**Project - Time Series Forecasting(Sparkling)**

| **Criteria** | **Points** |
| --- | --- |
| 1. Read the data as an appropriate Time Series data and plot the data. | 2 |
| 2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. | 9 |
| 3. Split the data into training and test. The test data should start in 1991. | 2 |
| 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. | 16 |
| 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. | 4 |
| 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. | 11 |
| 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. | 2 |
| 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. | 3 |
| 9. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.  Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present. | 5 |

**Table list**

Table 1.Dataset Sample(head)--5

Table 2.Dataset Sample(tail)--5

Table 3.Dataset info--6

Table 4.Dataset null count--6

Table 5.Dataset Description--6

Table 6.Pivot table for dataset--9

Table 7.Training Data--15

Table 8.Test Data--16

Table 9.Train and Test data for Linear Regresssion--18

Table 10.RMSE on test data using linear regression--19

Table 11.Naive Approach--19

Table 12.RMSE for Naïve Approach--20

Table 13.Simple Average--21

Table 14.RMSE for Simple Average--22

Table 15.Trailing moving Averages--22

Table 16.RMSE on training data on RMSE--24

Table 17.RMSE--24

Table 18: RMSE for simple smoothning--27

Table 19: RMSE on Double Smoothing Model--28

Table 20:RMSE on triple smoothing--30

Table 21.AR model summary--34

Table 22.Summary showing ARMA--36

Table 23.ARIMA Model Summary--38

Table 24.Consolidated RMSE table for all models--42

Table 25.Summary for Full data--43

Figure list

Fig 1. Sparkling Wine Sales graph--7

Fig 2. Year on Year Sparkling Wine Sales plot--8

Fig 3.Monthly plot for Sparkling wine--8

Fig 4.Month Plot of given Time Series--9

Fig 5 Yearly wine sale across Month--10

Fig 6.Empirical Cumulative Distribution--11

Fig 7.Wine Sales per Month and Month on Month--12

Fig 8. Additive Model--13

Fig 9.Multiplicative Model--14

Fig 10.Graph for test Data using Linear Regression--19

Fig 11.Graph for Test data using Naïve Forecast--20

Fig 12. Graph for Test data using Simple Average--21

Fig 13. Graph of Trailing moving averages--23

Fig 14. Graph of Trailing moving averages on test and train data.--23

Fig 15. All the models performances--25

Fig 16.SimpleSmoothing Model--26

Fig 17.Graph on Double smoothing Model--28

Fig 18.Graph on Triple Smoothing Model.--29

Fig 19: adfuller test result--32

Fig 20:Adfuller test result after difference--33

Fig 21.Graph for AR Model--35

Fig 22. Graph showing ARMA model--37

Fig 23.Graph showing ARIMA Model--39

Fig 24.Autocorrelation--40

Fig 25.Table showing SARIMA Model summary--41

Fig 26.Graph showing SARIMA model--41

Fig 27.Correlation--44

Fig 28. Next 12 months prediction--46

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

**Answer:-**

We have loaded the all the required packages and loaded Sparkling Wine sales Data file using Pandas.

Dataset has 187 rows and 1 column as we have set YearMonth Column as index.

We have viewed first and last few rows using head() and tail() functions respectively.

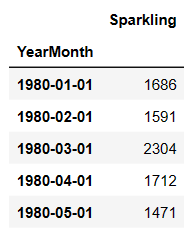


Table 1.Dataset Sample(head)

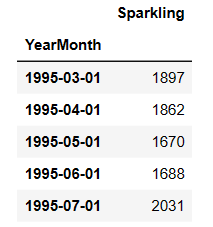


Table 2.Dataset Sample(tail)

We can view dataset information using info()

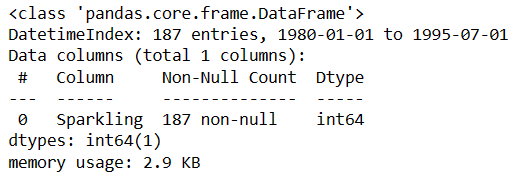


Table 3.Dataset info

There is one column with integer datatype. There are no null values.

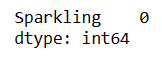


Table 4.Dataset null count

Dataset can be described using describe() function

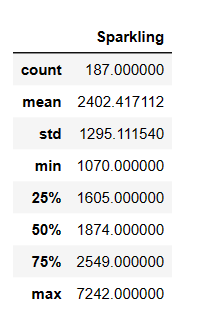


Table 5.Dataset Description

1. The Average Sparkling Wine sale is arounf 2400.
2. At Max 7242 wines were sold and at least 1070 wines were sold.
3. Total 187 times wine was sold.

We have plotted data for Sparkling Wine sales.

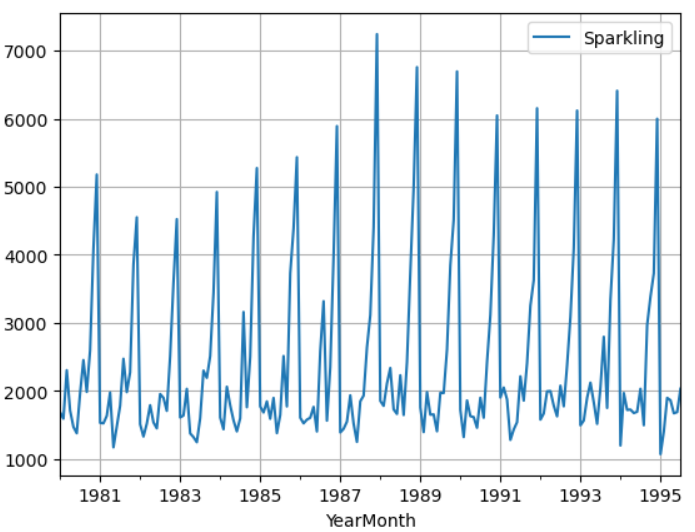


Fig 1. Sparkling Wine Sales graph

**Inference**

We observe both trend and multiplicative seasonality from the plot shown above.

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

**Answer:-**

We have plotted a box and whisker (1.5\* IQR) plot to understand the spread of the data and check for outliers in each year, if any.

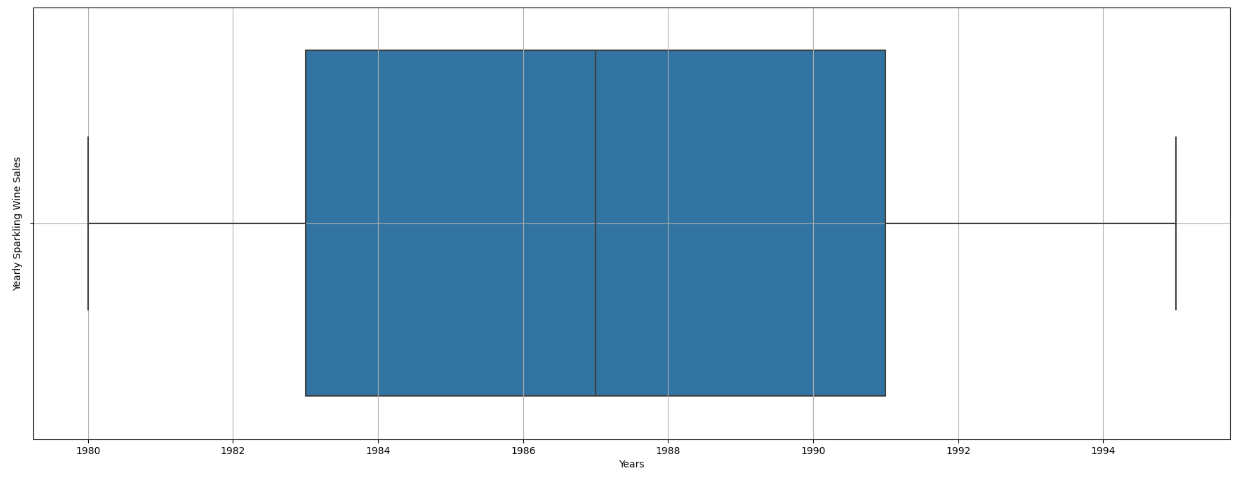


Fig 2. Year on Year Sparkling Wine Sales plot

**Inference:-**As we got to know from the Time Series plot, the boxplots over here also indicates a measure of trend being present. Also, we see that the Sale of Sparkling Wine has some outliers for certain years.

Since this is a monthly data, let us plot a box and whisker (1.5\* IQR) plot to understand the spread of the data and check for outliers for every month across all the years, if any.

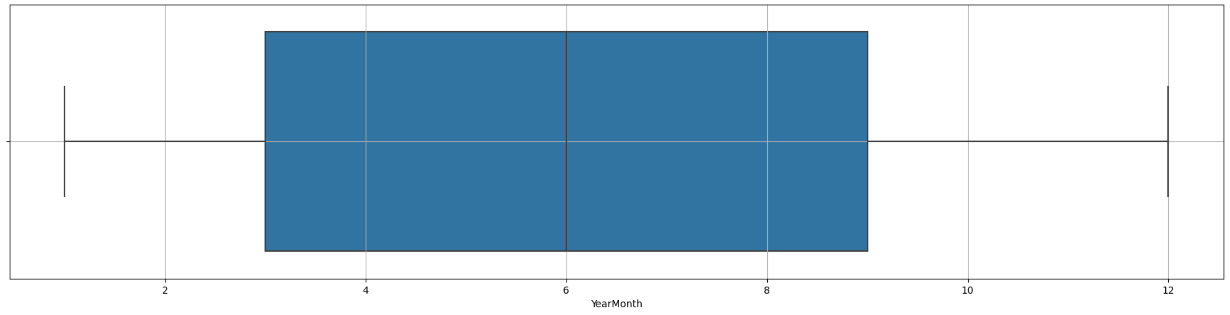


Fig 3.Monthly plot for Sparkling wine

The boxplots for the monthly production for different years does not show any outliers.

We have plotted monthplot of the give Time Series.

