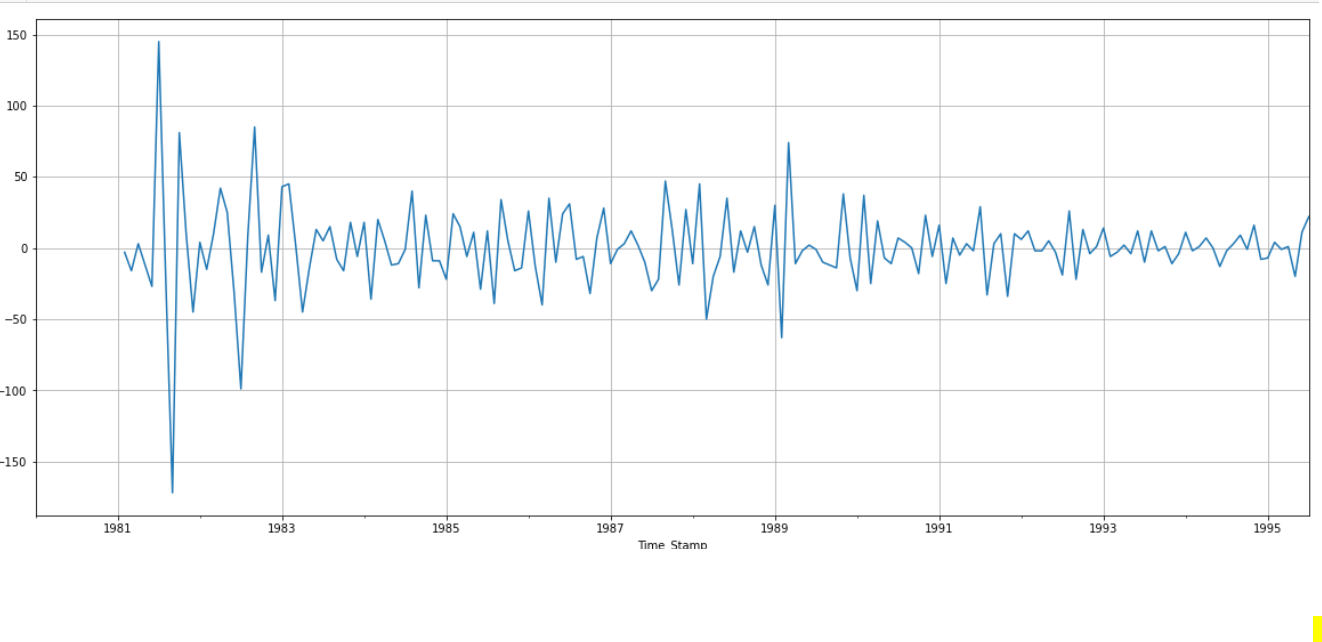




We see that there is a trend and a seasonality. So, now we take a seasonal differencing and check the series.

