BUSINESS REPORT

Time series Data

Rose wine

**Index-**

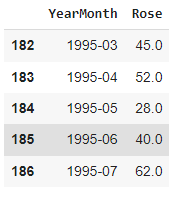
1. Read the data as an appropriate Time Series data and plot the data. [Page 2-6]
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. [Page 6-20]
3. Split the data into training and test. The test data should start in 1991. [Page 21-22]
4. Build all the 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. [Page 23-31]
5. Check for the stationarity of the data on which the model is being built 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. [Page 31-33]
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. [Page 34-42]
7. 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.

[Page 43]

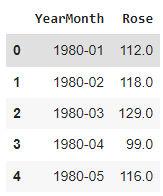
1. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands. [Page 44-49]
2. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. [Page 50-51]

**Q1. Read the data as an appropriate Time Series data and plot the data.**

Have a look at head and tail of the data set.



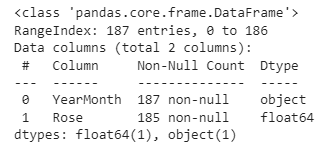
This is tail of data set this ends at 1995-07



Data set starts with the first month of 1980. Sales data is available from 1980 to 1995.

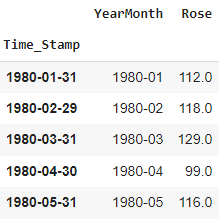
Let’s check the shape of the data.

Shape: The data set has a shape of (187, 2), indicating that it contains 187 rows and 2 columns. Each row represents a specific time point, and the columns are 'YearMonth' and 'Rose', respectively.

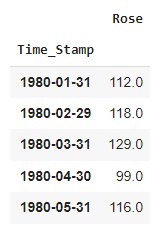


Data types of both the columns are object and int type for ‘YearMonth’ and ‘Rose’ respectively.

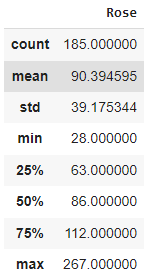
We have made a new date-time column and set it as an index to make this data time-series data.



After dropping YearMonth column we have final data set as -

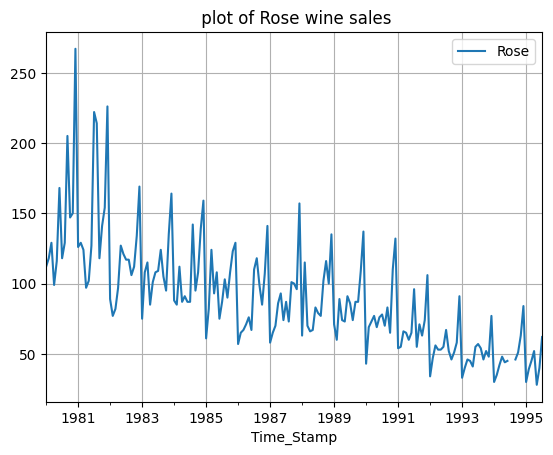


Let’s check the distribution of sales of sparkling wine-



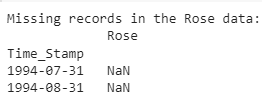
* Count: There are 185 data points in the dataset.
* Mean: The average value of the dataset is 90.394595.
* Standard deviation (std): It measures the dispersion or spread of the dataset. In this case, it is 39.175344, indicating that the data points are spread out around the mean.
* Minimum (min): The smallest value in the dataset is 28.
* 25th percentile (25%): 25% of the data points are below or equal to 63.
* 50th percentile (50% or median): 50% of the data points are below or equal to 86.
* 75th percentile (75%): 75% of the data points are below or equal to 112.
* Maximum (max): The largest value in the dataset is 267.

Distribution of sales of sparkling wine over few years of time is like below:



A pattern of decrease in sales can easily be seen along the months, and passing by years.

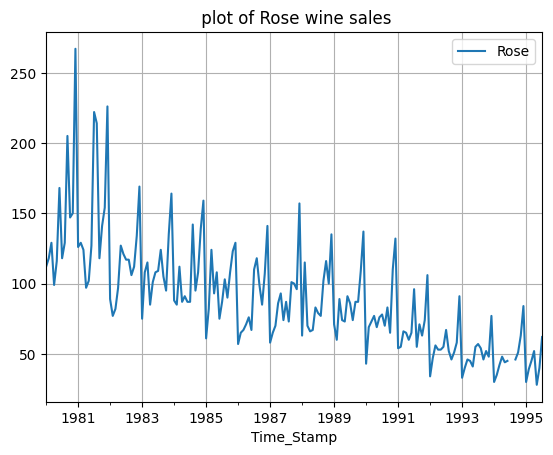
We have 2 missing records in data-



We use linear interpolation to impute these values-

He these are replaced as-



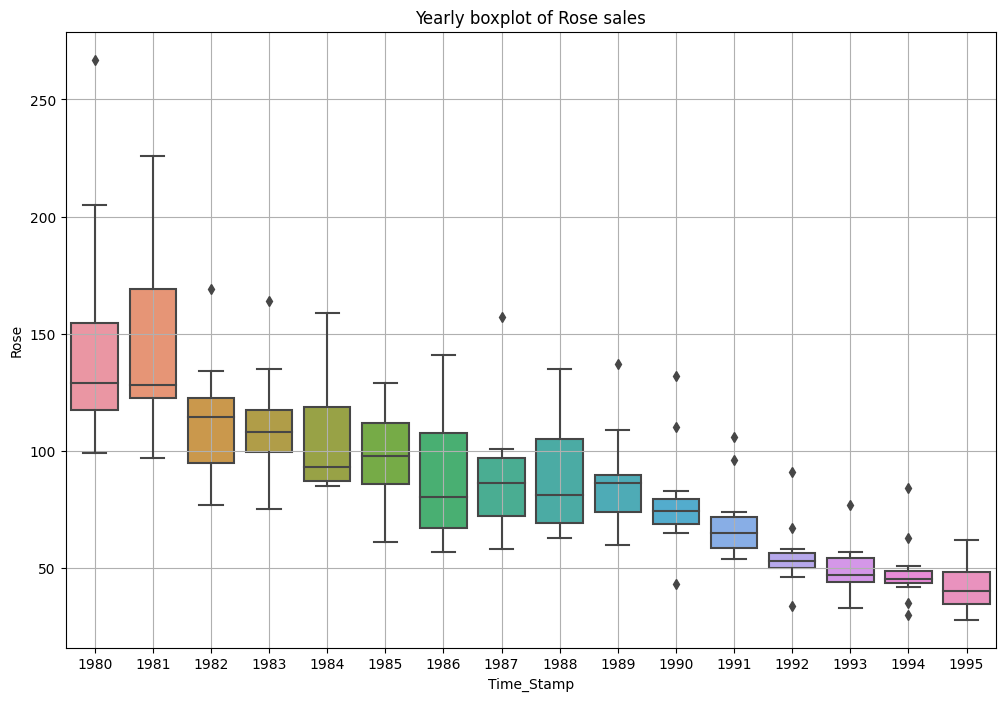


A decreasing trend in sales can be seen as years pass by.

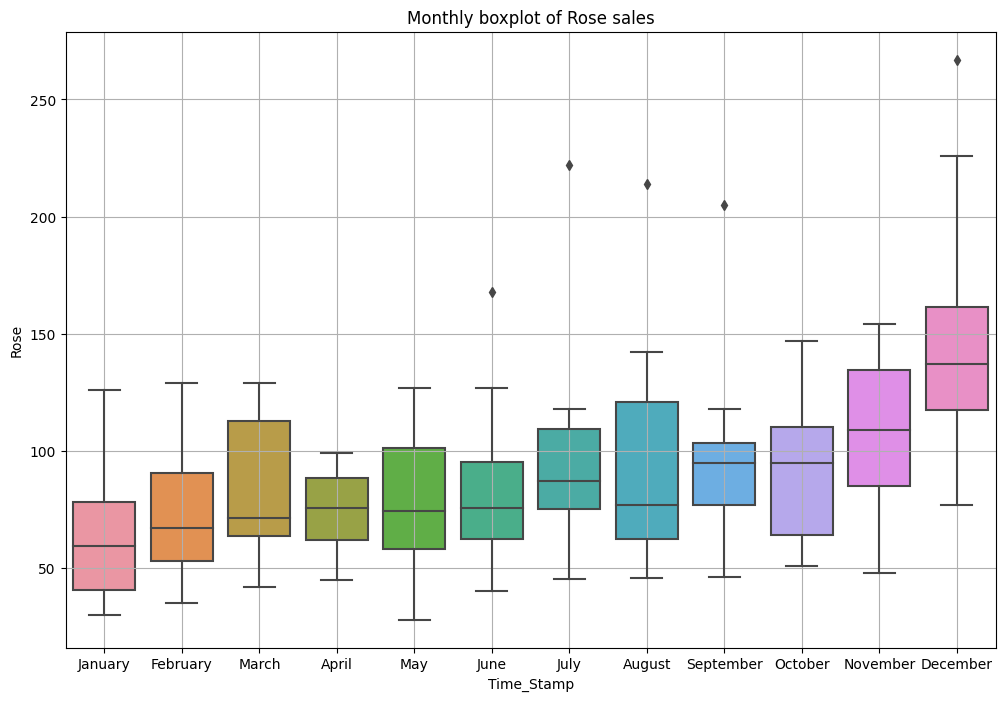
**2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

