BUSINESS REPORT

Time series Data

Sparkling Wine

**Index-**

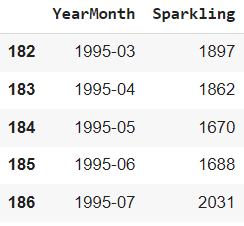
1. Read the data as an appropriate Time Series data and plot the data. [Page 2-5]
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. [Page 6-18]
3. Split the data into training and test. The test data should start in 1991. [Page 19-20]
4. Build all 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 20-29]
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 29-31]
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 31-39]
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 40]

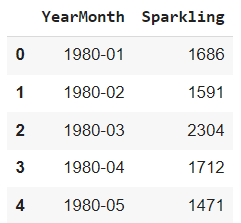
1. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands. [Page 40-48]
2. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. [Page 48-49]

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

Have a look at head and tail of data set.



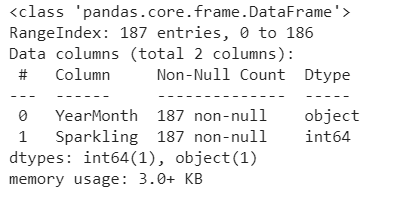
This is tail of data set this ends at 1995-07



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

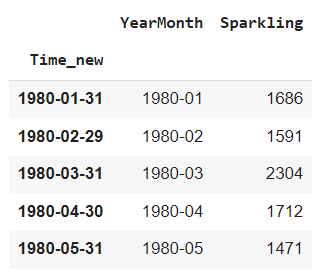
Let’s check the shape of the data.

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 'Sparkling', respectively.

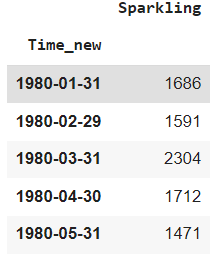


Data types of both the columns are object and int type for ‘YearMonth’ and ‘Sparkling’ respectively.

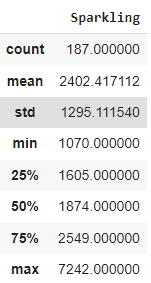
We have made a new date-time column and set it as an index to make this data time-series data.



After dropping YearMonth column we have final data set as -

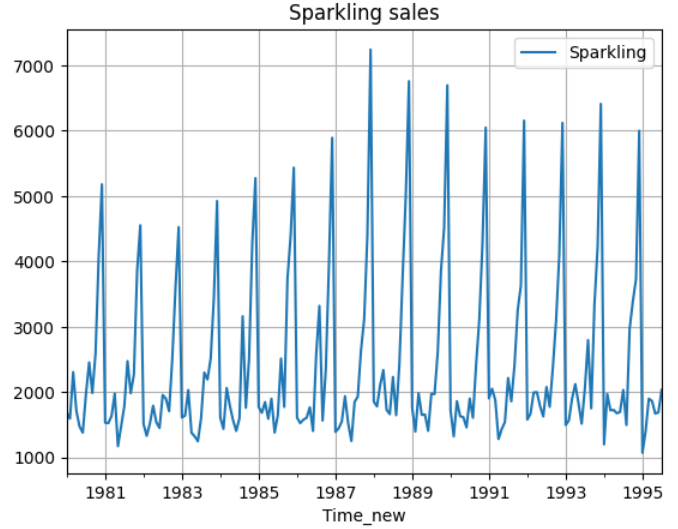


Let’s check the distribution of sales of sparkling wine-



* Mean: The average "Sparkling" sales is approximately 2402.42.
* Standard Deviation: The sales data has a standard deviation of approximately 1295.11, indicating a moderate amount of variability in the sales values.
* Minimum: The lowest recorded "Sparkling" sales during the given time period is 1070.
* 25th Percentile (Q1): 25% of the sales data points are below 1605, giving an indication of the lower range of sales.
* Median (50th Percentile or Q2): The middle value of the dataset is 1874, with 50% of the sales data points falling below this value.
* 75th Percentile (Q3): 75% of the sales data points are below 2549, giving an indication of the upper range of sales.
* Maximum: The highest recorded "Sparkling" sales during the given time period is 7242.

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

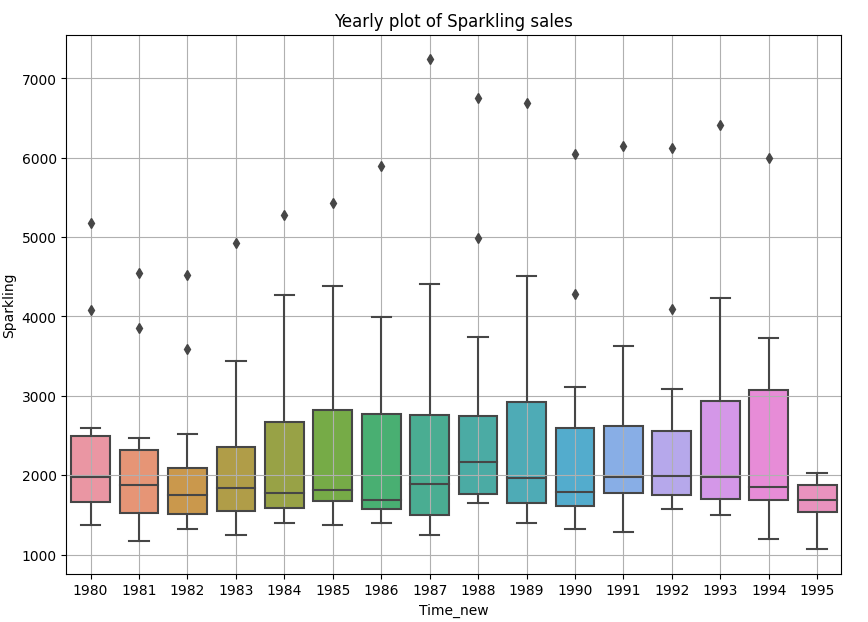


A pattern of increase and decrease in sales can easily be seen along the months, and passing by years.

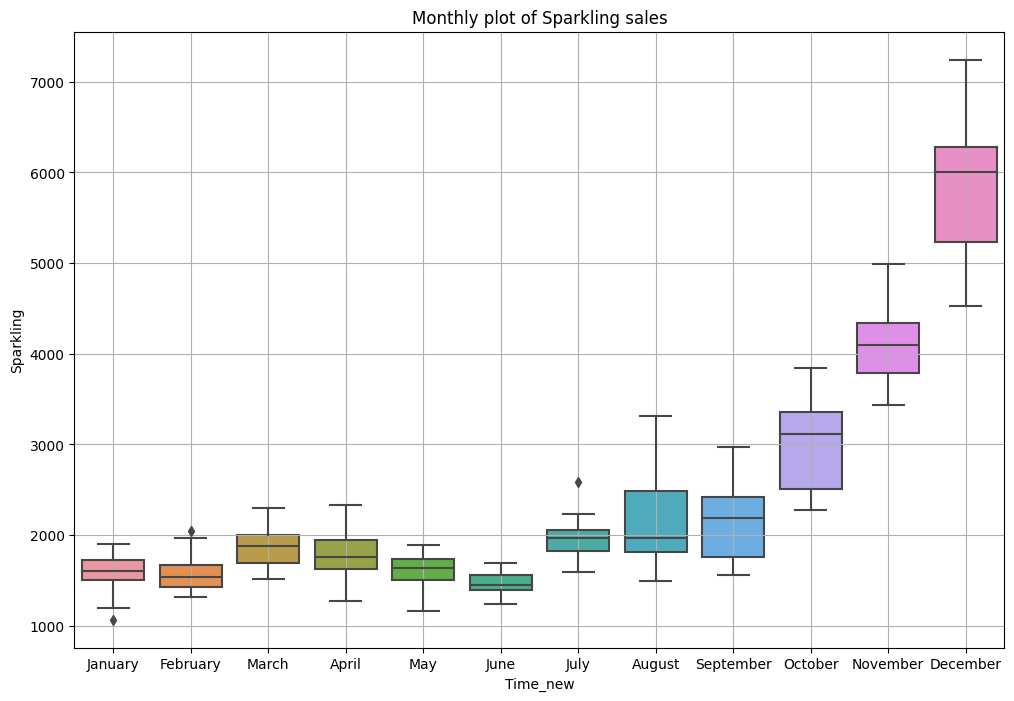
**2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

Let’s perform EDA-

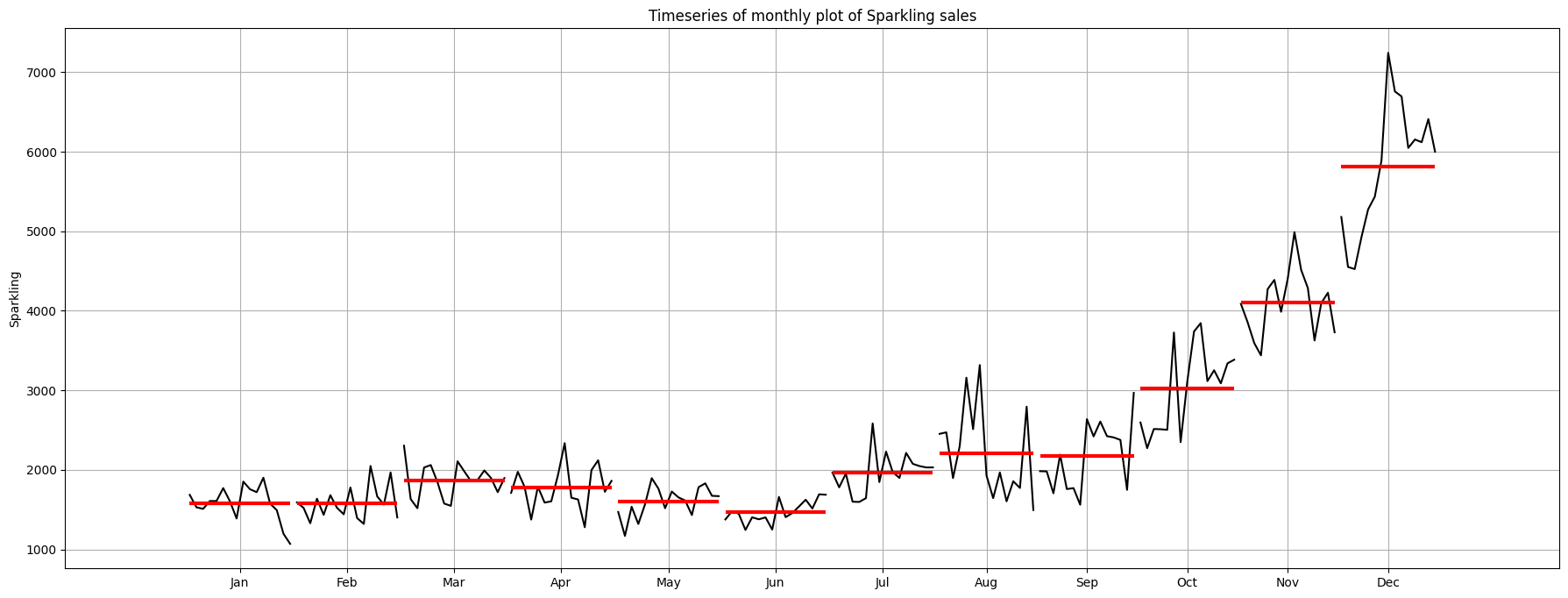
Box plot-



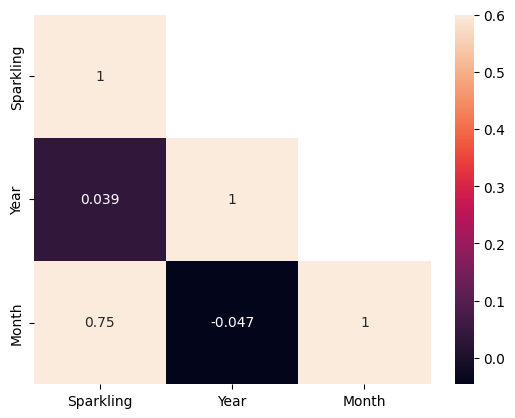
* Some outliers can be seen in data as this shows that some of the months have done exceptionally well and sparkling wine sales have gone up in those months.
* Mean of sales along the year has gone up and down starting from 1980 to 1995.
* A great variability can be seen in years starting from 1983 till 1989. This shows the variability in monthly sales along these years.



