TIME SERIES FORECASTING

BUSINESS REPORT

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

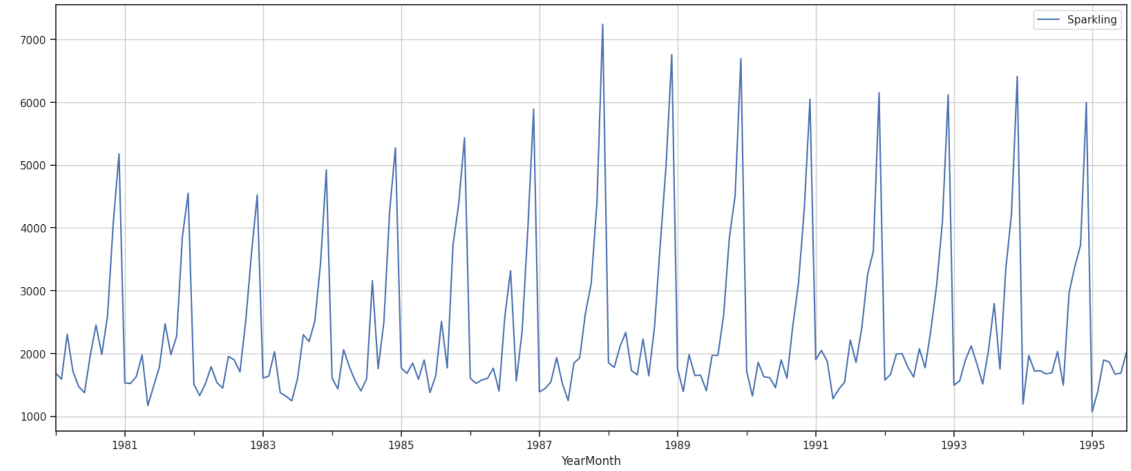
Data set for the Problem: Sparkling.csv and Rose.csv

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

Response:- first 5 rows of the data

	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

- The dataset has 187 rows.
- The shape of the dataset is (187,1)



We have seperated the YEAR and MONTH as 2 seperate columns, the new dataset looks like this

Sparkling	Year	Month

Sparkling	Year	Month	
YearMonth			
1980-01-01	1686	1980	1
1980-02-01	1591	1980	2
1980-03-01	2304	1980	3
1980-04-01	1712	1980	4
1980-05-01	1471	1980	5

The data has 3 int data types

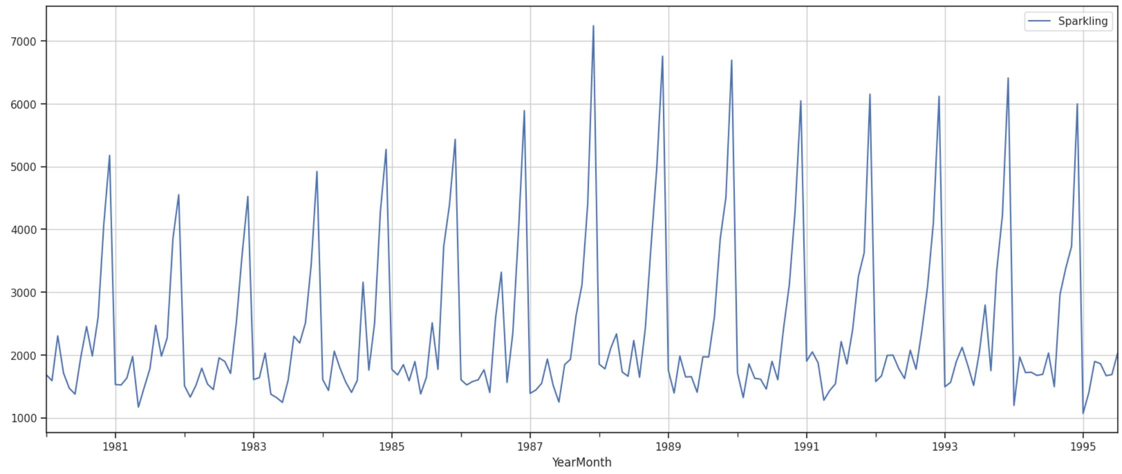
Sparkling	Year	Month	
count	187.000000	187.000000	187.000000
mean	2402.417112	1987.299465	6.406417
std	1295.111540	4.514749	3.450972
min	1070.000000	1980.000000	1.000000
25%	1605.000000	1983.000000	3.000000
50%	1874.000000	1987.000000	6.000000
75%	2549.000000	1991.000000	9.000000
max	7242.000000	1995.000000	12.000000

The above table gives the detailed description of the data

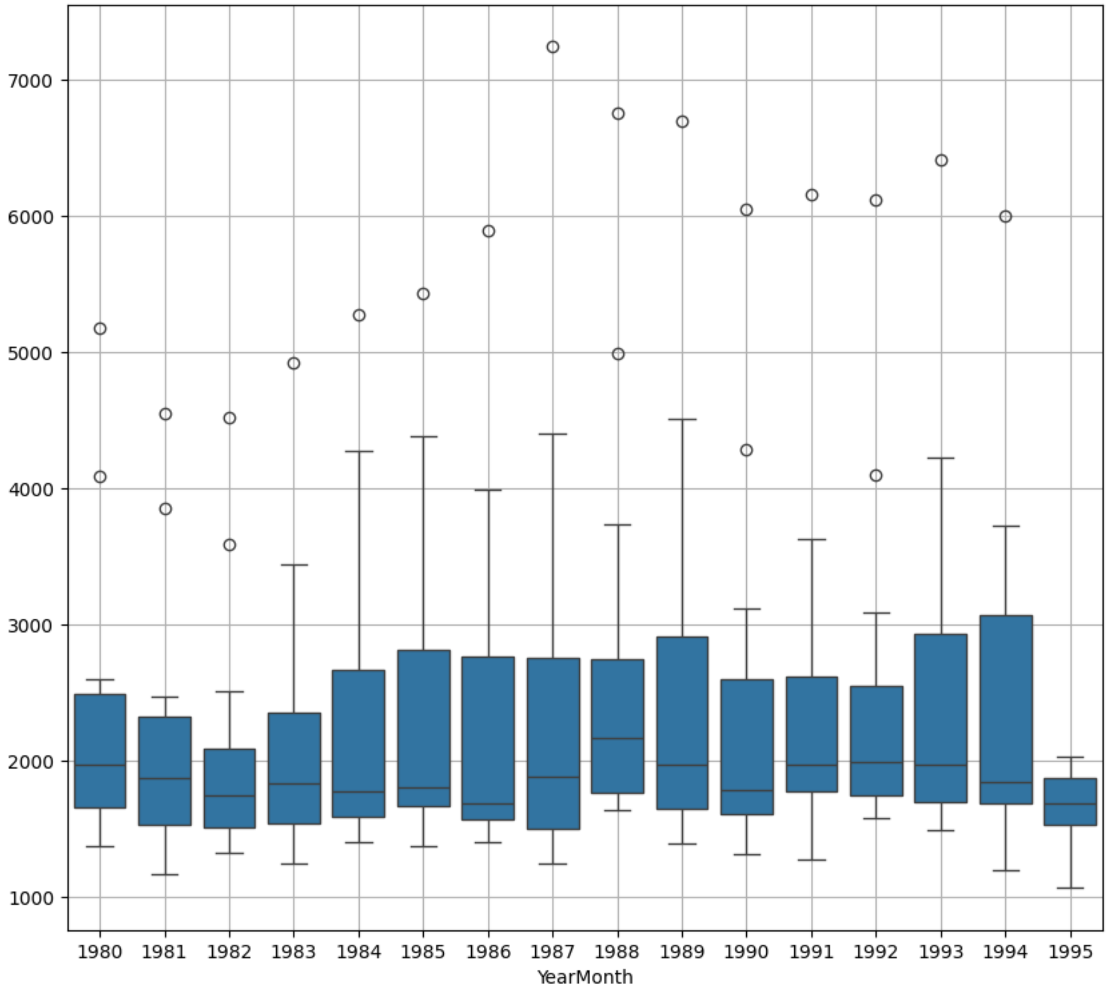
- The Sparkling, Year and Month has count =187
- The mean of the sparkling wine sales is 2402.417112
- The Min value is 1070
- The max value is 7242
- We can also see the mean of the data is almost near to 75%

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

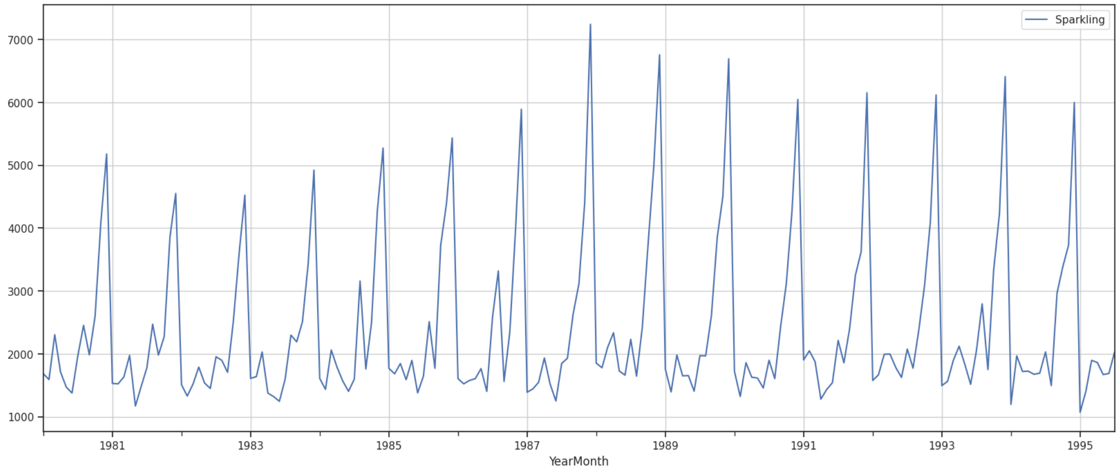
Response:-



from the graph we can see that there are outliers, but we dont have to treat them as they dont have significant effect on the results

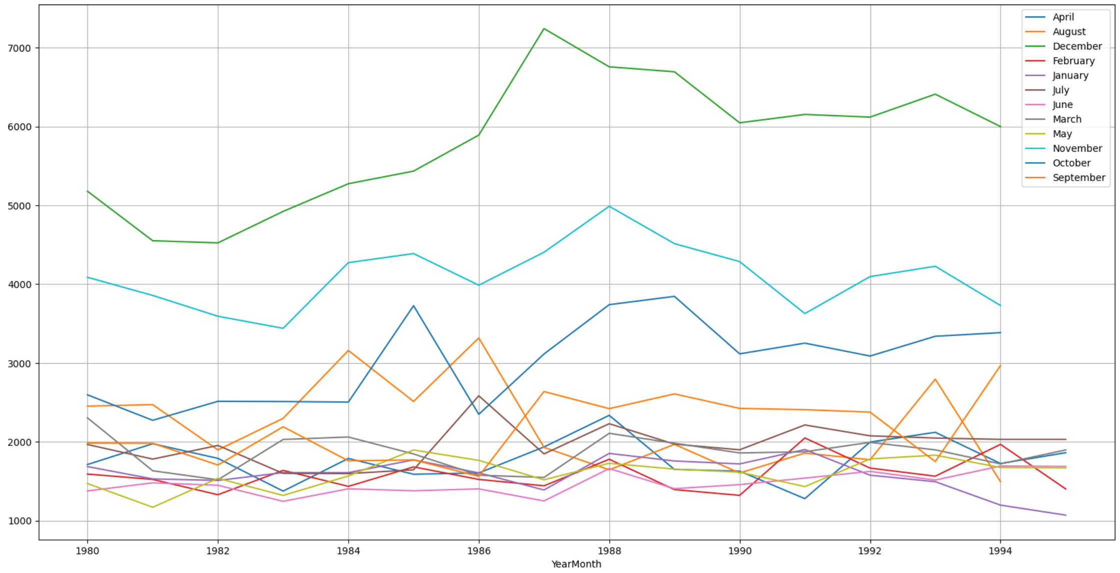


The lot shows the yearly sales of sparkling wine, The sales are almost consistent throughout the years 1980-1995, There is a peak in the sale during 1988 and 1999



The plot shows that sales are highest in the month of December and lowest in the month of January. Sales are consistent from January to July then from august the sales start to increase. Outliers are present in January, February and July

- Grapth of monthly sparkling across years.



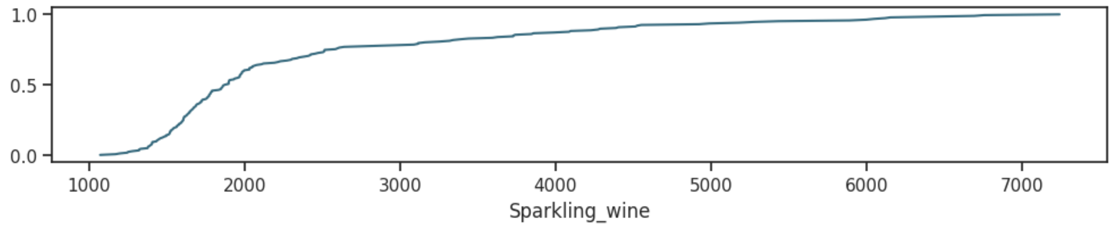
The image is a time series graph, which shows data points plotted over time. In this case, the x-axis is labeled "YearMonth" and appears to range from 1980 to 1994. The y-axis is labeled "Sales" and has tick marks from 1,000 to 7,000.

It appears that wine sales may have increased slightly over the period shown. There is some variability in sales from year to year, but the overall trend appears to be upwards.

It is important to note that this is just a small sample of data, and it is possible that the trend would not hold if we looked at a longer period of time.

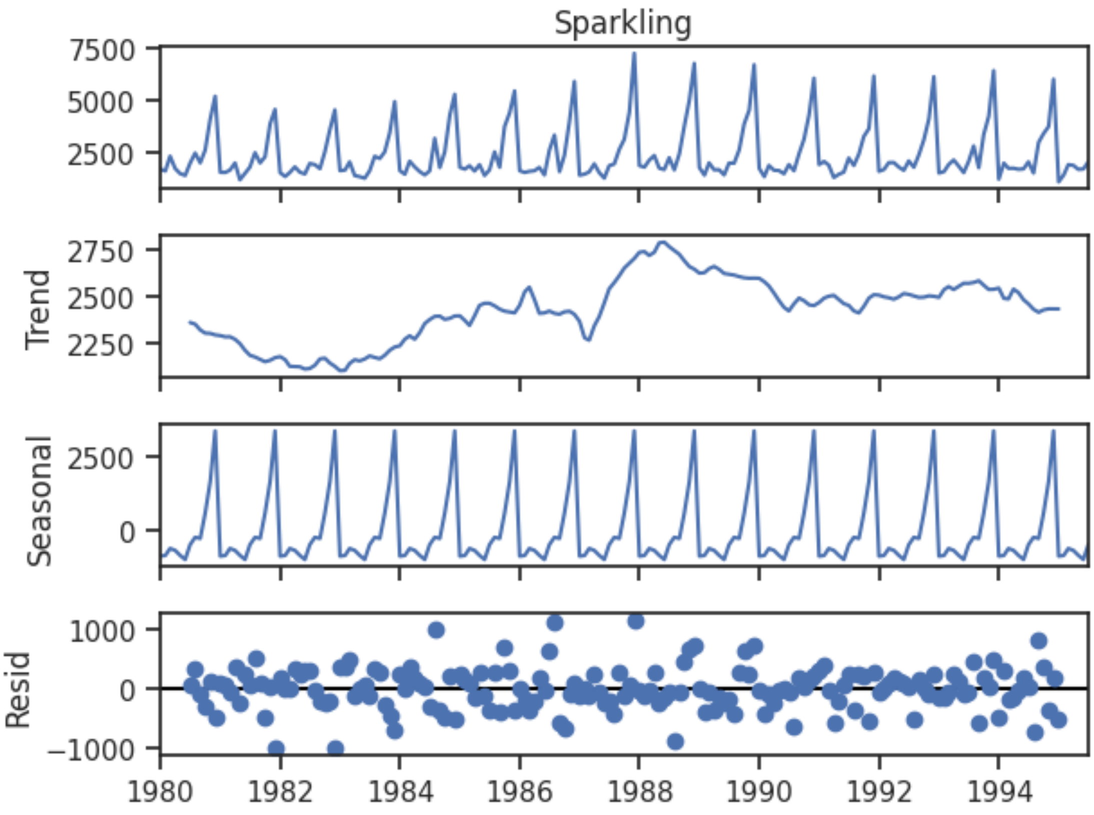
December has the highest sales over the yearsand the year 1988 wasthe year with highest number of wine sales.