Let’s perform EDA-

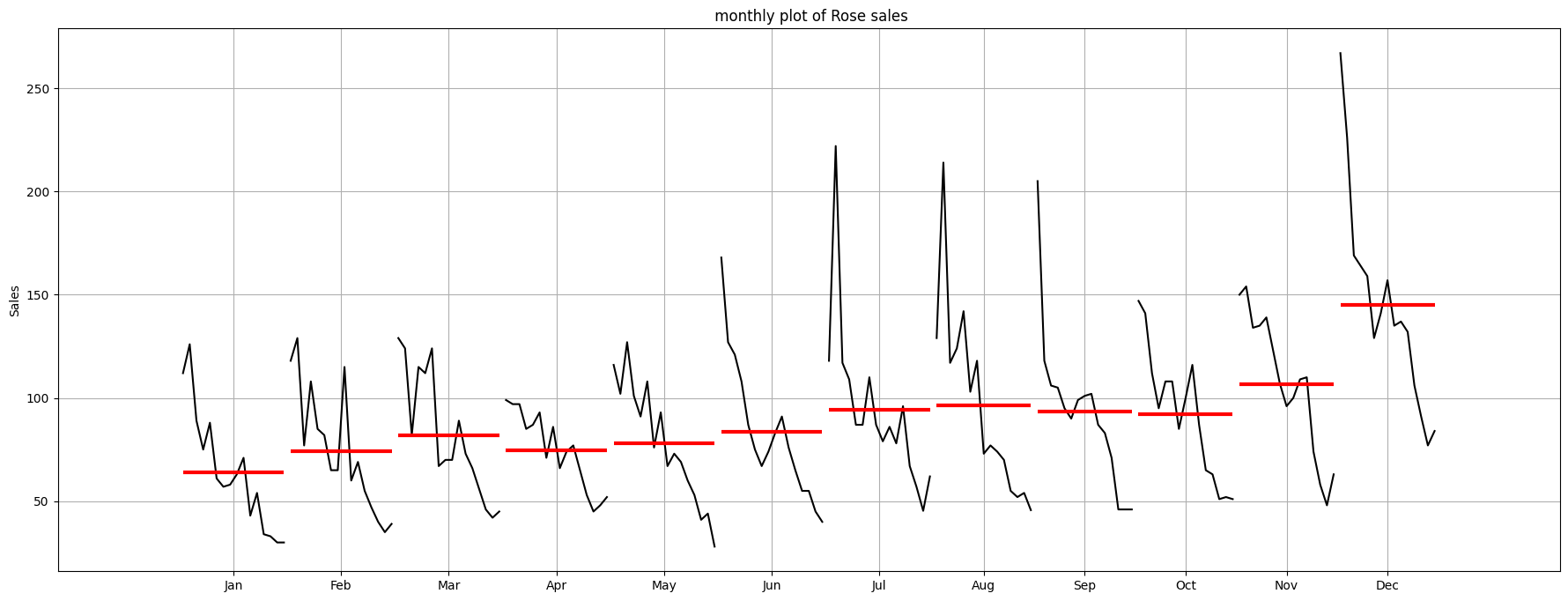
Box plot-



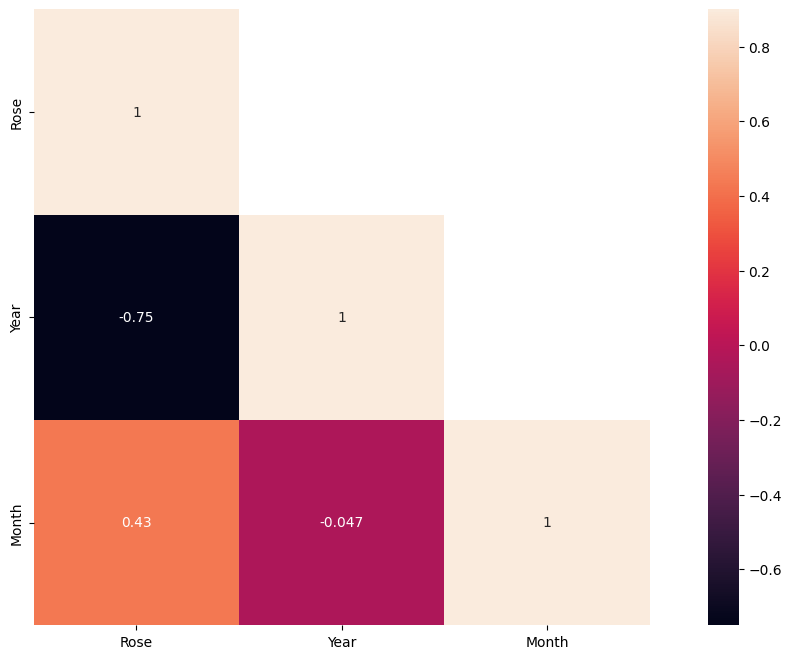
* Some outliers can be seen in data as this shows that some of the months have done exceptionally well and rose wine sales have gone up in those months.
* Mean of sales along the year has gone down starting from 1980 to 1995.
* As the years pass by , the overall yearly sales and the variability in monthly sales has decreased.



* There is a gradual upward trend in sales from January to October, indicating a consistent growth pattern.
* However, there are several standout years for the months of June, July, August, September, and December, where sales significantly surpass the sales of other years in those respective months.
* To maximize sales, it would be beneficial to pay special attention to November and December, as these months consistently demonstrate peak sales performance.

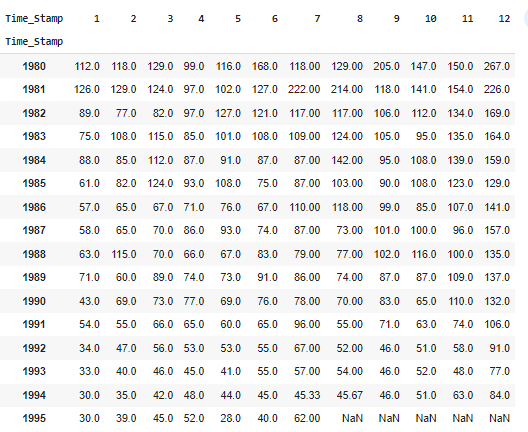


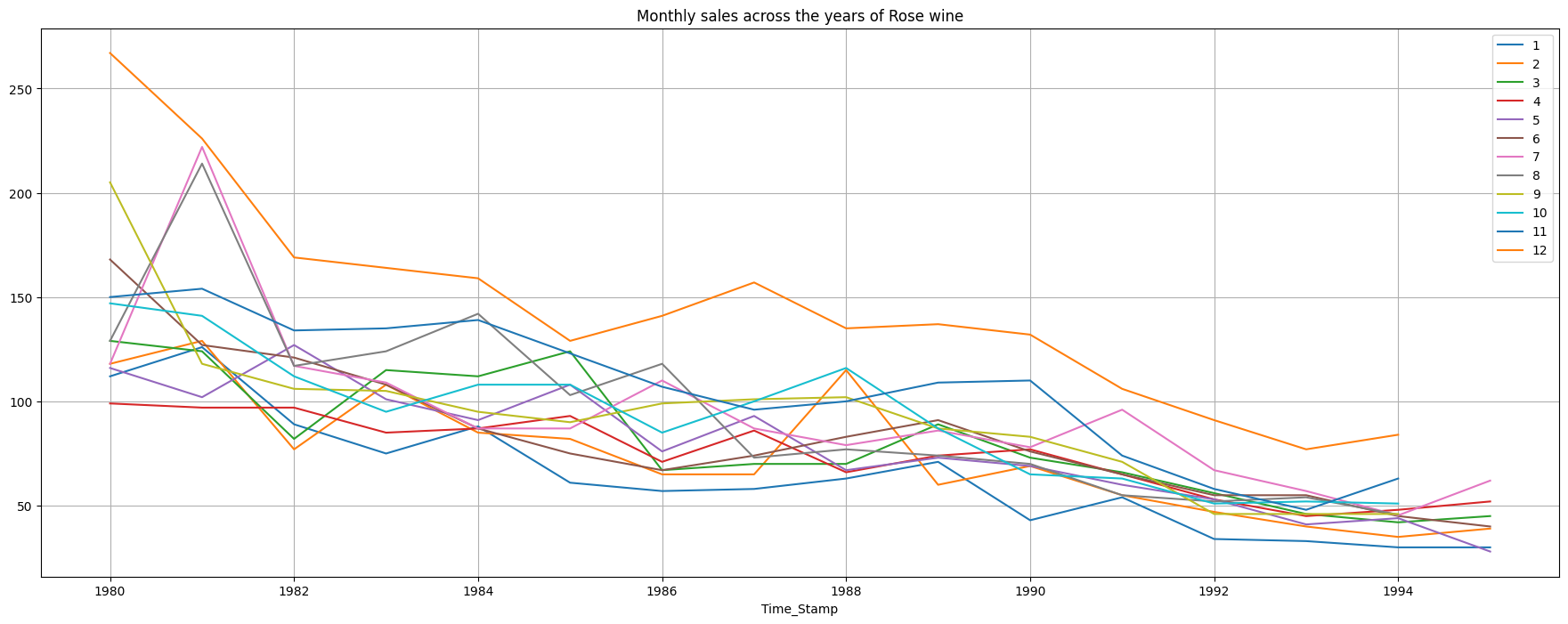
Heatmap-

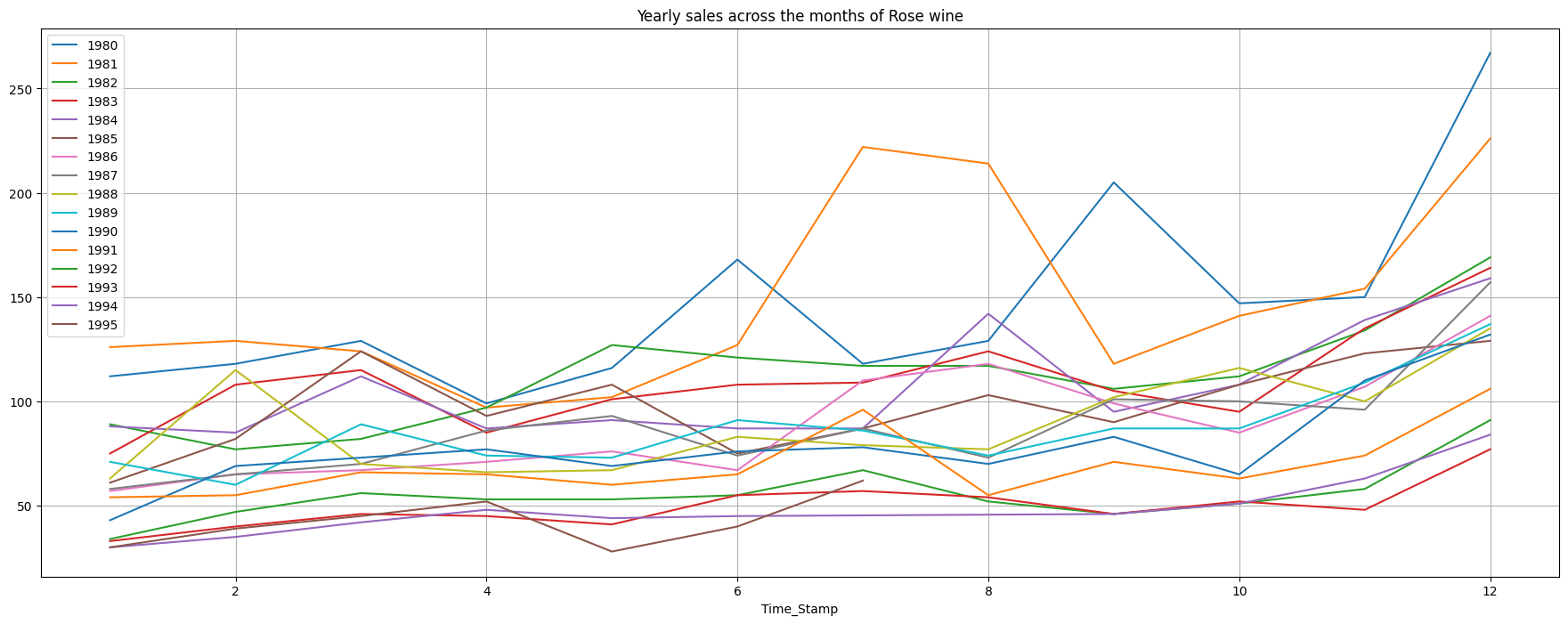


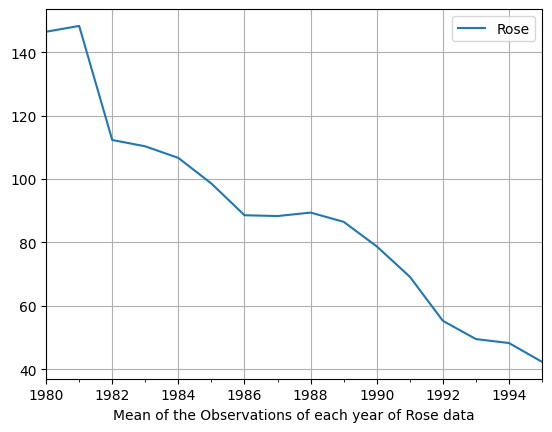
High correlation can be seen between year and rose wine sales. It’s a negative correlation as the years pass by the sales have gone down.

