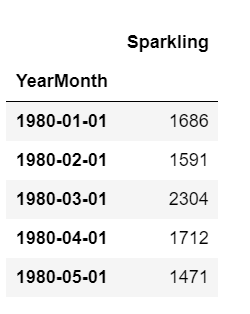
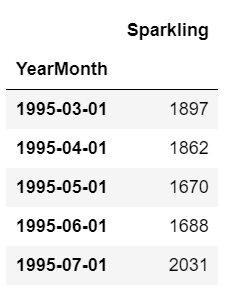
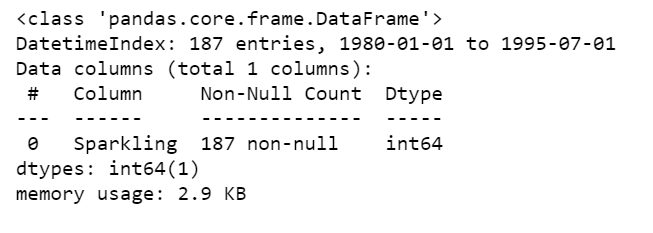
**TIME SERIES PROJECT SPARKLING**

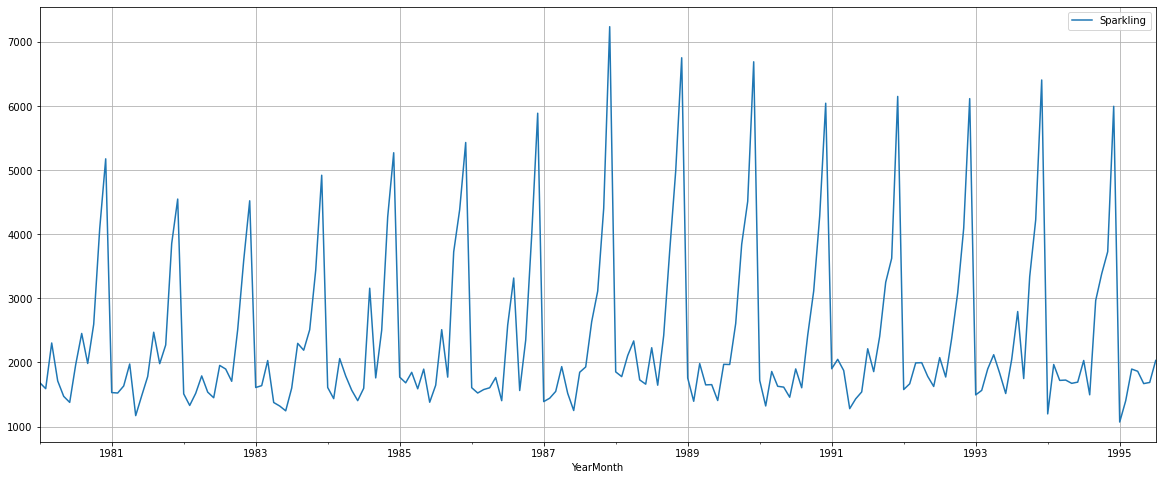
1. **Read the data as an appropriate Time Series data and plot the data.**

The above is the head of the dataset. The year and month column are made as the index of the series. And parse date is set as True. We have data from starting of 1980-01-01 and ending till 1995-07-01. So, we have 15 years and 6 months of data for this series.

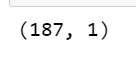


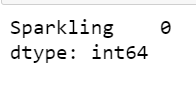
There are total of 187 records and the sparkling variable is in int64 format. There are no null values as all 187 records are available.

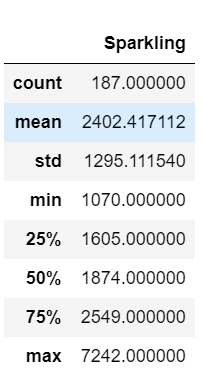


We can see that there is no trend present in the above series. There is a sure seasonality present. End of the year there is a rise and beginning of the year there is a sharp fall in sales.

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

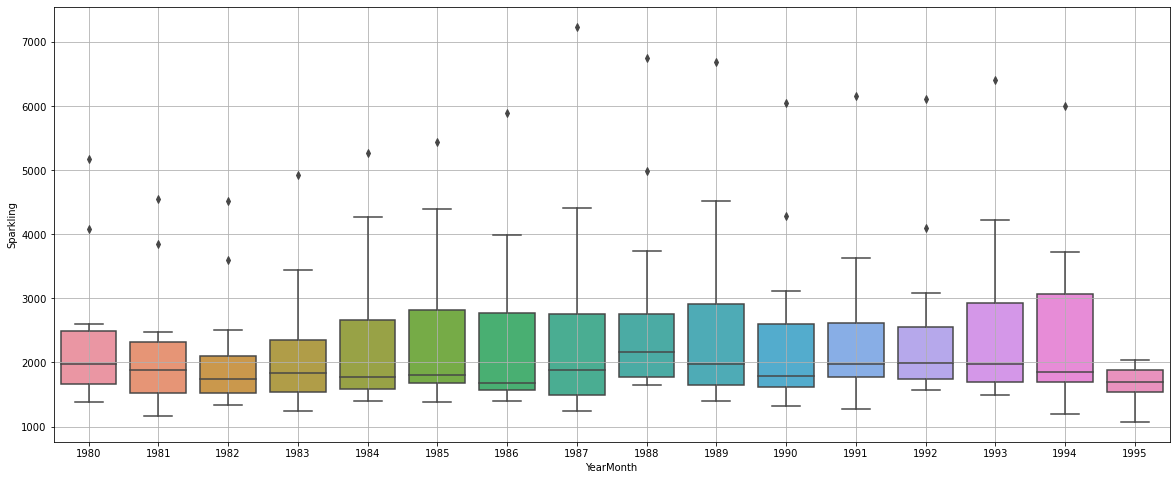
**Null Values:**





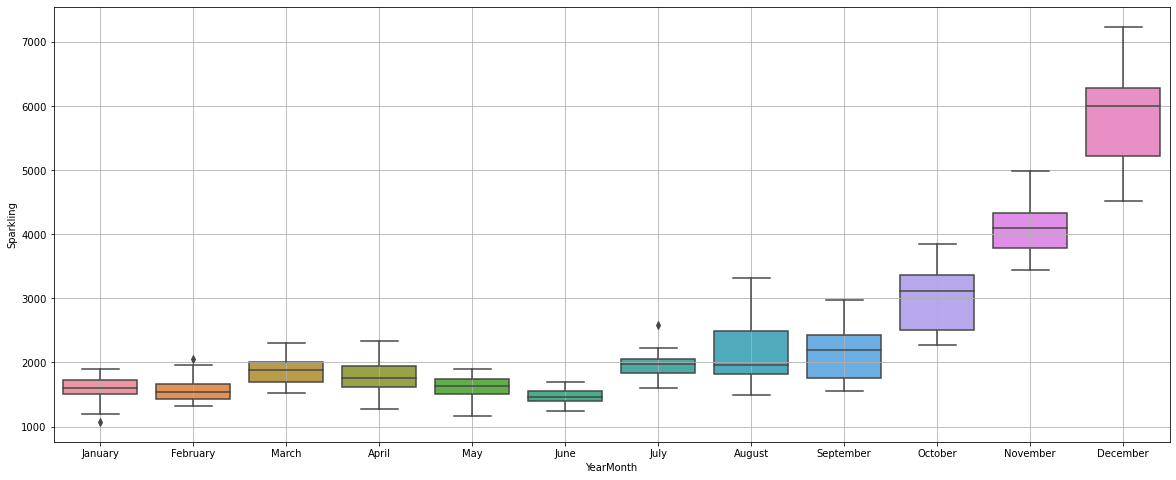
There are 187 records totally present in this sparkling series and there are no null values present. In the describe function we can see that the mean sales are around 2402.

**Yearly Plot:**

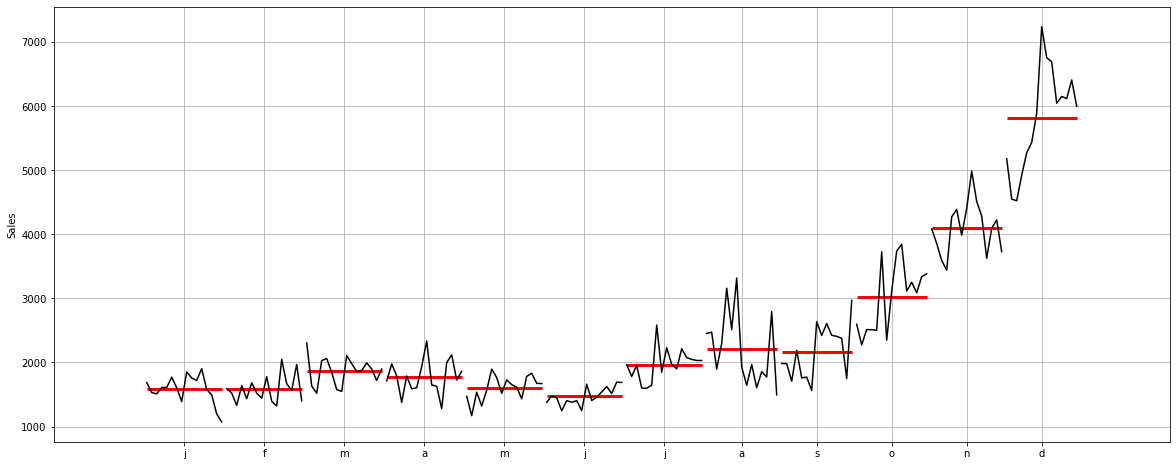


We can see that the sales are mostly consistent throughout years. In the beginning of the series ie during 1980 1981 and 1982 the sales were less and from then the sales is mostly consistent and similar when compared to other years.

**Monthly plot:**

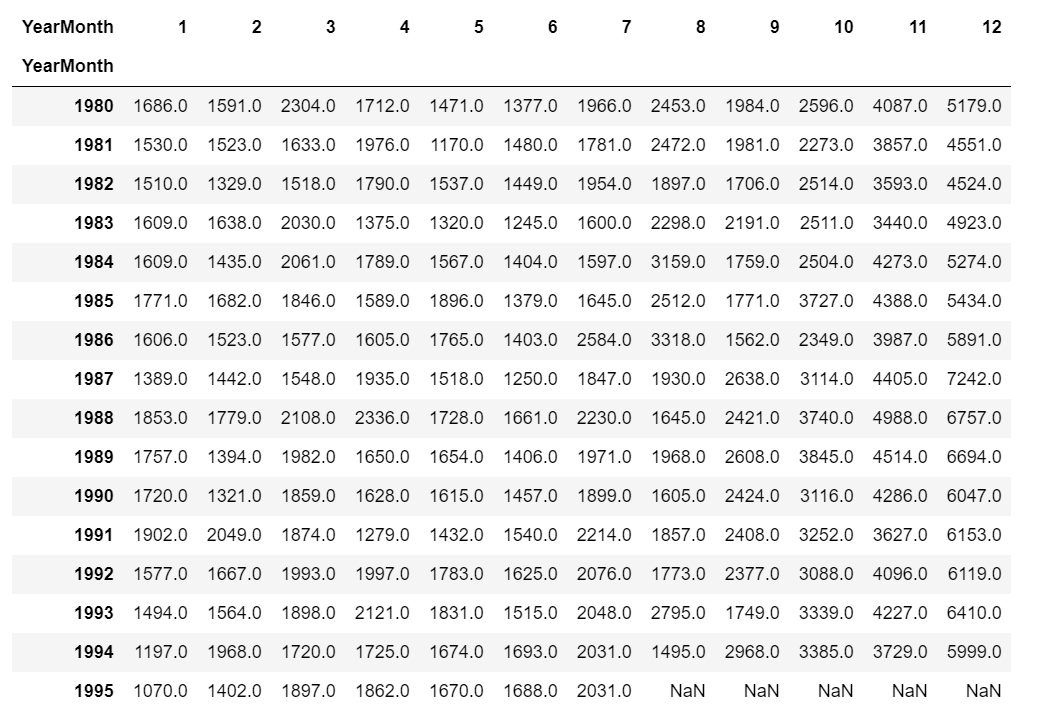
****

Here we can see that from January to July the sales are less. All these months have a similar sale. But from august to December the sales are increasing except for September. December has the highest sales of all the months.



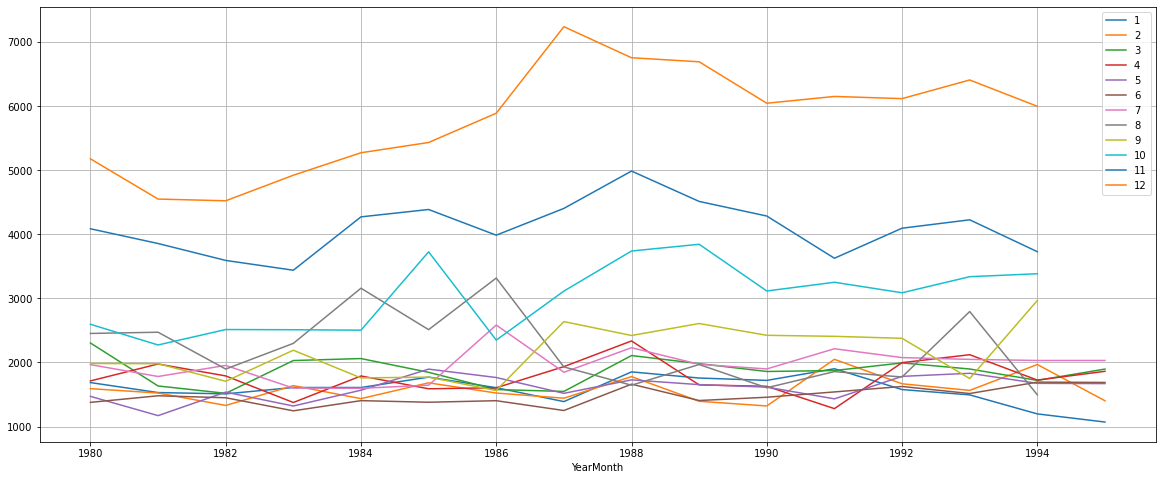
This graph shows the average sales of all the months that is represented by the red line. We can see that the from September the sales average is increasing. **This shows us that there is clear seasonality in our data.**

**Sales figures:**

****

It is clear that the sales are high in October, November, and December from the above sales figures. Sales has been increasing every year for these months from the earlier years the later years have more sales.

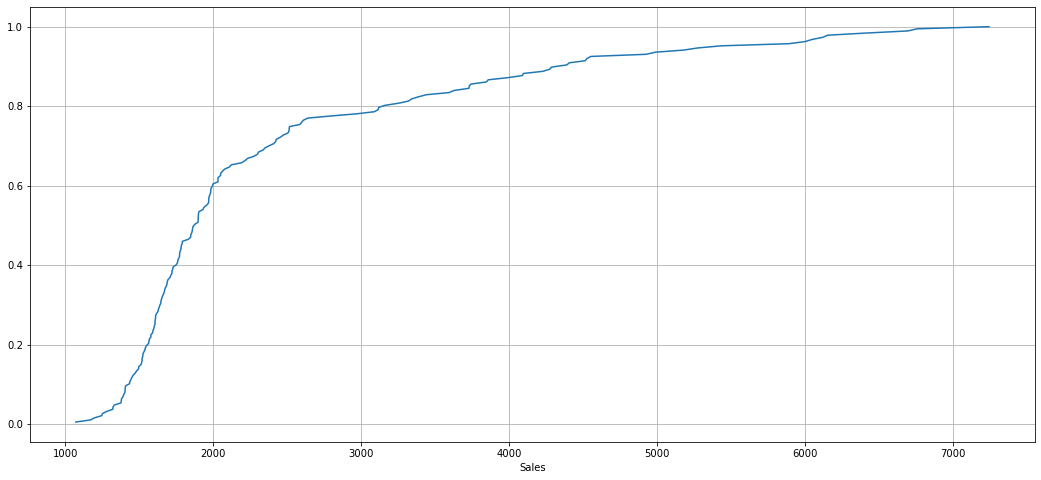
**Sales graph:**

****

We can say that December month sales are the best compared to rest of the months. After 1980 the sales are increasing till 1987. After which the sales have reduced but not drastically. November also has a similar pattern of sales.

After 1986 October also started having better sales in sparkling wine.