* From January to June, the sales data consistently remains below 2000 units for most of the years. However, there are a few notable observations. In January, one particular year stands out with significantly lower sales compared to the other years. In February, there is an outlier representing a year with better sales performance than the other February months.
* Another interesting anomaly occurs in July, where sales surge unexpectedly higher than expected. From July to September, there is a gradual increase in sales. However, starting from October, the sales show a remarkable growth trajectory, with December standing out as the best-performing month.
* Overall, the sales data exhibits a relatively consistent pattern from January to June, with a few exceptions and outliers. However, from October onwards, there is a notable upward trend in sales, culminating in a strong performance in December.

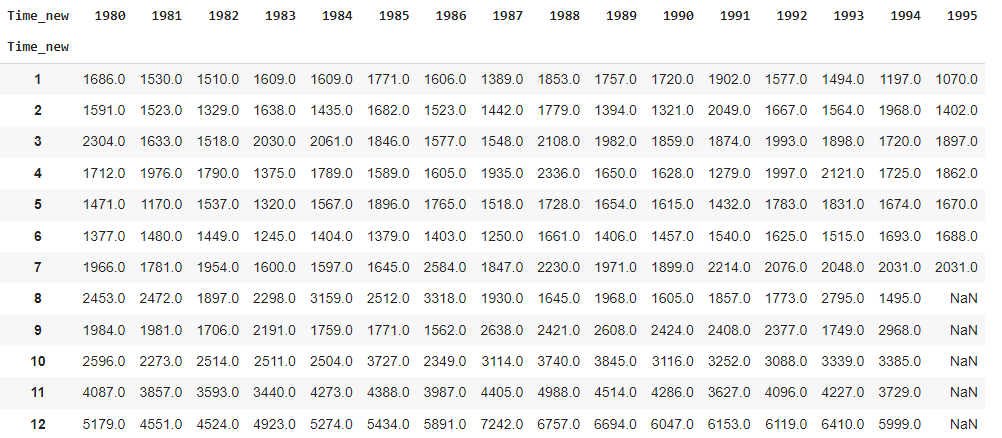


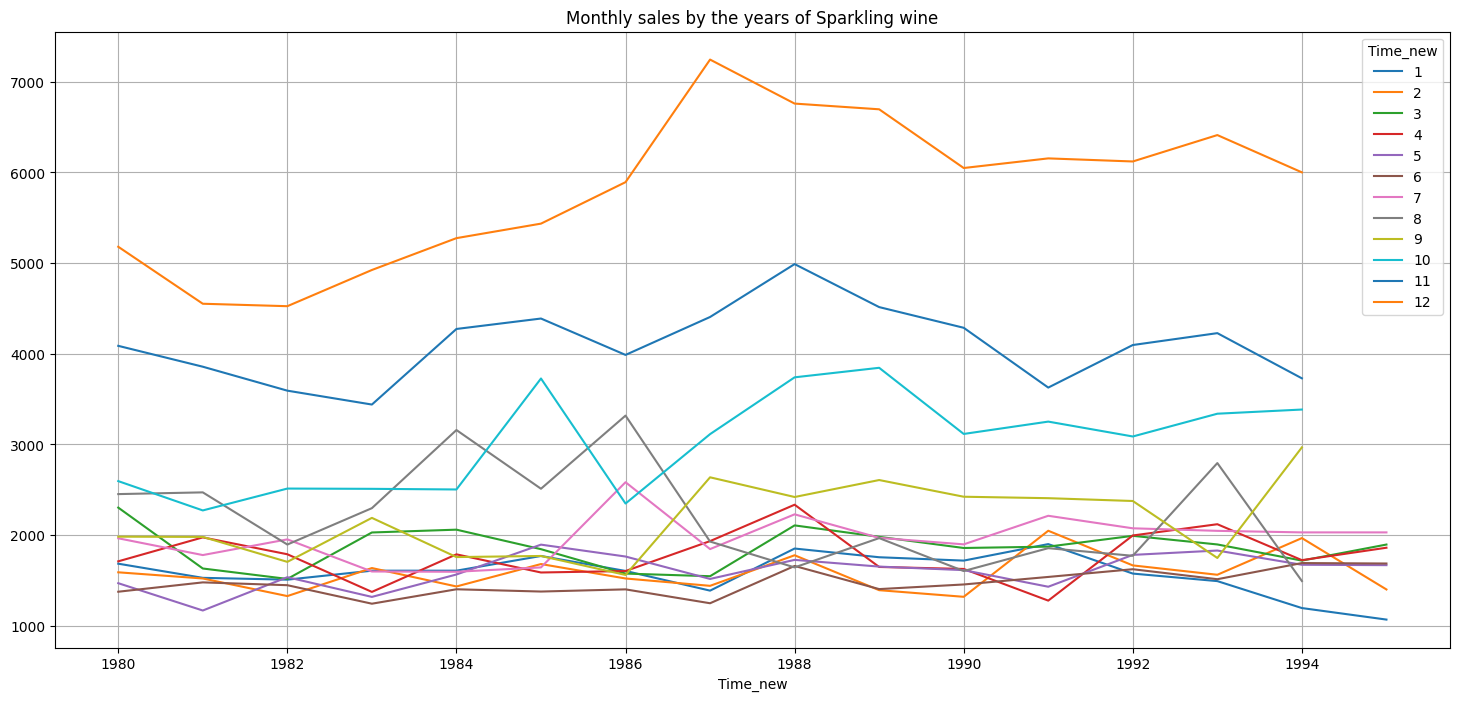
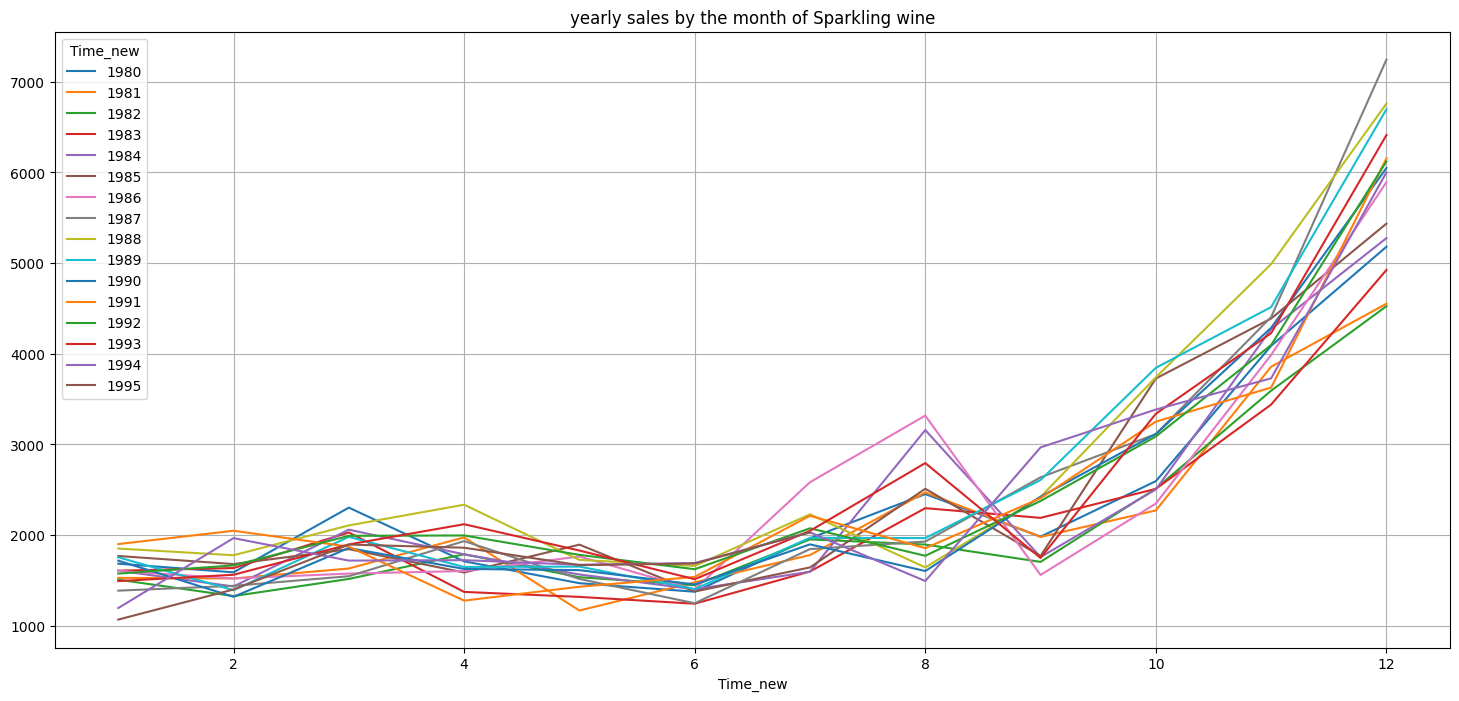
Heatmap-



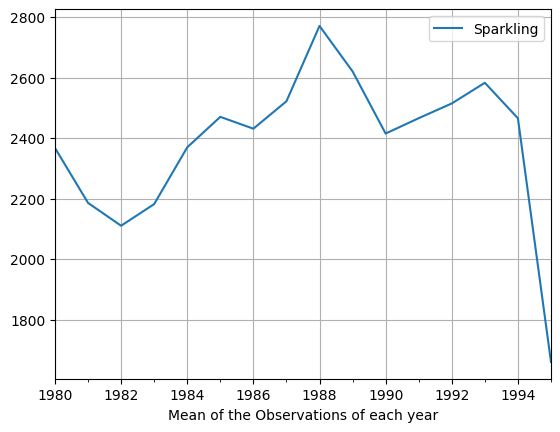
High correlation can be seen between month and sparkling wine sales.

**This is monthly sales by year.**



1. Overall Increasing Trend: There is a general increasing trend from 1982 to 1987 and there is dip in sales till 1990 and after that sales are somewhat consistent. 