- ECDF PLOT



The visualization depicts that over half of the sales fall below the 2000 mark. The highest recorded sales value is 7000. Approximately 80% of the sales fall below the threshold of 3000, indicating that the majority of sales are relatively modest, with only a small proportion exceeding this value.

- DECOMPOSITION



The graphs reveal that the highest point in sales occurred between 1988 and 1989. Subsequently, there has been a discernible decline in the trend over the following years. Additionally, the residual values are dispersed and do not follow a linear pattern, suggesting unpredictability or irregularity in the data. Both the long-term trend and seasonal patterns are evident, indicating fluctuations that occur regularly over time.

Question 3:-Split the data into training and test. The test data should start in 1991.

Response:-

Shape of datasets:

train dataset: (132, 3)

test dataset: (55, 3)

Rows of dataset:

First few rows of Training Data

Sparkling Year Month YearMonth

1980-01-01 1686 1980 1

1980-02-01 1591 1980 2

1980-03-01 2304 1980 3

1980-04-01 1712 1980 4

1980-05-01 1471 1980 5

Last few rows of Training Data

Sparkling Year Month YearMonth 1990-08-01 1605 1990 8

1990-09-01 2424 1990 9

1990-10-01 3116 1990 10

1990-11-01 4286 1990 11

1990-12-01 6047 1990 12

First few rows of Test Data

Sparkling Year Month YearMonth 1991-01-01 1902 1991 1

1991-02-01 2049 1991 2

1991-03-01 1874 1991 3

1991-04-01 1279 1991 4

1991-05-01 1432 1991 5

Last few rows of Test Data

Sparkling Year Month YearMonth 1995-03-01 1897 1995 3

1995-04-01 1862 1995 4

1995-05-01 1670 1995 5

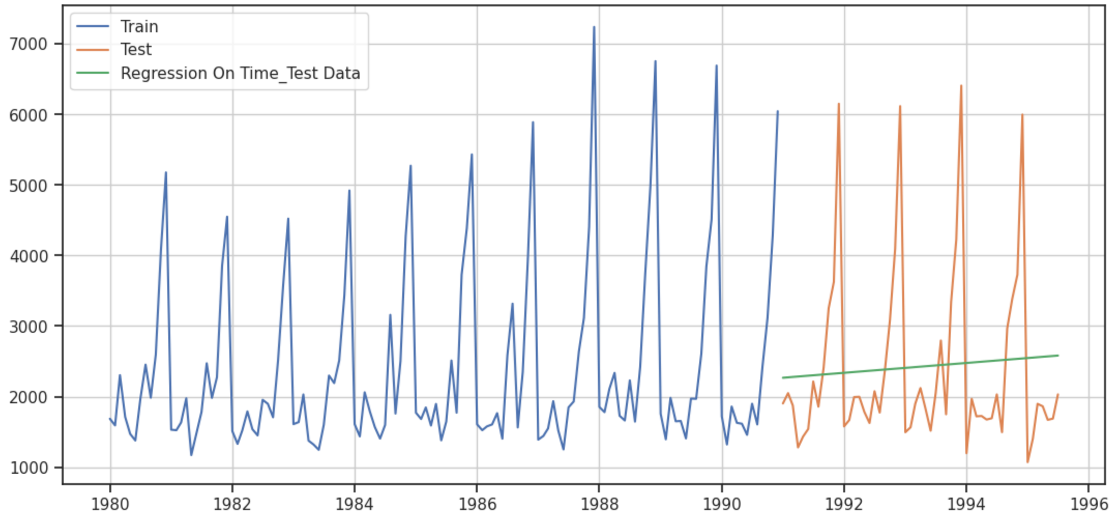
1995-06-01 1688 1995 6

1995-07-01 2031 1995 7

Question 4:- Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

Response:-

#Linear Regression

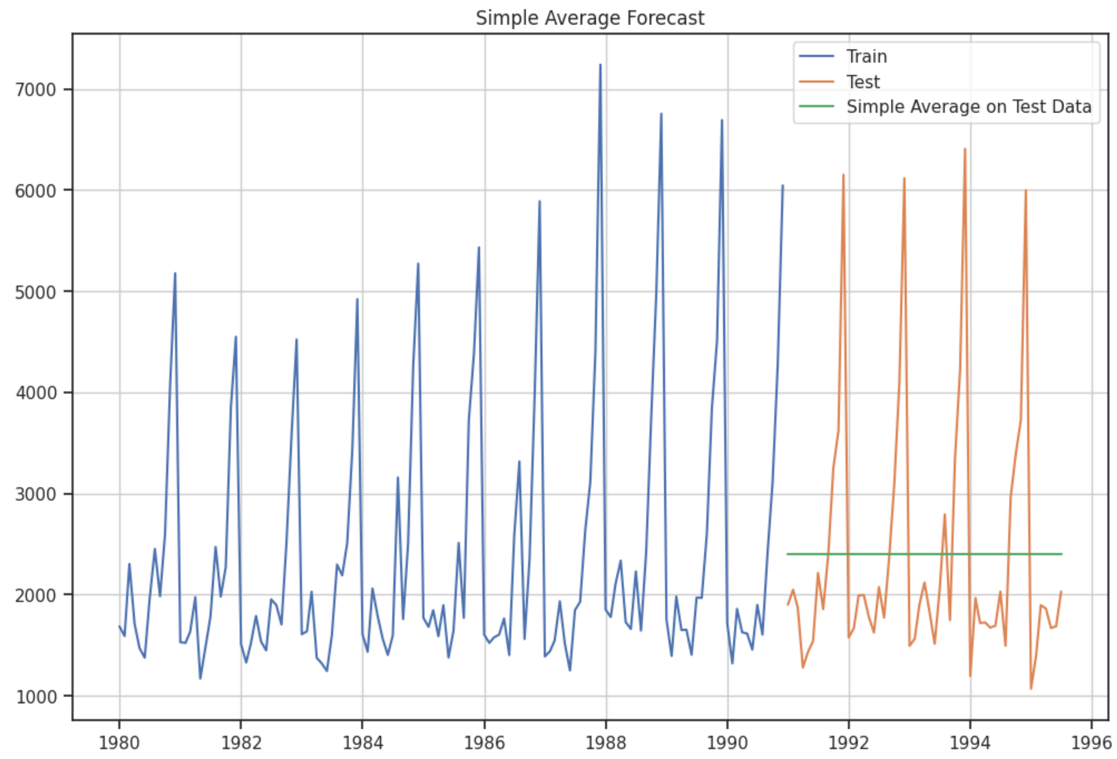


The green line represents the predictions generated by the model, while the orange line represents the actual test values. Upon observation, it's evident that there's a substantial discrepancy between the predicted values and the actual values. The predicted values, denoted by the green line, diverge significantly from the orange line, which represents the ground truth.

To quantify this disparity, the model's performance was assessed using the Root Mean Square Error (RMSE) metric. For this specific model, which employs Linear Regression, the calculated RMSE value is 1275.867052.

The RMSE essentially measures the average magnitude of the errors between predicted and actual values. In this context, an RMSE of 1275.867052 suggests that, on average, the model's predictions deviate from the true values by approximately 1275.867052 units. Such a high RMSE value indicates that the model's predictive accuracy is relatively poor, as it is failing to accurately capture the underlying patterns in the data.

#Simple Average



The green line on the graph represents the predictions generated by the model, while the orange line represents the actual values from the test dataset. Upon visual inspection, it's evident that the predicted values deviate significantly from the actual values, indicating a lack of accuracy in the model's predictions.

To quantify the model's performance, the Root Mean Square Error (RMSE) metric was utilized. RMSE measures the average magnitude of the errors between predicted and actual values, with lower values indicating better performance.

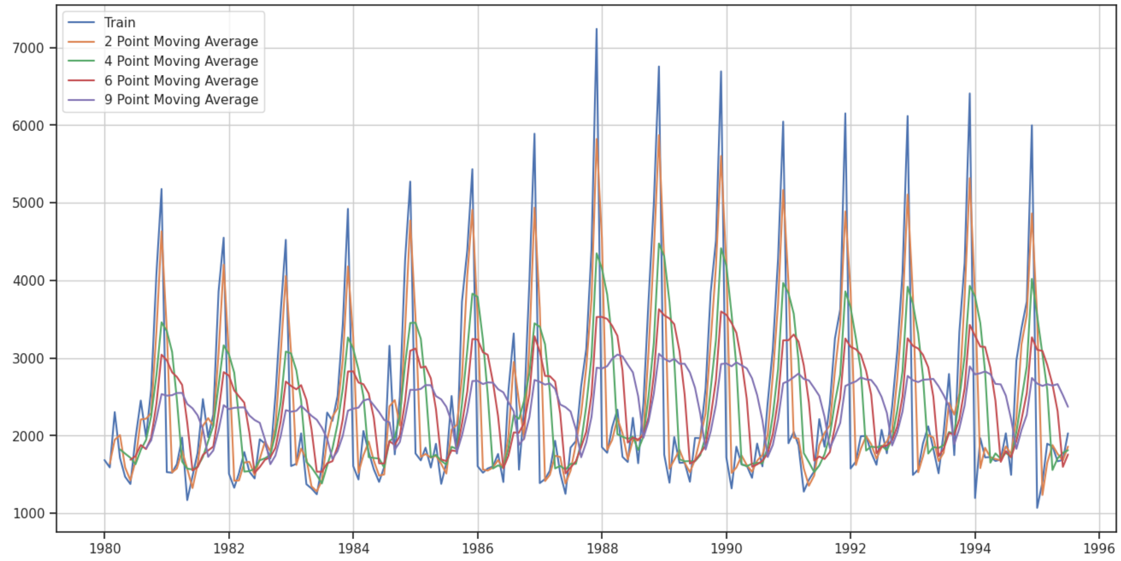
The RMSE values for two different models were calculated and compared:

1. Simple Average Model: The RMSE for this model is 1275.081804. This model likely takes the average of all the training data and uses it as the prediction for every instance in the test set. Despite its simplicity, it serves as a baseline for comparison.

2. Linear Regression Model: The RMSE for this model is slightly higher at 1275.867052. Linear regression attempts to fit a straight line to the data, predicting values based on the linear relationship between features and the target variable. However, despite its more complex approach, it still fails to make accurate predictions as indicated by the relatively high RMSE.

In summary, both models exhibit poor performance, with the linear regression model not significantly outperforming the simple average model. This suggests that the models are not adequately capturing the underlying patterns in the data, resulting in inaccurate predictions.

#Moving Average

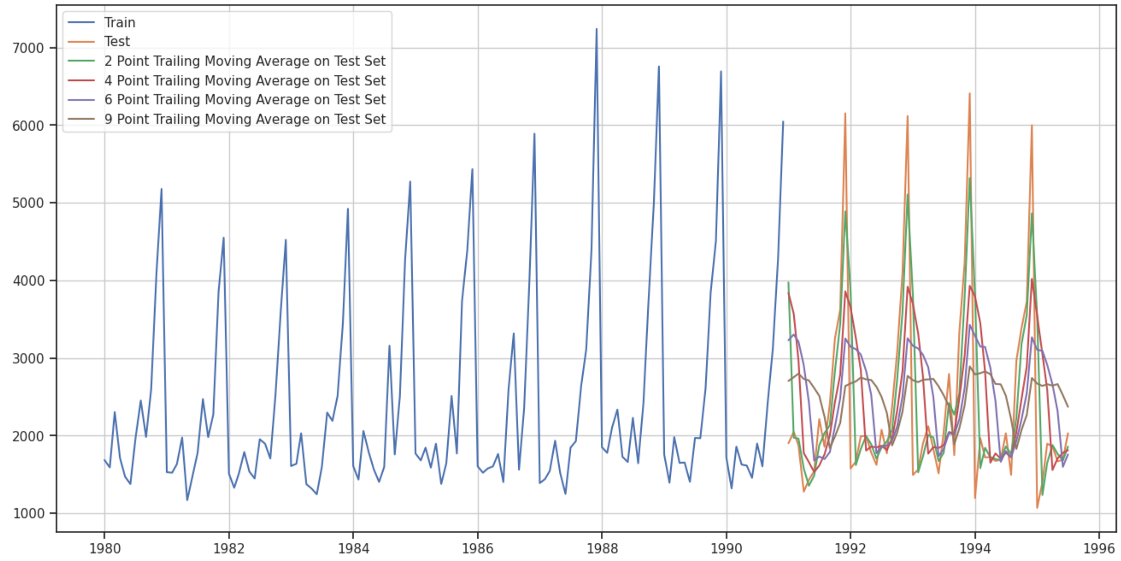


In our analysis, we constructed multiple moving average models employing rolling windows ranging from 2 to 9.

Moving average models offer an improvement over simple averaging techniques. Unlike simple averages that consider all available data equally, moving averages utilize a rolling window, which focuses solely on the preceding n observations to predict the next value, where n represents the size of the rolling window. This approach enables the model to adapt to recent trends in the data, providing more accurate predictions.

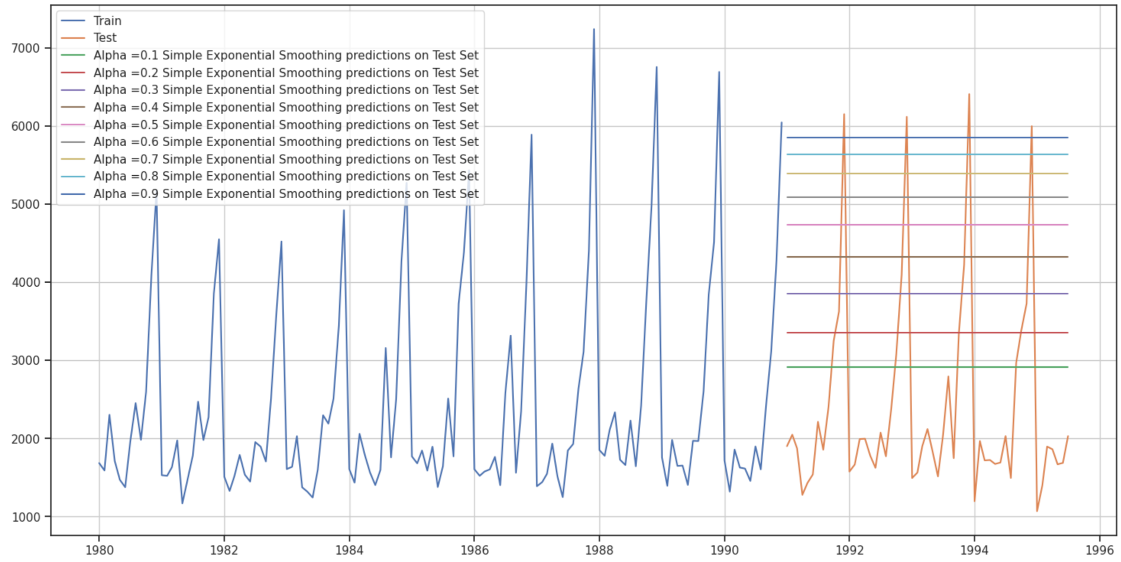
The choice of the rolling window size is crucial. A higher rolling window incorporates more historical data into the prediction, resulting in a smoother curve and potentially capturing long-term trends. Conversely, smaller rolling windows prioritize recent data, making the model more responsive to short-term fluctuations.

In summary, the use of moving average models with varying rolling window sizes allows us to capture different aspects of the data's behavior. By adjusting the rolling window, we can strike a balance between capturing short-term fluctuations and long-term trends, thereby improving the accuracy of our predictions compared to simple averaging methods.