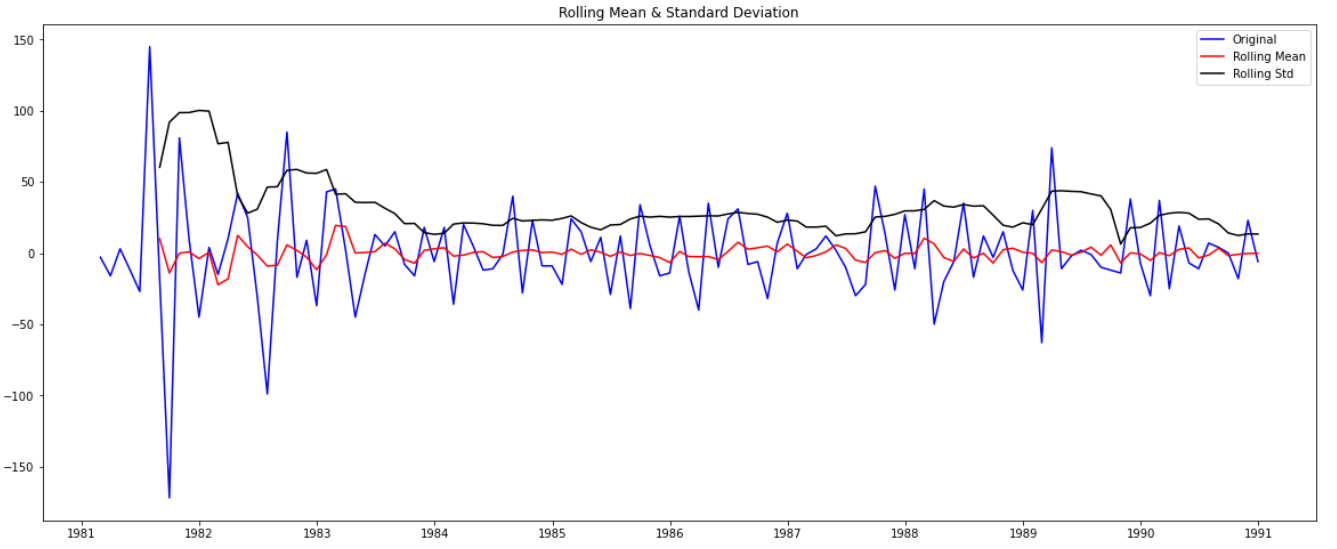


We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.