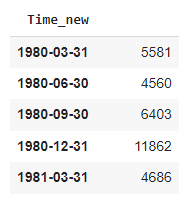
2. Seasonal Patterns: There seems to be a recurring seasonal pattern in sales. For example, there is a noticeable increase in sales towards the end of the year (months 9-12), suggesting a possible holiday season effect. This pattern is consistent across multiple years.



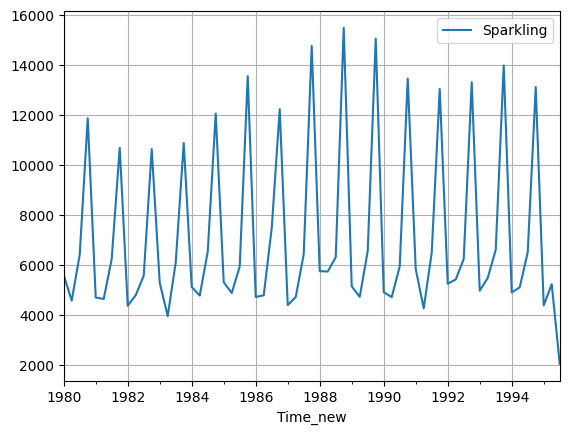
Mean of yearly sales, maximum sales is in year1988. Sales have increased from 1982 to 1988.

Let’s check data quarterly.

Sum of quarterly sales-

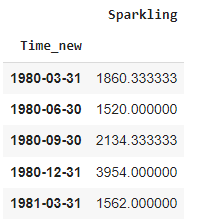


Quarter sales looks like-

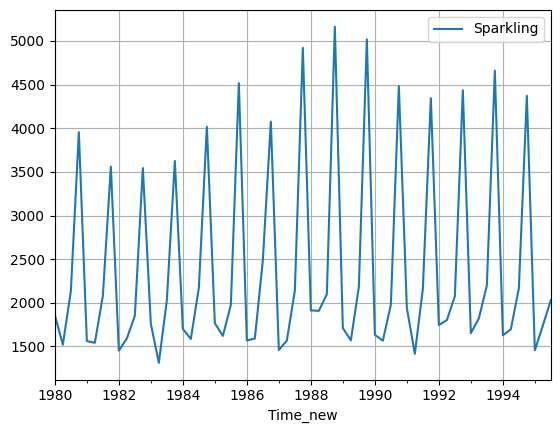


A pattern can be seen,with peak sales in a selective quarter.In most cases we have maximum sales in the last quarter of the year for all years.

Mean quarterly sales-

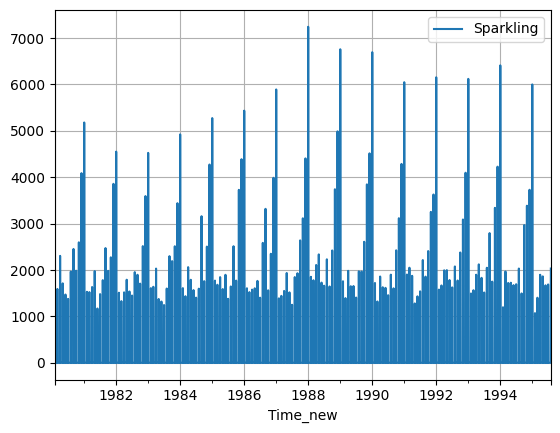


Shows similar pattern with maximum sales around last quarter along all years.



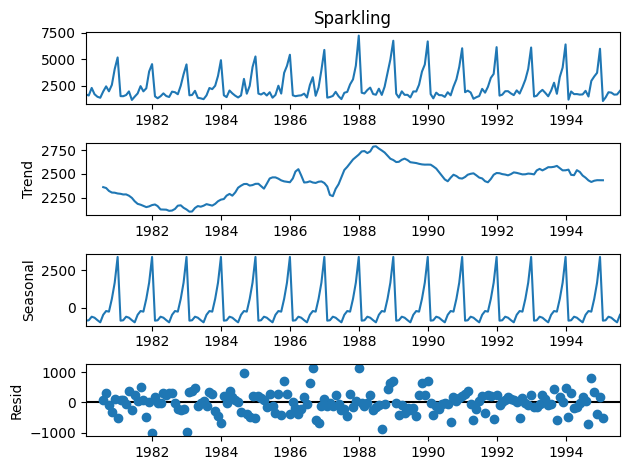
Let’s have a looks at daily sales too

It looks like this , with peaks on few days-



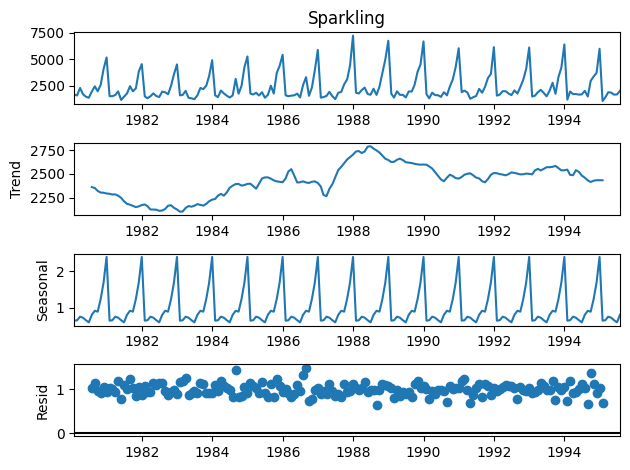
We are now going to decompose the Data-

Additive decomposition-

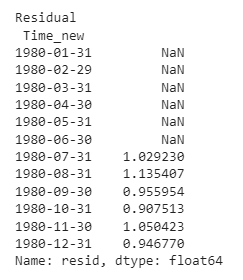
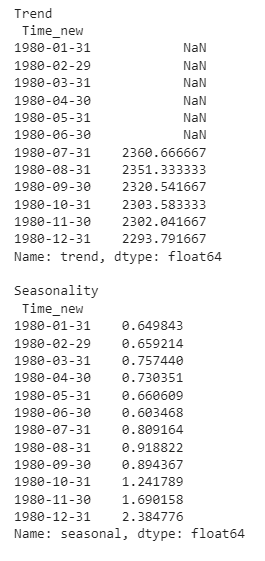


* We can see some trends in data. It increased from 1983 to 1986 , then fluctuated downwards a bit and started increasing from 1987 to 1989. Then the trend was stable till 1995.
* Seasonality can be seen in data easily. It has repetitive peaks in a few months along moving years.
* Residuals are along the zero line but are also deviating from zero till 1000.

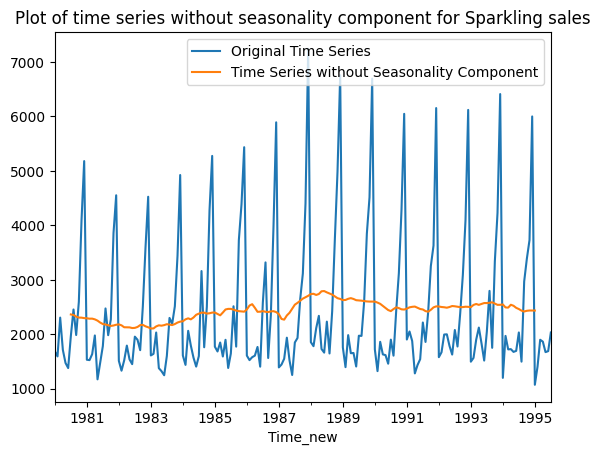
Multiplicative decomposition-



* 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.

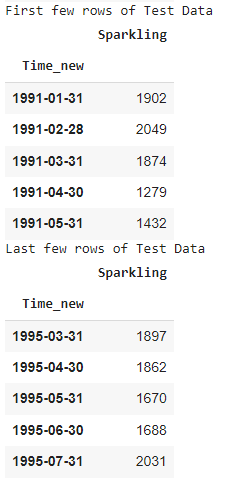
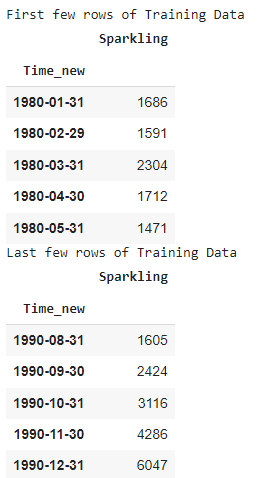


This is according to Multiplicative decomposition, residuals close to one can be seen.



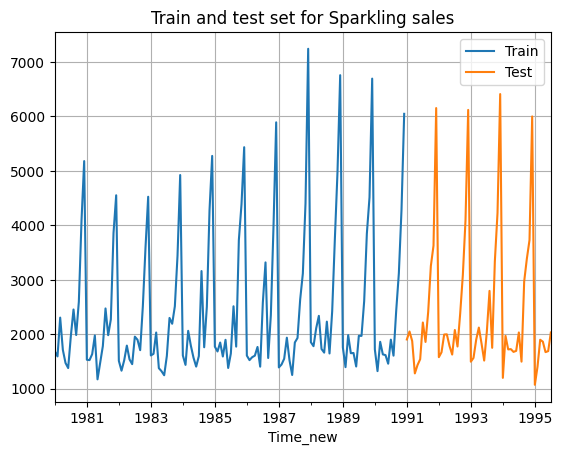
**3. Split the data into training and test. The test data should start in 1991.**

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)

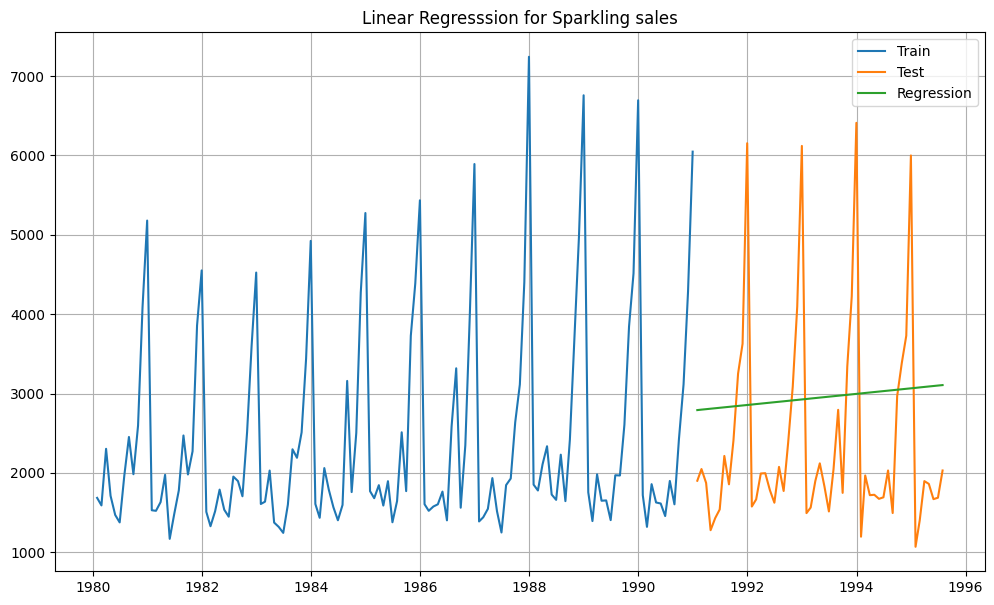
****

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.

