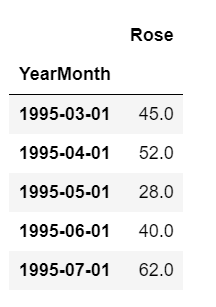
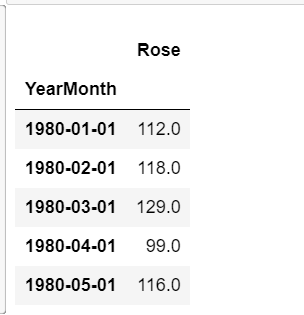
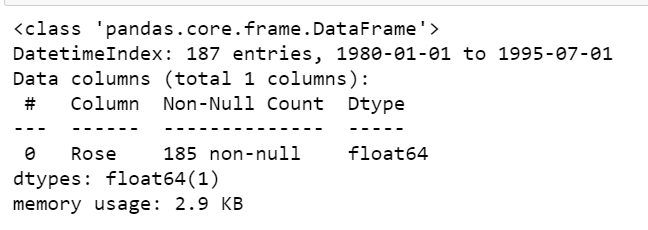
**Time Series Project:- Rose**

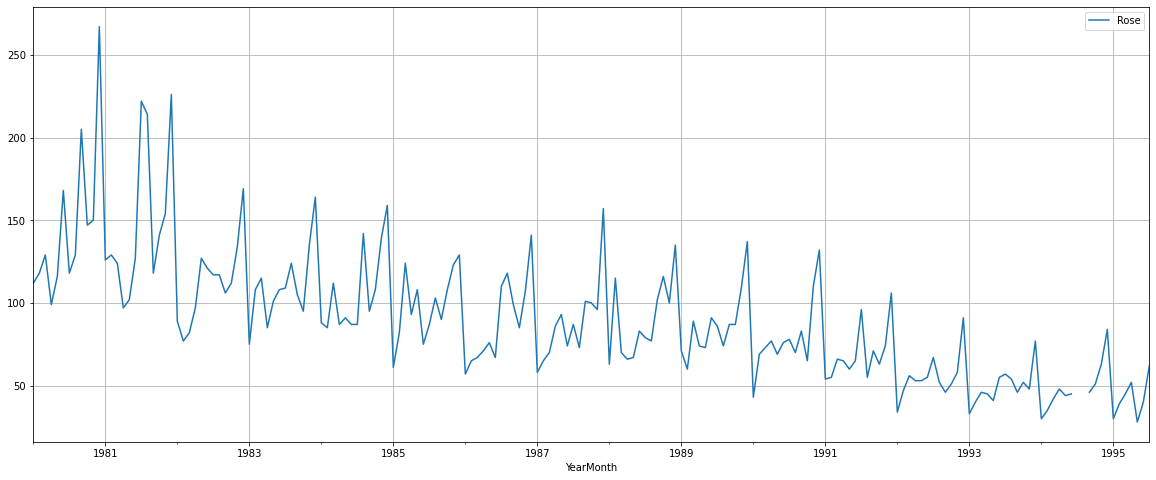
1. **Read the data as an appropriate Time Series data and plot the data.**



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 problem



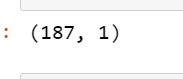
We can see the Rose wine data is in float data type. And out of 187 entries we have 185 entries so there are two missing data in the data set.



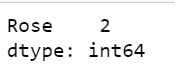
Above the data set is plotted and we can see that there is a downwards trend in the wine sales for the Rose wine. Looks like there is seasonality also but will check the monthly plot to confirm it.

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

**The shape of the dataset:**

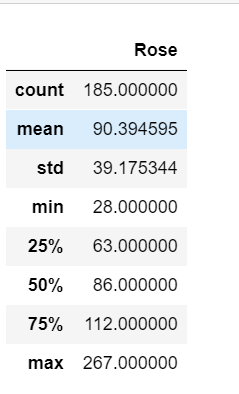
1. 

**Null value:**



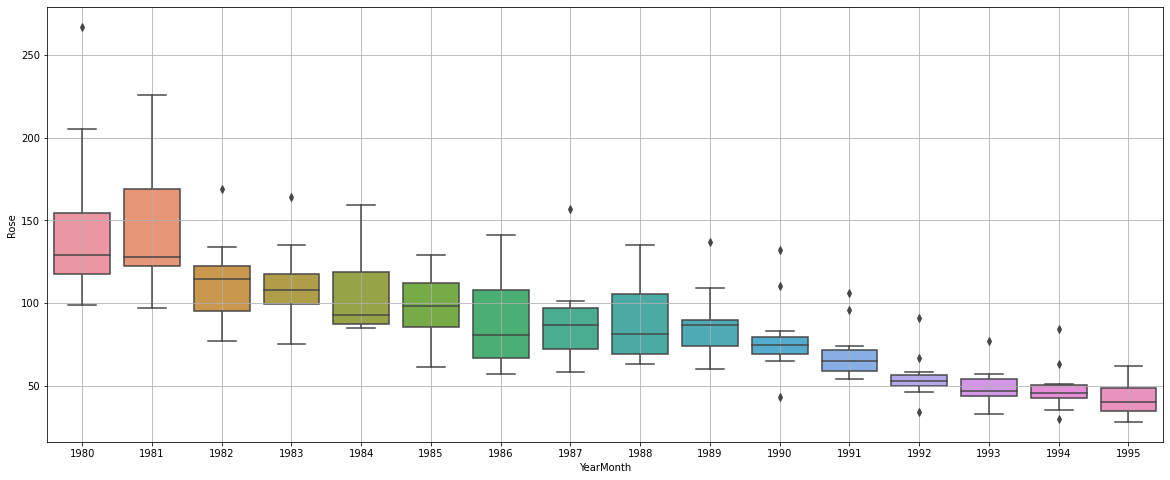
There are 2 null values in the dataset. These null values have to be treated in time series and will be done.

**Describe:**



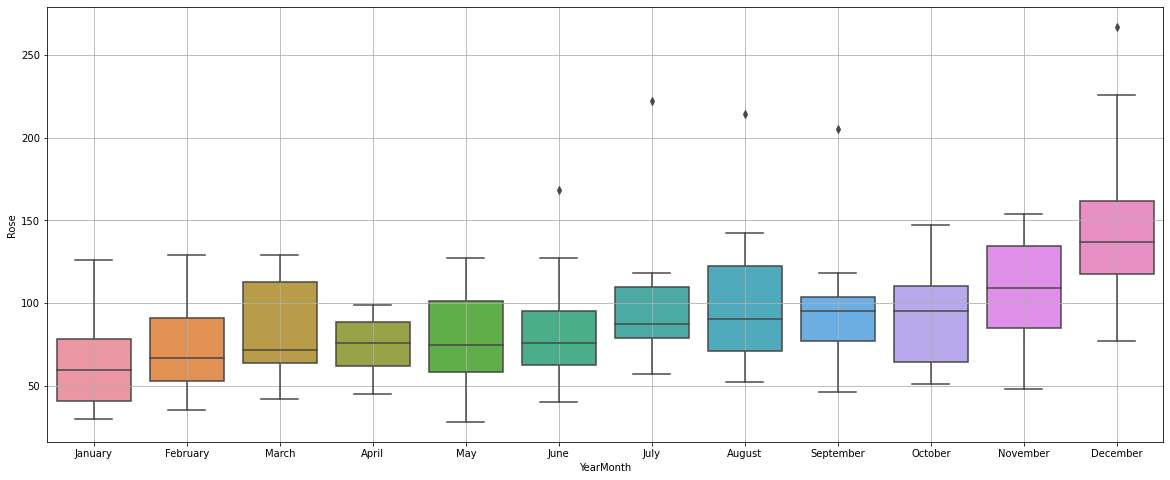
The average sales of Rose wine are 86. Maximum sales are 267 and minimum sales is 28. The mean and the median values are nearby.

**Yearly Plot:**

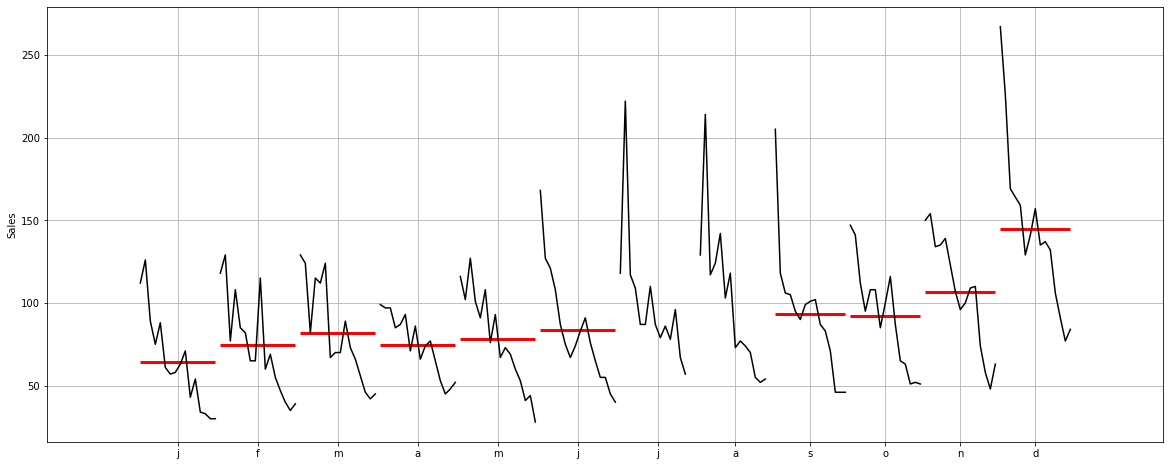


We can see that after the first two years the sales for rose wine has been decreasing steadily. There are couple of outliers also in most of the years. From the yearly plot we can confirm that there is a visible downward trend in the series.

Monthly plot for Rose wine:

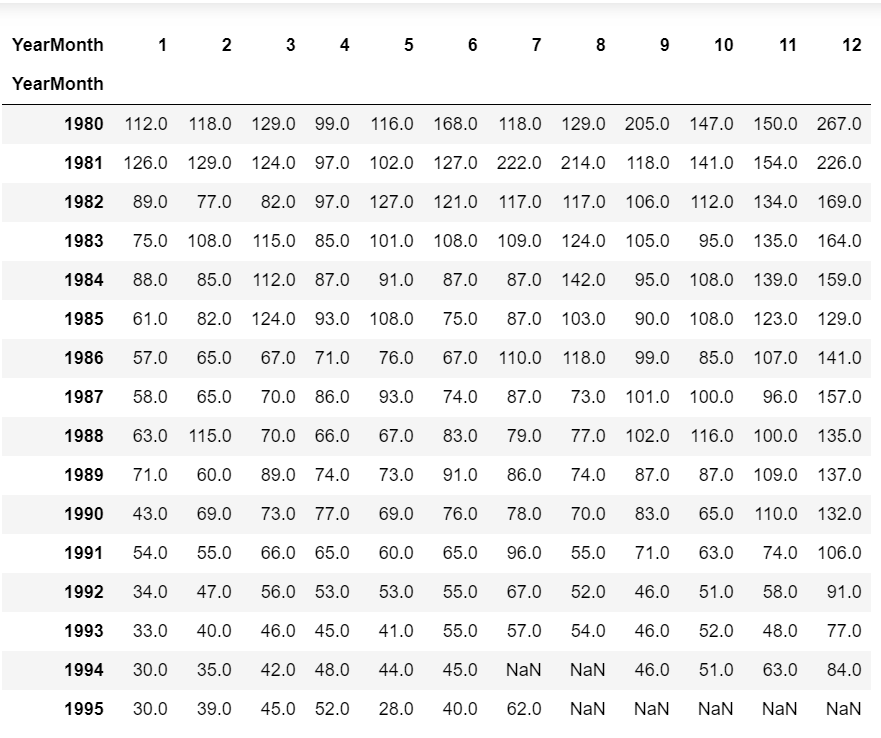


We can confirm from the plot that sales are high for wine during the end of the year. October November and December have most of the sales. April having the least sales. Probably during the festival season of Christmas and new year the wine sales are high. Seasonality can be seen only on the November and December months rest of the months all have a similar sale. There is not much of a difference in sales, and seasonality remains constant for these months apart from one or two months.



From the month plot we can see that sales have been continuously decreasing for every month. The sales average is high as seen earlier in the month of November and December. The mean is similar for all the months. Except for November and December, the mean sales are higher.

**Sales figures:**



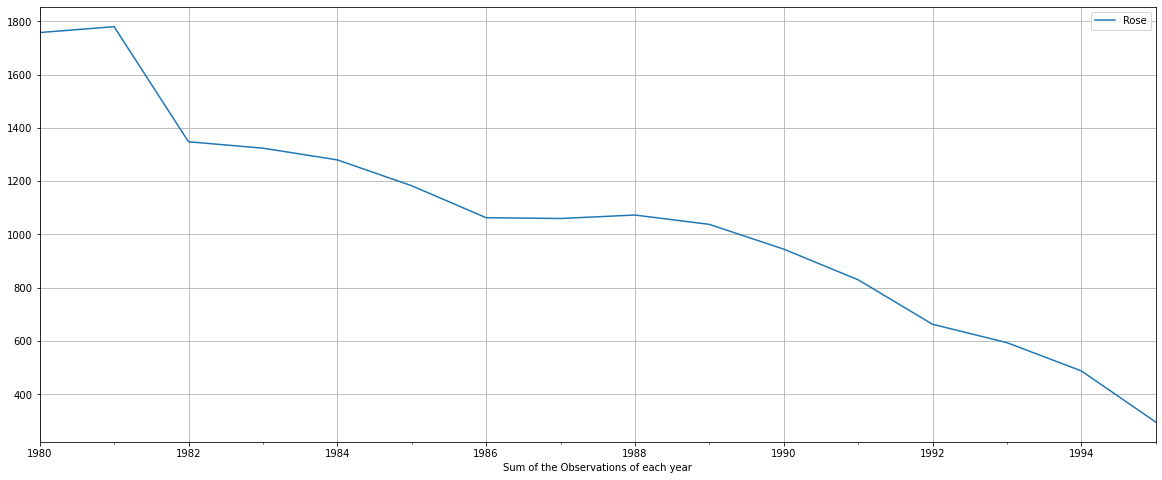
We can see above 1980 and 1981 had the most sales figures for all the months.

**Year and month sales graph:**

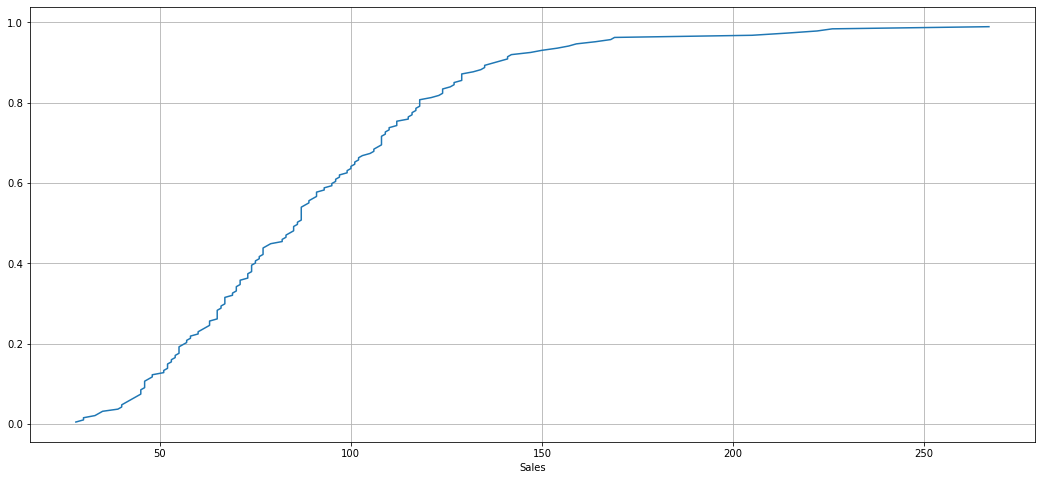


It is clearly visible that December sales are the highest out of all the years monthly sales. Between 1980 and 1982 July and august has most of the sales after December. After 1982 November was the month of all years where sales were high. However, as years past by we can see October and July overtaking the sales in some years. But the sales have dipped rapidly over the decade.