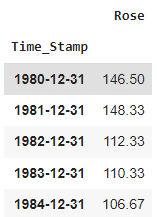
**This is monthly sales by year.**





1. 1. Overall decreasing Trend: There is a general decreasing trend from 1980 to 1994 
2. Seasonal Patterns: There seems to be a recurring seasonal pattern in sales. For example, there is a noticeable increase in sales towards the end of the year (months 10-12), suggesting a possible holiday season effect. This pattern is consistent across multiple years.

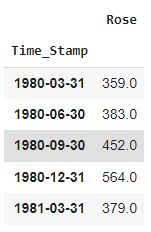




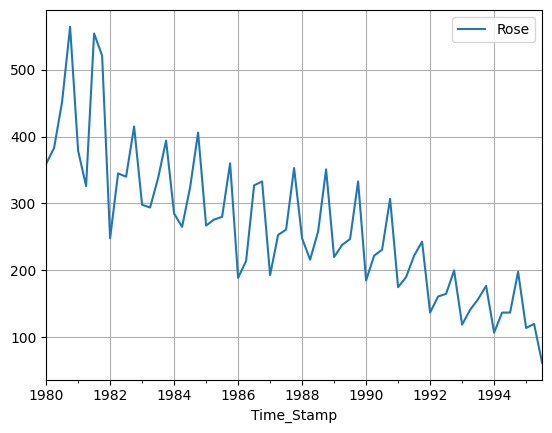
Mean of yearly sales, maximum sales is in year1980. Sales have decreased from 1980 to 1995.

Let’s check data quarterly.

Sum of quarterly sales-

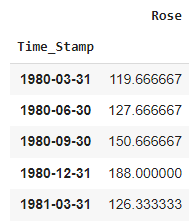


Quarter sales looks like-

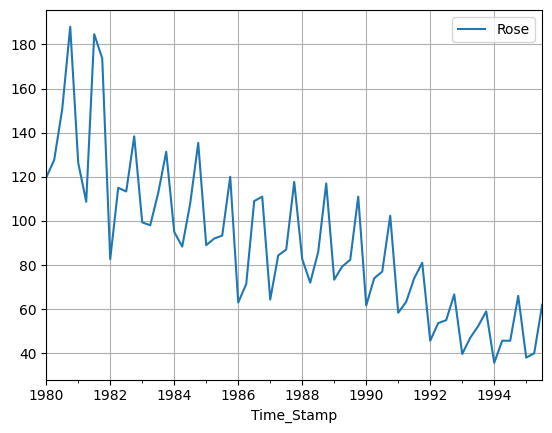


A pattern can be seen,with peak sales in a selective quarter.overall trend is decreasing

Mean quarterly sales-

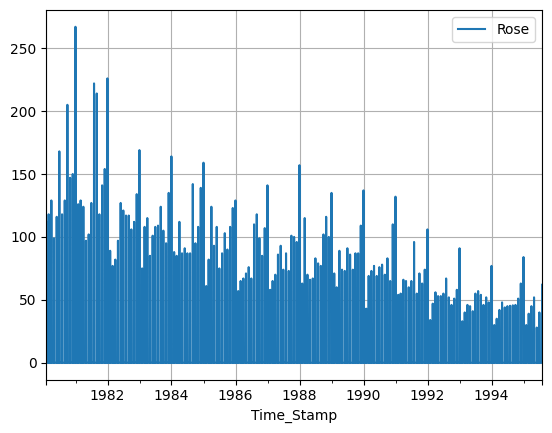


Shows a similar pattern with decreasing sales year on year.



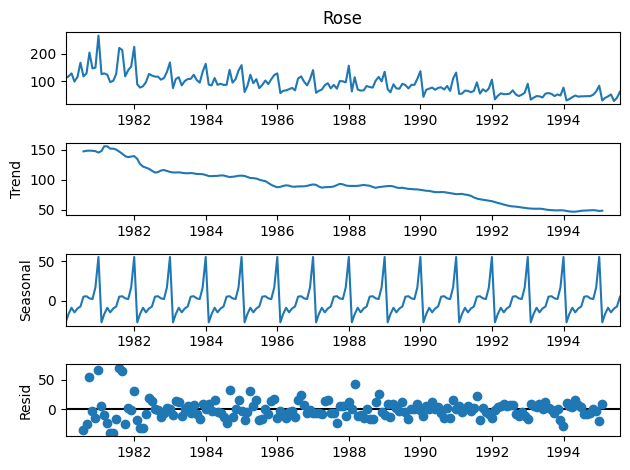
Let’s have a looks at daily sales too

It looks like this , with peaks on few days and overall decrease in sales -



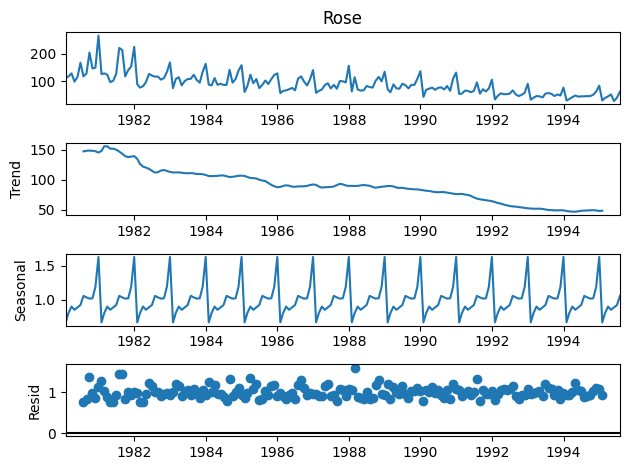
We are now going to decompose the Data-

Additive decomposition-

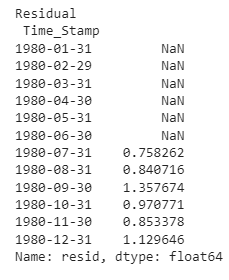
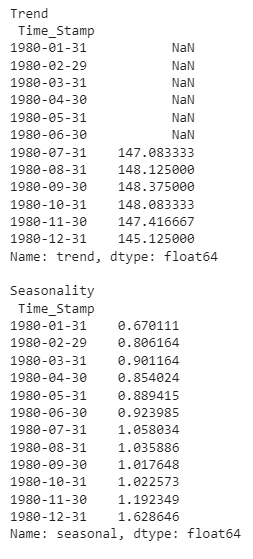


* We can see trends in data. It decreases from 1983 to 1994 , then fluctuated
* Seasonality can be seen in data easily. It has repetitive peaks in a few months along moving years.
* Residuals are along the zero line but are also deviating from zero till 50.

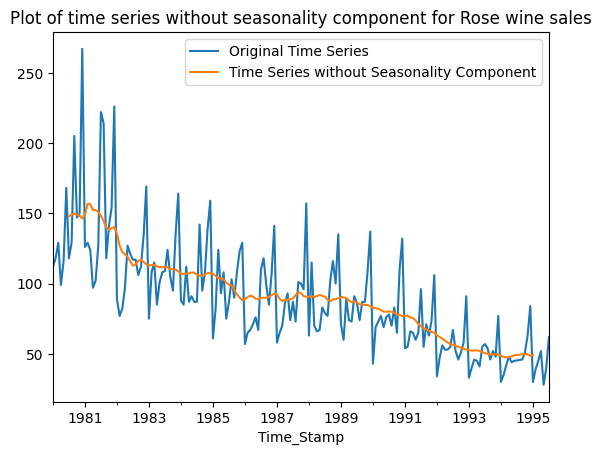
Multiplicative decomposition-



* Same trend and seasonality can be seen in this decomposition too, same as additive.
* For the multiplicative series, we see that a lot of residuals are located around 1. Thus Multiplicative Decomposition is the right way to decompose the time series.

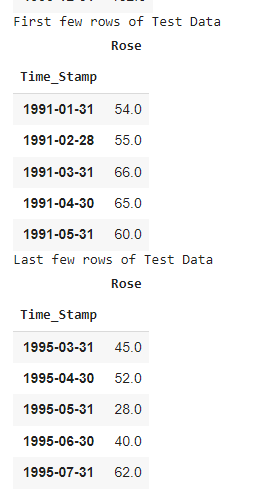
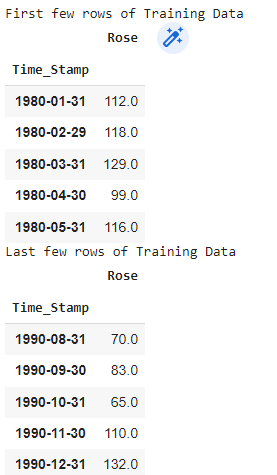


This is according to Multiplicative decomposition, residuals close to one can be seen.



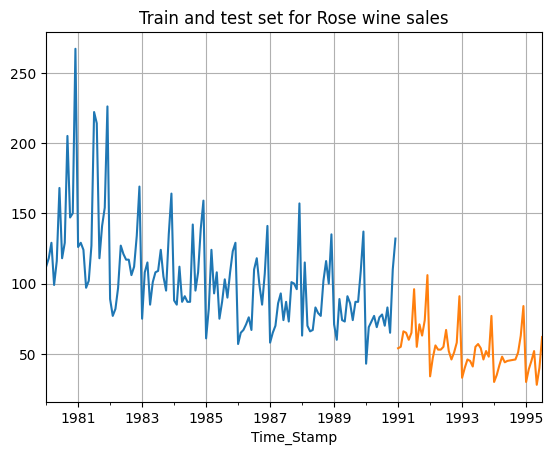
**3. Split the data into training and test. The test data should start in 1991.**

After splitting the Data , here are few rows of test and train data.

