PROJECT

ON

TIME SERIES FORECASTING

By SHAJIL FERNANDEZ

16-12-2023

Table of Contents:

Problem 1 – Sparkling Data:

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

1.2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition…………………………………………………...…... Pg - 2

1.3. Split the data into training and test. The test data should start in 1991……………………………………………………………………………...…Pg - 6

1.4. 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…………………………………………………………………………..… Pg -7

1.5. 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………………….…………… Pg -13

1.6. 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…. Pg -17

1.7. 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...………………………………………………………………………...…… Pg -21

1.8. 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……………………………...……………………..…… Pg -21

1.9. 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……………………………………………………………………………….. Pg -24

Problem 2 – Rose Data:

2.1. Read the data as an appropriate Time Series data and plot the data... Pg - 25

2.2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.……………………………………………...….. Pg - 26

2.3. Split the data into training and test. The test data should start in 1991………………………………………………………………………………………Pg - 30

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.…………………………………………………………………………………. Pg -31

2.5. 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.……………….……………… Pg -36

2.6. 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…. Pg -40

2.7. 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...……………………………………………………………………...…… Pg -45

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……………………………..……………….…… Pg -45

2.9. 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……………………………………………………………………………... Pg -48

List of Figures:

SPARKLING DATA:

Figure 1: Plotting the time series data ……………………................................……………. Pg -1

Figure 2: Yearly BoxPlot …………………………………………………………………..…….….... Pg -2

Figure 3: Monthly BoxPlot ……………………………………………....…………….…….… Pg -3

Figure 4: MonthPlot ………....……………………………………………………………….…..... Pg -4

Figure 5: Plot showing monthly wine sales across years ……………….……………......… Pg -5

Figure 6: Decomposition with multiplicative seasonality ……………………..……….…..… Pg -6

Figure 7: Train and Test data …………………………………………….…...…………….…... Pg -6

Figure 8: Linear regression model on train and test data ………………………………...…... Pg -7

Figure 9: Naïve forecast model on train and test data …………………………….………. Pg -8

Figure 10: Simple average model on train and test data …………………….…….……..…. Pg -9

Figure 11: SES model on train and test data (Alpha=0.0702)……………………….…..…….. Pg -10

Figure 12: SES model on train and test data (Alpha=0.03)……………………….………….. Pg -10

Figure 13: DES model on train and test data ……………………………….....……..…….. Pg -11

Figure 14: TES model on train and test data (Alpha=0.1119, Beta=0.0498, Gamma=0.3617)…………………………………………………………….………………. Pg -12

Figure 15: TES model on train and test data (Alpha=0.4, Beta=0.1, Gamma=0.2)………….. Pg -13

Figure 16: ADF test on train set ……………………………………….……………..……….…... Pg -14

Figure 17: Differencing (d=1) on train set ………………………...…………………..………….. Pg -15

Figure 18: Log transformation on train set……………………………….…………….………. Pg -15

Figure 19: Differencing (d=1) on log transformed data ……………………………..…..… Pg -16

Figure 20: SARIMAX results for ARIMA automated version (AIC)…………………….…..…. Pg -17

Figure 21: Predictions on test data ARIMA automated version (AIC)………………………..Pg -18

Figure 22: ACF ……………………………………………………………………………………Pg -19

Figure 23 PACF ……………………………………………………………………...……….. Pg -19

Figure 24: SARIMAX results for ARIMA manual version (ACF and PACF)……………….….. Pg -20

Figure 25: Predictions on test data ARIMA manual version (ACF and PACF)……………..… Pg -20

Figure 26: Predictions on test data (future 12 months)……………………………………..…. Pg -22

Figure 27: Plot showing 95% confidence interval …………………………………………..…. Pg -23

Figure 28: Plot showing 90% confidence interval ………………………………………….... Pg -23

ROSE DATA:

Figure 29: Plotting the time series data ……………………............................……….…. Pg -25

Figure 30: Yearly BoxPlot ………………………………………………………….…..…........ Pg -26

Figure 31: Monthly BoxPlot …………………………………………....………….....……. Pg -27

Figure 32: MonthPlot …………...……………………………………………………….…..... Pg -28

Figure 33: Plot showing monthly wine sales across years ………………….……...…......…. Pg -29

Figure 34: Decomposition with multiplicative seasonality ……………………..…………..…. Pg -30

Figure 35: Train and Test data …………………………………………...…………….…..... Pg -30

Figure 36: Linear regression model on train and test data ………………………………..….... Pg -31

Figure 37: Naïve forecast model on train and test data ……………………………...…..…. Pg -32

Figure 38: Simple average model on train and test data ………………..…………..…...…. Pg -32

Figure 39: SES model on train and test data (Alpha=0.0987)……………………….…….….. Pg -33

Figure 40: SES model on train and test data (Alpha=0.07)……………………….……….….. Pg -34

Figure 41: DES model on train and test data ……………………………….....…….……….. Pg -35

Figure 42: TES model on train and test data (Alpha=0.0439, Beta=2.2383, Gamma=0.0005)…………………………………………………………………………………… Pg -36