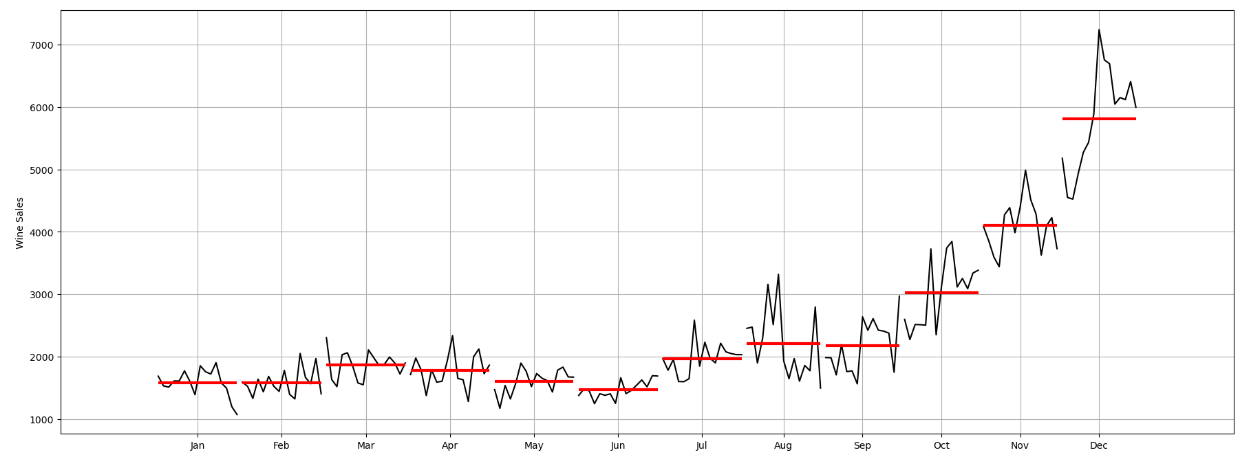


Fig 4.Month Plot of given Time Series

**Inference:-**

We can see that the trend is increasing rapidly from August to December every year.Trend is constant from Jan to July.We can say that wine sale increases in Winter season.

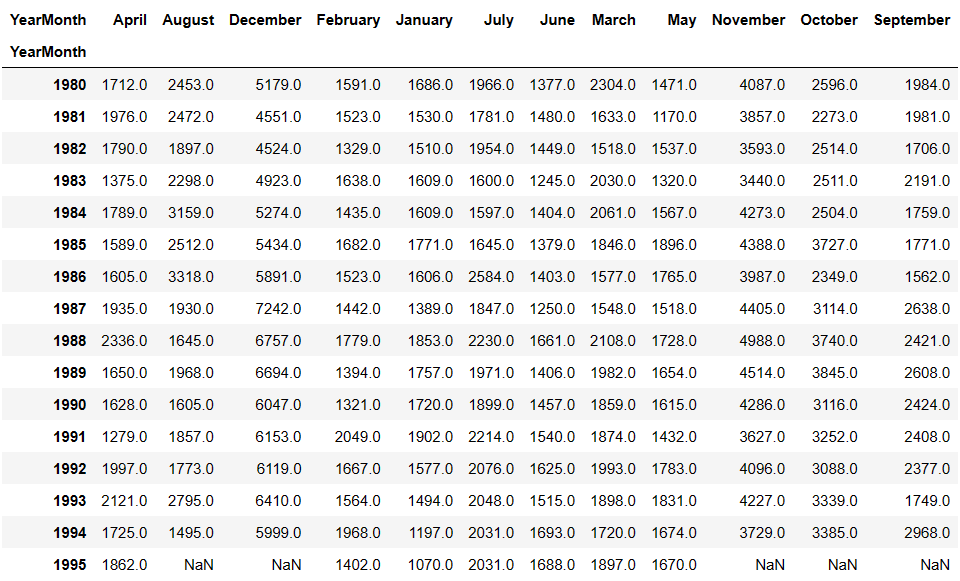


Table 6.Pivot table for dataset

Above table shows pivoted data for Sparkling Wine sales dataset.

Lets plot this data and draw inference out of it.

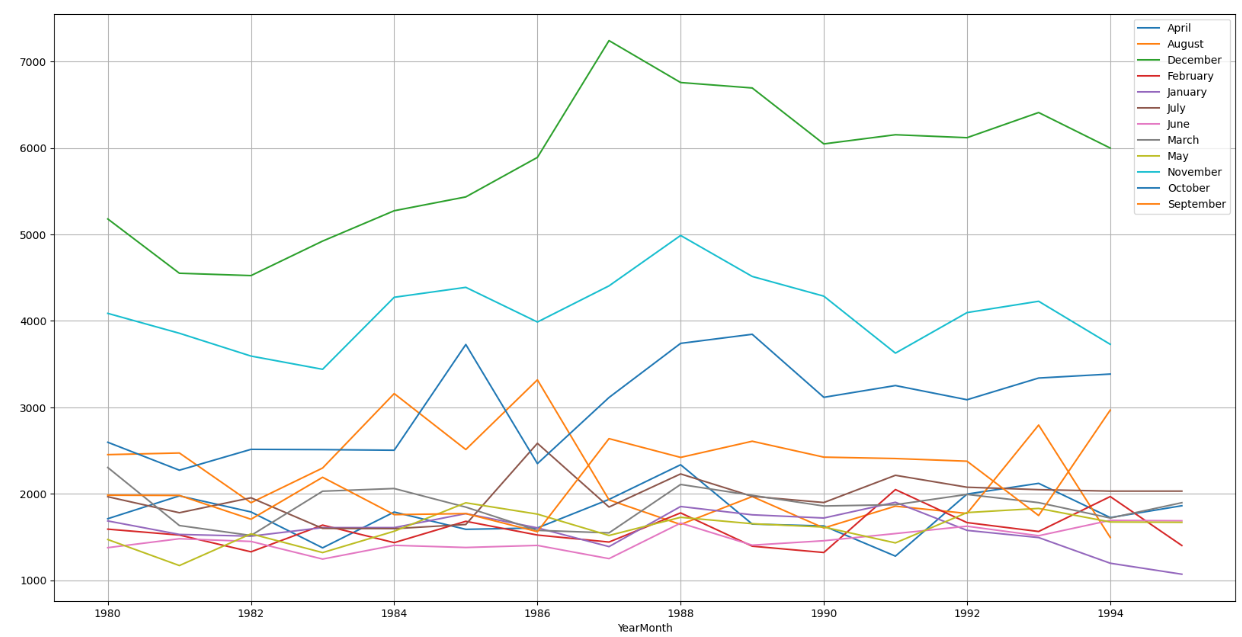


Fig 5 Yearly wine sale across Month

**Inference:-**

We can see every year wine sales drastically increases in the month of December. There is gradual increase in the sales from October to December.

We have plotted Empirical Cumulative Distribution graph for Wine Sales.

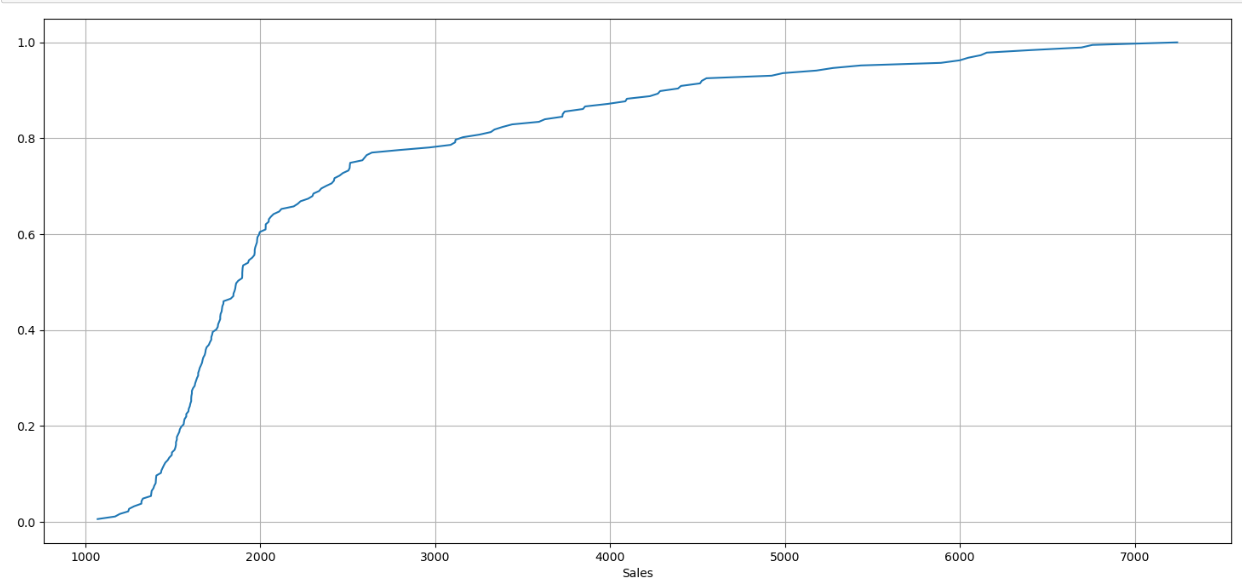


Fig 6.Empirical Cumulative Distribution

**Inference:-**

This particular graph tells us what percentage of data points refer to what number of Sales.

We have plotted Plot the average Sales per month and the month on month percentage change of Sales.