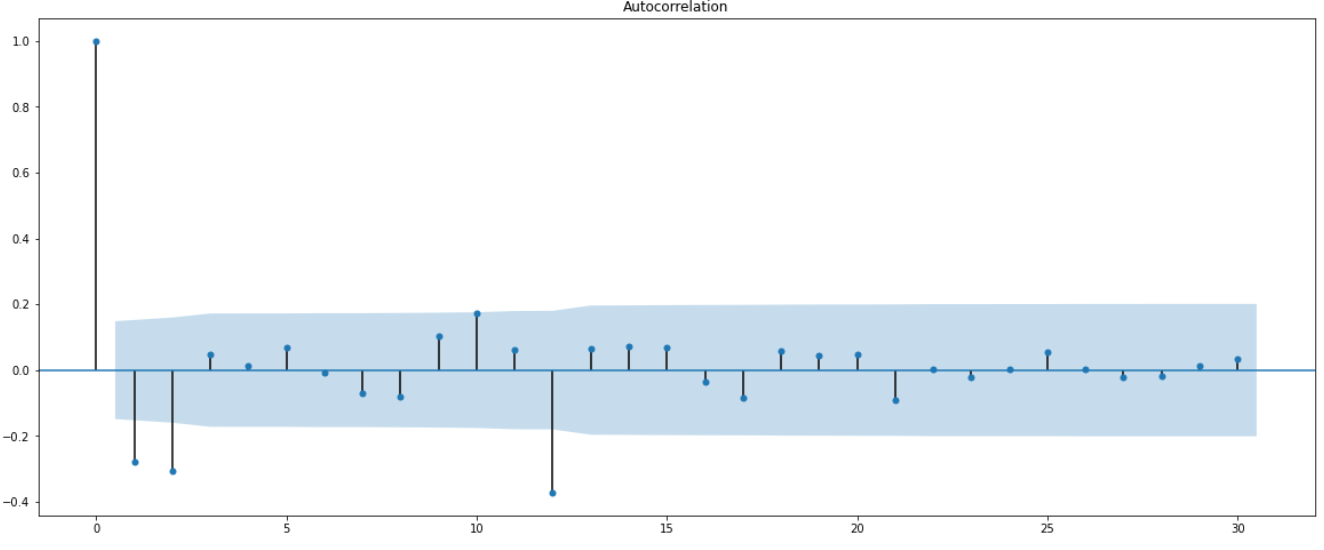


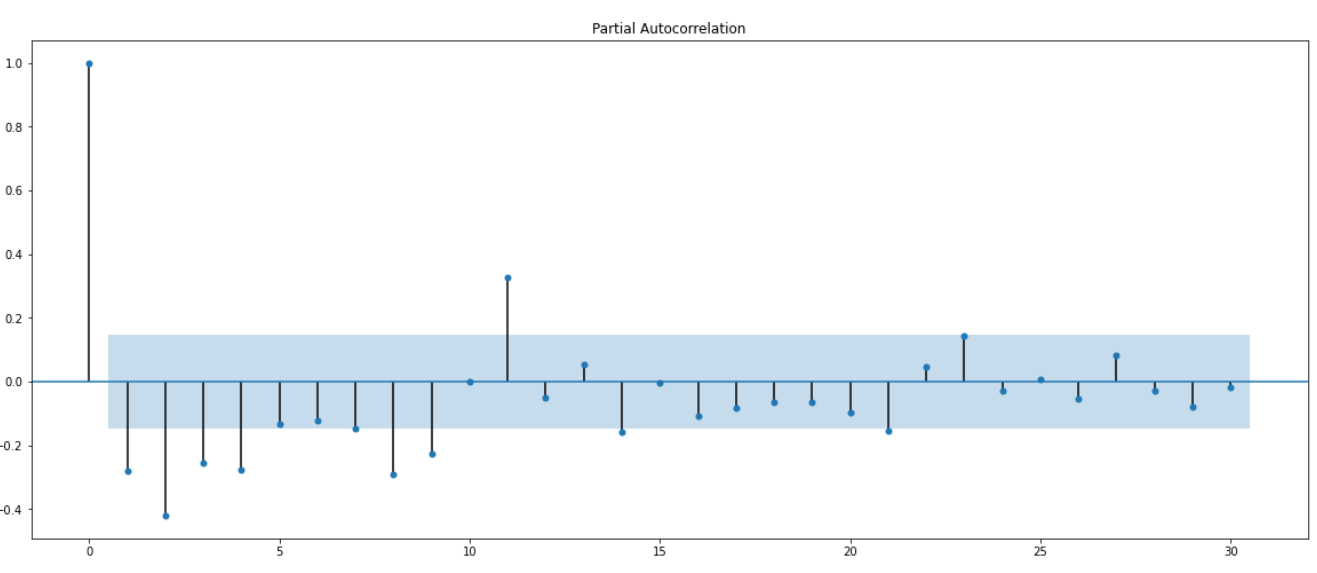
Now we see that there is almost no trend present in the data. Seasonality is only present in the data.

Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.







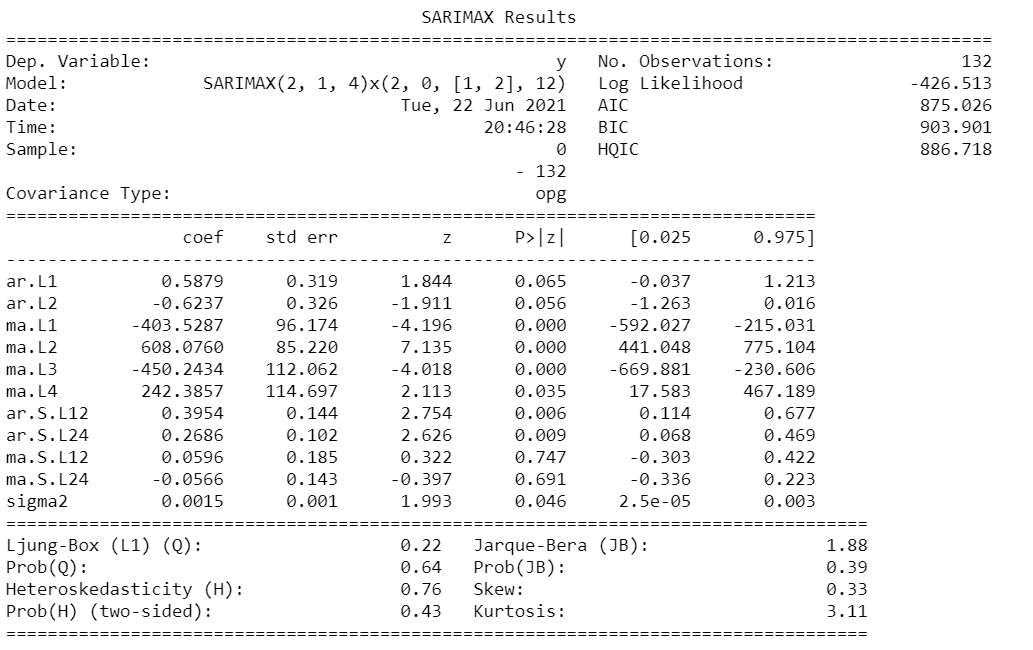


Here, we have taken alpha=0.05.