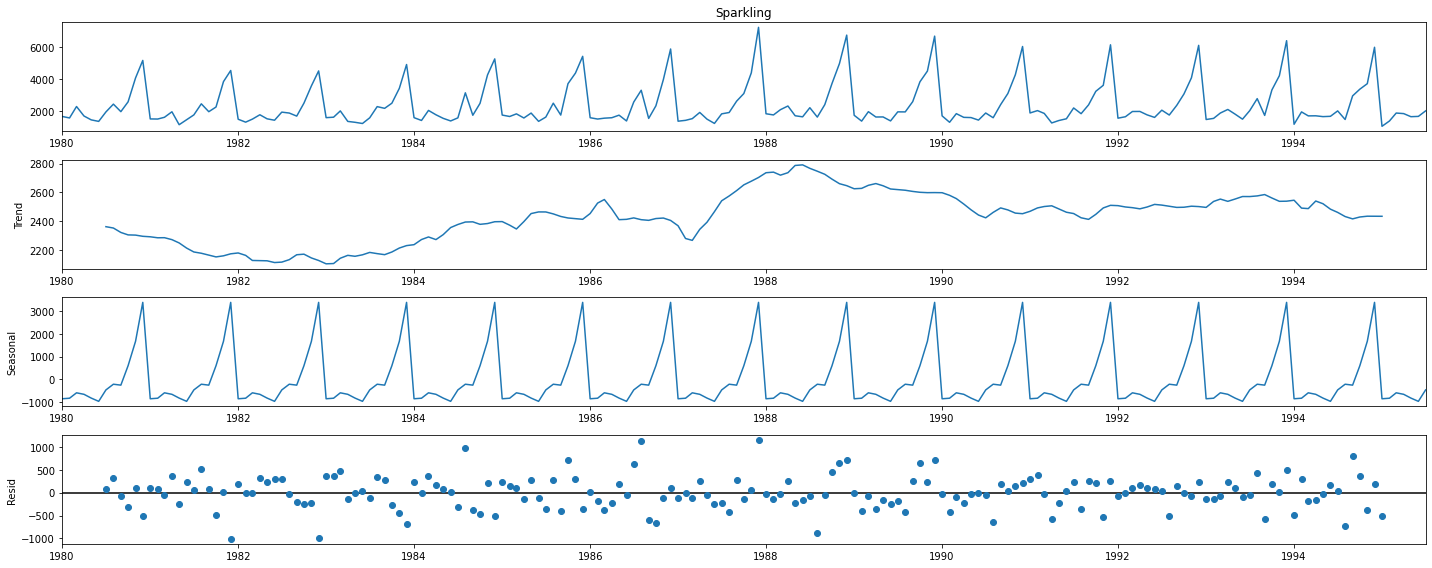
**Empirical Graph:**



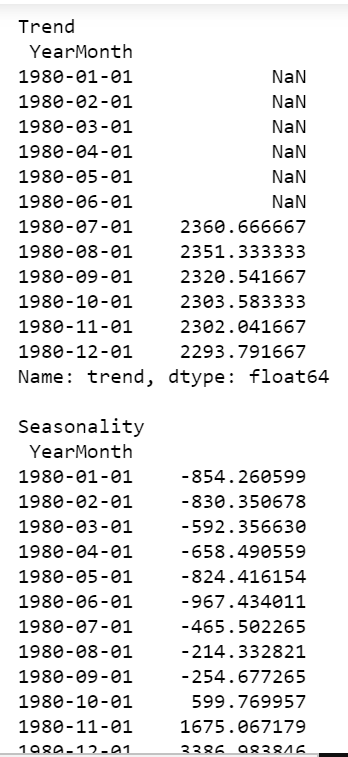
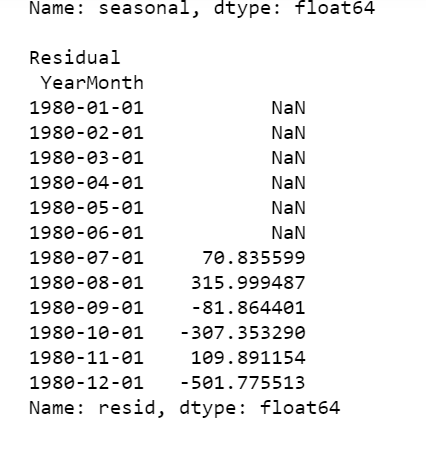
We can see that the first 60% of the data contributed less sales around 2000 but then the sales have picked up and the graph shows us that. 40% of the data provided us the sales from 200 0 to 7000.

**Decomposition:**

**Additive:**

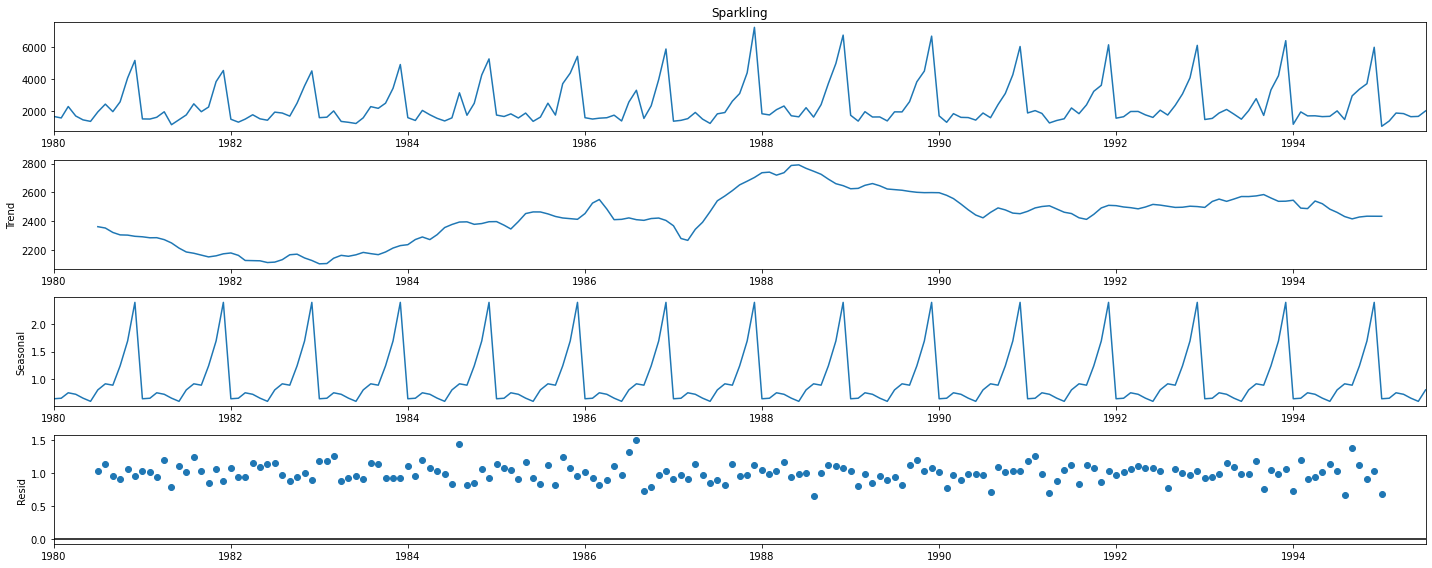
****

We can see here that there is seasonality in the series. And the error is spread across as none of them are in the zero.

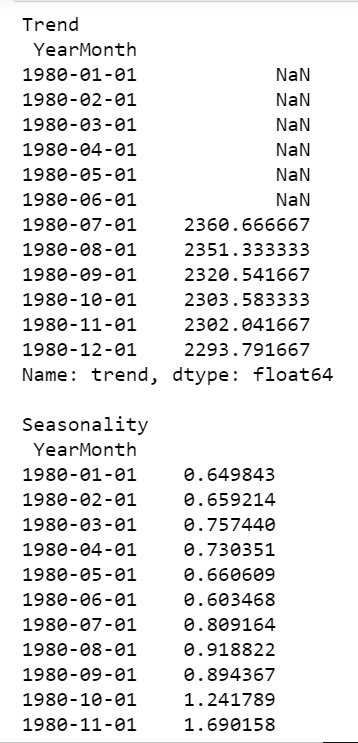
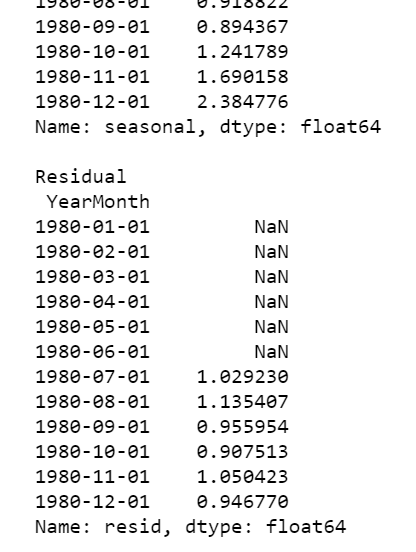
 

Above is the data for the decomposition for the additive.

**Multiplicative:**

****

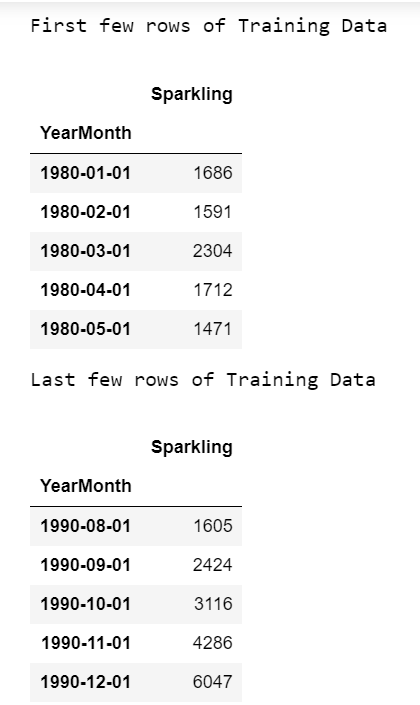
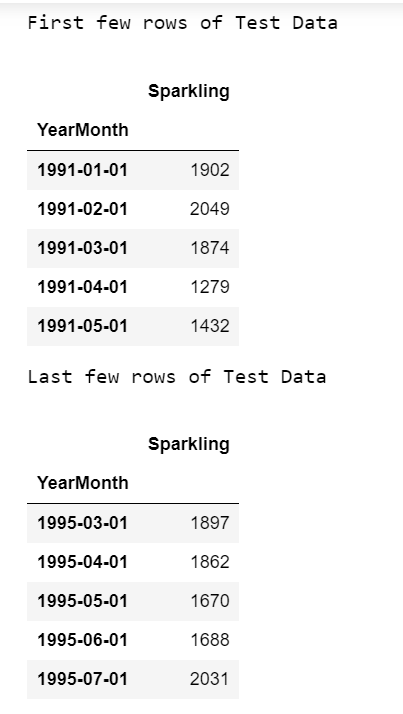
Here also the errors are spread but not as bad as the additive model. There is no trend in the series.

The above is the data of multiplicative decomposition.

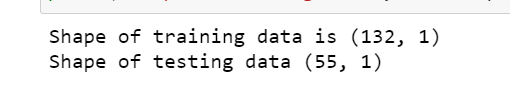
1. **Split the data into training and test. The test data should start in 1991.**

The data is split into training and testing. The training data contains data from 1980 to 1990. And the testing data contains data from 1991 to 1995. Above we can see the head and tail function of both the training and testing data.



Above is the split of the training and testing data. The orange is the testing data.



The training data has 132 records, and the testing data has 55 records.

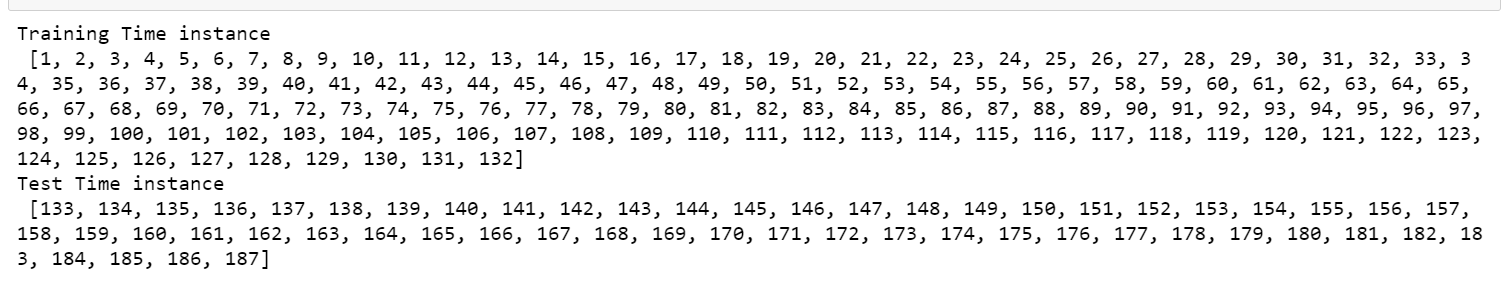
1. **Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.**

We will build the below models:

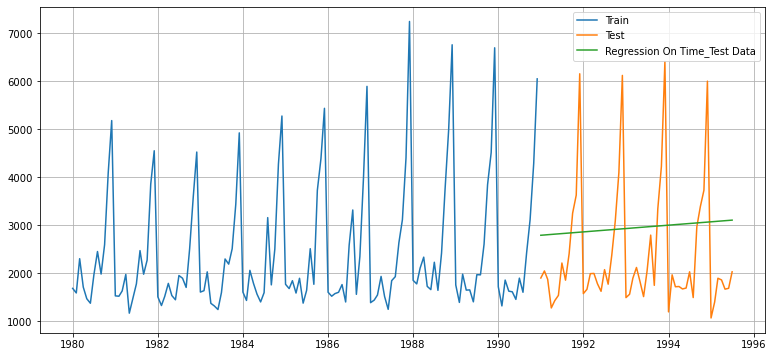
* Linear Regression model.
* Naïve model
* Simple average
* Moving average
* 3 types of exponential smoothing models.

**Linear Regression:**

For linear regression we have split the data into train and test and the number of rows are split based on the train and test.

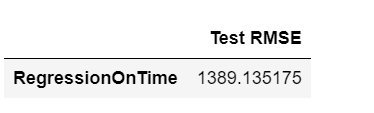


Below we can see the prediction of the linear regression model.

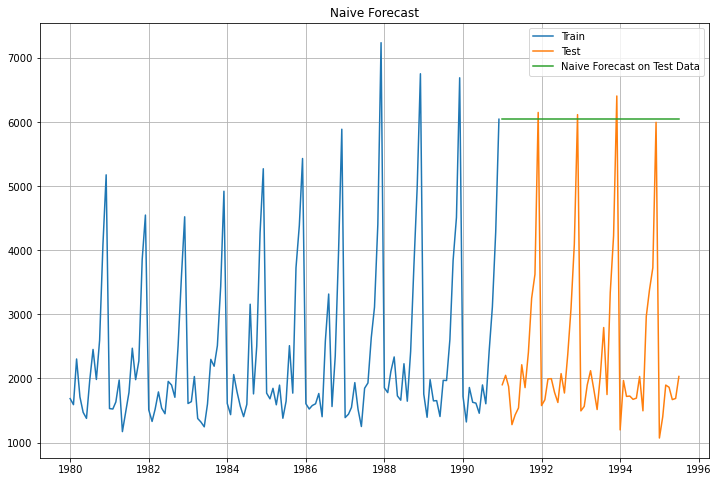


RMSE:

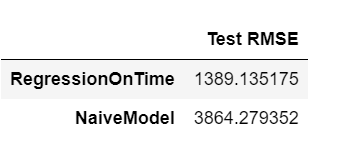
The RMSE for linear regression is 1389.13.



**Naïve Model:**

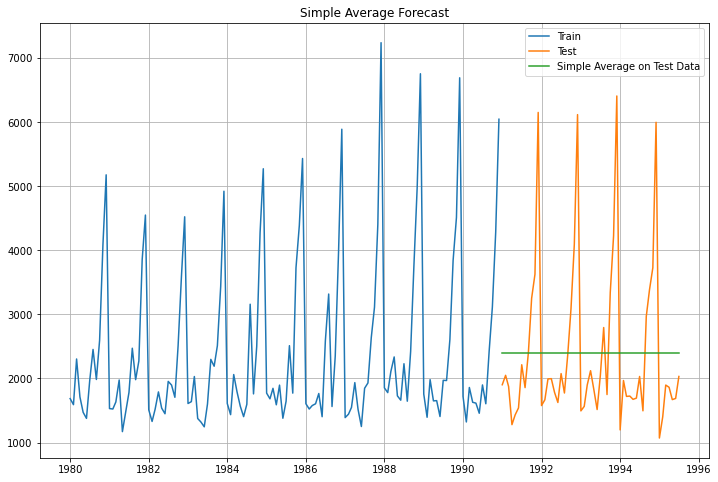


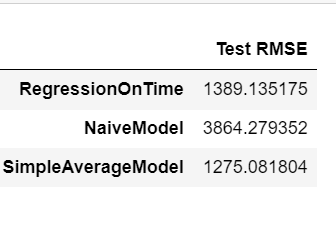
RMSE:



We can see that the RMSE for naïve based model is much higher than the linear regression model.