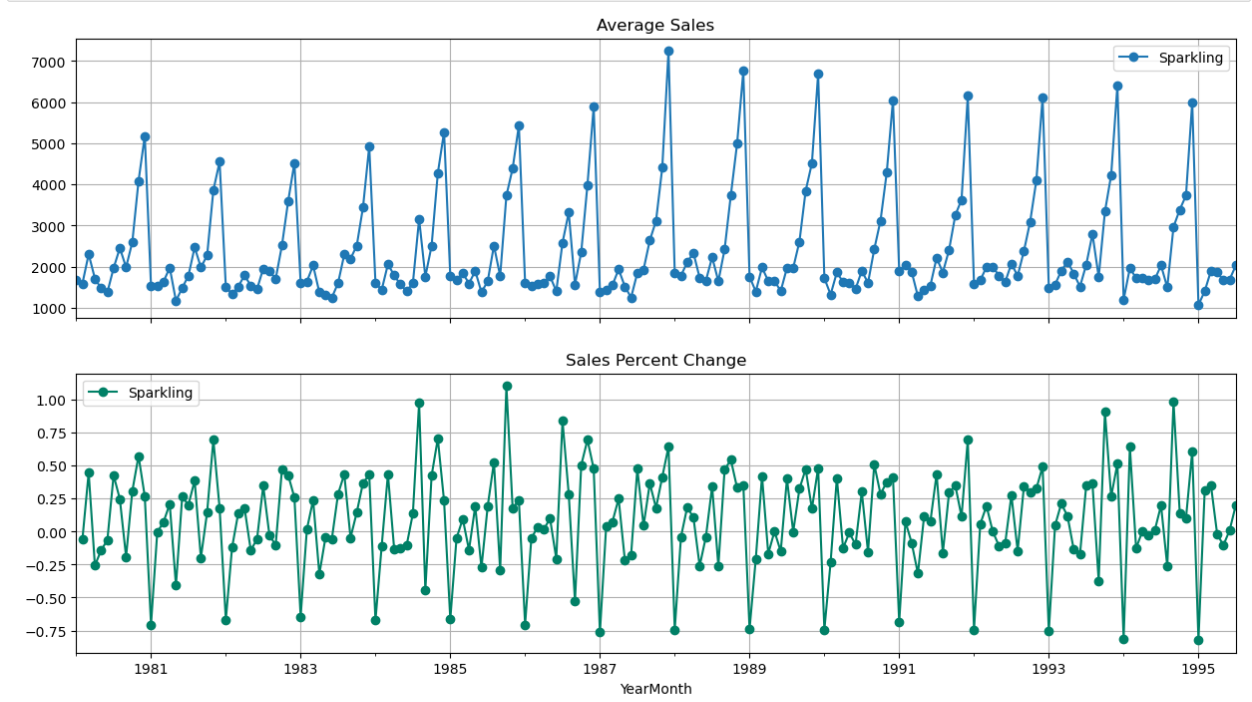


Fig 7.Wine Sales per Month and Month on Month

**Inference:-**

The above two graphs tells us the Average Sales and the Percentage change of Sales with respect to the time.

We have decomposed the Series and plotted Different Components.

**Additive Decomposition**

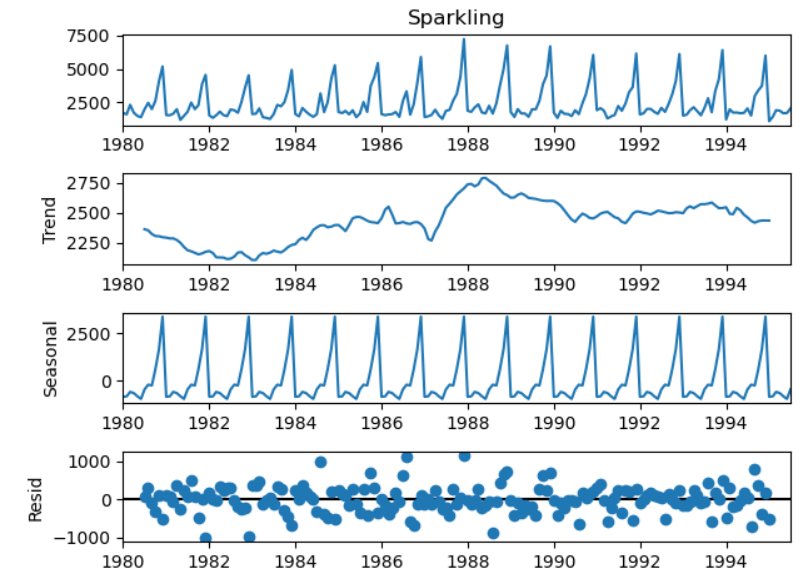
****

Fig 8. Additive Model

**Inference:-**

As per the 'additive' decomposition, we see that there is no finite trend in the earlier years of the data. There is a seasonality in the data. We see that the residuals are located around 0 from the plot of the residuals in the decomposition.

**Multiplicative Decomposition**

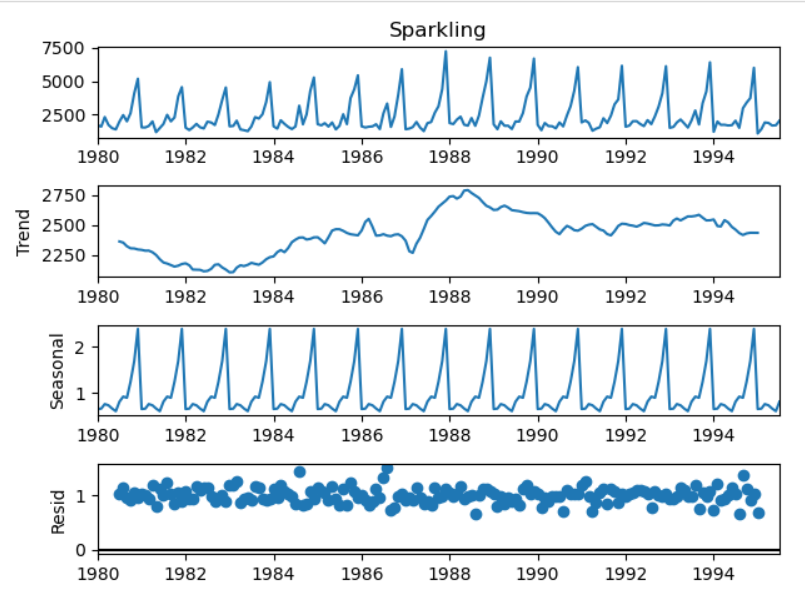
****

Fig 9.Multiplicative Model

**Inference:-**

As per the Multiplicative decomposition, we see that there is no finite trend in the earlier years of the data. There is a seasonality in the data.

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.

Also it is evident that there is a 6 month seasonality in the data from the above plots.

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

**Answer:-**

We have split the data into Training data and Test Data.

Records older than 1991 are selected in Training data. And recent data after or equal to 1991 is selected as Test Data.

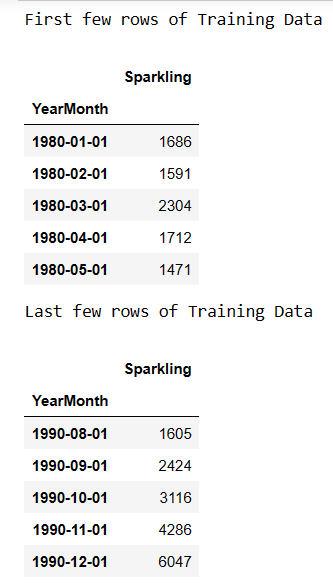


Table 7.Training Data

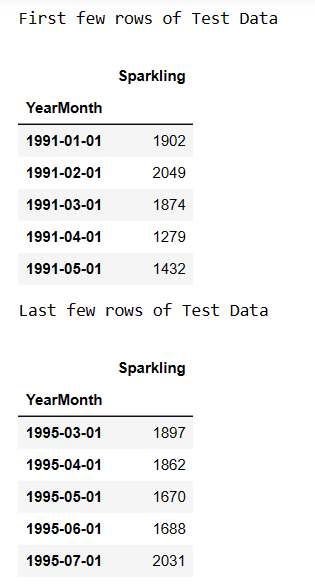


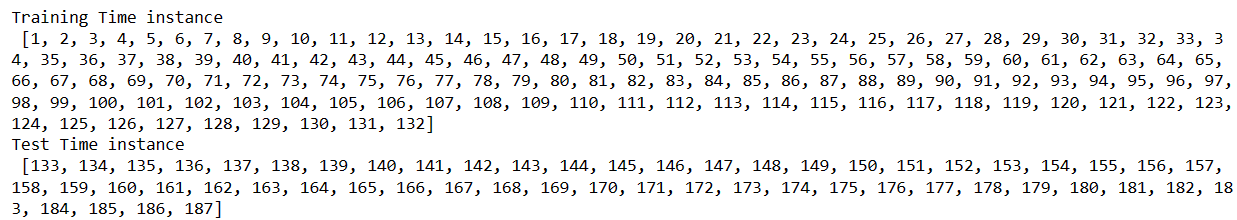
Table 8.Test Data

**Ques 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.**

**Answer:-**

**Building different models and comparing the accuracy metrics.**

1. **Linear Regression Model**

* For this particular linear regression, we are going to regress the Sparkling variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.
* 
* We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

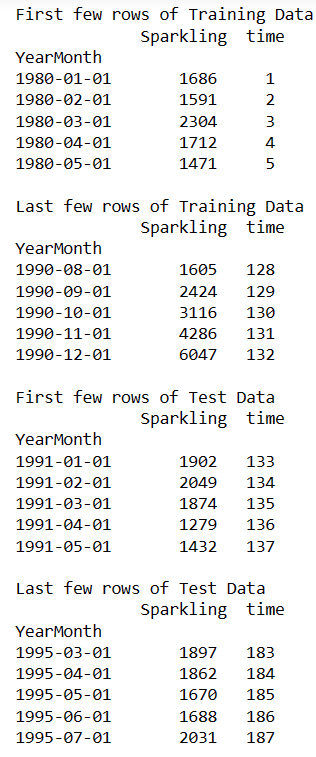


Table 9.Train and Test data for Linear Regresssion

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

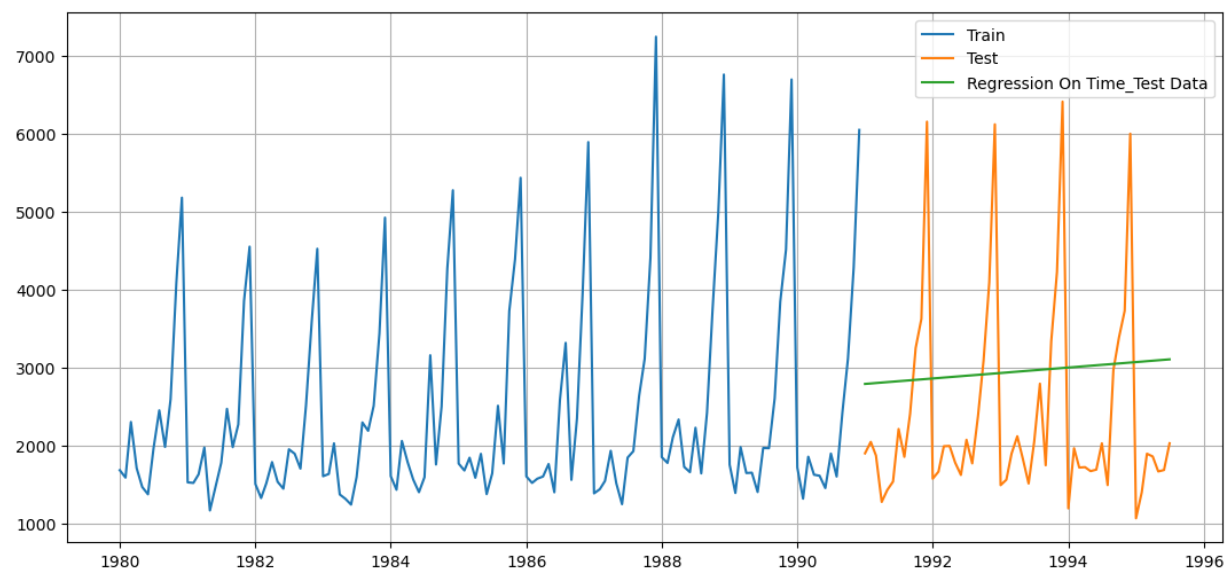


Fig 10.Graph for test Data using Linear Regression

**Inference:-**

Green line is our forecast using the linear expression method. And actual test data is the orange graph. Green line which is our evaluated forecast does not match the actual observation. So we can say that linear regression did not do a very good forecast.

