

**Sparkling 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:  [Sparkling.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.

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

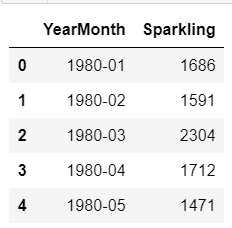
First of all we will import all the necessary libraries like Pandas, Numpy, seaborn, os to set the path and Statsmodels as base library for all the models.

After setting the path we will import Sparkling dataset

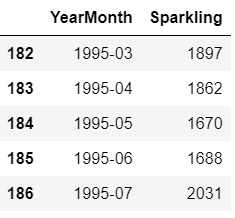
We will read the dataset First by using pandas library.

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.



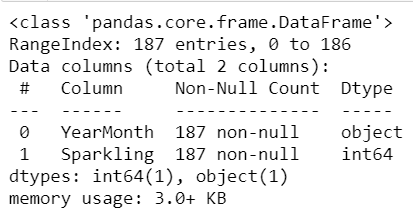
No null values are present in the dataset.

**Shape of the Data:**

(187, 2)

So Dataset has 187 rows and 2 Columns.

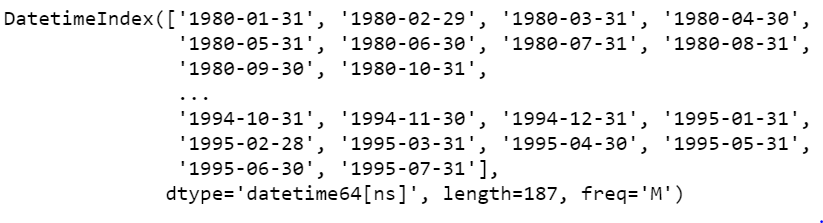
**Data Info:**



We can see that YearMonth column is of Object Data type while Sparkling is having int data type.

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.

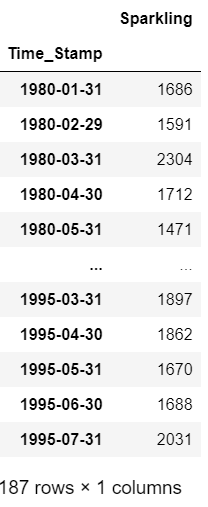


Total 187 entries needs to be added to Original Dataframe while we need to remove YearMonth Column.

And need to make This new Column as Index.

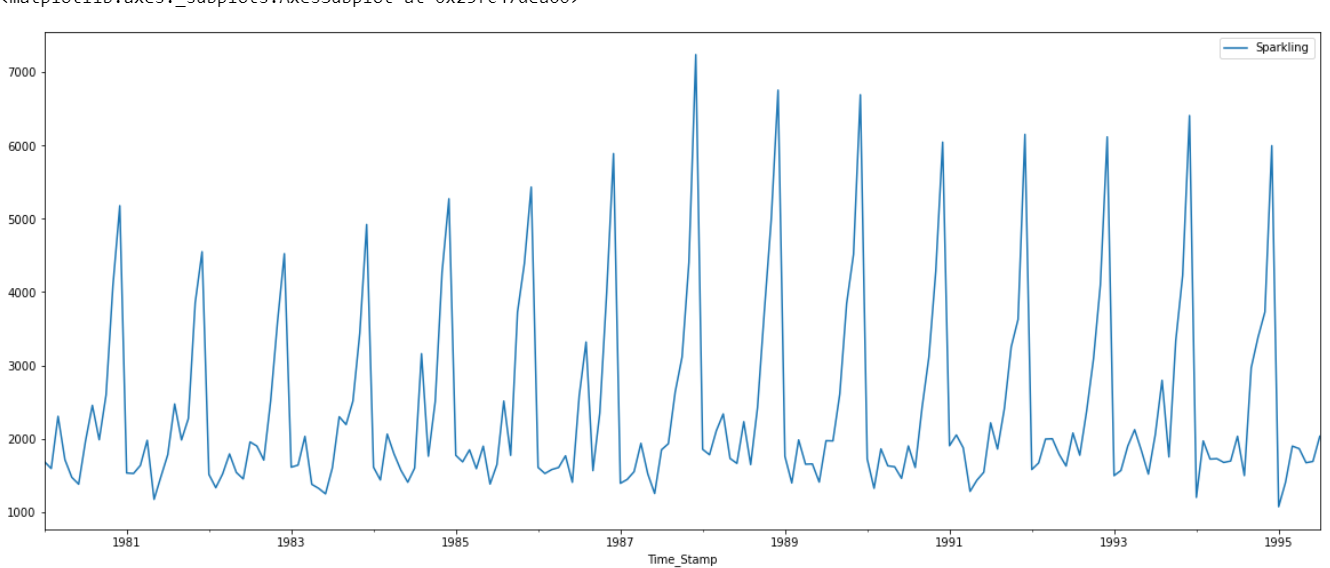


We can see the timestamp is incorporated in dataset, we will make this as index. We only need the column Sparkling Sales. Other Time\_Stamp will be index column.



We will plot the dataset of Sparkling wine sales.

Lets plot the time series



By looking and Above plot we must say there is seasonality present.

In above plot we can see some spikes at the end of every year, but we are little bit Confused about trend, like from 1982 to 1987 there is Some positive trend, but from there onwards downward trend.Lets do further EDA to get more insights.

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

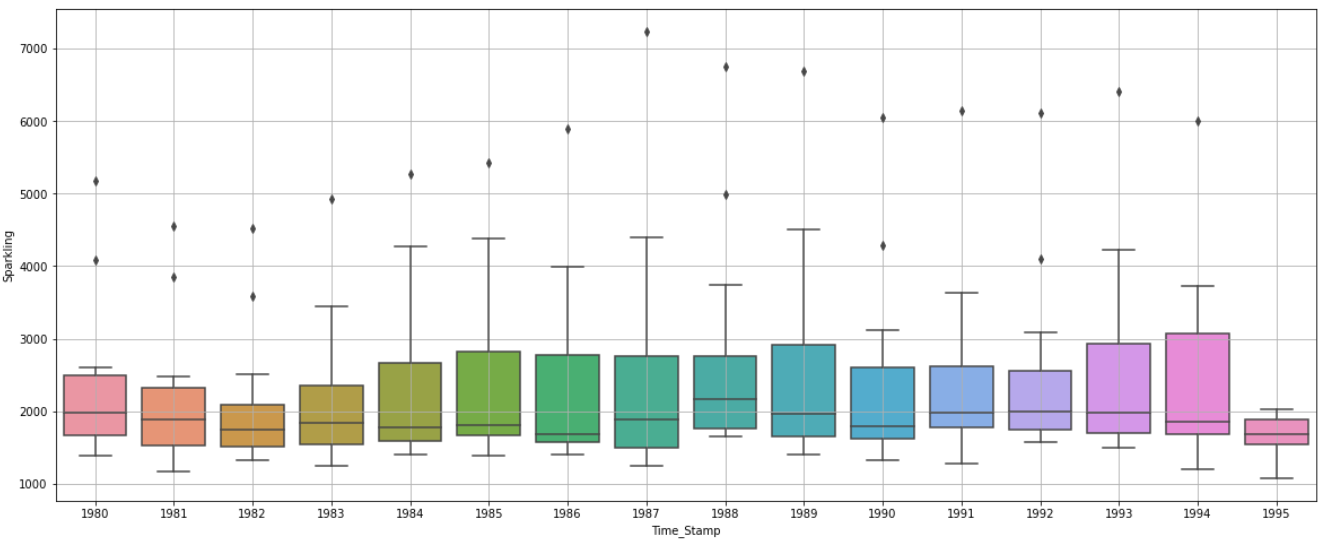
**Description:**



from above describe Function we can see that Highest Sales of Sparkling wine is 7242.Mean and Median are having much difference.

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

**Yearly plot:**

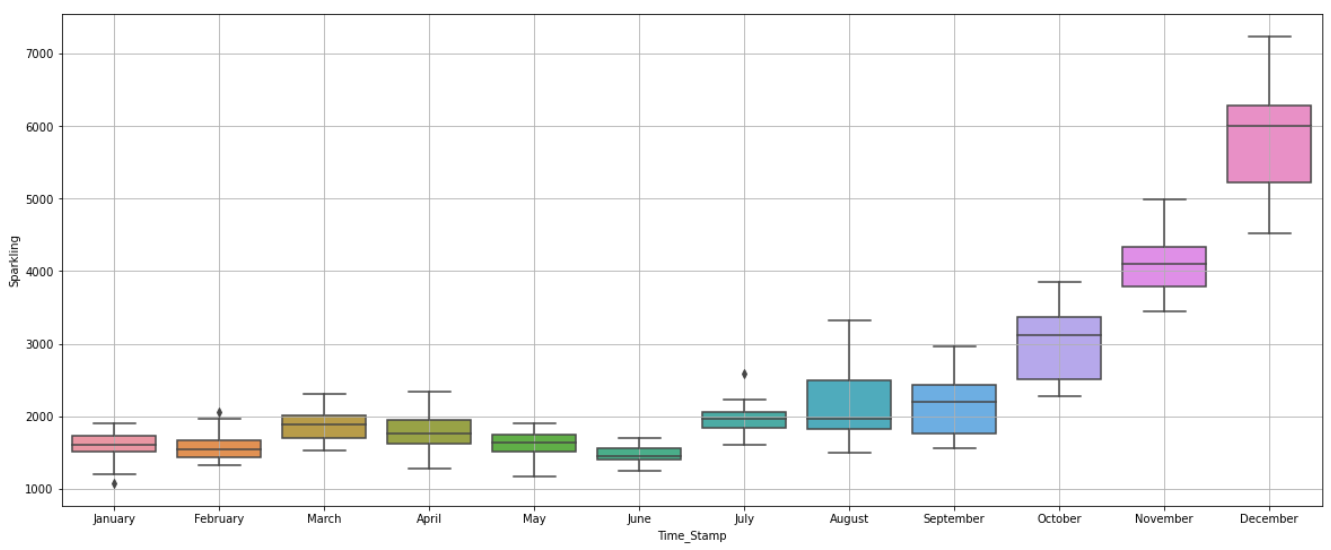


We can see from above graph the median value for all years is remains around 2000. It means the sales of Sparkling Wine is Almost same over the years, that is not Positive or Negative trend.

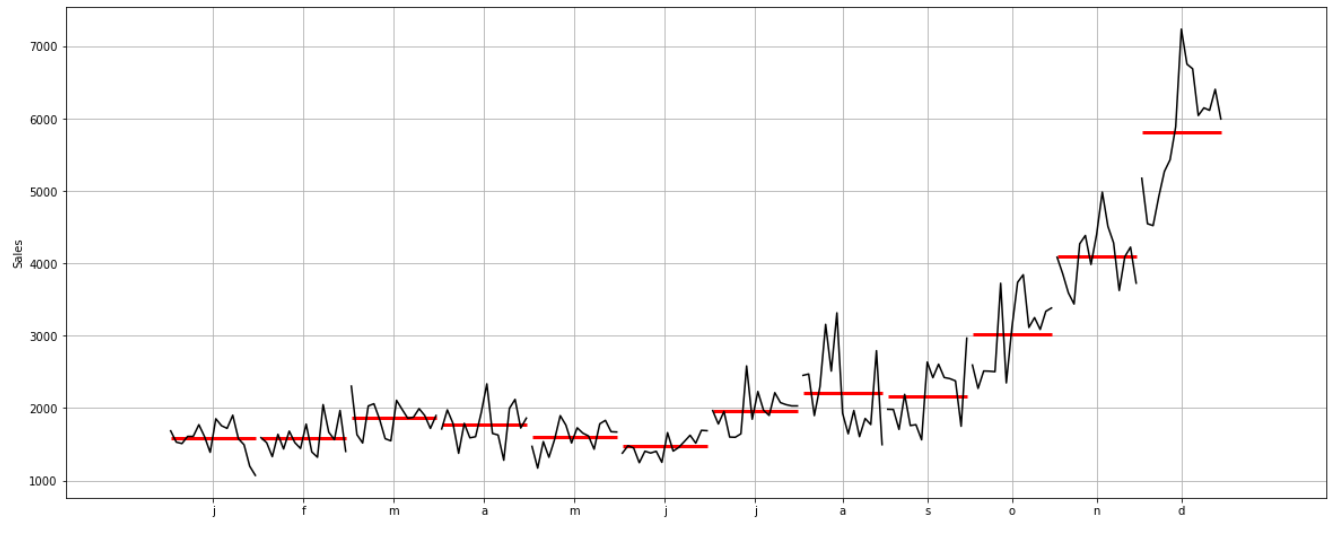
Highest Sales of Sparkling wine is occurred in year 1987, and in 1995 one of the month has brought Lowest sale of wines.