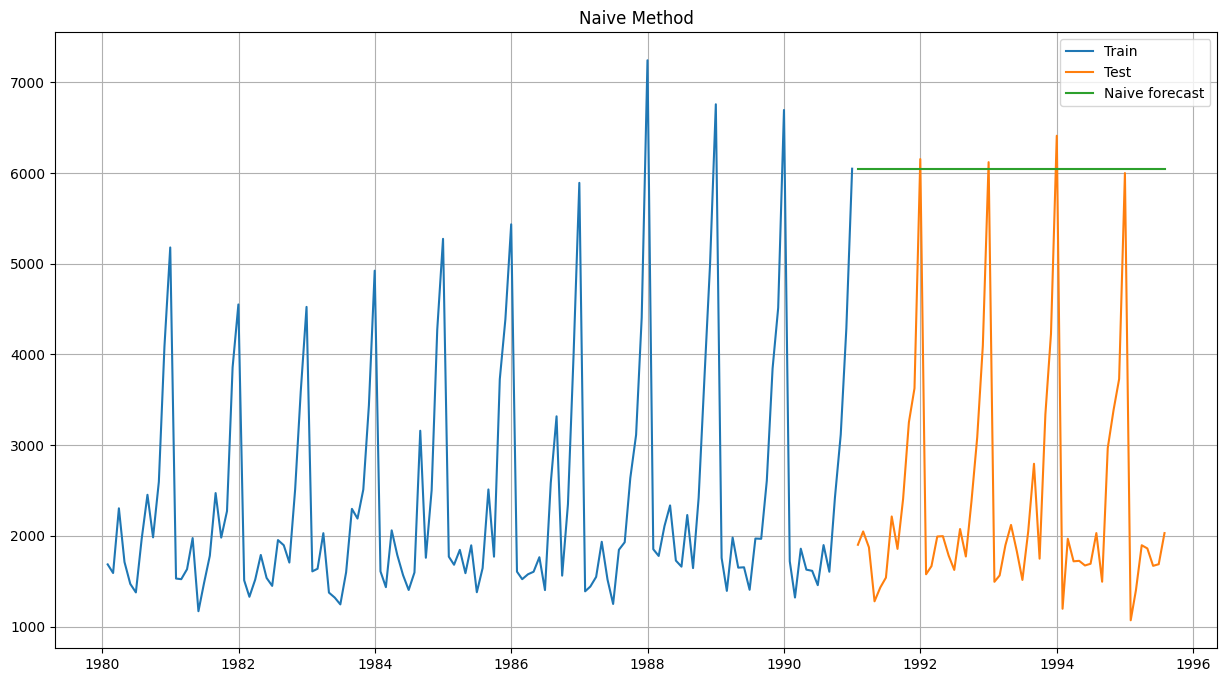
**Linear regression-**



For RegressionOnTime forecast on the Test Data,  **RMSE is 1389.14**

It is just a straight line trying to predict patterns.

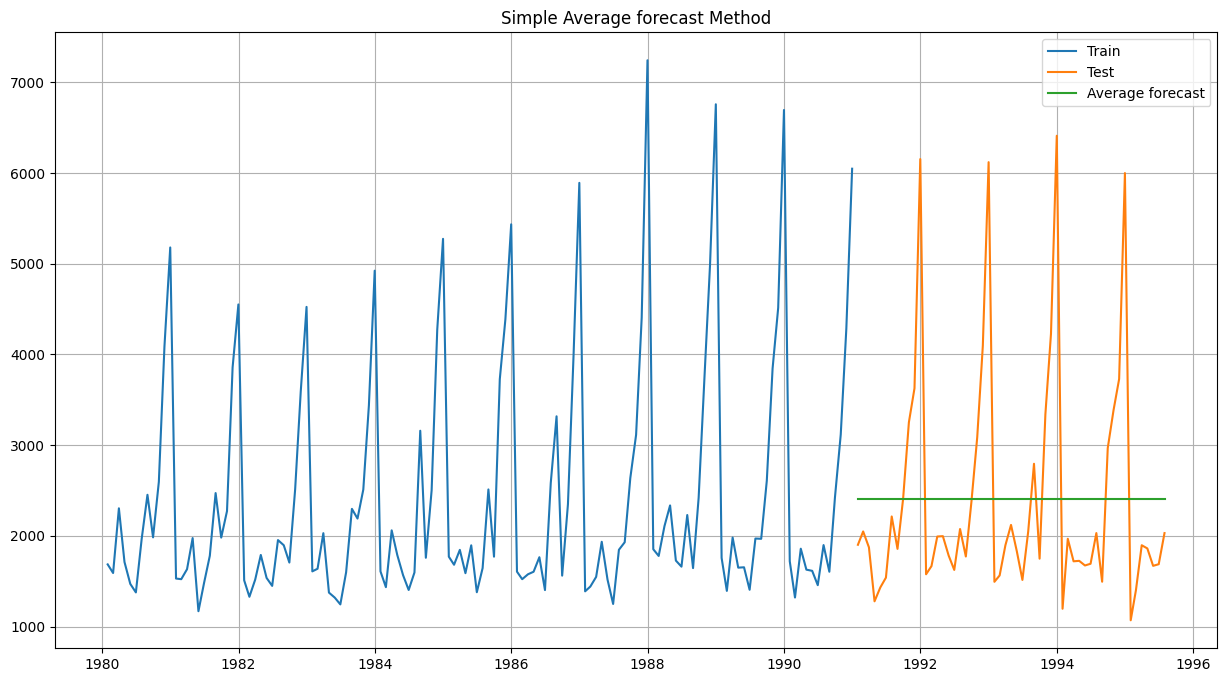
**Naive forecast-**

****

For Naive forecast on the Test Data,  **RMSE is 3864.28**

This is less than the previous model.

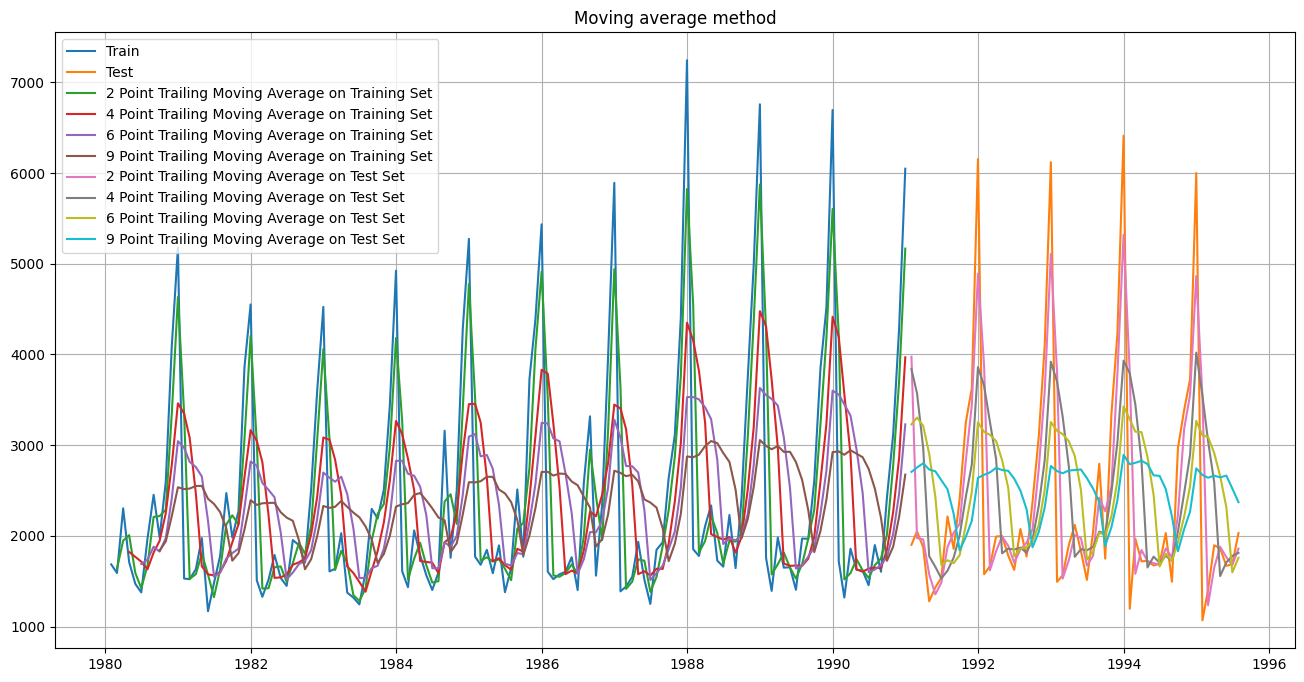
**Simple average forecast-**

****

Simple Average forecast on the Test Data**, RMSE is 1275.08**

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-

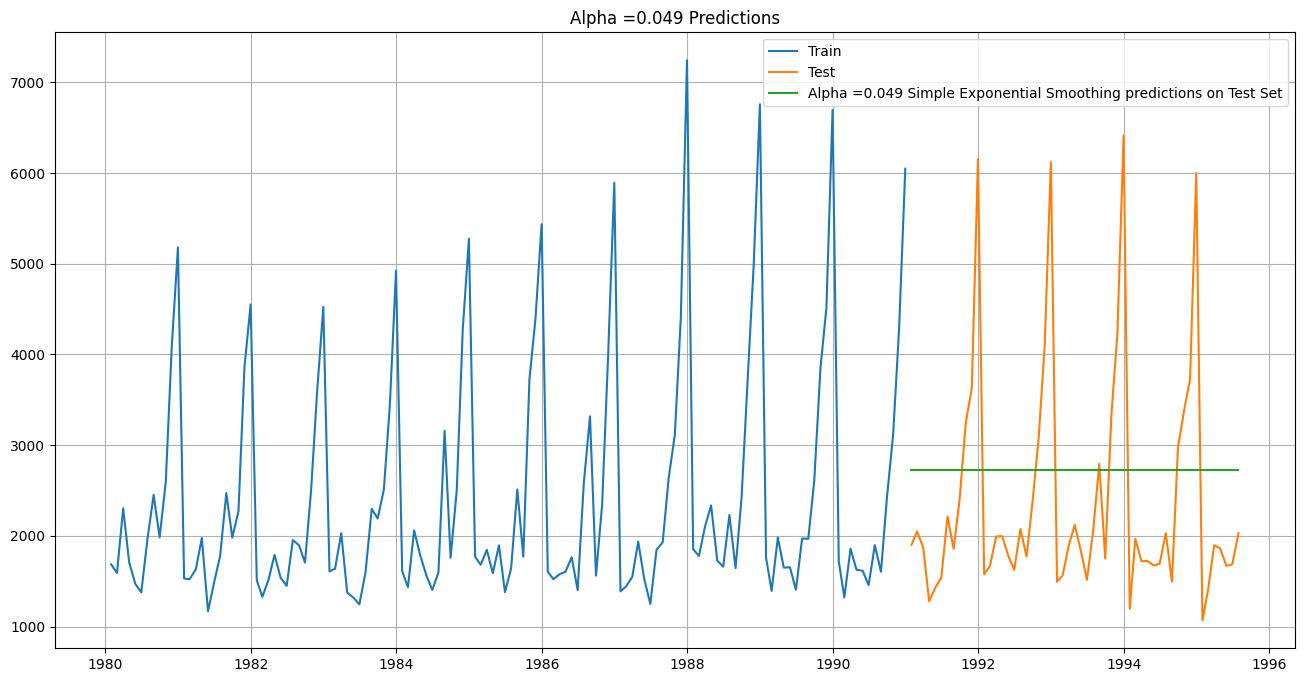
* The 2-point trailing moving average has an RMSE value of 813.400684. It calculates the average of every two consecutive monthly data points. A lower RMSE value indicates a better fit of the moving average to the monthly sales data.
* The 4-point trailing moving average has an RMSE value of 1156.589694. It calculates the average of every four consecutive monthly data points. The higher RMSE value suggests a less accurate fit of the moving average to the monthly sales data compared to the 2-point trailing moving average.
* The 6-point trailing moving average has an RMSE value of 1283.927428. It calculates the average of every six consecutive monthly data points. Again, a higher RMSE value compared to the previous techniques indicates a less accurate fit of the moving average to the monthly sales data.
* The 9-point trailing moving average has an RMSE value of 1346.278315. It calculates the average of every nine consecutive monthly data points. Similar to the previous techniques, a higher RMSE value suggests a less accurate fit of the moving average to the monthly sales data.