****

Shape of train set (132, 1)

Shape of test set (55, 1)

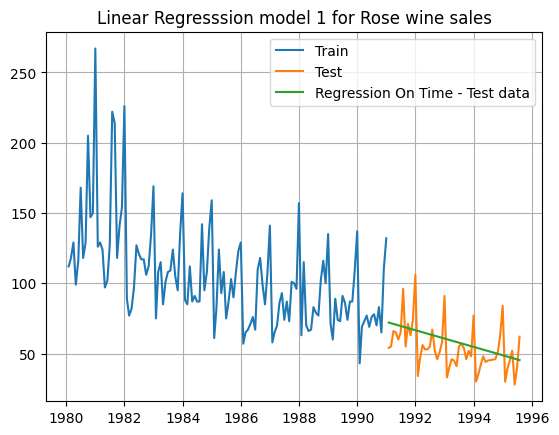
****

Data looks like this after splitting.

**4. Build all the 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.**

All the models are built and these RMSE are recorded and plots are made to visualize how well the model is predicting.

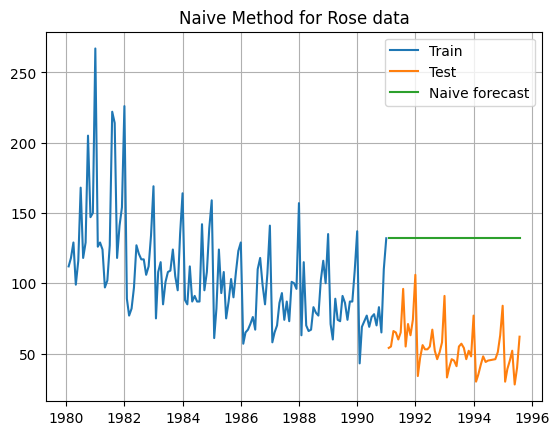
**Linear regression-**



For RegressionOnTime forecast on the Test Data, RMSE is 15.27

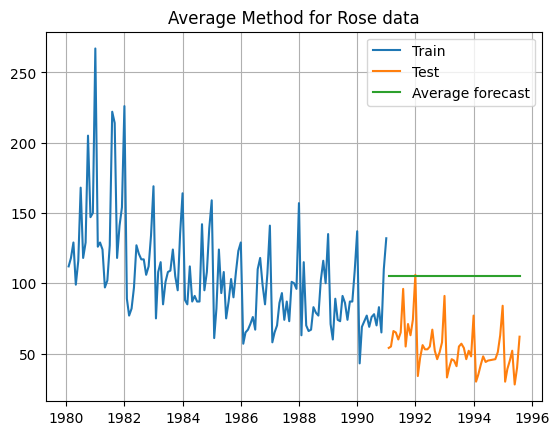
It is just a straight line trying to predict patterns.

**Naive forecast-**

****

For Naive forecast on the Test Data for Rose wine, RMSE is 79.72

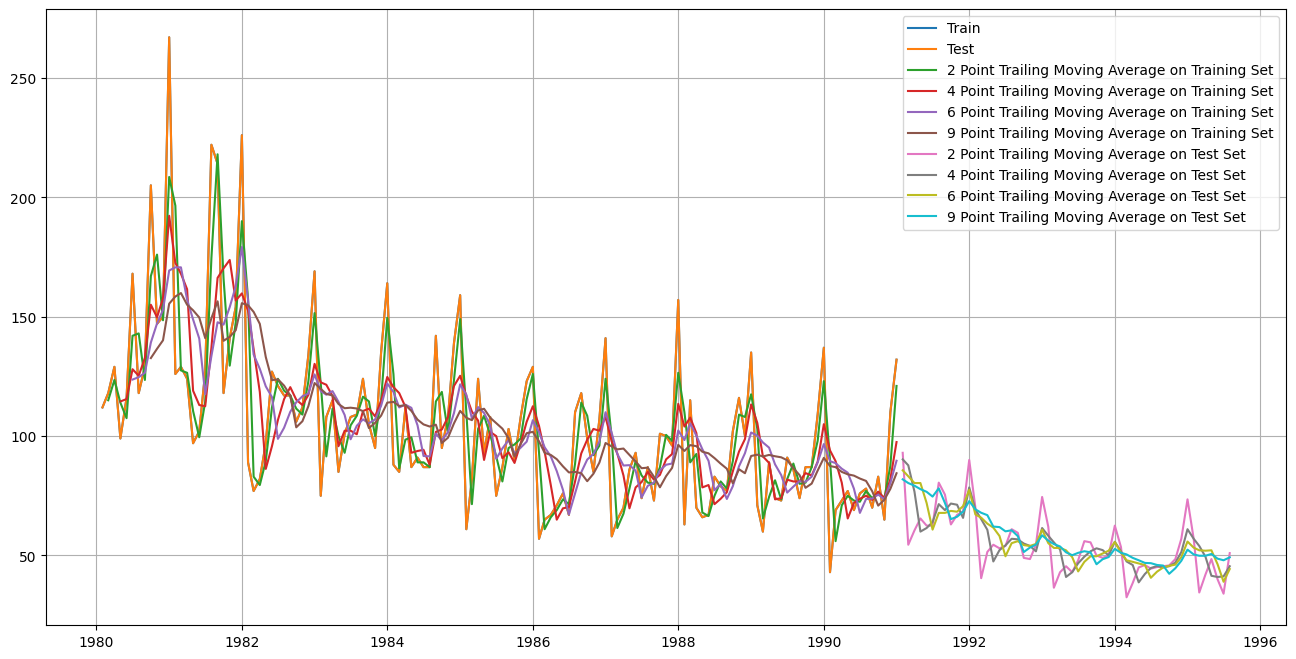
**Simple average forecast-**

****

For Simple Average forecast on the Test Data of Rose wine, RMSE is 53.46

Better than naive but still a straight line trying to capture patterns.

**Moving Average-**

****

Its able to capture some patterns on Test data lets check for RMSE for all the Moving average models-

* The 2-point trailing moving average represents the average of the current month's value and the value from the previous month. The calculated value of 11.529278 indicates the average of these two months' data.
* The 4-point trailing moving average is calculated by taking the average of the current month's value and the values from the three preceding months. The resulting value of 14.451395 represents the average of these four months' data.
* Similarly, the 6-point trailing moving average is obtained by averaging the current month's value with the values from the five preceding months. The given value of 14.566339 indicates the average of these six months' data.
* The 9-point trailing moving average is computed by taking the average of the current month's value and the values from the eight preceding months. The provided value of 14.727631 represents the average of these nine months' data.

In list we have RMSE -

2pointTrailingMovingAverage 11.529278

4pointTrailingMovingAverage 14.451395

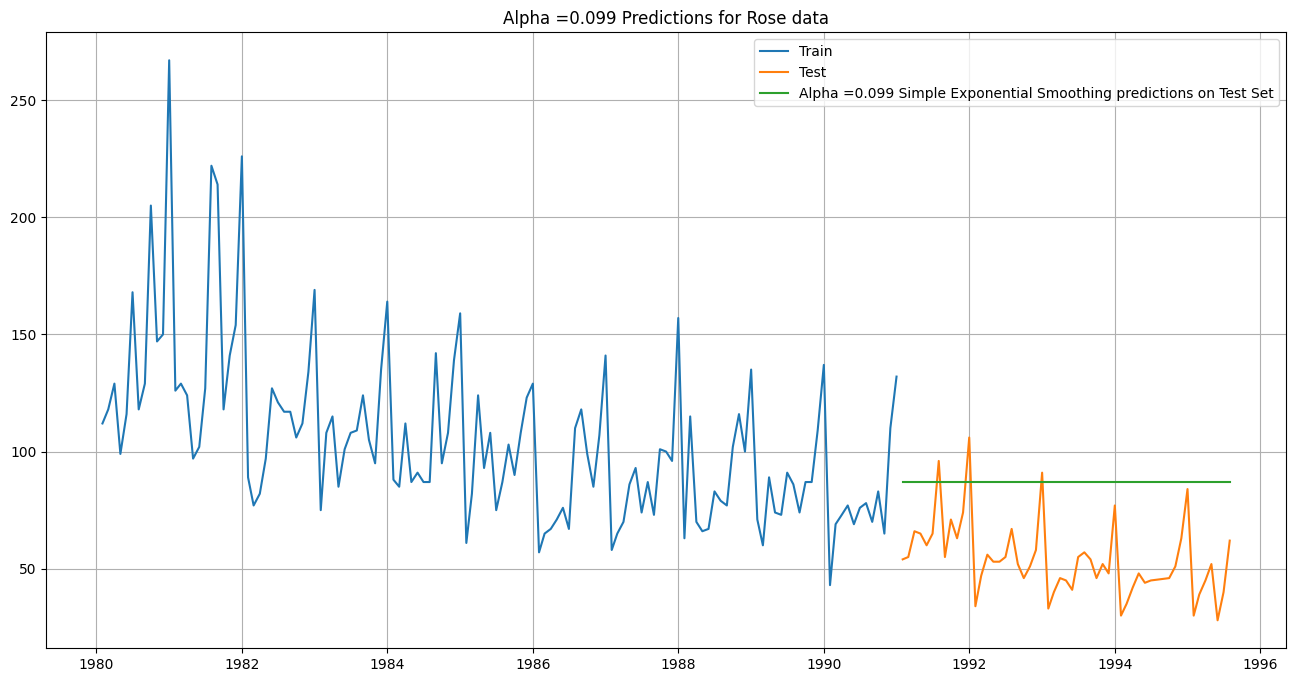
6pointTrailingMovingAverage 14.566339

9pointTrailingMovingAverage 14.727631

Lowest is for 2 point moving average.