**Simple average:**

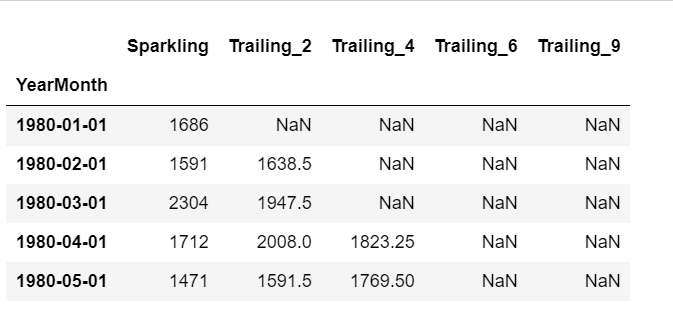
****

****

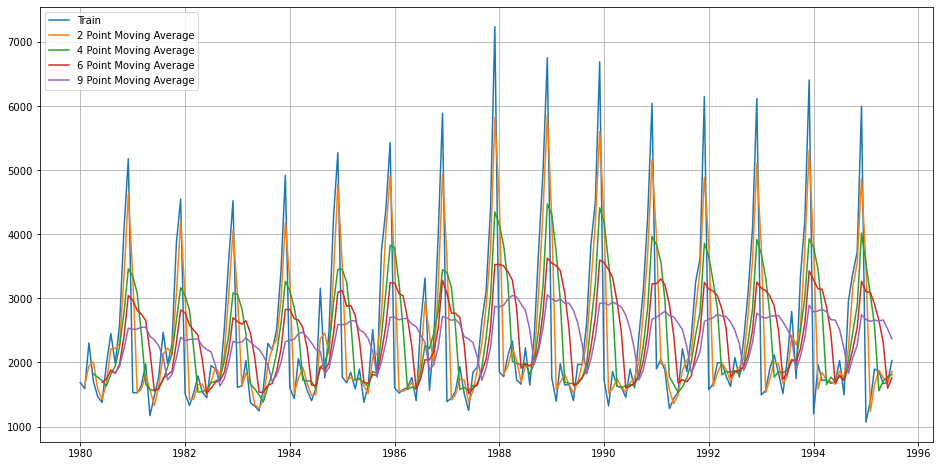
We can see above the graph and the RMSE for the simple average model. This is the best model out of the three models we have performed till now.

**Moving average:**

We will see the moving average of different trailing values.

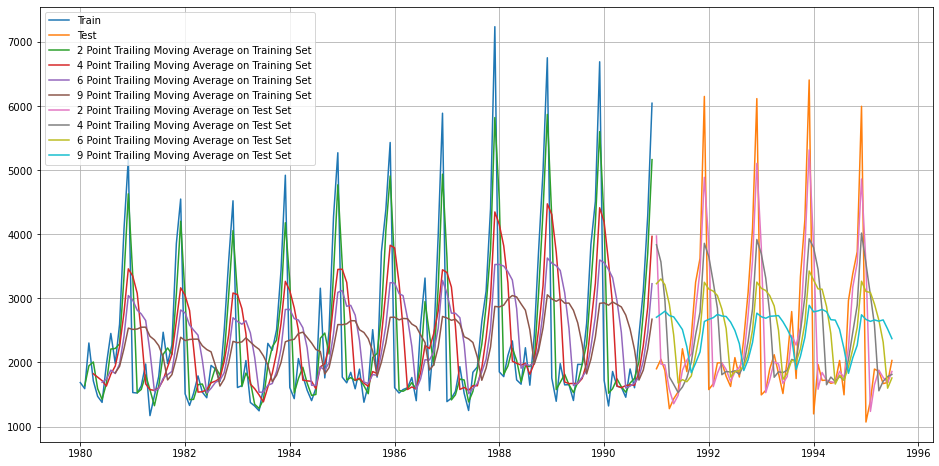


**Training data moving average graph:**



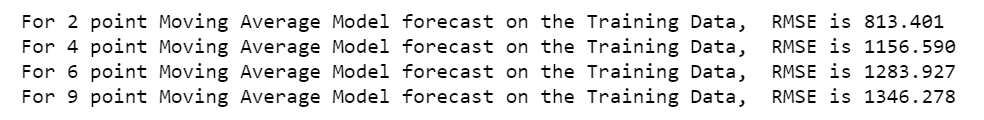
Above we can see the different moving averages graph and their performance with training data.

Testing data and training data moving average:

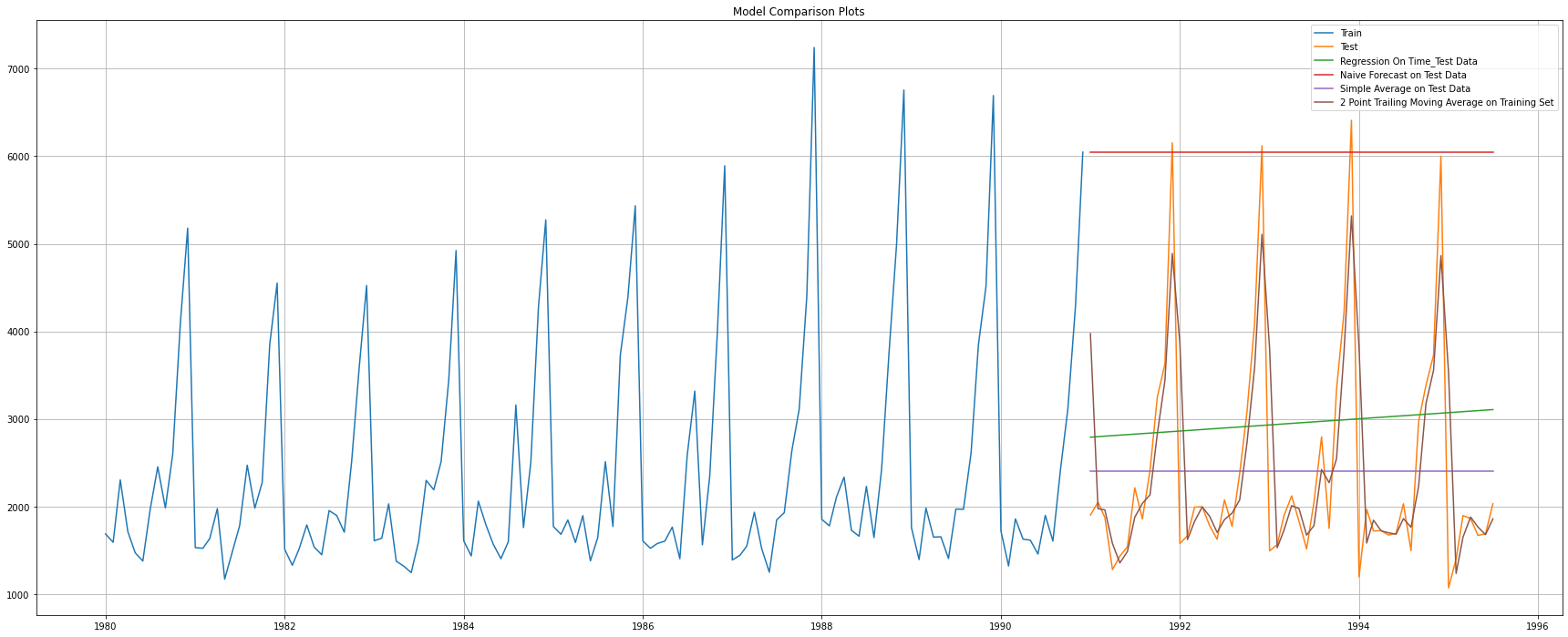


Above is the performance of different moving average on the training and test data compared to the original train and test data set.

RMSE:



The 2-point moving average is having the least RMSE of 813.401.

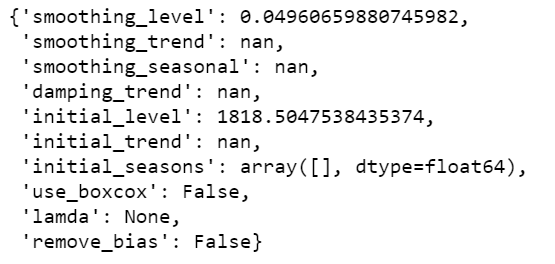


Plot that shows the performance of all model on the test data. We can see that the 2-point moving average is performing best among the linear, naïve, and simple average models.

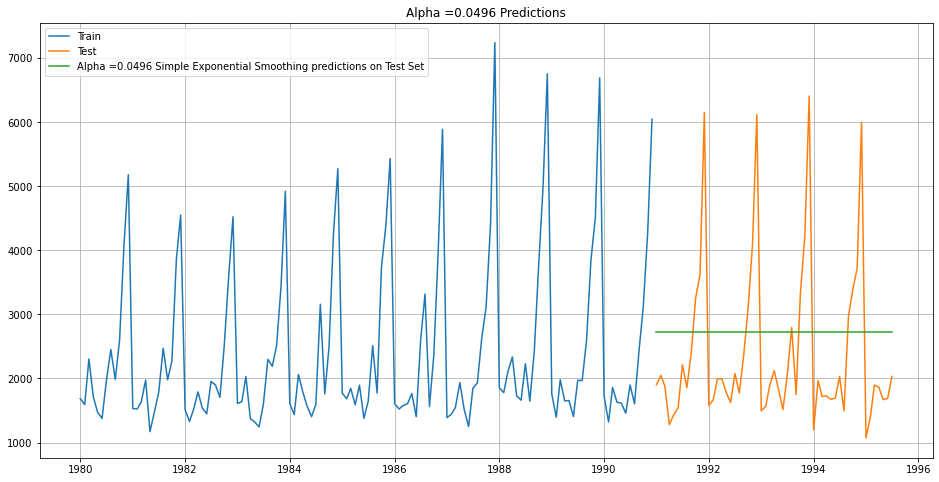
**Exponential smoothing models:**

**Simple smoothing:**

we use the fit function to find the best param for the simple smoothing. The simple smoothing finds only best level in the series.



The level provided by the model is 0.0496.



Based on the given level we forecast, and the green line is the forecasted line. This model captures only the level in the series.

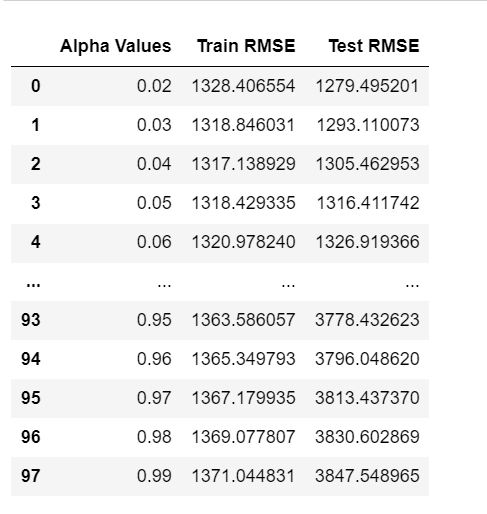
RMSE:



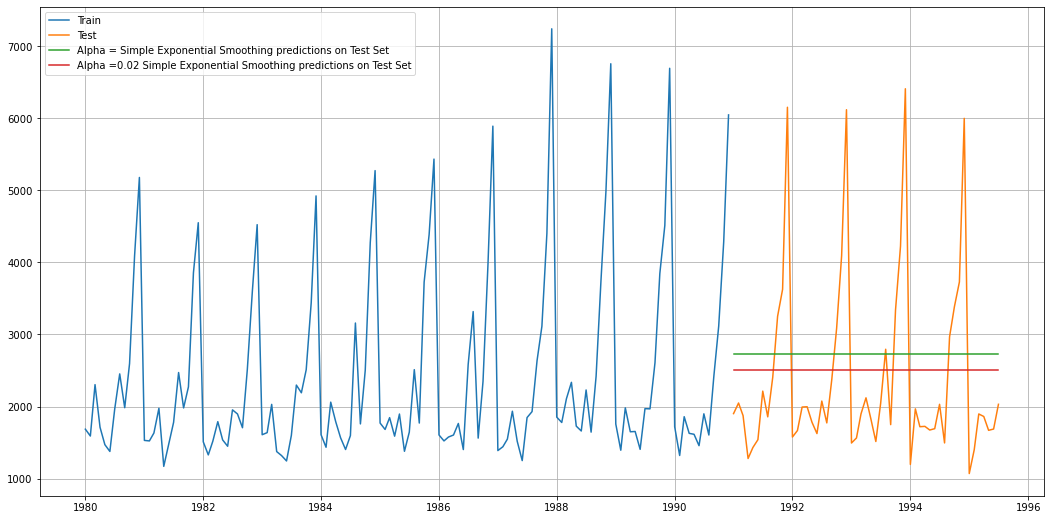
**Trying different iteration for the simple smoothing model:**

Here we are setting the alpha value from 0.2 to 1 and the model will increment each value by .1.

From the different alpha values below are the best values found by the model.



We can see that when alpha is 0.20 the test RMSE is 1279.49.

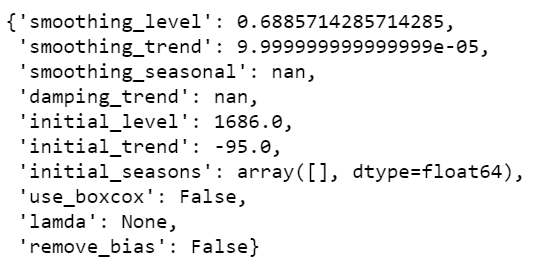


Above is the graph that shows both simply smoothing representation with the test data.

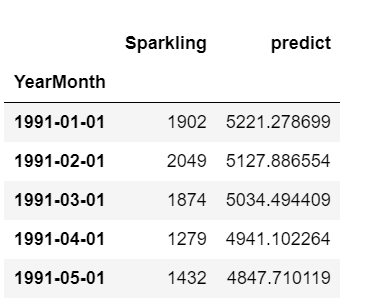
**Double Exponential smoothing:**

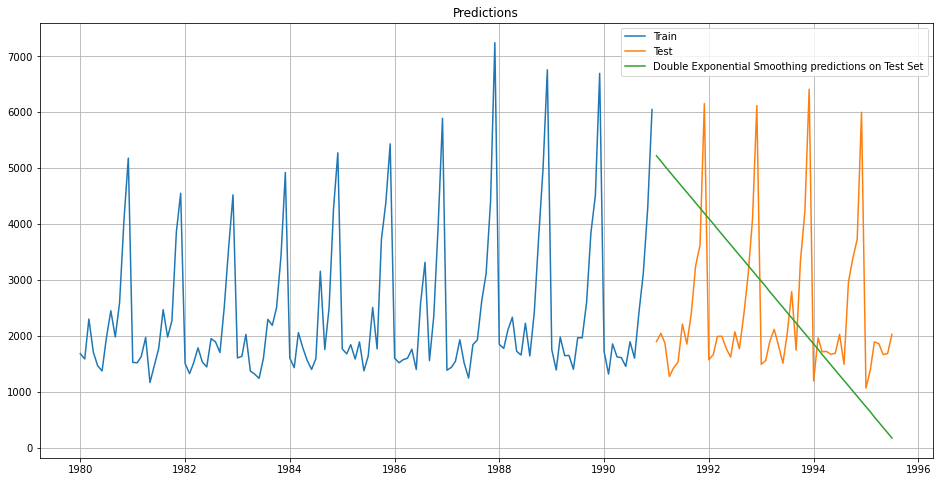
This model gives us the level and the trend in the series. Level is alpha and trend is beta.

We can see below the model when we use the fit function gives us alpha of 0.688 and trend as 9.99 e-05



Predicted values:





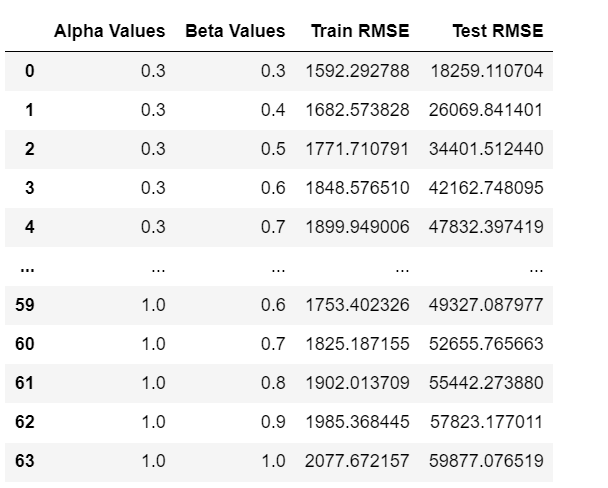
Above we can see the prediction using the alpha and beta we got earlier from the model.