The above is the plot on training and test set

#Simple Exponential Smoothing



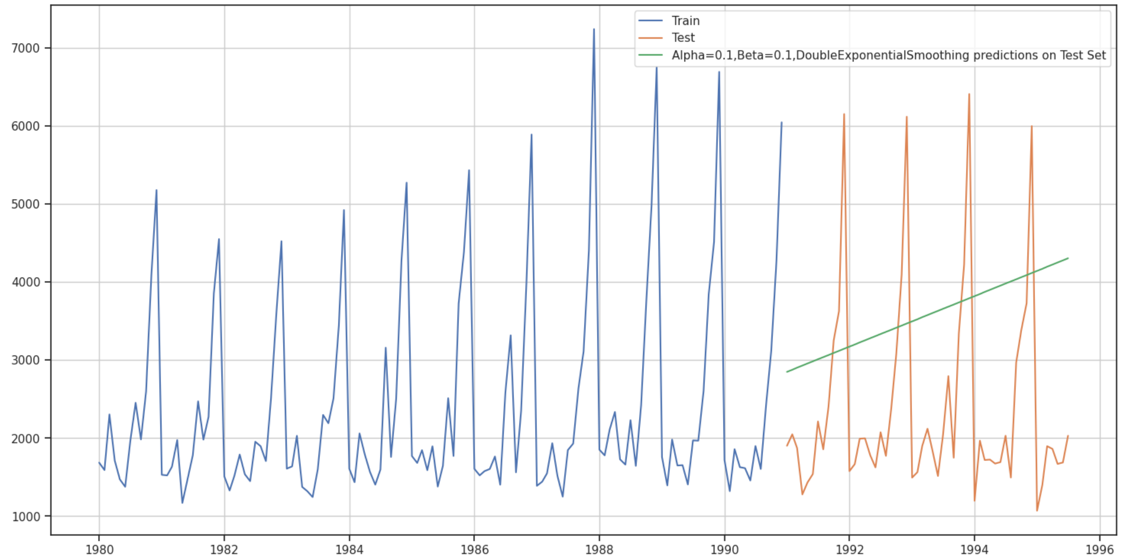
Alpha ValuesTrain RMSETest RMSE00.11333.8738361375.39339810.21356.0429871595.20683920.31359.5117471935.50713230.41352.5888792311.91961540.51344.0043692666.35141350.61338.8053812979.20438860.71338.8443083249.94409270.81344.4620913483.80100680.91355.7235183686.794285

The model's performance was assessed using the Root Mean Square Error (RMSE) metric across different alpha values. Alpha values represent the smoothing factor in exponential smoothing models, where smaller values assign more weight to recent observations, and larger values assign more equal weight to all observations.

Here's the report of the data:

Alpha Value	Test RMSE
0.1	1375.393398
0.2	1595.206839
0.3	1935.507132
0.4	2311.919615
0.5	2666.351413
0.6	2979.204388
0.7	3249.944092
0.8	3483.801006

From the results, it's apparent that as the alpha value increases, the RMSE also increases, indicating worsening performance. Lower alpha values (0.1 to 0.3) yield relatively lower RMSE values, suggesting better predictive accuracy. However, as the alpha value exceeds 0.3, the RMSE starts to increase more significantly, indicating a decrease in prediction accuracy. This suggests that a smaller alpha value, such as 0.1 or 0.2, might be more suitable for this dataset in terms of minimizing prediction errors.



Question 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.

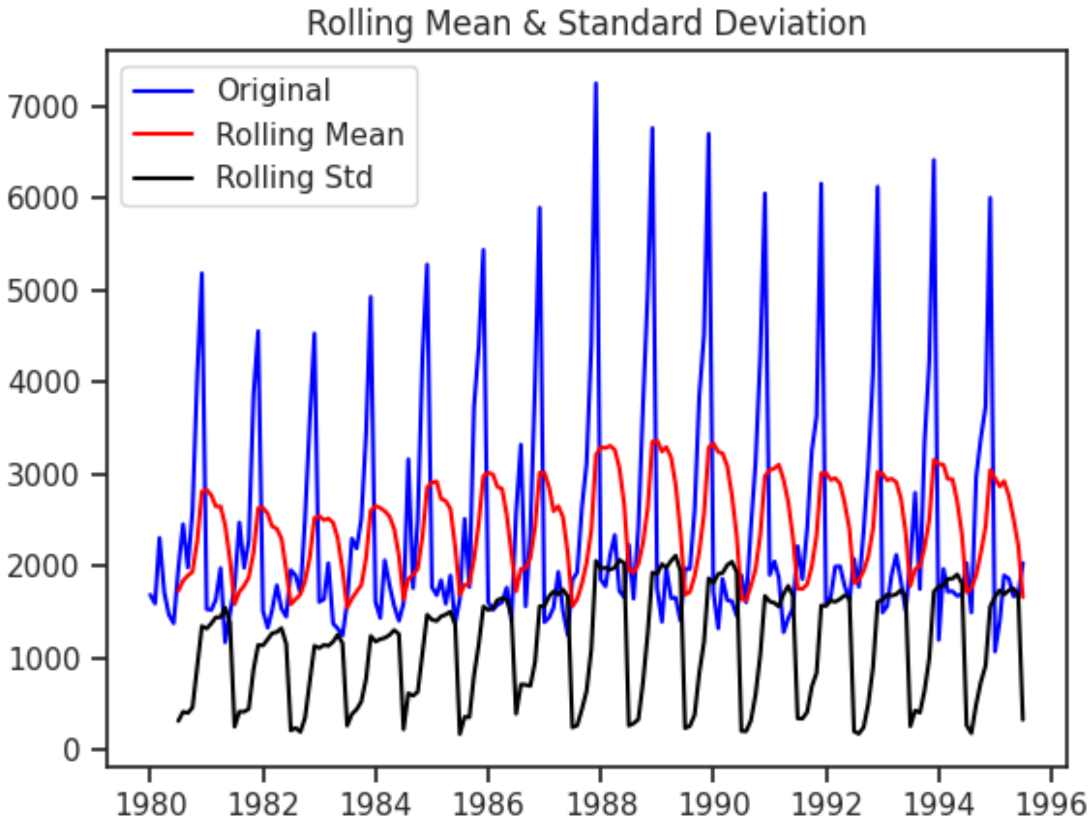
Response:-

The Augmented Dickey-Fuller test is employed to determine the presence of a unit root in a time series, which in turn indicates whether the series is non-stationary.

In its simplest form, the hypothesis for the ADF test can be stated as follows:

- Null Hypothesis (H0): The time series possesses a unit root and is therefore non-stationary.
- Alternative Hypothesis (H1): The time series does not possess a unit root and is therefore stationary.

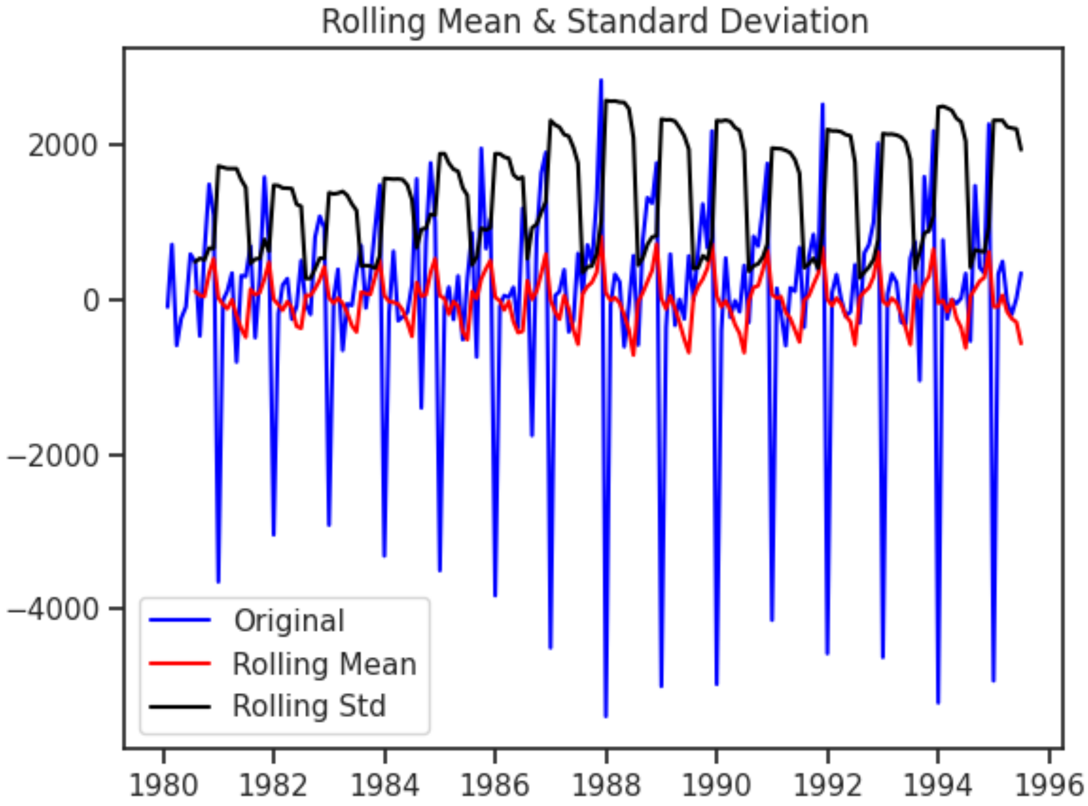
For the purpose of building ARIMA models, it's essential for the series to be stationary. Consequently, we aim for the p-value of the ADF test to be less than the chosen significance level, denoted as α.



The results of the Dickey-Fuller Test indicate a p-value of 0.601061. This suggests that the time series data is likely non-stationary, meaning it exhibits trends or seasonality that may affect its behavior over time.

To address the non-stationarity of the series, the differencing approach was employed. This involved using the .diff() function on the existing series with the default argument of 1, indicating differencing of order 1. This process calculates the difference between consecutive observations, effectively removing trends or seasonality present in the data.

Additionally, since differencing of order 1 generates the first value as NaN, these NaN values were dropped from the dataset. This step ensures that the differenced series remains continuous and ready for further analysis.



The results of the Dickey-Fuller Test yielded a p-value of 0.000000, significantly lower than the conventional significance level of 0.05. Consequently, the null hypothesis, which posited that the series is not stationary at difference = 1, was rejected. This rejection implies that the series indeed became stationary after the differencing process was applied.

Rejecting the null hypothesis indicates strong evidence that the time series data exhibits stationarity following the differencing. The null hypothesis is typically rejected when the p-value is less than the chosen significance level, in this case, 0.05.

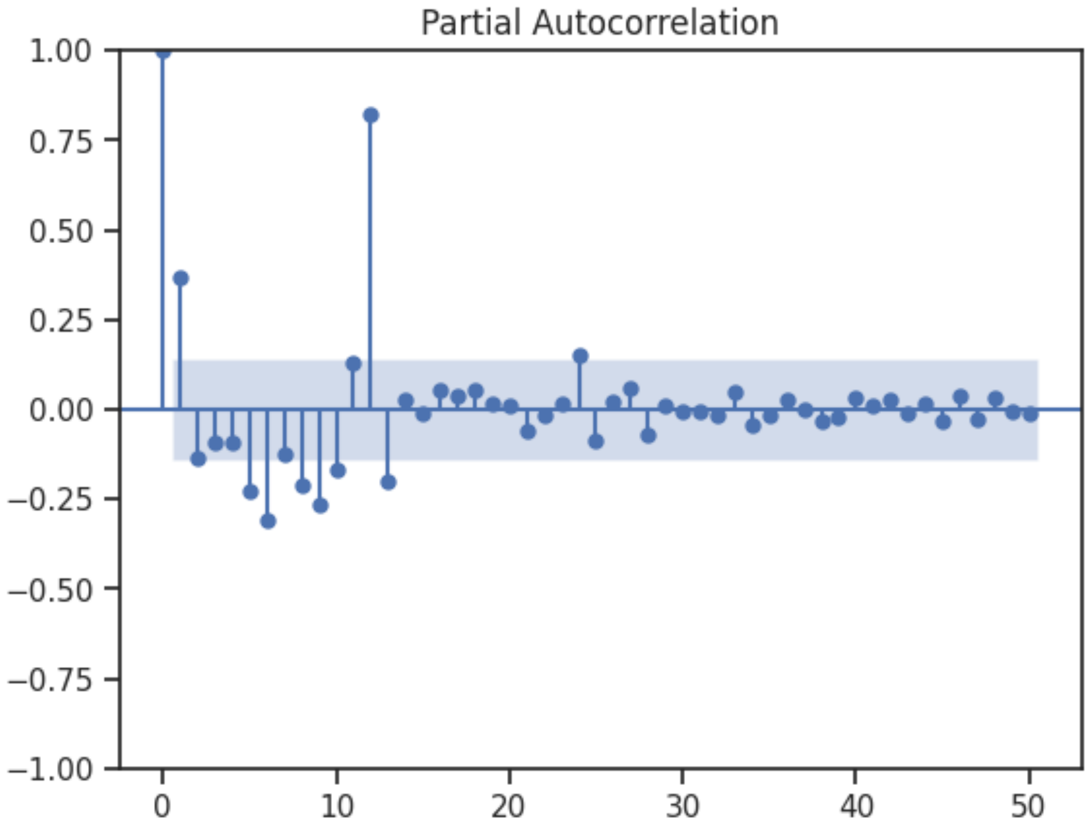
Moreover, visual inspection of the rolling mean plot revealed a straight line, which further supports the notion of stationarity. Additionally, observing the series from both directions showed similar patterns, indicating stationarity regardless of the direction of observation.

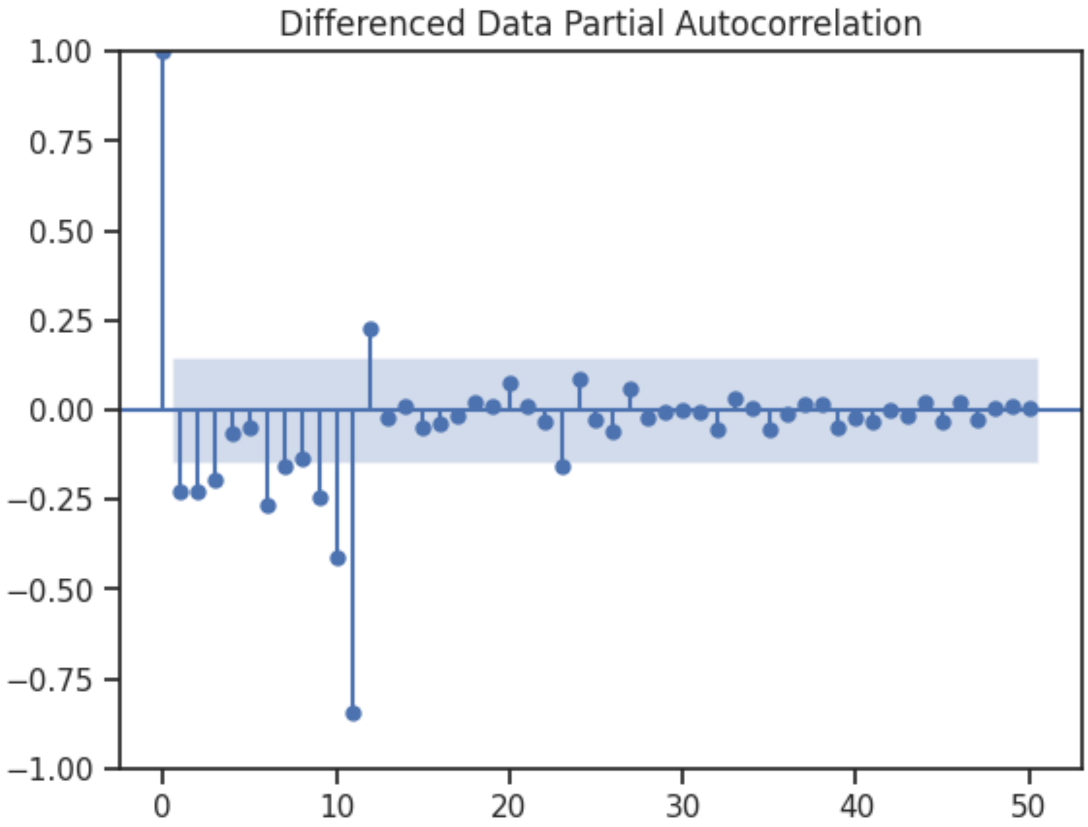
In summary, the Dickey-Fuller Test, coupled with visual analysis, confirms that the differencing approach successfully rendered the series stationary. This transformation enhances the reliability of subsequent analyses and modeling efforts, as stationarity simplifies the data's behavior and allows for more accurate predictions.