We will do model evaluation by calculating Root mean square error on test data.

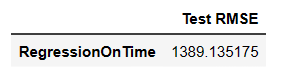


Table 10.RMSE on test data using linear regression

1. **Naive Approach**

* For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today therefore the prediction for day after tomorrow is also today.

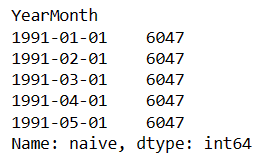


Table 11.Naive Approach

* Now that our training and test data has been modified, let us go ahead use naïve Approach to build the model on the training data and test the model on the test data.

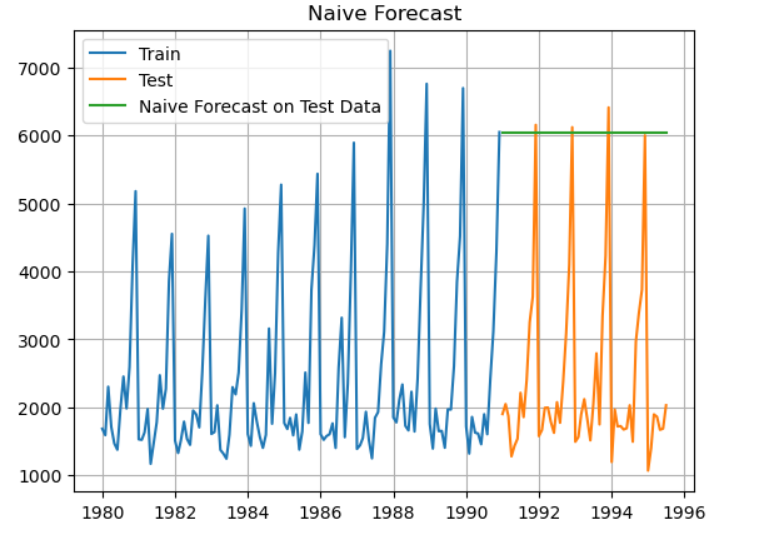


Fig 11.Graph for Test data using Naïve Forecast

**Inference:-**

Green line which is our forecast does not match with our actual test observation. So this model has not performed well.

We will do model evaluation by calculating Root mean square error on test data.

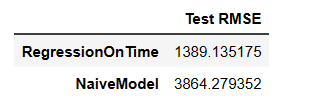


Table 12.RMSE for Naïve Approach

1. **Simple Average**

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

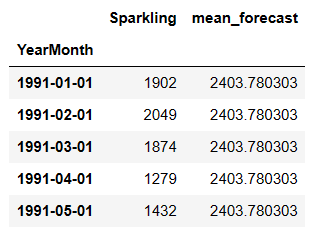


Table 13.Simple Average

* Now that our training and test data has been modified, let us go ahead use Simple Average to build the model on the training data and test the model on the test data.

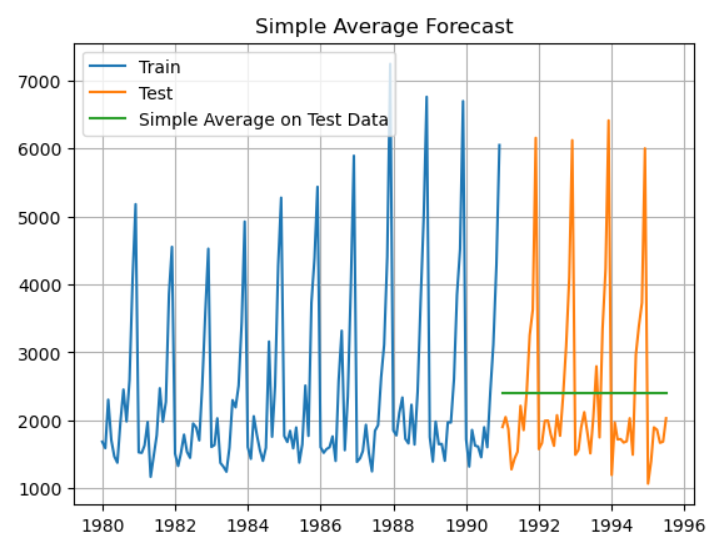


Fig 12. Graph for Test data using Simple Average

**Inference:-**

Forecast does not match with actual test observation. This did perform better than naïve approach but this model cannot be considered good model.

Model Evaluation

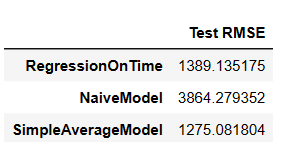


Table 14.RMSE for Simple Average

1. **Moving Average Model**

* For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.
* For Moving Average, we are going to average over the entire data
* Trailing moving averages

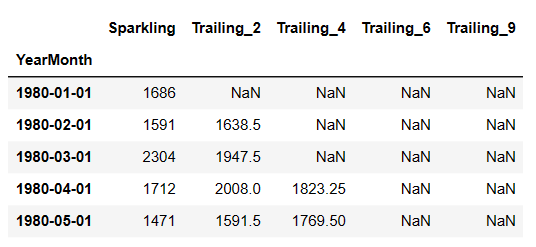


Table 15.Trailing moving Averages

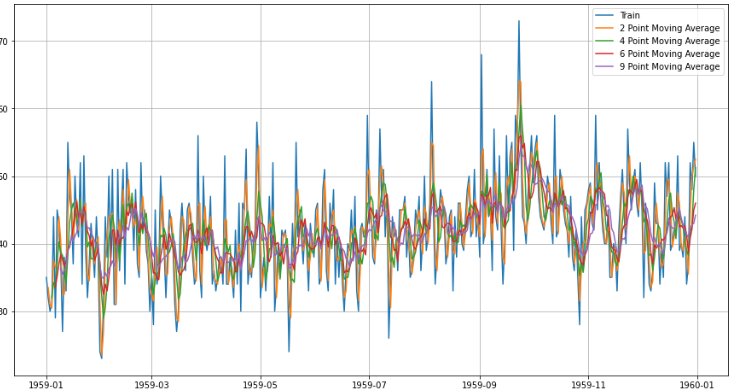
* 

Fig 13. Graph of Trailing moving averages.

* Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.
* 

Fig 14. Graph of Trailing moving averages on test and train data.

* We can see
* We will do model evaluation by calculating Root mean square error on test data.

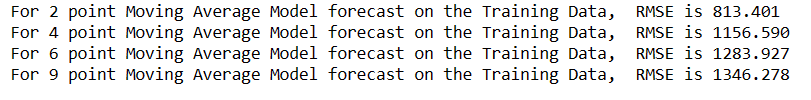


Table 16.RMSE on training data on RMSE

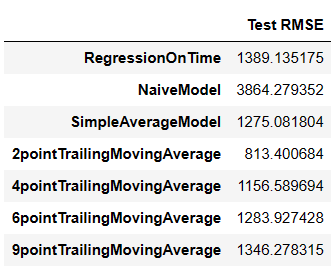
****

Table 17.RMSE

* We can see that 2 point training Moving Average is performing better than other trailing Moving Averages.
* Lets compare all the above models using graph.