In list we have RMSE -



**Simple Exponential smoothing -**

**At alpha = 0.049**

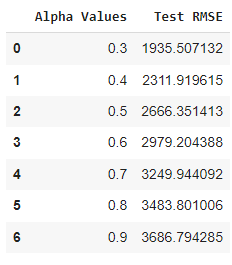
****

We have RMSE as -

For Alpha =0.049 Simple Exponential Smoothing Model forecast on the Test Data, **RMSE is 1316.03**

We tried for for other alpha values-

We get RMSE as -



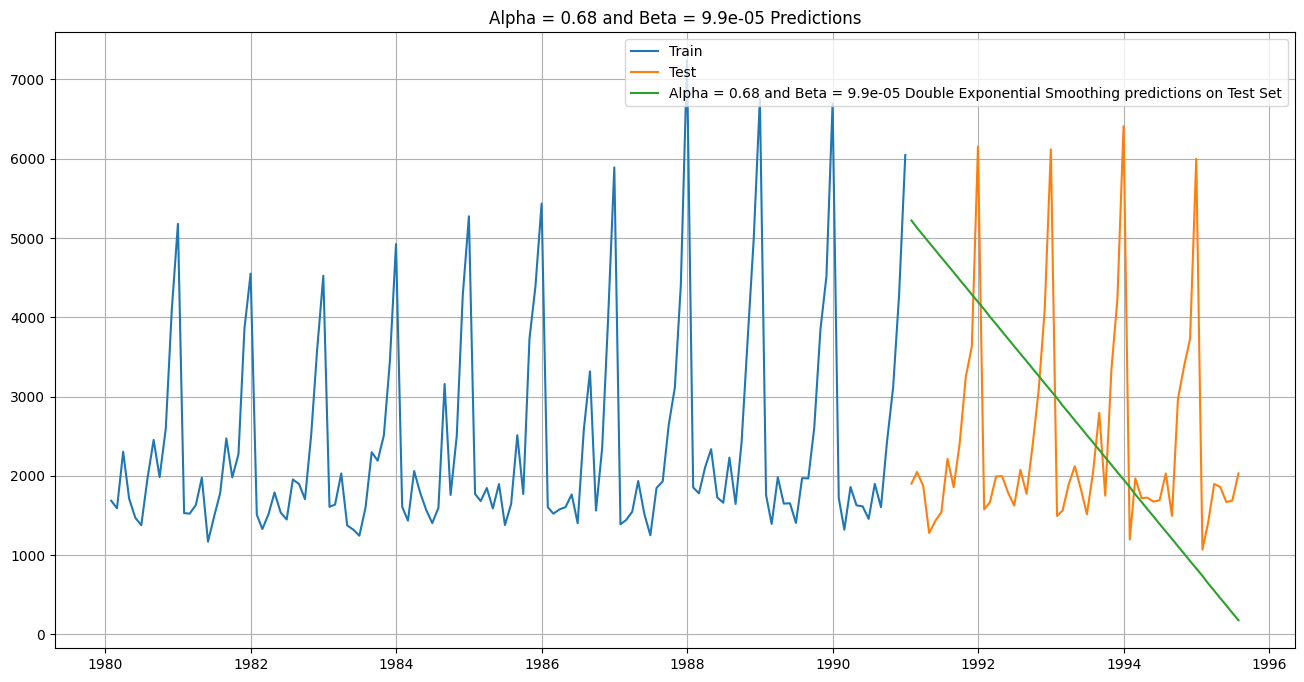
For alpha = 0.3 we have the lowest RMSE.

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.

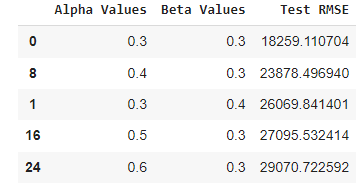
**Double Exponential Smoothing -**

For Alpha = 0.68 and Beta = 9.9e-05

Double Exponential Smoothing Model forecast on the Test Data, **RMSE is 2007.24**



We check for other values of alpha and beta too

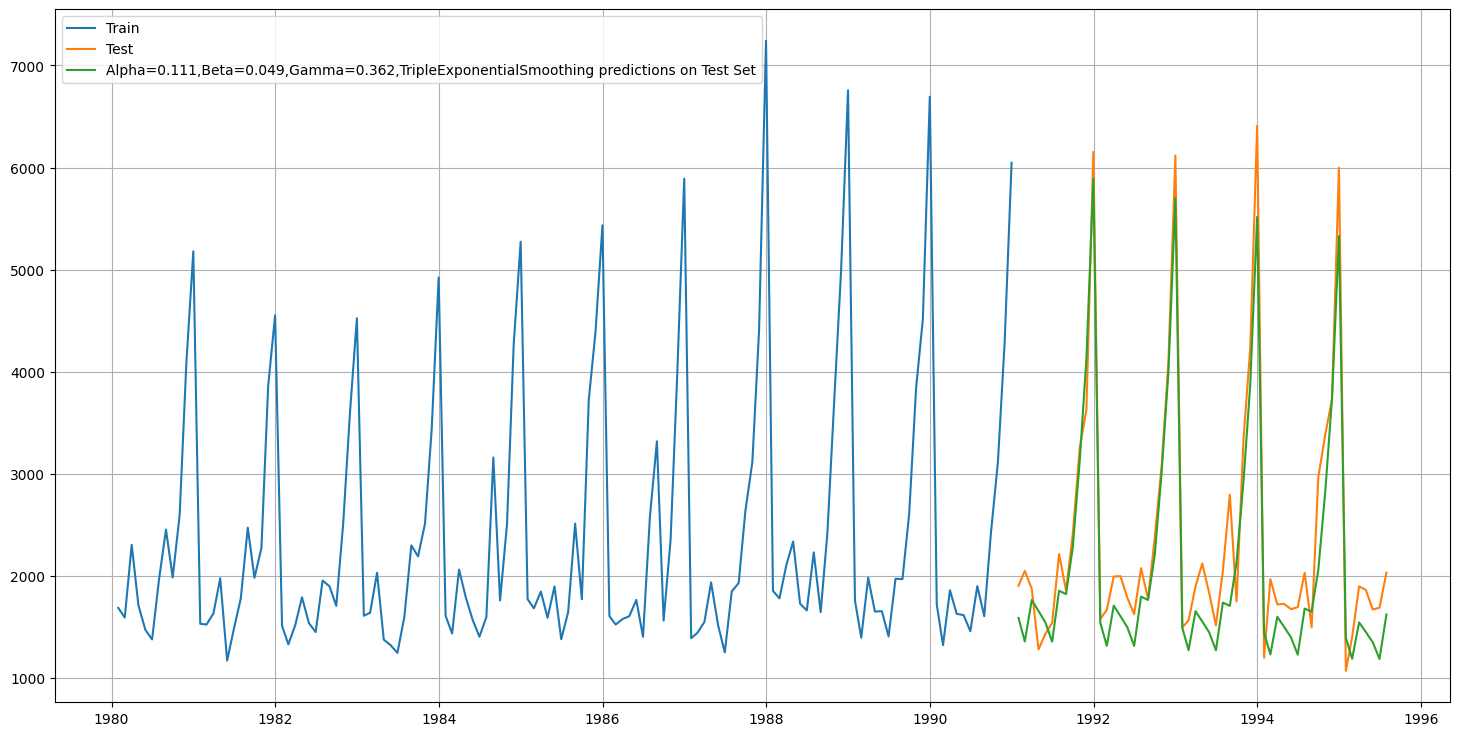


We get the lowest RMSE for Alpha 0.3 and beta 0.3 for double exponential smoothing.

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

It suggests that the combination of alpha = 0.3 and beta = 0.3 captures both the underlying level and trend in the data effectively, resulting in more accurate forecasts.

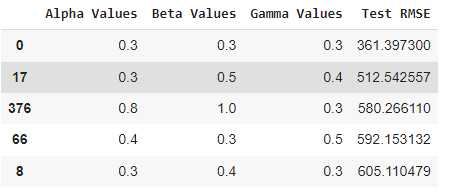
**Triple Exponential smoothing-**



For **Alpha=0.111,Beta=0.049,Gamma=0.362**, Triple Exponential Smoothing Model forecast on the Test Data, **RMSE is 403.71**