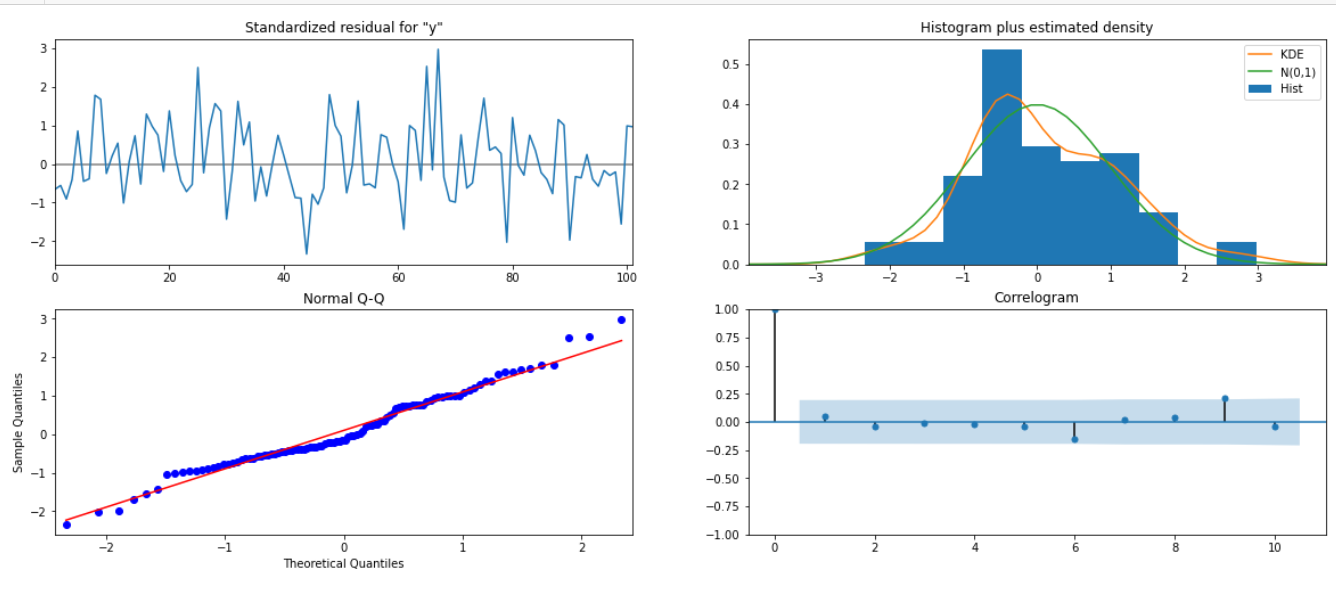
We are going to take the seasonal period as 12. We will keep the p(1) and q(1) parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period). By looking at the plots we see that the ACF and the PACF do not directly cut-off to 0.

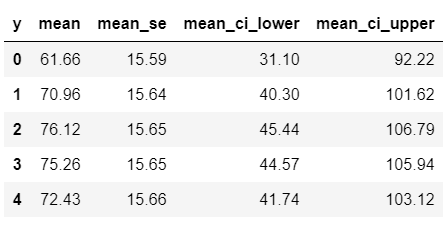
This is a common problem while building models by looking at the ACF and the PACF plots. But we are able to explain the model.



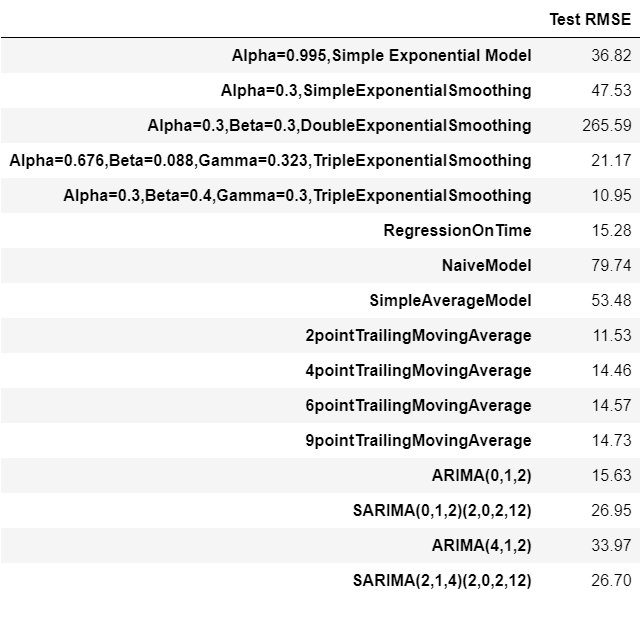
Results Dignostics for Manual SARIMA:



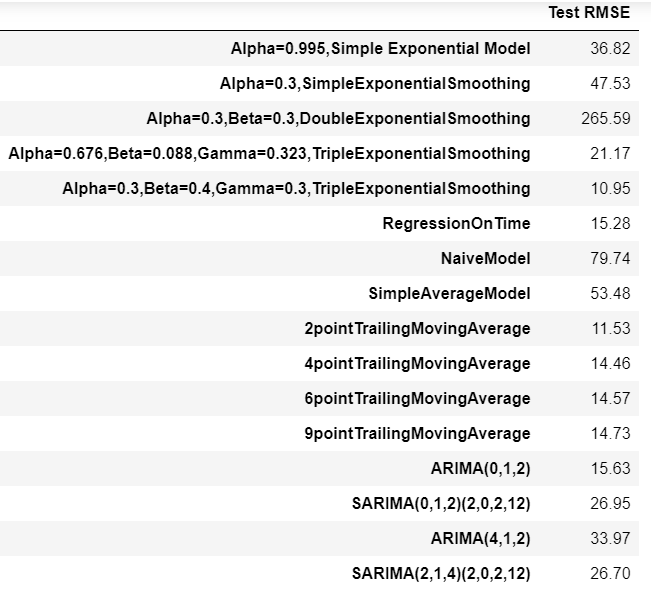
### **Predict on the Test Set using this model and evaluate the model.**



RMSE for Manual SARIMA is 26.696148760797275



### 8.Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.



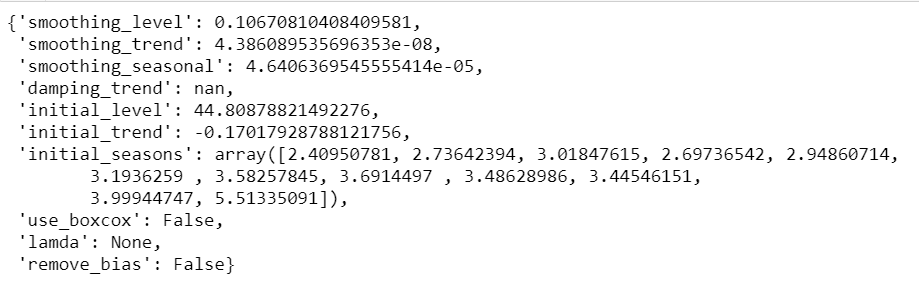
## 9.Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

If we see all the RMSE, We can Say that Triple exponential Smoothing can perform Well for this series, This series has level, Seasonality and Trend also.

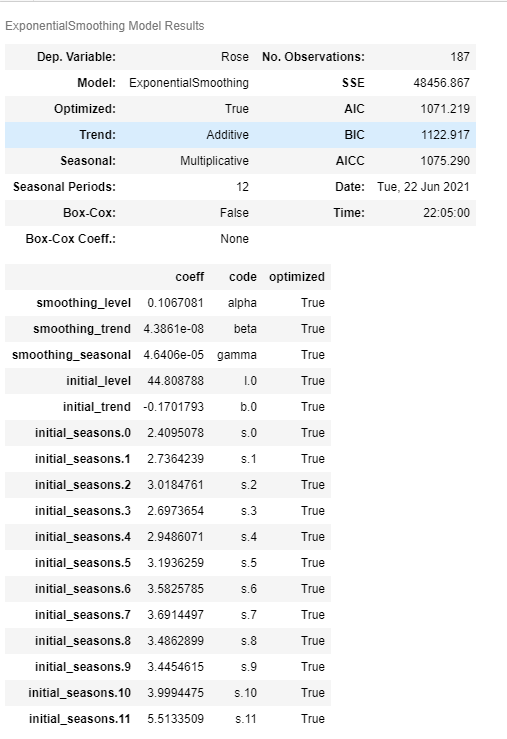
So we will Select Triple Exponential Smoothing model, we will give train data as Whole Data. That is, Test and Train data earlier, Now all data will be used for Training for Purpose.