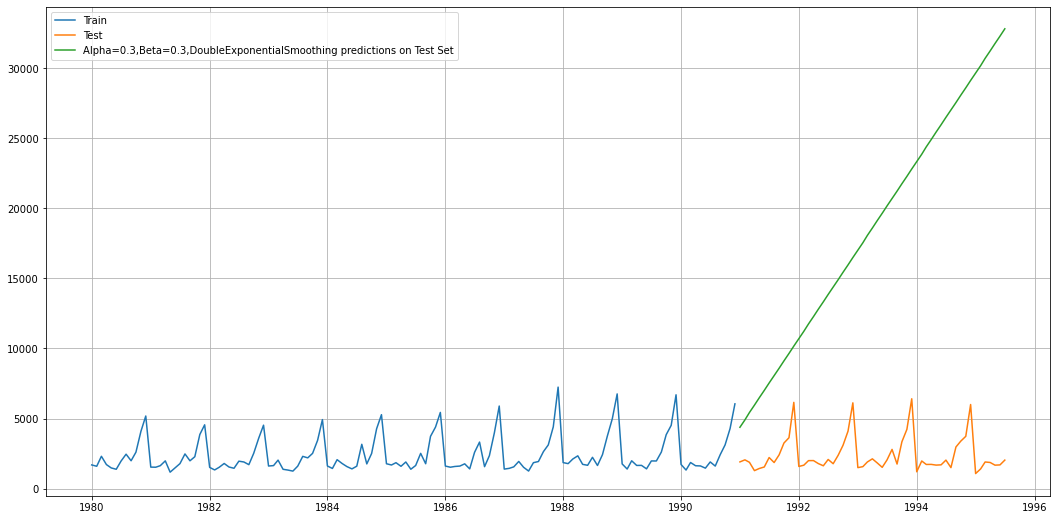
RMSE is : 2007.239.



We can see if this model can give a better RMSE using different alpha and beta values.

I have used the alpha values from 0.3 to 1.1 and beta values of 0.3 to 1.1 to see if we can get a better test RMSE and below are the results of it.

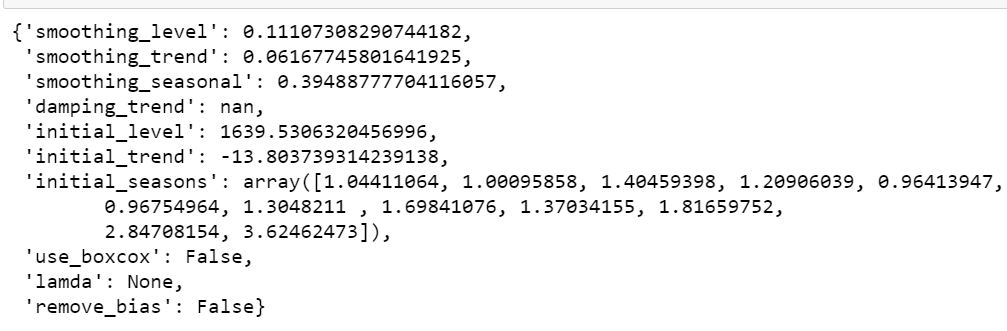




The RMSE for this model is very high. We will continue forward checking with triple exponential smoothing and check the RMSE.

**Triple Exponential smoothing:**

This model captured the level, trend, and seasonality in the data.



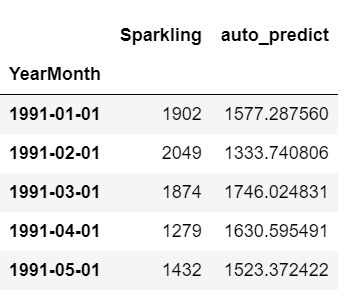
Above are the parameters we get when we use the fit function and the best param is given.

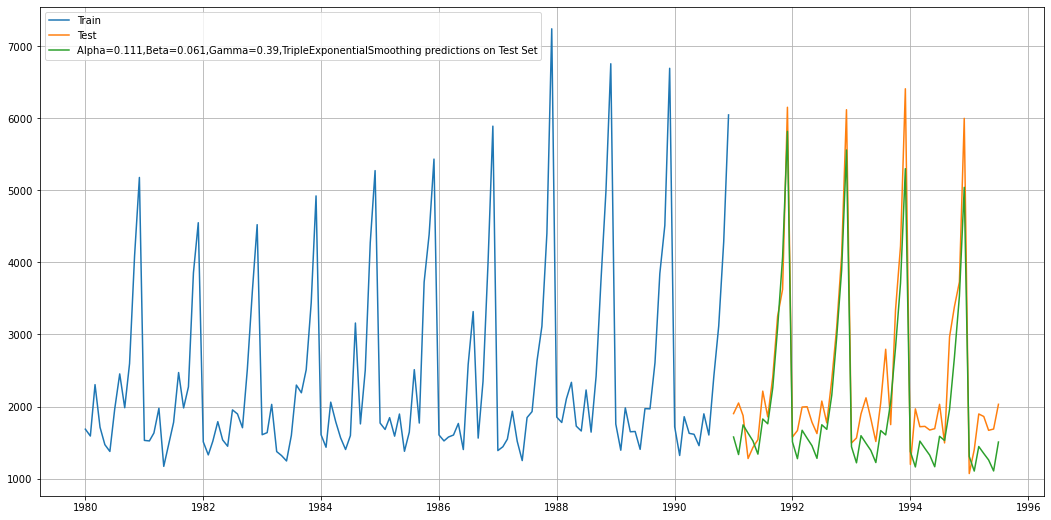
The level is 0.111. This is the alpha.

Trend is 0.06161.This is the beta.

Seasonality is 0.3948. This is the gamma.

**Predicted Values:**





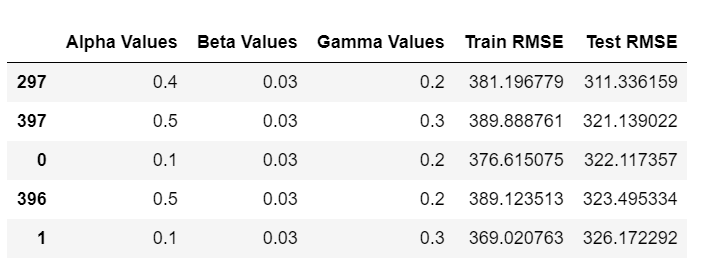
Using the best param we can see that we have a better model compared to all the model we have done till now. We can see the predicted values the green line is almost close to our test data. This shows that the prediction is better compared to other models. We will see if we can improve this by changing the alpha, beta, and gamma values.

RMSE: 469.432.



**Different parameters:**

I am setting the alpha values from 0.100 to 1. Beta values also from 0.03 to 1 and gamma value from 0.2 to 1. Let’s run the model with these parameters and see if we can reduce the RMSE.

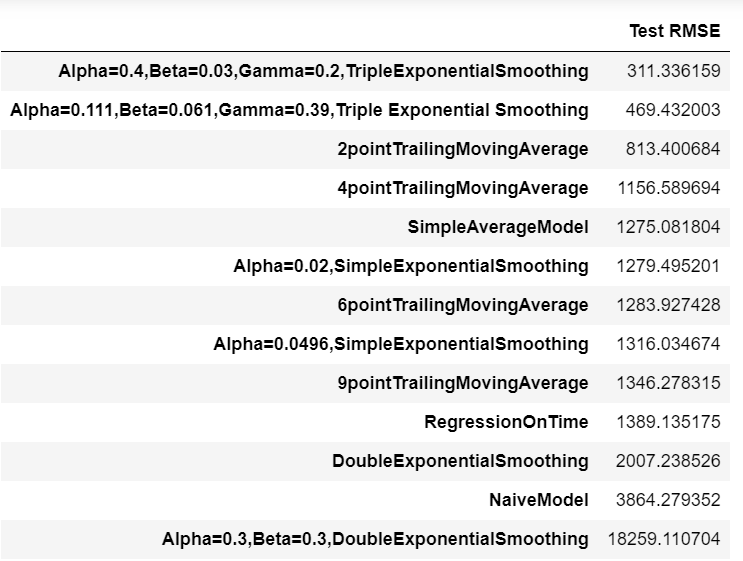


We can see that the RMSE we got is 311.336. for alpha as 0.4 beta as 0.03 and gamma as 0.2.

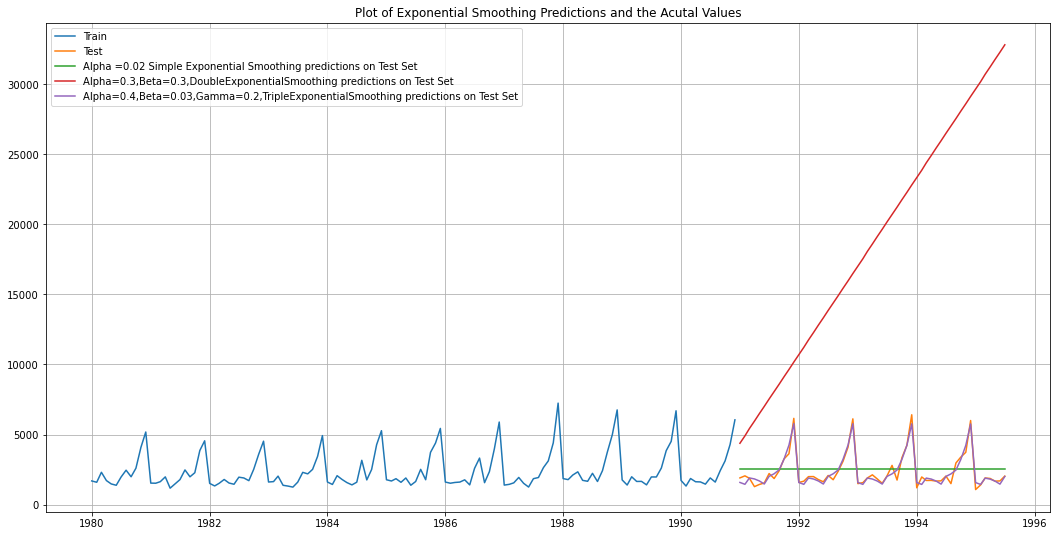
This is the best model so far, we have found with the least RMSE of 311.336.

**Performance on TEST data:**





**ALL model plot:**



We can see the purple line is performing the best that is it is almost similar to the test data.

1. **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.**

**Hypothesis for stationarity:**

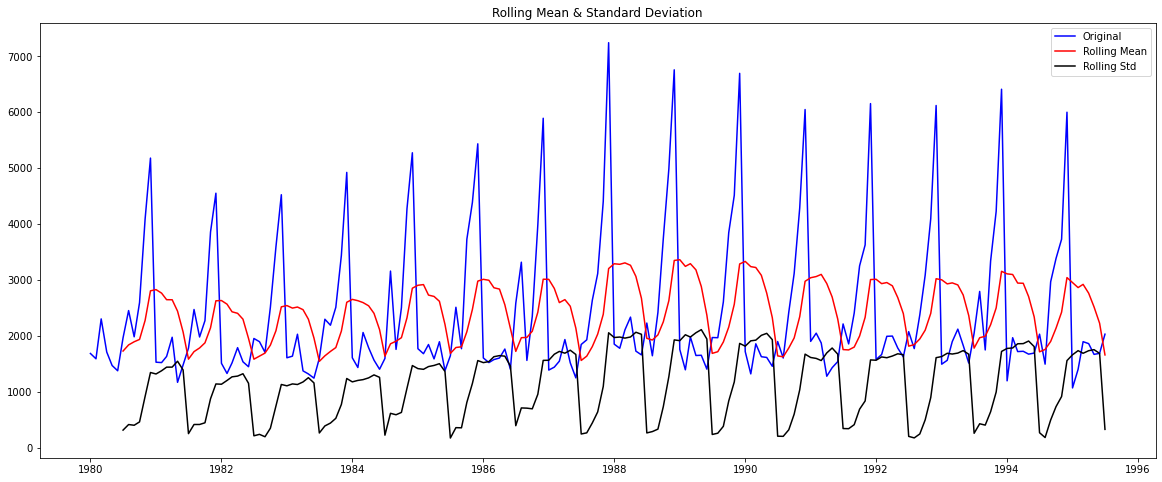
**HO:** The series is nonstationary

**H1:** The series is stationary.

The stationarity is checked for 0.05 significance.

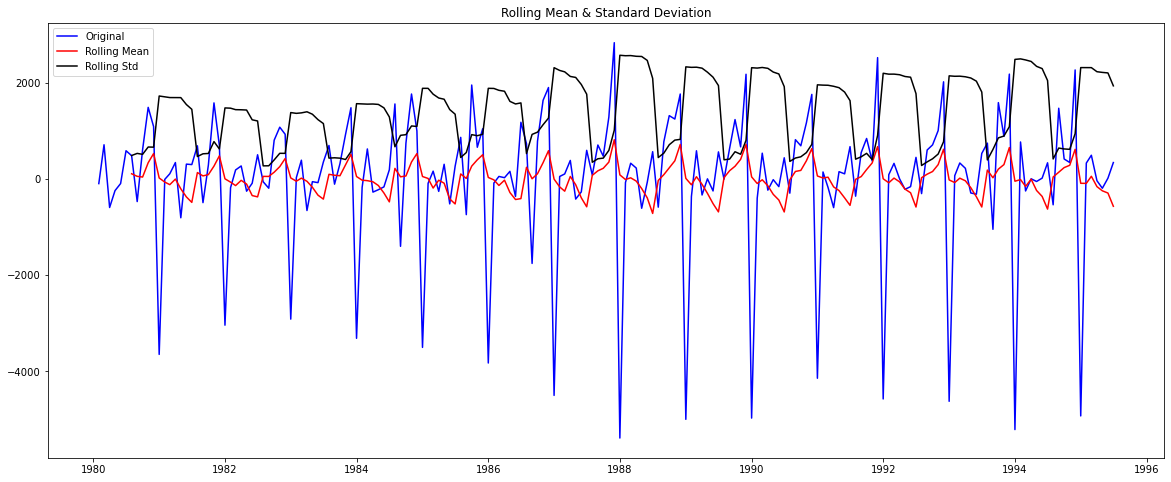
When the P value is less than 0.05, we accept the alternate hypothesis.

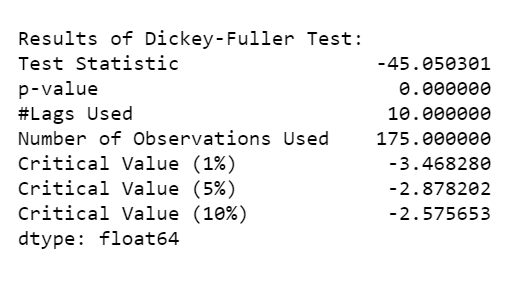
We use the ADF test to check the stationarity of the series and it found that the model is not stationary.





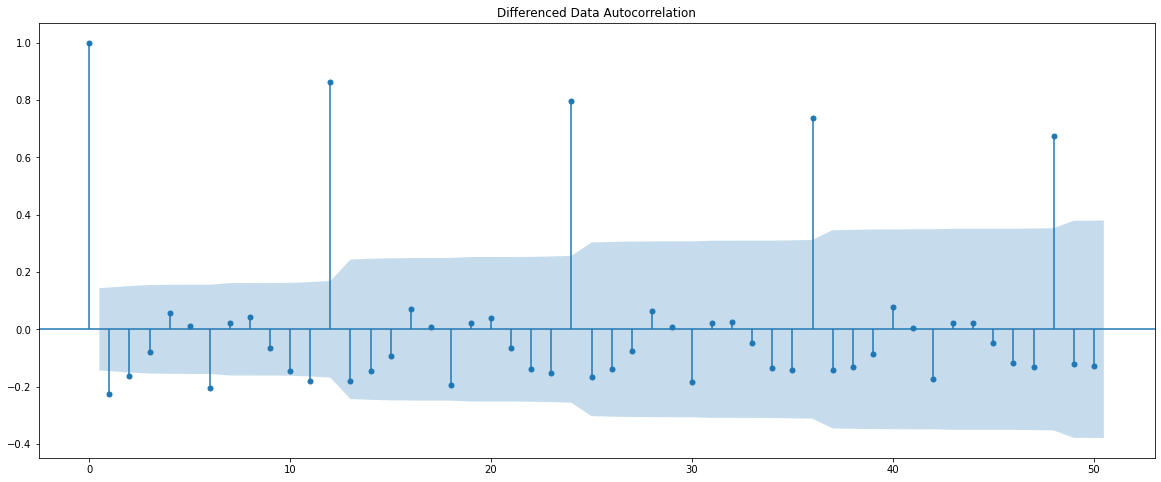
To make the series stationary we have to take difference and see and what difference the series is becoming stationary.