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-

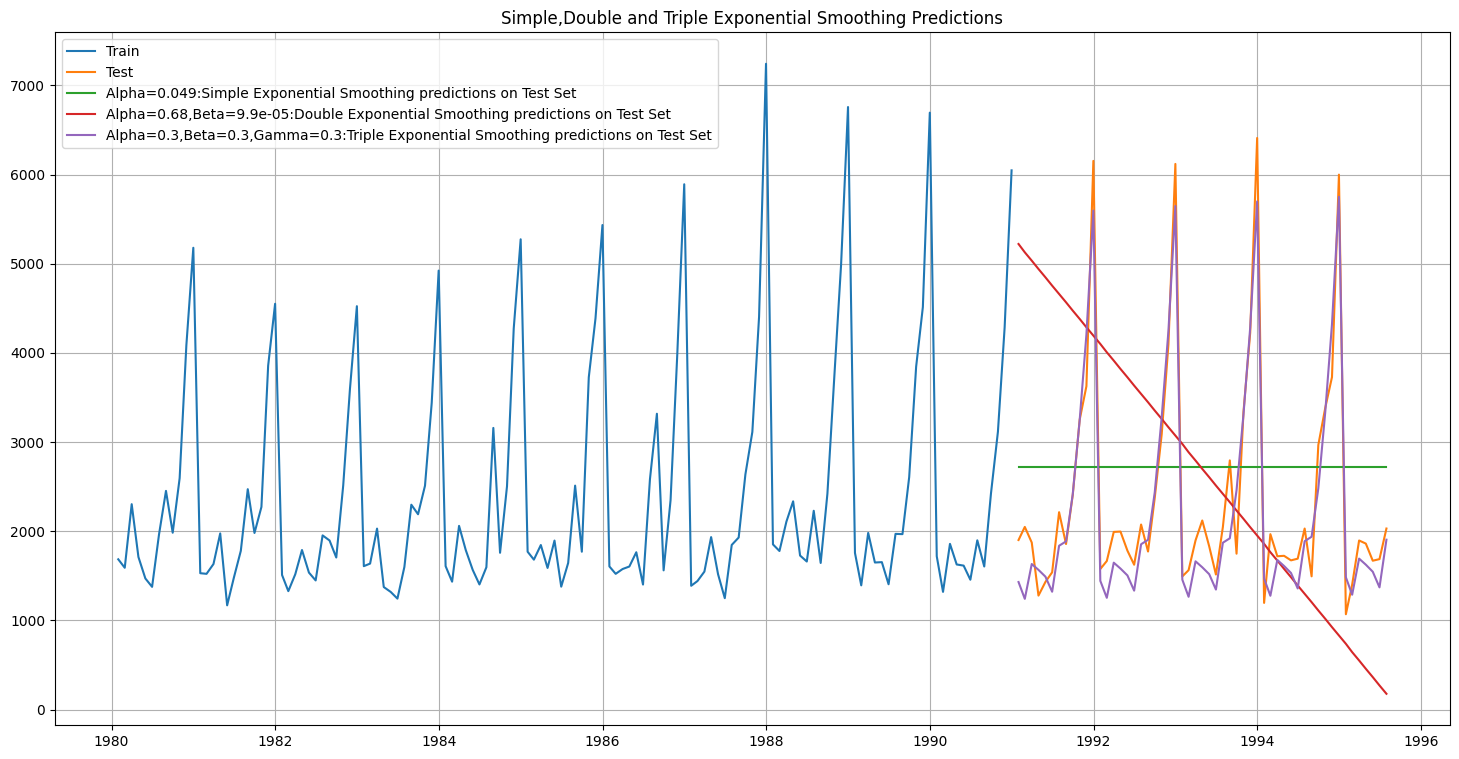


**For Alpha=0.3,Beta=0.3,Gamma=0.3 we are getting RMSE as 361.39**

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

* Setting Alpha=0.3, Beta=0.3, and Gamma=0.3 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.**

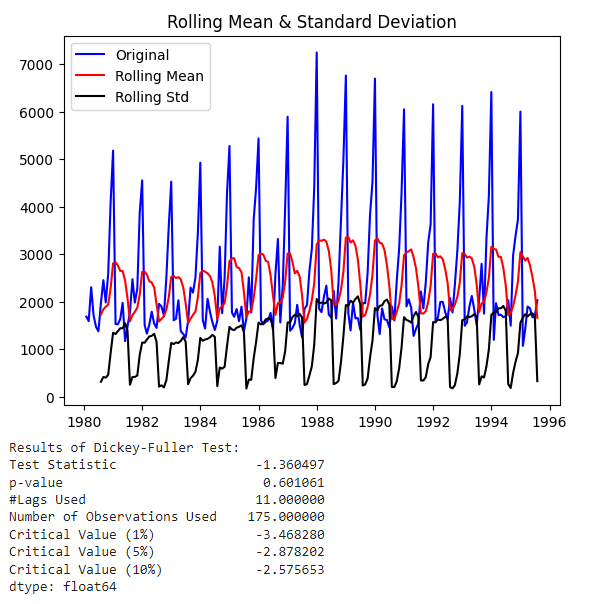
We will use DF Test-

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.

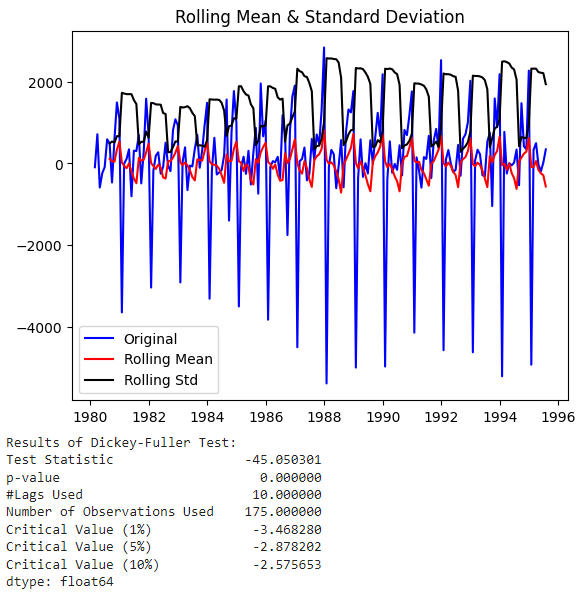
So if p-value < 0.05 then null hypothesis(Time series is non-stationary) is rejected else the Time series is non-stationary is failed to be rejected .

Let’s check-



Since the p-value is 0.601 at 5% critical value, which is greater than 0.05, the null hypothesis is not rejected. Hence the time series is non-stationary.

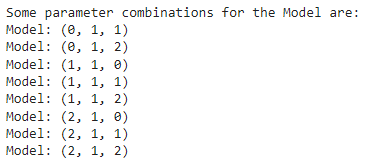
Let us take a difference of order 1 and check whether the Time Series is stationary or not.



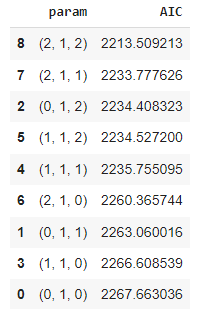
We see that at 𝛼 = 0.05 the Time Series is stationary.

**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-**

****

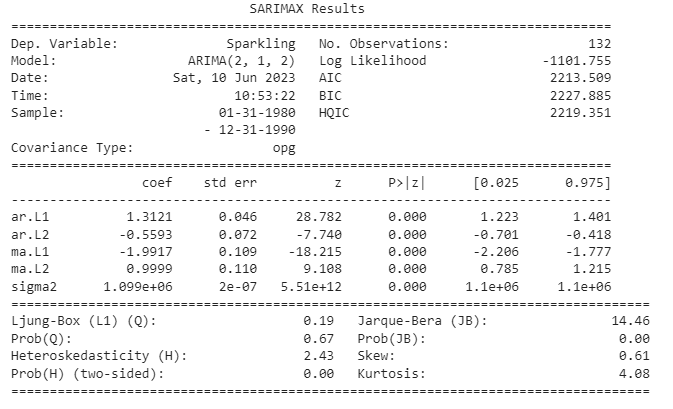
We have to check AIC for all the above combinations-



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

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(2, 1, 2) model is:

Δy(t) = 1.3121 \* Δy(t-1) - 0.5593 \* Δy(t-2) - 1.9917 \* ε(t-1) + 0.9999 \* ε(t-2) + ε(t)

where:

* Δy(t) represents the first-order difference of the dependent variable at time t (Sparkling).
* Δy(t-1) and Δy(t-2) are the first-order differences of the dependent variable at time t-1 and t-2, respectively.
* 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.
* 1.3121, -0.5593, -1.9917, and 0.9999 are the estimated coefficients of the autoregressive (AR) and moving average (MA) terms.

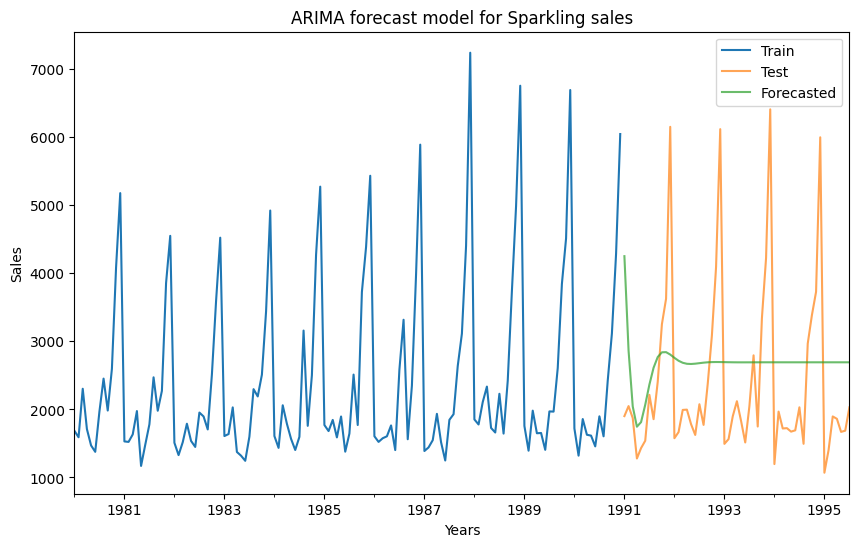
Based on the provided results, the coefficients of ar.L1, ar.L2, ma.L1, and ma.L2 are estimated for the model. The significance and impact of these coefficients can be evaluated by considering their magnitude and statistical significance.

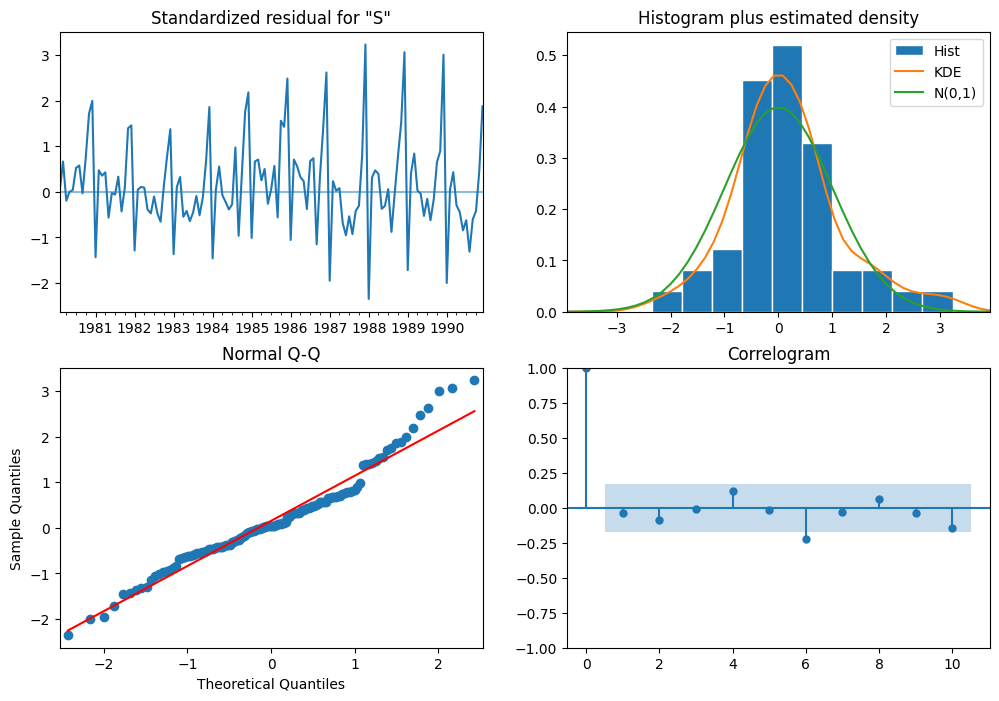
From the results, it can be observed that all the AR and MA coefficients have a p-value of 0.000, indicating that they are statistically significant. Additionally, the magnitudes of these coefficients can also provide insights into their importance.