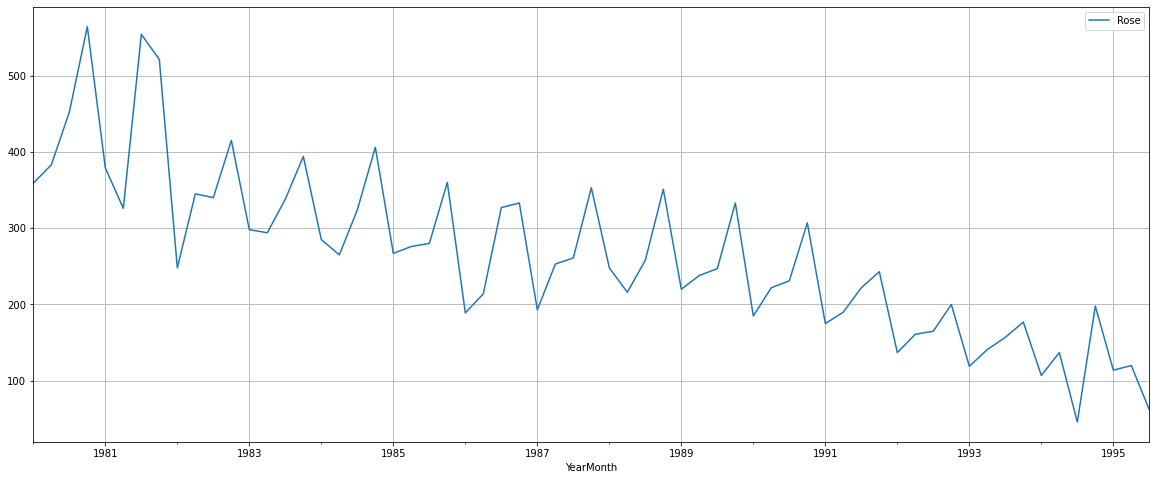
Yearly sales figure:



The above graph shows the sales quantity for each year. We can see that during 1980 and 1981 the sales for Rose wine were around 1800 units. From 1980 to 1982 there is a steep dip of around 24% drop in sales. By 1990 the sales were reduced to almost 50% of what they were selling in 1980.

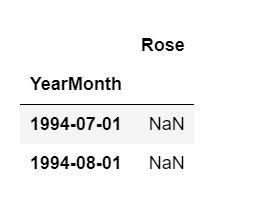
the above plot shows the % of data point for each sale. 90% of the data contributed 150 units of sales. So, rest 10% of the data contributed sales from 150 to 280 units. We can also say that 10% of the data only contributed to high sales probably the early 1980 and 1982. After that the sales were very less for the rest 8 years of data.

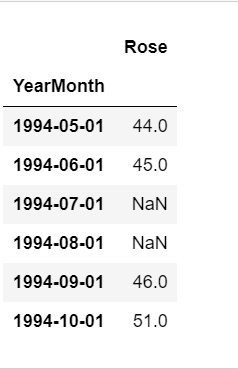
Quarterly data:



We can see that the quarterly data shows that during the early 80’s the high ups and downs every quarter, the first quarter showed a rise and 2nd quarter were falling.

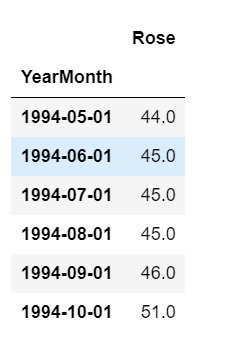
**Null value impute:**

****

****

As seen earlier there are two null values present in the dataset. We have to treat these null values in time series because in time series null values cannot be present.

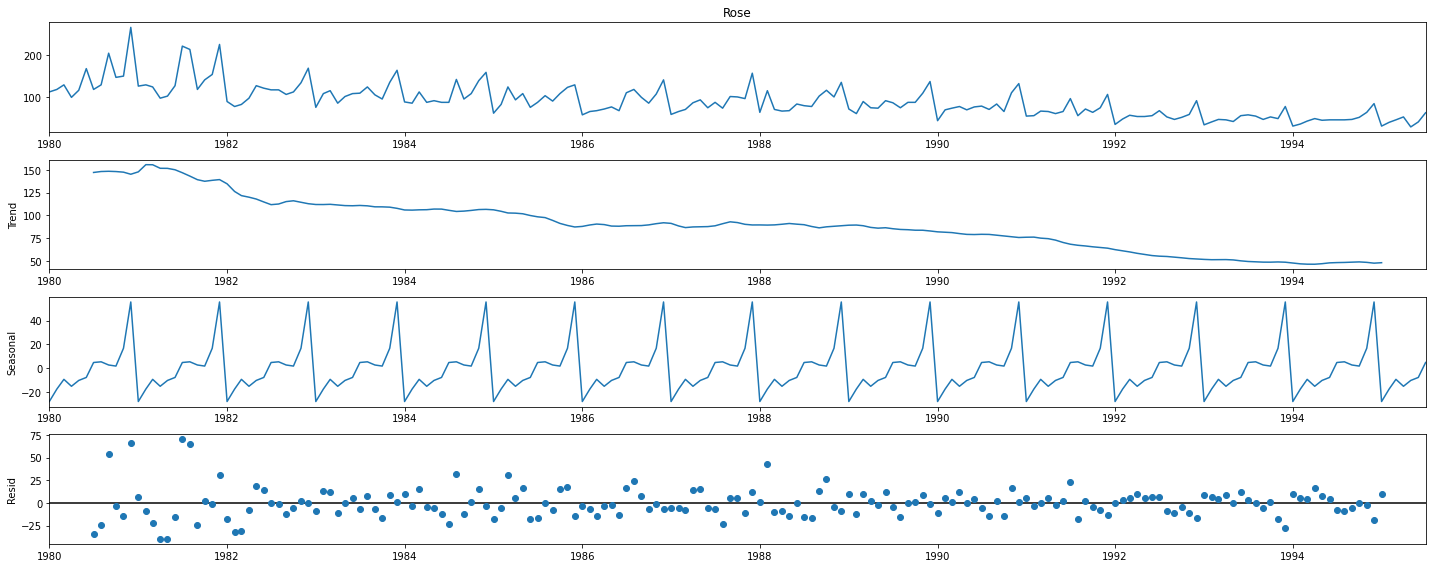
I am using the forward fill to impute the null values.



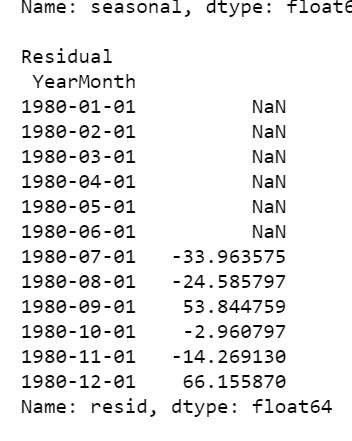
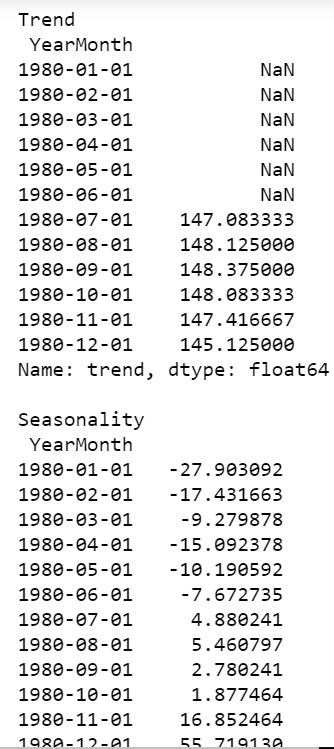
This is the dataset after the forward fill treatment of the null values. The null values are replaced with 45.

**Decomposition:**

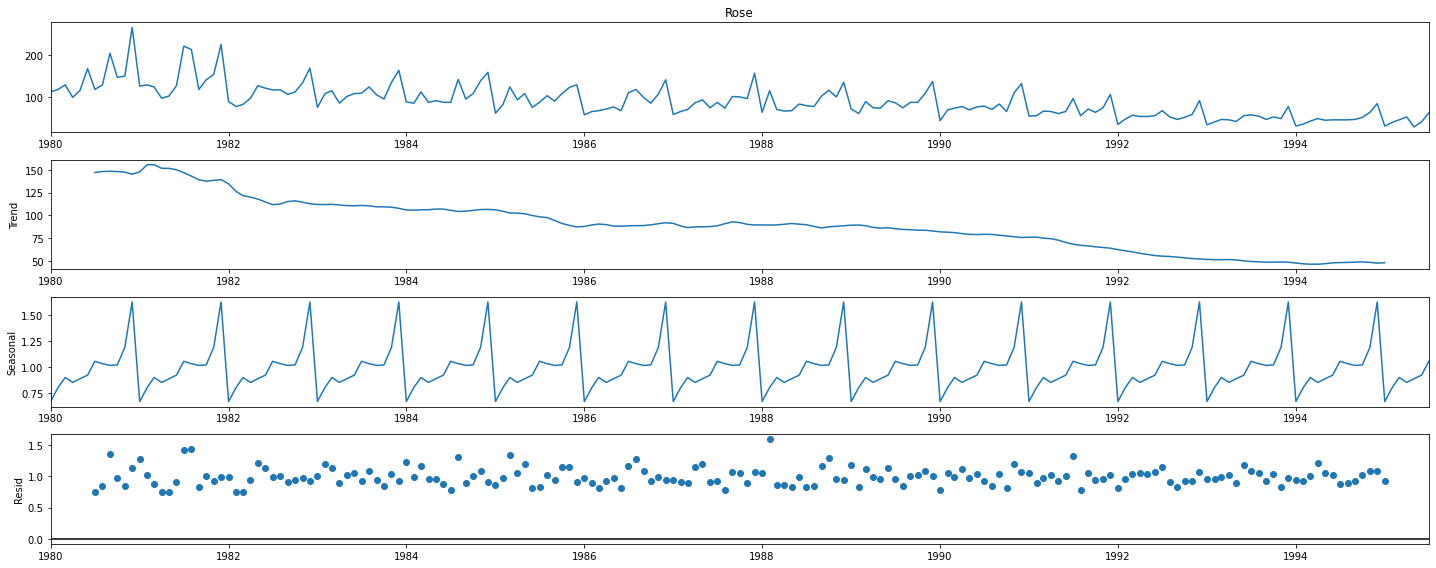
**Additive method:**

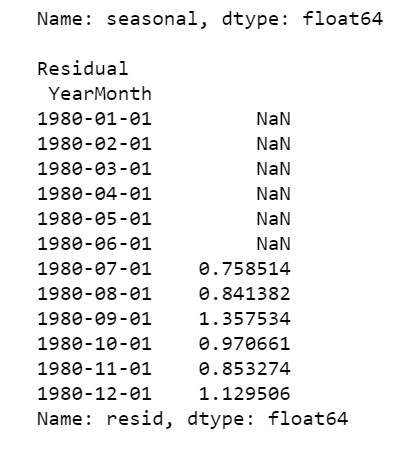
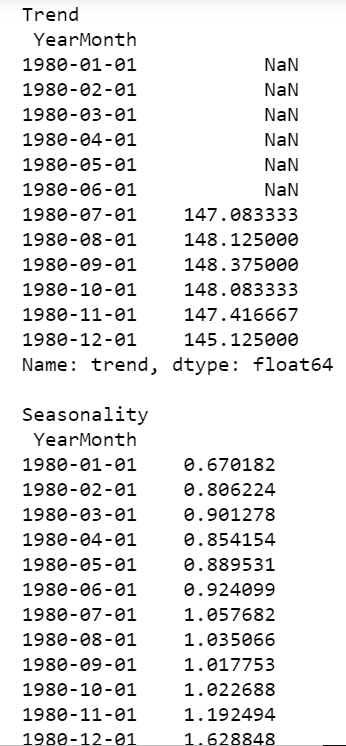
****

We can see after the decomposition there is a downward trend and seasonality in the dataset. Also, we can see the residuals are not near zero there are spread in the beginning and towards the end there are close to zero. We can also see that the unexplained data the errors are high in the residual column.

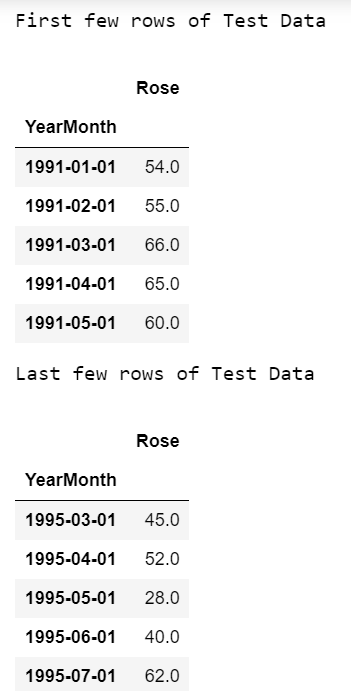
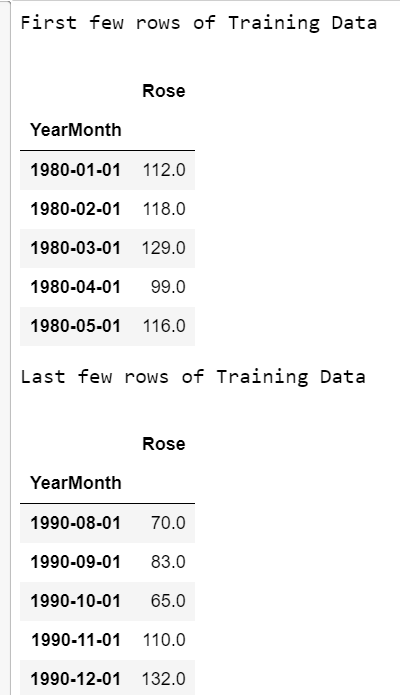


**Multiplicative method:**

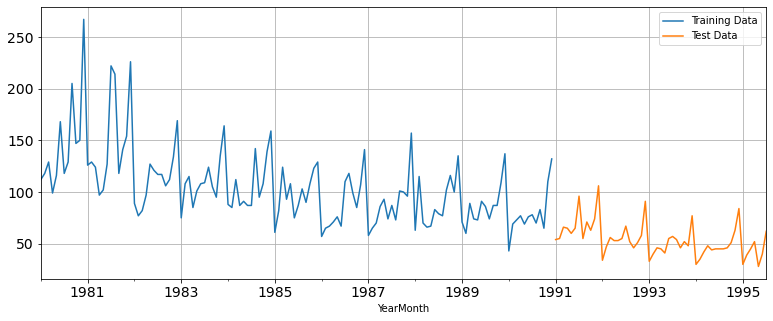
****



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.





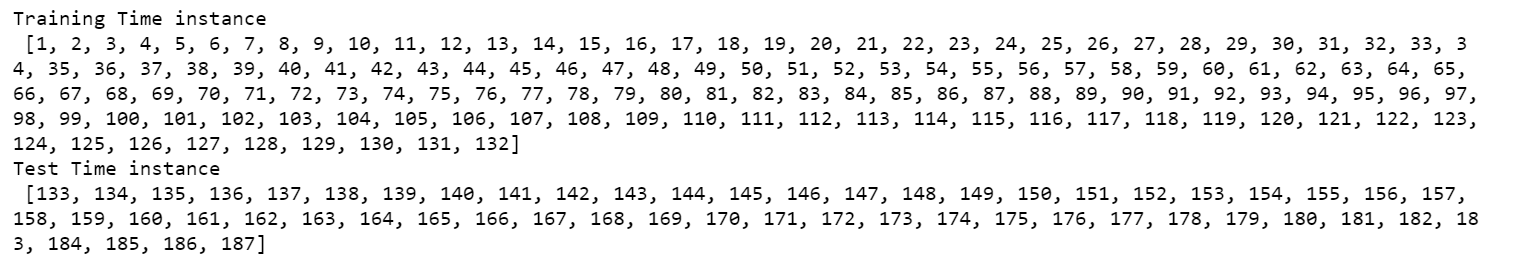
1. **Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.**

We will build the below models:

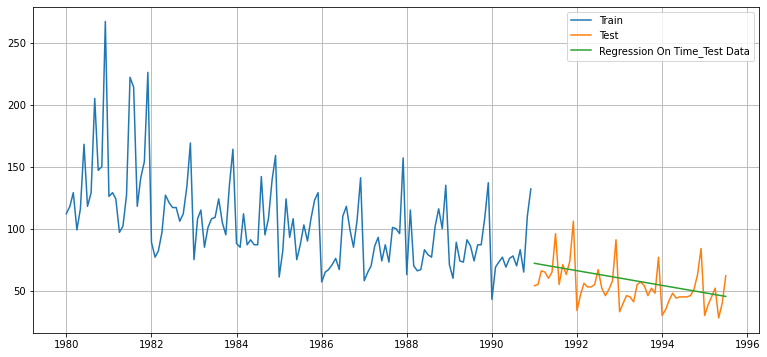
* Linear Regression model.
* Naïve model
* Simple average
* Moving average
* 3 types of exponential smoothing models.

**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.

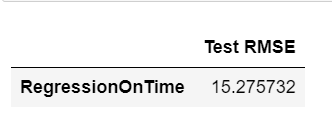


Below we can see the prediction of the linear regression model.

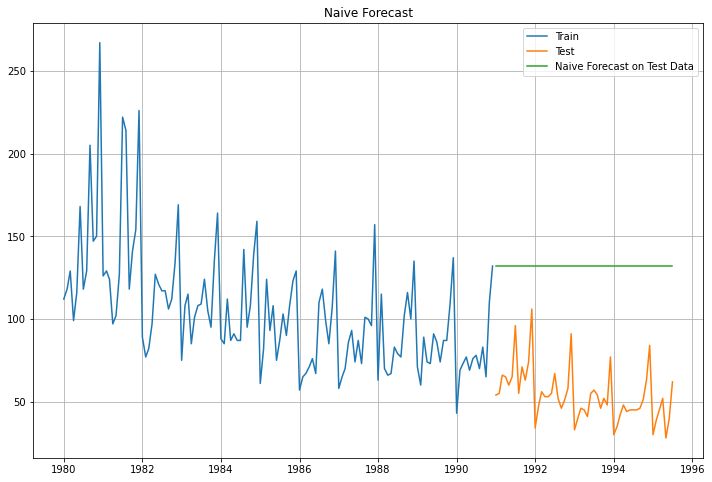
****

RMSE:

The RMSE for linear regression is 15.27.