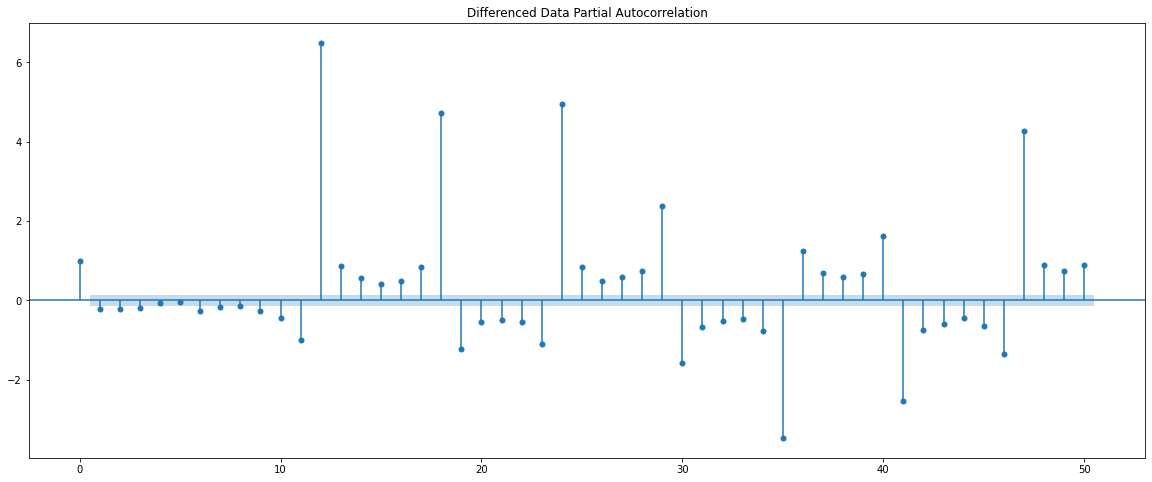




When using the difference as 1 we have got the P value of 0 which is very less to 0.05 therefore, we can reject the null hypothesis of the series is nonstationary. With taking a difference of 1 our series is stationary now.

1. **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.**

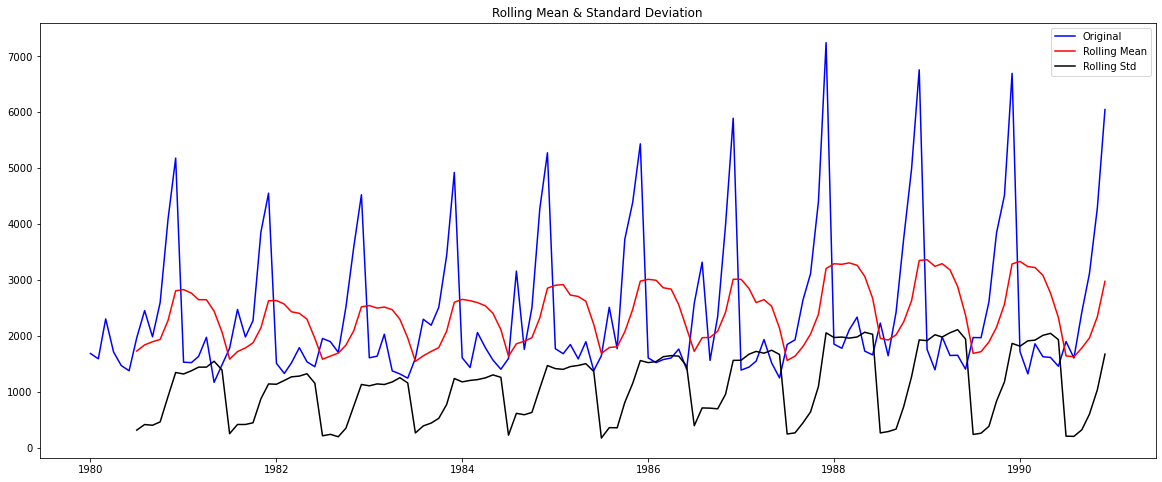


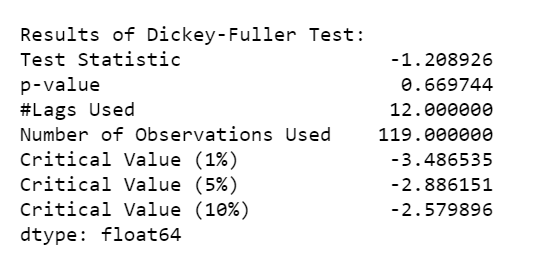


Above are the ACF and PACF plots. This will help in determining the values of P and Q in the arima and sarima parameters. However, we will build an automated model of the sarima and the arima for this task.

We have to define a range of values for p and q based on the above plot and I have defined the values for a range of 0 to 3 for arima and the model will run with all the values in the range for p and q and give us all the combinations from 0 to 3 along with the value of d. The d value I have selected here is from range of 1,2. Since we got the stationarity using the difference of 1 the d I have selected as 1.

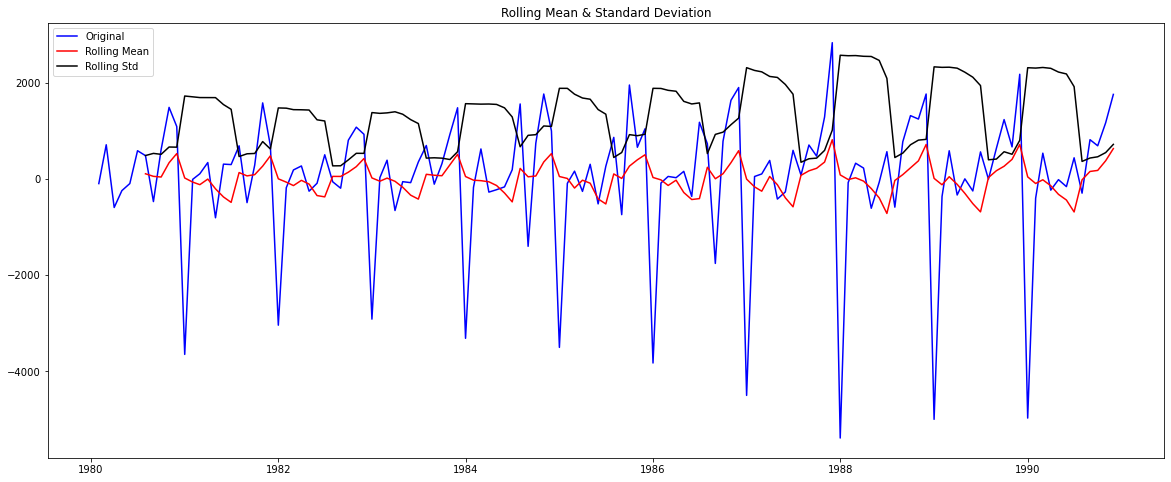
**Checking stationarity on the training data:**

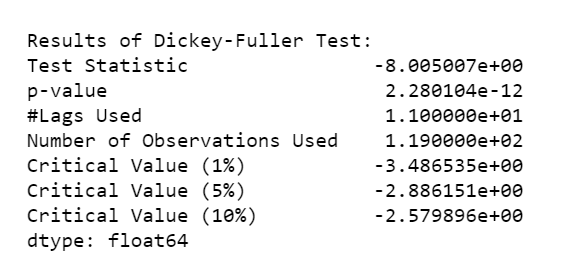
****

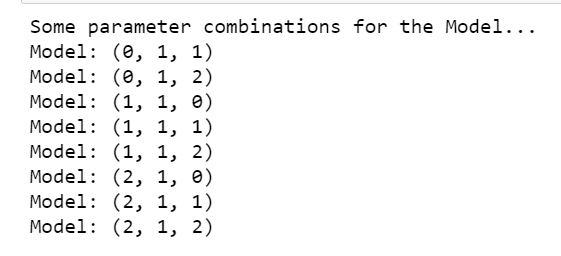
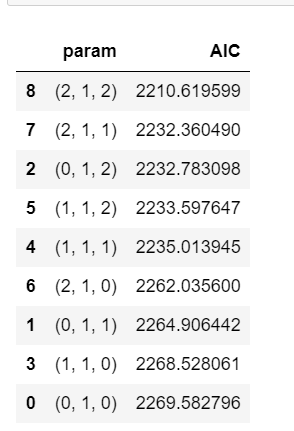
****

The series is not stationary since the P value is more than our significance of 0.05. So we will take a difference and make series stationary.

**Taking a diff of 1 to make the data stationary:**

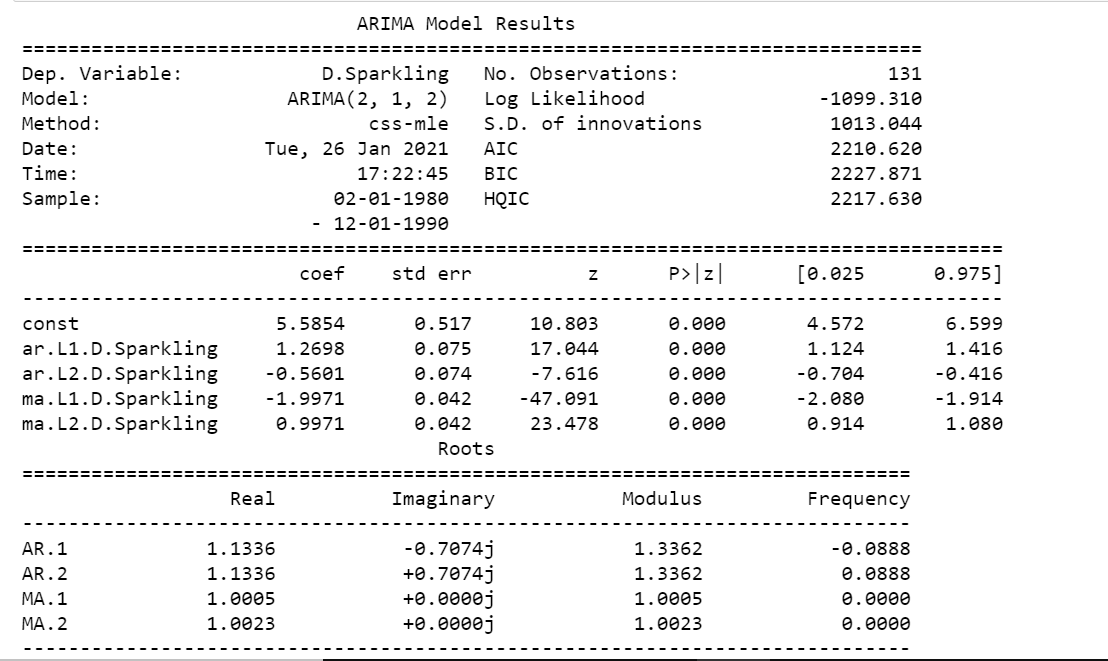
****

****

**** ****

Above is the some of the parameter combination and the best AIC value for the parameters.

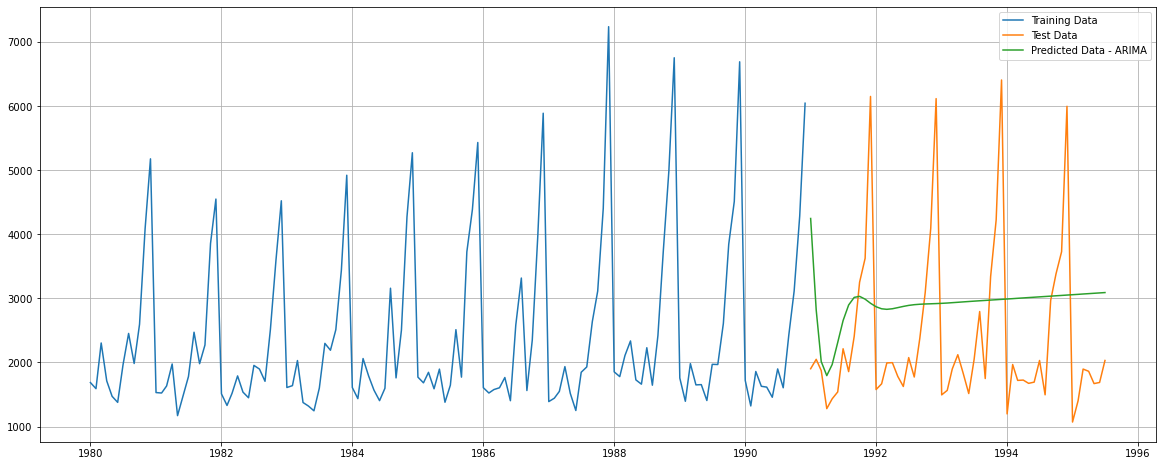
Above we can see the automated model has given us different combination of p,d,q values from the range of values we had assigned. And from those combination the best values are 2,1,2 and the AIC for this combination is 2210.65.

****

All the constants have p value as significant.

**Automated ARIMA RMSE:**





Above is the predicted arima in comparison with the test data.

**Automated SARIMA**:

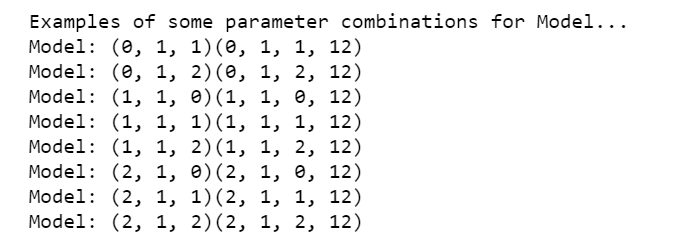
Sarima calculates the seasonality in the series. Since our series has seasonality, we will be doing the SARIMA in 12- and 24-months basis.

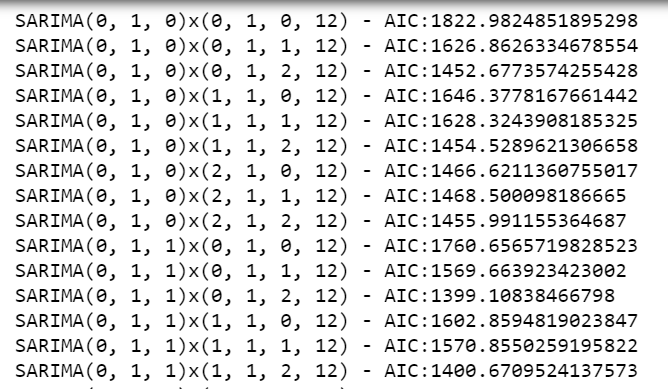
**12 months SARIMA:**

Like we did for ARIMA sarima also has the PDQ vlaues to be entered. And based on the above ACF and PACF plots we have given a range of values for P and Q and for D also.

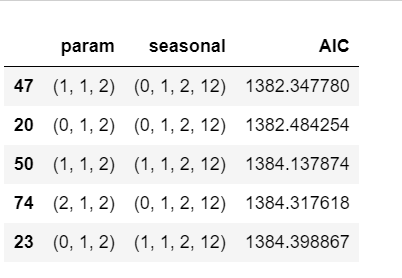
Like we did for ARIMA sarima also has the PDQ vlaues to be entered. And based on the above ACF and PACF plots we have given a range of values for P and Q and for D also.

We have give the range for p,q from (0,3) and D range from (1,2) and d range is (1,2)

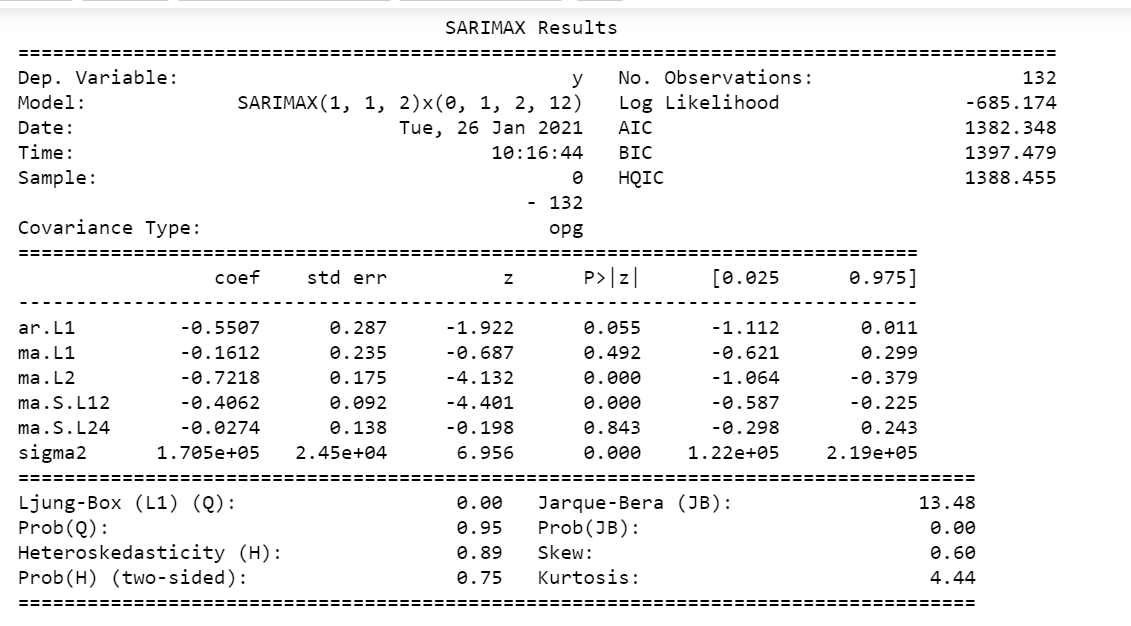




Then the SARIMA model combines different combinations and find the least AIC values from the range of values.