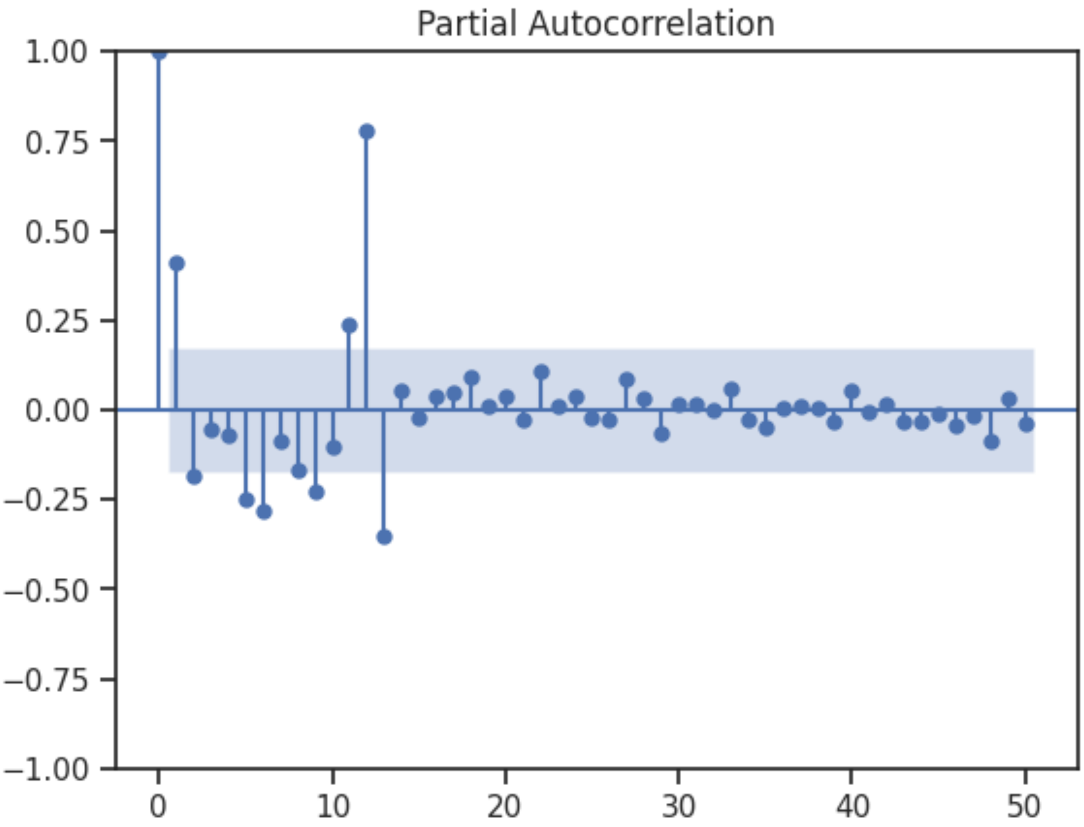
Question 6:-Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

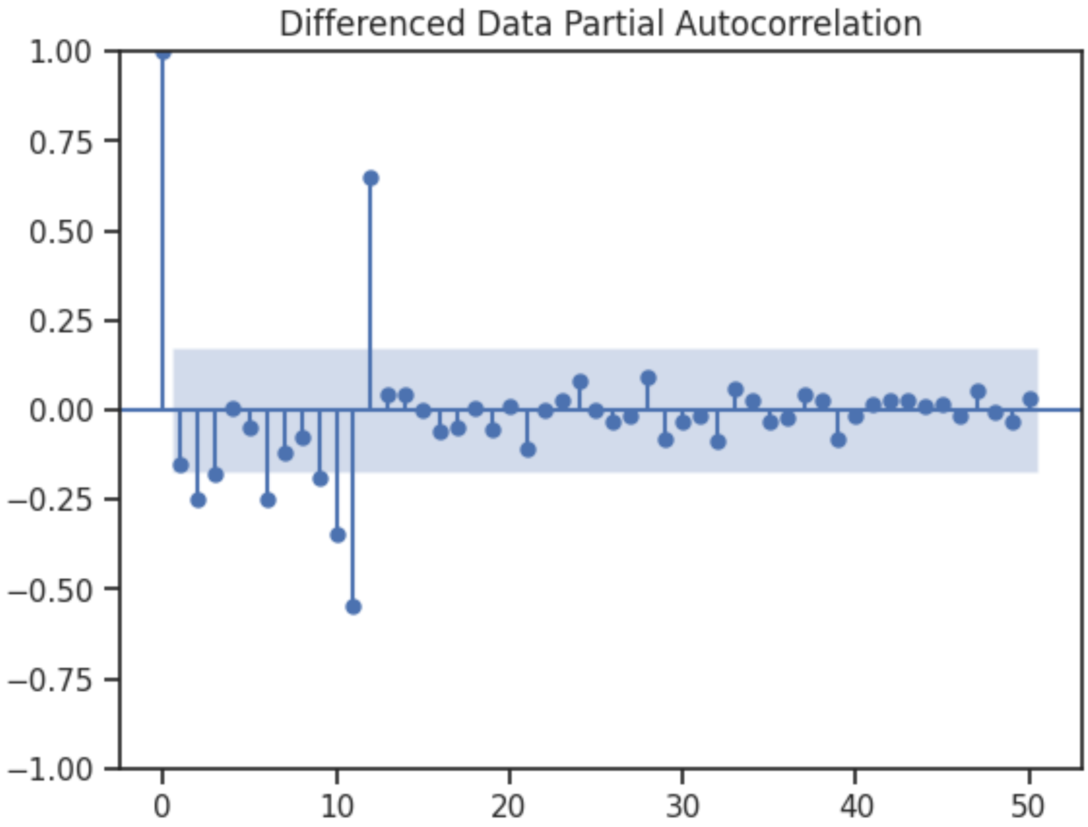
Response:-

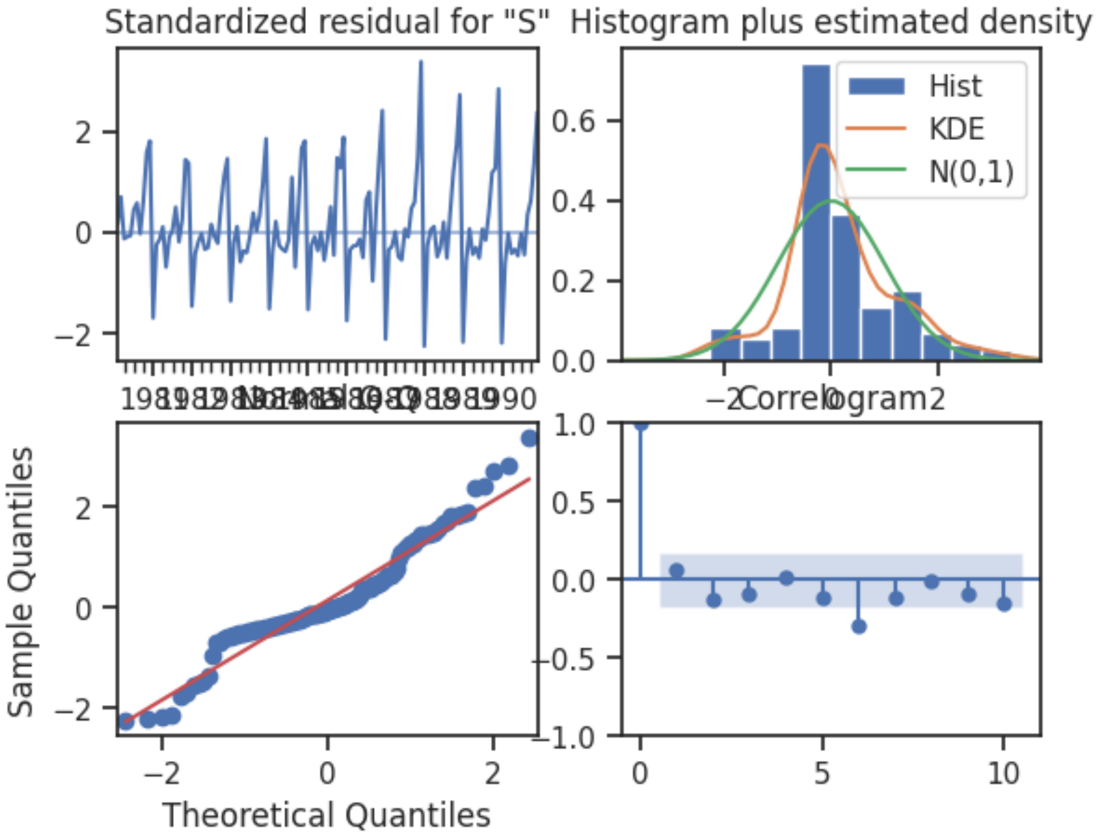
```
 SARIMAX Results
==============================================================================
======== Dep. Variable: Sparkling No. Observations: 132 Model: ARIMA(2, 1, 2) Log
Likelihood -1101.755 Date: Sun, 25 Feb 2024 AIC 2213.509 Time: 22:03:07 BIC 2227.885
Sample: 01-01-1980 HQIC 2219.351 - 12-01-1990 Covariance Type: opg
==============================================================================
======== coef std err z P>|z| [0.025 0.975] -----------------------------------------------------------
------------------ ar.L1 1.3121 0.046 28.782 0.000 1.223 1.401 ar.L2 -0.5593 0.072 -7.740
0.000 -0.701 -0.418 ma.L1 -1.9917 0.109 -18.215 0.000 -2.206 -1.777 ma.L2 0.9999 0.110
9.108 0.000 0.785 1.215 sigma2 1.099e+06 2e-07 5.51e+12 0.000 1.1e+06 1.1e+06
==============================================================================
============= Ljung-Box (L1) (Q): 0.19 Jarque-Bera (JB): 14.46 Prob(Q): 0.67 Prob(JB):
0.00 Heteroskedasticity (H): 2.43 Skew: 0.61 Prob(H) (two-sided): 0.00 Kurtosis: 4.08
==============================================================================War
nings: [1] Covariance matrix calculated using the outer product of gradients (complex-
step). [2] Covariance matrix is singular or near-singular, with condition number 3.27e+27.
Standard errors may be unstable.
```




Differenced Data Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50


Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50


Differenced Data Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50



Question 7:-Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Response:-

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612454
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
Simple Average Model	1275.081804
Linear Regression	1275.867052
6pointTrailingMovingAverage	1283.927428
Auto_ARIMA	1299.979749

	Test RMSE
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1304.927405
ARIMA(3,1,3)	1319.936734
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670

Question 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.

Response:-

Sparkling_Predictions

1995-08-01. 1988.782193

1995-09-01 2652.762887

1995-10-01 3483.872246

1995-11-01 4354.989747

1995-12-01 6900.103171

1996-01-01 1546.800546

1996-02-01 1981.361768

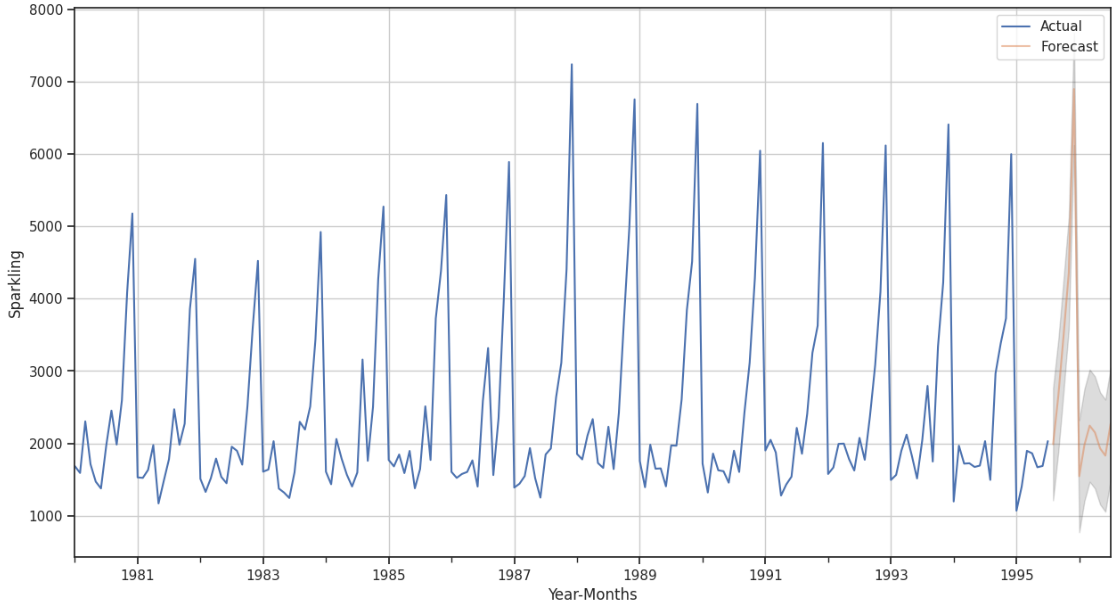
1996-03-01 2245.459724

1996-04-01 2151.066942

1996-05-01 1929.3558151

996-06-01 1830.619260

1996-07-01 2272.156151



After assessing the performance of all the models we constructed, it is evident that the triple exponential smoothing or the Holt-Winters model consistently yields the lowest Root Mean Square Error (RMSE). Consequently, this model emerges as the most optimal choice for making sales predictions.

Therefore, the sales forecasts generated by this preferred model are deemed to be the most reliable and accurate among all the models considered.

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

Response:-

1. **Sales Projection for Sparkling Wine:** The model forecasts that the sales of Sparkling wine for the upcoming year will at least match those of the previous year, if not surpass them. Moreover, there's a possibility of peak sales for the next year exceeding the current year's figures.

2. **Consistent Popularity of Sparkling Wine:** Despite experiencing a slight decrease in sales, Sparkling wine remains a consistently favored choice among customers. Its popularity, which peaked in the late 1980s, has sustained over time.

3. **Impact of Seasonality:** Seasonal fluctuations notably influence the sales of Sparkling wine. Sales tend to be sluggish during the first half of the year, gaining momentum from August through December.

4. **Recommendation for Marketing Strategies:** It's advised for the company to implement marketing campaigns during the initial months of the year, especially

from March to July when sales typically dip. By capitalizing on this slower period, the company can potentially stimulate sales and maintain market engagement.

5. **Promotional Pairing Strategy:** An effective approach could involve pairing Sparkling wine with another less popular wine, such as "Rose wine," as part of a special promotional offer. This tactic aims to incentivize customers to explore the less popular wine while also boosting sales of Sparkling wine.

In summary, the analysis suggests that the company should focus on leveraging seasonal trends, implementing targeted marketing campaigns during slow sales periods, and exploring promotional pairings to stimulate sales growth and maintain market competitiveness.