Lets check for Monthly plot

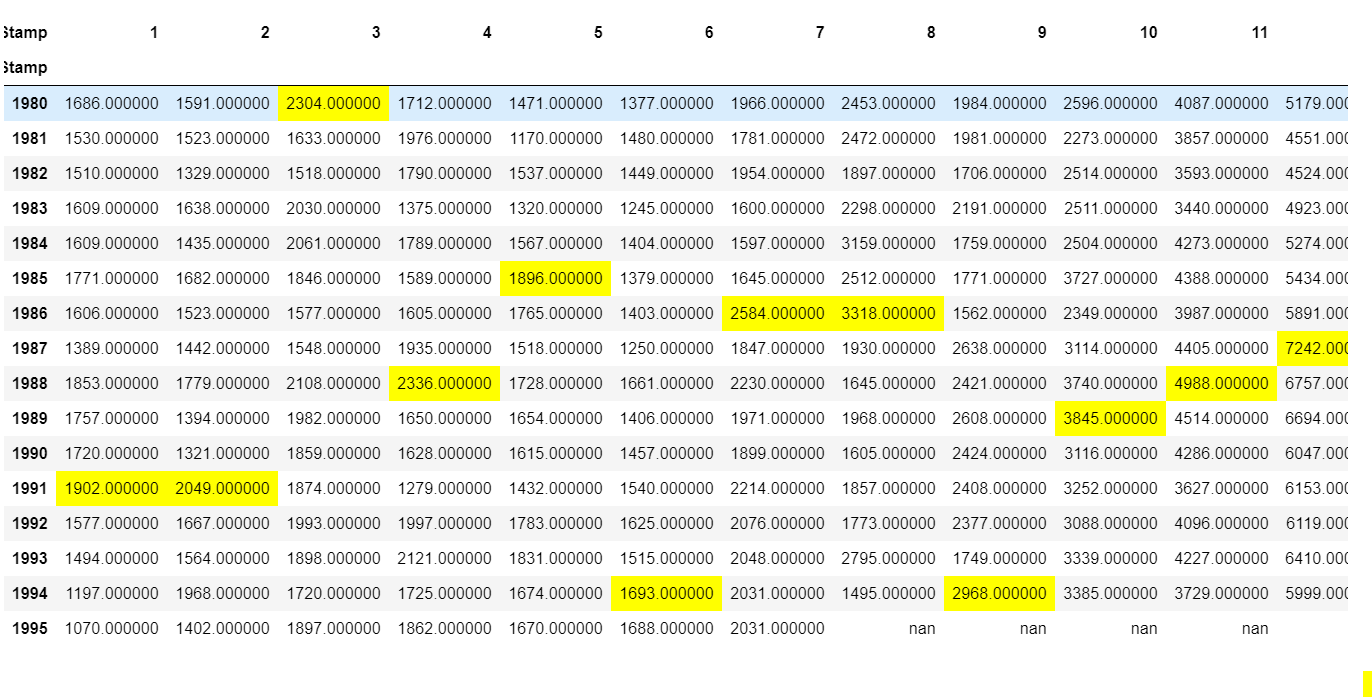
**Monthly plot:**



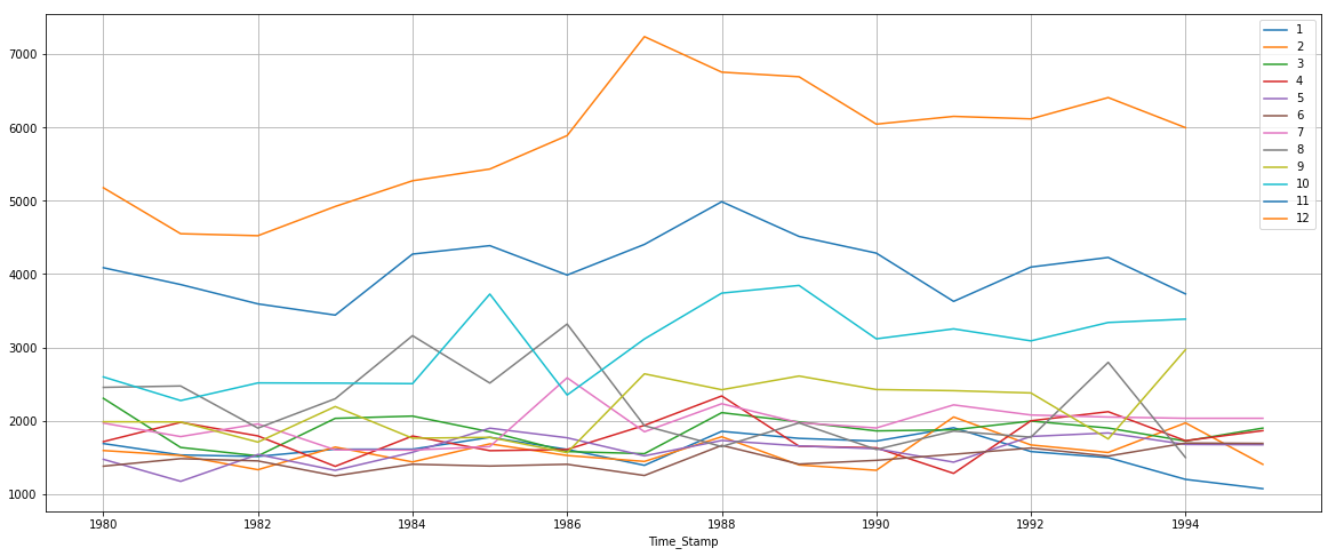
Yes finally we get Something notable, In above plot Sales has been Increasd from August to December. Till this point Sales are around 2000. so We need Higher stock or Production from August to December.October, November , December has sales more than 2000 and it is almost 3000,4000,6000.



**Highlight The maximum sales:**

We can see above, The maximum Sales of Month in a All years.

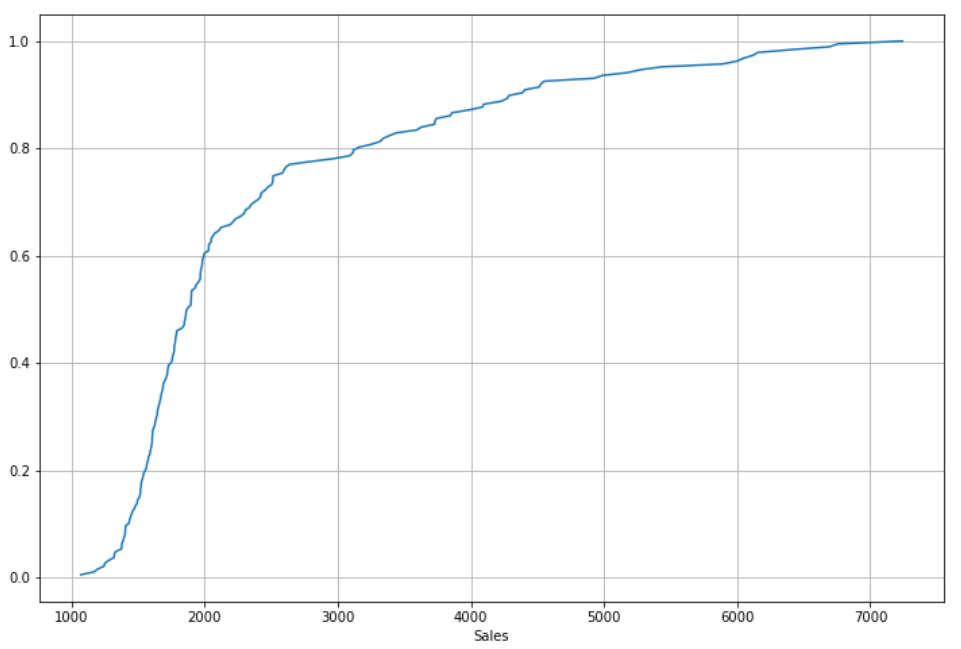
We can see above, The maximum Sales of Month in a all years.



From 9th month sales are going High thats what we can see in above plot, This is true for all the mentioned time series. In The Year the 1987 December the Sales are Highest it can be seen.

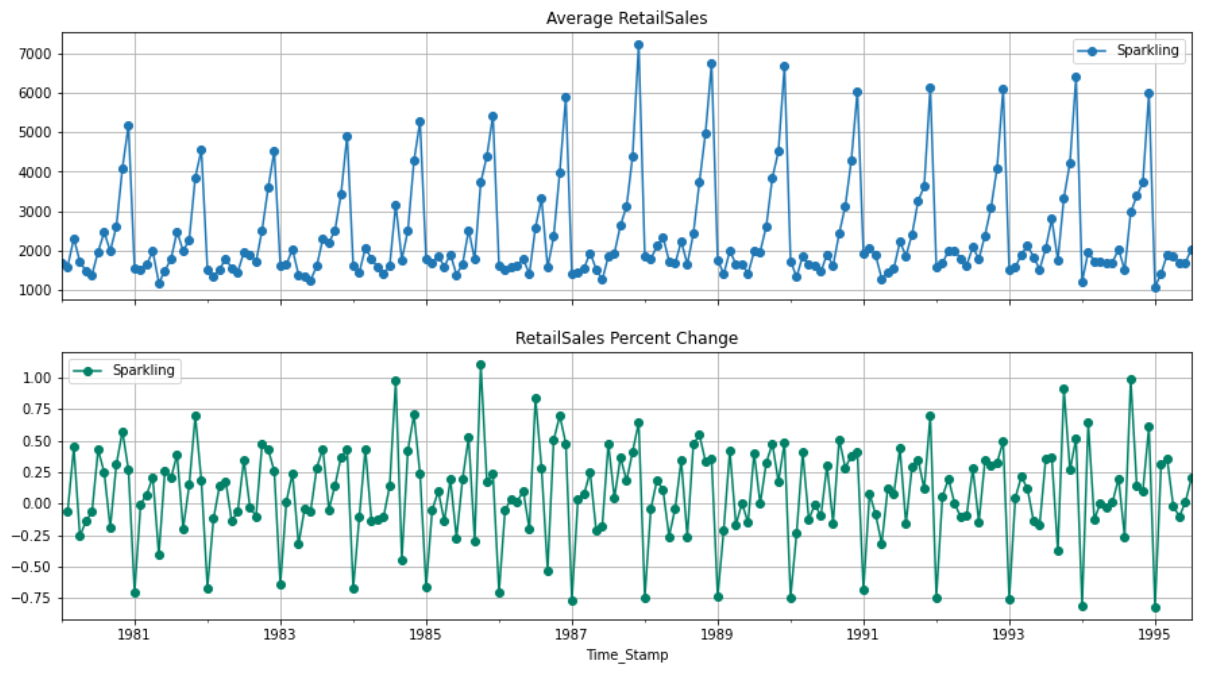
The orange line is for December month, It is Alienated from Other lines.

**Empirical Cumulative Distribution.**



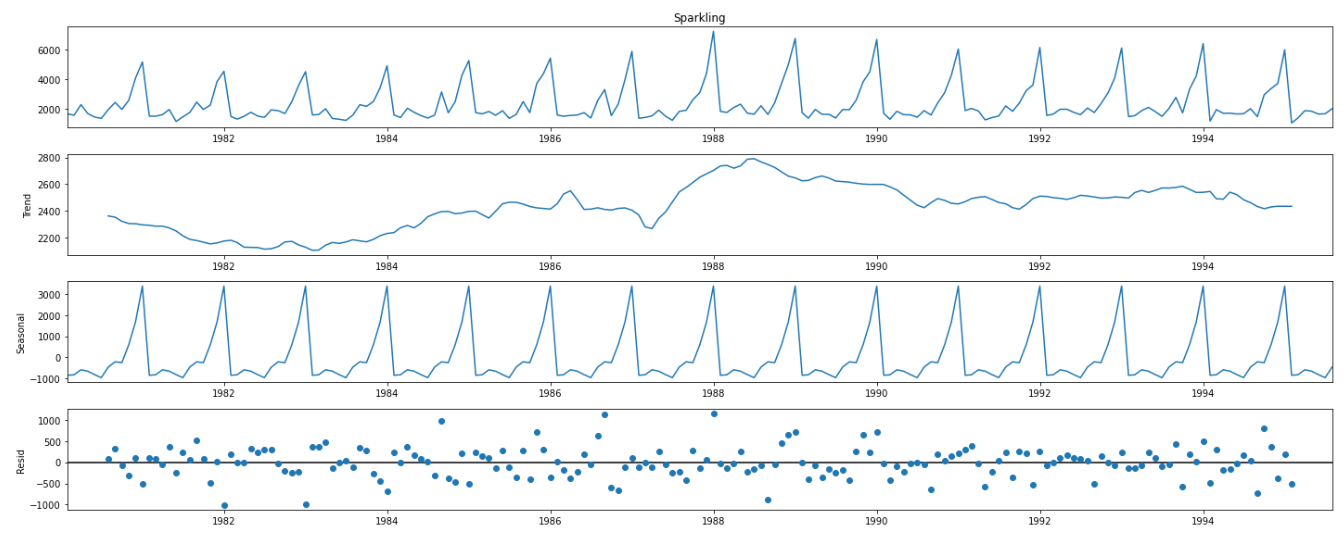
This particular graph tells us what percentage of data points refer to what number of Sales. 60% of the sales are below 2000. Maximum sales is greater than 7000. Only 20% of whole data tells us that sales are more than 3000-6000..

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



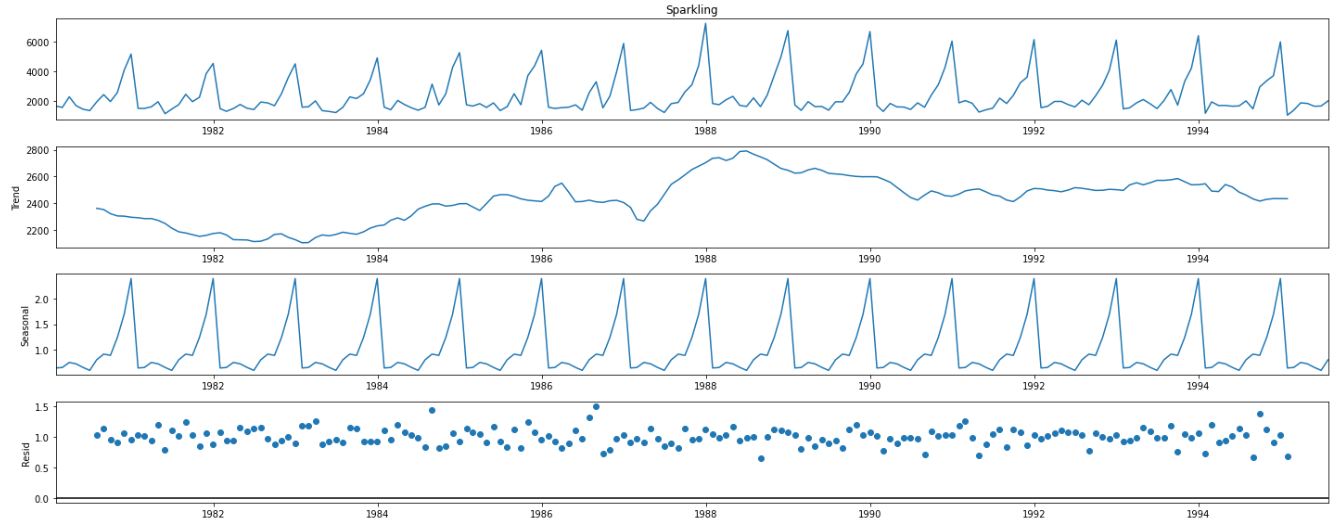
### **Decompose the Time Series and plot the different components.**

**Additive Decomposition:**



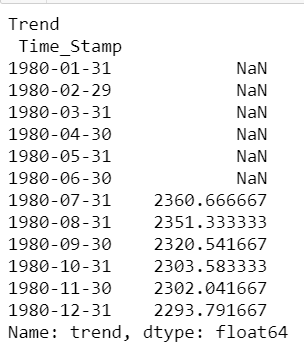
We were not Clear about the trend and Seasonality, but here one thing we can notice There is strong Seasonality, Positive Negative trends. One more thing to Notice in Additive Decomposition the Residuals are showing patterns, it is not permissible actually not a good choice, but before saying anything lets move to Multiplicative Decomposition.

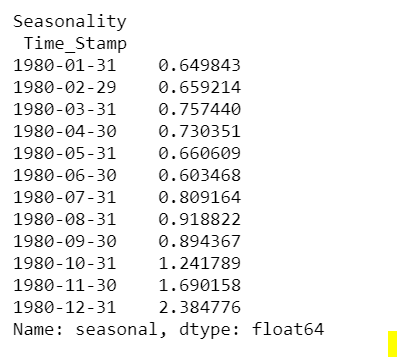
**Multiplicative Decomposition:**

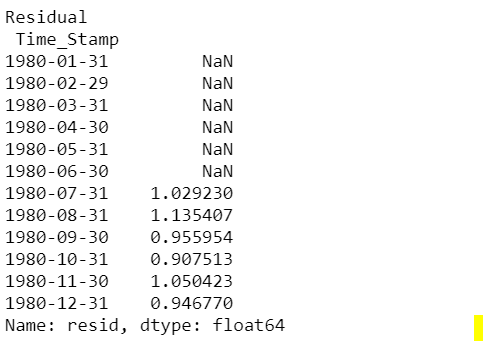


For the multiplicative decomposition series, we see that a lot of residuals are located around 1.