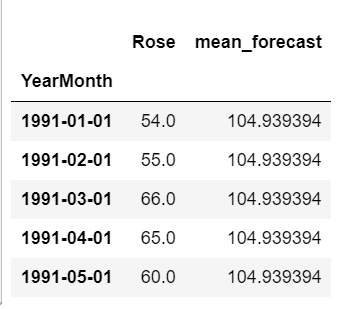
**Naïve model:**



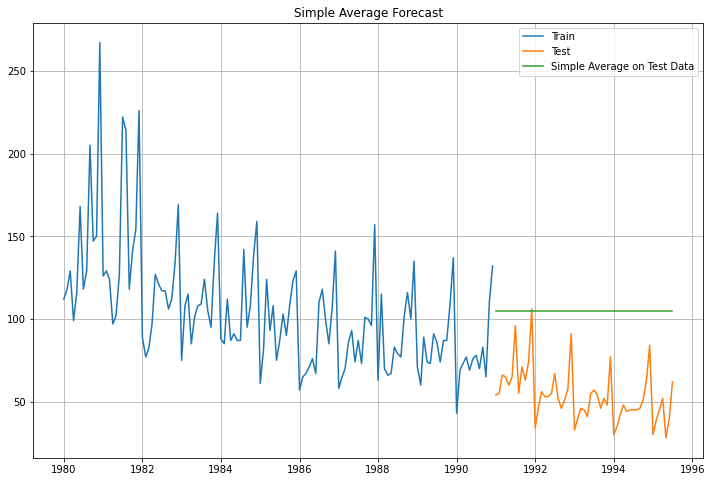


Above we can see the graph for the naïve based model, and we can see the RMSE score is extremely poor. The score is 79.73. Compared to Linear the naïve based model performs extremely badly.

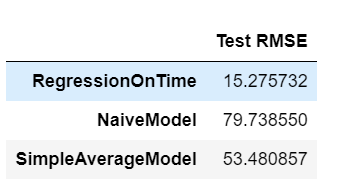
**Simple average:**

****

for simple average we take the mean forecast and keep that value as constant and predict the RMSE.



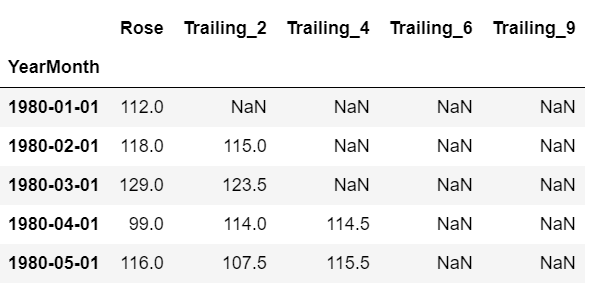
We can see the best fit line is not giving a good line. It’s a flat line.

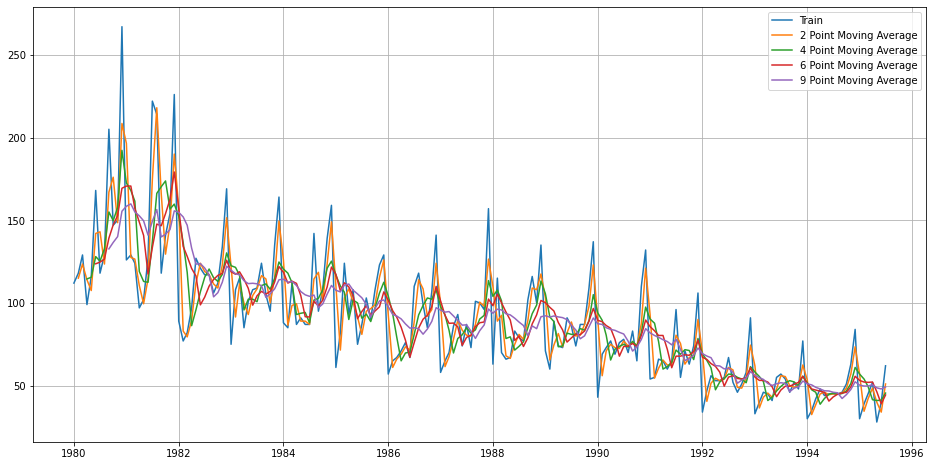


RMSE score for simple average model is 53.48.

**Moving average:**

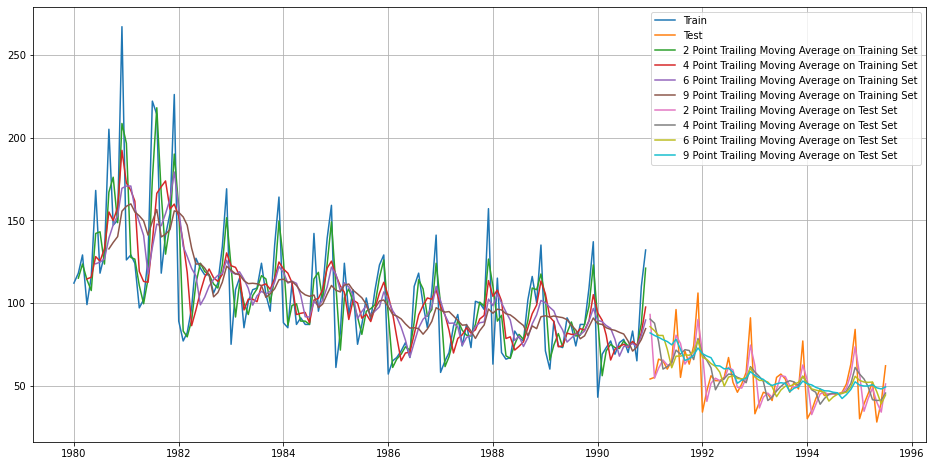
We will see the moving average of different trailing values.





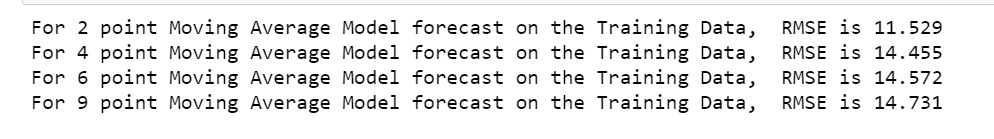
We can see the plot of different point moving average above with comparison of original train data.

**Moving average on the test data:**

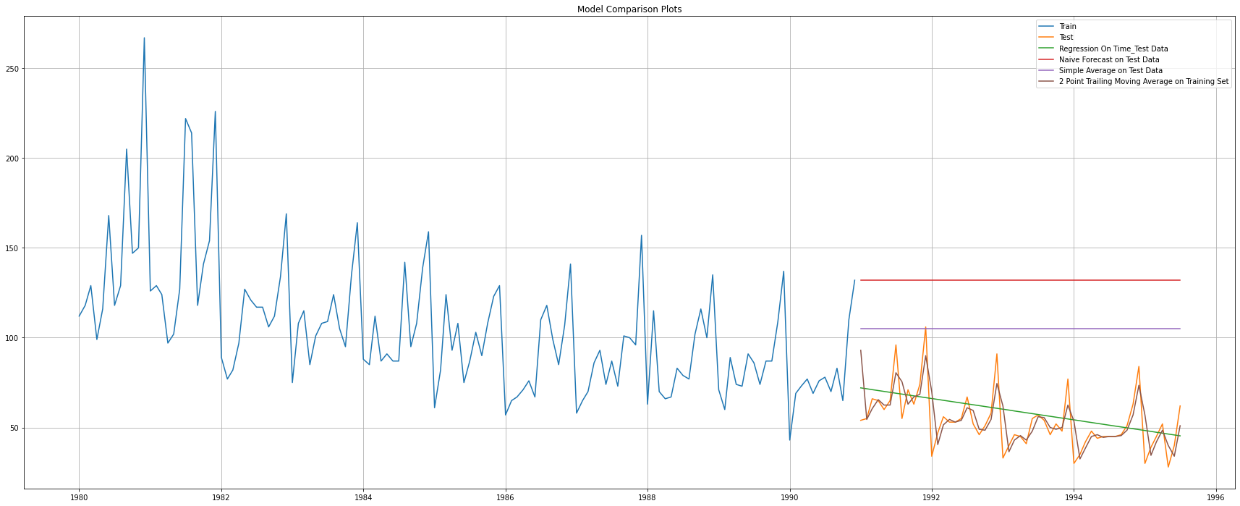
****

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 has given the least RMSE of 11.529.

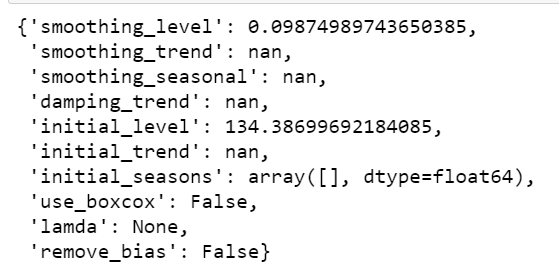


Plot that shows the performance of all model on the test data. We can see that the 2-point moving average is performing best among the linear, naïve, and simple average models.

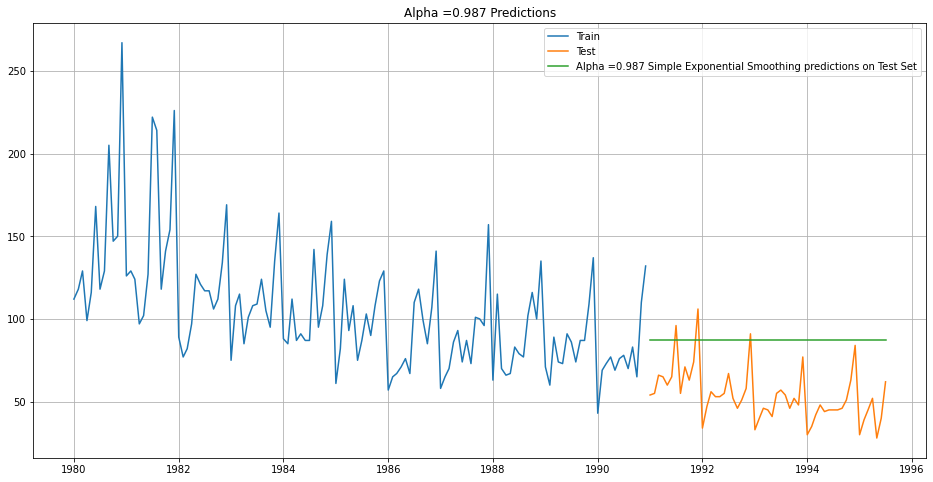
**Exponential smoothing models:**

**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 0.0987



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

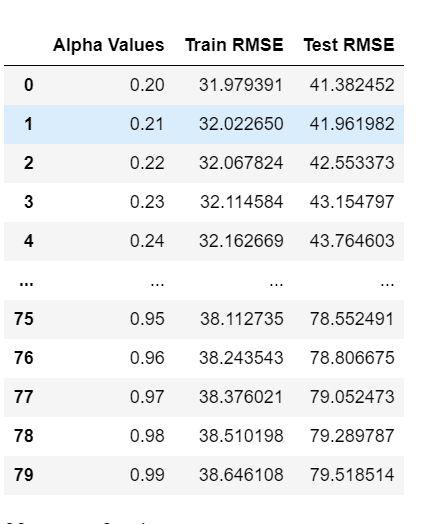


The RMSE is very high for this model at 36.817

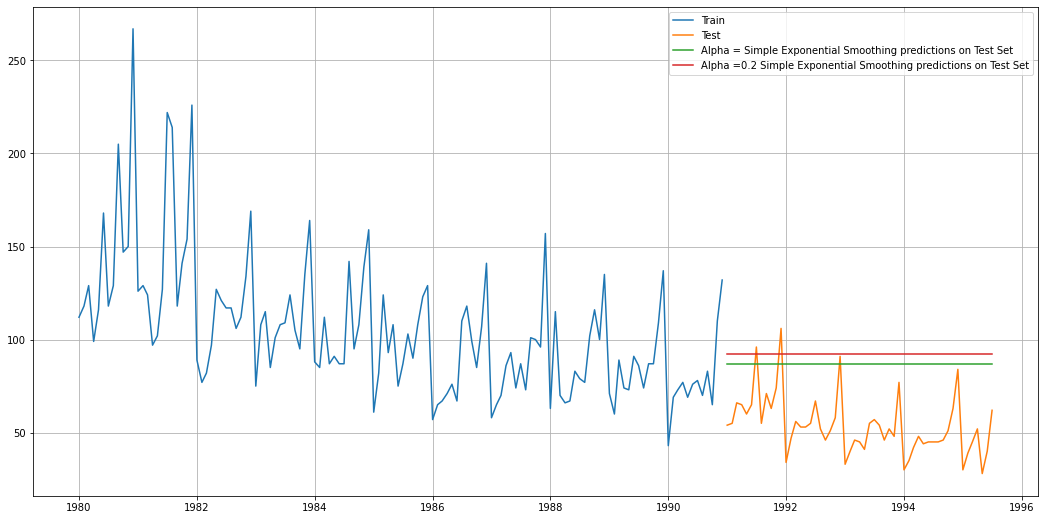
**Trying different iteration for the simple smoothing model:**

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 41.38.

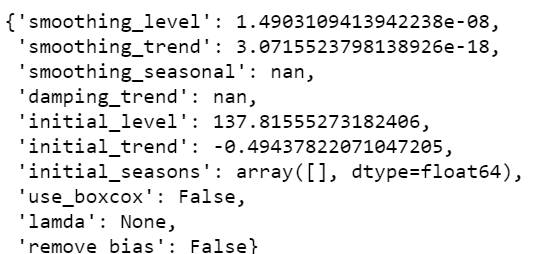


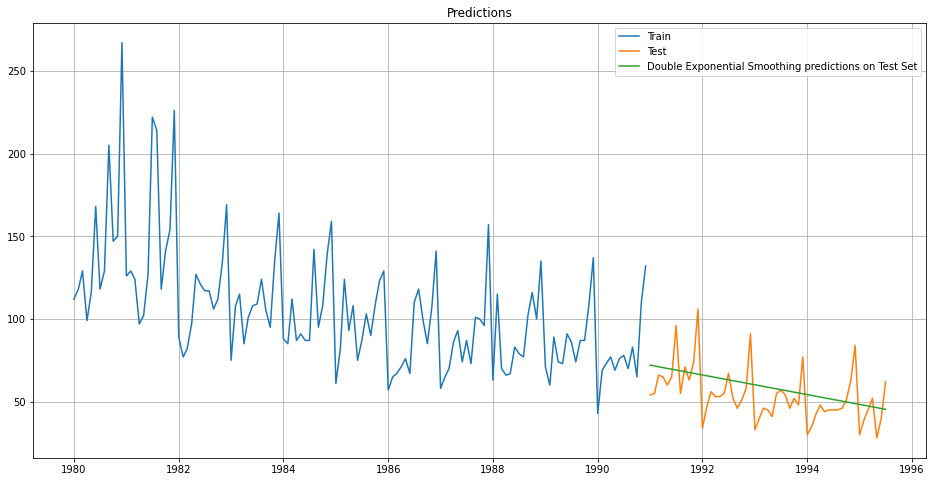
Above is the graph that shows both simply smoothing representation with the test data.

**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 1.49e-08 and trend as 3.07e-18.

****



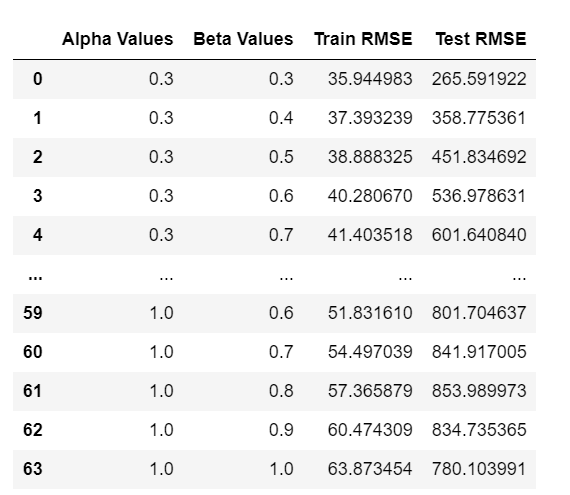
Above we can see the prediction using the alpha and beta we got earlier from the model.

RMSE is :



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

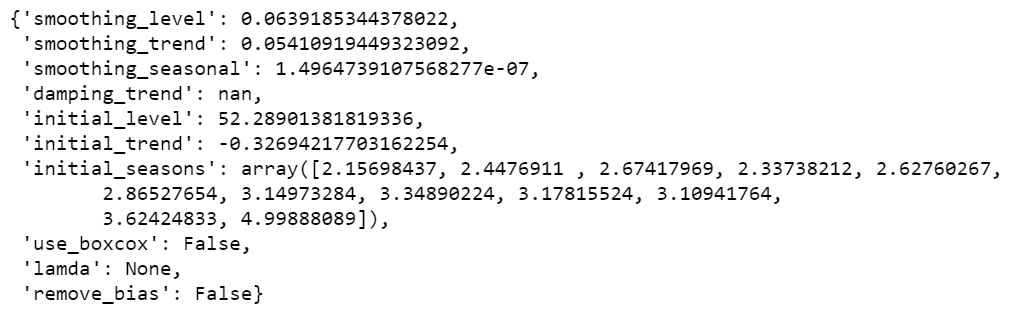
I have used the alpha values from 0.3 to 1.1 and beta values of 0.3 to 1.1 to see if we can get a better test RMSE and below are the results of it.