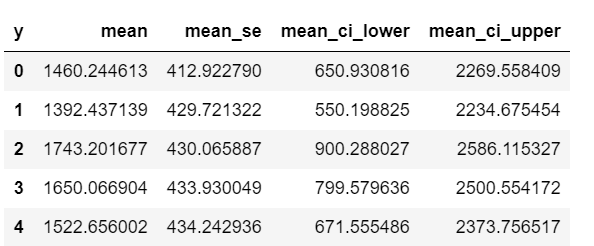


These are the best combination with least AIC



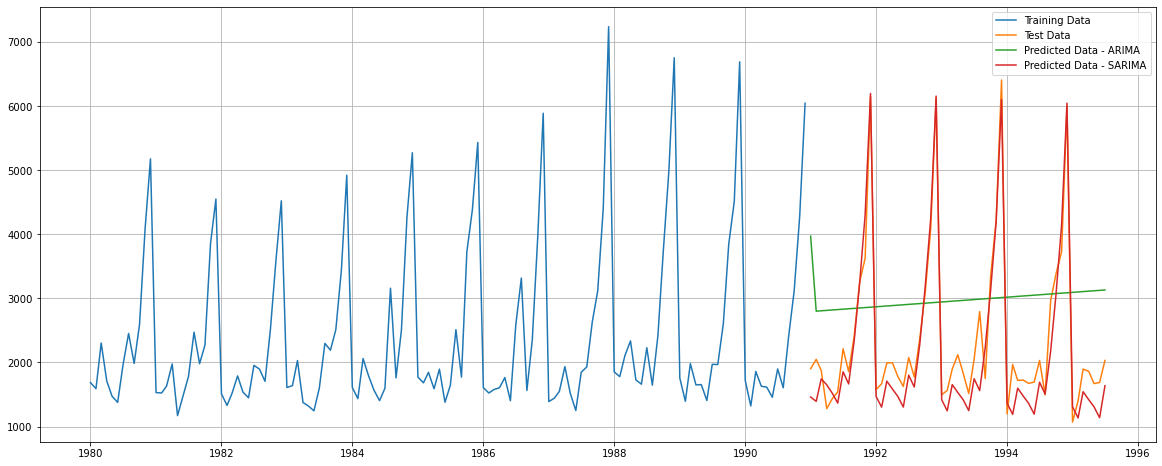
This table shows the different constant values. And based on P value we can know which all are significant if the P value is less than 0.05 then we can conclude those constants are significant.

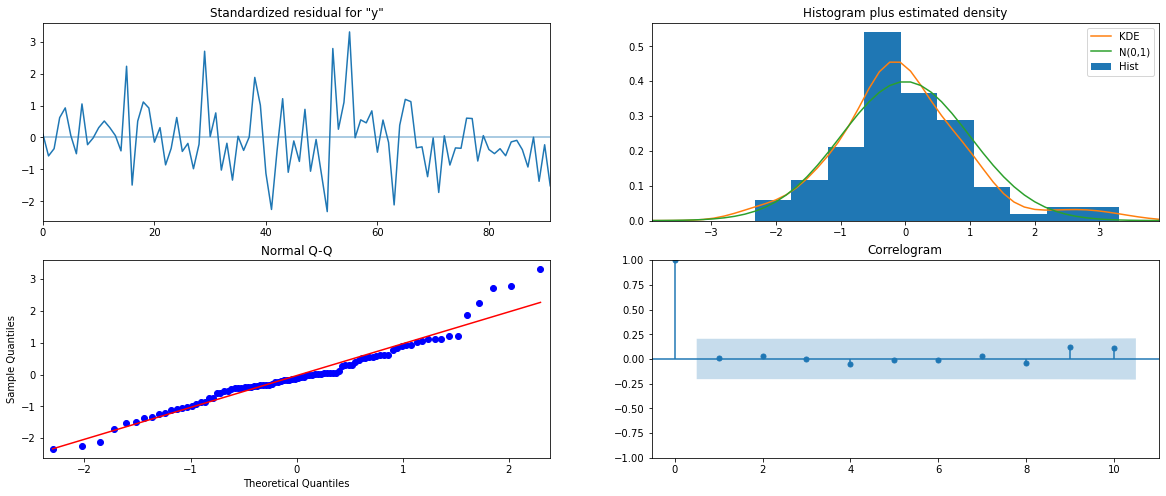
**Predicted mean for lower and upper confidence interval:**



RMSE SARIMA 12 months:



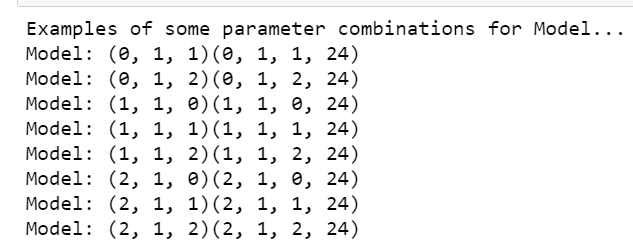


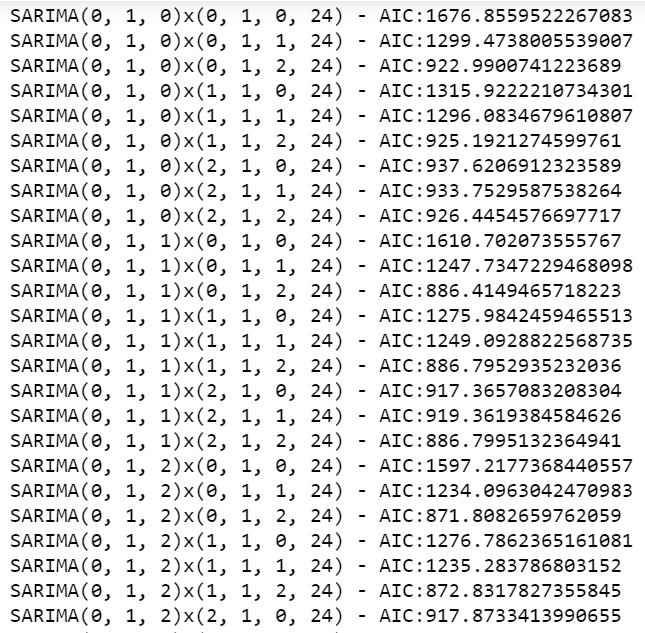


**SARIMA 24.**

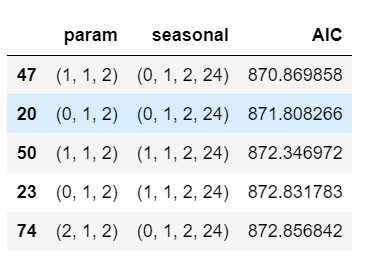
We will also see how the model performs with SARIMA as 24 months.

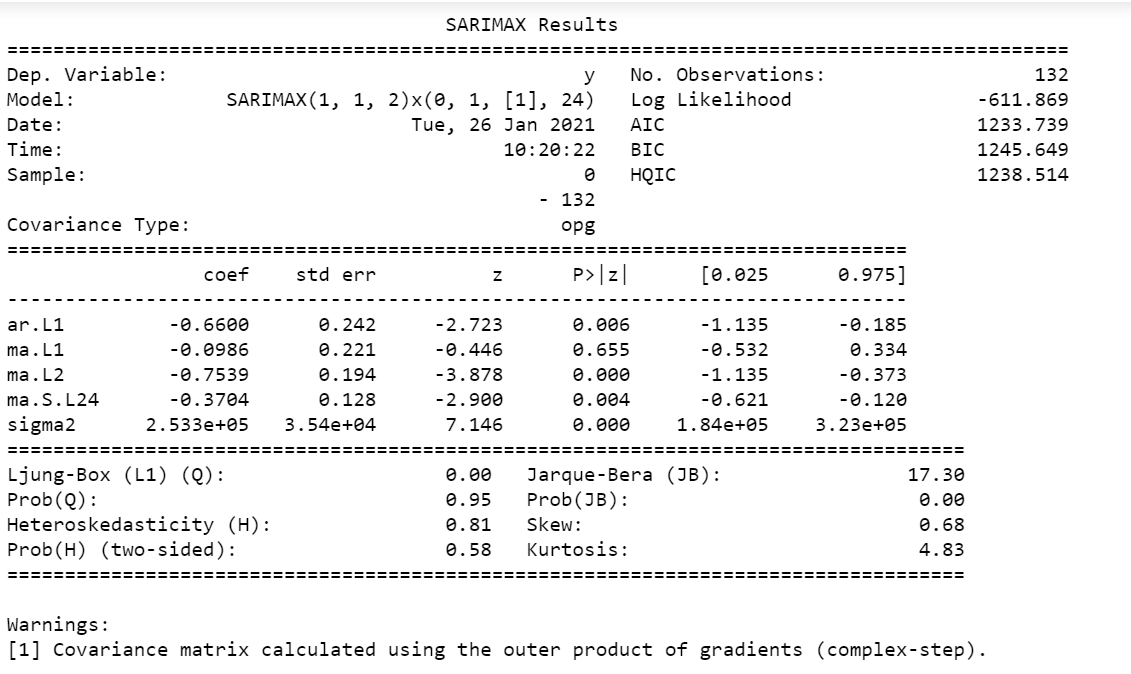
**Parameter combinations:**

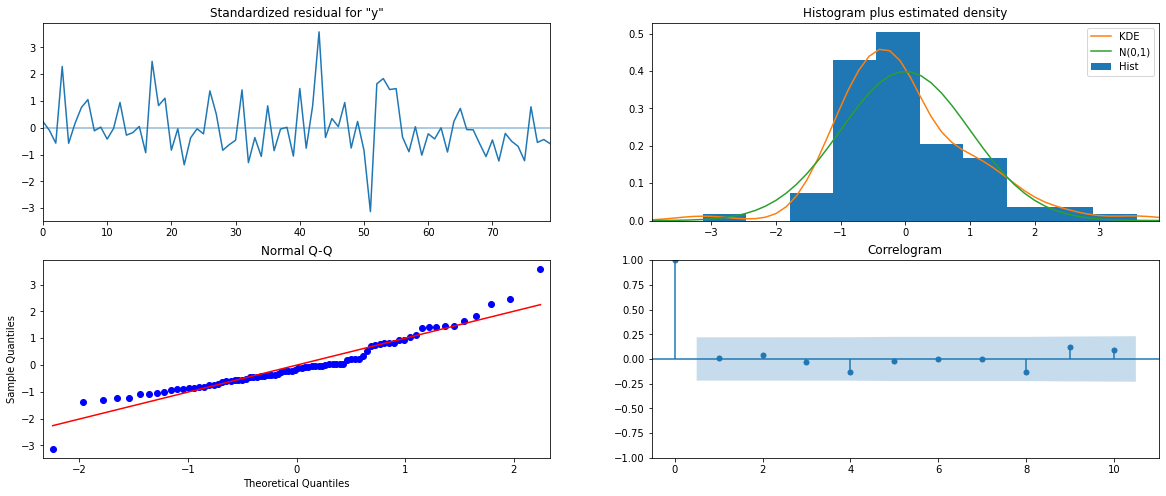


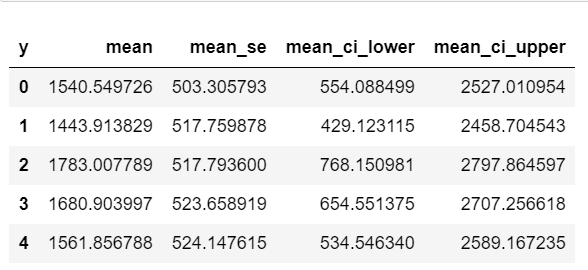


Best combination that gave least AIC for SARIMA as 24 monhs:





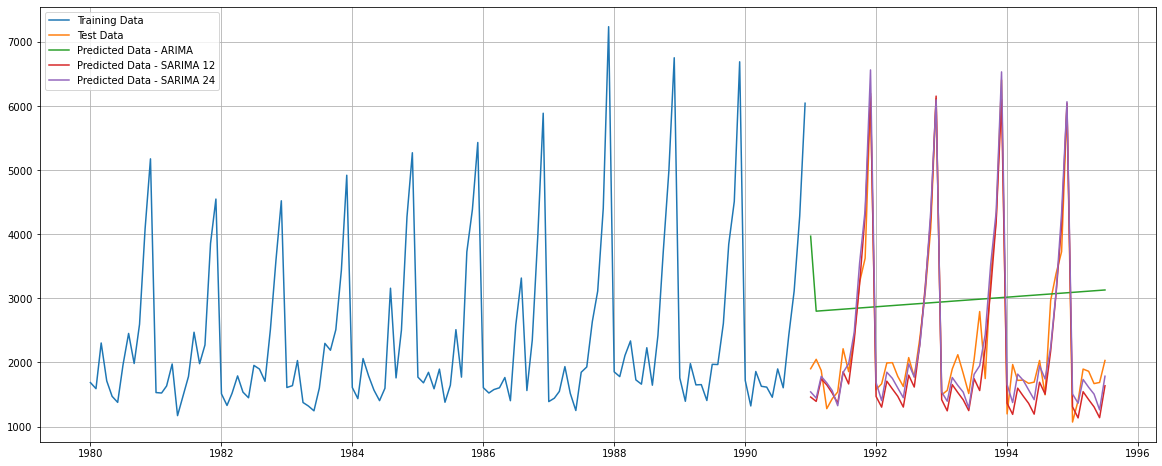




**RMSE SARIMA 24:**

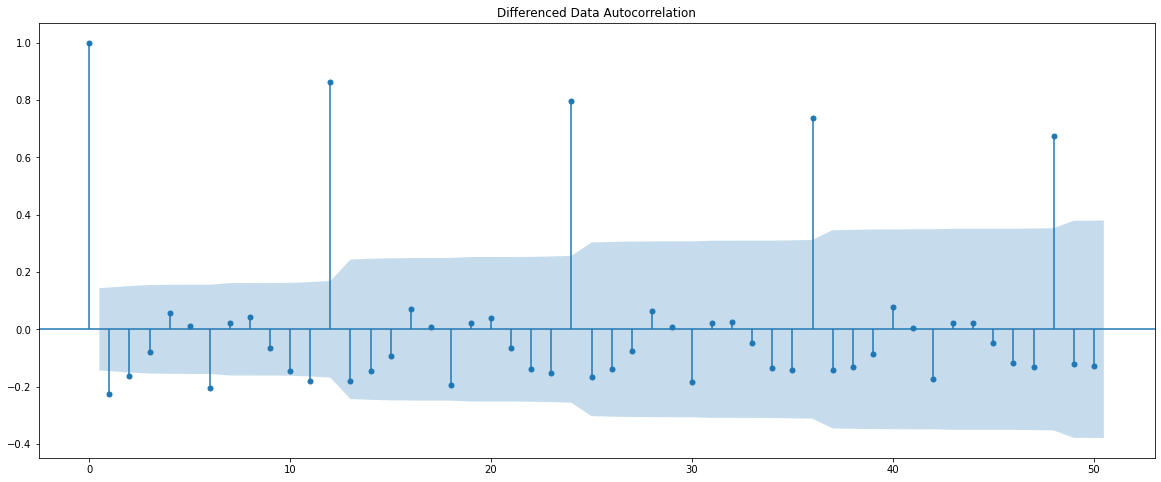


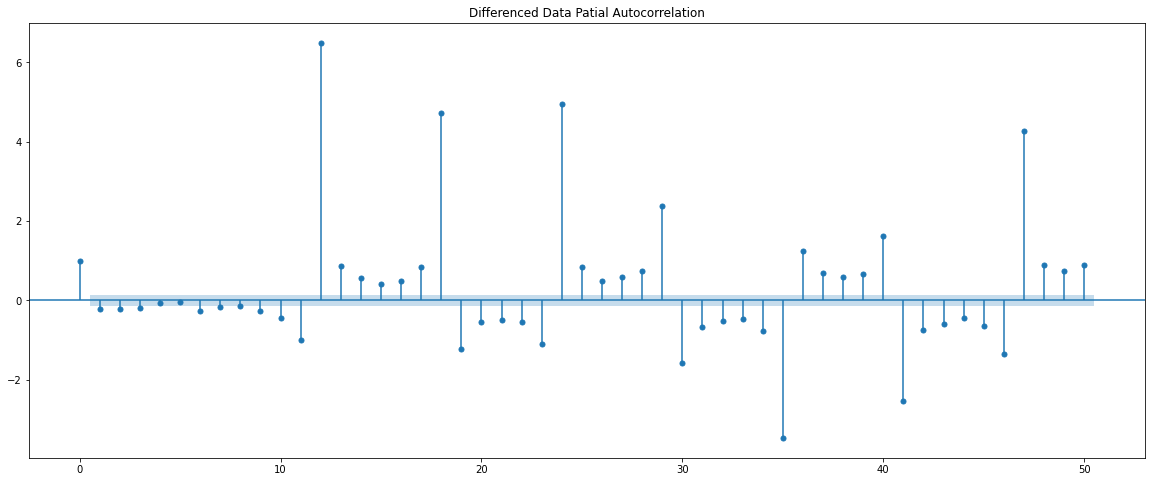
**All the models forecasted graph:**

****

1. **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.**

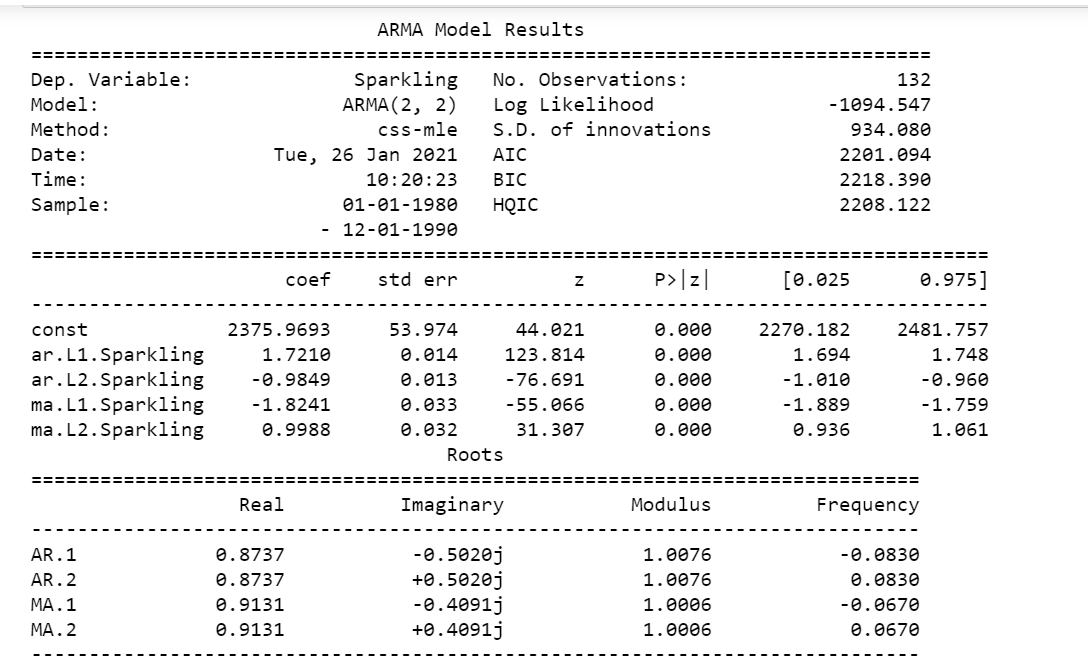
[Manual ARIMA and SARIMA:](#_top)

****

****

Above we can see the ACF and PACF graphs. The ACF gives us the q value and PACF gives us the p value. As per the graph the acf we can see after two lags the 3rd lag is inside the confident interval band. So, q will be 2.