Figure 43: ADF test on train set ………………………………………………..…….…….. Pg -37

Figure 44: Differencing (d=1) on train set ……………………...……………………..….……… Pg -38

Figure 45: Log transformation on train set……………………………………….……...……. Pg -39

Figure 46: Differencing (d=1) on log transformed data …………………………………….…. Pg -39

Figure 47: SARIMAX results for ARIMA automated version (AIC)…………………………..…. Pg -41

Figure 48: Predictions on test data ARIMA automated version (AIC)…………………………..Pg -41

Figure 49: ACF …………………………………………………………………………………Pg -42

Figure 50: PACF ……………………………………………………………….....……….. Pg -43

Figure 51: SARIMAX results for ARIMA manual version (ACF and PACF)……………….….. Pg -43

Figure 52: Predictions on test data ARIMA manual version (ACF and PACF)………………. Pg -44

Figure 53: Predictions on test data (future 12 months)…………………………………….…. Pg -46

Figure 54: Plot showing 95% confidence interval ………………………………….……….…. Pg -47

Figure 55: Plot showing 90% confidence interval ………………………………………….... Pg -48

List of Tables:

SPARKLING DATA:

Table 1: Top 5 rows of the dataset………………………………………………………….…..……. Pg - 1

Table 2: Statistical summary of the dataset ……………………….…………….………………... Pg - 1

Table 3: Monthly wine sales across years ……………………………………..…………………. Pg - 2

Table 4: AIC values ……………………………………………………….…….………….……….... Pg - 2

Table 5: Data frame showing built models with RMSE values ……………………..………..….. Pg - 3

Table 6: 95% confidence interval …………………………………….……………………….…. Pg - 3

Table 7: 90% confidence interval ………………………………………………………….……..... Pg - 3

ROSE DATA:

Table 8: Top 5 rows of the dataset………………………………………………………….……..… Pg - 25

Table 9: Statistical summary of the dataset ……………………….……………..…………..…... Pg - 26

Table 10: Monthly wine sales across years ………..…………………………..……….……… Pg - 28

Table 11: AIC values ……………………….………………………..……………..……….... Pg - 40

Table 12: Data frame showing built models with RMSE values …………………………..…... Pg - 45

Table 13: 95% confidence interval ………………………………….……………...………… Pg - 46

Table 14: 90% confidence interval ……………………………………………...…..…….…... Pg - 47

Problem 1:

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.

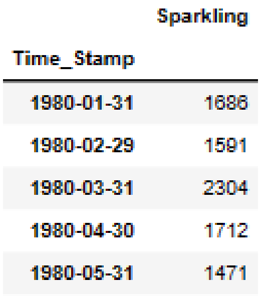
Sparkling data:

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

Ans:

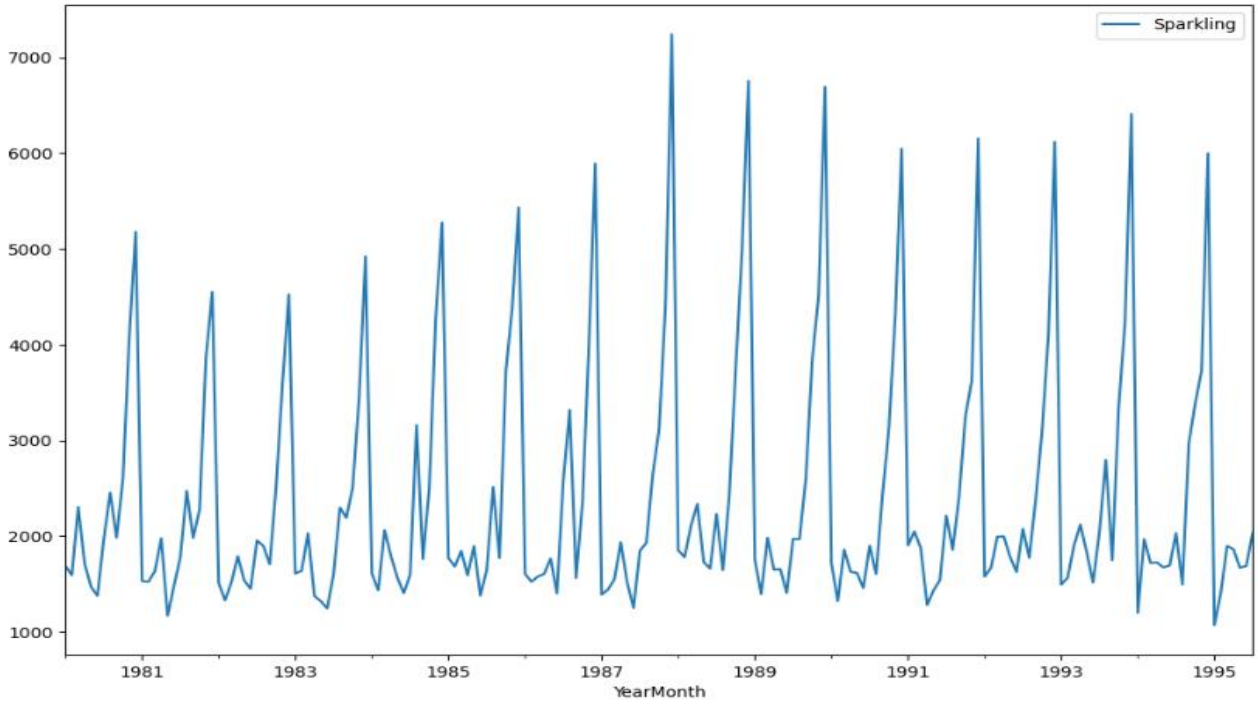
- There are 187 rows and 2 columns in the given dataset.
- Creating a timestamp column and adding it to the dataframe.
- Removing 'YearMonth' column and updating the timestamp as index.