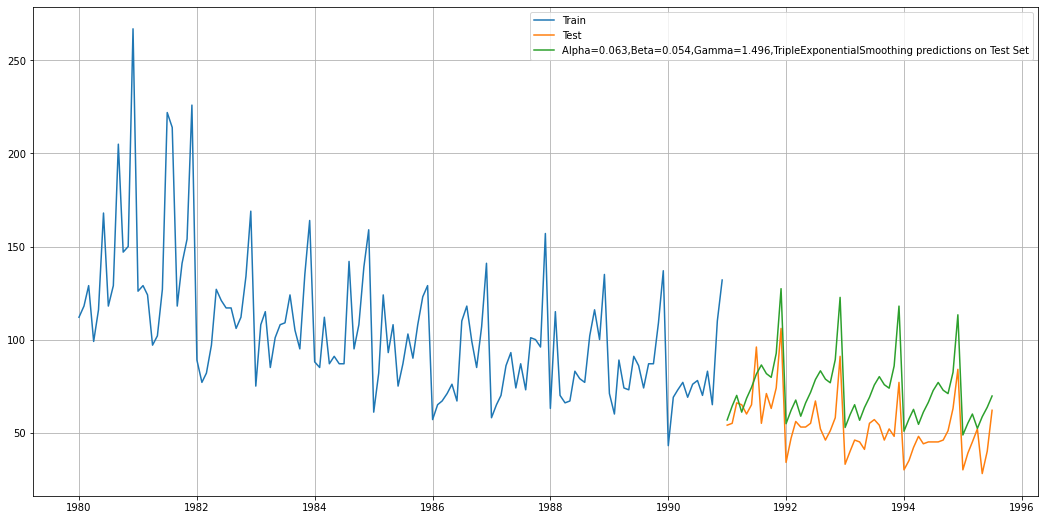
The RMSE are extremely high compared to previous parameters.

**Triple Exponential smoothing:**

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

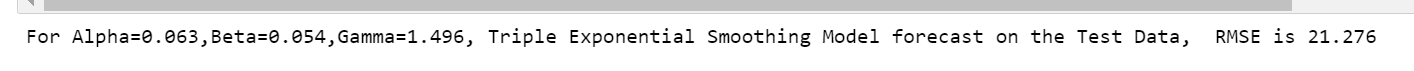
****

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



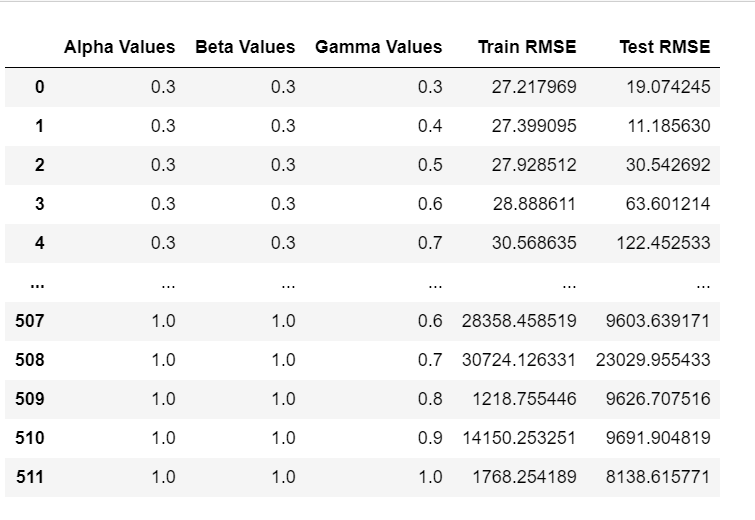
Using the best param we can see that we have a better model compared to all the model we have done till now. We can see the predicted values the green line is almost close to our test data. This shows that the prediction is better compared to other models. We will see if we can improve this by changing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

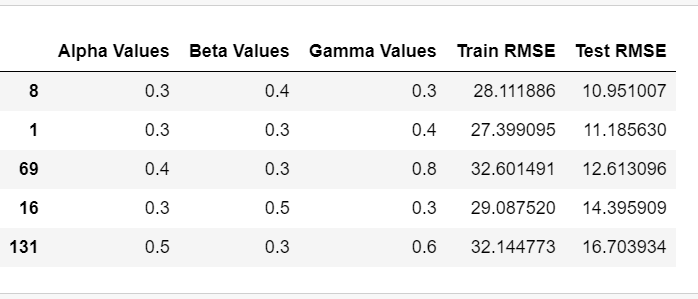
The RMSE is 21.76



**Different parameters:**

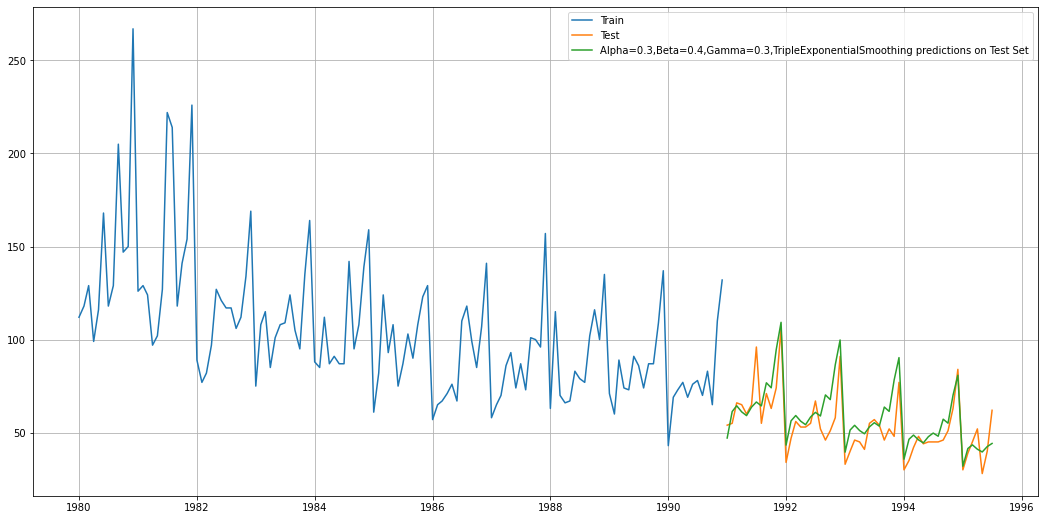
I am setting the alpha values from 0.3 to 1. Beta values also from 0.3 to 1 and gamma also the same. Lets run the model with these parameters and see if we can reduce the RMSE.



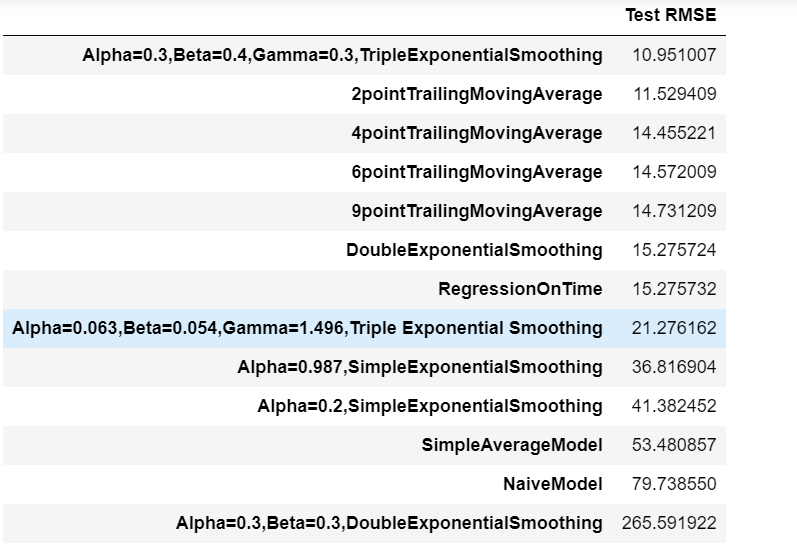


We can see that the RMSE we got is 10.95 for alpha as 0.3 beta as 0.4 and gamma as 0.3

This is the best model so far we have found with the least RMSE of 10.95.

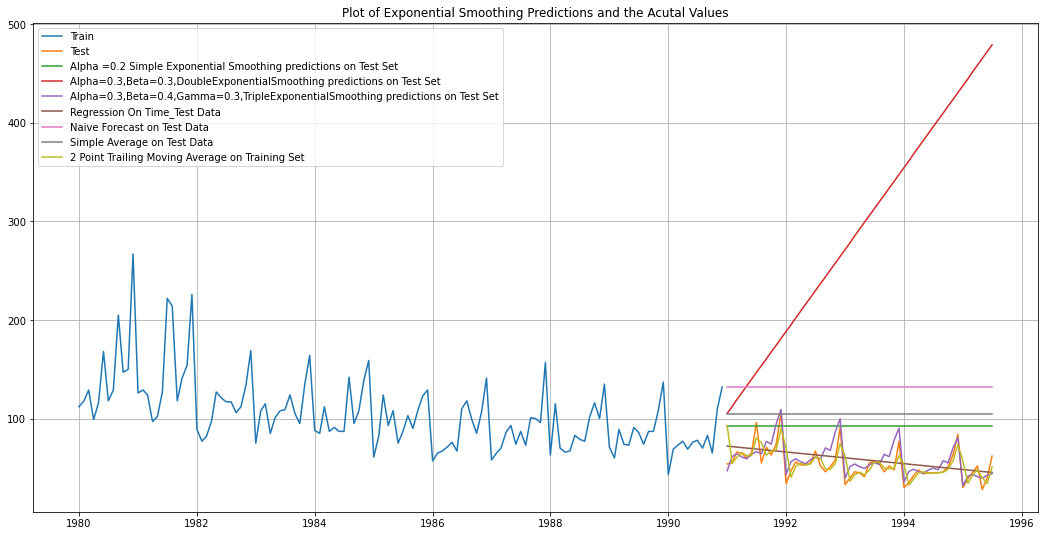


Here we can see that the predicted graph (green) is performing much better than the earlier parameters.



Out of all the model this is the best performing model till now. We will see how the series performs when we do arima and sarima.

**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:**

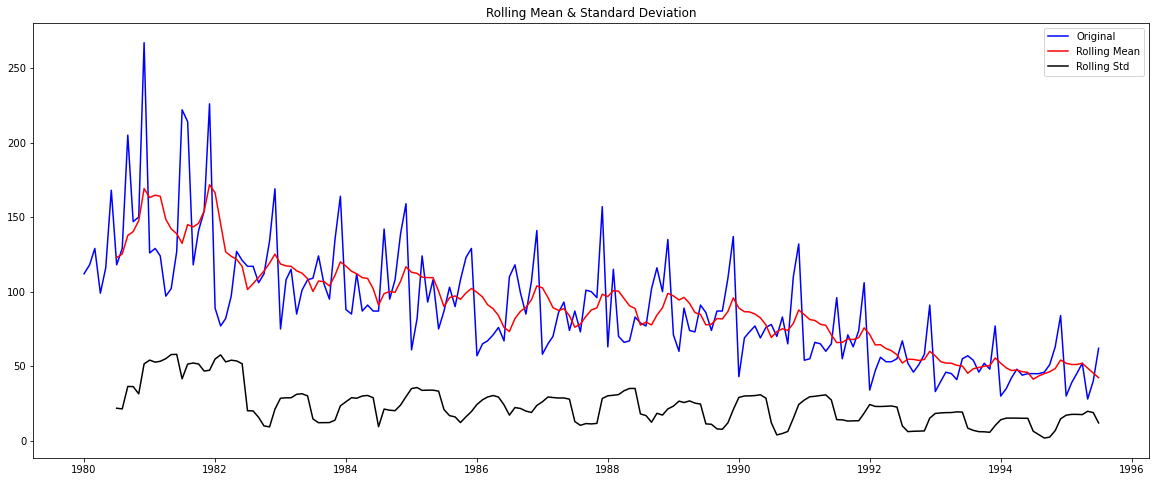
**HO:** The series is nonstationary

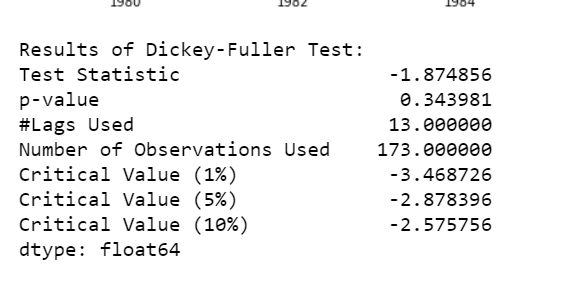
**H1:** The series is stationary.

The stationarity is checked for 0.05.

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

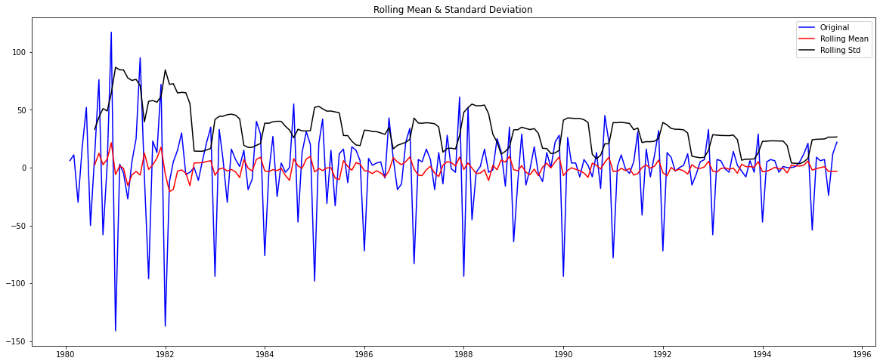
We use the ADF test to check the stationarity of the series and we found that the model is not stationary.

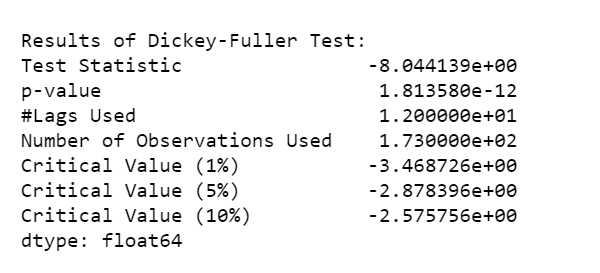




We can see that the P value is 0.34 which is higher than 0.05 so we cannot reject the null hypothesis which is the series is nonstationary. So as per the above the series we have is nonstationary.

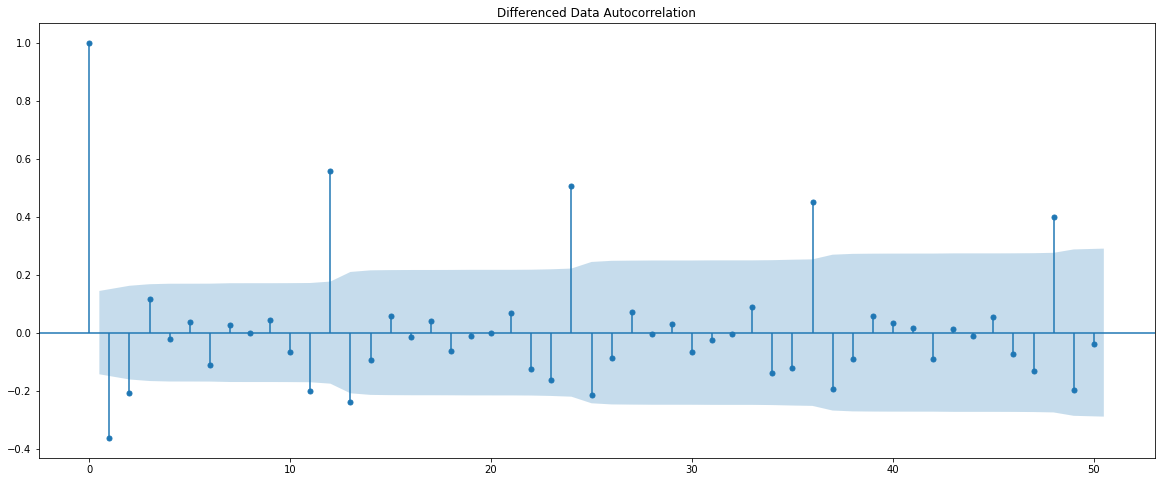
To make the series stationary we have to take difference and see and what difference the series is becoming stationary.