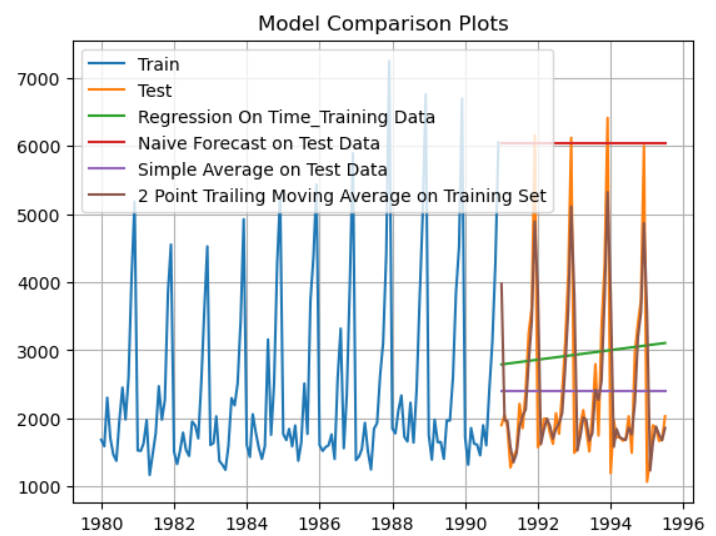


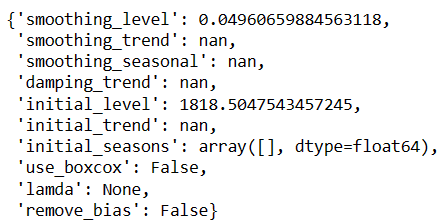
Fig 15.All the models performances

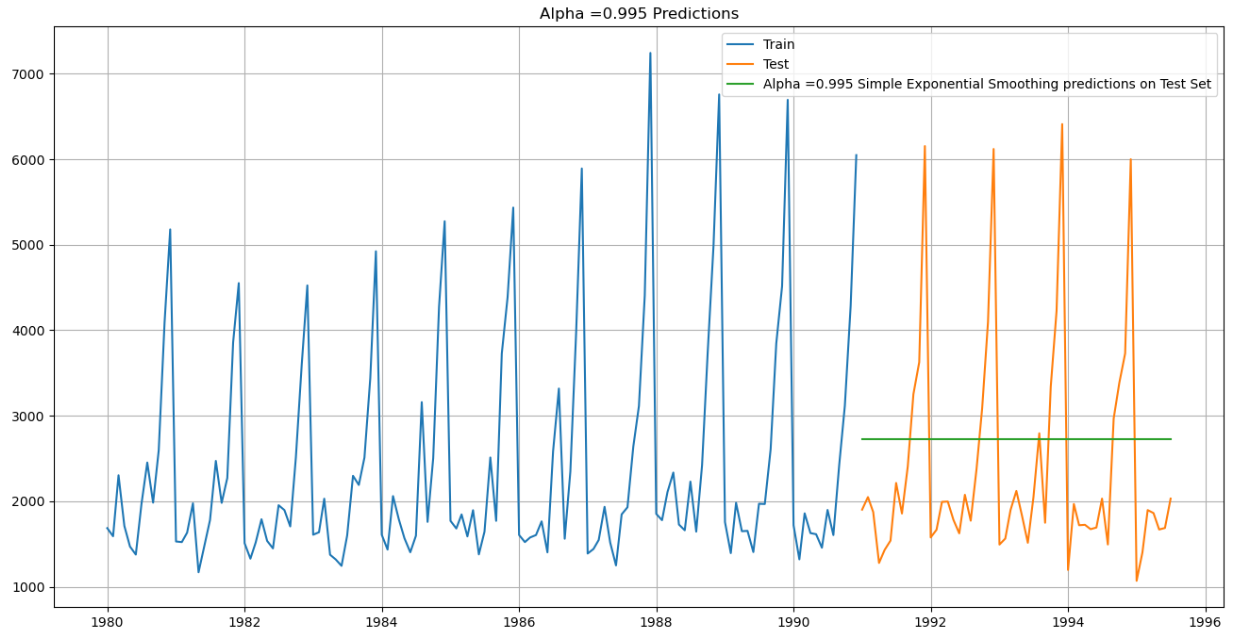
**Inferences:-**

We can see red line shows Naïve Forecast, Green line shows linear regression on test data,purple line shows simple Average and brown shows 2 point training moving averages.

So far 2 point Moving Average is performing better than all other models.

1. **Simple Exponential Smoothing**

* SimpleExpSmoothing class must be instantiated and passed the training data.
* The fit() function is then called providing the fit configuration, the alpha value, smoothing\_level. If this is omitted or set to None, the model will automatically optimize the value.
* Simple Exponential smoothing has only level but no trend and no seasonality.
* We have called SimpleExponential smoothing method with initialization method as estimated.
* Training data was then fitted to simpleexponential model.
* 

 Fig 16.SimpleSmoothing Model

**Inference:-**

Forecast is a straight line which does not match with the trend and seasonality of the test observation. We cannot say that this model is performing well.

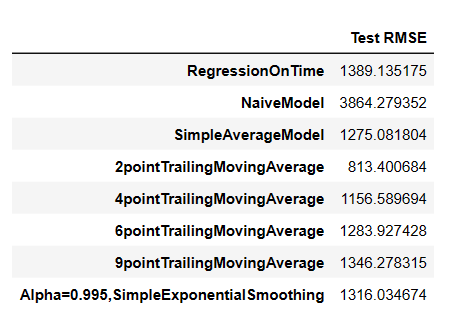
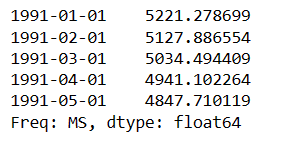


Table 18: RMSE for simple smoothing

1. **Double Exponential Smoothing**

* Two parameters α and β are estimated in this model. Level and Trend are accounted for in this model.
* Double ExpSmoothing class must be instantiated and passed the training data.
* The fit() function is then called providing the fit configuration, the alpha value, smoothing\_level. If this is omitted or set to None, the model will automatically optimize the value.
* 

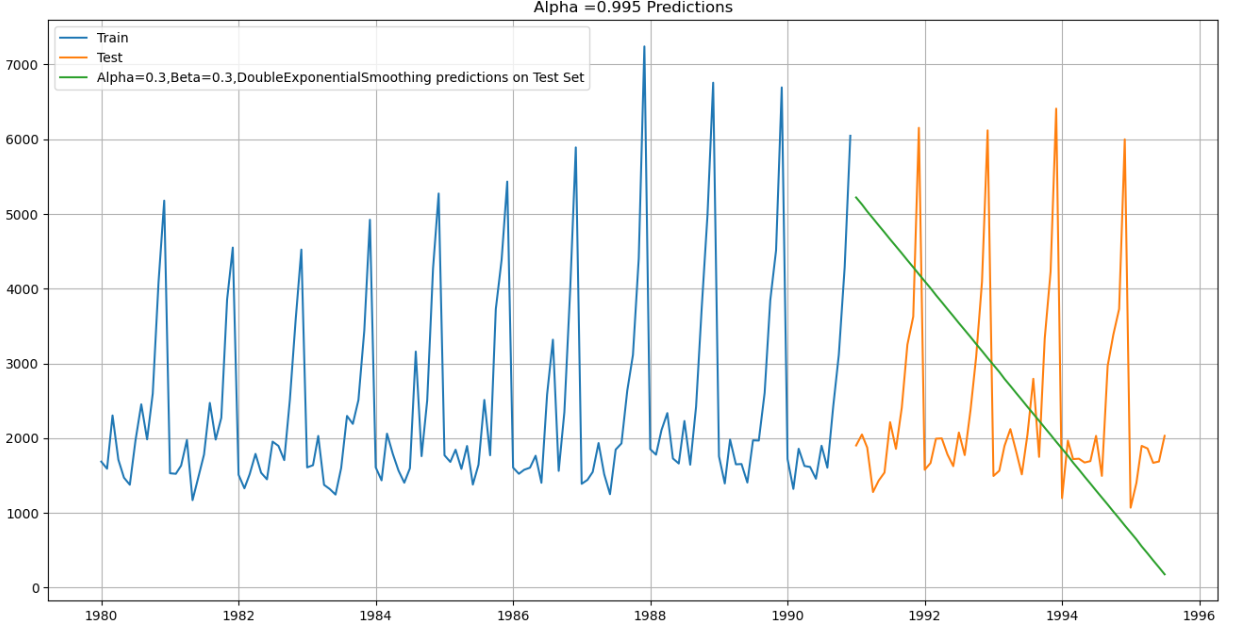


Fig 17.Graph on Double smoothing Model

**Inference:-**

Forecast is a straight line which does not match with the trend and seasonality of the test observation. We have trend here but does not have seasonality.We cannot say that this model is performing well.

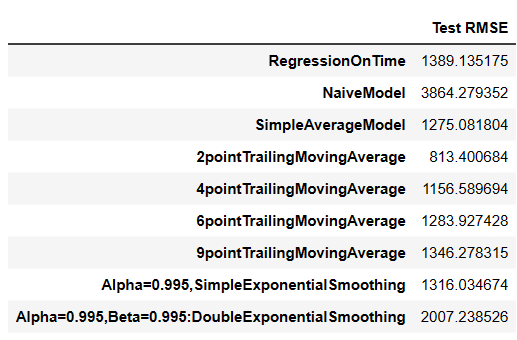
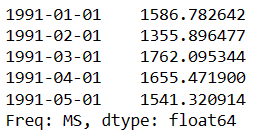


Table 19: RMSE on Double Smoothing Model

1. **Triple Smoothing Model**

* Three parameters α, β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model
* The Triple Smoothing fit of the model is by the best parameters that Python thinks for the model. It uses a brute force method to choose the parameters.
* ****

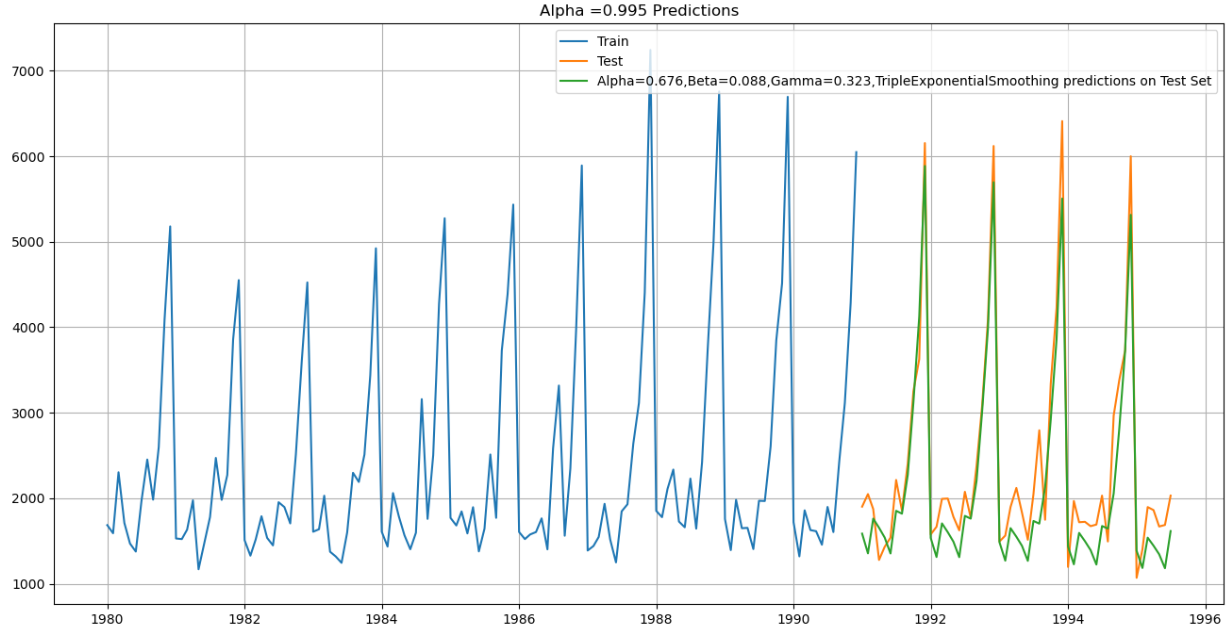
****

Fig 18.Graph on Triple Smoothing Model.