Table 1: Top 5 rows of the dataset



	Sparkling
Time_Stamp	
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Figure 1: Plotting the time series data



Insights:

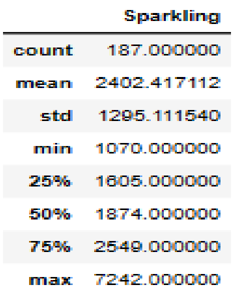
- The above plot does not show any presence of trend over the time period.

- The above time series plot has a presence of seasonality.

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

Ans:

Table 2: Statistical summary of the dataset

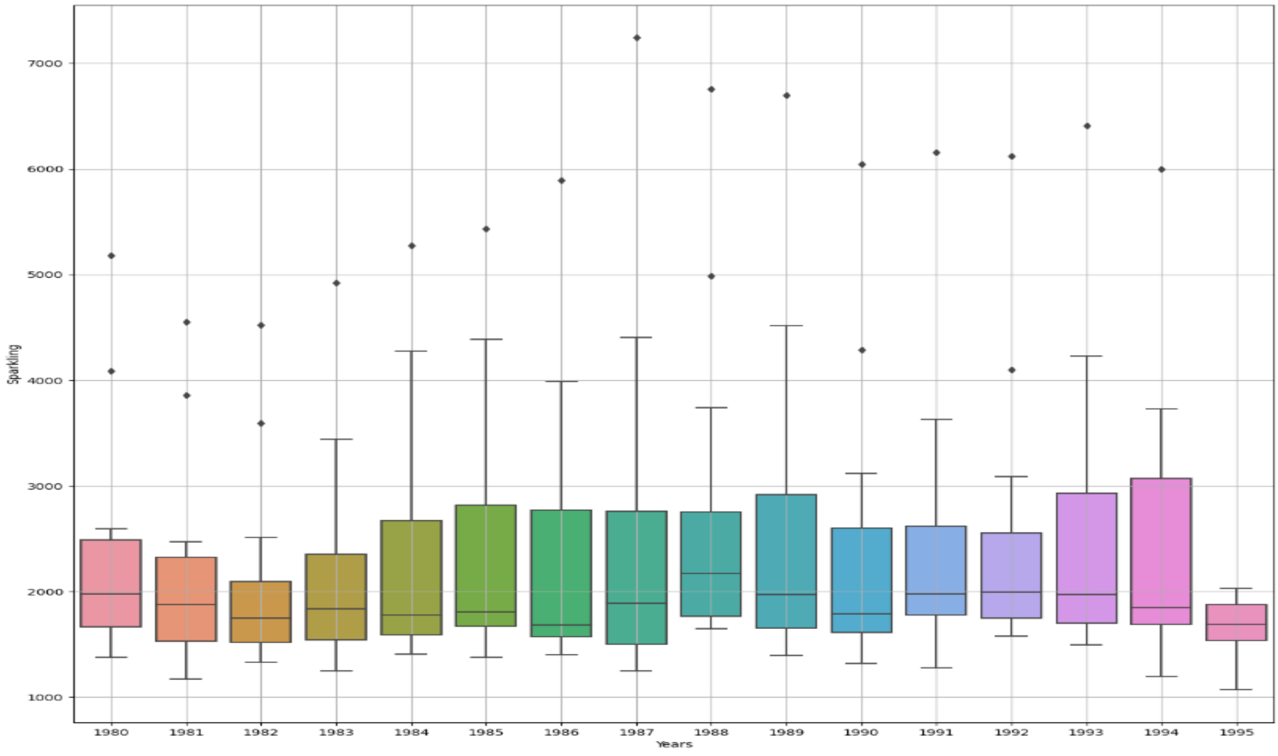


	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

Observations:

- Mean sales of Sparkling wine is 2402.42.
- About 75% of data have sale below 2549.