Almost Same Trend And Seasonality, Also residual is showing patterns so we will choose Additive decomposition as its alwyas easy to go with Additive Model, why to choose complicate models? when both are giving almost same results.





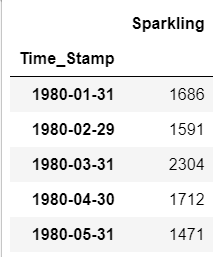


# 3. Split the data into training and test. The test data should start in 1991.

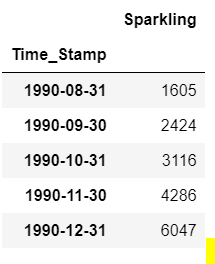
Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

Train Data:

First Five rows:

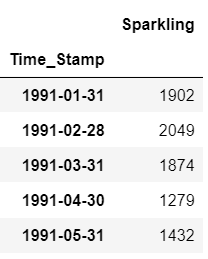


Last Five rows:

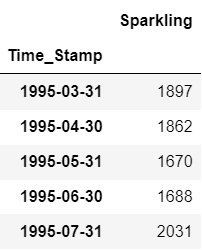


Test Data:

First 5 rows:

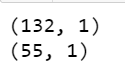


Test Data Last 5 rows:



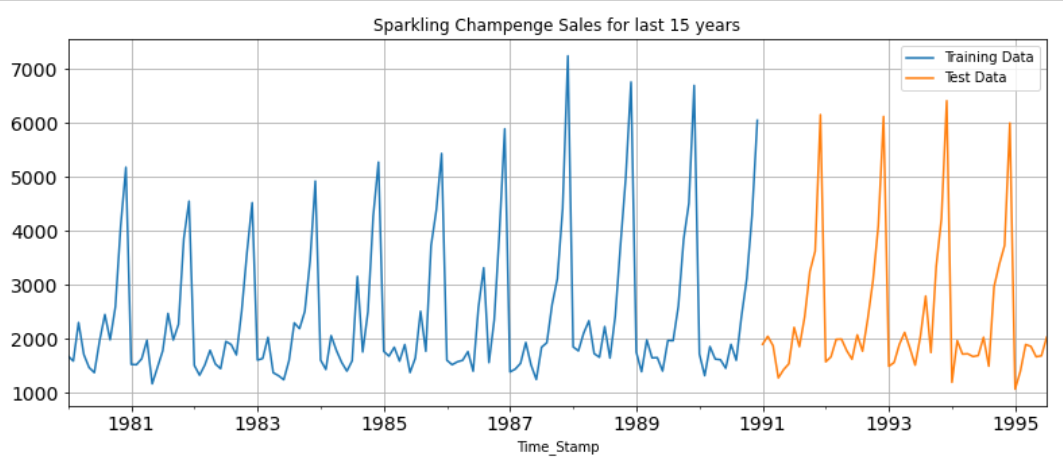
Shape of Train Data: 132 Rows and 1 Column

Shape of Test Data : 55 Rows and 1 Column



test data starts at Jan-1991 and It is till July-1995

**Train and Test data plot:**



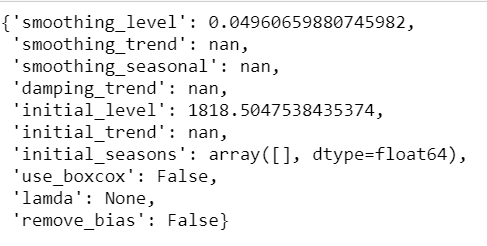
# 4.Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

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

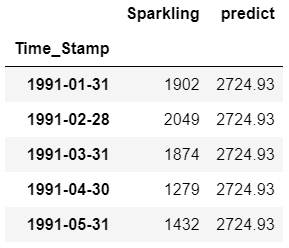
We have created the test and train data for the upcoming Models.

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

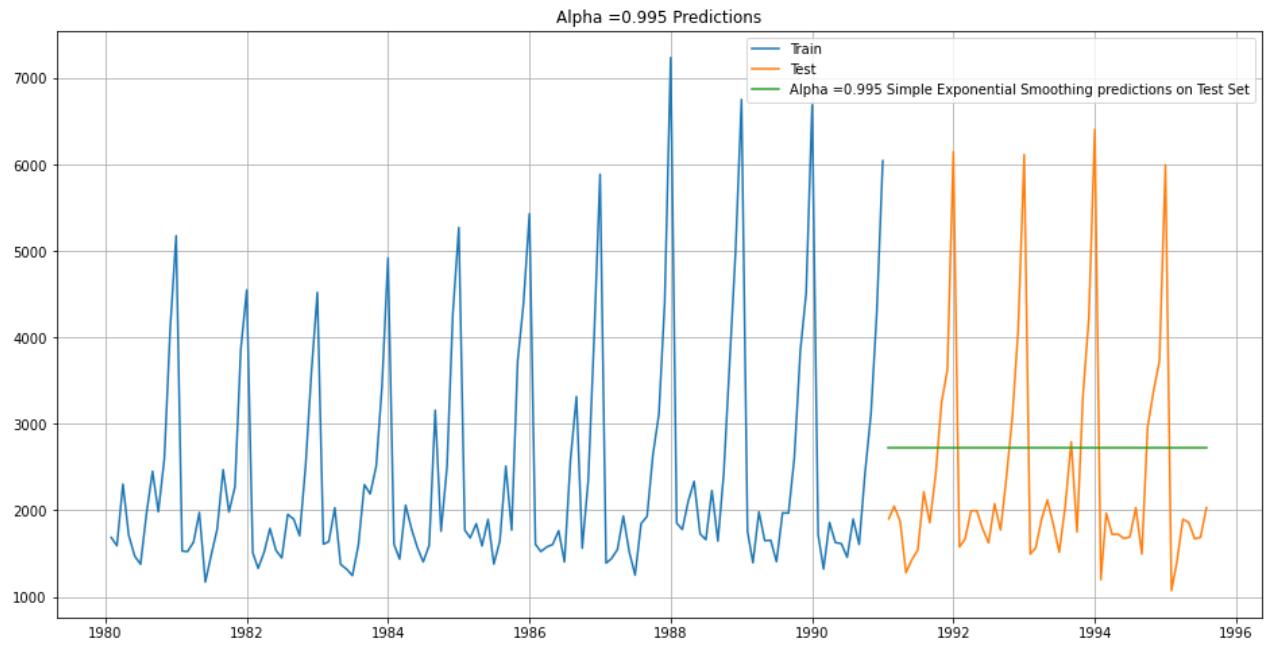
The parameters of Model Simple Exponential Smoothing



Now lets predict on the test data



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 1316.035

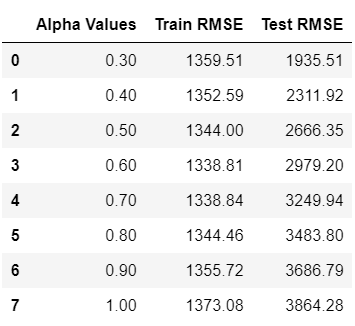


Setting different alpha values. the higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

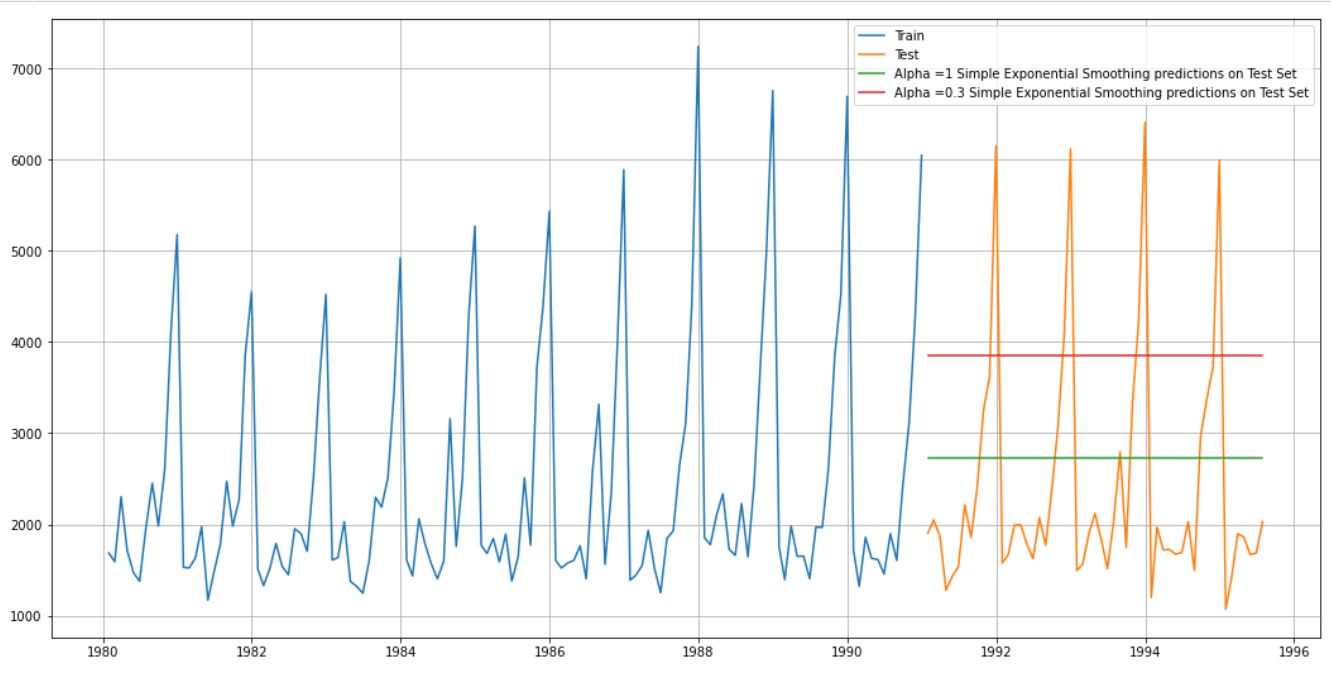
First we will define an empty dataframe to store our values from the loop



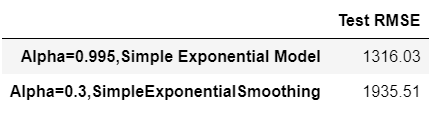
##### ***Model Evaluation:***



Plotting on both the Training and Test data



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

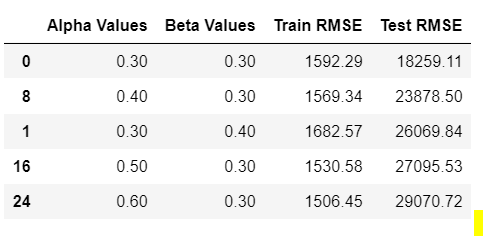


#### **Model 2 : Double Exponential Smoothing (Holt's Model)**

Two parameters 𝛼 and 𝛽 are estimated in this model. Level and Trend are accounted for in this model.

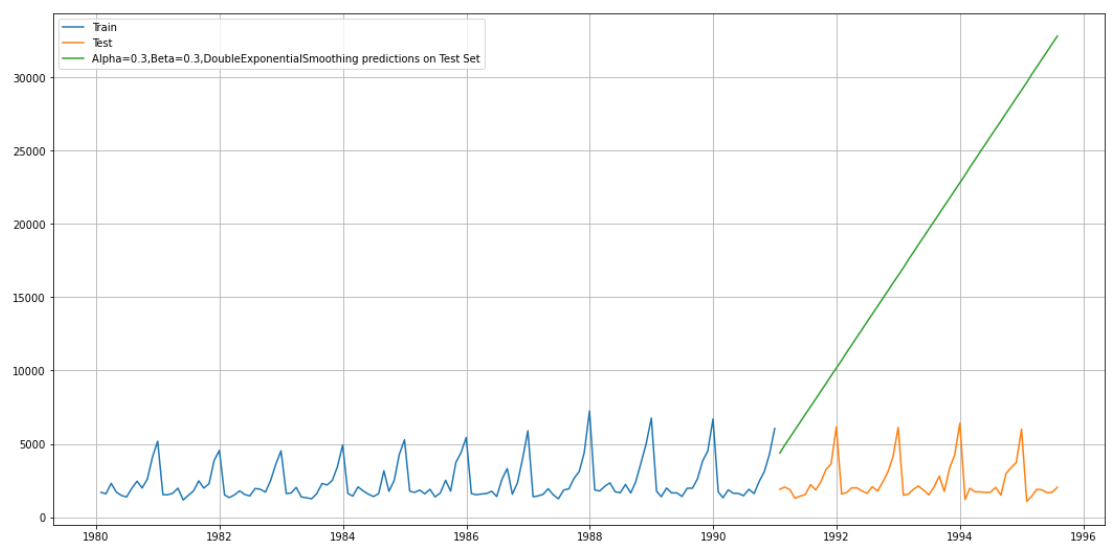
First We will create a model and then we will predict on test data.

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.

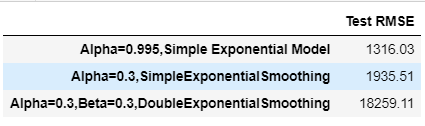


Test RMSE is lowest for Alpha=0.3 and Beta =0.3. hence we will select these values for better performance.

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)**

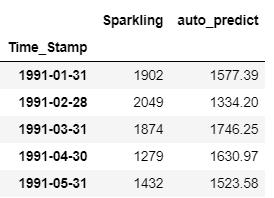
Three parameters 𝛼 , 𝛽 and 𝛾 are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Autofit parameters for the Triple Exponential smoothing:

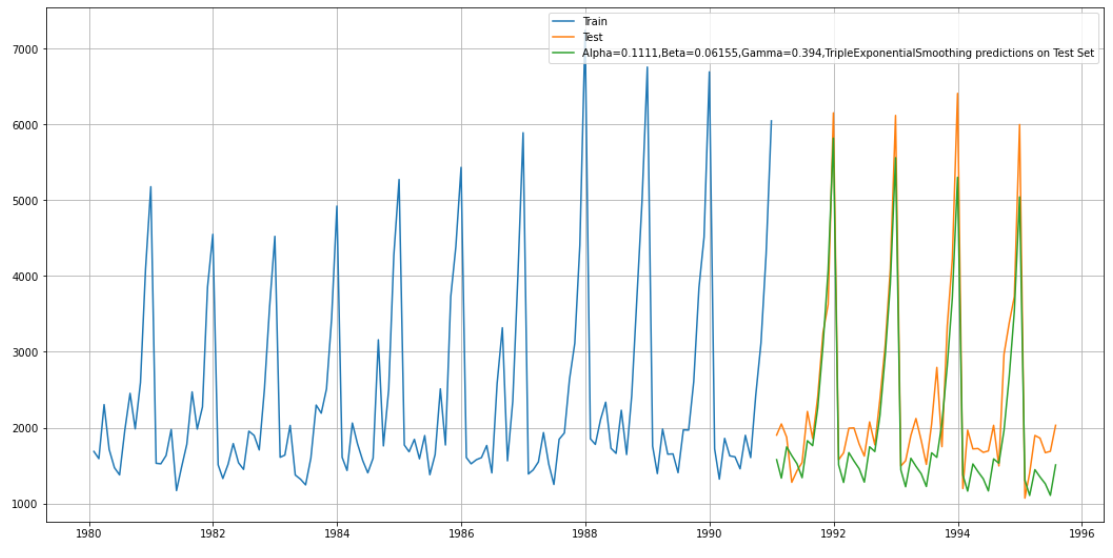


The above fit of the model is by the best parameters that Python thinks for the model. It uses a brute force method to choose the parameters.