In this case, the coefficients with the highest magnitudes are ar.L1 (1.3121) and ma.L1 (-1.9917). These coefficients indicate a strong influence of the lagged values and the error terms on the current value of the dependent variable.

Therefore, in terms of prediction, the ar.L1 and ma.L1 coefficients can be considered as the most important parameters in this ARIMA(2, 1, 2) model, as they have the largest magnitudes and are statistically significant.

Forecast looks like this-



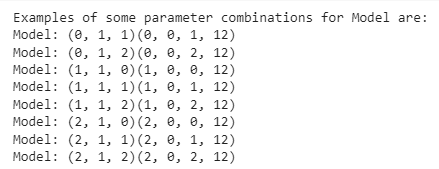


Histogram plus estimated density plot shows that the distribution of residuals is a bit deviated from normal , but can be considered as normal , the same is shown by the Q\_Q plot too. A bit of a deviation is these at one of the tails.

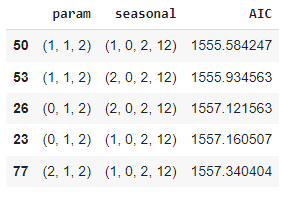
Correlogram shows that there are no remaining patterns or dependencies in the residuals.

**RSME** of ARIMA(2,1,2) is = **1299.979749**

**SARIMA-**

****

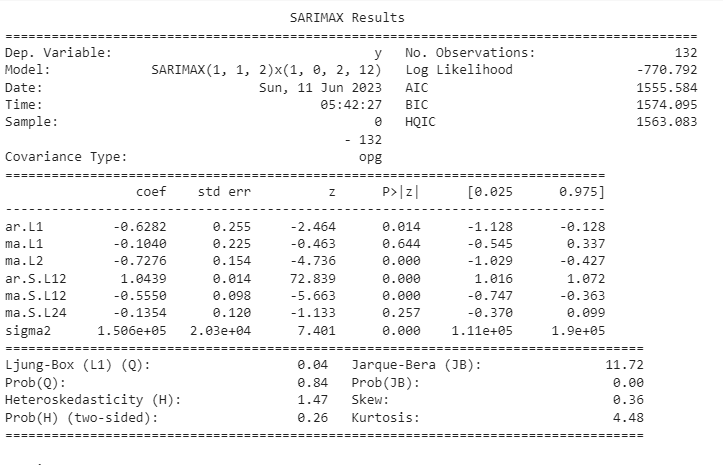
We have to check AIC for all the above combinations-



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

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 -



The equation for the SARIMAX(1, 1, 2)x(1, 0, 2, 12) model is:

Δy(t) = -0.6282 \* Δy(t-1) - 0.1040 \* ε(t-1) - 0.7276 \* ε(t-2) + 1.0439 \* Δy(t-12) - 0.5550 \* ε(t-12) - 0.1354 \* ε(t-24) + ε(t)

where:

Δy(t) represents the first-order difference of the dependent variable at time t.

Δy(t-1) is the first-order difference of the dependent variable at time t-1.

Δy(t-12) is the first-order difference of the dependent variable at time t-12.

ε(t) represents the residual error at time t.

ε(t-1) and ε(t-2) are the lagged residual errors at time t-1 and t-2, respectively.

ε(t-12) and ε(t-24) are the lagged residual errors at time t-12 and t-24, respectively.

-0.6282, -0.1040, -0.7276, 1.0439, -0.5550, and -0.1354 are the estimated coefficients of the autoregressive (AR) and moving average (MA) terms.

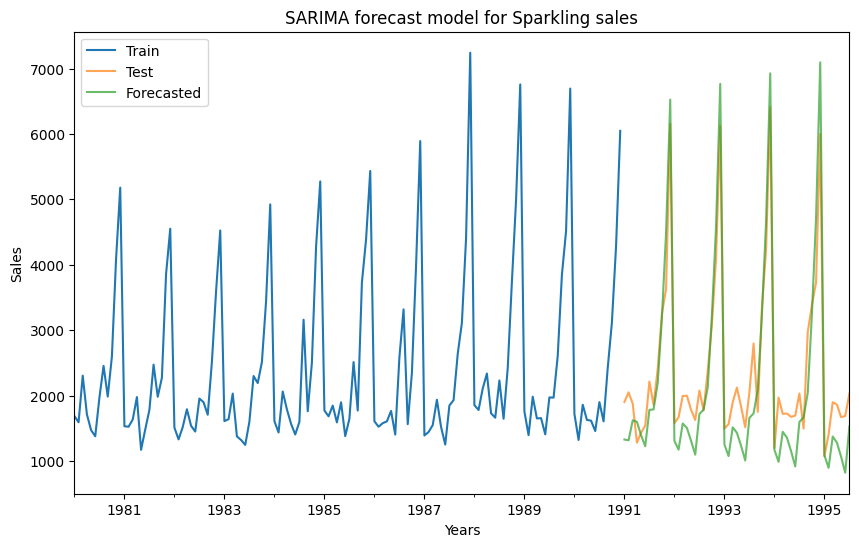
From the provided results, it can be observed that the coefficient ar.L1 (-0.6282) and ma.L2 (-0.7276) are statistically significant with p-values less than 0.05. These coefficients represent the autoregressive and moving average components of the model, respectively, and indicate their influence on the current and lagged values of the dependent variable.

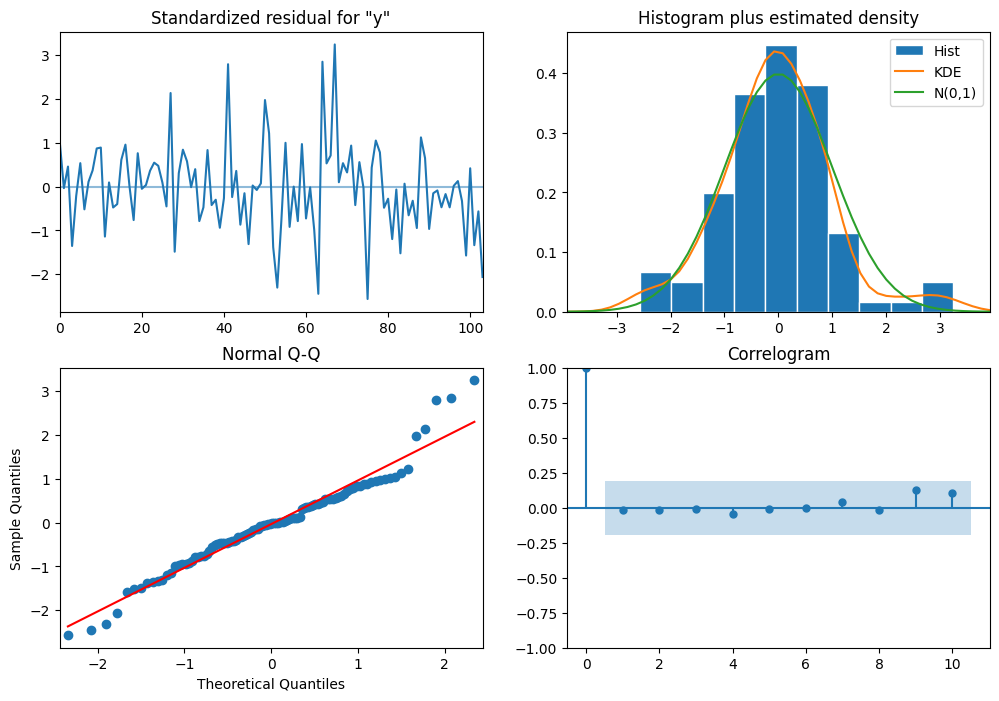
The coefficient ma.L1 (-0.1040) is not statistically significant with a p-value of 0.644, suggesting that it may not have a significant impact on the model.

The seasonal components of the model, represented by the coefficients ar.S.L12 (1.0439), ma.S.L12 (-0.5550), and ma.S.L24 (-0.1354), also play a role in capturing the seasonal patterns at a lag of 12 and 24 months.

In terms of prediction and model interpretation, the significant coefficients (ar.L1, ma.L2, ar.S.L12, and ma.S.L12) should be given more importance as they have a statistically significant impact on the model's performance.

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)= 528.6593092642535**

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.**

****

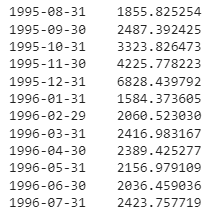
**8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.**