**Simple Exponential smoothing -**

**At alpha = 0.099**

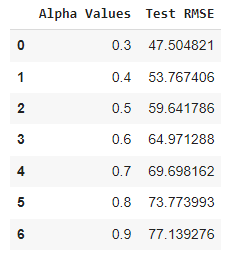
****

We have RMSE as -

For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data of Rose wine, RMSE is 36.80

We tried for for other alpha values-

We get RMSE as -



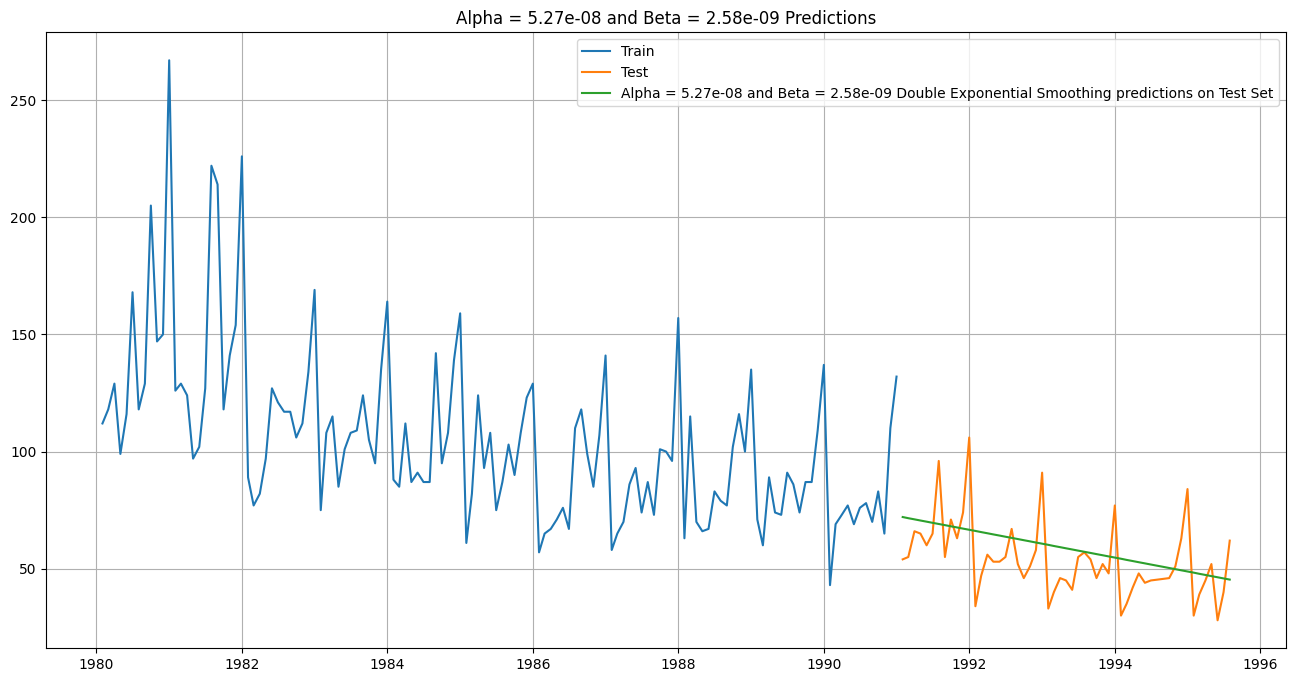
For alpha = 0.3 we have the lowest RMSE.

By setting alpha to 0.3, we are giving more weight to recent observations while forecasting, indicating that recent data points have a stronger influence on the forecasted values.

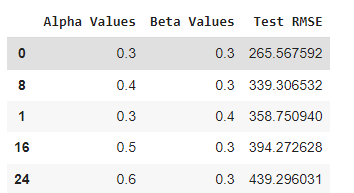
**Double Exponential Smoothing -**

For Alpha = 5.27e-08 and Beta = 2.58e-09

Double Exponential Smoothing Model forecast on the Test Data, RMSE is 15.27



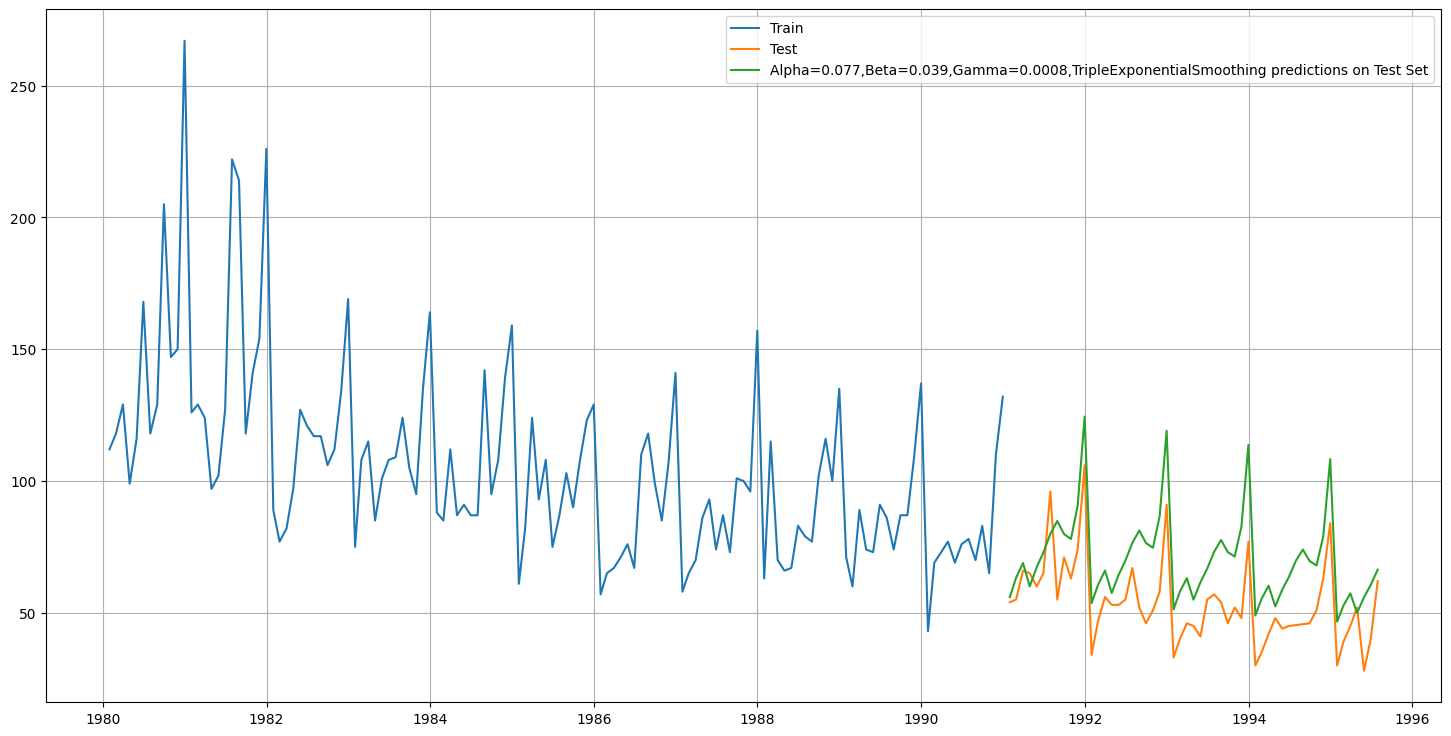
We check for other values of alpha and beta too



We get the Highest RMSE for Alpha 0.3 and Beta 0.3 for double exponential smoothing.

By setting alpha and beta to 0.3, we are giving equal importance to both recent observations and recent trends while forecasting.

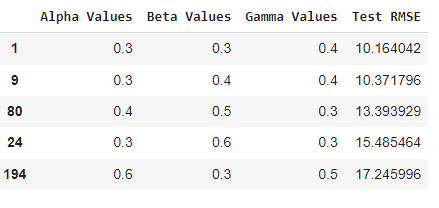
**Triple Exponential smoothing-**

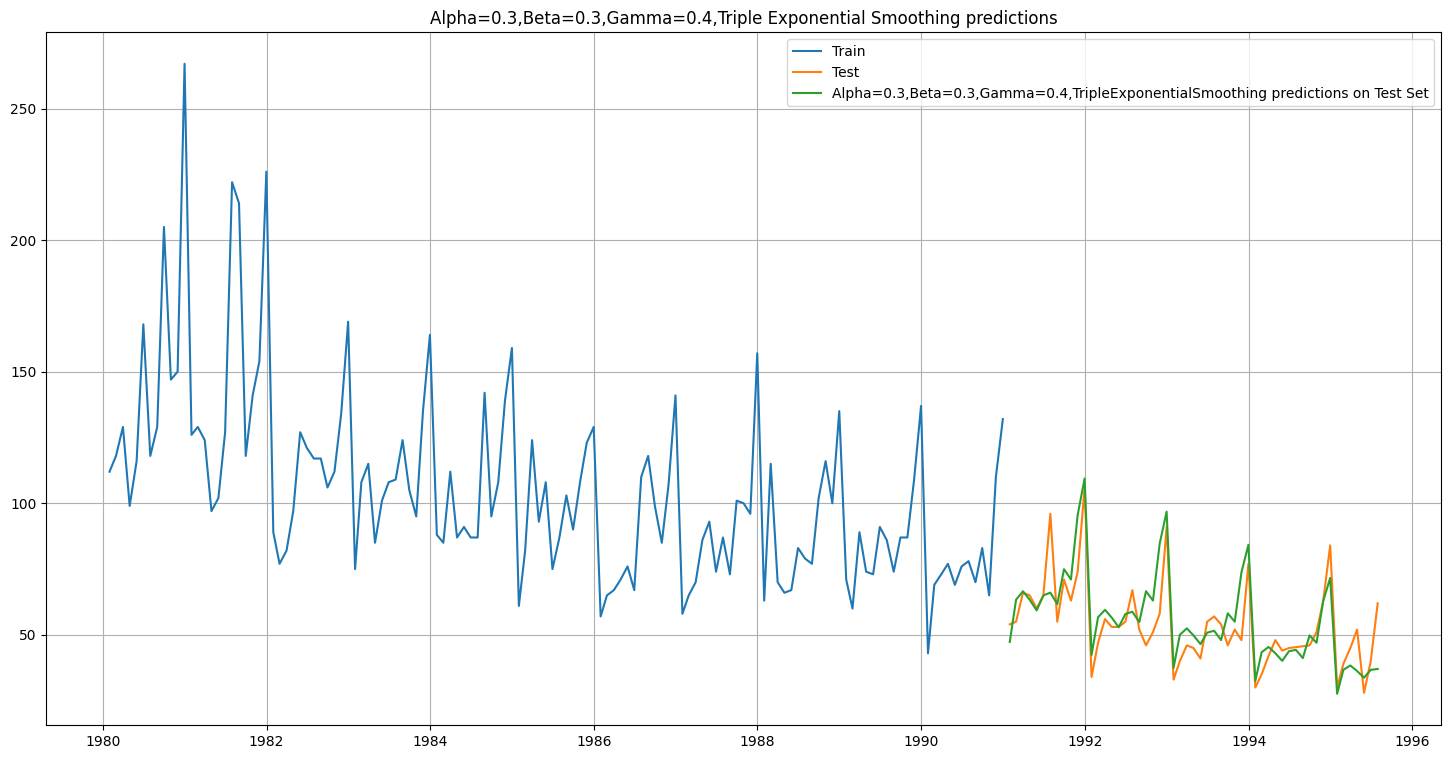


For **Alpha=0.077,Beta=0.039,Gamma=0.0008**, Triple Exponential Smoothing Model forecast on the Test Data**, RMSE is 19.11**

We are getting quite low RMSE , the pattern is finally following the trend in test data.