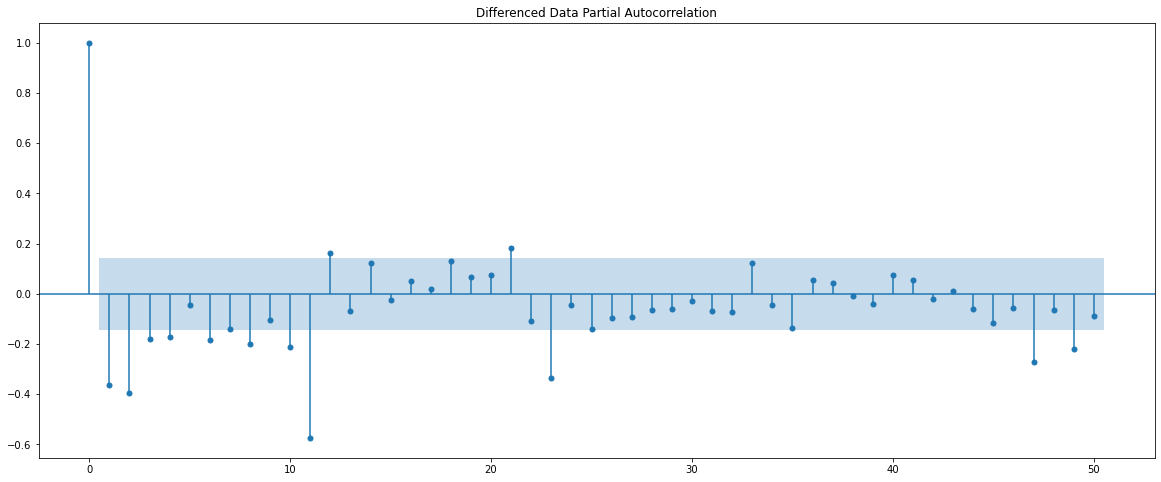




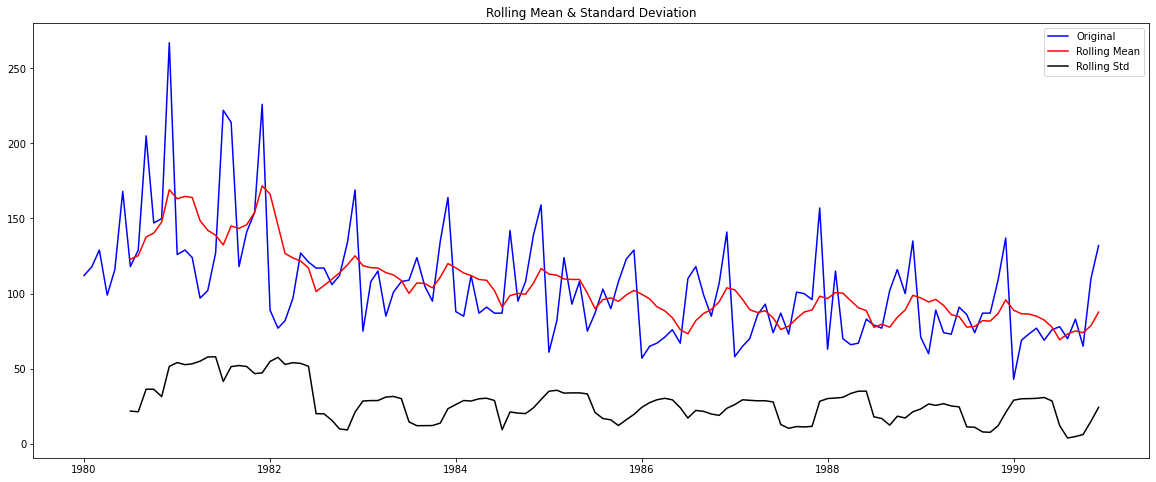
When using the difference as 1 we have got the P value of 1.813e-12 which is very less to 0.05 therefore we can reject the null hypothesis of the series is nonstationary. With taking a difference of 1 our series is stationary now.

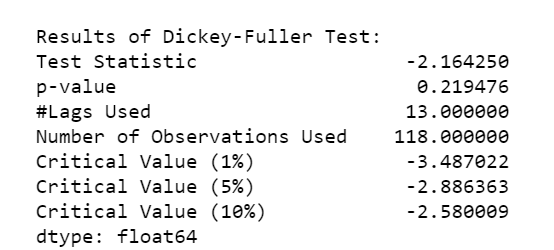
1. **Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**

****

****

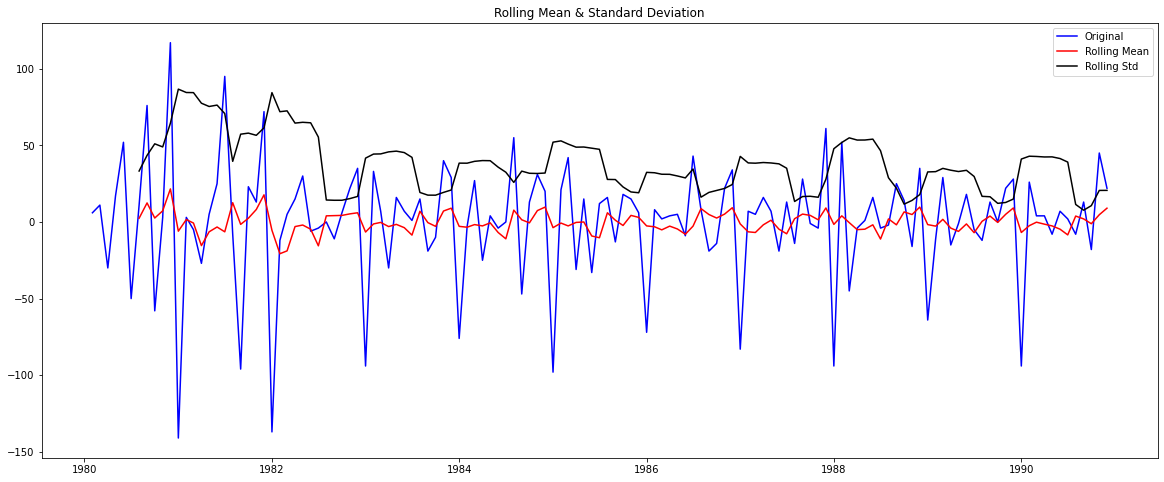
**Checking stationarity on training data:**

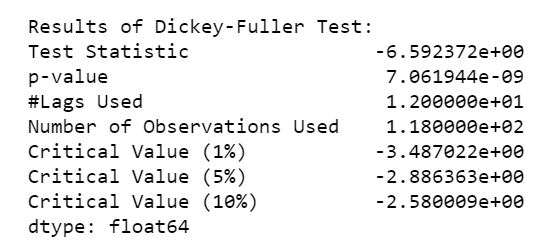
****

****

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:**

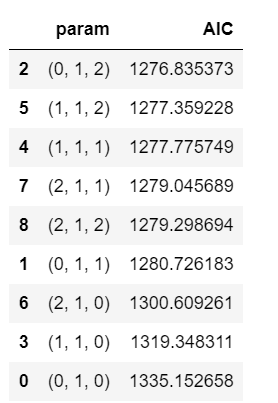
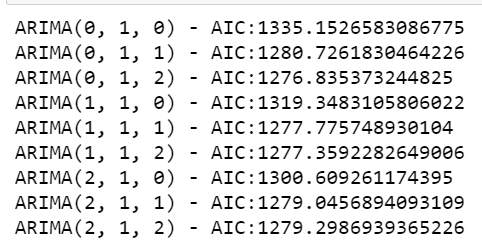
****

****

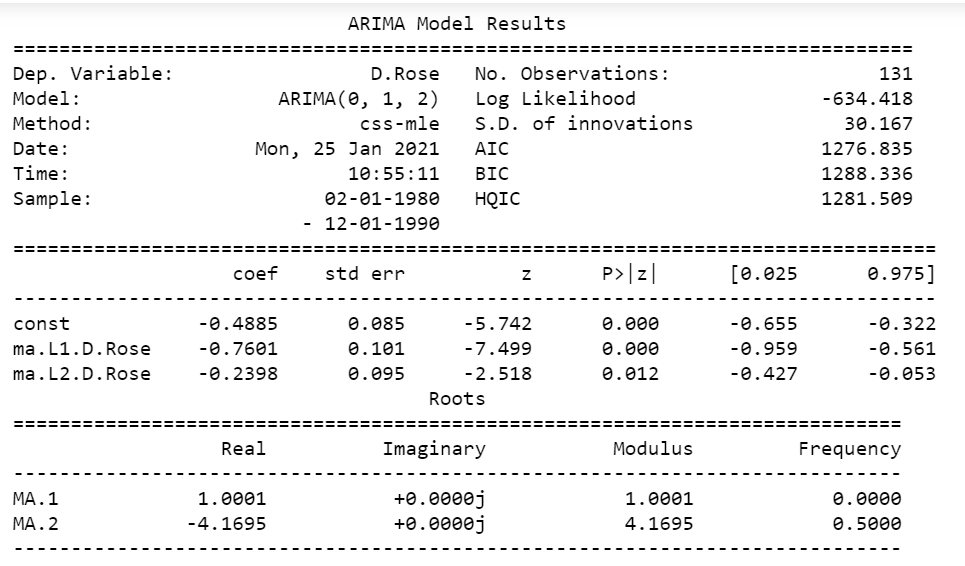
Above are the ACF and PACF plots. This will help in determining the values of P and Q in the arima and sarima parameters. However, we will build an automated model of the sarima and the arima for this task.

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.

The automated model will run different combination of the p,d,q and the best combination that gives the least AIC value is selected.



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 0,1,2 and the AIC for this combination is 1276.85.



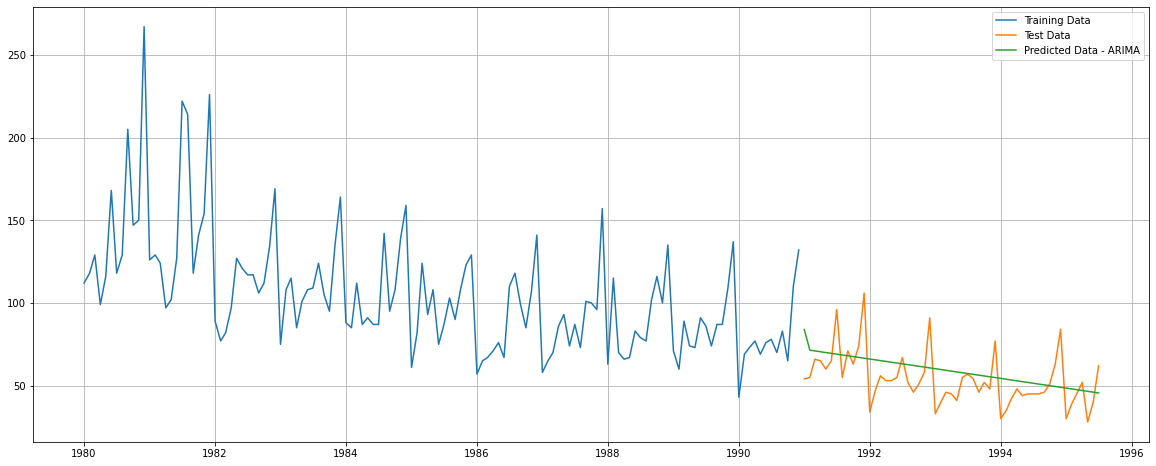
Above we can see the constant and their P values. The ma is the moving average and two ma are there for level 1 and 2. In 0,1,2 2 represents the ma so there are two ma. Ma.L2 p value is more than 0.012 so it is not significant can be dropped or since its not too much of a difference we can keep it also.

**RMSE automated arima:**

The values are predicted using the p,d,q values. They are predicted on the test data and the RMSE scored is calculated. Below is the RMSE score.



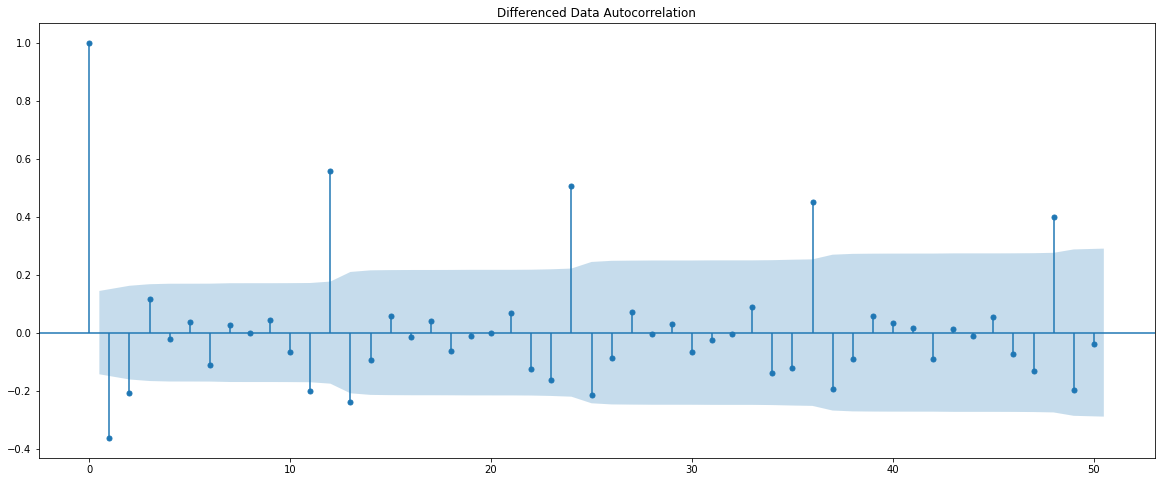
The RMSE for the automated arima is 15.62.



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.

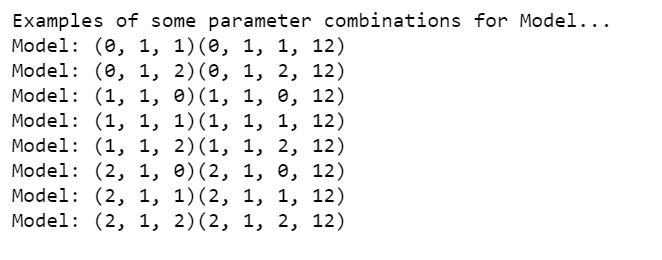


We can clearly see that every 12 lags the lags are showing a seasonality. So we will do a automated SARIMA with 12 and 24 months.

**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.

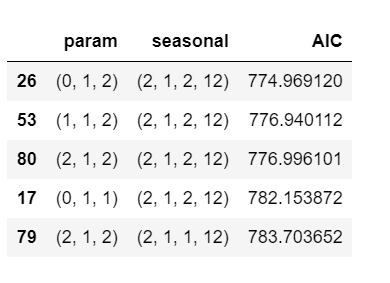
We have give the range for p,q from (0,3) and D range from (1,2) and d range is (1,2)



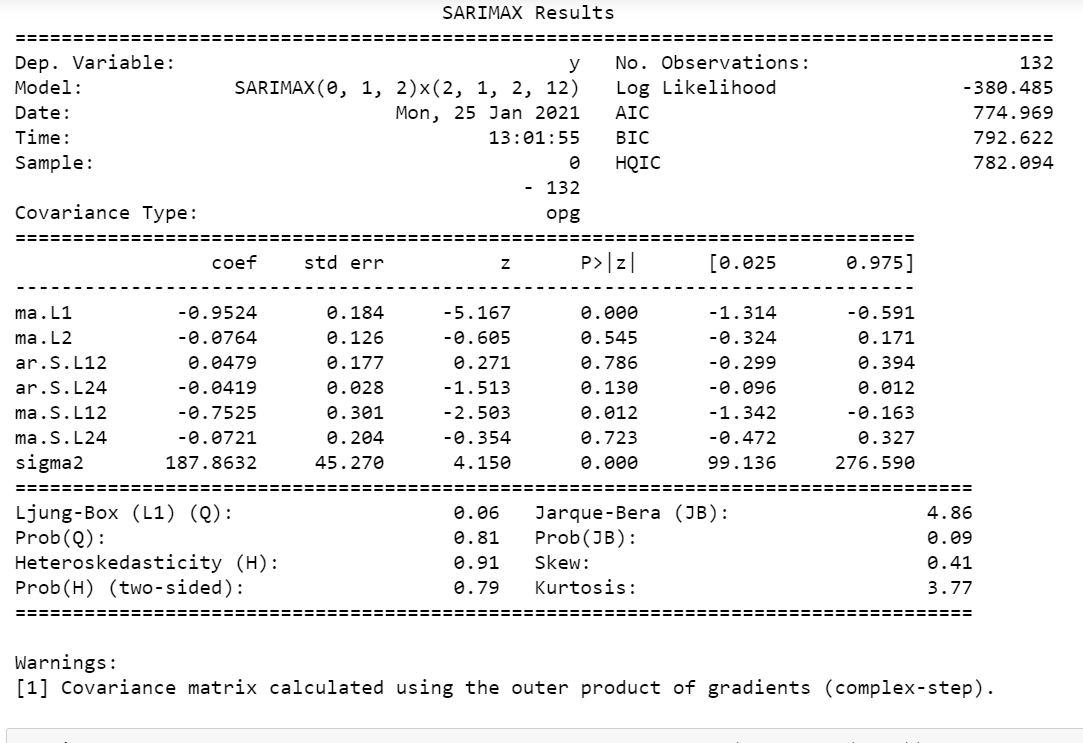
Above are the differenent combination that we get.