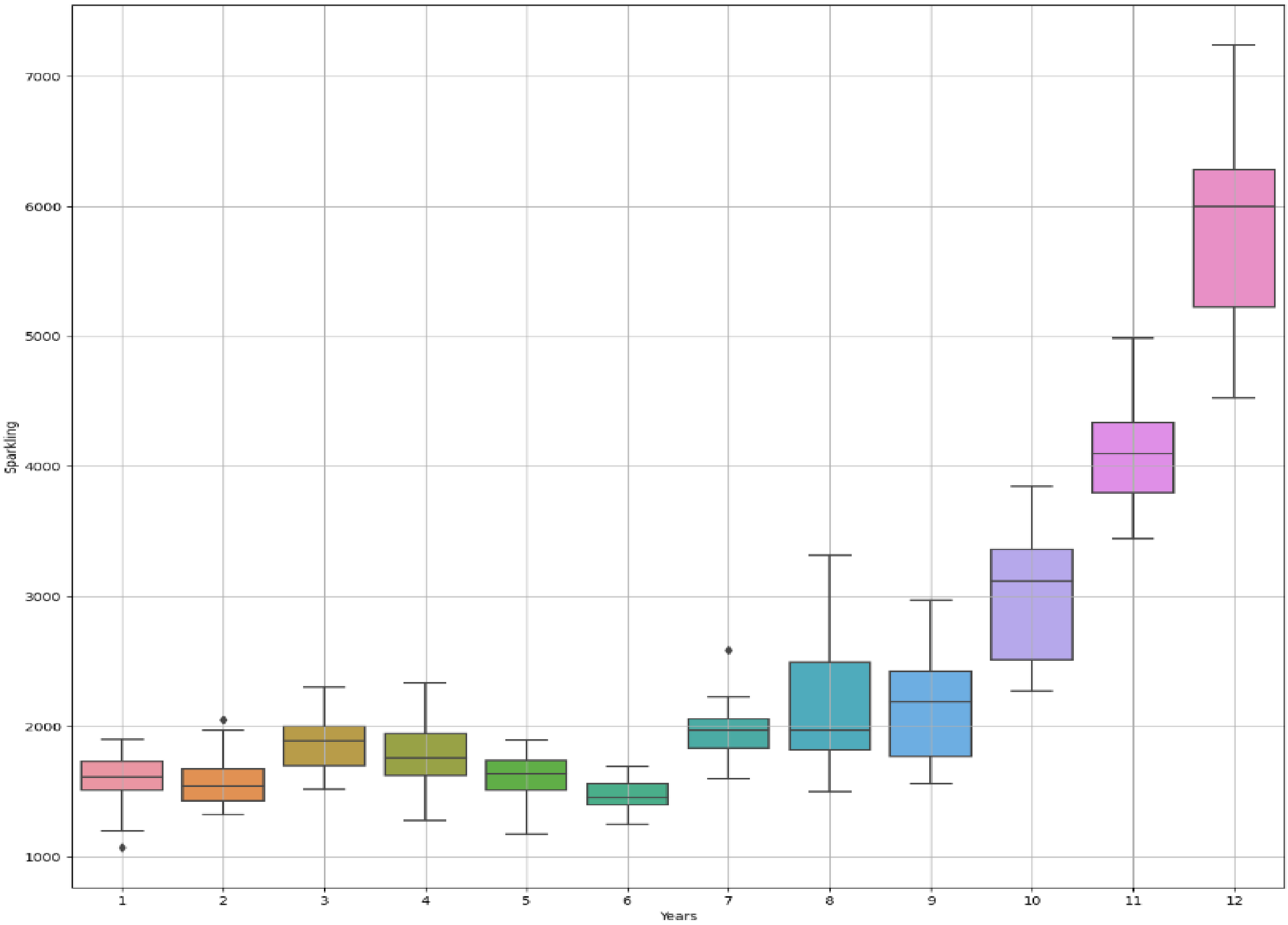
Figure 2: Yearly BoxPlot



Observations:

- We can observe a slight variation in the median across years.
- Based on the above boxplot, there is no presence of Trend.