**Inferences:-**

Above graph has all the factors trend,level and seasonality. The forecast is matching the actual test observations.We can say that triple exponential smoothing is best among all the smoothing models.

Triple Exponential Smoothing has performed the best on the test as expected since the data had both trend and seasonality.

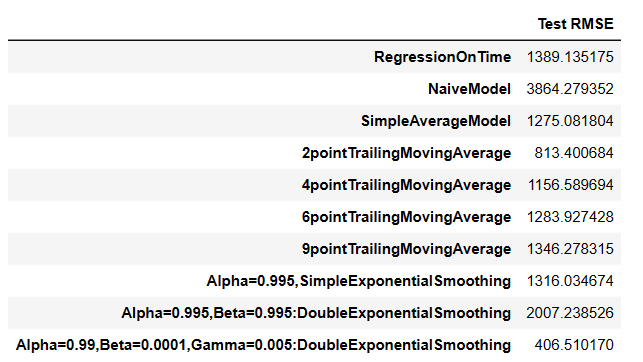


Table 20:RMSE on triple smoothing

We see that the multiplicative seasonality model has not done that well when compared to the additive seasonality Triple Exponential Smoothing model.

**Ques 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**

**Answer:-**

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

Null Hypothesis H0: Time Series is non-stationary.

Alternate Hypothesis 𝑯𝒂: Time Series is stationary.

So Ideally if p-value < 0.05 then null hypothesis: TS is non-stationary is rejected else the TS is non-stationary is failed to be rejected.

We have performed adfuller test to check for stationarity of data. We got below result.



Here p value is greater than 0.05 so data is non stationary.

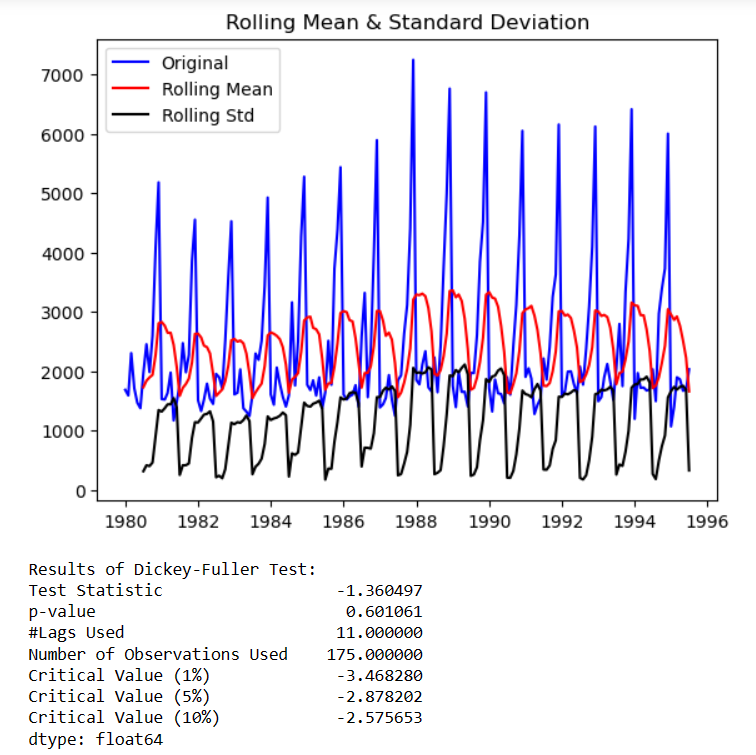


Fig 19: adfuller test result

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

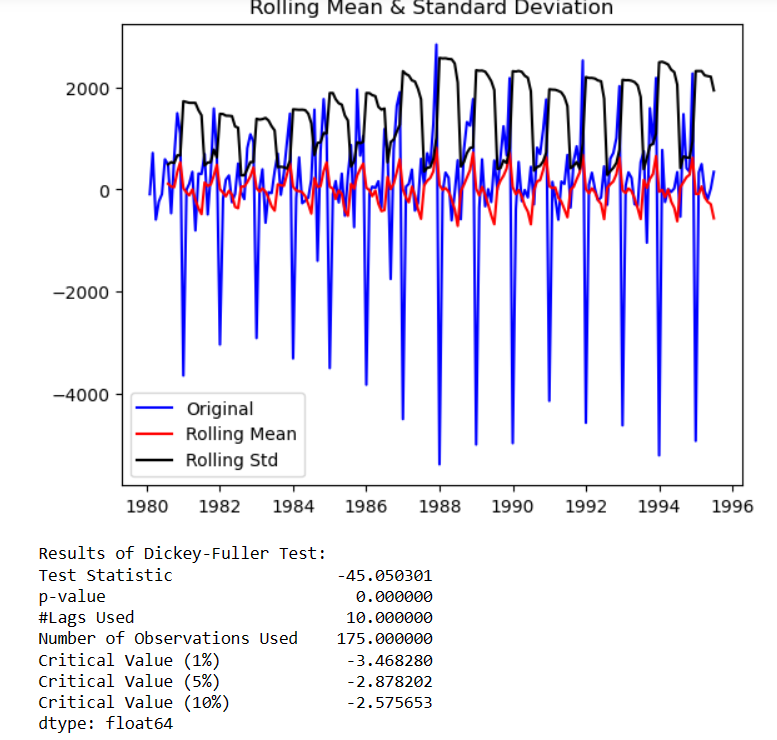
****

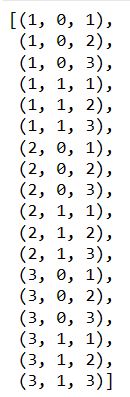
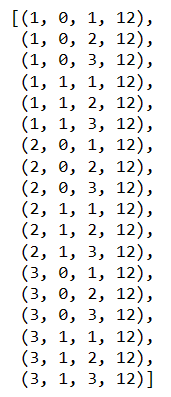
Fig 20:Adfuller test result after difference

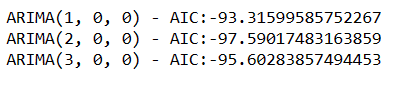
* We can see that p value has reduced to 0 which means that now data is stationary.
* We see that at α = 0.05 the Time Series is indeed stationary.

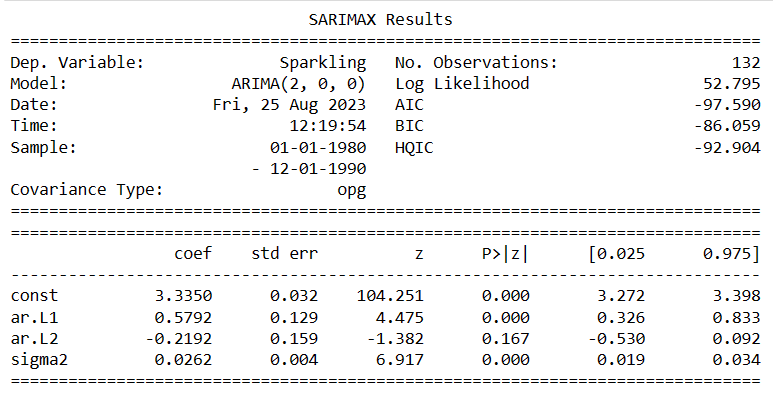
**Ques 6:- Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**

**Answer:-**

1. **AR Model**

* We have set the range for p d and q.
* Below shows the parameter combination for model.
* 
* 
* We are building the AR model with best ‘p’ parameter



 Table 21.AR model summary



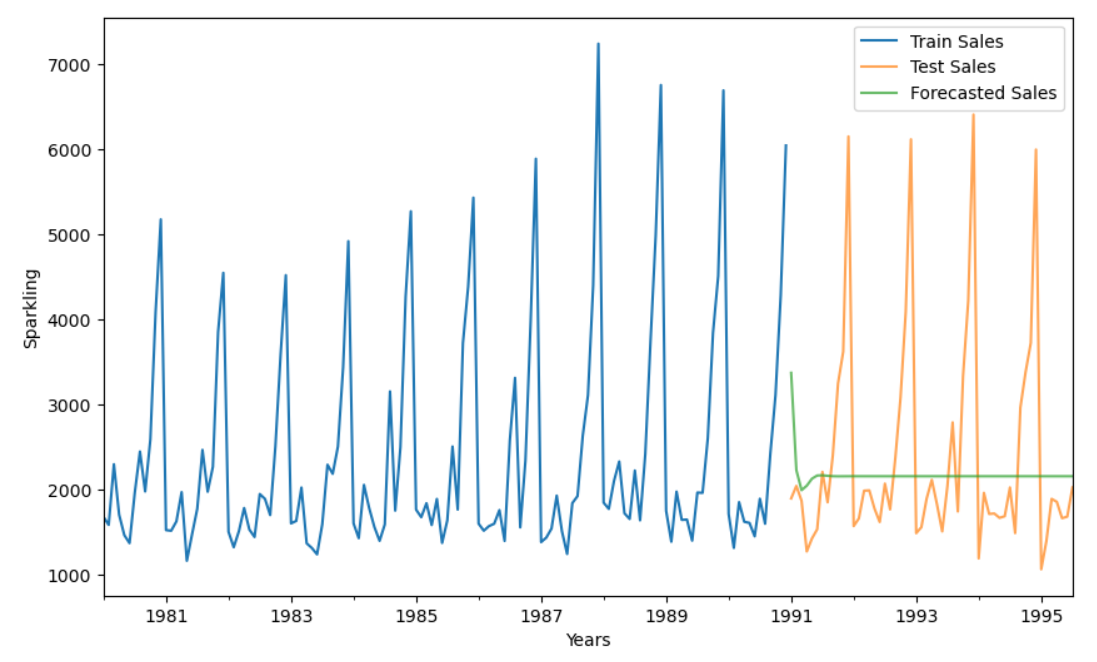


Fig 21.Graph for AR Model

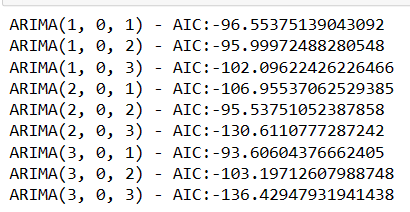
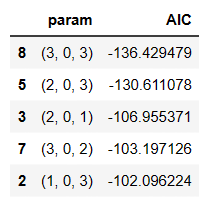
* We can see from above graph that the green line which is our forecast does not match with our actual observation so this plot does not give us desired result.