Let’s improve this more with trying more on hyperparameters-



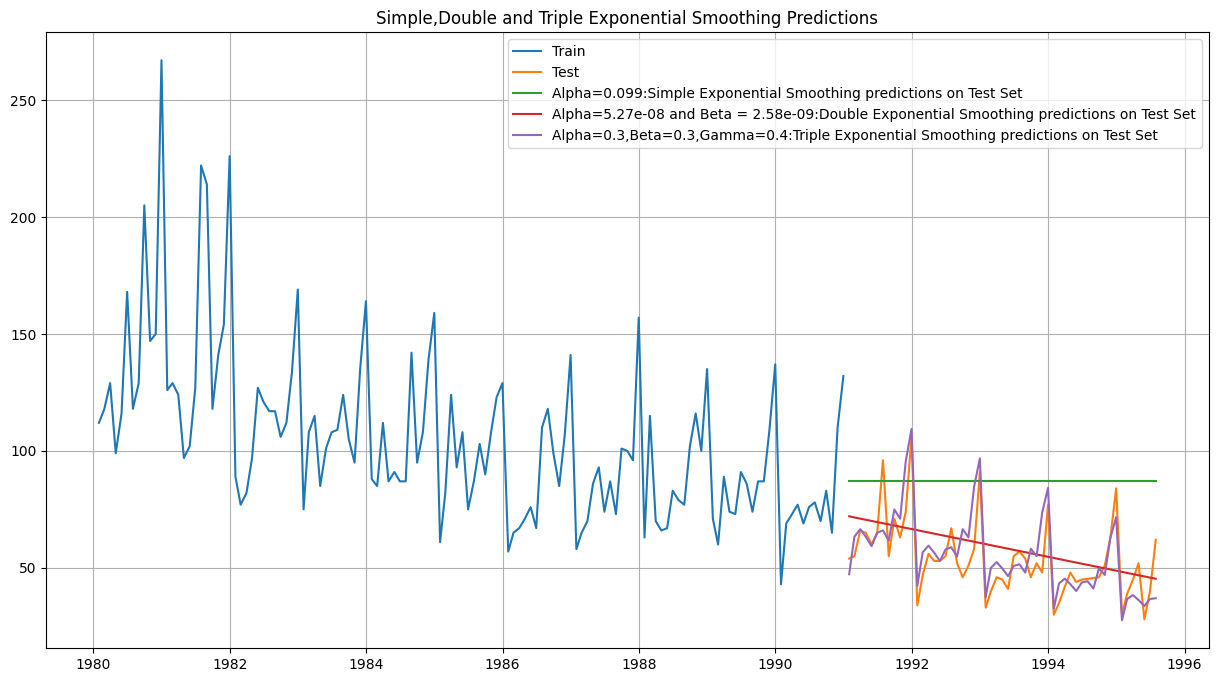


**For Alpha=0.3,Beta=0.3,Gamma=0.4 we are getting RMSE as 10.16**

This is the best RMSE score we get till now in all the models-

* Setting Alpha=0.3, Beta=0.3, and Gamma=0.4 in triple exponential smoothing means that recent observations, trend changes, and seasonal patterns have a significant impact on the forecast.
* The model is designed to be responsive to recent changes in the data while considering the overall level, trend, and seasonality.
* This balanced approach aims to capture recent dynamics in the data for more accurate forecasting.

Overall image of all models-



**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.**

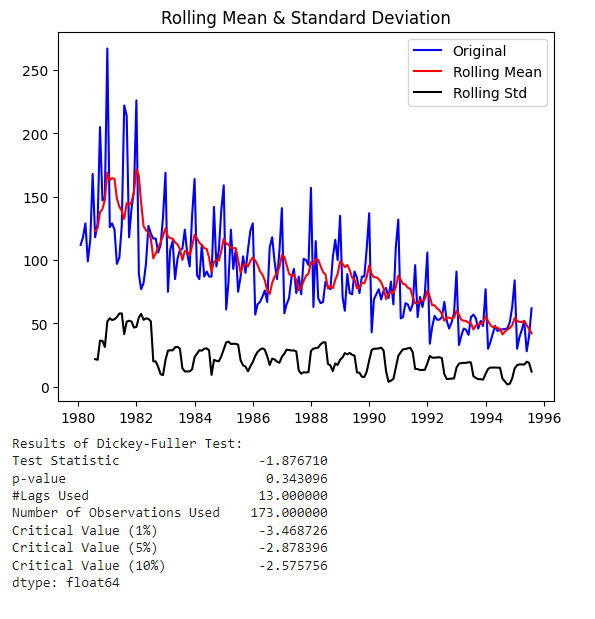
We will use DF Test-

Dickey-Fuller Test - Dicky Fuller Test on the timeseries is run to check for stationarity of data.

Null Hypothesis 𝐻0 : Time Series is non-stationary. Alternate Hypothesis 𝑯𝒂 : Time Series is stationary.

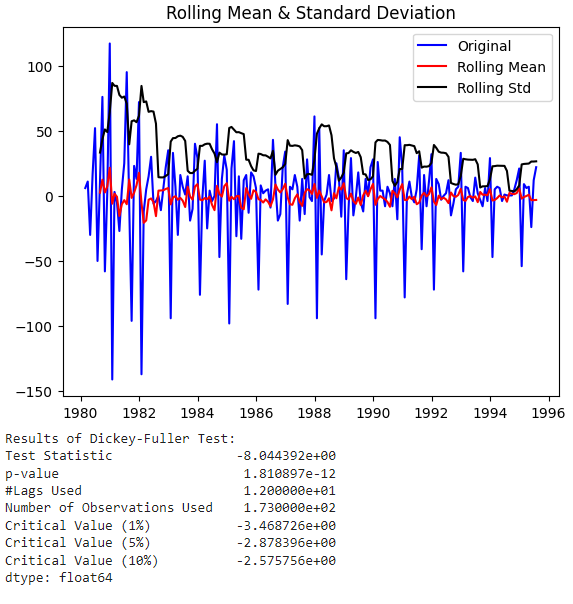
So if p-value < 0.05 then null hypothesis(Time series is non-stationary) is rejected else the Time series is non-stationary is failed to be rejected .

Let’s check-



Since the p-value is 0.343 at 5% critical value, which is greater than 0.05, the null hypothesis is not rejected. Hence the time series is non-stationary.

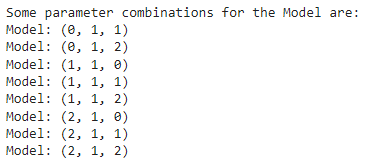
Let us take a difference of order 1 and check whether the Time Series is stationary or not.



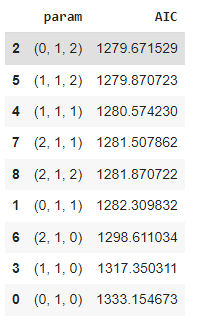
We see that at 𝛼 = 0.05 the Time Series is stationary.

**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.**

**ARIMA-**

****

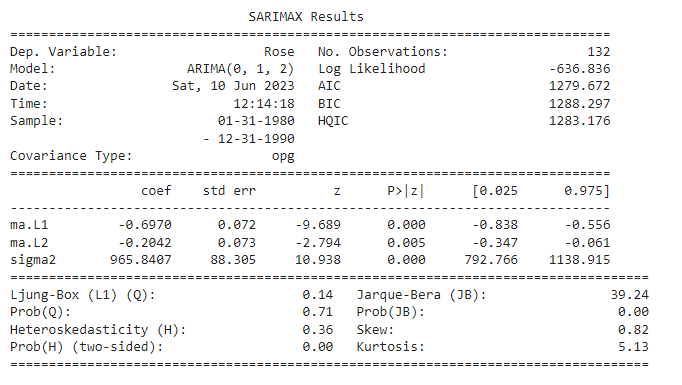
We have to check AIC for all the above combinations-



We get lowest AIC for (0,1,2)

The AIC value is used to compare different models and select the one that provides the best balance between goodness of fit and complexity. Lower AIC values indicate better-fitting models.

We will train model for these parameters -



The equation for the ARIMA(0, 1, 2) model is:

Δy(t) = -0.6970 \* ε(t-1) - 0.2042 \* ε(t-2) + ε(t)

where:

* Δy(t) represents the first-order difference of the dependent variable at time t (Rose).
* ε(t) represents the residual error at time t, which follows a normal distribution with mean zero and constant variance.
* ε(t-1) and ε(t-2) are the lagged residual errors at time t-1 and t-2, respectively.
* -0.6970 and -0.2042 are the estimated coefficients of the moving average (MA) terms.

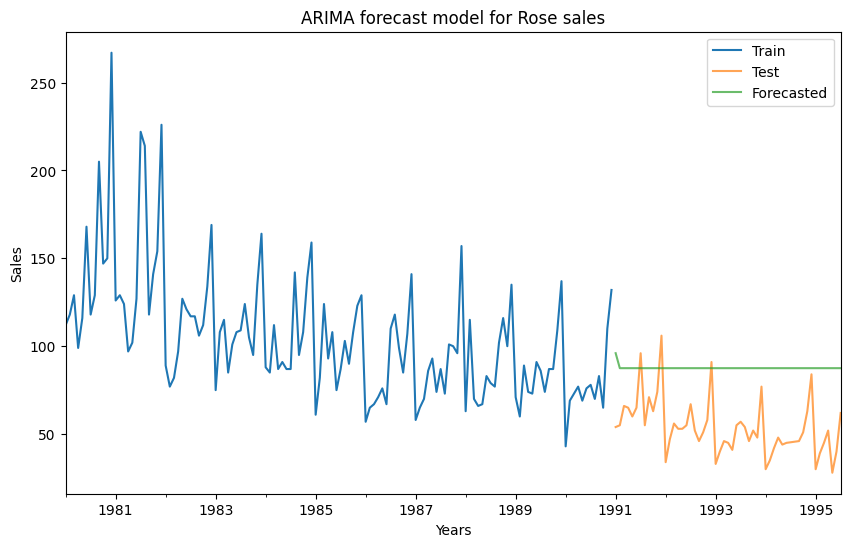
Based on the provided results, the coefficients of ma.L1 and ma.L2 are estimated for the model. These coefficients represent the influence of the lagged residual errors on the current value of the dependent variable.