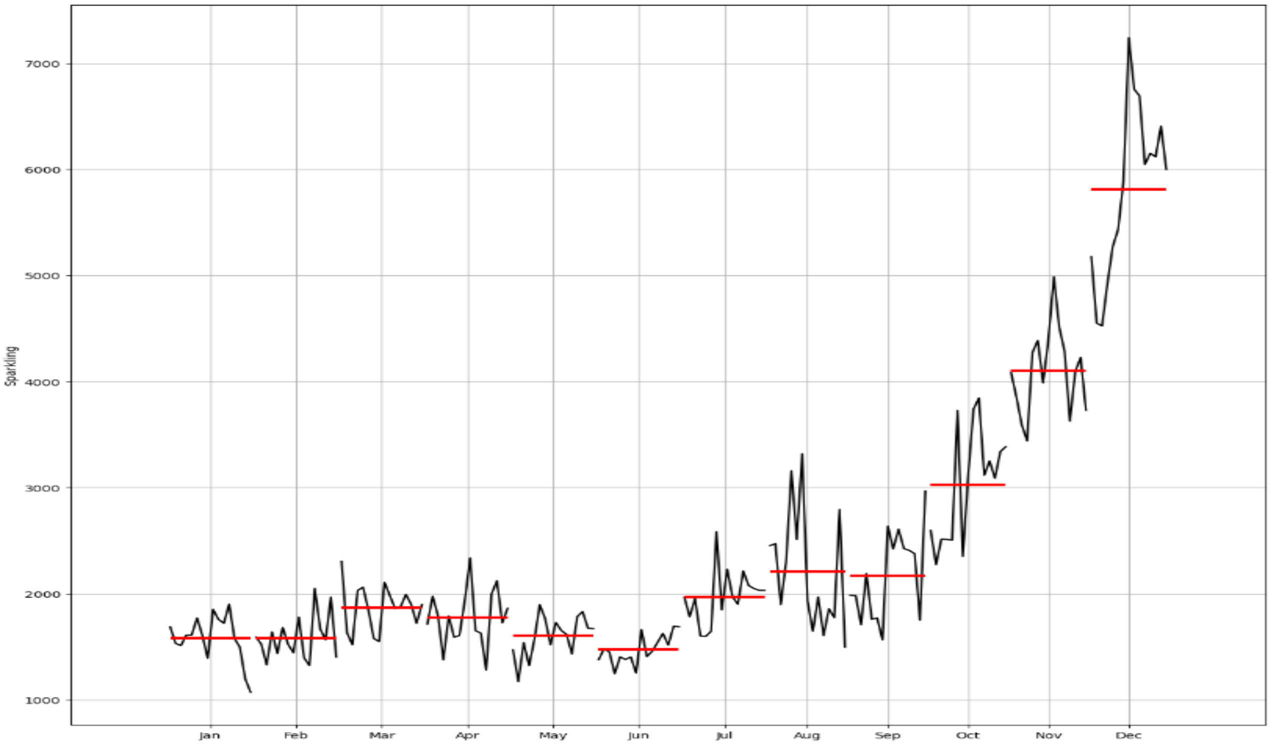
Figure 3: Monthly BoxPlot



Observations:

- We can note a peak in sales during the last quarter of a year.
- Maximum sale is observed in the month of December.

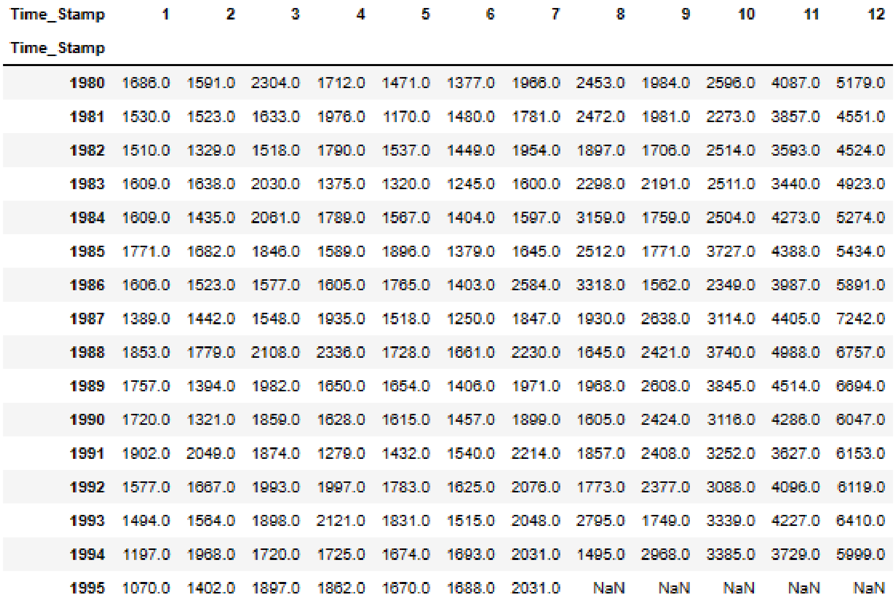
Figure 4: MonthPlot



Observation:

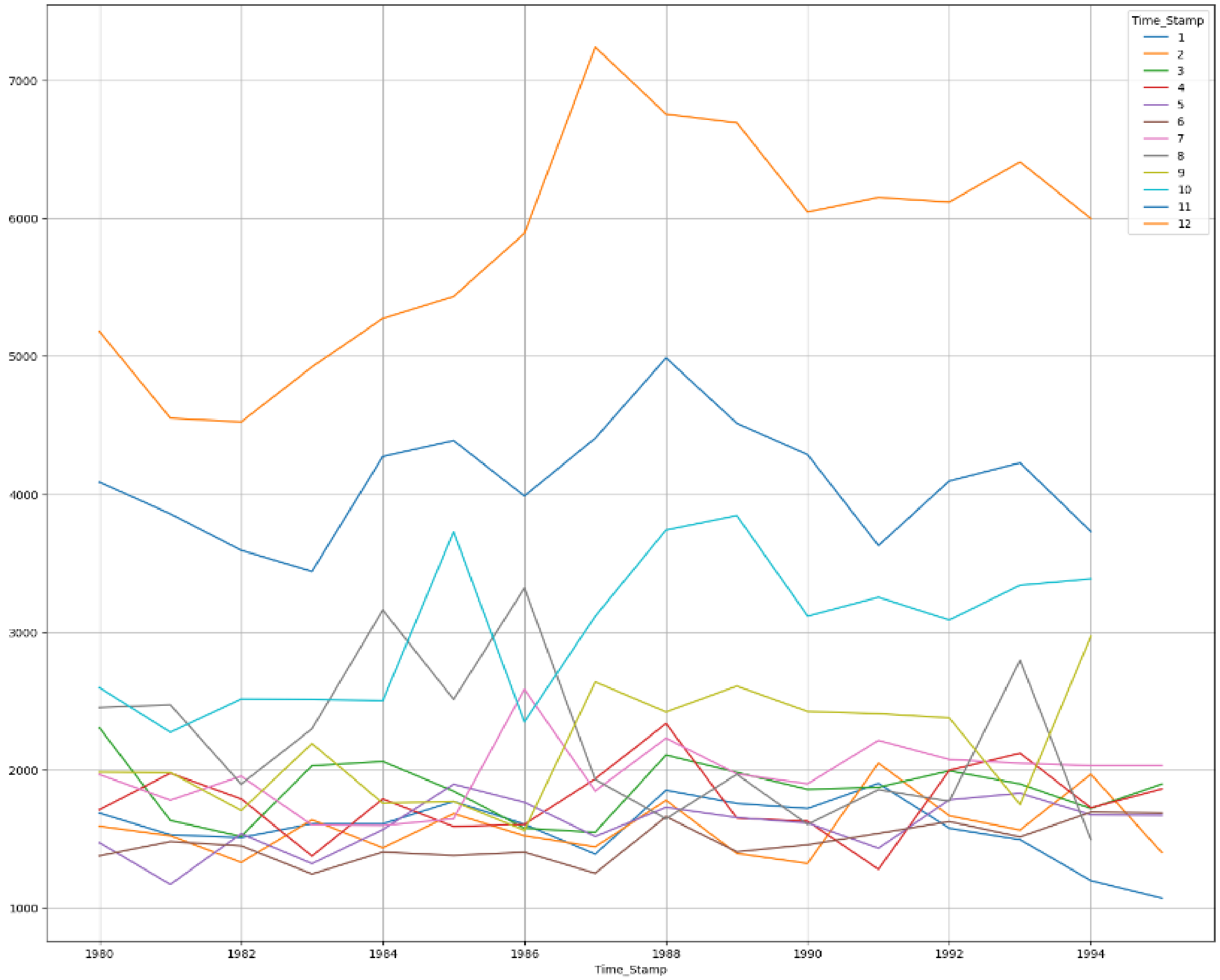
- June sees lowest sales across years.