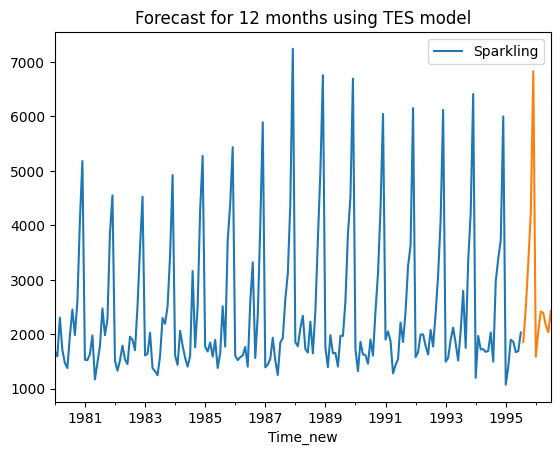
We will take 3 models-

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

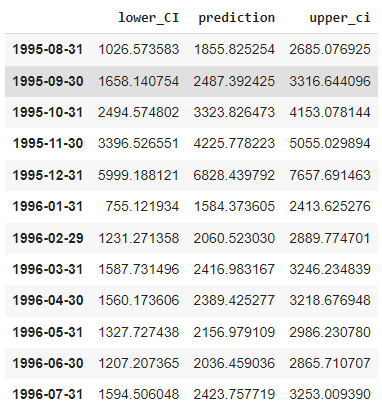
RMSE on full data is = 421.95

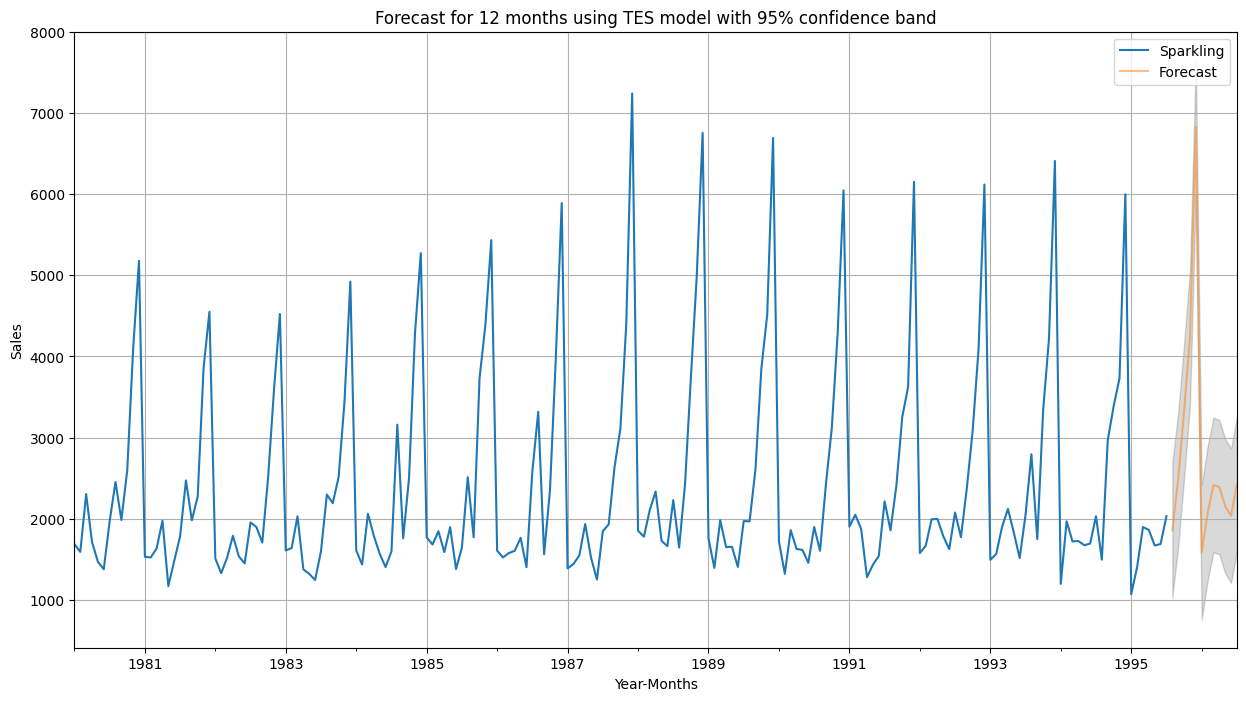
12 months into the future-





With 95% confidence interval few top records are displayed-

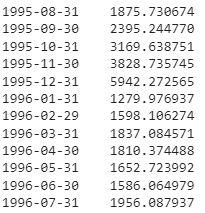


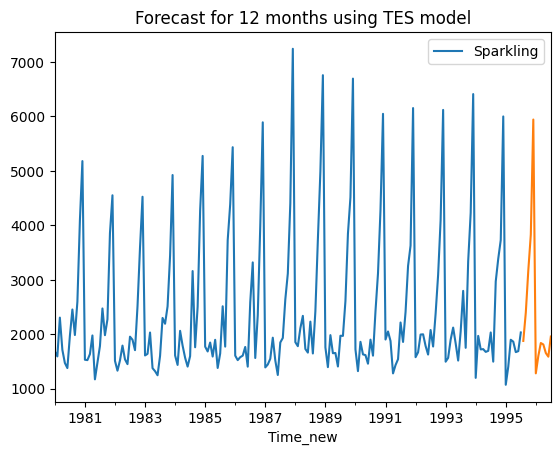


1. **TES - Alpha=0.111,Beta=0.049,Gamma=0.363,TripleExponentialSmoothing**

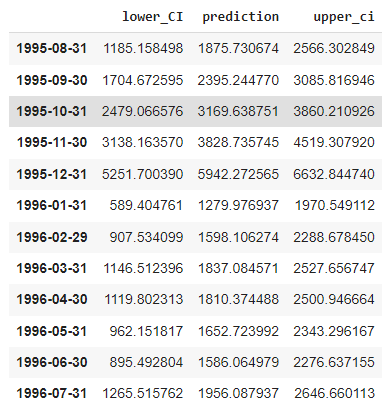
RMSE on full data is = 351.413

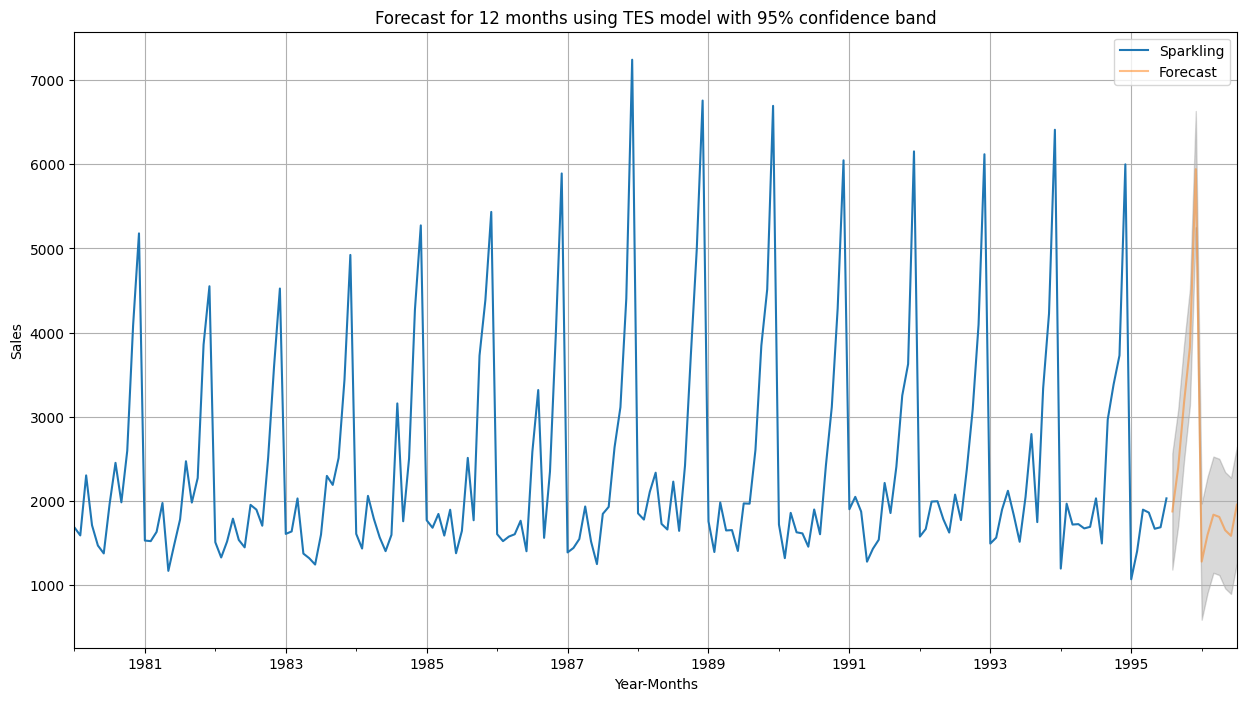
12 months into the future-





With 95% confidence interval few top records are displayed-

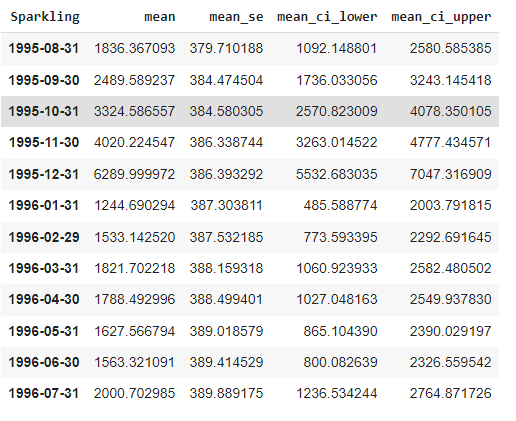


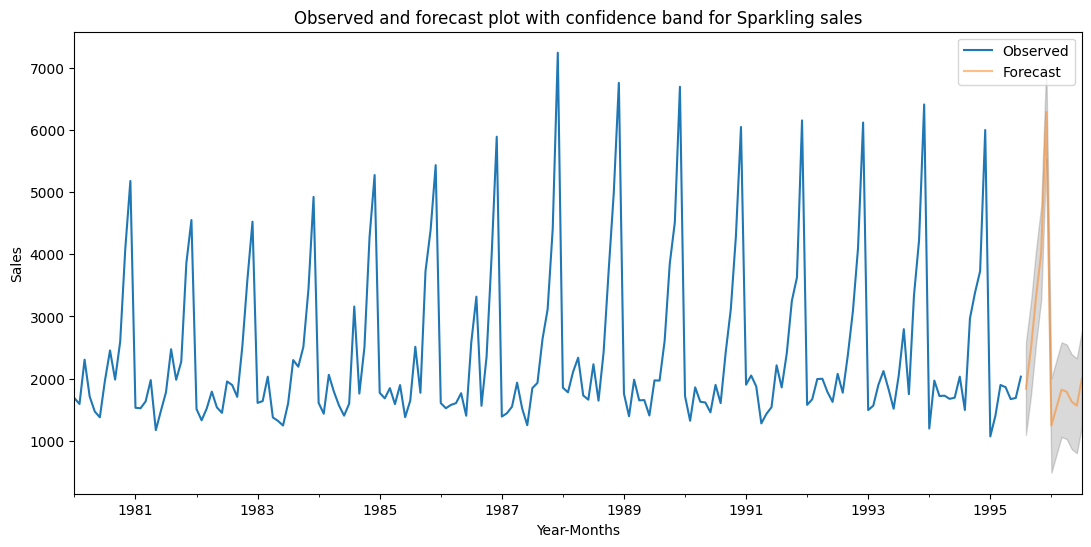


### SARIMA (1,1,2)(1,0,2,12) forecast

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=0.111,Beta=0.049,Gamma=0.363**
2. **TES - alpha = 0.3, beta = 0.3, gamma = 0.3**

**Few measures for the company-**

* Optimize production and inventory management: Leverage the seasonal component (Gamma) to adjust production levels and ensure sufficient stock availability during high-demand seasons. This helps avoid stockouts and capitalize on increased consumer interest.
* Plan seasonal promotions and marketing initiatives: Utilize the insights from the seasonal component to plan targeted promotions, discounts, and advertising campaigns during peak seasons. This can help drive sales and create a sense of urgency among consumers.

**Seasonality**:

**End-of-year Peak**: Sales reach their peak in December with a value of 5942.27. This suggests a seasonal pattern where customers tend to make more purchases during the holiday season. This could be due to factors like gift-giving, year-end promotions, or increased consumer spending.

**Beginning-of-year Dip**: Sales are relatively lower in January with a value of 1279.98. This dip in sales after the holiday season could be attributed to factors such as decreased consumer spending, post-holiday fatigue, or a focus on saving after the expenses of the previous months.

**Business Approach:**

**Capitalize on Peak Season**: The business should anticipate and plan for the increased demand during the end-of-year peak season. This includes ensuring sufficient inventory levels, optimizing production capacity, and implementing targeted marketing campaigns to attract holiday shoppers. Offering special promotions, discounts, or bundling deals can also help drive sales during this period.

**Mitigate the Beginning-of-year Dip**: The business should be prepared for the lower sales volume at the beginning of the year. This can involve adjusting production levels, managing inventory efficiently, and implementing cost-saving measures to offset the decrease in revenue. Additionally, the business can consider introducing new products or services that align with customers' post-holiday needs or resolutions.