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

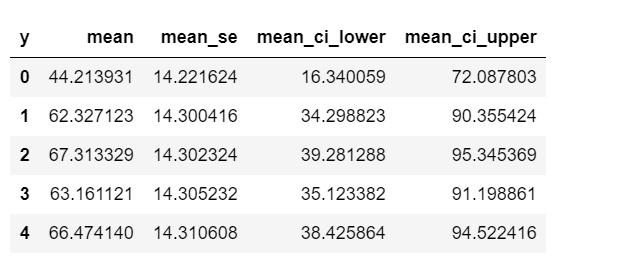


These are the best combination with least AIC



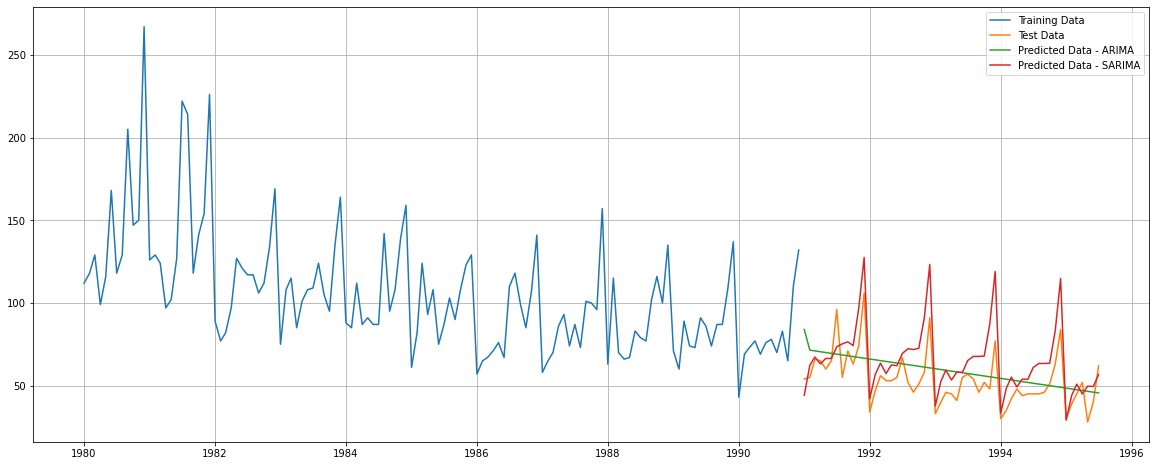
This table shows the different constant values. And based on P value we can know which all are significant if the P value is less than 0.05 then we can conclude those constants are significant.

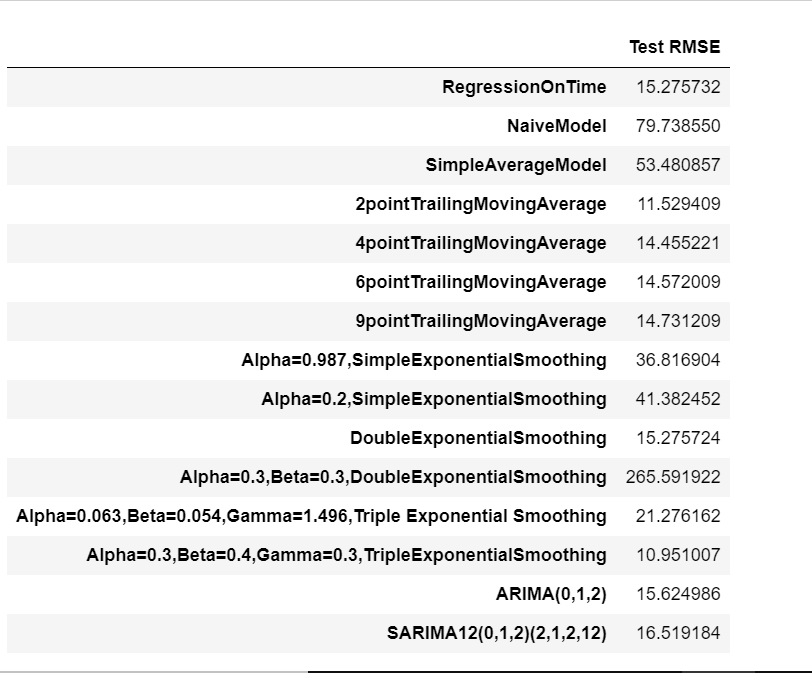
**Predicted mean for lower and upper confidence interval:**



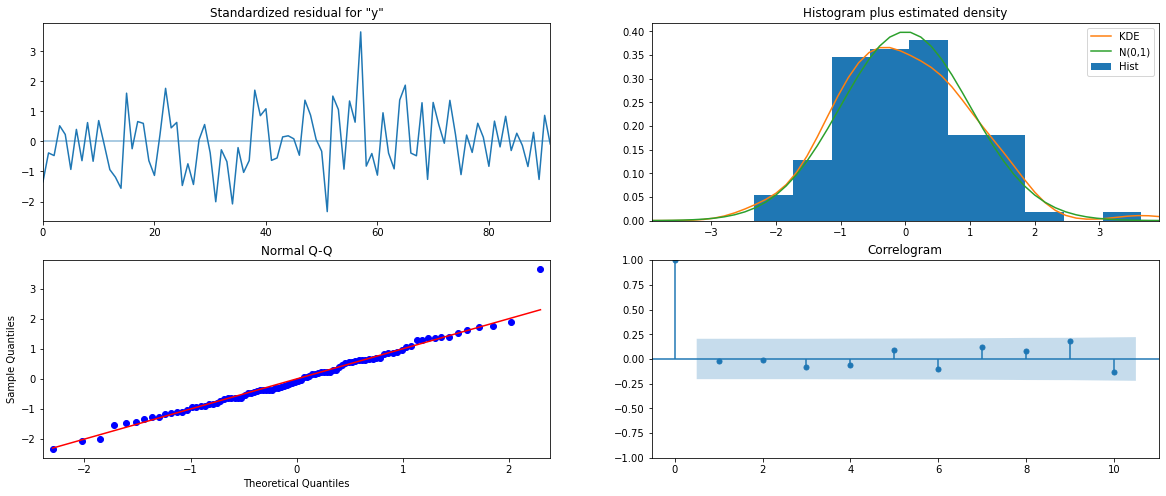
RMSE SARIMA 12 months:



The RMSE score for SARIMA 12 months is 16.519. The arima model performed better than SARIMA 12. 



Above is the RMSE of all the model that we have done so far.

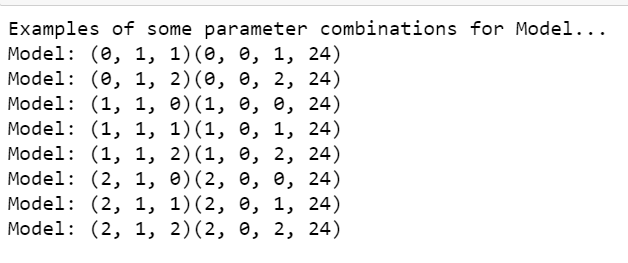


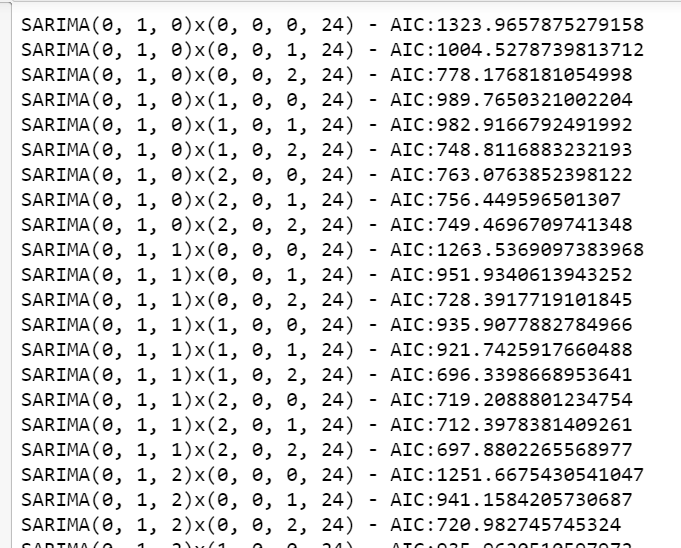
We can see the histogram look ok for the model. And the errors are almost on the line only few are outside the line.

**SARIMA 24.**

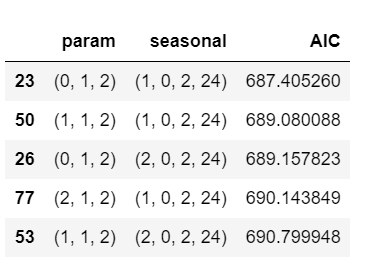
We will also see how the model performs with SARIMA as 24 months.

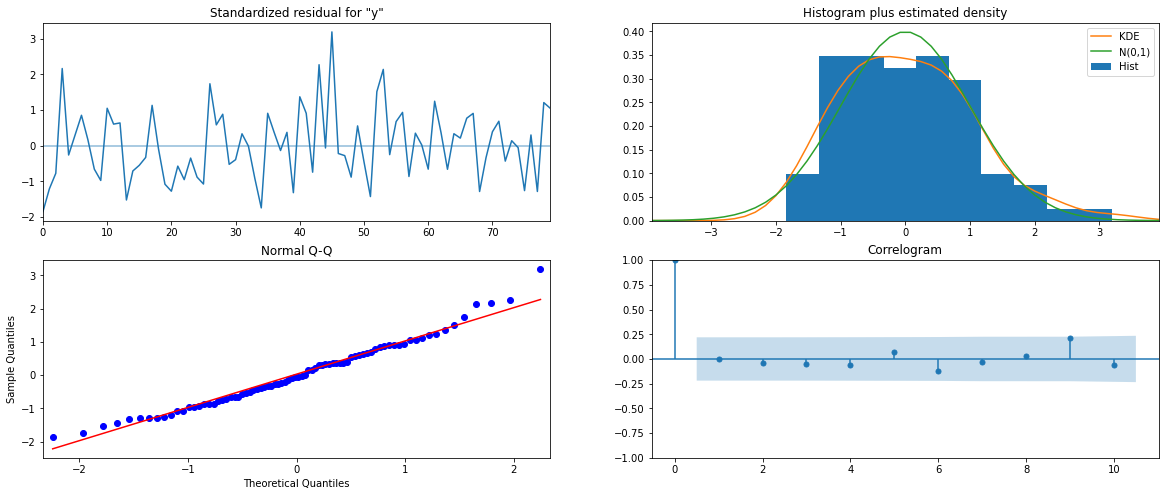
**Parameter combinations:**



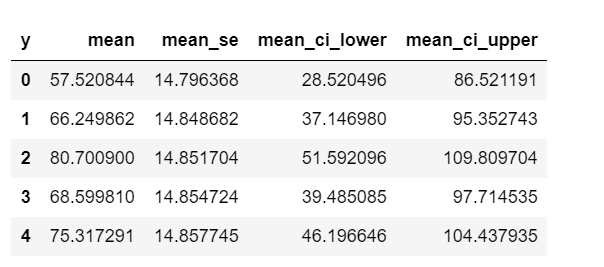


Best combination that gave least AIC for SARIMA as 24 monhs:

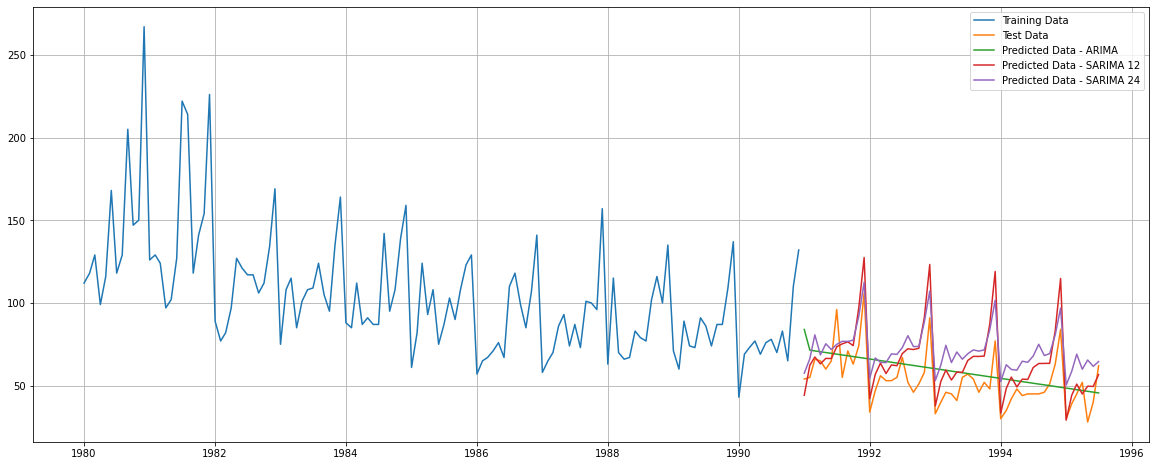




Forecasted mean lower and upper confidence interval values.

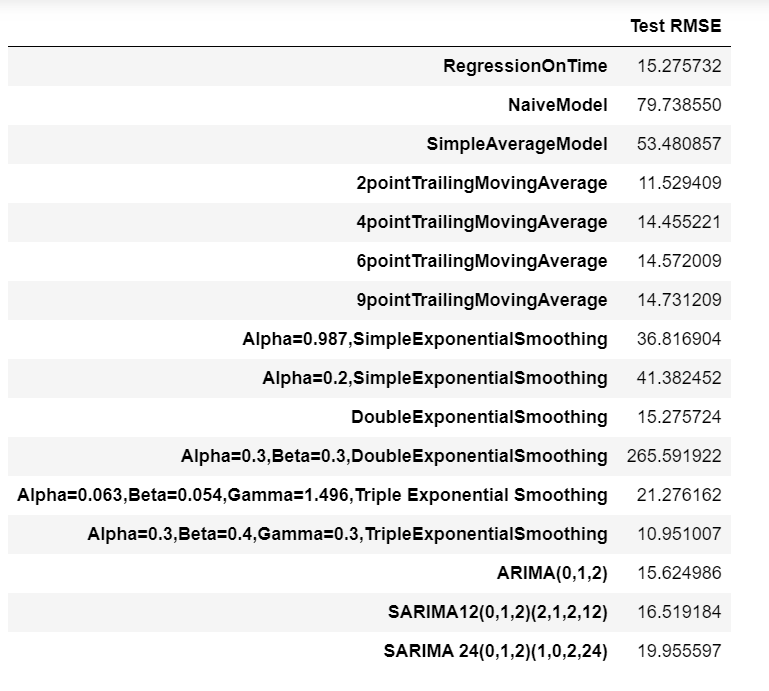


All the models forecasted graph:



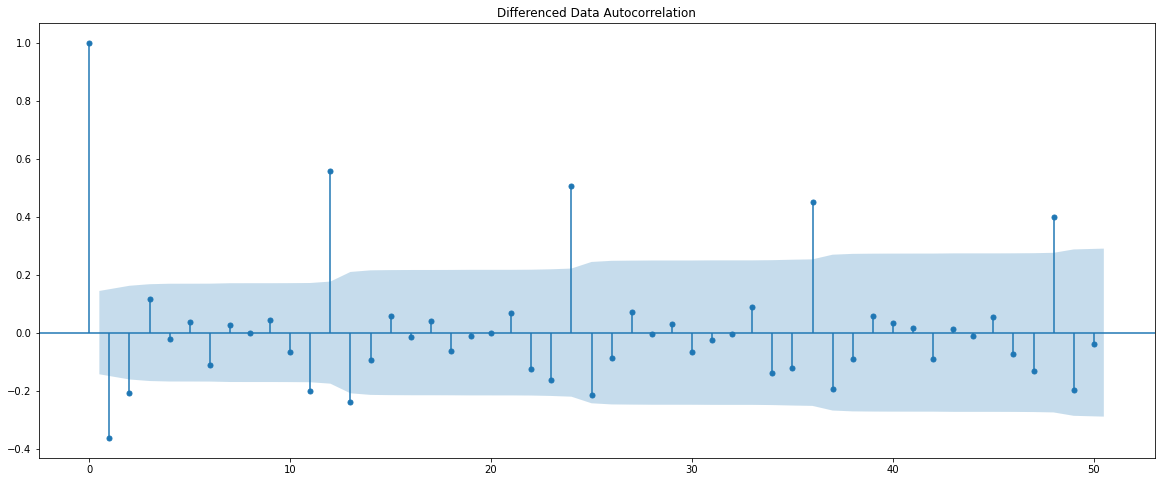
RMSE SARIMA 24 months:

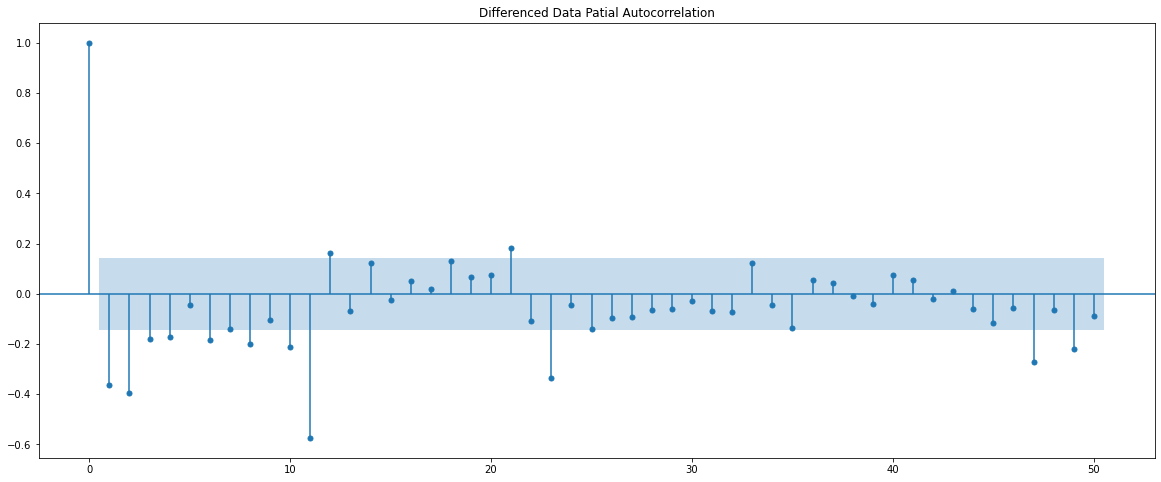




1. **Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**

[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.