Table 3: Monthly wine sales across years



Time_Stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_Stamp												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

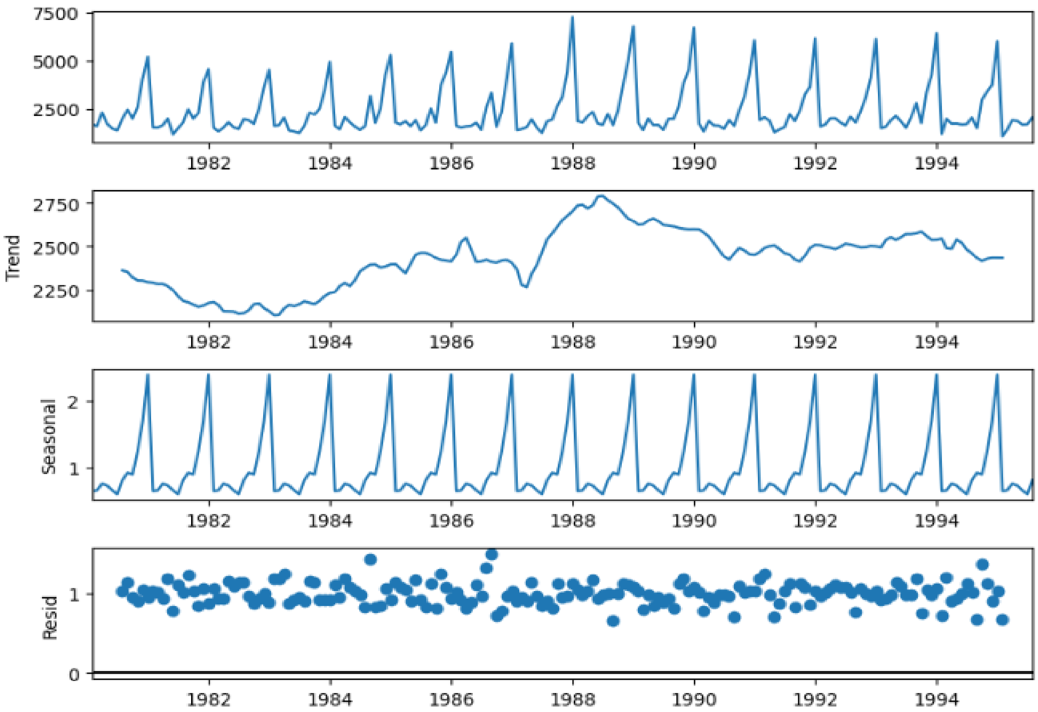
Figure 5: Plot showing monthly wine sales across years



Observations:

- Last quarter of the year shows the maximum sales.
- 1987 December, witnessed the highest sales in sparkling wine.

Figure 6: Decomposition with multiplicative seasonality



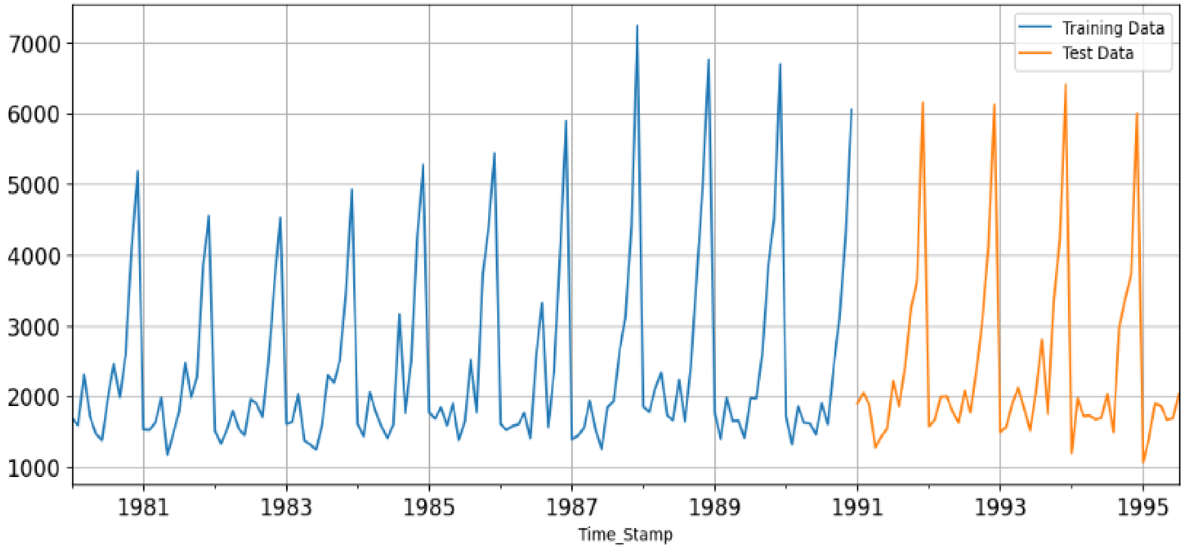
Observations:

- Seasonality is observed, and there is no trend even though there are some variations in the trend.
- Residuals are almost flat and there is no pattern observed.

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

Ans:

Figure 7: Train and Test data



Observations:

- Test data is considered from 1991 onwards.
- There are 132 and 55 values in the train and test set.

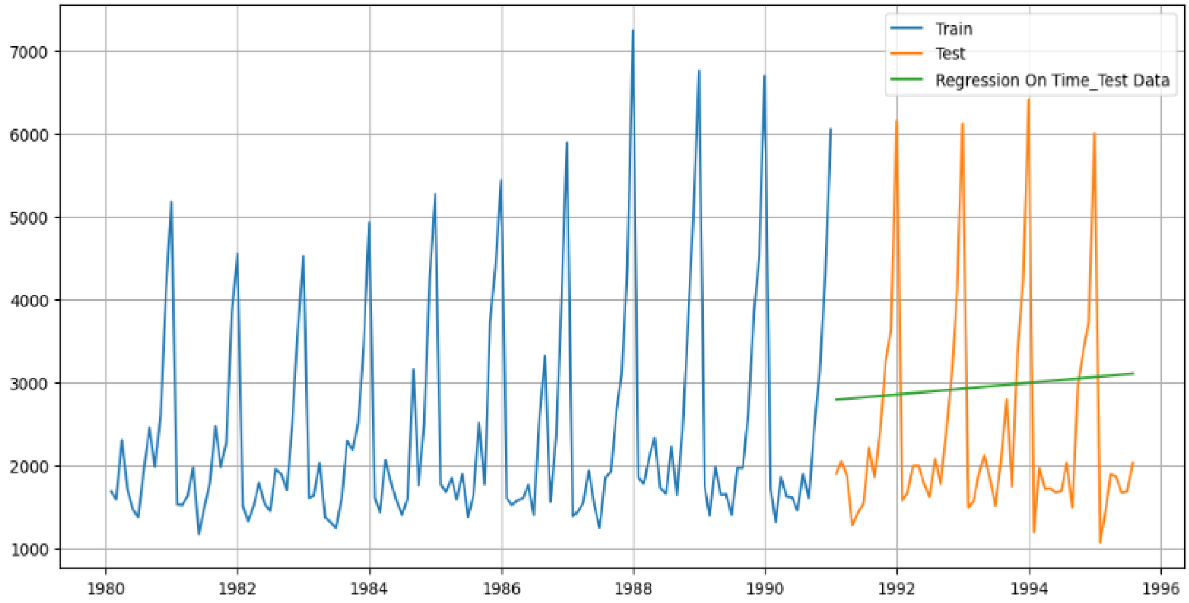
1.4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Ans:

a) Model 1: Linear Regression:

- We are generating the numerical time instance and adding the same to the train and test set.
- Initializing and Fitting the Linear regression model to the train set.