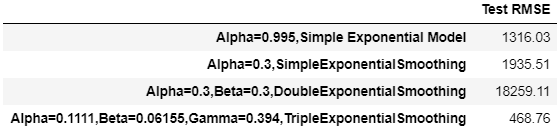
Lets predict on the test data:



Plotting on both the Training and Test using autofit



For Alpha=0.1111,Beta=0.06155,Gamma=0.394, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 468.758

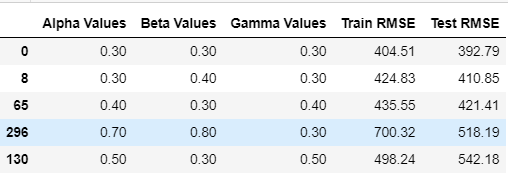


Lets define an empty dataframe to store our values from the loop



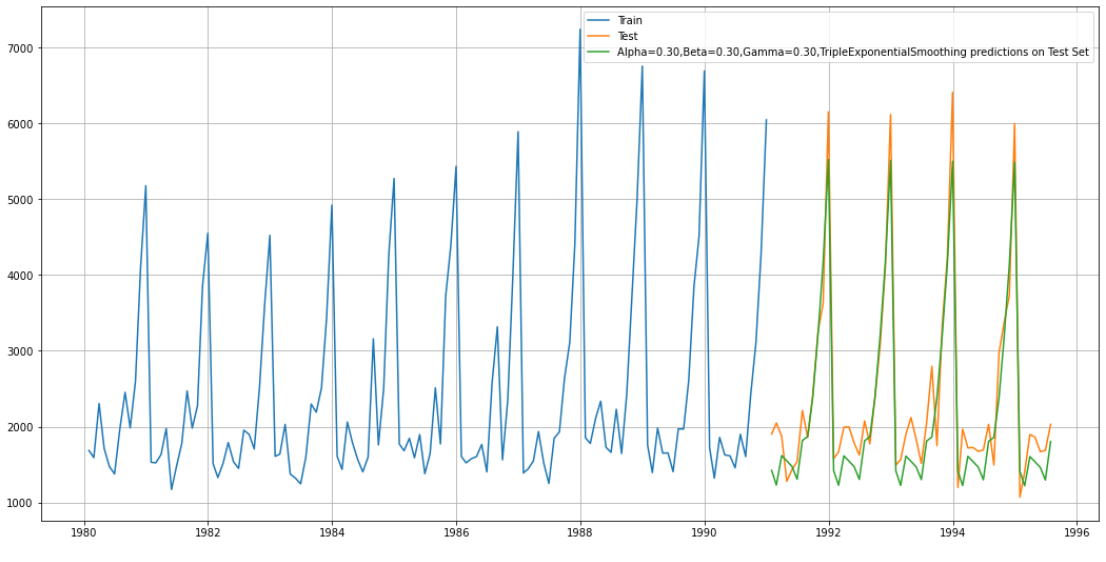
We will append all the results to it, by changing values of alpha, beta , gamma.

After Sorting the first five rows are

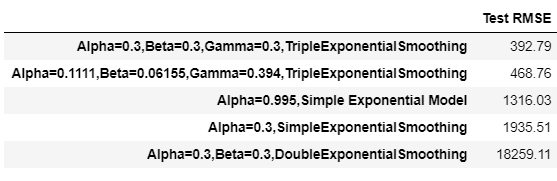


So this model is giving RMSE 392.79 lowest so far, and Ideal values are 0.3,0.3,0.3 for Alpha, Beta , Gamma.

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

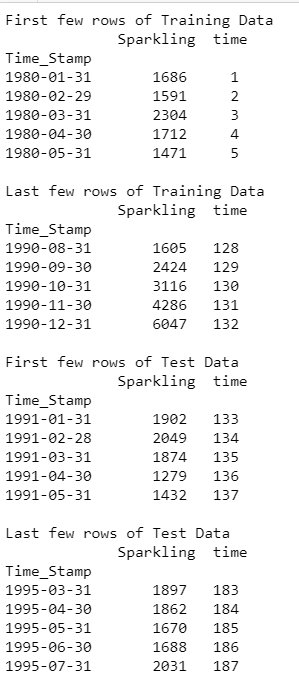
Lets Build Other Models Also.

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

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

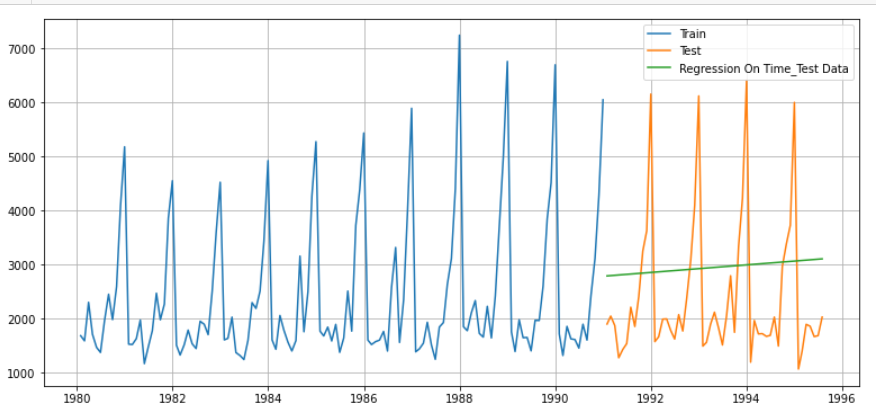


We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.



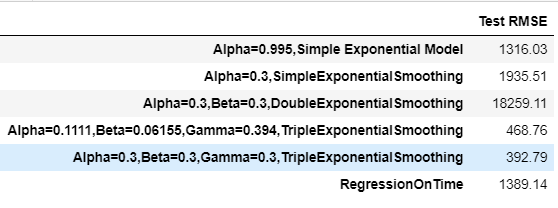
Now our training and test data has been modified, let us go ahead use Linear regression to build the model on the training data and test the model on the test data.

Prediction plot:



#### **Model Evaluation**

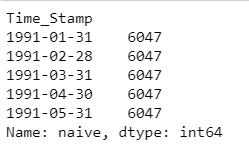
For RegressionOnTime forecast on the Test Data, RMSE is 1389.35



#### **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:



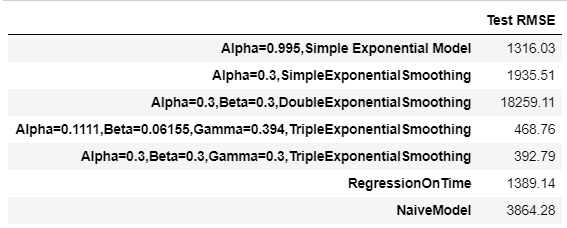
#### **Prediction Plot:**

#### 

#### 

#### **Model Evaluation**

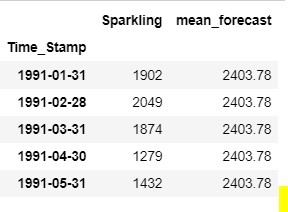
For RegressionOnTime forecast on the Test Data, RMSE is 3864.279



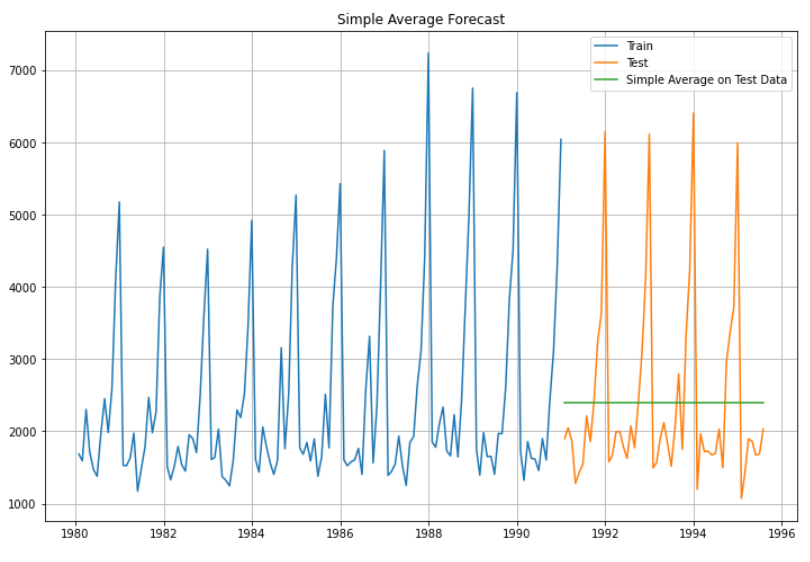
#### **Model 6: Simple Average**

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

Forecast:

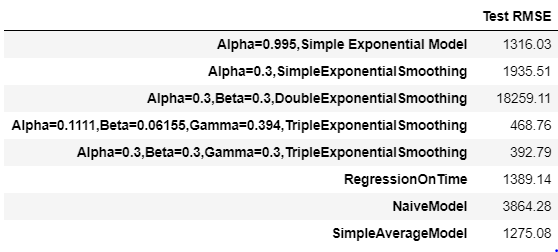


**Prediction Plot vs Test Data**:



#### **Model Evaluation**

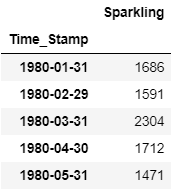
For Simple Average forecast on the Test Data, RMSE is 1275.082



#### **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.

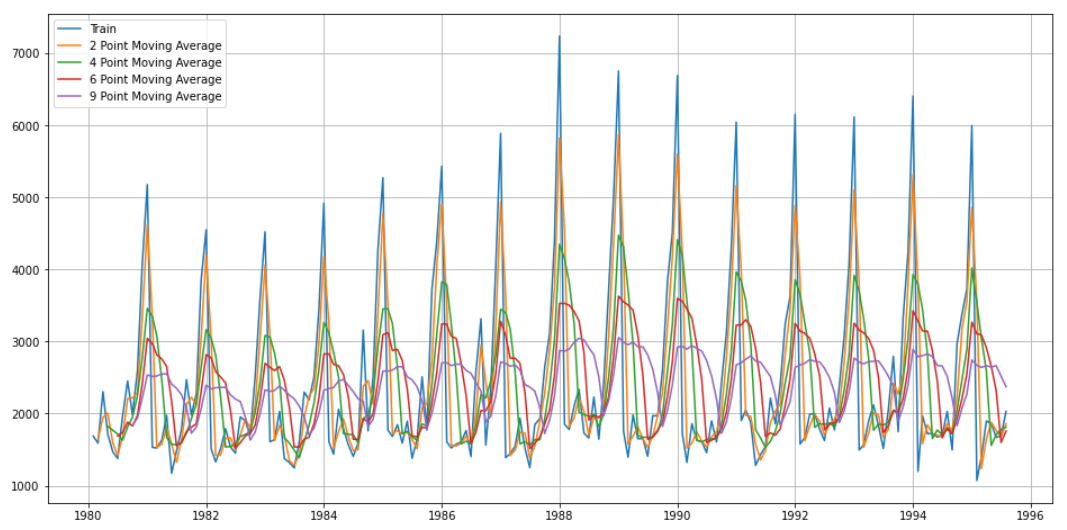
For Moving Average, we are going to average over the entire data.