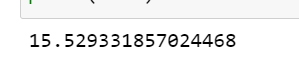
As per the PACF graph after the 4 th lag the 5th lag is inside the confidence interval band. So as pe the graphs the p value is 2 and q is 4. However, I am considering p as 2 and q as 3 for the manual arima calculation. And d as 1.



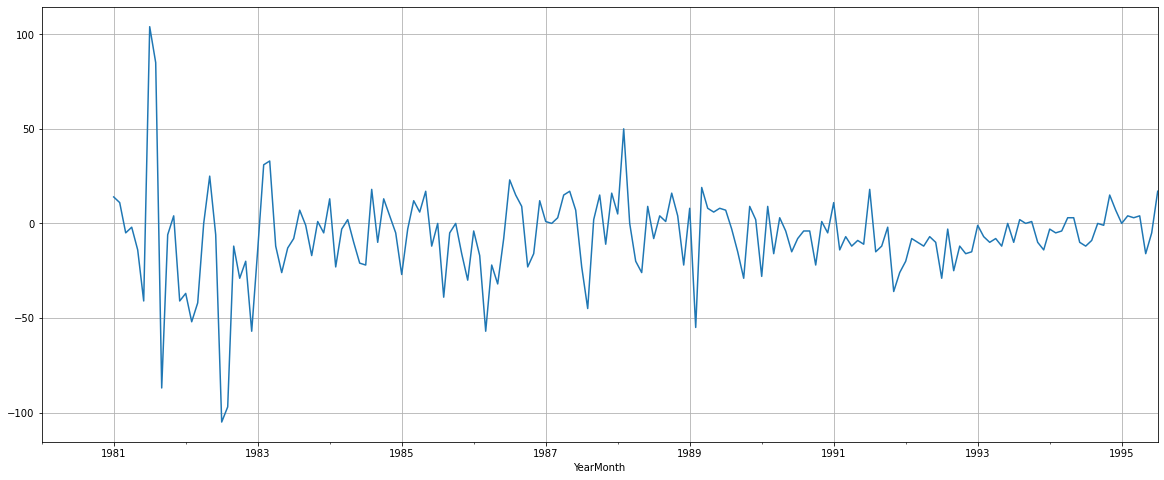
We can see in manual arima most of the constant are having higher p value. We can see the ar auto regression is having 3 constants. And all three are having high P value and they are not significant.

RMSE ARIMA manual:

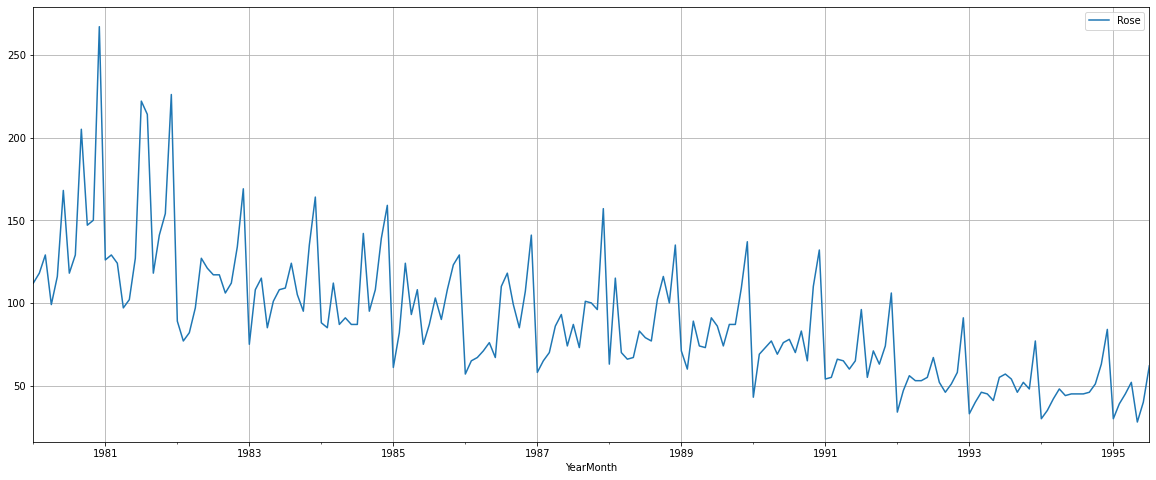
The RMSE is



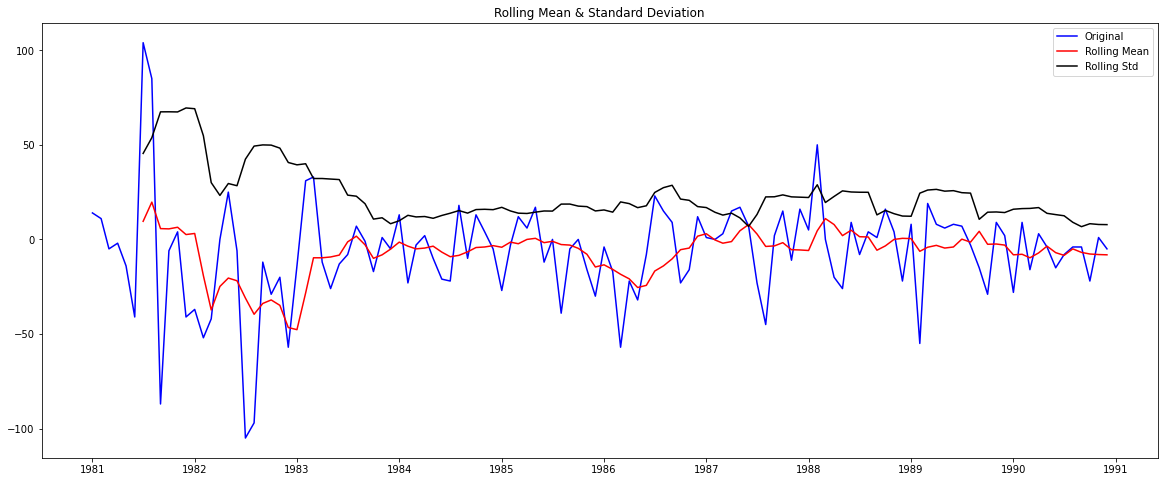
**SARIMA Manual:**

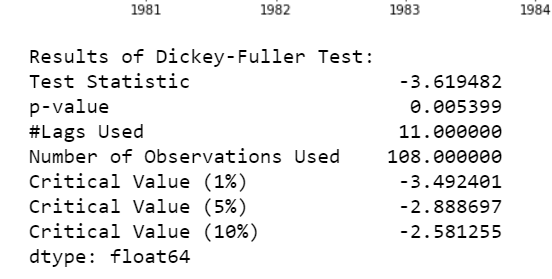
****

In the above graph we have used the difference of 12 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. And we applied a 12 month differentiate on this.





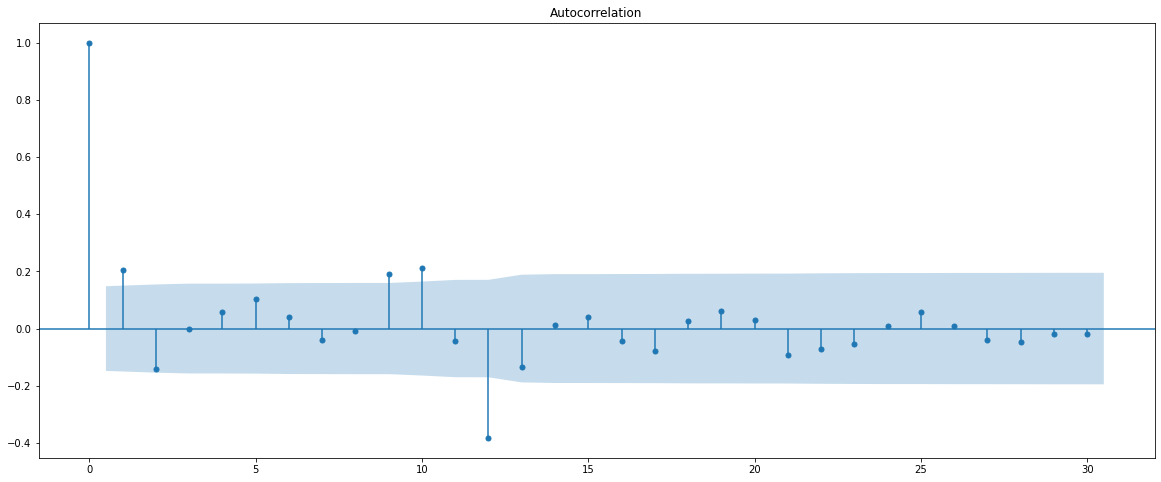
On the original series will diff as 12 we get a p value is 0.005.

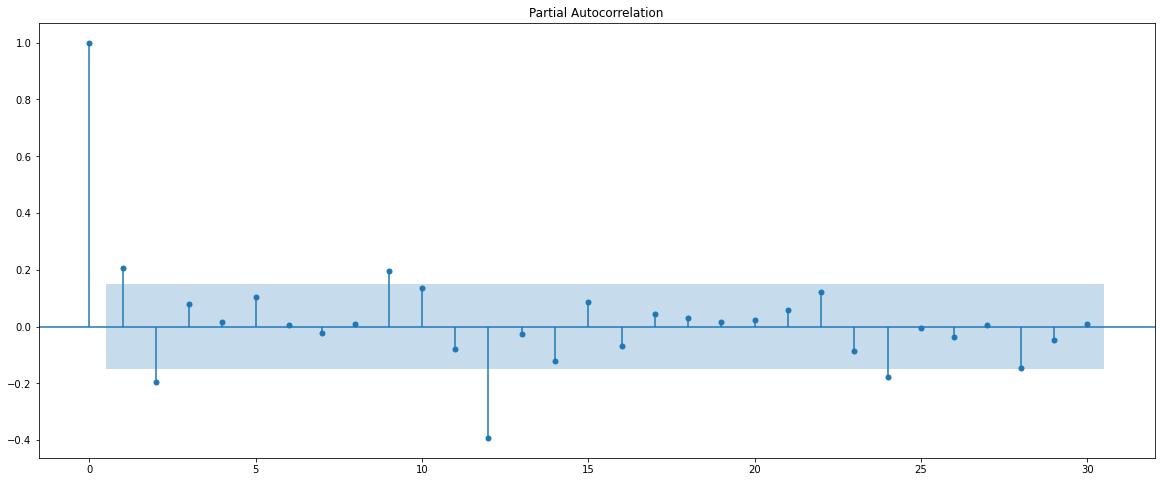
Ho: the data is non stationery

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.

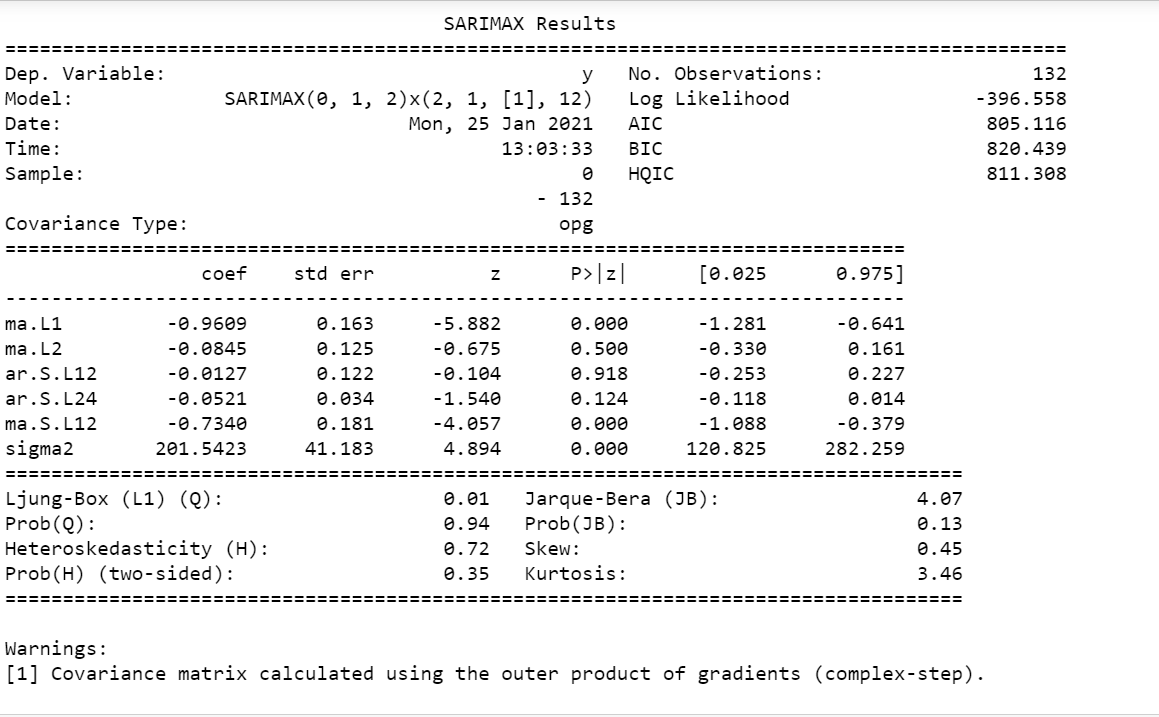
Below the ACF and PACF plot for SARIMA as 12 months



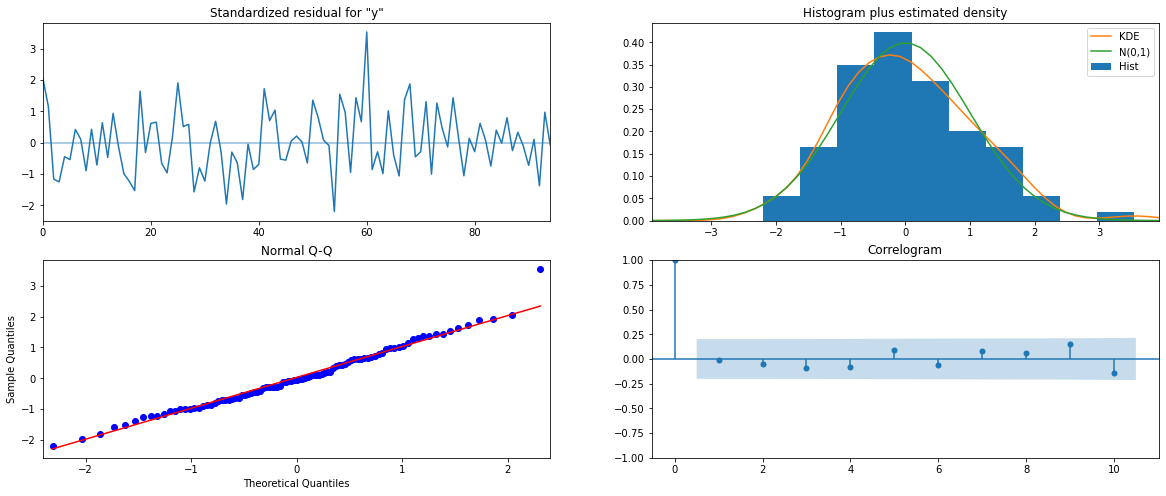


We can see the difference in the lags here when compared to the [original ACF and PACF](#ACF)

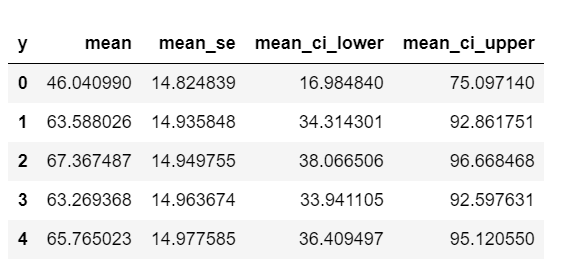
We can see that the p value which was earlier 4 is now 2. After the 2nd lag the 3rd lag is inside the confidence interval area. The q value is 1.



Above is the p values of the SARIMA 12 months constants.

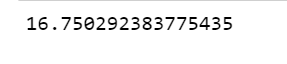


Above we can see the KDE is almost close to the green line. And very few values are outside the red line in the normal Q-Q image. All the lags are inside the CI area.