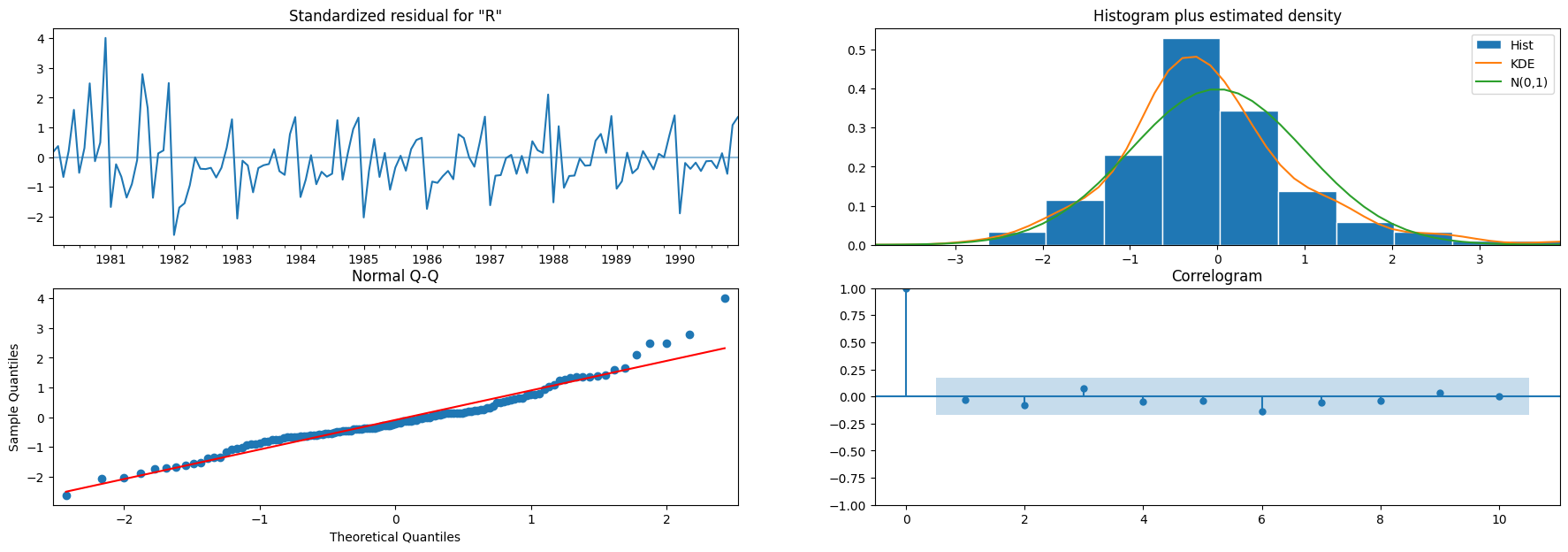
Both ma.L1 and ma.L2 have p-values less than 0.05, indicating that they are statistically significant. The magnitude of these coefficients (-0.6970 and -0.2042) provides insights into the strength of their impact on the dependent variable.

Additionally, the sigma2 value (965.8407) represents the estimated variance of the residual errors in the model.

Therefore, in terms of prediction, the ma.L1 and ma.L2 coefficients can be considered as the most important parameters in this ARIMA(0, 1, 2) model, as they have significant effects on the current value of the dependent variable.

Forecast looks like this-



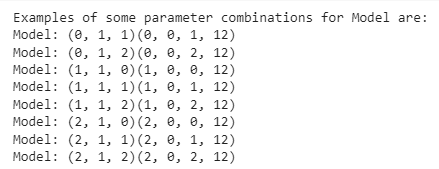


Histogram plus estimated density plot shows that the distribution of residuals is a bit deviated from normal , but can be considered as normal , the same is shown by the Q\_Q plot too. A bit of a deviation is these at one of the tails.

Correlogram shows that there are no remaining patterns or dependencies in the residuals.

**RSME** of ARIMA(2,1,2) is = **37.3064**

**SARIMA-**

****

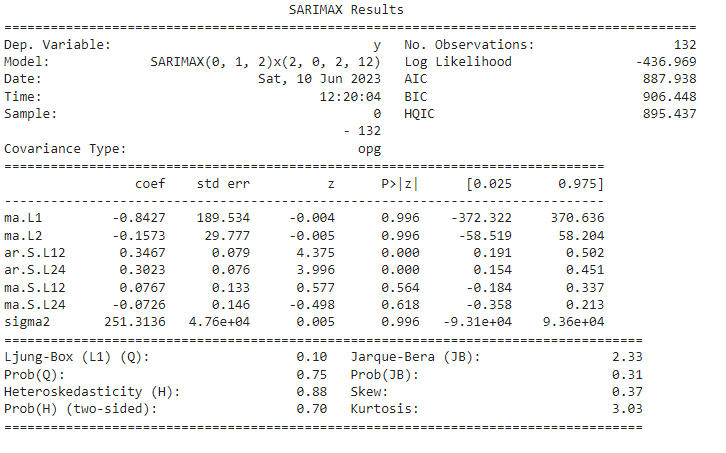
We have to check AIC for all the above combinations-



We get lowest AIC for (0,1,2)(2,0,2,12))

The AIC value is used to compare different models and select the one that provides the best balance between goodness of fit and complexity. Lower AIC values indicate better-fitting models.

We will train model for these parameters -



For the MA terms:

ma.L1 coefficient (-0.8427): This represents the impact of the first lag of the residual error on the current value of the dependent variable. However, the coefficient has a high standard error (189.534) and a very high p-value (0.996), indicating that it is not statistically significant and may not have a meaningful impact on the model.

ma.L2 coefficient (-0.1573): This represents the impact of the second lag of the residual error on the current value of the dependent variable. Similar to ma.L1, the coefficient has a high standard error (29.777) and a very high p-value (0.996), indicating its lack of statistical significance.

For the AR and seasonal AR terms:

ar.S.L12 coefficient (0.3467): This represents the impact of the first seasonal lag (lag 12) of the dependent variable on the current value. The coefficient has a statistically significant impact, with a low p-value (0.000).

ar.S.L24 coefficient (0.3023): This represents the impact of the second seasonal lag (lag 24) of the dependent variable on the current value. Similar to ar.S.L12, the coefficient has a statistically significant impact, with a low p-value (0.000).

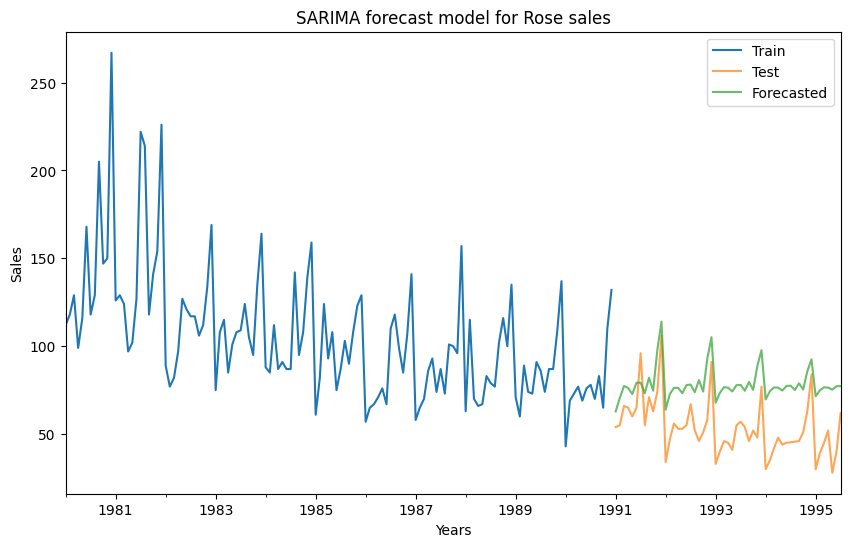
For the seasonal MA terms:

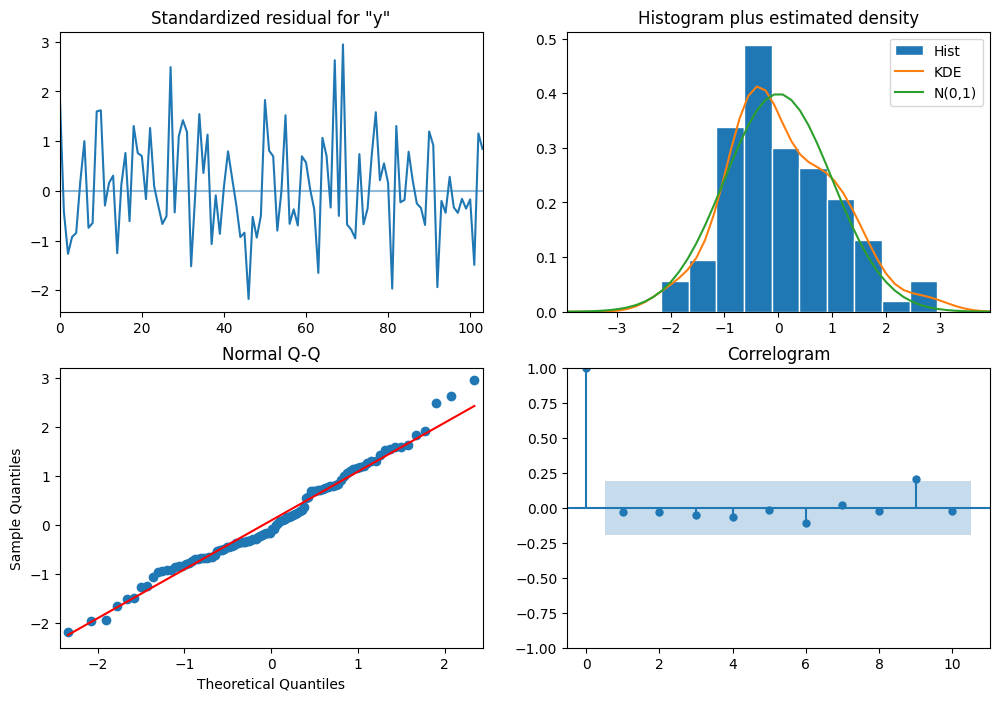
ma.S.L12 and ma.S.L24 coefficients: These represent the impacts of the first and second seasonal lags of the residual error, respectively, on the current value of the dependent variable. However, both coefficients have high p-values (0.564 and 0.618), indicating that they are not statistically significant.

The sigma2 value (251.3136) represents the estimated variance of the residual errors in the model.

In summary, for this SARIMAX model, the most significant coefficients are ar.S.L12 and ar.S.L24, which indicate the seasonal effects on the dependent variable. The MA terms (ma.L1 and ma.L2) and seasonal MA terms (ma.S.L12 and ma.S.L24) are not statistically significant and may not have a meaningful impact on the model.

Forecast looks like this-





Histogram plus estimated density plot shows that the distribution of residuals is a bit deviated from normal , but can be considered as normal , the same is shown by the Q\_Q plot too. A bit of a deviation is these at one of the tails.

Correlogram shows that there are no remaining patterns or dependencies in the residuals.

**RMSE for SARIMA (1,1,2)(1,0,2,12)= 26.928367334275265**

This is better than the Arima model.