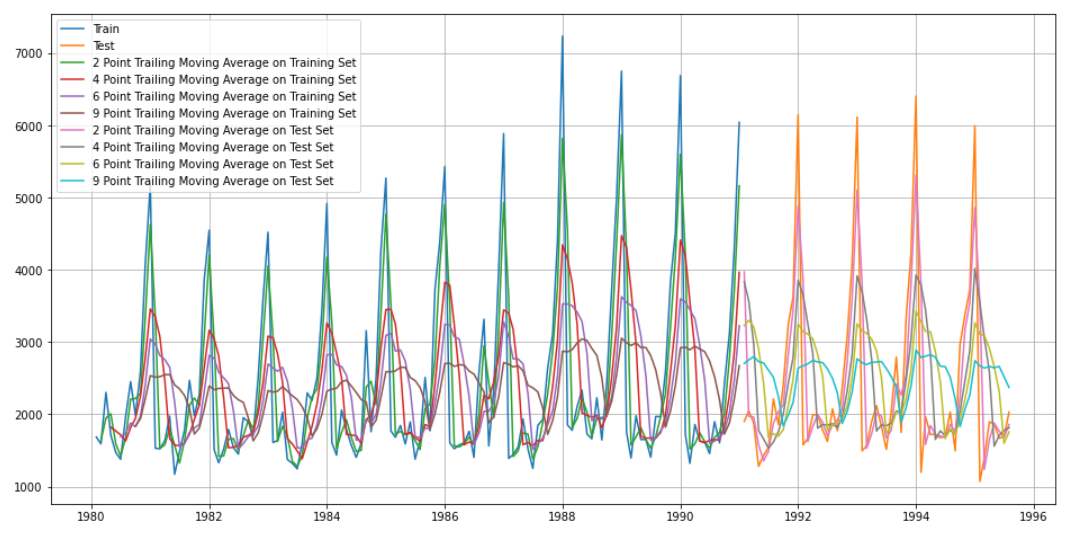
For Trailing Average,



**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 813.401

For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590

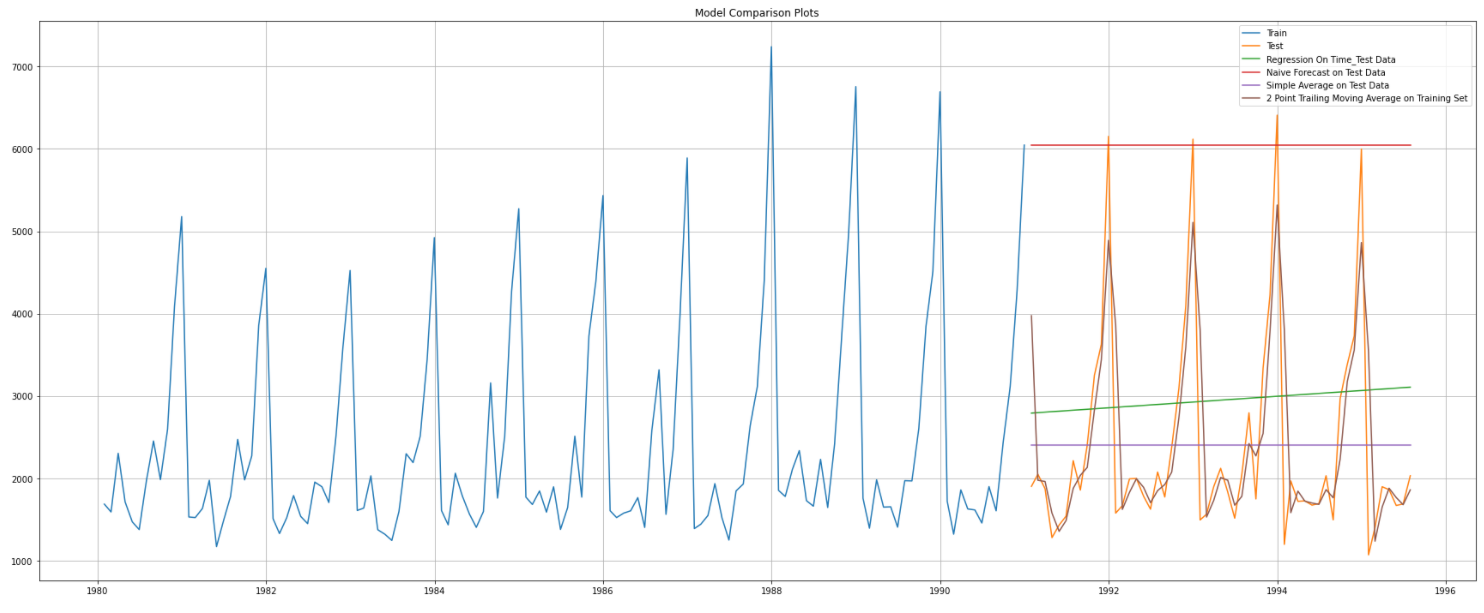
For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927

For 9 point Moving Average Model forecast on the Training Data, RMSE is 1346.278

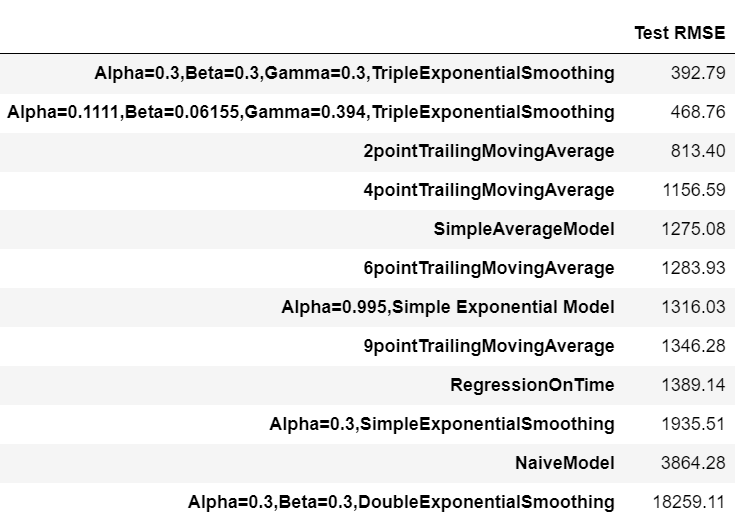
Test RMSE results for All models:



**Plotting on both Training and Test data**



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



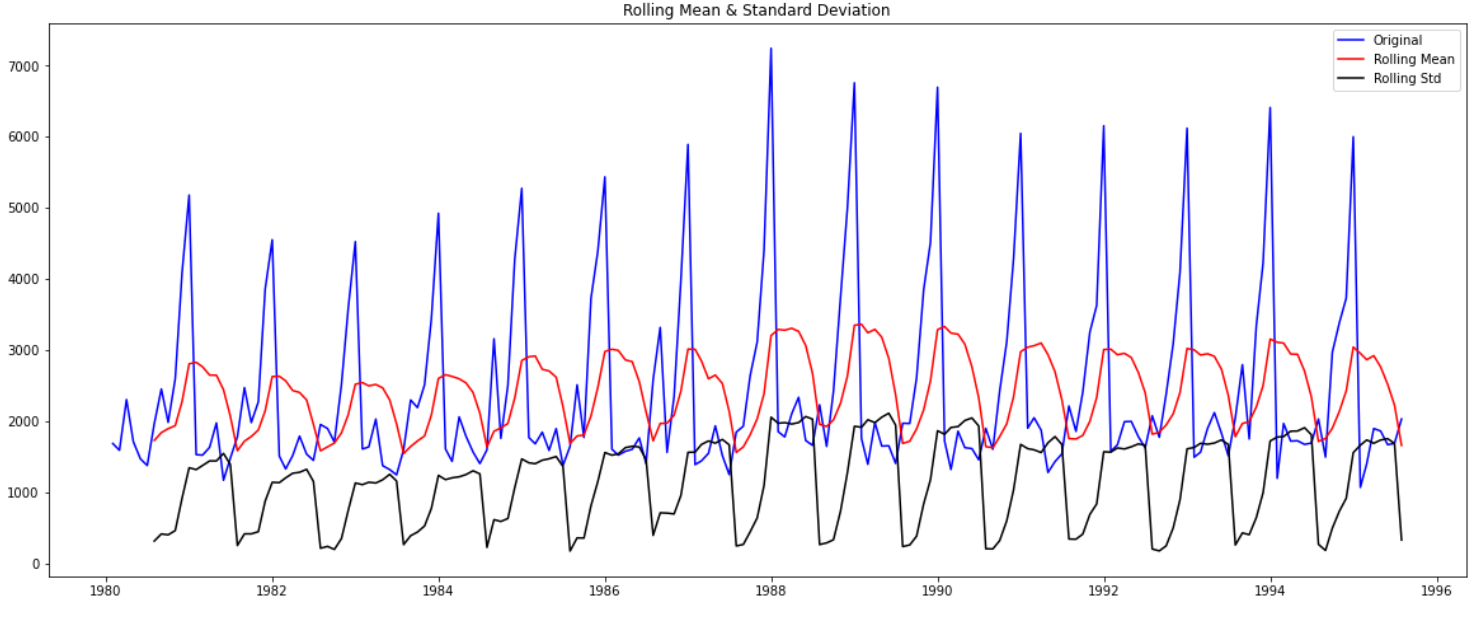
# 5.Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and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

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

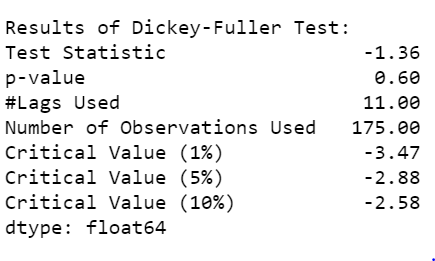
Dicky Fuller Test

Null Hypothesis H0- Series is not Stationary

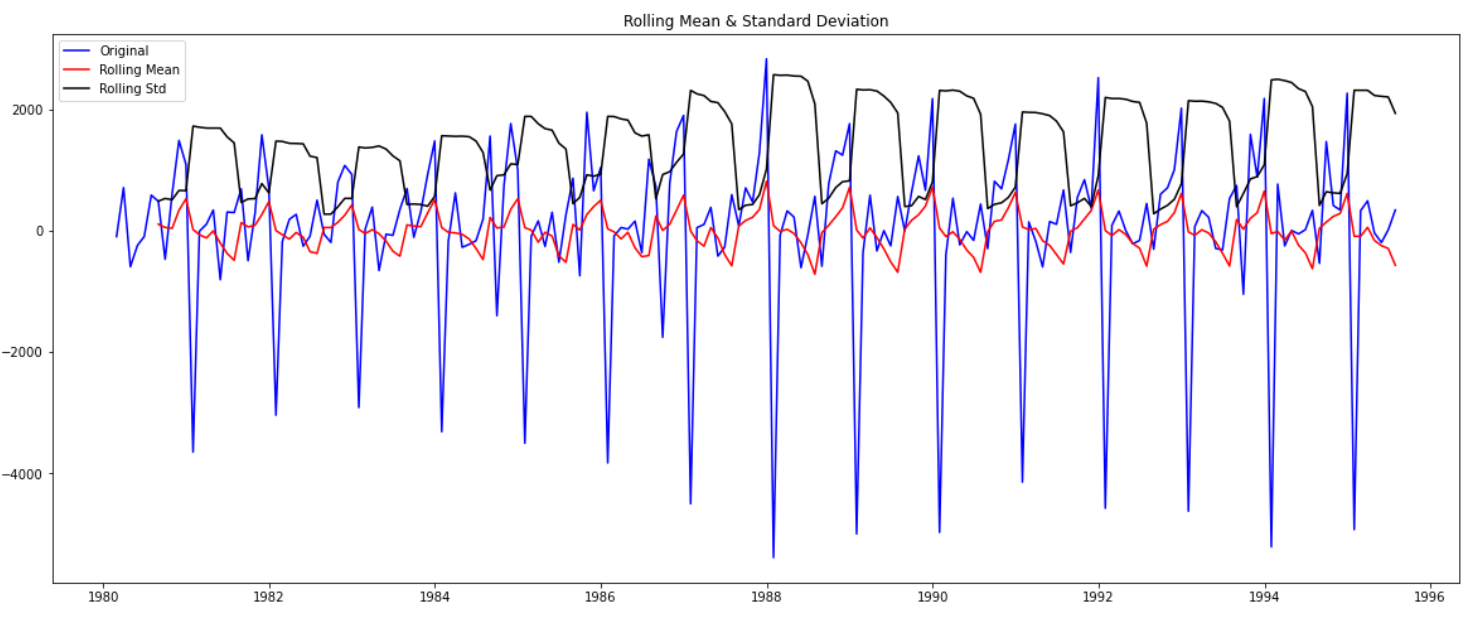
Alternative Hypothesis H1- Series is Stationary

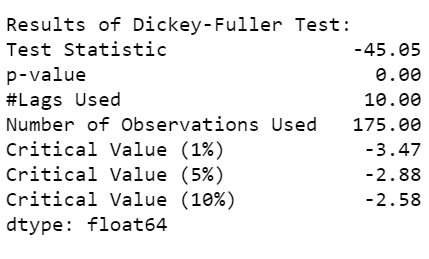


P value is higher that 0.05 Hence Null Hypothesis is True, we will take Order 1 differencing to make series stationary.



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



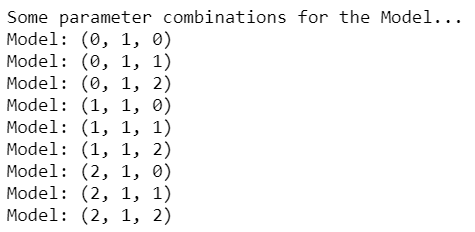


We see that at P value is less than 0.05 the Time Series is indeed stationary.

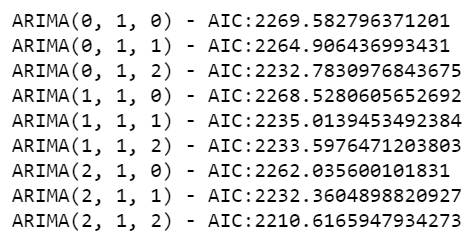
### 6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE. 8

#### **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.

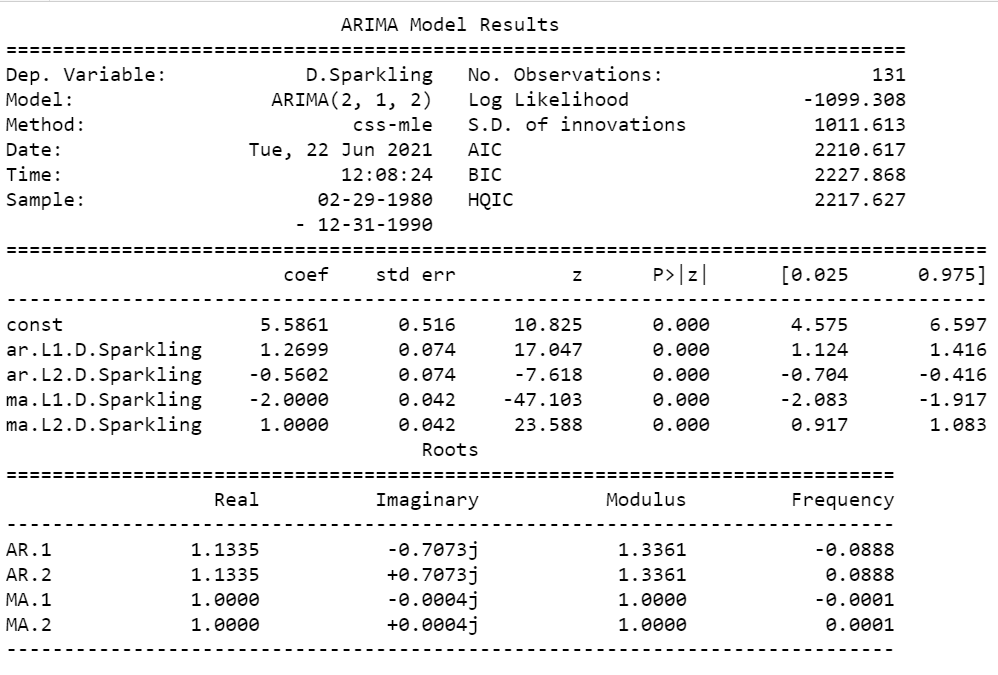


We will Apply all values of p and q. check which combination is giving us the low AIC score.