We are predicting here for Next 12 Months.

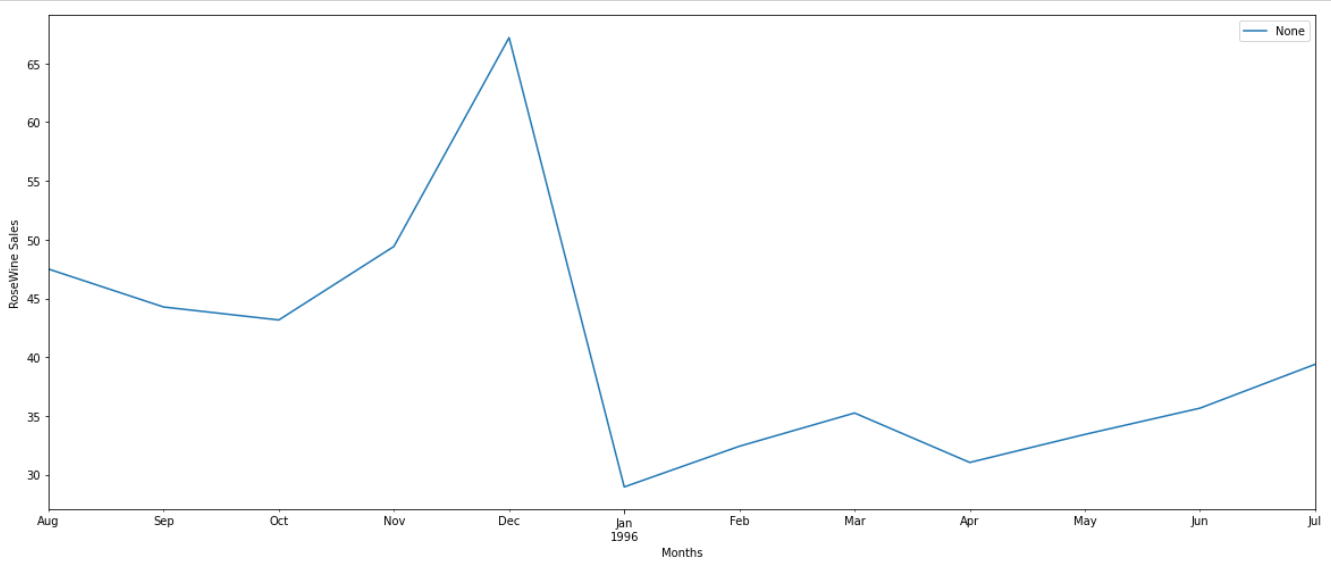
Parameters:



**Final Model results:**



Sales for the Next 12 Months,



Rose wine Sales and Particular Month :



## 10.Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present. 5

To predict Sales in 12 Months we have studied, Analysed and applied different Models on the existing Data.

While Doing the EDA part we got to know that

1.There is downward trend, and also series has Seasonality, Trend and Level.

2.There are two missing values, we have filled that with the forward filling method.

3.The sales are High in the month of December,it is might due to Festivals like New year, Christmas.

4.The High Stock should be Ready from the month of August to December.

5.As per Predicted, The Maximum sales are happening in the month of December, and it is around 67, so it will be okay if we got 67 wines in stock in The Dec-1995.

6.Here we come with interseting insights from Monthly plot, From January to October Median of Sales is almost Same. Only Sales are increasing in November and December.In This December Only Sales has Crossed figure of 250 wines.

7.It needs to be analysed further, Why the Sales are Decreasing every passing year. Quality issue, branding, Advertisement issue that can be analysed from different data.

To Predict the Results, We have applied Double,Triple ExponentialSmoothing.Also applied Regression, NaiveModel, SimpleAverageModel, TrailingMovingAverage, Then automated ARIMA, SARIMA by comparing AIC values and Manual ARIMA, SARIMA by Partial Autocorrelation Function and Auto-correlation Function plots.

After Applying all the Models, we found that RMSE value for Triple exponential smoothing was a lowest, so we have decided to go with this model. We have applied the whole Data earlier it was Test and Train, we have applied full data, fit the model and then predicted the Next 12 Months Sales and plotted the same.

THE END