1. **ARMA MODEL**

* ARMA Model building to estimate best 'p' , 'q' ( Lowest AIC Approach )
* mproving AutoRegressive Models through Moving Average Forecasts.
* ARMA models consist of 2 components:-

AR model: The data is modeled based on past observations.

MA model: Previous forecast errors are incorporated into the model.

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

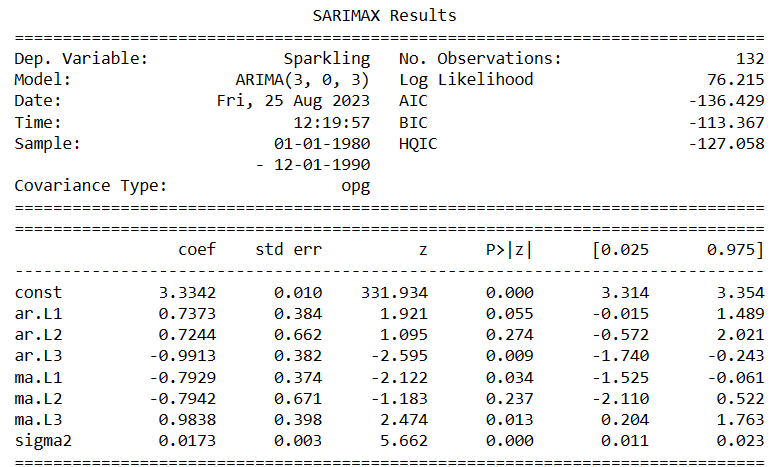


Table 22.Summary showing ARMA

* 

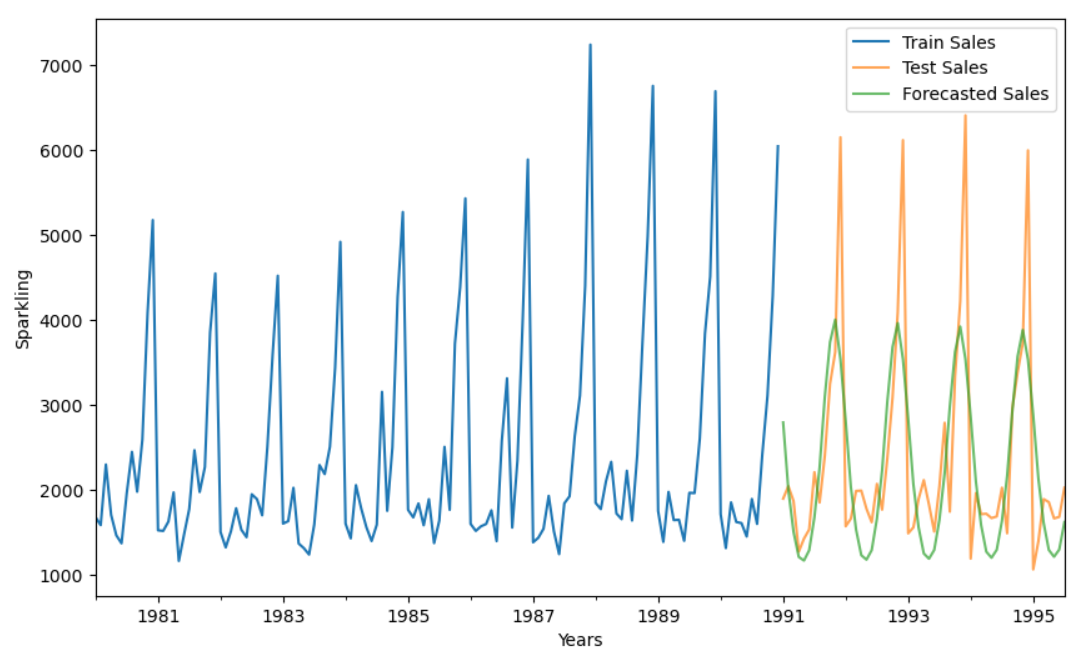
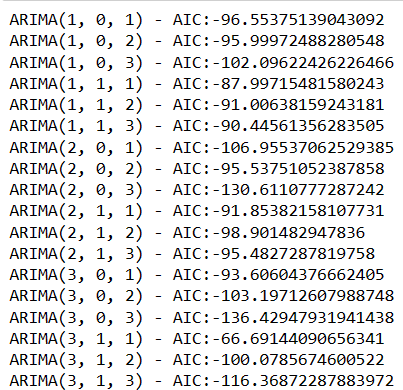
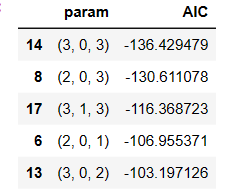


Fig 22. Graph showing ARMA model

* This model has better performance as compared to AR model. RMSE for ARMA is less than AR.As of now this is the good performing model.

1. **ARIMA MODEL**

* ARIMA Model building to estimate best 'p' , 'd' , 'q' paramters ( Lowest AIC Approach )
* ****
* Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value.
* 
* Building ARIMA model with best parameters p,d,q

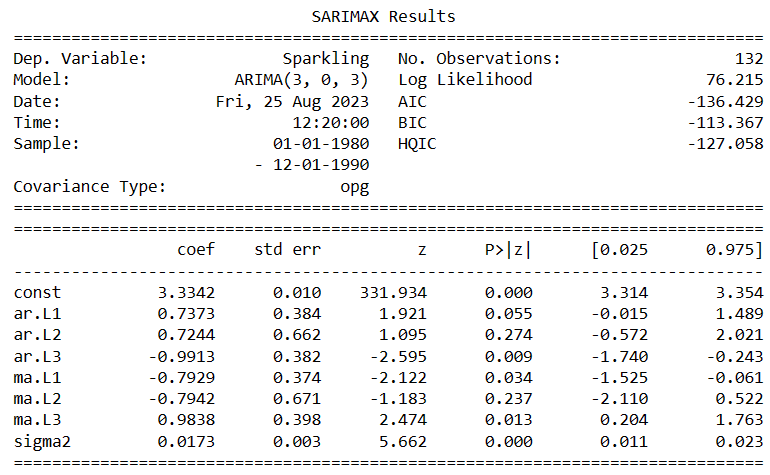


Table 23.ARIMA Model Summary

* 



Fig 23.Graph showing ARIMA Model

* This Model has performed same as ARMA model but better than AR model.
* The forecast is following the pattern of actual observation.
* RMSE for this model is better than AR model but same as ARMA.

1. **SARIMA Model**

* Finding Seasonality = 12 from ACF/PACF plots

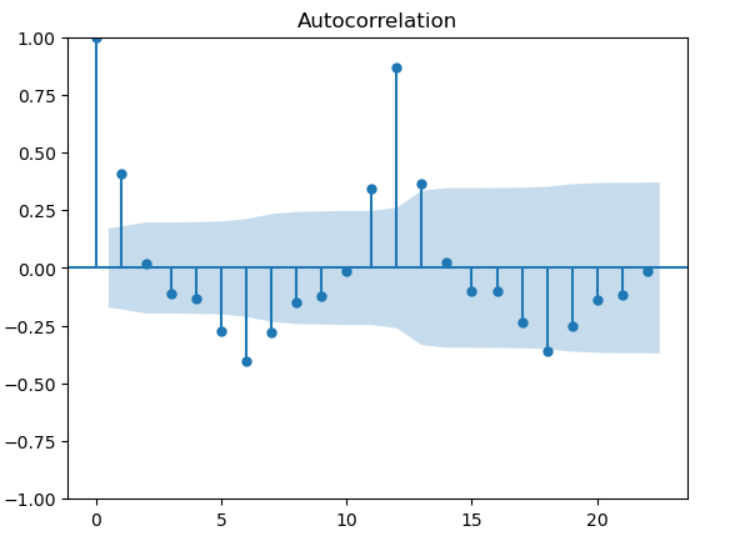


Fig 24.Autocorrelation

* Above figure shows that there is no proper correlation between residuals.
* SARIMA Model building to estimate best parameters
* 



Table 24.Table showing SARIMA Model summary

* We have extracted the predicted and true values of our time series.
* 

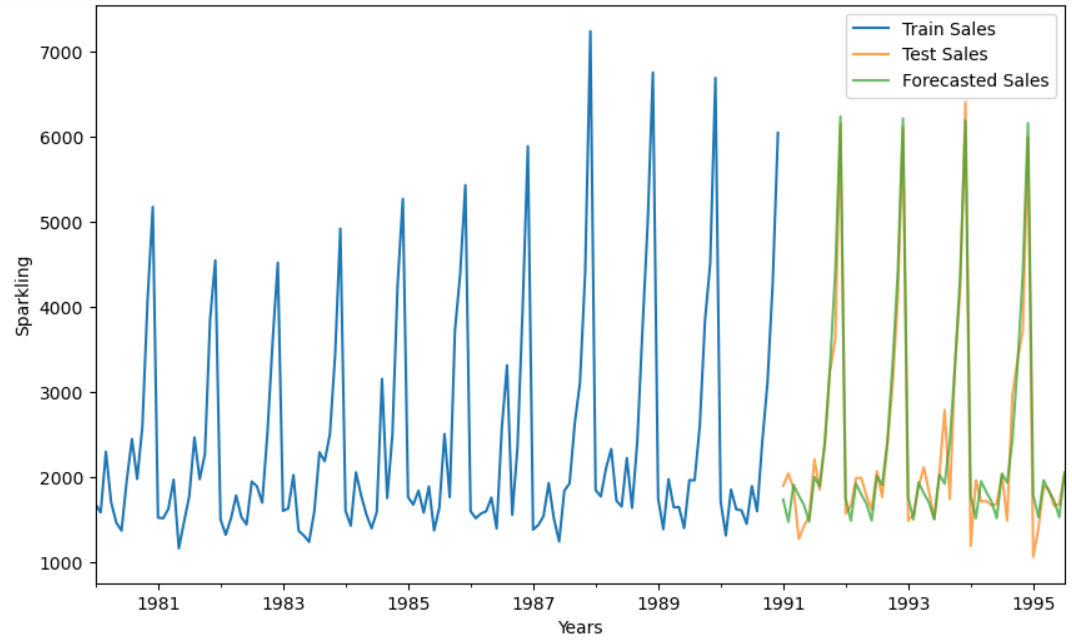
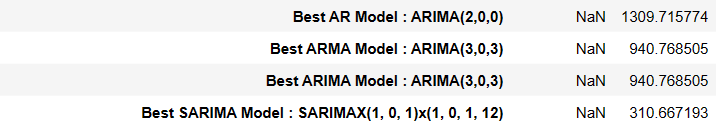


Fig 26.Graph showing SARIMA model



* So far SARIMA model has lowest RMSE which is good.
* Forecast is overlapping with the actual observations keeping in mind the trend and seasonality.