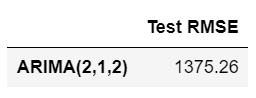


Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value

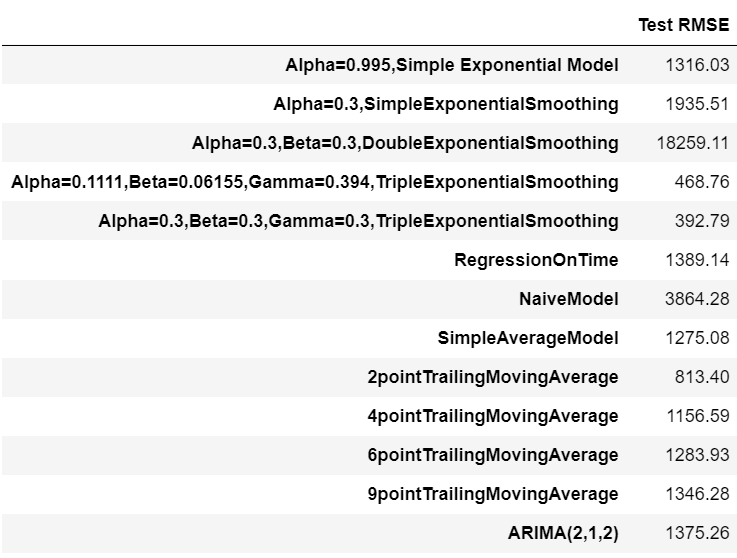




**Predict on the Test Set using this model and evaluate the model.**



Performance of Models so far:

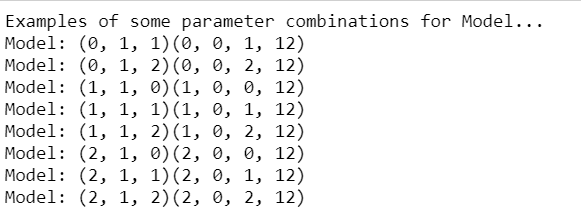


Still this point Triple exponential has performed the best.

**Automated Version of SARIMA**

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

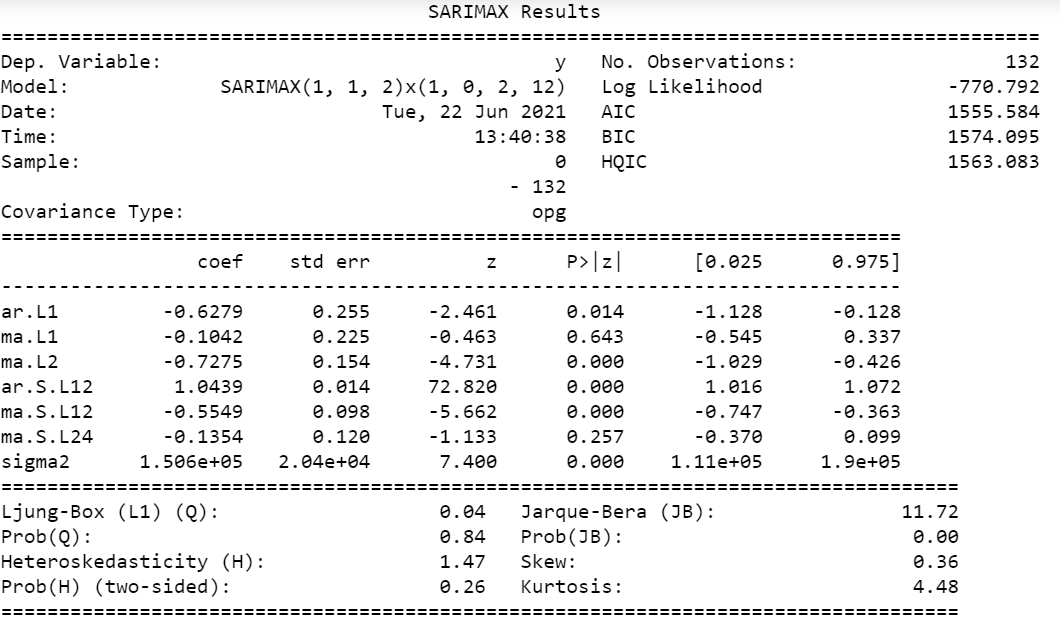
It is extension of ARIMA model by considering Seasonality factor.



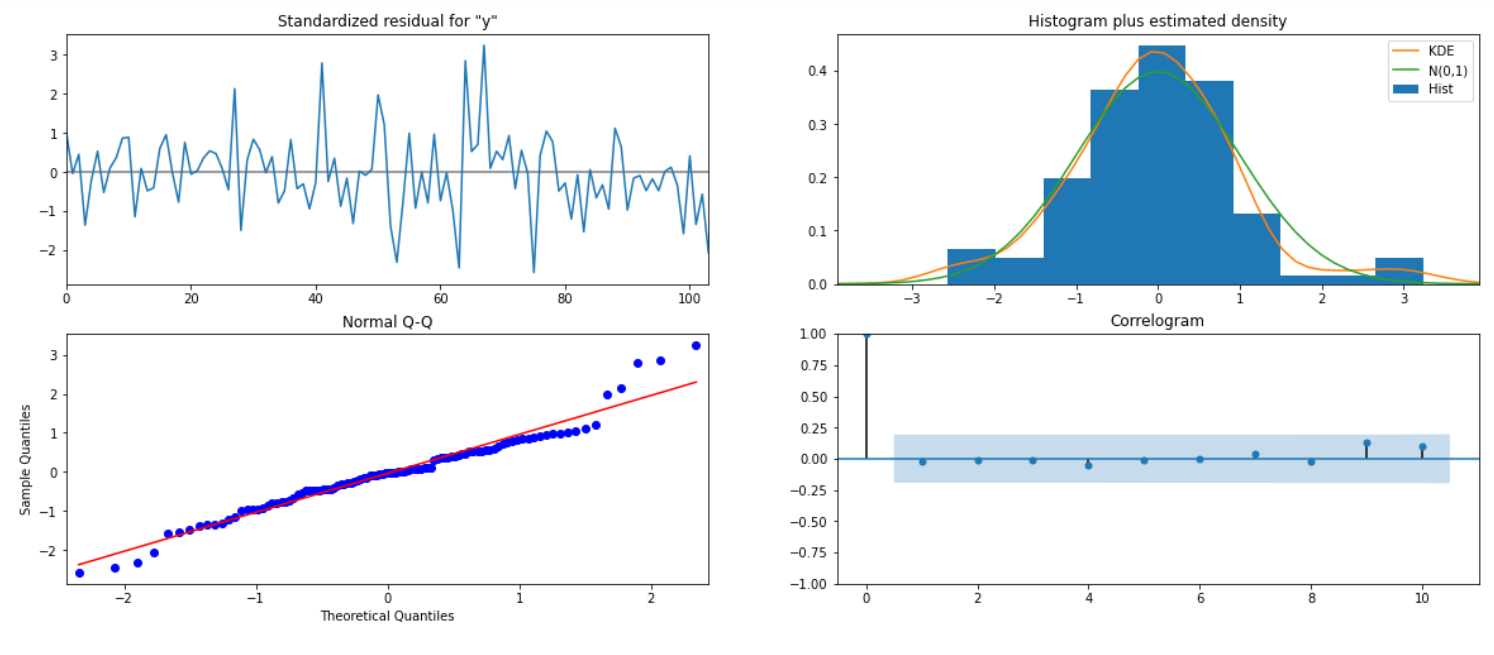
Top 5 Parameters AIC score:



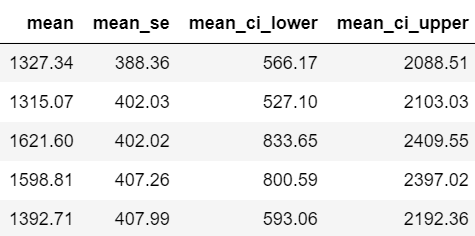
From the above result we will choose (1,1,2)(1,0,2,12).



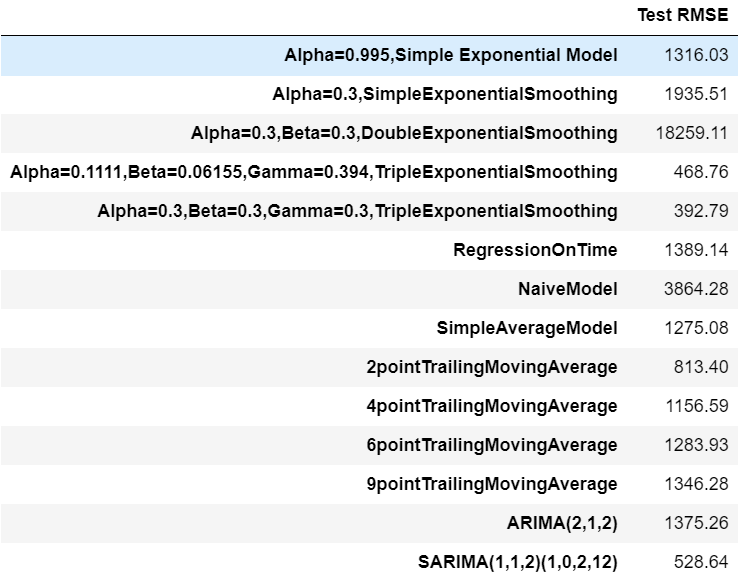
Results:



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



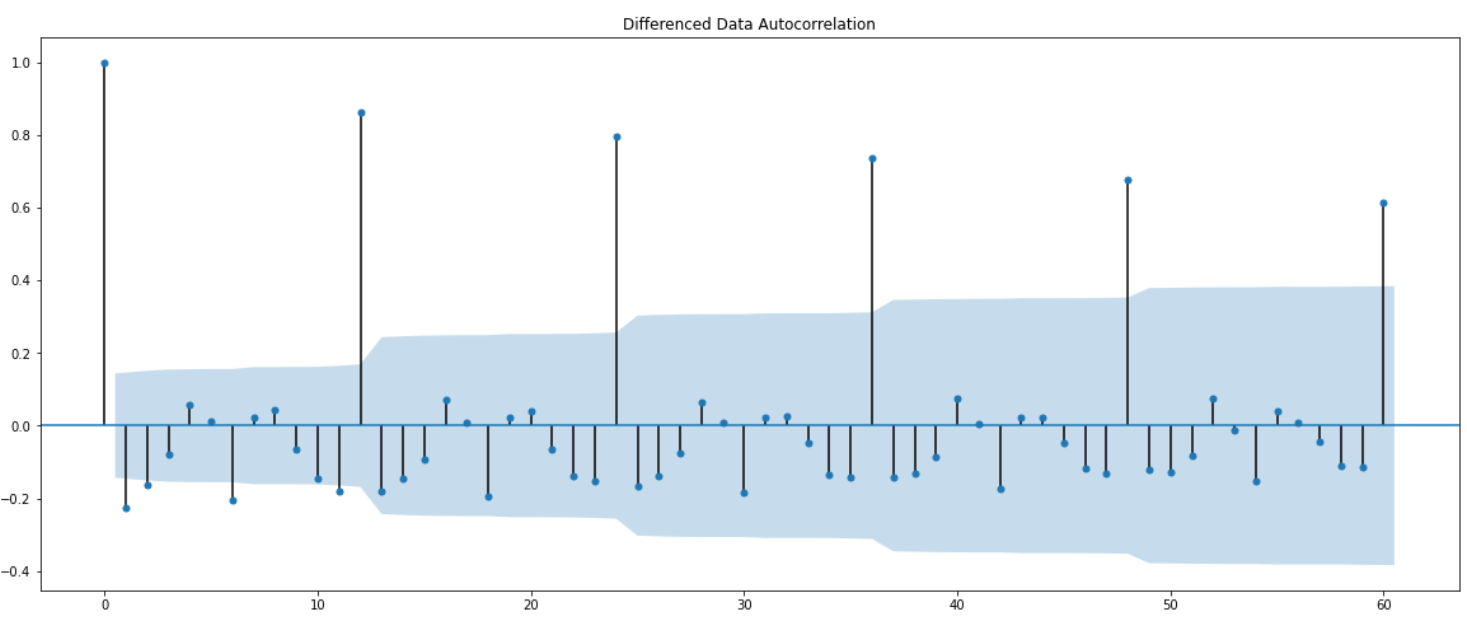
RMSE for SARIMA model is 528.64.

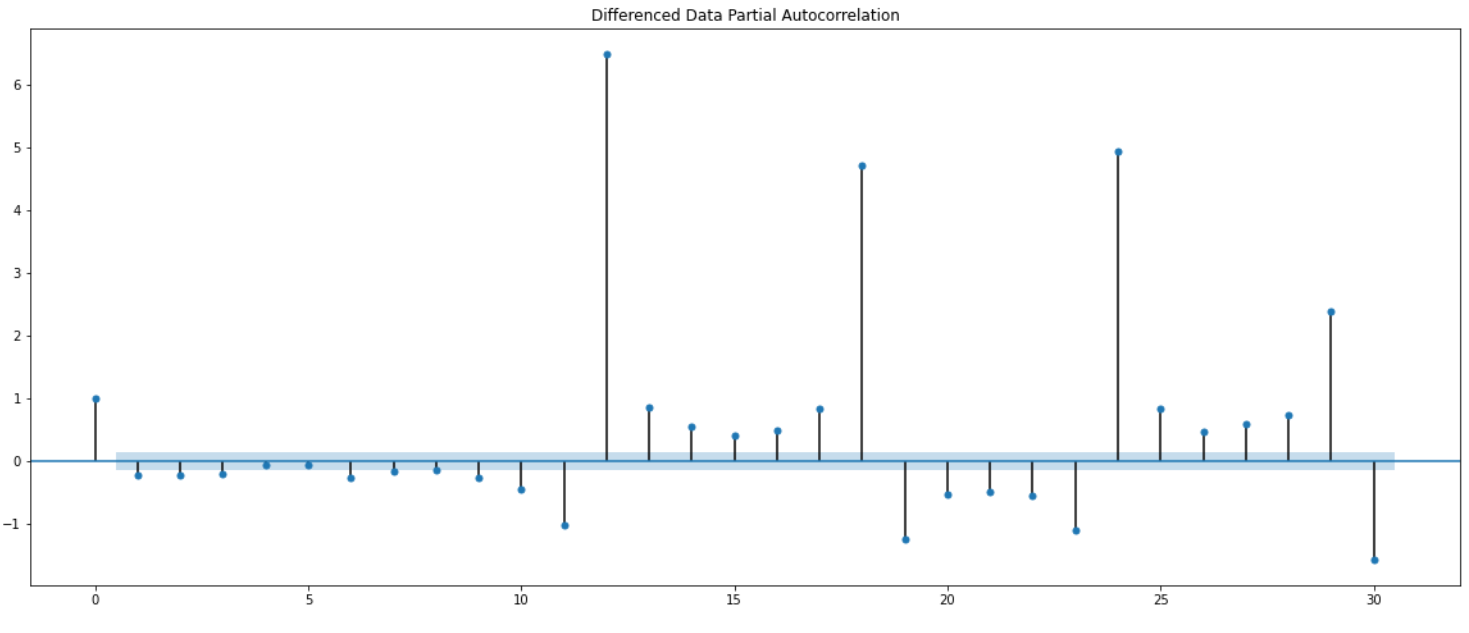


### 7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

#### **Manual ARIMA**

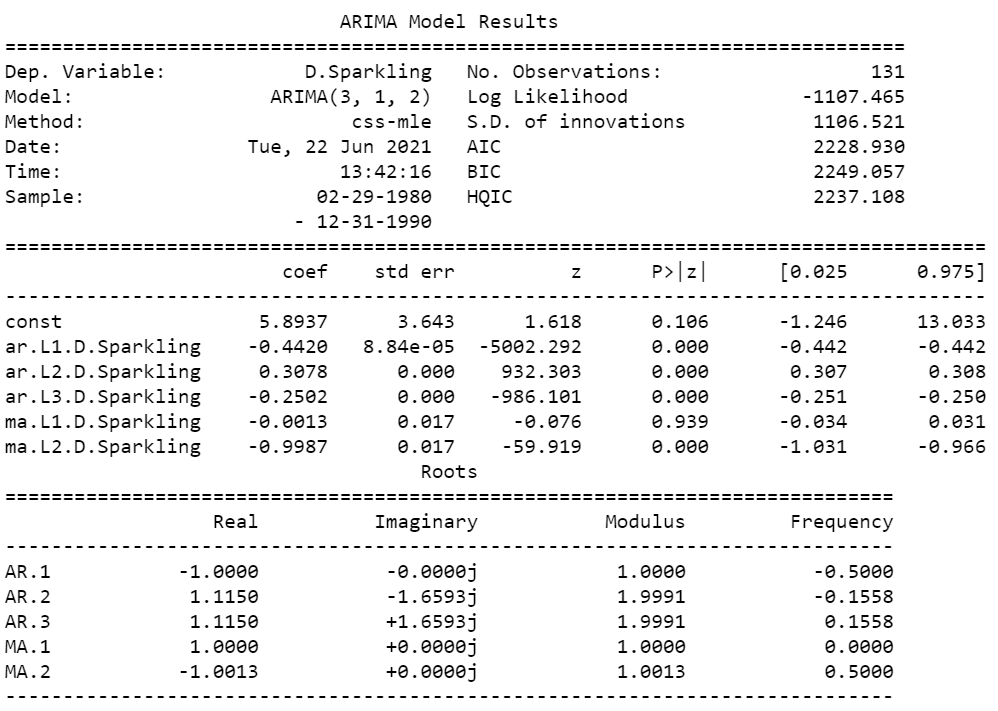
Let us look at the ACF and the PACF plots





Here, we have taken alpha=0.05.