Business Report for "Rose.csv"

Problem 2

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

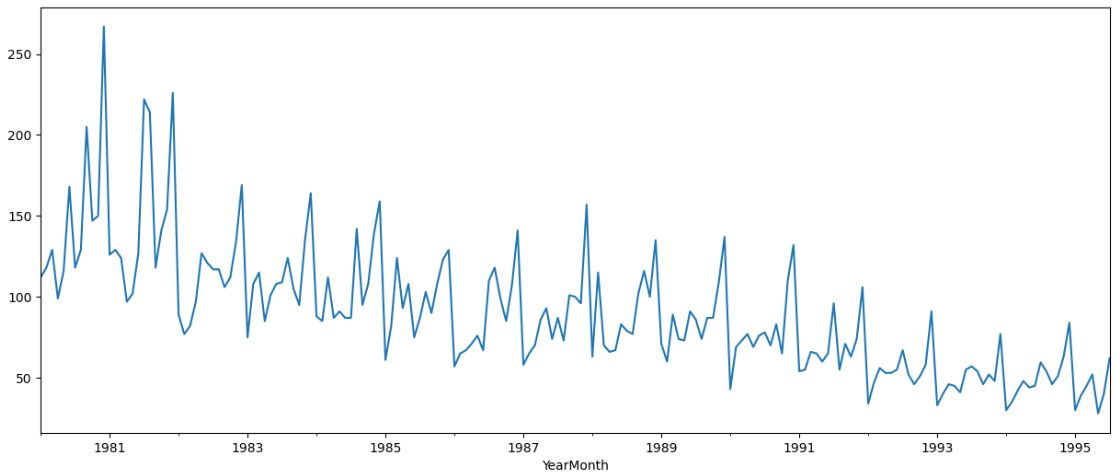
Data set for the Problem: [Rose.csv]

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

   Response:-

   - The dataset has 187 rows and 3 columns
   - The data has 2 int datatype and 1 float datatype
   - first 5 rows of the data

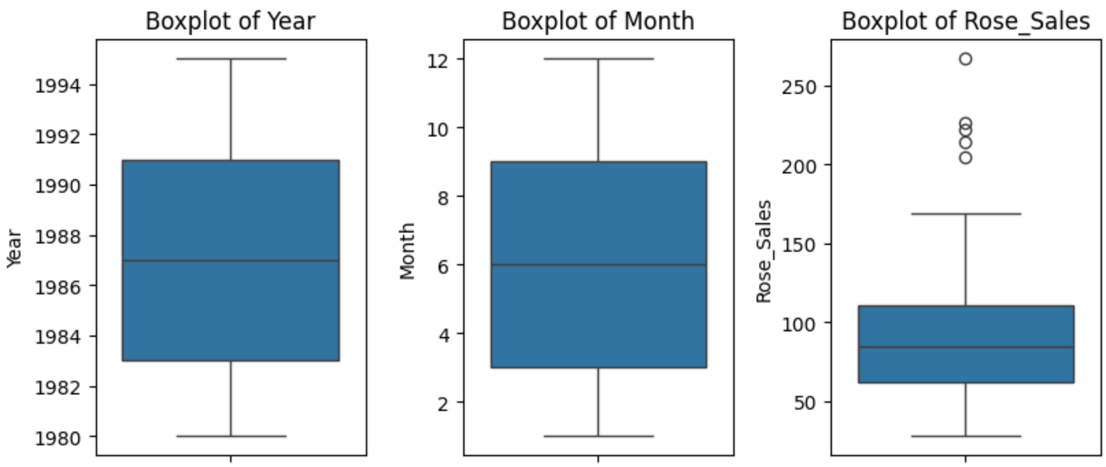
Rose	Year	Month	
YearMonth			
1980-01-01	112.0	1980	1
1980-02-01	118.0	1980	2
1980-03-01	129.0	1980	3
1980-04-01	99.0	1980	4
1980-05-01	116.0	1980	5



	Rose_Sales	Year	Month
count	185.000000	187.000000	187.000000
mean	90.394595	1987.299465	6.406417
std	39.175344	4.514749	3.450972
min	28.000000	1980.000000	1.000000
25%	63.000000	1983.000000	3.000000
50%	86.000000	1987.000000	6.000000
75%	112.000000	1991.000000	9.000000
max	267.000000	1995.000000	12.000000

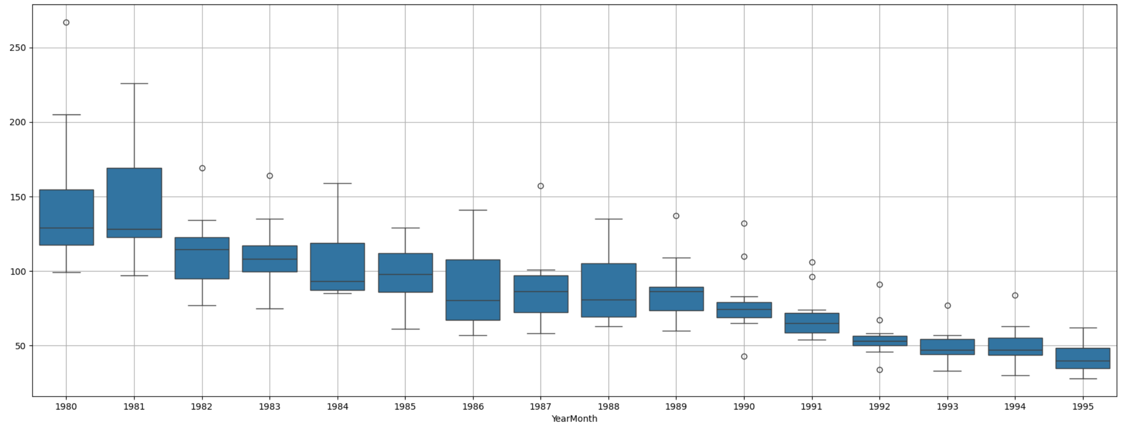
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Response:-

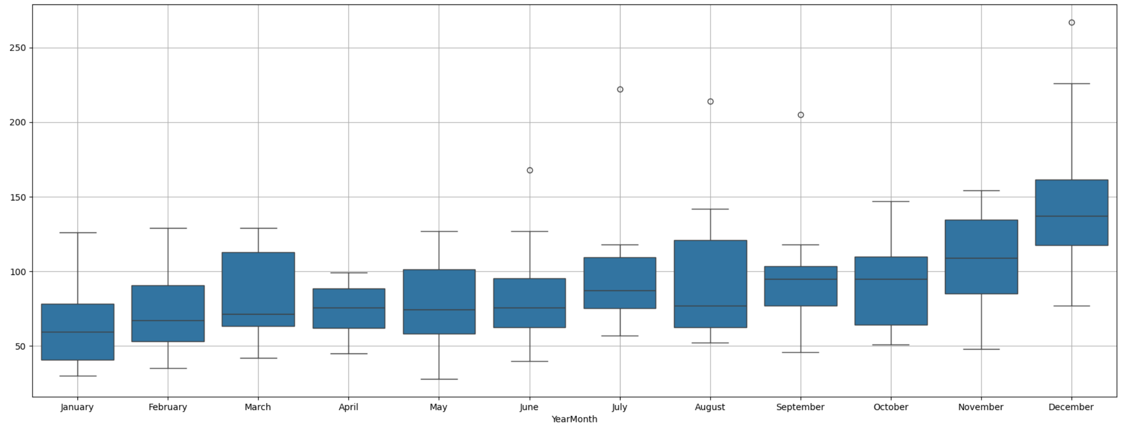


The box plot shows:

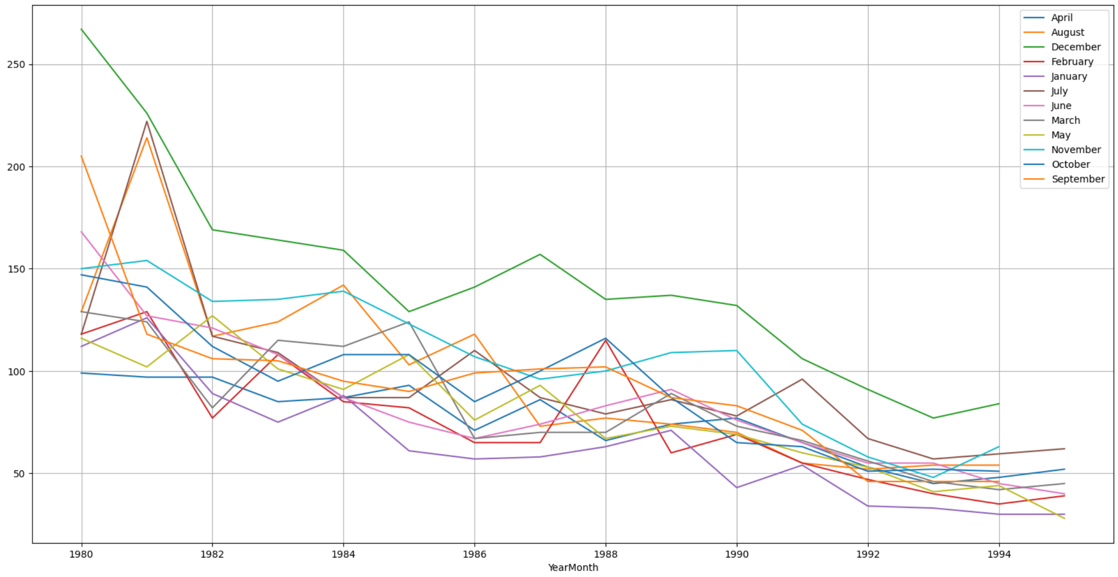
- Sales boxplot has outliers we can treat them but we are choosing not to treat them as they do not give much effect on the time series model.

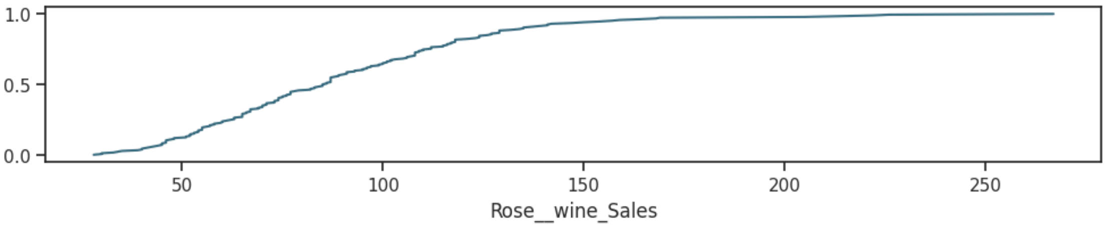


This yearly box plot shows there is consistency over the years and there was a peak in 1980-1981. Outliers are present in almost all years.



The plot shows that sales are highest in the month of December and lowest in the month of January. Sales are consistent from January to July then from august the sales start to increase. Outliers are present in June, July, august, September and December.

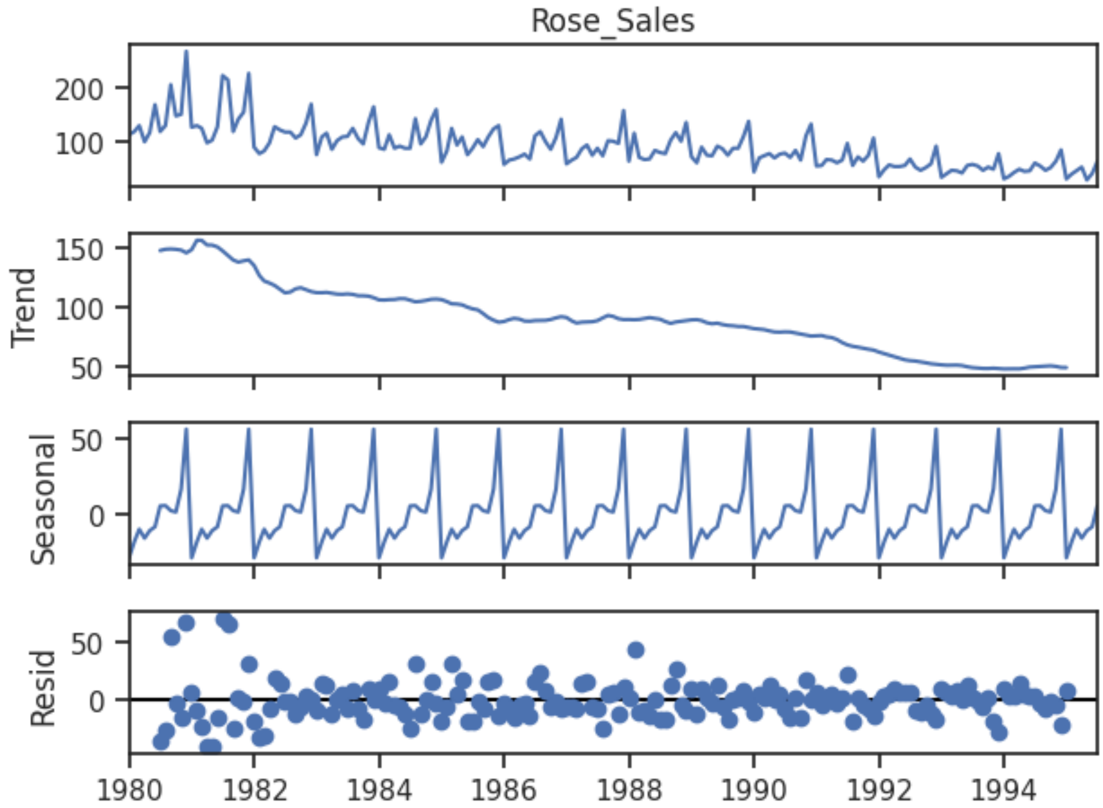




This plot shows:

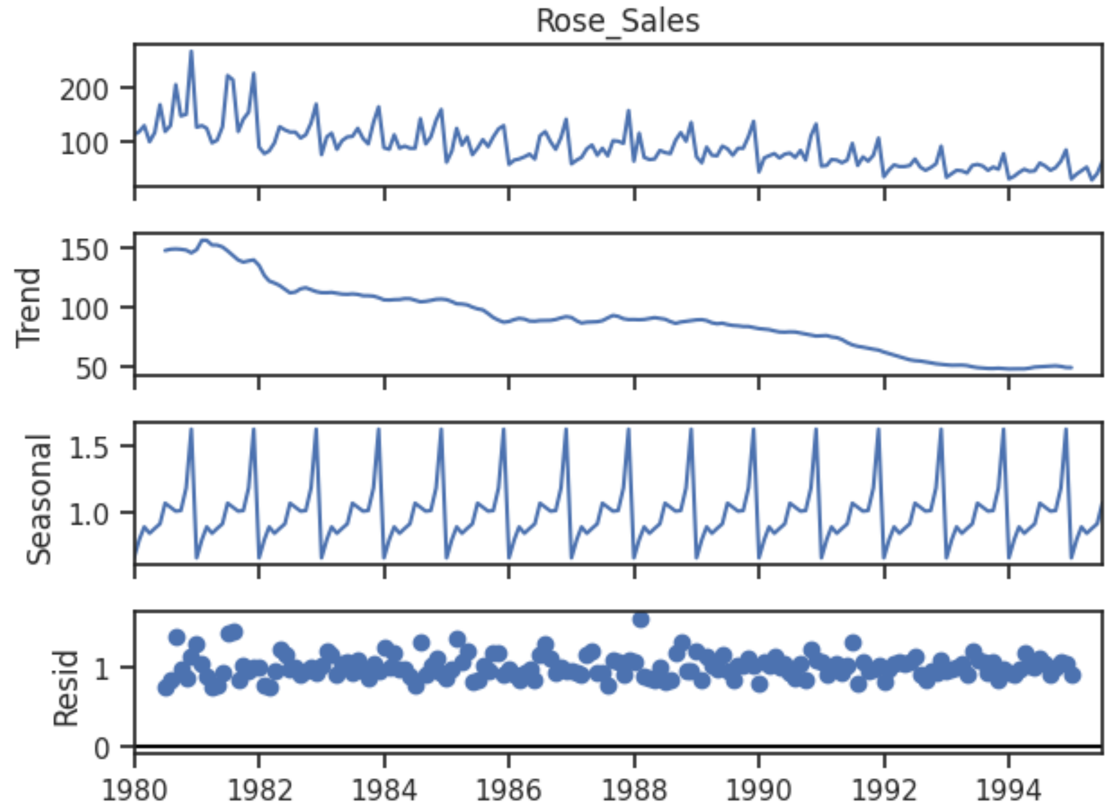
50% sales has been less 100

Highest vales is 250
Aprox 90% sales has been less than 150



The plots reveal the following observations:

- The peak year is identified as 1981.
- There is a noticeable decline in trend post-1981.
- The residual plot exhibits dispersion and lacks linearity.
- Both trend and seasonality components are discernible from the data.