**7. 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.**

****

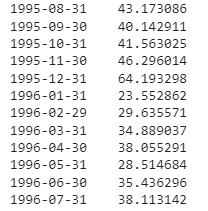
**8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.**

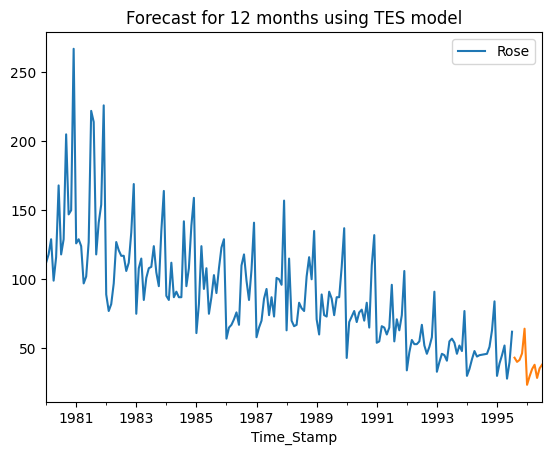
We will take 3 models-

1. **TES - alpha = 0.3, beta = 0.3, gamma = 0.4 forecast**

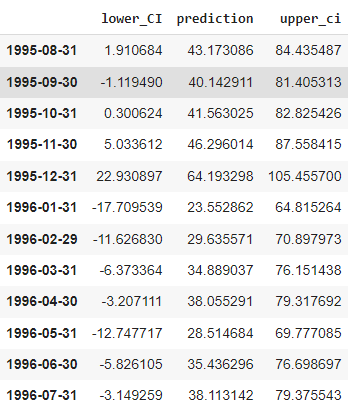
RMSE on full data is = 21.01323616895947

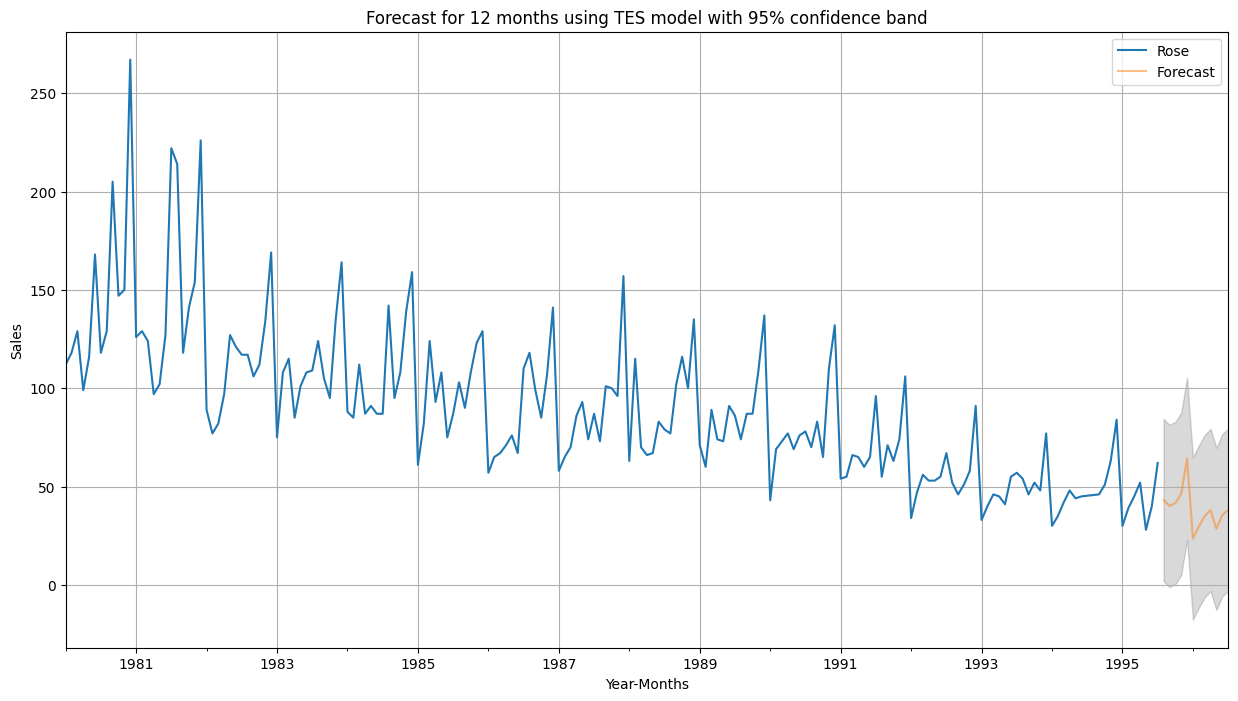
12 months into the future-





With 95% confidence interval few top records are displayed-



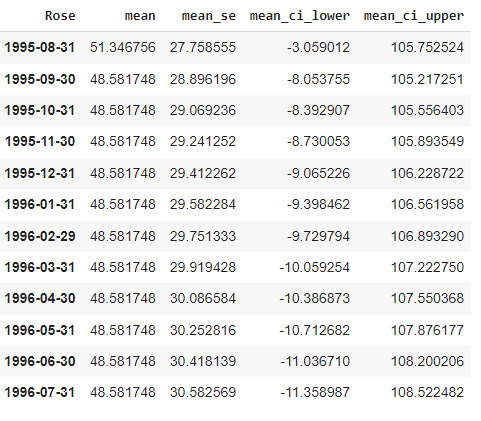


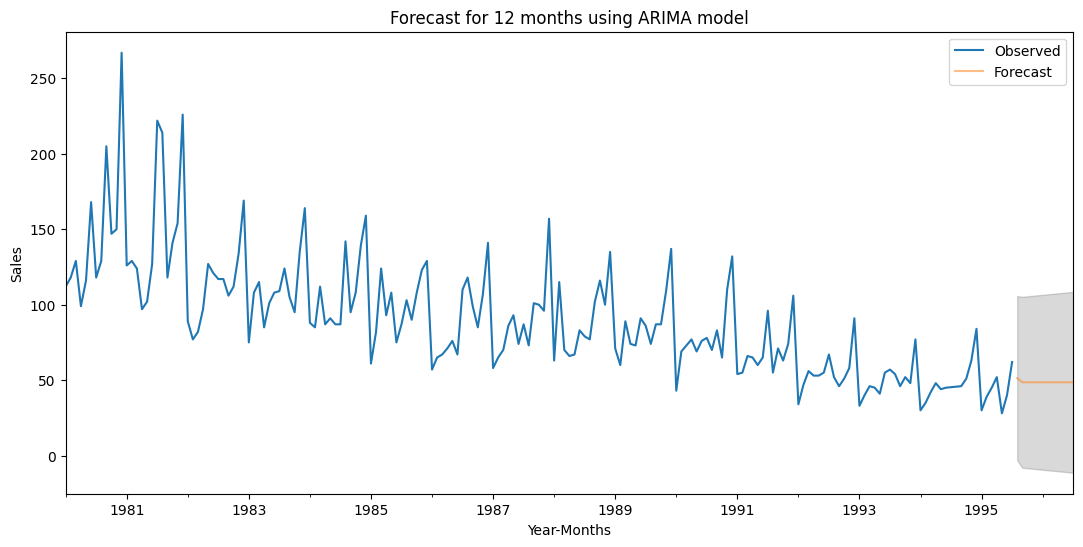
1. **ARIMA (0,1,2) forecast**

RMSE on full data is = 28.9757008

12 months into the future-

With 95% confidence interval few top records are displayed

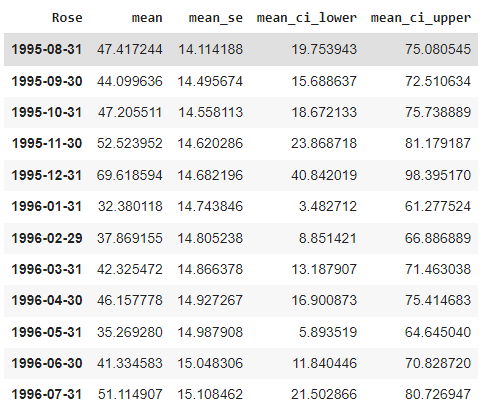


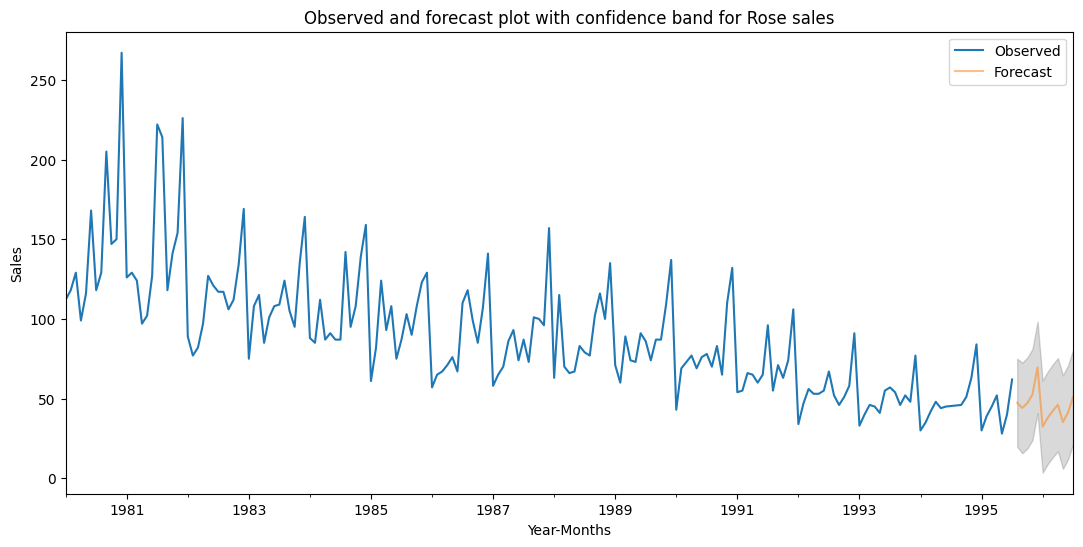


### SARIMA (0,1,2)(2,0,2,12) forecast

RMSE of the Full Model 539.9853944534857

12 months into the future-





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

In this report we have tried different methods to find out the patterns that sprinkling sales follow, we have seen in this report that some months have great going in terms of sales thus shows seasonal impact on sales.

From the models that are used for forecasting 12 months into the future the triple exponential smoothing seems to be doing the best in all of them. Their RMSE is also low on full data and was lowest when it was just applied on test data.

We should use below 2 models for the prediction of future sales.

1. **TES - alpha = 0.3, beta = 0.3, gamma = 0.4**
2. **2-point trailing moving average**

**Few measures for the company-**

**Seasonal Variation**

* The sales data shows a recurring pattern where sales tend to peak around December (64.19) and gradually decrease in the following months.
* This suggests a seasonal demand for the product, possibly related to holiday or festive seasons.
* The company can plan their production, marketing, and inventory management strategies accordingly, focusing more resources during peak seasons to maximize sales.

**Fluctuations and Trends**

* There are fluctuations in sales from month to month, indicating that there might be other factors influencing customer behavior.
* The company should investigate the underlying causes of these fluctuations to identify potential drivers and adapt their marketing or product strategies accordingly.
* Additionally, it appears that sales experienced a dip in January (23.55) and May (28.51) compared to surrounding months, which could be an opportunity for the company to implement targeted promotions or incentives to stimulate sales during these periods.