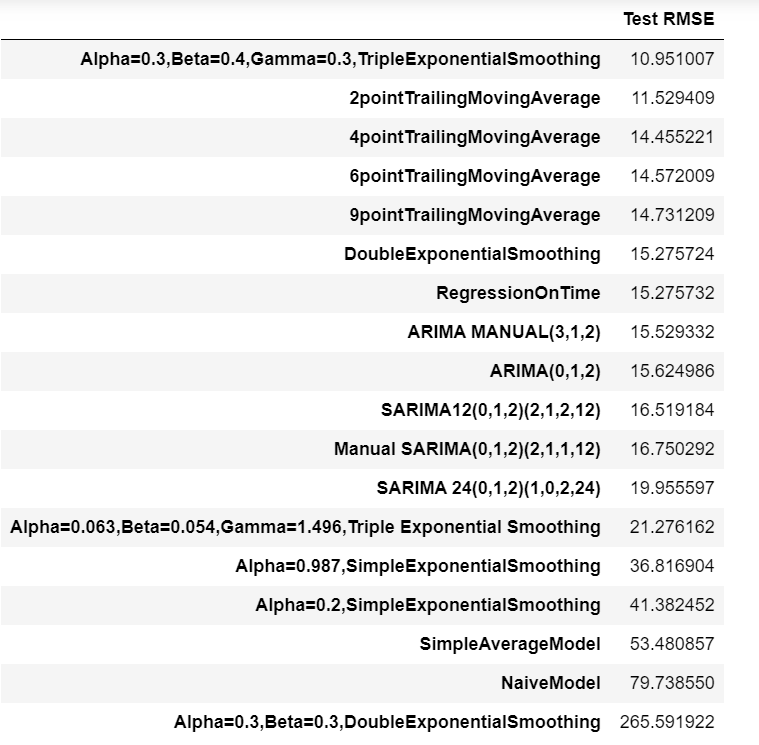


These are the upper and lower predicted confidence intervals.

The RMSE for the SARIMA 12 is



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 the table above with the RMSE scores of all the models that was performed I n this time series.

From the table we can see that the best model with the least RMSE score is the triple exponential smoothening model.

The parameters for this model are:

alpha = 0.3

Beta = 0.4

Gamma = 0.3

The RMSE for this model is 10.95. followed by 2-point trailing methods are others.

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.**

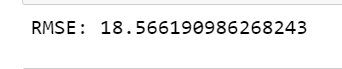
The most optimum model to be built is using the triple exponential method since that gave the least RMSE value. So we will built the full data based on this. The original data frame is built using the parameter of the triple exponential smoothing.

The smoothing level is set to 0.3

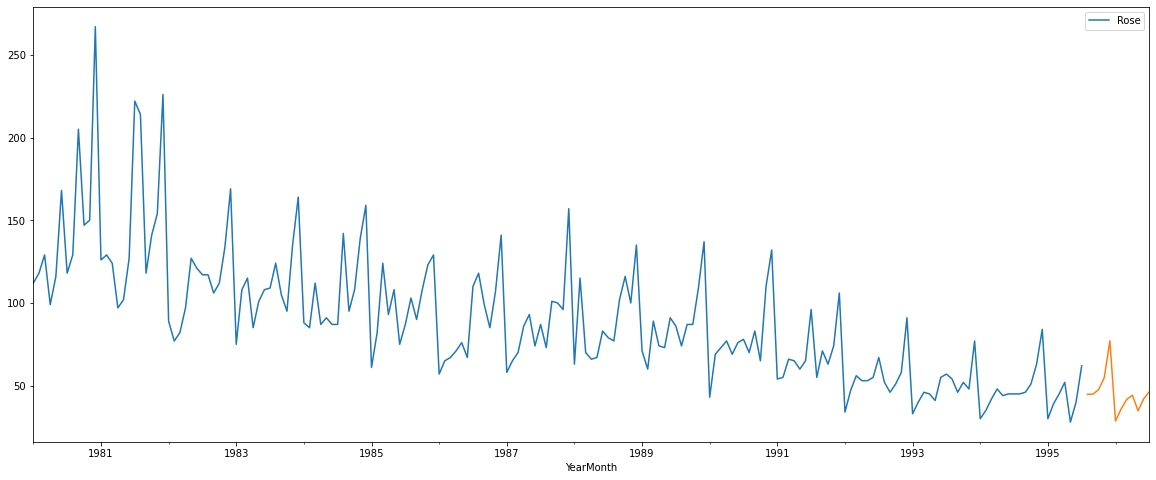
Smoothing trend = 0.4

Smoothing seasonal = 0.3

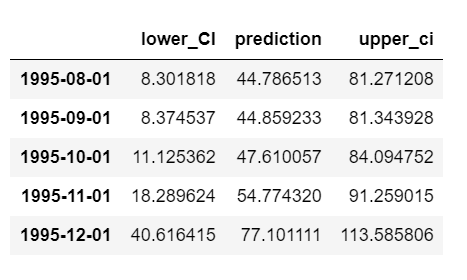
The RMSE for this model is



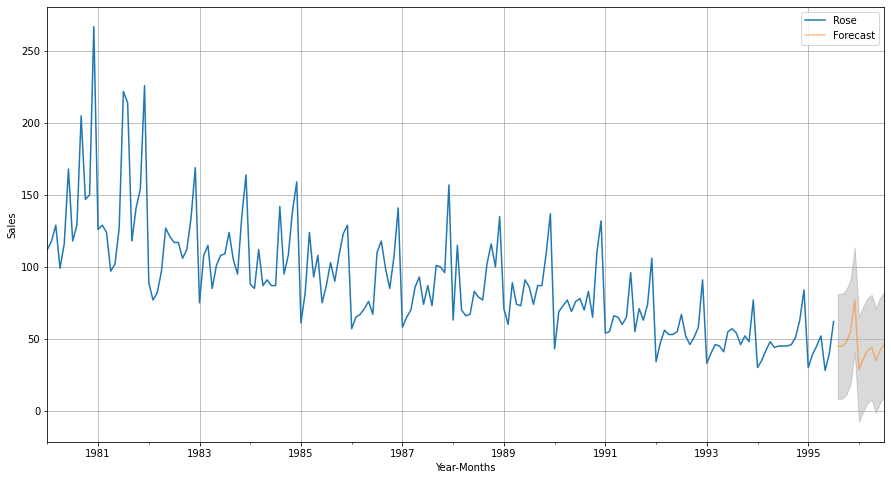
**Prediction Plot.**



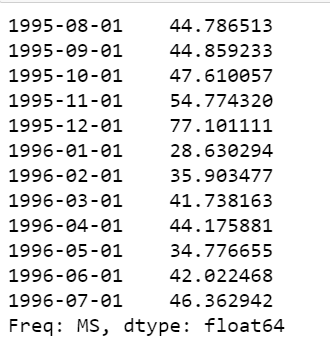
I have predicted the forecast for 17 months. Since the task was to predict for months and our data had 5 months of missing values, I predicted for total 17 months. We can see the prediction is almost like the 1995 and data. Similar to the last years data.



These are the prediction lower, upper and the prediction values. So, this table shows us the range that the prediction can go higher or lower.

Above is the forecasted area in orange and the original series in blue. The grey shades are the upper and lower CI and the orange line is the predicted figures.

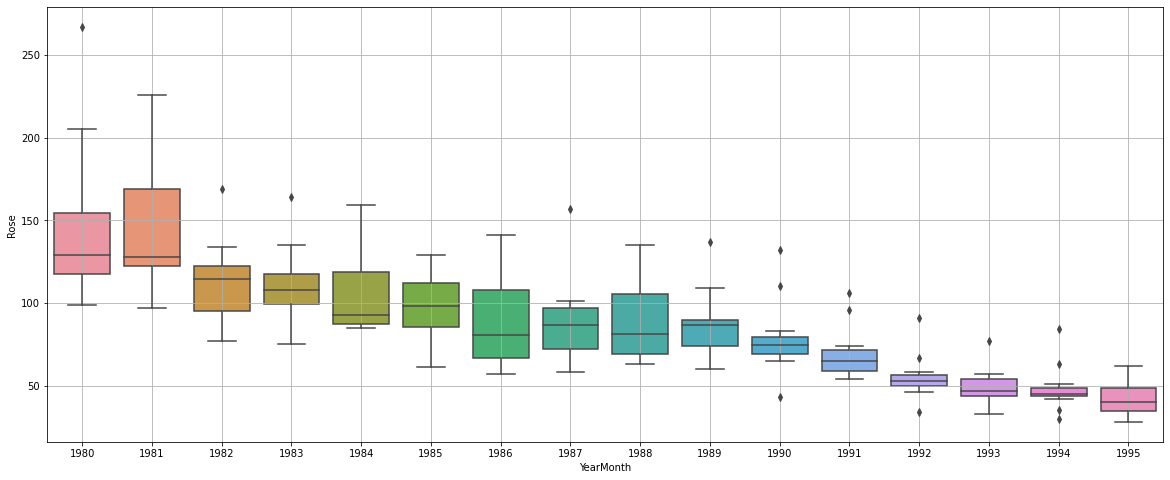
**Prediction sales:**



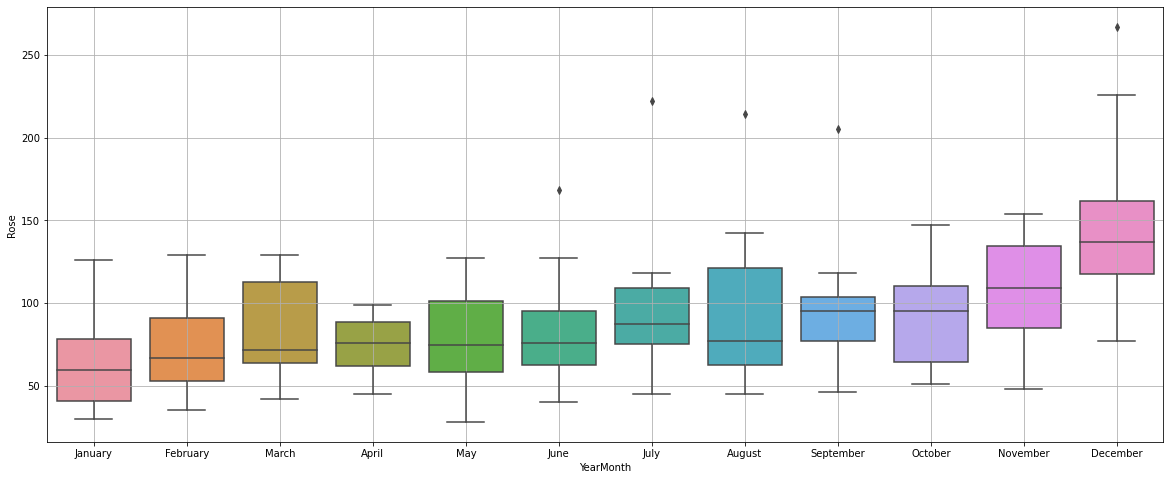
1. **Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**

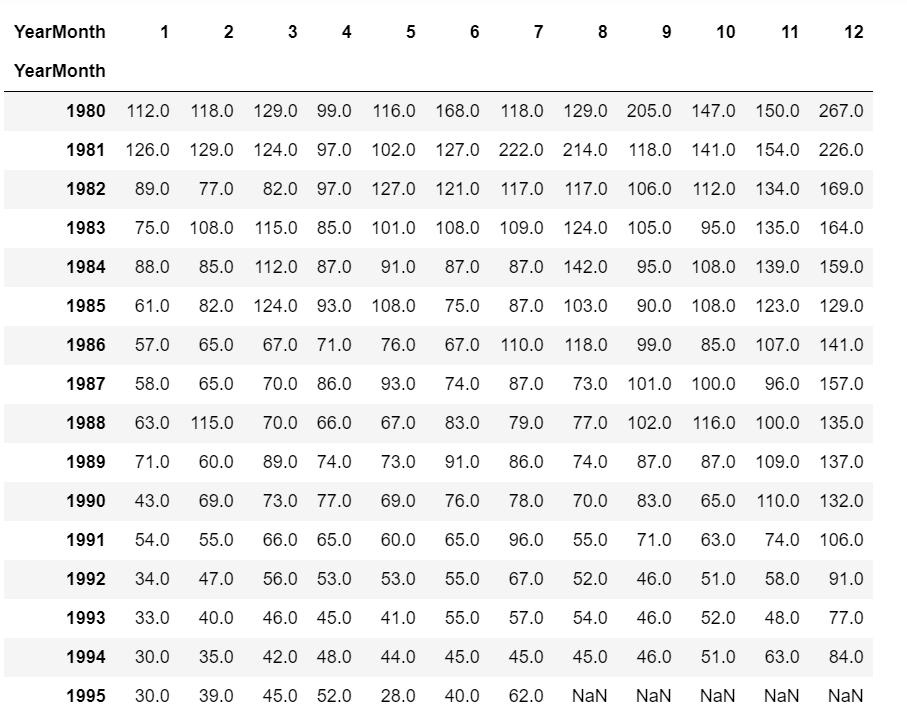
We have used various models to analyses the Rose time series and from the various model we have got the triple exponential model which gave us the best RMSE score. The lower the RMSE the better the model. RMSE means the error in that model is less.

The Rose wine series had a downwards trend in the series and there was seasonality also. Apart from the first two years there was a steep downward trend. The sales was reducing drastically.



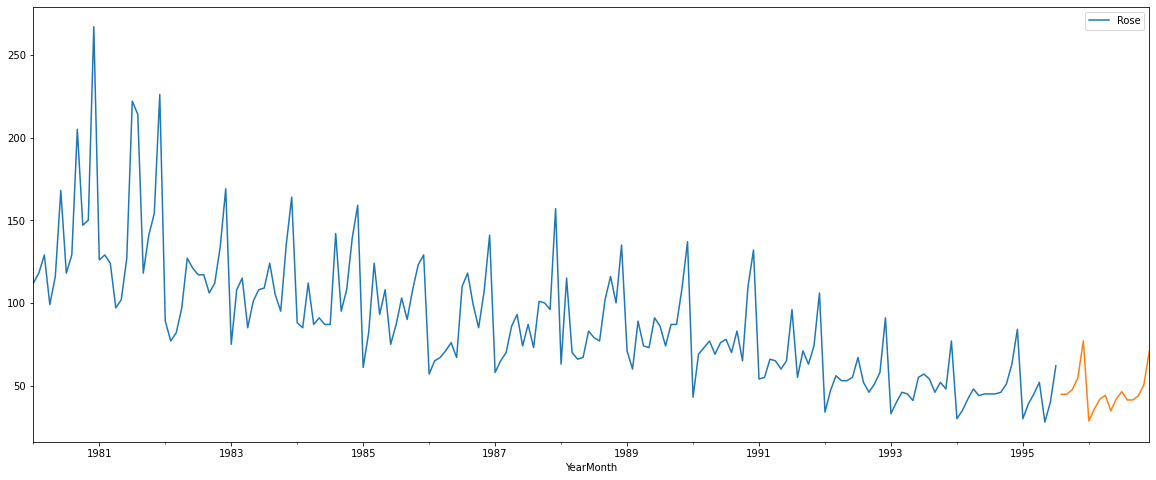
From the above yearly bar plot we can clearly see that the first two years had incredibly good sales. And since that every month the sales has dropped.

The above monthly plot shows the seasonality in the data. In October, November, and December the sales are highly compared to other months.

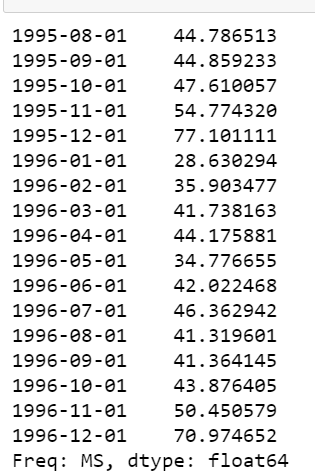


Above is the table that shows the sales of every month from 1980 to 1995. We can see that the peak season of December in 1980 had a sale of 267 units of wine. And after 14 years in 1994 December the sales are just 84 units which is a whopping 217% reducing in sales for rose wine.

After running all the models to find the best RMSE score the triple exponential is the best model. We have to fit the finding of this model on our original data frame so that we can predict the future sales of the Rose wine.



Above is the prediction we have done for total of 17 months that is from 1995 -08 to 1996 -12. We can see that the prediction is almost similar to the previous years.



**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 predicted sales are also following a downward trend.
* During the end of the year the sales is increasing in the forecasted sales also.
* Company has to analyze why their sales have been reducing consistently. Is it that their wine is not good in the market? Have they changed any ingredients that sales have been declining?
* They should see what wine is selling more in the market and see why they are selling more. What is the difference between Rose wine and the one which is selling more?
* The company can also be prepared in not stocking up more wine in production since the forecast says the sale will be less.
* Since they are selling more towards the end of the year, they can give some special offers or do some advertisements during the holiday seasons to increase the sale more during October, November, and December.
* They can also re brand themselves and launch their products so that customers can be attracted. Since they are in the market for more than 15 years, they can focus on attracting customer with their brand image.
* They can also see who all their loyal customers are and can promote their products to them more.