The plots depict the following observations:

- The peak year is identified as 1981.
- A discernible trend decline is evident post-1981.
- The residuals exhibit dispersion but appear to follow an approximate straight line.
- Both trend and seasonality components are observable in the data.
- Residual values range from 0 to 1 for the multiplicative model, contrasting with the range of 0 to 50 for the additive model.
- Consequently, the multiplicative model is favored due to its more stable residual plot and narrower range of residuals.

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

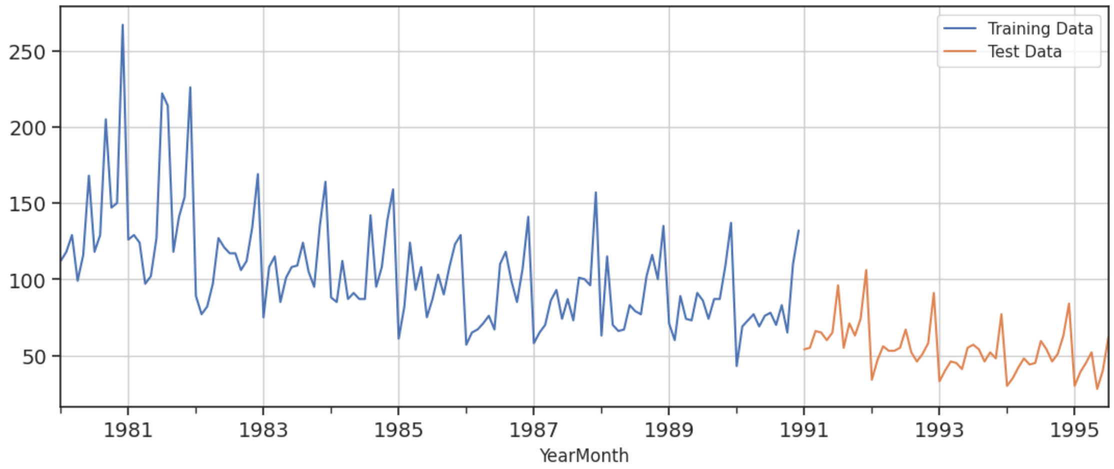
Response:-

Shape of datasets:

train dataset:(132, 3)

```
 test dataset: (55, 3)

 Rows of dataset:

 First few rows of Training Data

 Year Month Rose_Sales YearMonth

1980-01-01 1980 1 112.0

1980-02-01 1980 2 118.0

1980-03-01 1980 3 129.0

1980-04-01 1980 4 99.0

1980-05-01 1980 5 116.0

Last few rows of Training Data

 Year Month Rose_Sales

YearMonth

 1990-08-01 1990 8 70.0

1990-09-01 1990 9 83.0

1990-10-01 1990 10 65.0

1990-11-01 1990 11 110.0

1990-12-01 1990 12 132.0

First few rows of Test Data

 Year Month Rose_Sales

YearMonth

1991-01-01 1991 1 54.0

1991-02-01 1991 2 55.0

1991-03-01 1991 3 66.0

1991-04-01 1991 4 65.0

1991-05-01 1991 5 60.0

 Last few rows of Test Data

 Year Month Rose_Sales YearMonth
```

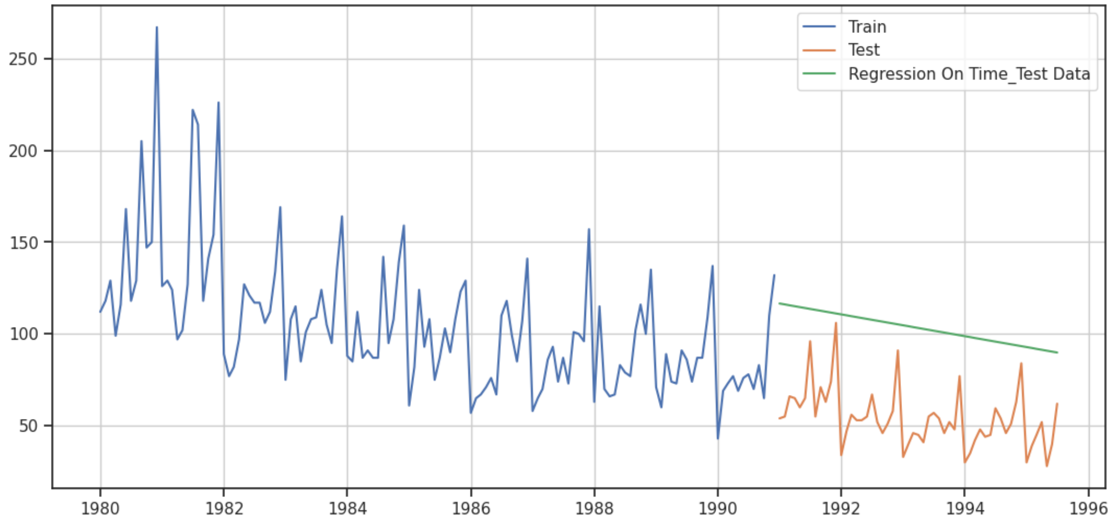
```
1995-03-01 1995 3 45.0
1995-04-01 1995 4 52.0
1995-05-01 1995 5 28.0
1995-06-01 1995 6 40.0
1995-07-01 1995 7 62.0
```



Question 4:- Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

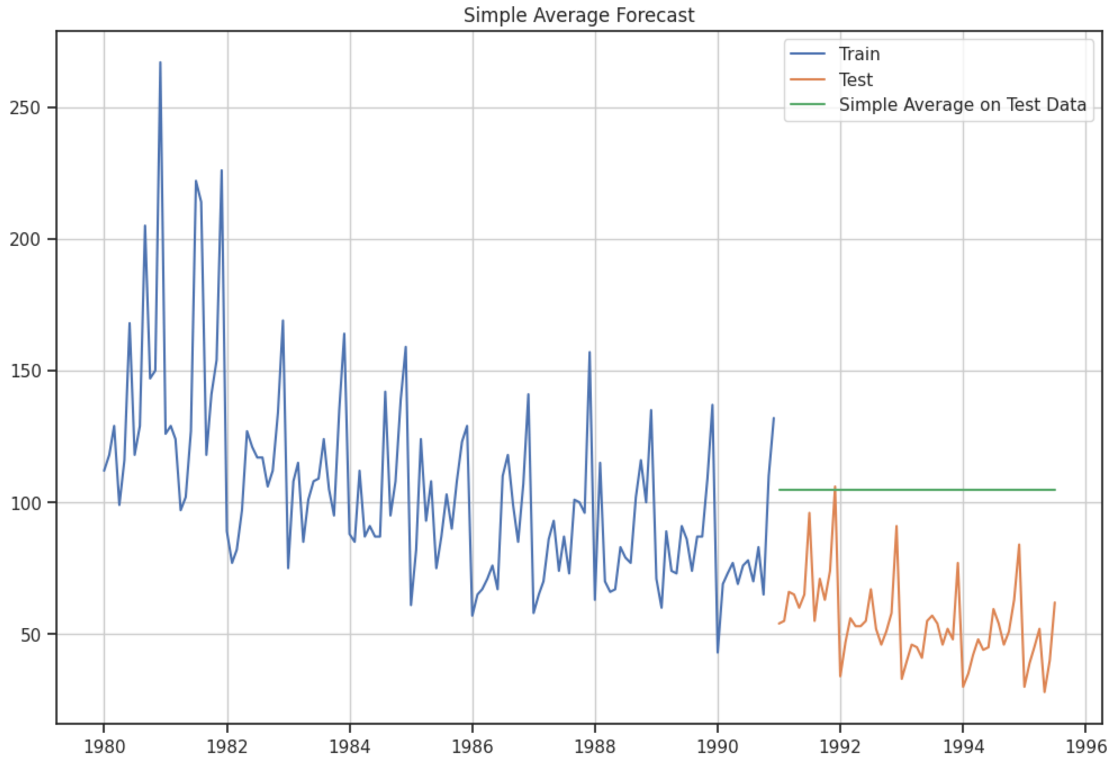
Response:-

Model 1: Linear Regression



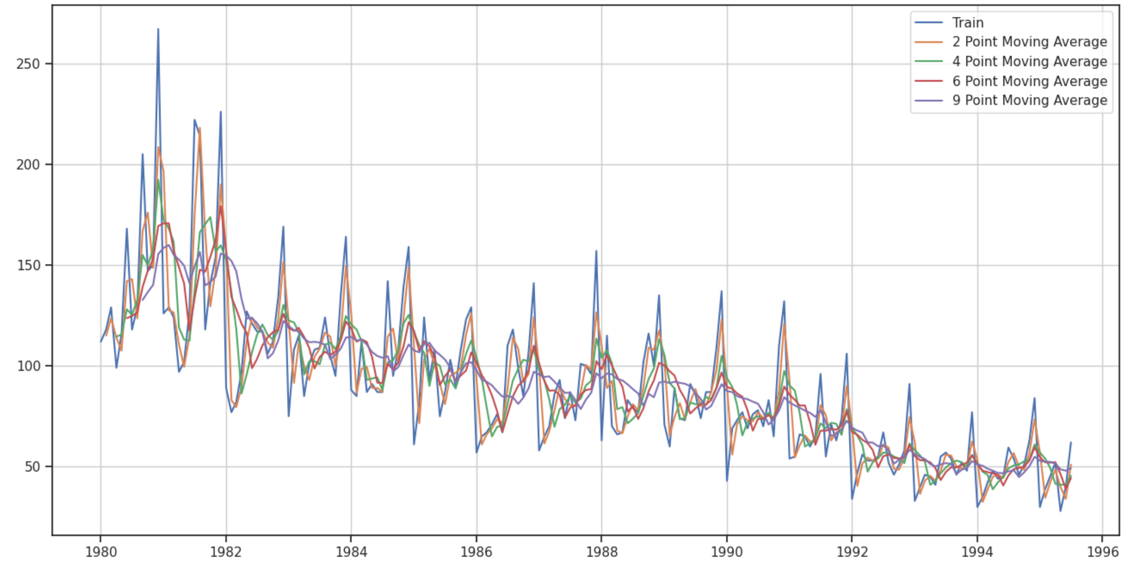
The model's performance was assessed using the Root Mean Square Error (RMSE) metric. The calculated RMSE for the Linear Regression model is 51.080941. This indicates that the predicted values deviate considerably from the actual values, suggesting a lack of accuracy in the model's predictions.

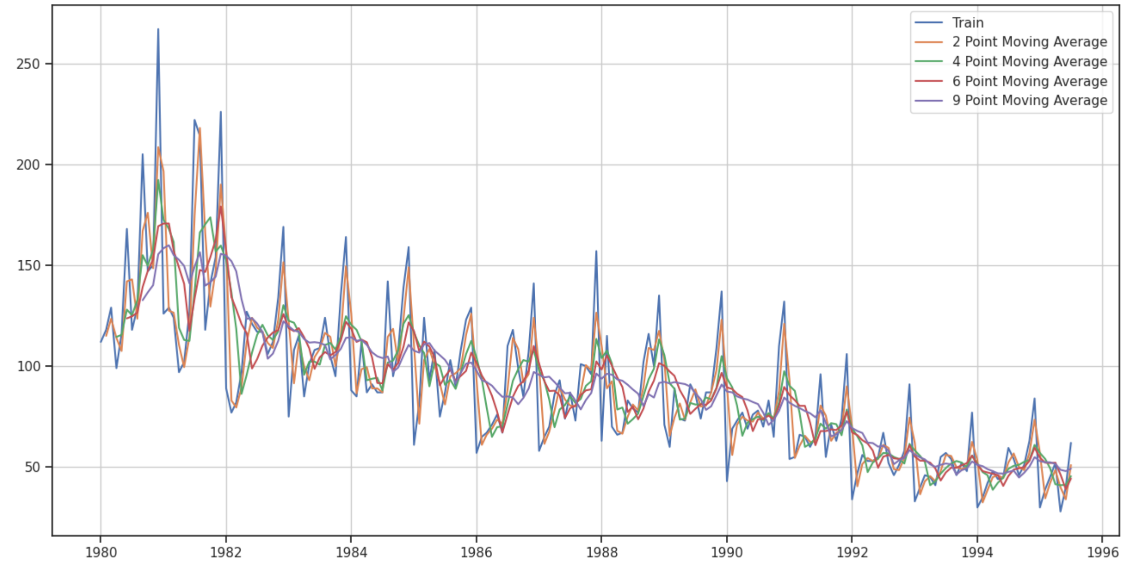
Simple average



The model's performance was assessed using the Root Mean Square Error (RMSE) metric. The RMSE calculated for the Simple Average Model is 53.049755. This indicates that the predicted values deviate significantly from the actual values, suggesting a lack of accuracy in the model's predictions.

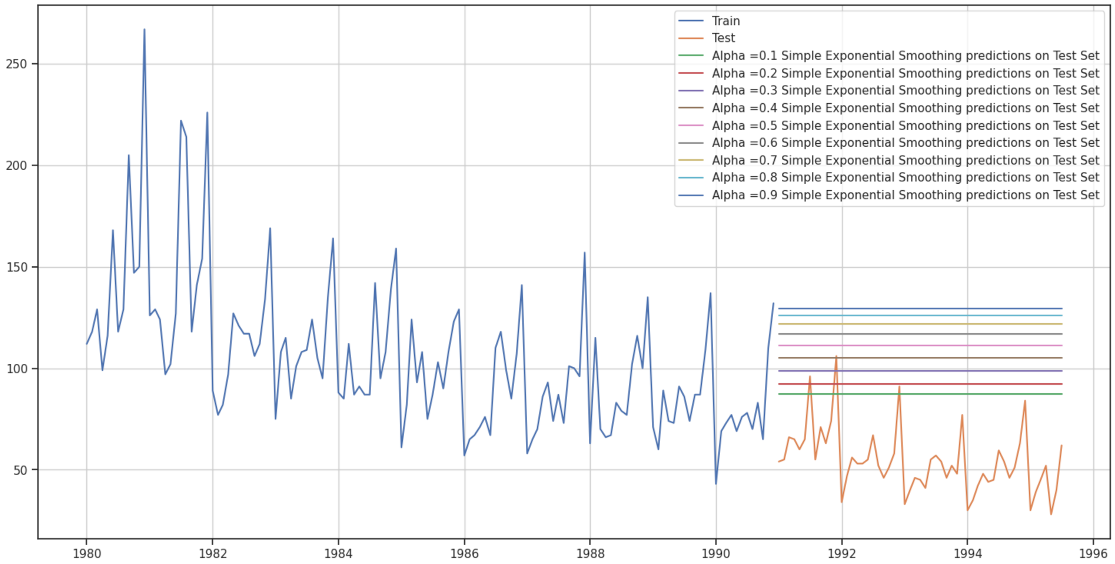
moving average





The model's performance was evaluated using the Root Mean Square Error (RMSE) metric. Multiple moving average models were constructed with rolling windows ranging from 2 to 9. Unlike a simple average, a rolling average considers only the preceding n values to make predictions, where n represents the rolling window size. This approach accounts for recent trends and is typically more accurate. A higher rolling window results in a smoother curve as it incorporates more values, enhancing the model's ability to capture underlying patterns in the data.