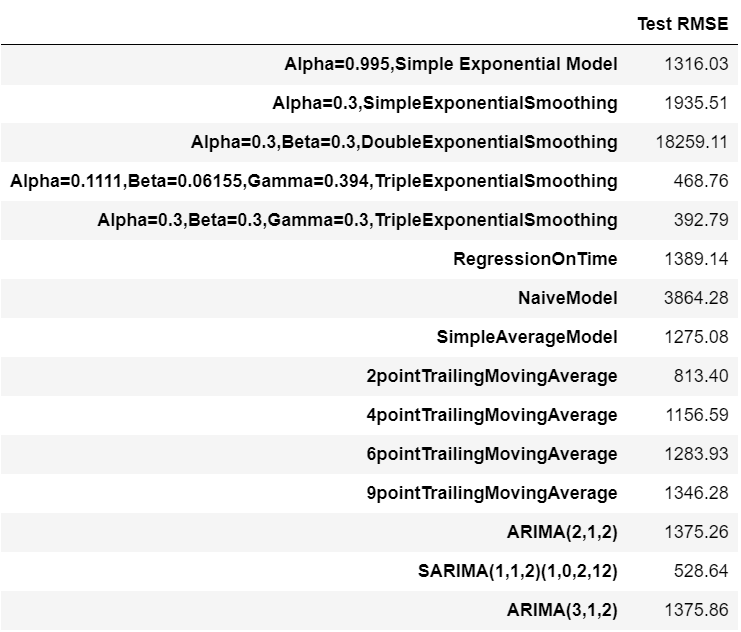
The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0. By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 3 and 2.



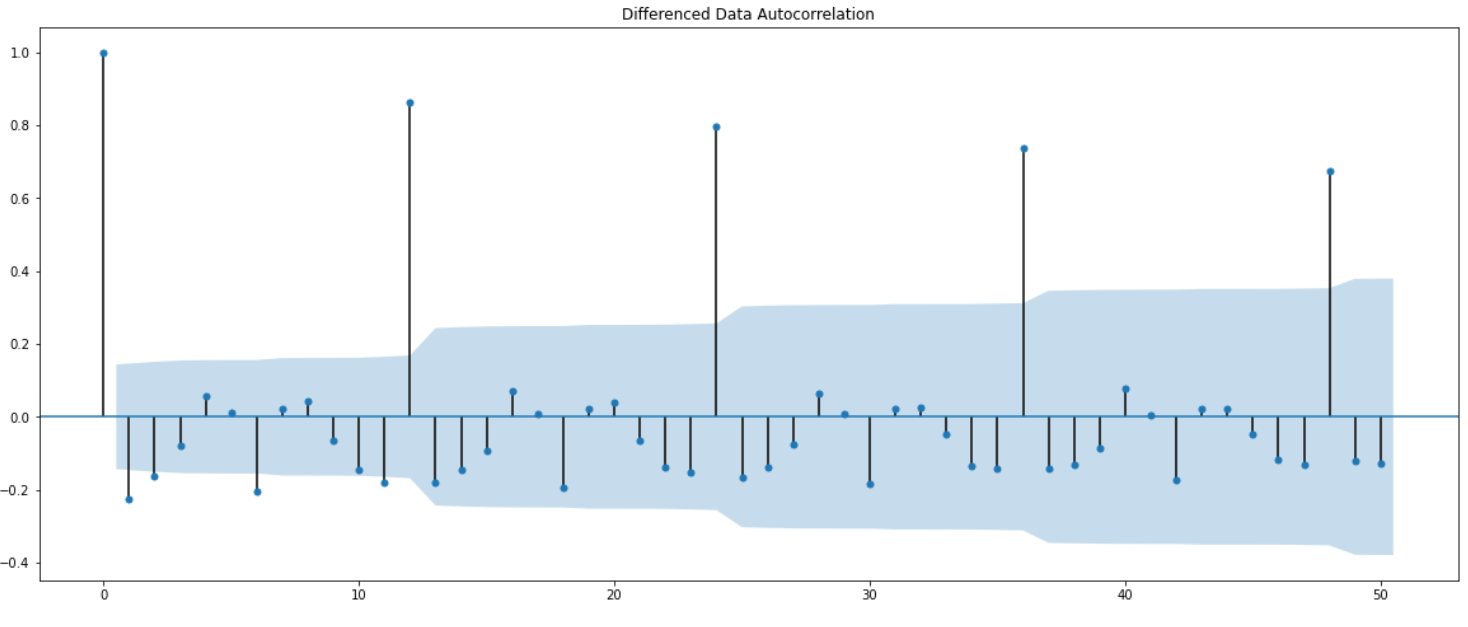
Predict on the Test Set using this model and evaluate the model.

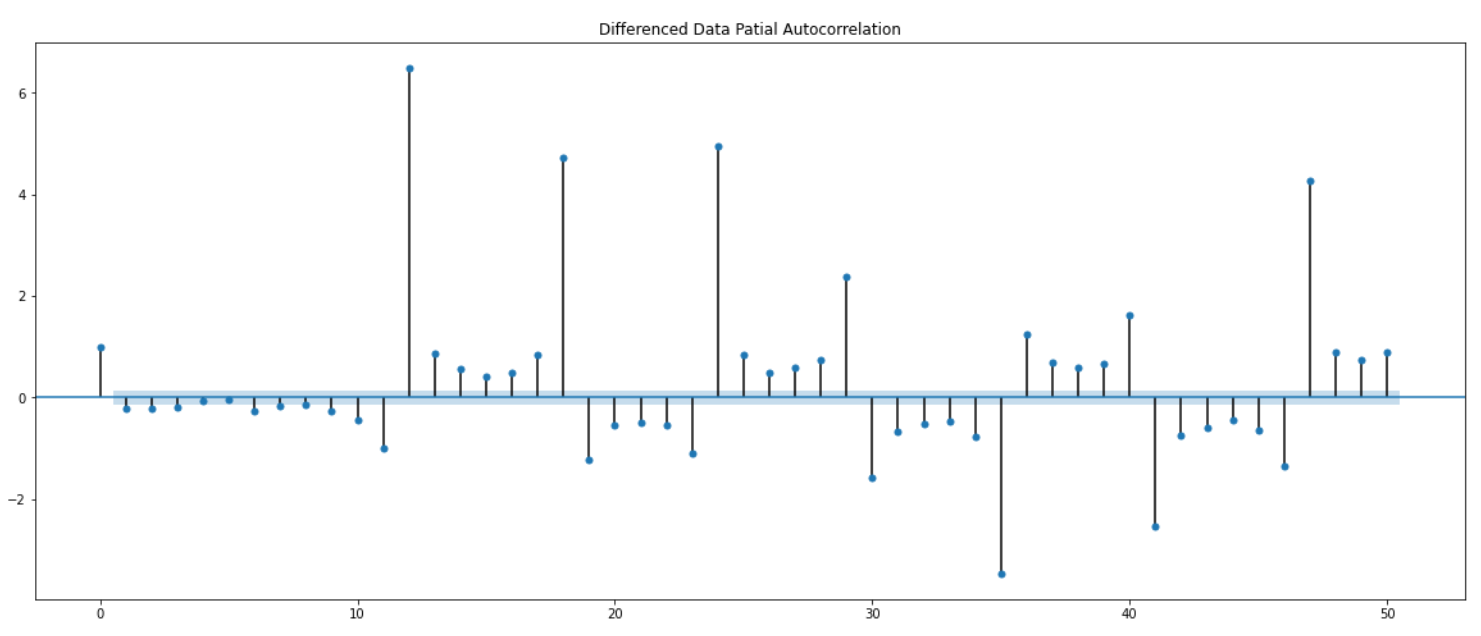
RMSE for Manual ARIMA model is 1375.80.

The Overall Results till now,



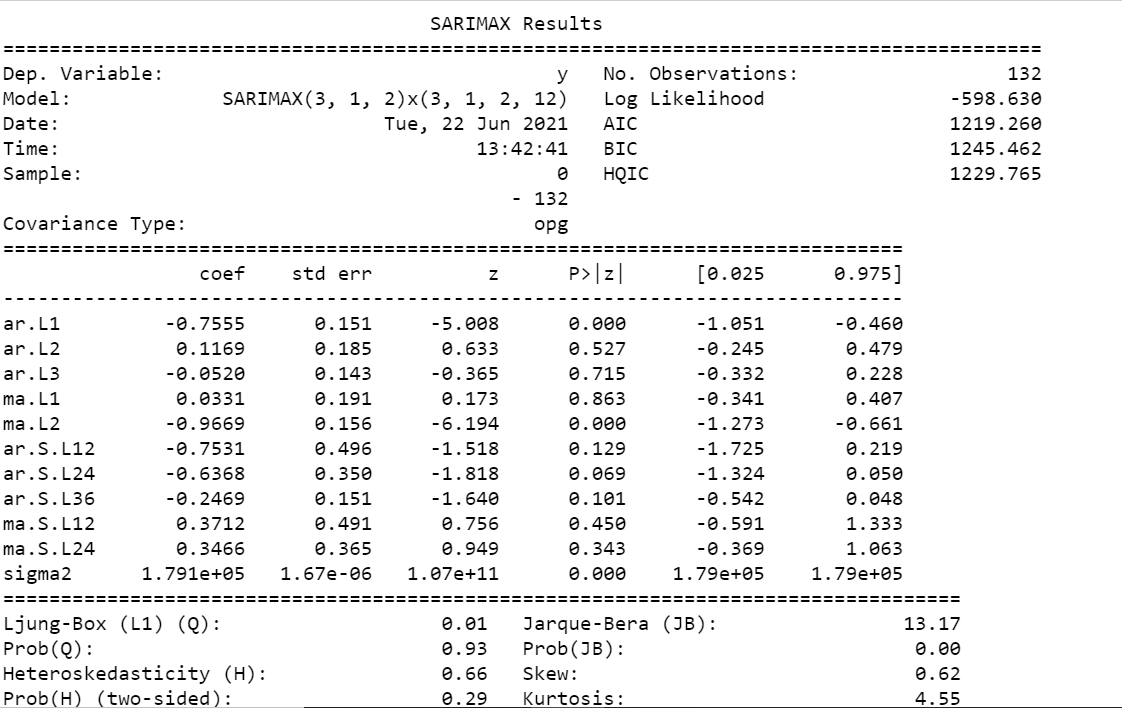
#### **SARIMA Model : Manually looking at ACF and PACF**



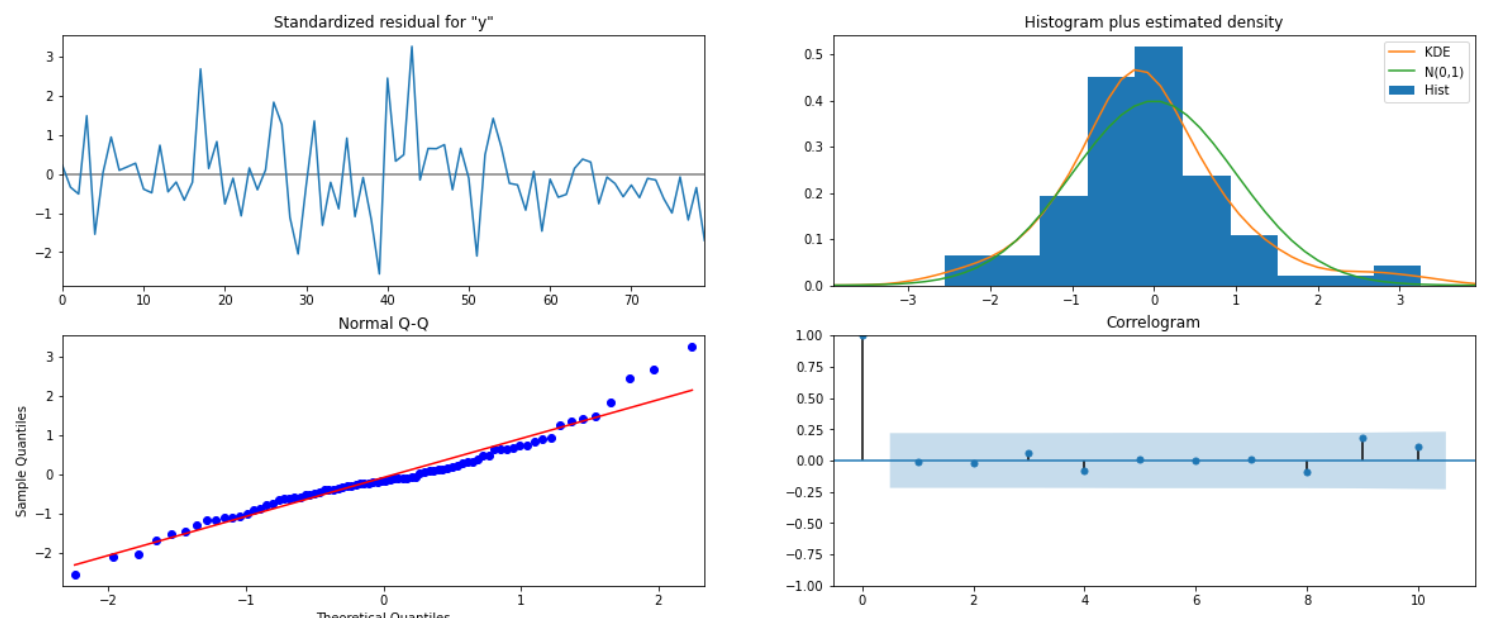


The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF 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.