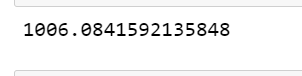
As per the PACF graph after the 3rd lag the 4th lag is inside the confidence interval band. So as pe the graphs the p value is 2 and q is 3. However, I am considering p as 2 and q as 2 for the manual arima calculation. And d as 1.



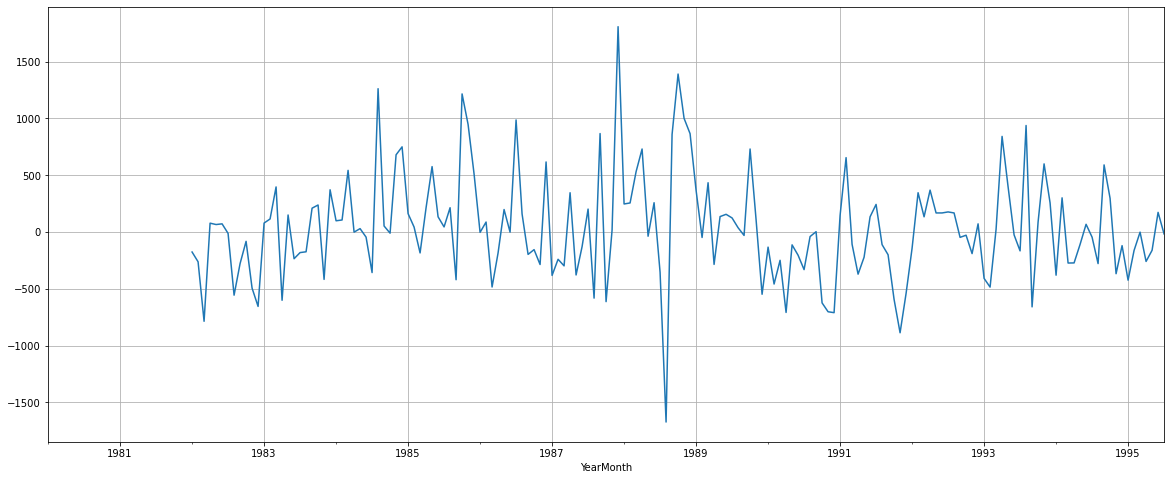
We can see that all the constants are significant because their P values are less than 0.05.

**RMSE ARIMA manual:**

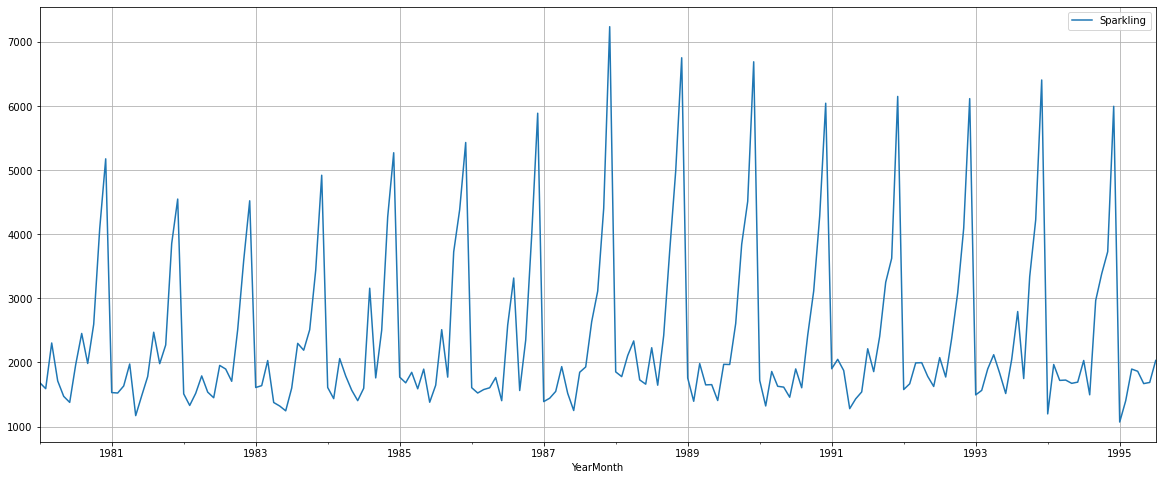
The RMSE is



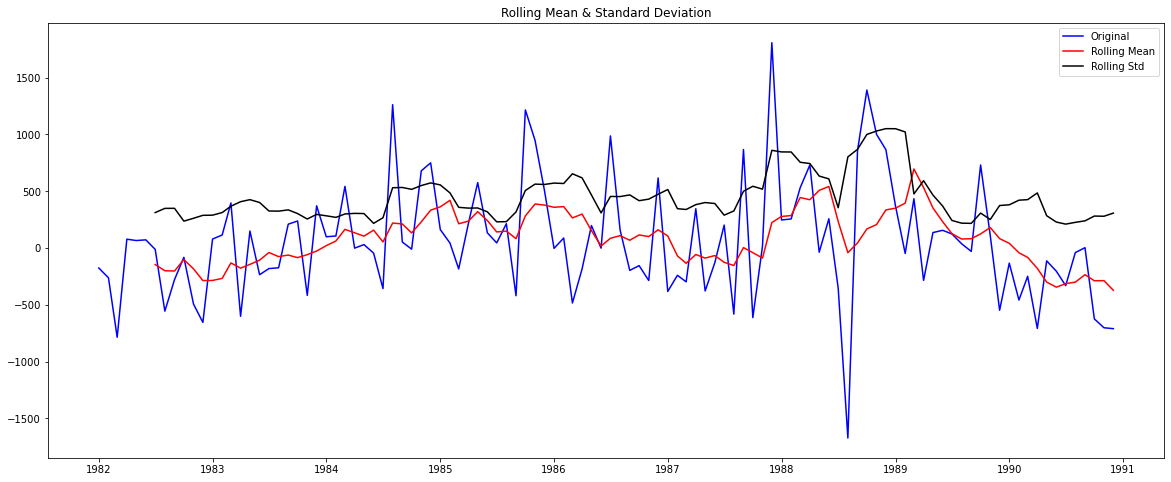
**SARIMA Manual:**

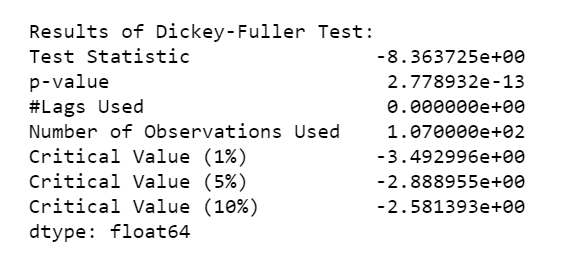


In the above graph we have used the difference of 24 seasonality which we found earlier as the best out of the 12- and 24-months seasonality. This is done to make the series stationery and if there are any trend it will eliminate them.



Above is the original series.





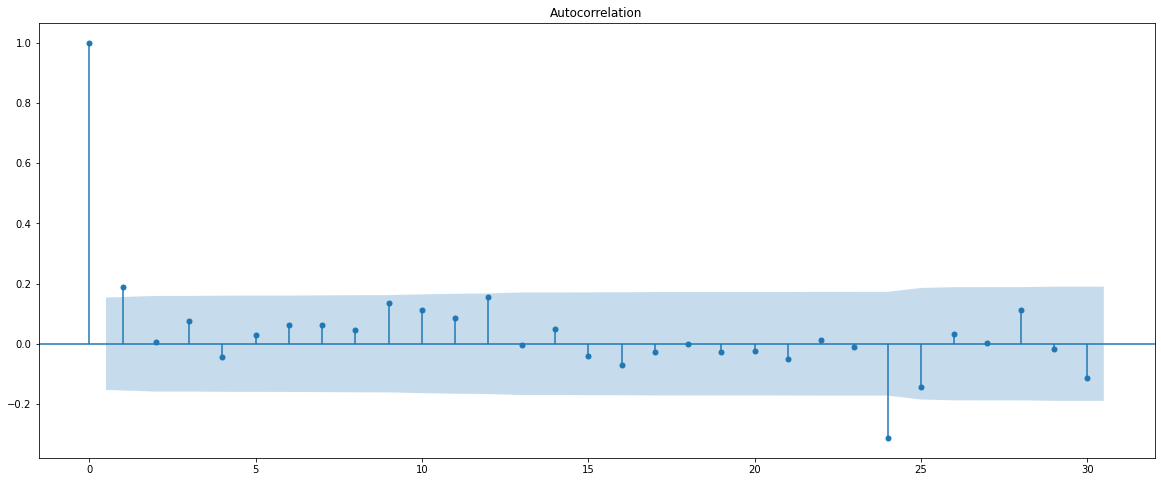
On the original series will diff as 24 we get a p value is 0.005.

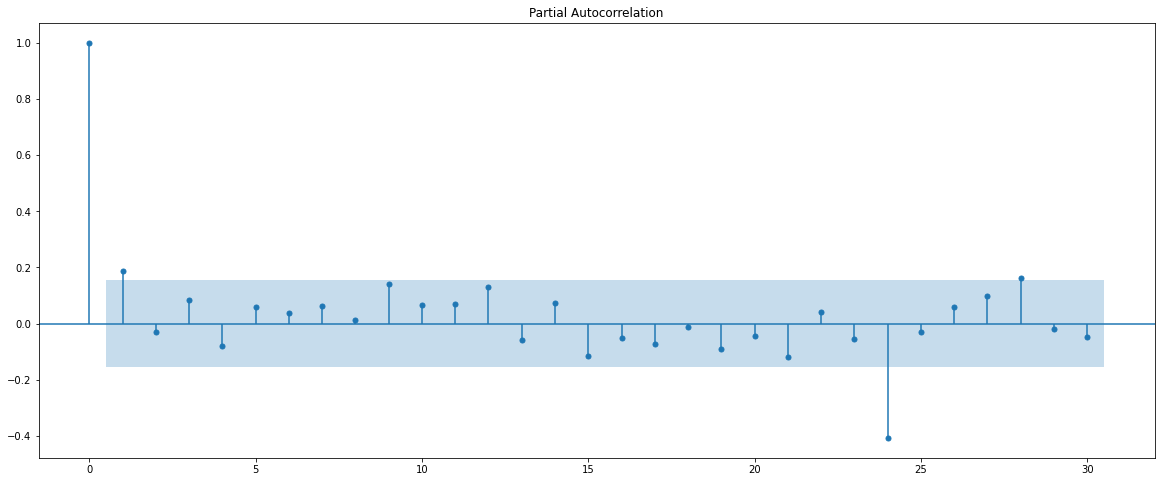
Ho: the data is non stationery

H1: the data is stationery.

As per the hypothesis p value is less than 0.05. so, we reject the null saying the series is stationery.

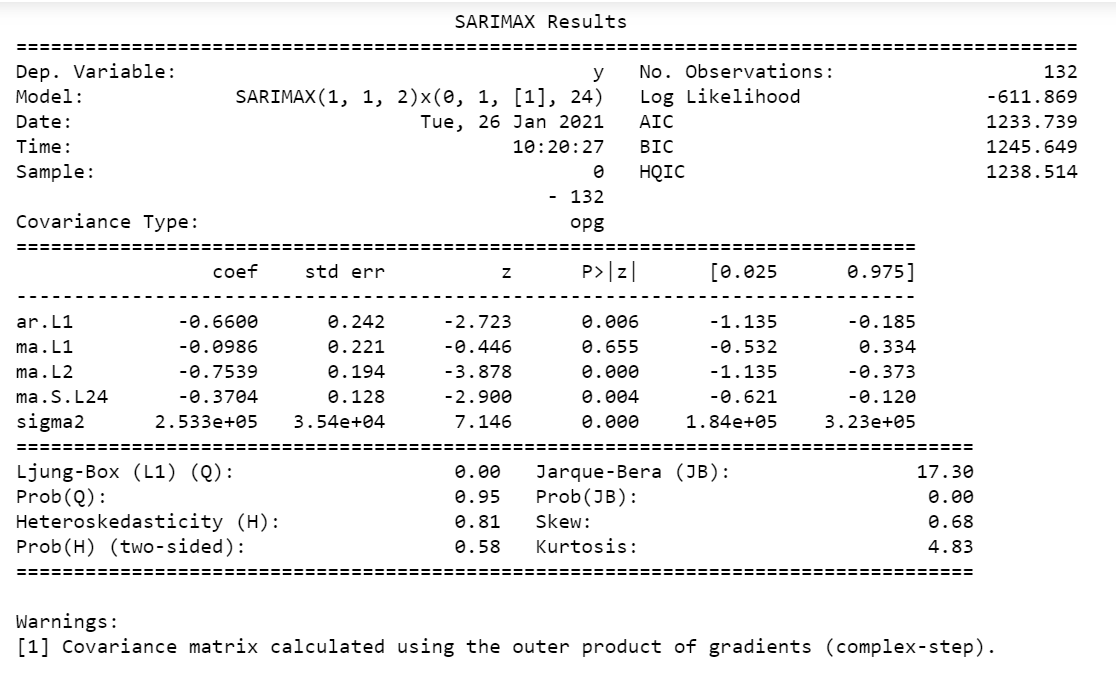
Below the ACF and PACF plot for SARIMA as 24 months.





We can see the difference in the lags here when compared to the [original ACF and PACF](#ACF)

We can see that the p value which was earlier 3 is now 1. The q value is 1.