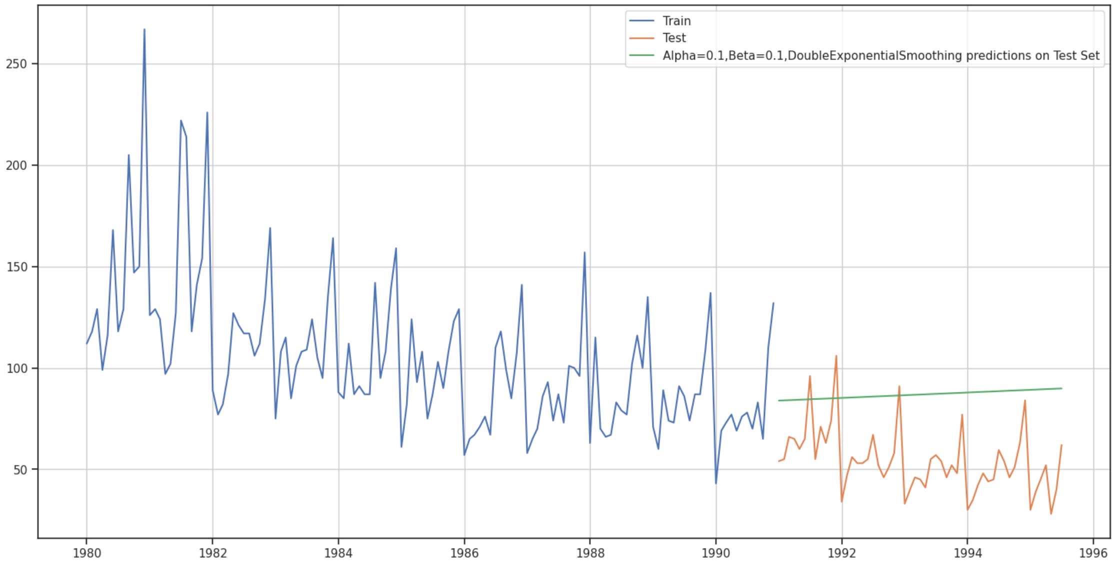
Simple expontial smoothing



Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha=0.1,SimpleExponentialSmoothing 36.429535

Double exponential smoothing



Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing36.510010

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.

Response:-

To assess the stationarity of the entire time series data, the Augmented Dickey-Fuller (ADF) test was conducted. This test is designed to detect the presence of a unit root in the time series, which indicates non-stationarity.

The hypothesis for the ADF test can be summarized as follows:

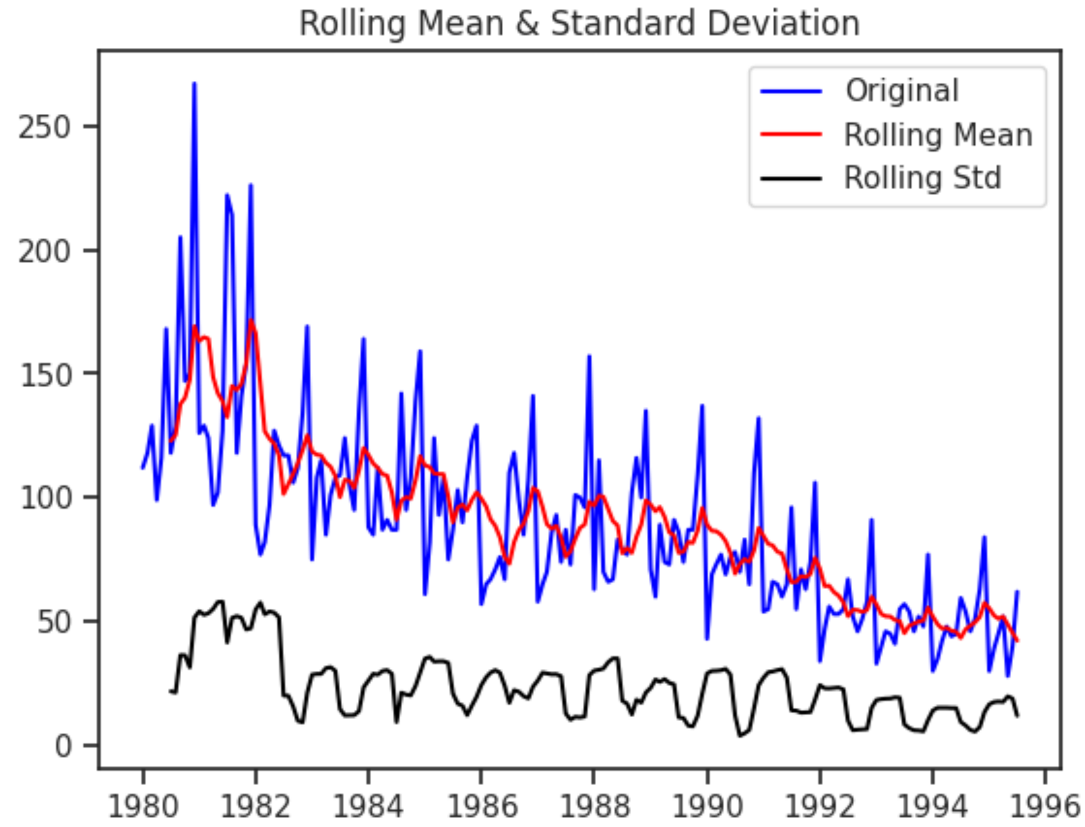
- Null Hypothesis (H0): The time series contains a unit root and is therefore non-stationary.

- Alternative Hypothesis (H1): The time series does not contain a unit root and is therefore stationary.

For the purpose of building ARIMA models, stationarity is desired. Therefore, we aim for the p-value of the ADF test to be less than the chosen significance level, denoted as α.

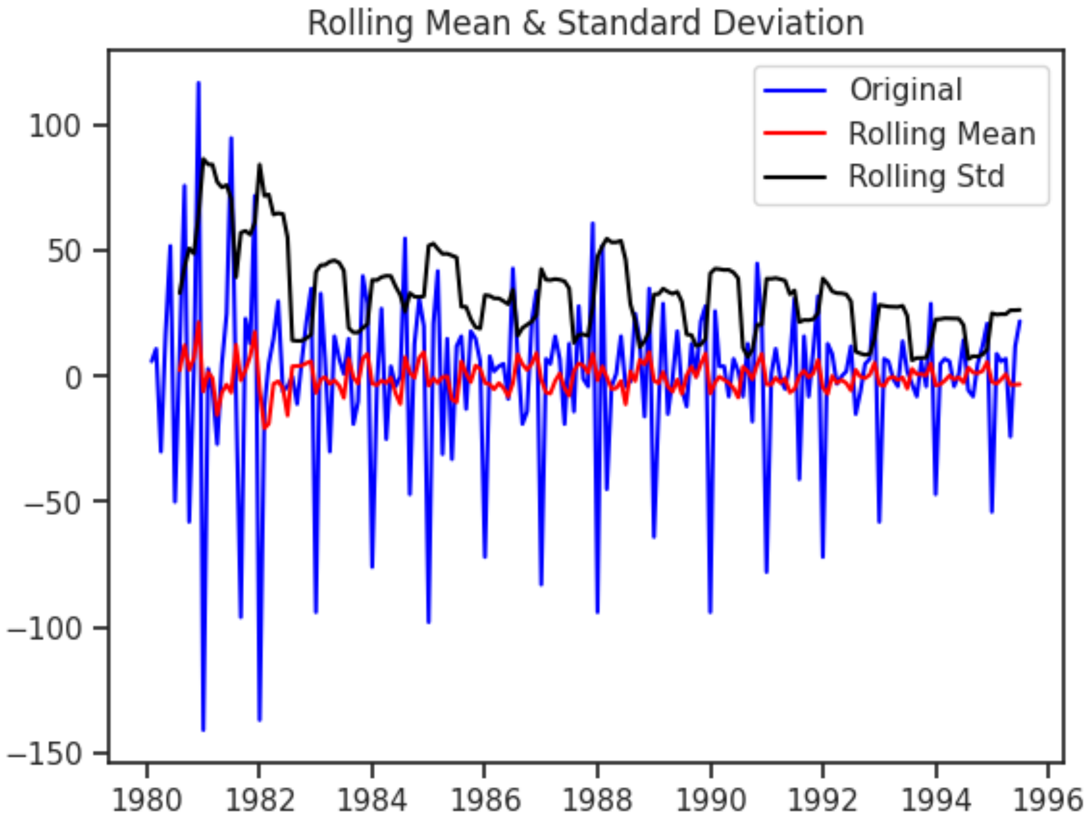
Upon analysis, it was found that at a 5% significance level, the p-value of the ADF test indicated that the time series is non-stationary. This suggests that we fail to reject the null hypothesis, indicating the presence of a unit root and confirming the non-stationarity of the series.

In summary, based on the results of the ADF test, the entire time series data is deemed to be non-stationary, which may require further preprocessing steps before building ARIMA models for accurate forecasting.



Results of Dickey-Fuller Test: Test Statistic -1.892338 p-value 0.335674 #Lags Used 13.000000 Number of Observations Used 173.000000 Critical Value (1%) -3.468726 Critical Value (5%) -2.878396 Critical Value (10%) -2.575756 dtype: float64

Since we couldn't reject the null hypothesis, it suggests that the series is inherently non-stationary. To address this, we applied the differencing approach to make the series stationary. Specifically, we utilized the .diff() function on the existing series, defaulting to a differencing order of 1. This process involved subtracting each observation from its preceding one. Additionally, as the first value after differencing would result in NaN, we removed these NaN values to ensure continuity in the data. This approach aimed to eliminate trends and seasonality, making the series suitable for further analysis and modeling.



```
Results of Dickey-Fuller Test: Test Statistic -8.032729e+00 p-value 1.938803e-12
#Lags Used 1.200000e+01 Number of Observations Used 1.730000e+02 Critical
Value (1%) -3.468726e+00 Critical Value (5%) -2.878396e+00 Critical Value (10%)
-2.575756e+00 dtype: float64
```

Rejecting the null hypothesis that the series is non-stationary at difference = 1 indicates that the series has become stationary following the differencing process.

4. 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.

   Response:-

```
 SARIMAX Results
==============================================================================
============= Dep. Variable: Rose_Sales No. Observations: 132 Model: ARIMA(2,
1, 3) Log Likelihood -631.348 Date: Sun, 25 Feb 2024 AIC 1274.695 Time: 18:46:04
BIC 1291.946 Sample: 01-01-1980 HQIC 1281.705 - 12-01-1990 Covariance Type:
opg
==============================================================================
============= coef std err z P>|z| [0.025 0.975] ------------------------------------------
---------------------------------- ar.L1 -1.6779 0.084 -20.034 0.000 -1.842 -1.514 ar.L2 -
```

```
0.7288 0.084 -8.702 0.000 -0.893 -0.565 ma.L1 1.0447 0.644 1.622 0.105 -0.217
2.307 ma.L2 -0.7718 0.134 -5.775 0.000 -1.034 -0.510 ma.L3 -0.9046 0.584 -1.549
0.121 -2.049 0.240 sigma2 858.8436 541.924 1.585 0.113 -203.308 1920.995
========================================================================
==================== Ljung-Box (L1) (Q): 0.02 Jarque-Bera (JB): 24.45 Prob(Q):
0.88 Prob(JB): 0.00 Heteroskedasticity (H): 0.40 Skew: 0.71 Prob(H) (two-sided):
0.00 Kurtosis: 4.57
========================================================================
====================
```

1. **Model Specification:**
   - The SARIMAX model is specified as SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12), indicating the use of seasonal and non-seasonal autoregressive and moving average components.
   - This suggests that the model incorporates three autoregressive terms, one moving average term, and three seasonal autoregressive terms, with a seasonal periodicity of 12 months.
2. **Model Fit:**
   - The log likelihood value is -377.200, indicating the maximized log-likelihood of the model.
   - The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used to evaluate model fit, with lower values indicating better fit. The AIC is 774.400 and the BIC is 799.618.
3. **Parameter Estimates:**
   - The coefficient estimates provide insights into the strength and significance of each component in the model.
   - Notably, the ma.L1 coefficient has a significant impact on the model, indicating the importance of the moving average term in capturing the data's behavior.
4. **Residual Diagnostics:**
   - The residuals of the model are examined for autocorrelation and heteroskedasticity using the Ljung-Box test and the Heteroskedasticity test, respectively.
   - The Ljung-Box (Q) statistic tests the residuals for autocorrelation at lag 1 and returns a value of 0.30, suggesting no significant autocorrelation.

- The Heteroskedasticity (H) test evaluates whether the variance of the residuals is constant over time. A p-value of 0.77 indicates no evidence of heteroskedasticity.

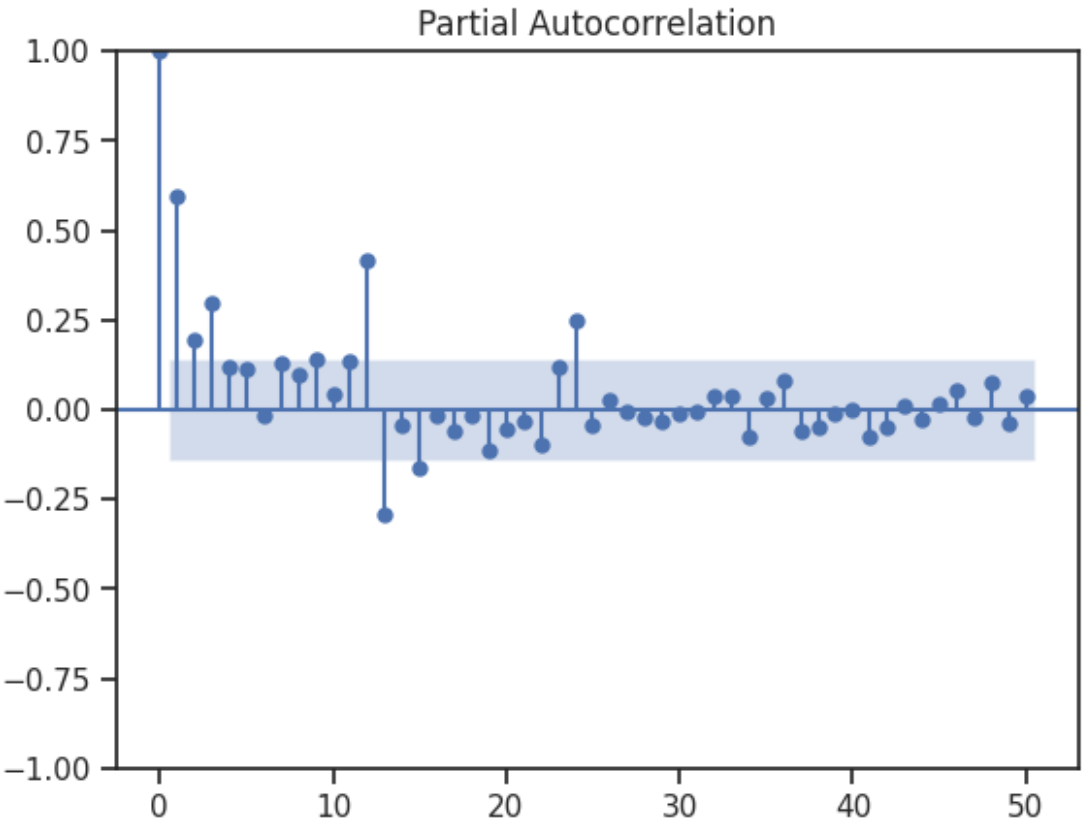
5. **Additional Tests:**

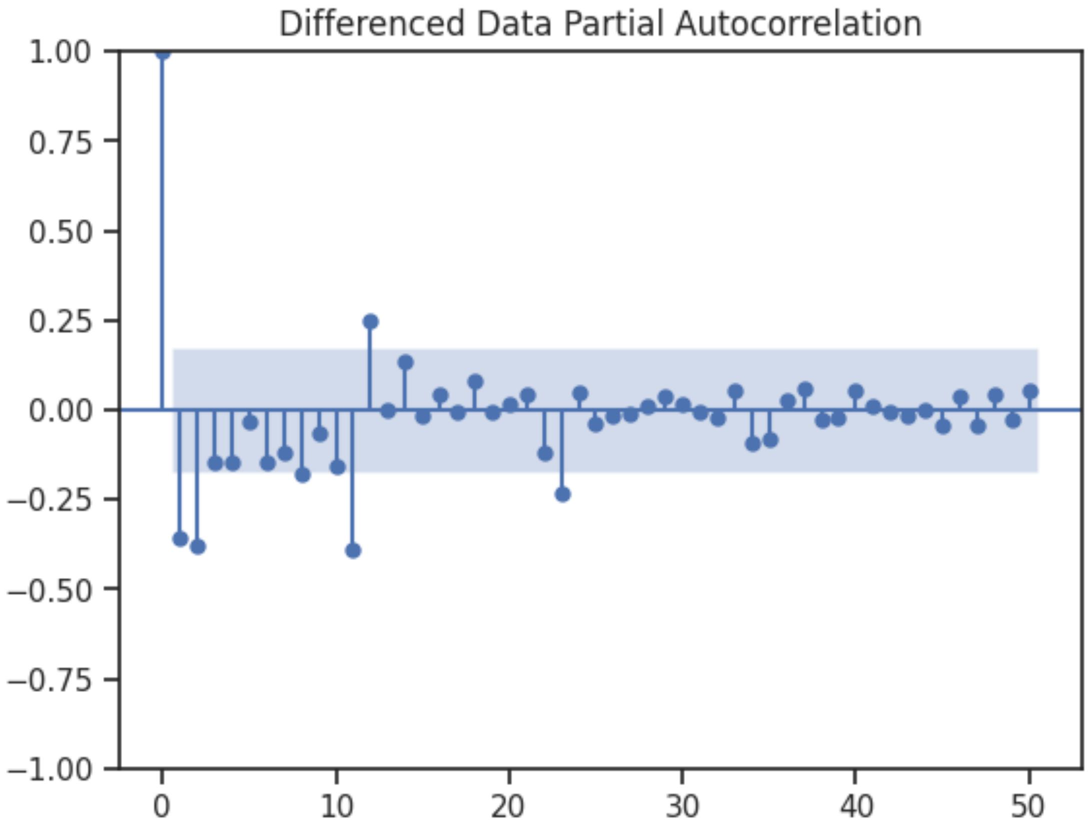
   - The Jarque-Bera (JB) test assesses the normality of the residuals. With a p-value of 0.44, there is no evidence to reject the null hypothesis of normality.