**Ques 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**

**Answer:-**

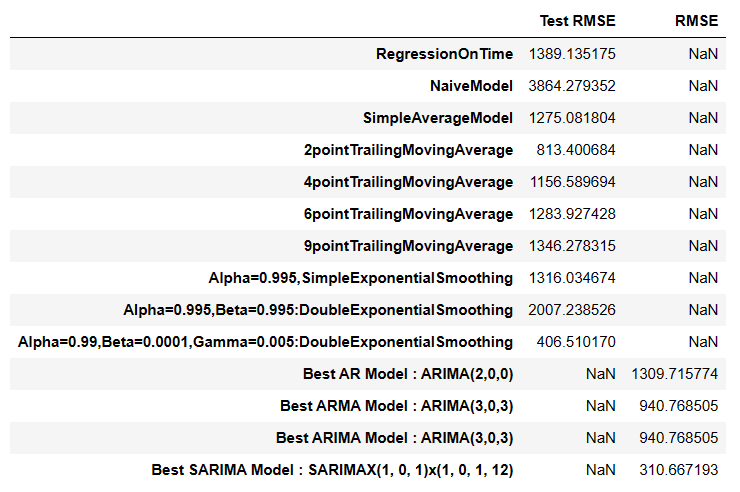
****

Table 24.Consolidated RMSE table for all models

# **Ques 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**

# **Answer:-**

Building the most optimum model on the Full Data

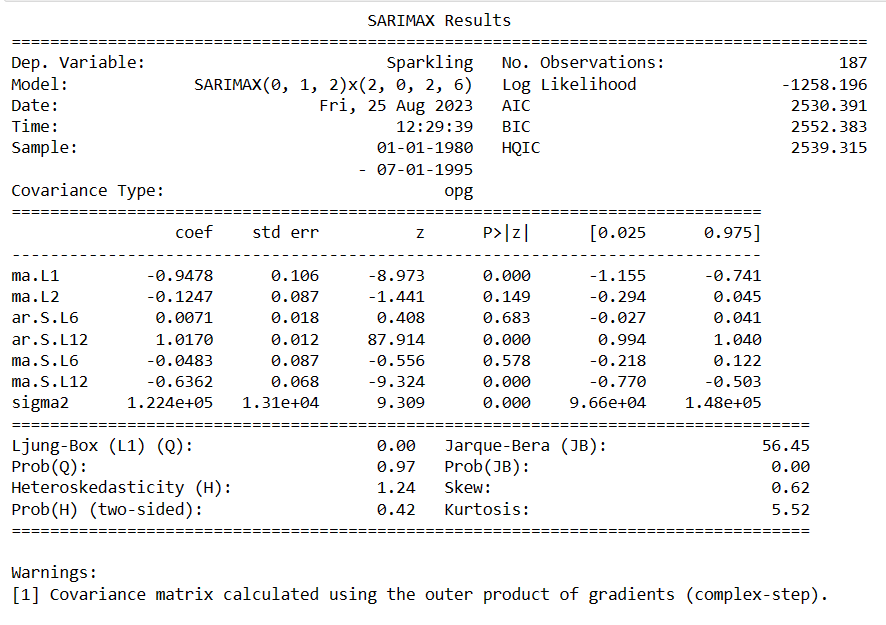


Table 25.Summary for Full data

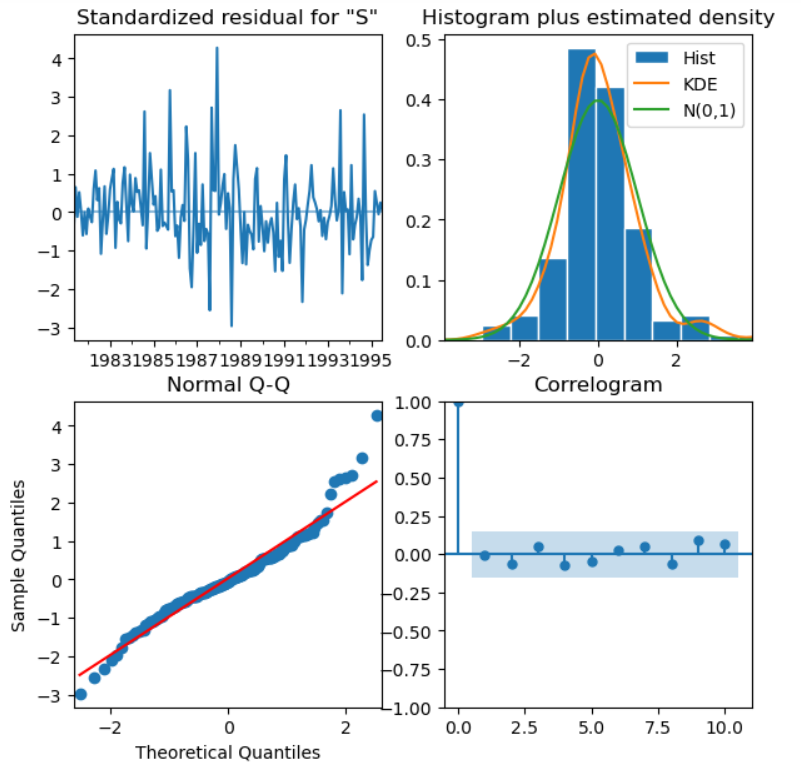


Fig 27.Correlation

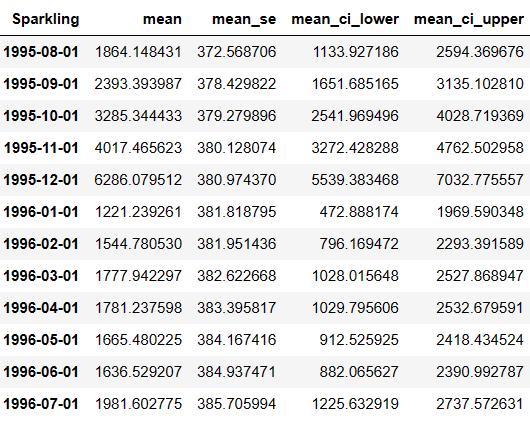
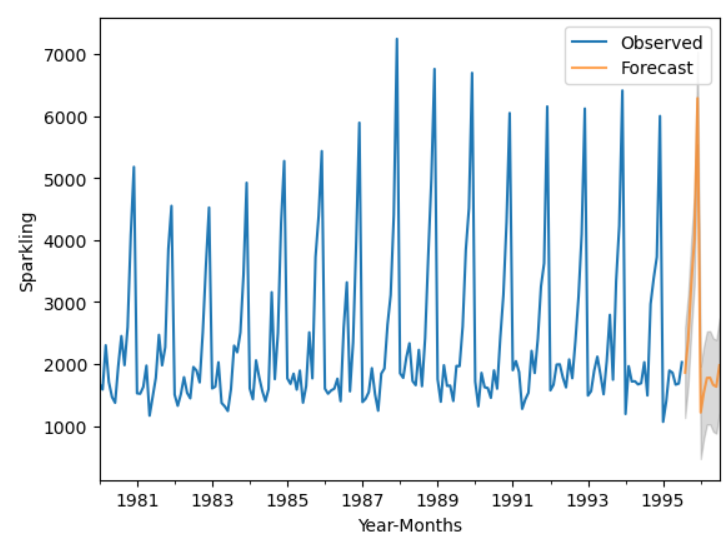
* **Standardized residuals plot**: The top left plot shows 1-step-ahead standardized residuals. If model is working correctly, then no pattern should be obvious in the residuals which is clearly not visible from the plot as well .
* **Histogram plus estimated density plot**: *T*his plot shows the distribution of the residuals. The orange line shows a smoothed version of this histogram, and the green line shows a normal distribution. If the model is good these two lines should be the same. Here there are small differences between them, which indicate that our model is doing just well enough.
* **Normal Q-Q plot**: The Q-Q plot compare the distribution of residuals to normal distribution. If the distribution of the residuals is normal, then all the points should lie along the red line, except for some values at the end, which is exactly happening in this case.
* **Correlogram plot**: The correlogram plot is the ACF plot of the residuals rather than the data. 95% of the correlations for lag >0 should not be significant (within the blue shades). If there is a significant correlation in the residuals, it means that there is information in the data that was not captured by the model, which is clearly not in this case
* Evaluating the model on the whole and predict 12 months into the future (till the end of next year)
* 
* 
* plot the forecast along with the confidence

Fig 28. Next 12 months prediction

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

# **Answer:-**

* We have loaded Sparkling.csv.
* We have checked for all the information regarding Dataset like datatypes,description.
* We have then performed EDA to check for any missing values and outliers.
* After EDA, we have split the data into train and test.
* Train is our sample data on which we will perform the algorithm.
* Test is our predicted data.
* All the data before 1991 was moved to train data and after or equal to 1991 was moved to test data.
* We have built different models and predicted the RMSE on test by Model evaluation.
  1. Linear Regression model
  2. Naïve Model
  3. Simple average model
  4. Moving Average (MA)
  5. Simple exponential model
  6. Double exponential Smoothing (Holt's Model)
  7. Triple Exponential Smoothing (Holt - Winter's Model)
* Built the automated ARIMA/SARIMA MODEL.
* We have also checked if the model is stationary or not using adfuller test.
* Required actions were taken to make sure the model is stationary as ours was found to be non stationary.
* We have predicted next 12 months data using our current and past data.
* Sparkling Wine sales was increased from the month of October to December.
* Out of all the model SARIMA model has performed best with lowest RMSE 310.
* From this we can say that wines sales increases rapidly in winter season.
* Suggestions for the company to be taken for future sales.
  + In-store discounts and Events can attract the customers to your store and can

increase sales.

* + Use of modern marketing techniques to reach more customers.
  + Reduction in prices.
  + Increase in quality.
  + Increase in advertising.
  + Advertise in local and rural areas.