We have selected the order as 1,1,2 this is the best order for automated sarima and we are using the same here and the seasonality we are using here is 1,1,1.

p.= 1

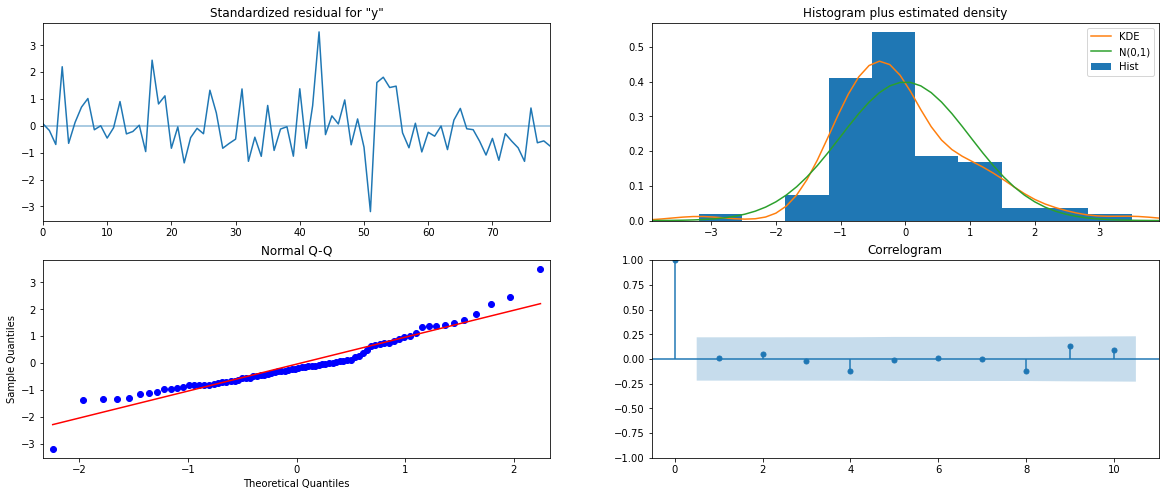
q=2

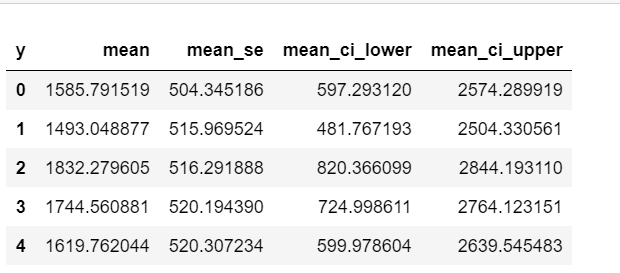
d=1

P=1

Q=1

D=1



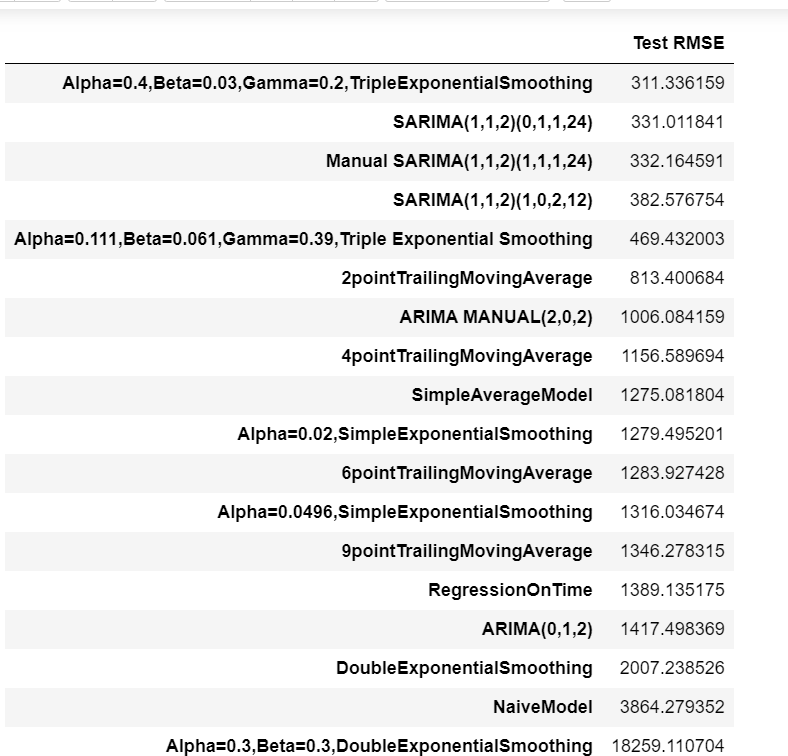


These are the upper and lower predicted confidence intervals.

RMSE:



1. **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.**



We can see above all the models RMSE scores. And the best RMSE for this series is the triple exponential smoothing with RMSE of 311.336. then followed by the SARIMA automated and manual.

The best parameters for the triple exponential smoothing are alpha 0.4, beta 0.03 and gamma 0.2.

1. **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.**

The most optimum model to be built is using the triple exponential method since that gave the least RMSE value. So, we will built the full data based on this. The original data frame is built using the parameter of the triple exponential smoothing.

The smoothing level is set to 0.4.

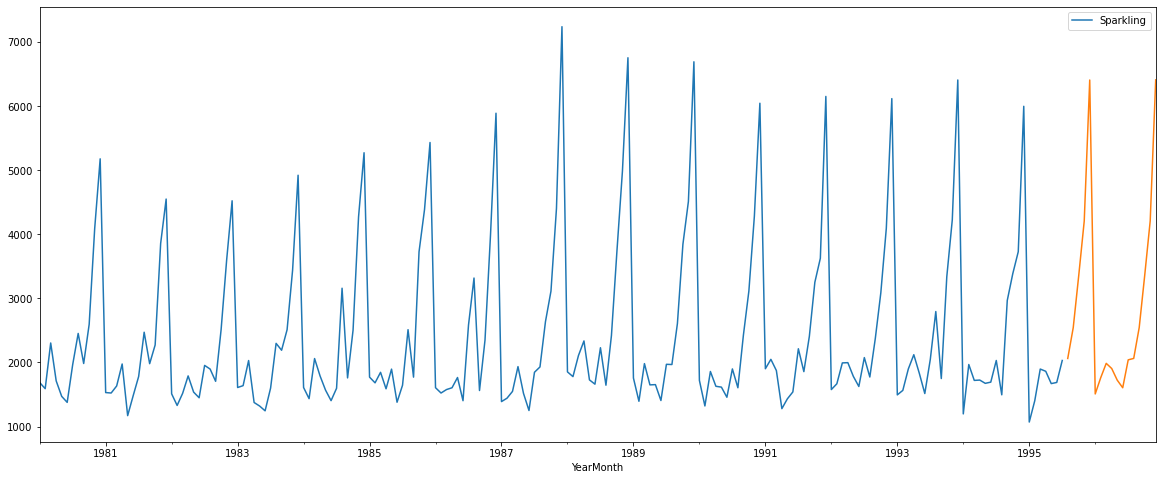
Smoothing trend = 0.003.

Smoothing seasonal = 0.2.

The RMSE for this model is.

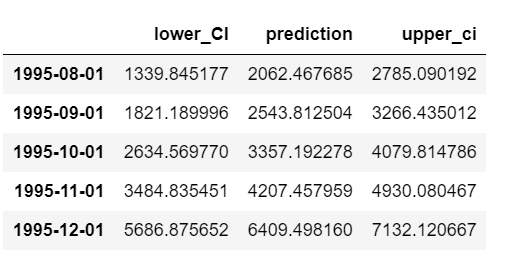


**Prediction Plot:**

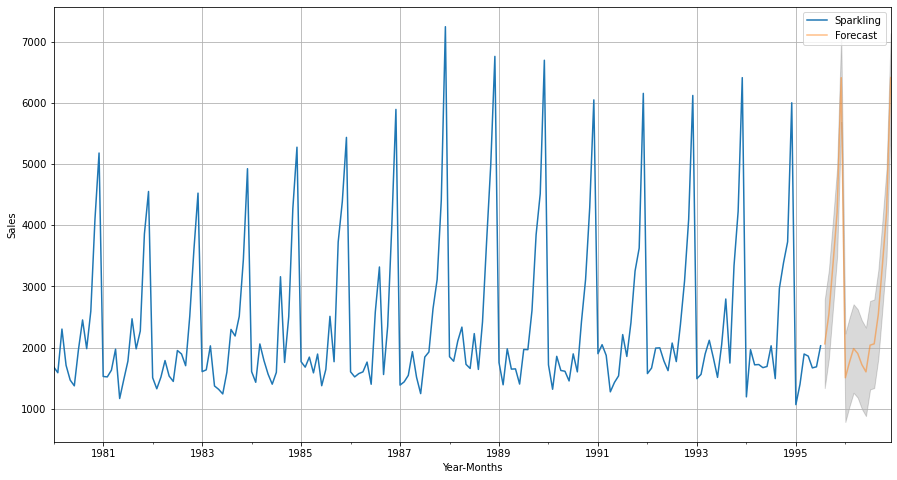


We can see that the prediction of 17 months in the above plot. Since our requirement was to predict for 12 months are there was no sales figures for the last 5 months of 1995 from august to December, I have included those months also for my prediction.

Also, the predicted graph is similar to the previous months. So, there will be a increase in sale in November and December.

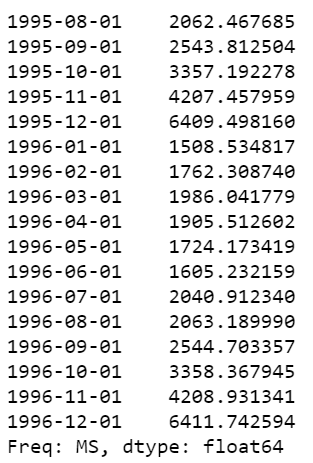


This table show the predicted values in the middle, the lower and upper confidence interval predictions also. So the sales can go minimum and maximum to the mentioned figures.



Above is the graph that represents the above seen table for upper, lower, and predicted sales. The grey areas are the lower ci and upper ci. So, this is the confidence interval area where the sales can be at minimum or maximum.

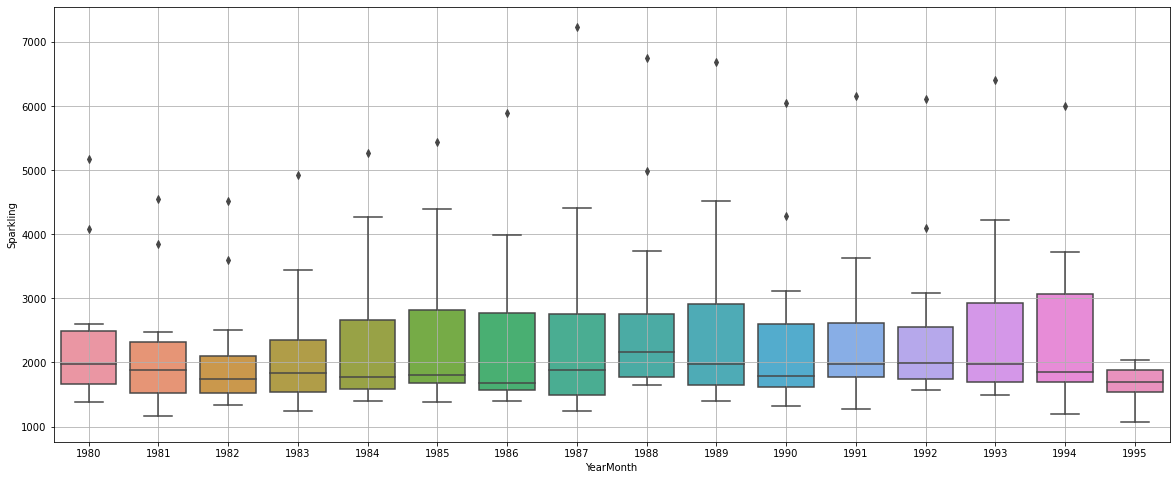
**Sales Prediction for 17 months:**

****

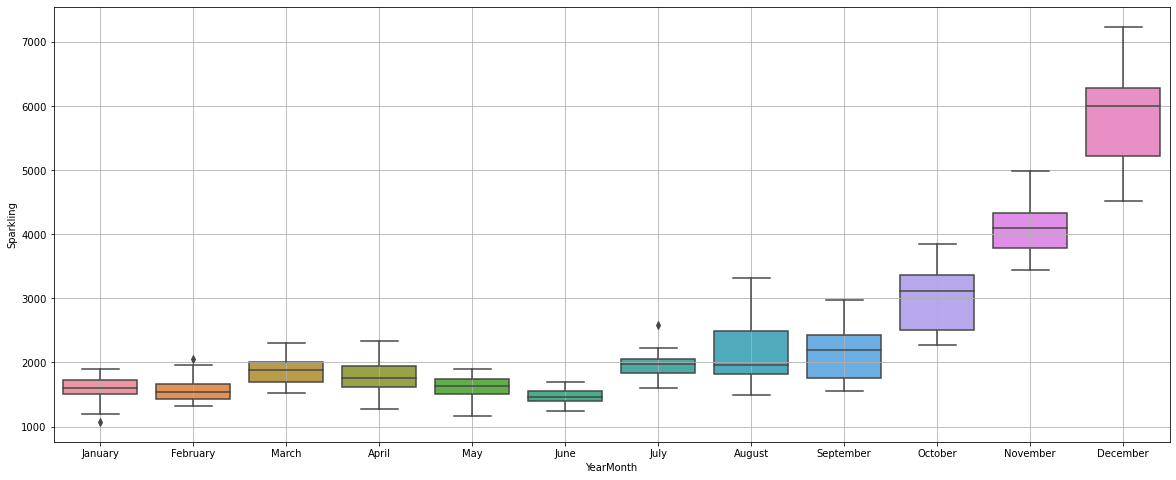
1. **Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**

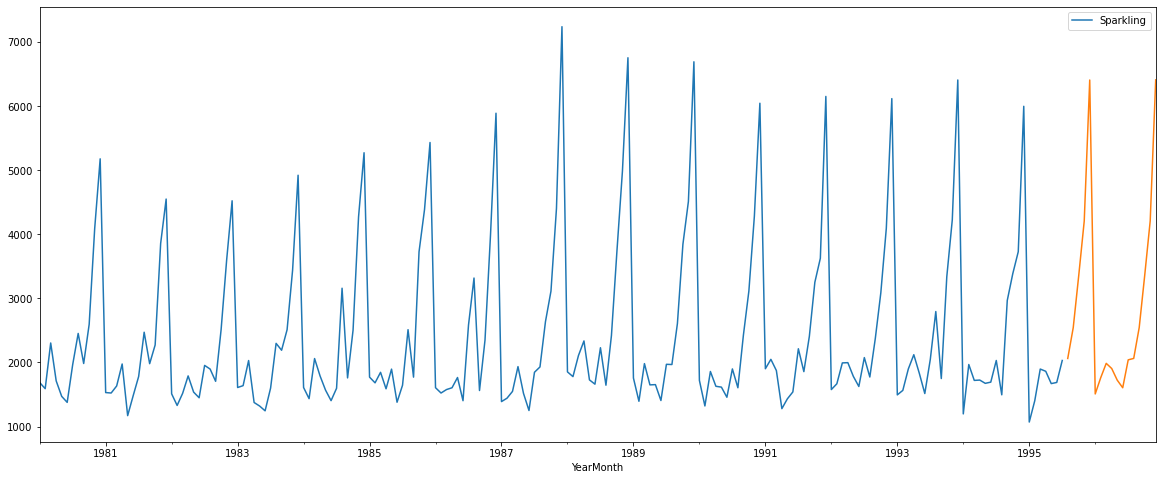
We have used various models to analyses the sparling time series and from the various model we have got the triple exponential model which gave us the best RMSE score. The lower the RMSE the better the model. RMSE means the error in that model is less.

The sparkling wine series had no trend in the series and there was seasonality. Towards the end of the year, we could see an increase in sales of wine for all the years.



From the above yearly bar plot we can clearly say that they are not trend in the data. There are few outliers in the data.

we can clearly see a seasonality in the data from September to December we can see that the sales are increasing and in January to July the sales remains constant with not much of variation.



From the above prediction we can see that it will follow a similar seasonality and the sales will increase during November, December.



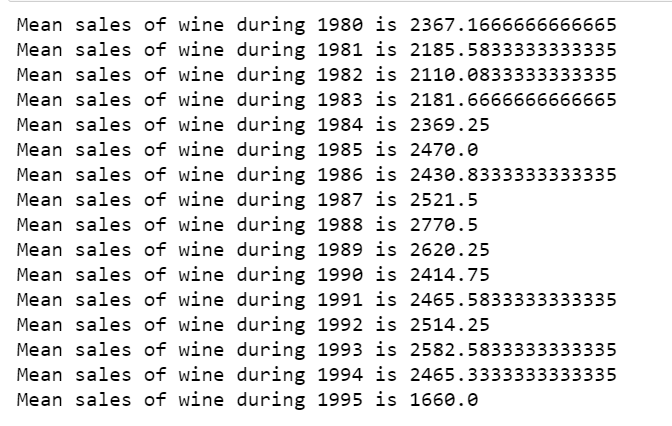
These are the predicted values for the sparkling. We can see that the sales is increasing as per the prediction also.

The average sales for 1996 as per the prediction are.



**Measures for future sales:**

* The prediction sales for 17 months are given above. We can see the lower sales prediction, upper sales prediction, and the actual sales prediction also.
* We can see that the prediction of sales shows a mean sales of 2593 volume.
* Since the predicted sales is good the company can plan to produce more wine and try to implement some advertisement to increase the sales.
* During the high sale months more ads can be target to the specify class of people who are purchasing sparkling wine.
* Since we know the upper and lower prediction sales company can prepare for the less and high sales months.
* The mean sales of our original series are 2402 and our mean sales now including the 17 months prediction is 2593. So, the sales are increasing which is good for the company.
* Below are the mean sales for each year from 1980.



* We can see that the sales have been increasing steadily for sparkling wine since 1991.