   - The model's skewness and kurtosis provide additional information about the distribution of residuals.

In summary, the SARIMAX model appears to provide a reasonable fit to the data, as indicated by the diagnostic tests and parameter estimates. However, further evaluation and validation may be necessary to ensure the model's reliability for making accurate sales predictions.

```
 SARIMAX Results
=====================================================
========================= Dep. Variable: y No. Observations: 132
Model: SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12) Log Likelihood -377.200 Date: Sun, 25 Feb
2024 AIC 774.400 Time: 18:50:34 BIC 799.618 Sample: 0 HQIC 784.578 - 132
Covariance Type: opg
=====================================================
=========== coef std err z P>|z| [0.025 0.975] ------------------------------------------
---------------------------------- ar.L1 0.0464 0.126 0.367 0.714 -0.201 0.294 ar.L2 -
0.0060 0.120 -0.050 0.960 -0.241 0.229 ar.L3 -0.1808 0.098 -1.838 0.066 -0.374
0.012 ma.L1 -0.9370 0.067 -13.903 0.000 -1.069 -0.805 ar.S.L12 0.7639 0.165 4.640
0.000 0.441 1.087 ar.S.L24 0.0840 0.159 0.527 0.598 -0.229 0.397 ar.S.L36 0.0727
0.095 0.764 0.445 -0.114 0.259 ma.S.L12 -0.4969 0.250 -1.988 0.047 -0.987 -0.007
ma.S.L24 -0.2191 0.210 -1.044 0.296 -0.630 0.192 sigma2 192.1390 39.627 4.849
0.000 114.471 269.807
=====================================================
================ Ljung-Box (L1) (Q): 0.30 Jarque-Bera (JB): 1.64 Prob(Q):
0.58 Prob(JB): 0.44 Heteroskedasticity (H): 1.11 Skew: 0.33 Prob(H) (two-sided):
0.77 Kurtosis: 3.03
=====================================================
================ Warnings:
```


Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50



```
 SARIMAX Results
======================================================================
============= Dep. Variable: Rose_Sales No. Observations: 132 Model: ARIMA(2,
1, 2) Log Likelihood -635.935 Date: Sun, 25 Feb 2024 AIC 1281.871 Time: 18:50:37
BIC 1296.247 Sample: 01-01-1980 HQIC 1287.712 - 12-01-1990 Covariance Type:
opg
======================================================================
============= coef std err z P>|z| [0.025 0.975] ------------------------------------------
---------------------------------- ar.L1 -0.4540 0.469 -0.969 0.333 -1.372 0.464 ar.L2
0.0001 0.170 0.001 0.999 -0.334 0.334 ma.L1 -0.2541 0.459 -0.554 0.580 -1.154
0.646 ma.L2 -0.5984 0.430 -1.390 0.164 -1.442 0.245 sigma2 952.1601 91.424
10.415 0.000 772.973 1131.347
======================================================================
================== Ljung-Box (L1) (Q): 0.02 Jarque-Bera (JB): 34.16 Prob(Q):
0.88 Prob(JB): 0.00 Heteroskedasticity (H): 0.37 Skew: 0.79 Prob(H) (two-sided):
0.00 Kurtosis: 4.94
======================================================================
==================
```

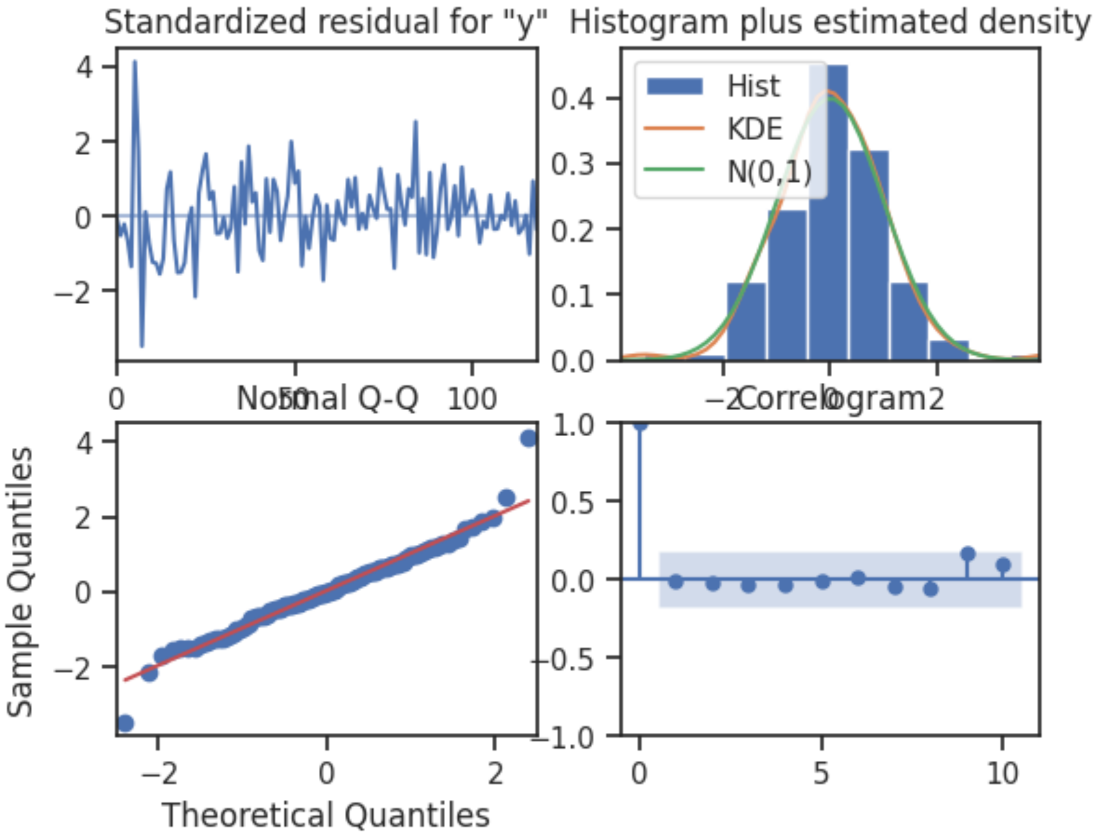
1. **Model Specification:**

- The SARIMAX model is specified as ARIMA(2, 1, 2), indicating the use of autoregressive and moving average components with differencing of order 1.
- This suggests that the model incorporates two autoregressive terms, one moving average term, and differencing of order 1 to achieve stationarity.

2. **Model Fit:**
    - The log likelihood value is -635.935, indicating the maximized log-likelihood of the model.
    - The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used to evaluate model fit. The AIC is 1281.871 and the BIC is 1296.247.
3. **Parameter Estimates:**
    - The coefficient estimates provide insights into the strength and significance of each component in the model.
    - Notably, the ma.L2 coefficient is statistically significant at the 5% level, indicating the importance of the second moving average term in capturing the data's behavior.
4. **Residual Diagnostics:**
    - The residuals of the model are examined for autocorrelation and heteroskedasticity using the Ljung-Box test and the Heteroskedasticity test, respectively.
    - The Ljung-Box (Q) statistic tests the residuals for autocorrelation at lag 1 and returns a value of 0.02, suggesting no significant autocorrelation.
    - The Heteroskedasticity (H) test evaluates whether the variance of the residuals is constant over time. A p-value of 0.00 indicates evidence of heteroskedasticity.
5. **Additional Tests:**
    - The Jarque-Bera (JB) test assesses the normality of the residuals. With a p-value of 0.00, there is evidence to reject the null hypothesis of normality.
    - The model's skewness and kurtosis provide additional information about the distribution of residuals.

In summary, while the SARIMAX model appears to provide a reasonable fit to the data, some diagnostic tests suggest potential issues such as heteroskedasticity and non-normality of residuals. Further evaluation and refinement may be necessary to ensure the model's reliability for making accurate sales predictions.

```
 SARIMAX Results
=============================================================================
========================== Dep. Variable: y No. Observations: 132 Model:
SARIMAX(2, 1, 2)x(2, 1, 2, 12) Log Likelihood -538.016 Date: Sun, 25 Feb 2024 AIC
1094.031 Time: 18:50:40 BIC 1119.044 Sample: 0 HQIC 1104.188 - 132 Covariance
Type: opg
=============================================================================
============= coef std err z P>|z| [0.025 0.975] -------------------------------------------------
---------------------------------- ar.L1 -0.5491 0.228 -2.408 0.016 -0.996 -0.102 ar.L2 -
0.0744 0.099 -0.753 0.452 -0.268 0.119 ma.L1 -0.1703 0.216 -0.787 0.431 -0.594
0.254 ma.L2 -0.6694 0.228 -2.937 0.003 -1.116 -0.223 ar.S.L12 -1.0134 0.524 -1.935
0.053 -2.040 0.013 ar.S.L24 -0.1002 0.175 -0.572 0.568 -0.444 0.243 ma.S.L12
0.2908 31.458 0.009 0.993 -61.365 61.947 ma.S.L24 -0.7079 22.376 -0.032 0.975 -
44.565 43.149 sigma2 430.3732 1.33e+04 0.032 0.974 -2.57e+04 2.66e+04
=============================================================================
===================== Ljung-Box (L1) (Q): 0.02 Jarque-Bera (JB): 27.15 Prob(Q):
0.90 Prob(JB): 0.00 Heteroskedasticity (H): 0.33 Skew: 0.26 Prob(H) (two-sided):
0.00 Kurtosis: 5.28
=============================================================================
===================== Warnings: [1] Covariance matrix calculated using the outer
product of gradients (complex-step).
```



5. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Response:-

	Test RMSE
Alpha=0.2,Beta=0.7,Gamma=0.2,TripleExponentialSmoothing	8.992350
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
(2,1,2)(2,1,2,12),Manual_SARIMA	14.974153
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535716

	Test RMSE
Auto_ARIMA	36.418573
Alpha=0.1,SimpleExponentialSmoothing	36.429535
ARIMA(3,1,3)	36.473225
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	37.192623
Linear Regression	51.080941
Simple Average Model	53.049755

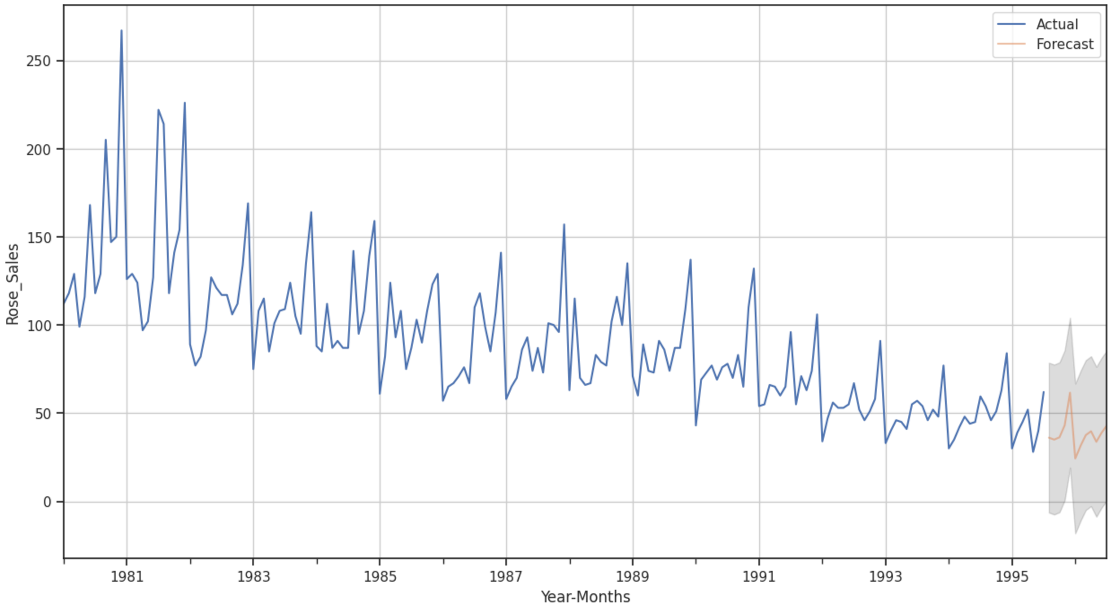
1. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Response:-

Rose_Sales_Predictions	
1995-08-01	36.096841
1995-09-01	34.999961
1995-10-01	36.289937
1995-11-01	43.126839
1995-12-01	61.593978
1996-01-01	24.293852
1996-02-01	31.406019
1996-03-01	37.545514
1996-04-01	39.735393
1996-05-01	33.753457
1996-06-01	38.868148
1996-07-01	43.093112

After carefully evaluating all the models we constructed, it is evident that the triple exponential smoothing, also known as the Holt-Winters model, consistently demonstrates the lowest Root Mean Square Error (RMSE). As a result, this model emerges as the most optimal choice for generating sales predictions.

Therefore, the forecasts produced by this preferred model are considered the most accurate and reliable among all the models examined. This implies that the Holt-Winters model offers the best predictive capability for forecasting sales, providing valuable insights for the company's strategic decision-making processes.



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

Response:-

1. **Trend Analysis of Rose Wine Sales:** Examination of the wine sales data highlights a consistent downward trajectory in the popularity of Rose wine within the company's offerings. This decline has persisted for over a decade, indicating a sustained shift in consumer preferences away from this variety.
2. **Future Projections:** Projections generated by the most optimal model reinforce the expectation that this downward trend for Rose wine will persist in the foreseeable future. Therefore, proactive measures are necessary to address this ongoing decline.
3. **Seasonal Sales Dynamics:** Sales of wine exhibit notable fluctuations in response to seasonal changes. Typically, there's an uptick in sales during festive seasons, while demand tends to wane during peak winter months, notably in January.
4. **Campaign Recommendations:** To counteract subdued sales during non-peak periods, the company should consider launching targeted campaigns aimed at boosting Rose wine consumption. Specifically, focusing efforts during the lean

period from April to June could yield the most significant results, as sales are typically at their lowest during this timeframe.

5. **Optimal Timing for Campaigns:** It's essential to strategically time marketing initiatives. Campaigns during peak sales periods, such as festivals, may not yield substantial improvements as demand is already high. Conversely, campaigns during peak winter months, notably January, are discouraged due to reduced consumer interest in purchasing wine during colder weather conditions.

6. **Revamping Strategies:** To address the underlying reasons behind the declining popularity of Rose wine, the company should conduct a thorough analysis. This includes examining production methods, branding, and marketing strategies to identify areas for improvement and potentially revitalize interest in the product.

In summary, the analysis underscores the importance of proactive measures to mitigate the declining sales trend of Rose wine. By strategically timing campaigns and reassessing product strategies, the company can better position itself to address market challenges and potentially regain market share in the wine industry.