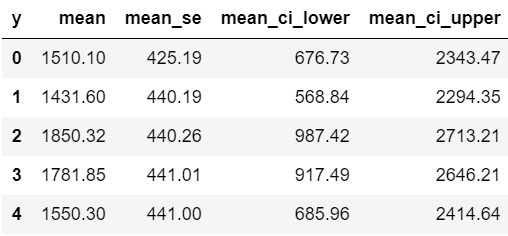
This is a common problem while building models by looking at the ACF and the PACF plots. But we are able to explain the model.



Results Dignostics for Manual SARIMA:

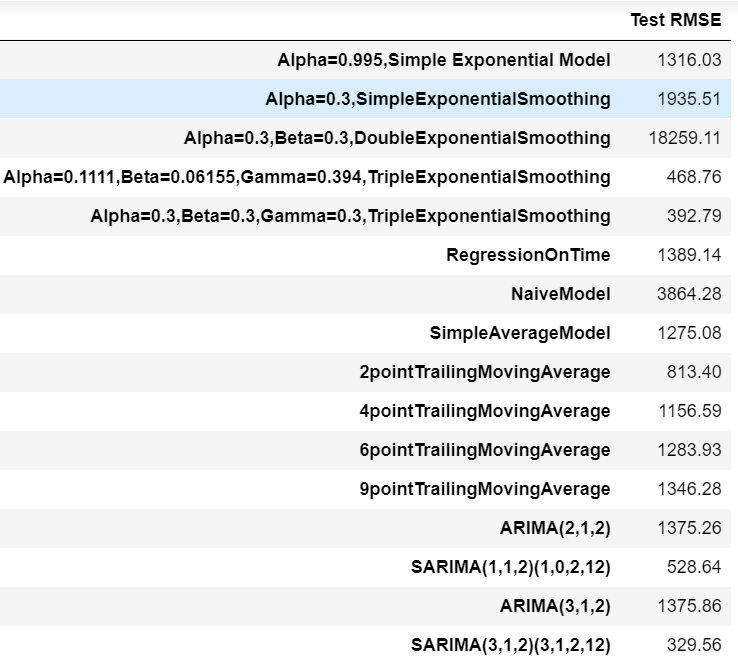


### **Predict on the Test Set using this model and evaluate the model.**



RMSE for Manual SARIMA is 329.55

### 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.



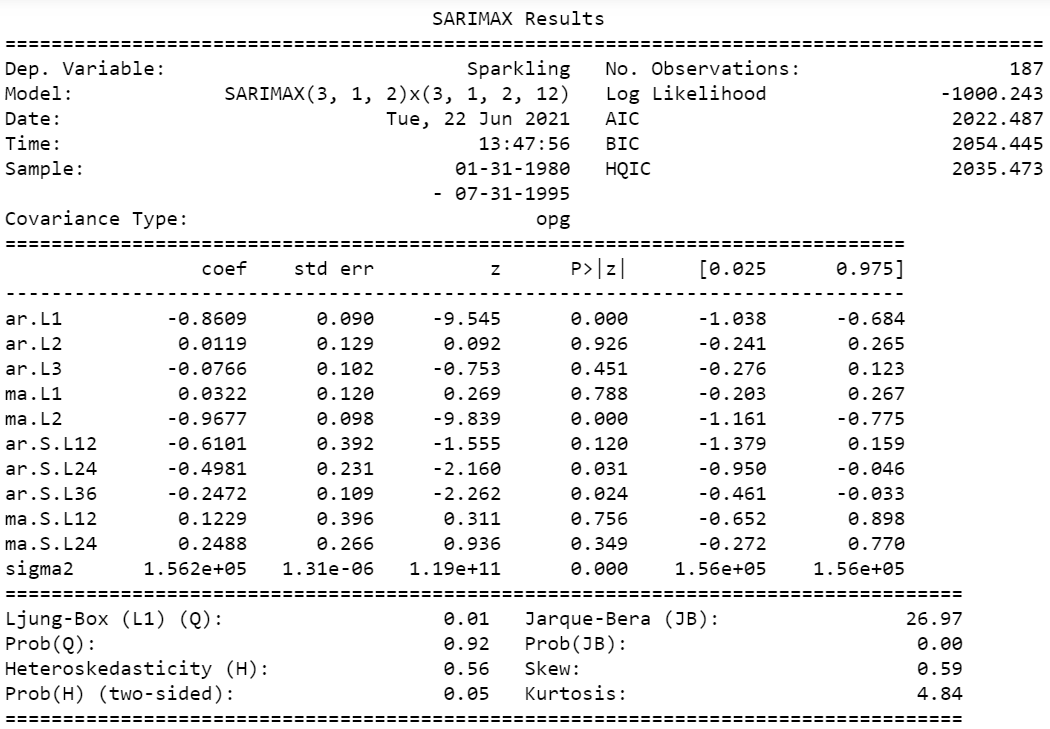
## 9.Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

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

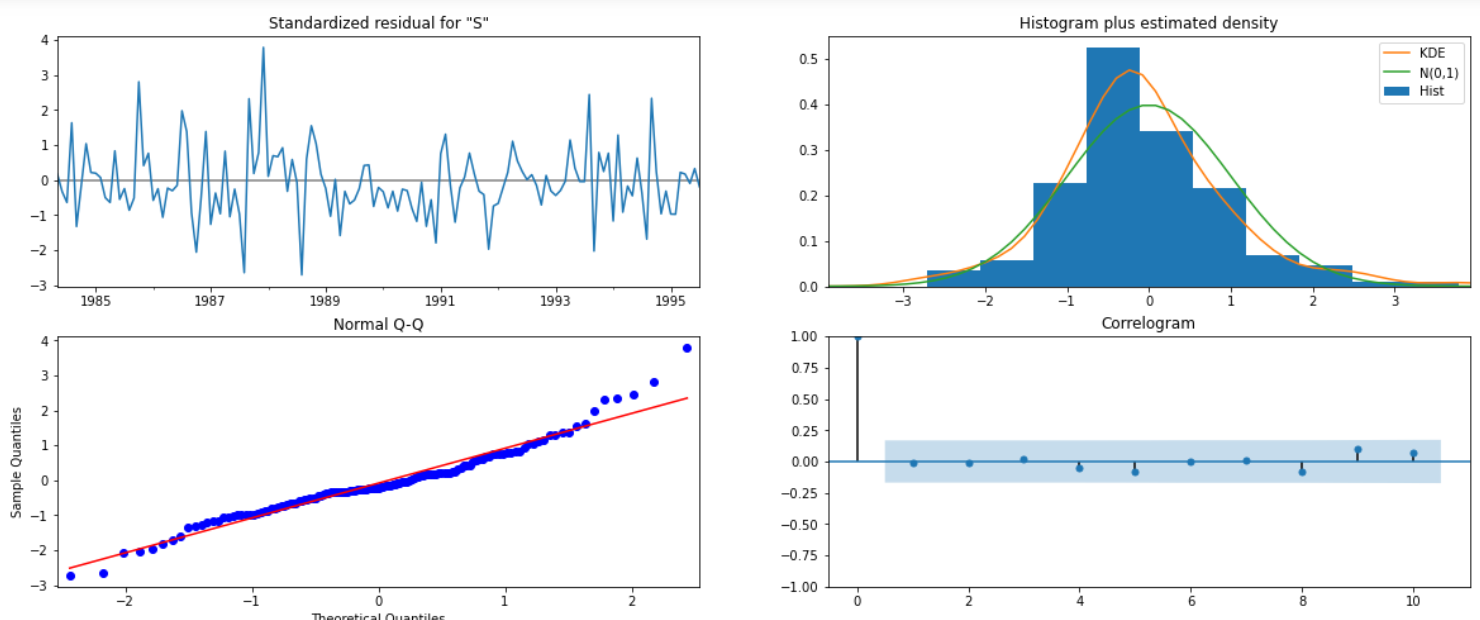
So we will Select SARIMA model, we will give train data as Whole Data. That is, Test and Train data earlier, Now all data will be used for Training for Purpose.

We are predicting here for Next 12 Months.

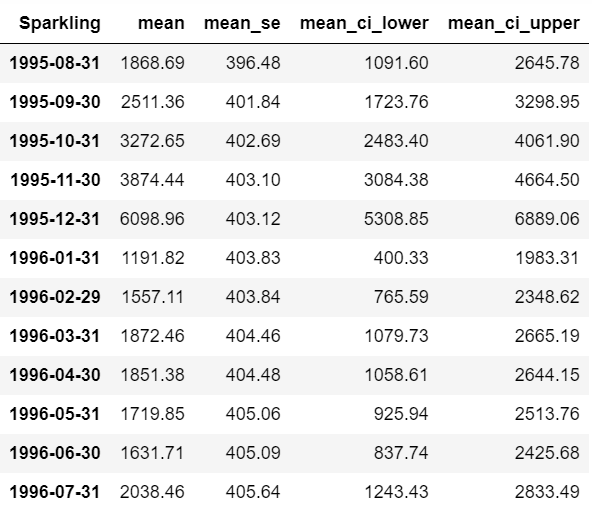
#### **Building the most optimum model on the Full Data.**



**Results Dignostics:**

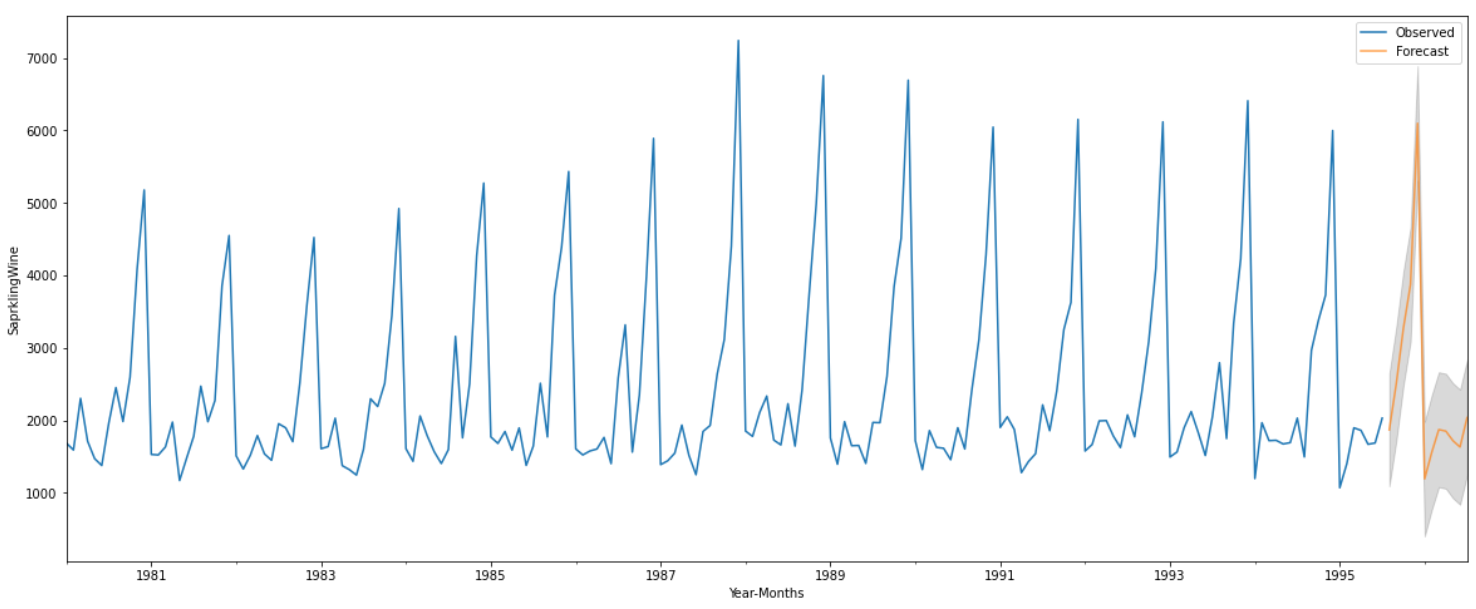


***predict 12 months into the future.***



RMSE of the Full Model 578.9516037053722

***plot the forecast along with the confidence band***



***In The above plot Orange line shows Sales of Sparkling wines in 12 Months.***

## 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 Analyse the sales of Sparkling Wine we got hidden insights by performing Exploaratory Data Analysis. We performed EDA then we get to know

1.From 1981 to 1987 trend is upward, but from 1987 is both upward and downward.

2.Median Value of Sales are remains around 2000.

3.Highest Sales of Sparkling wine is occurred in year 1987, and in 1995 one of the month has brought Lowest sale of wines.

4.From Monthly plot, Sales has been Increasd from August to December. Till this point Sales are around 2000. so We need Higher stock or Production from August to December. October,November,December has sales more than 2000 and it is almost equal to 3000,4000,6000.

5.In month of december atleast we need 7242 Sparkling wines in stock as this is the highest number reached.

After Decomposition it is clear that Series has Trend and Seasonality both.

So we Applied the models like Simple, Double, Triple Exponential Smoothing to the Train data till 1991. and then tested it on Test data. Out of these Triple exponential smoothing seems to be quite effective as the series has all three factors Alpha, Beta, gamma. Season, trend and level. The RMSE is low for Triple exponential smoothing.

Further we have applied models like Linear Regression and Naive Bayes, per as we can see in the Dignostics it failed to predict sales, and RMSE is also high for these models.

After We moved to the models like Simple and Trailing Moving Average.This time it was little better than Linear Regression and Naive Bayes.

After That Finally we Applied ARIMA and SARIMA Automated by AIC scores and Manual ARIMA SARIMA by observing ACF and PACF plots. PACF has given p value while ACF has given as q value.

The series has Seasonality hence The best suited Model was SARIMA, Only 329 RMSE on the Test Data, and When Applied on Full data it has given RMSE 578, Which is still good as compare to Others.

So we Finalize This Model. Here we have taken P as 3 and Q as 2, D=1 as we got stationary series after one order differencing.

Finally, We Predict Sales for Next 12 Months, and in December we will need 6099 Sparkling wines atleast to fulfill the requirement.

In Future Company Should be ready with the Stock of the Year,Company should hire staff as per the Requirement. If Company is producing wines for December then they must be think of manpower required. Excess Production might hamper the Business, so as predicted Company should Target.

If Company want to Increase Revenue, then they have to Think for all year instead of focussing only on December.

**THE END**