Figure 8: Linear regression model on train and test data



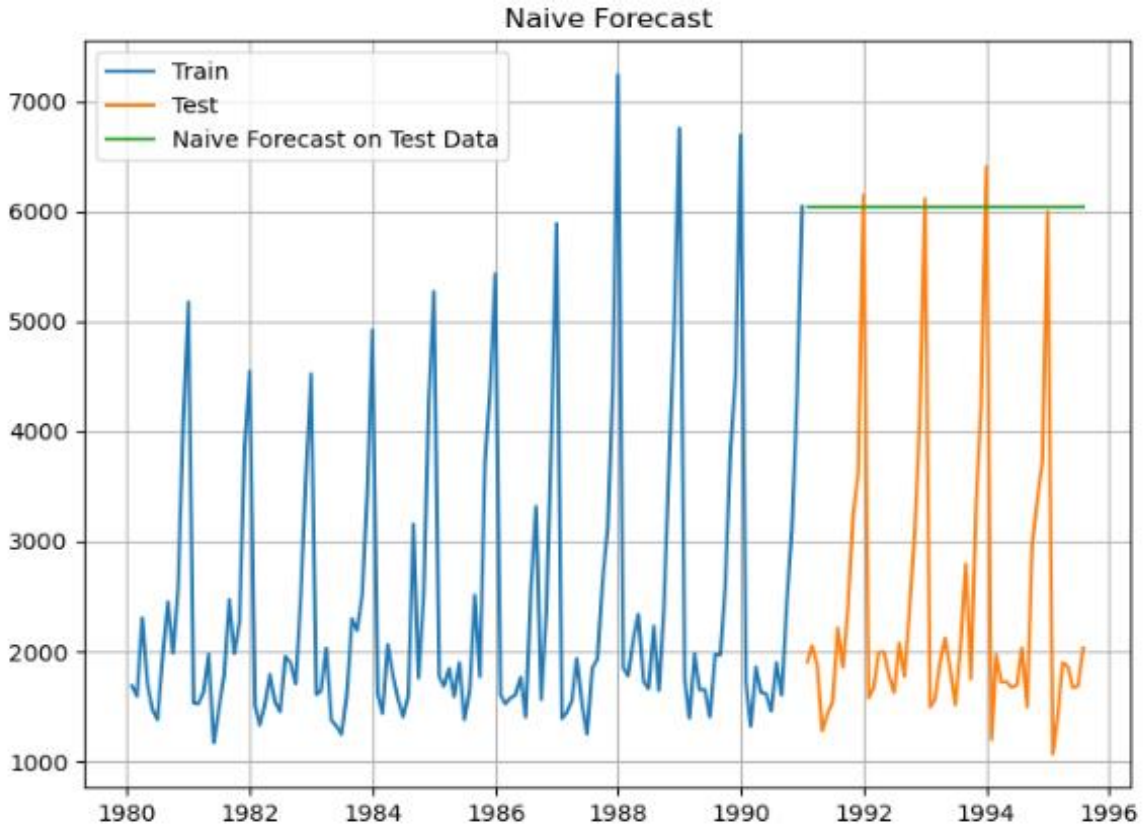
Observations:

- The above plot shows a slight increase in trend on the predicted values.
- RMSE on the test data for Linear Regression is 1389.14.

b) Model 2: Naïve Forecast:

- Naïve model focuses on the latest value to predict future sales.

Figure 9: Naïve forecast model on train and test data



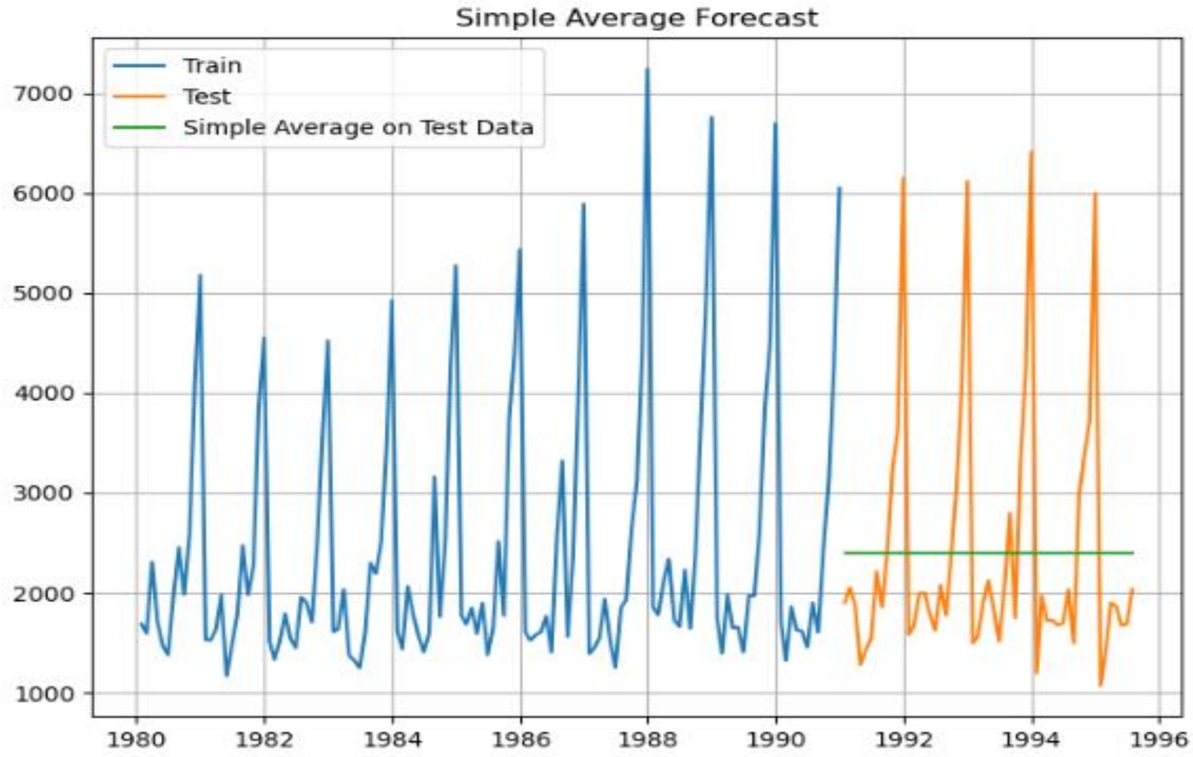
Observations:

- The above plot shows a flat trend, as it predicts values based on the latest value.
- RMSE on the test data for Naïve forecast model is 3864.28.

c) Model 3: Simple Average:

- Simple Average method uses the mean of the training values to predict future sales.

Figure 10: Simple average model on train and test data



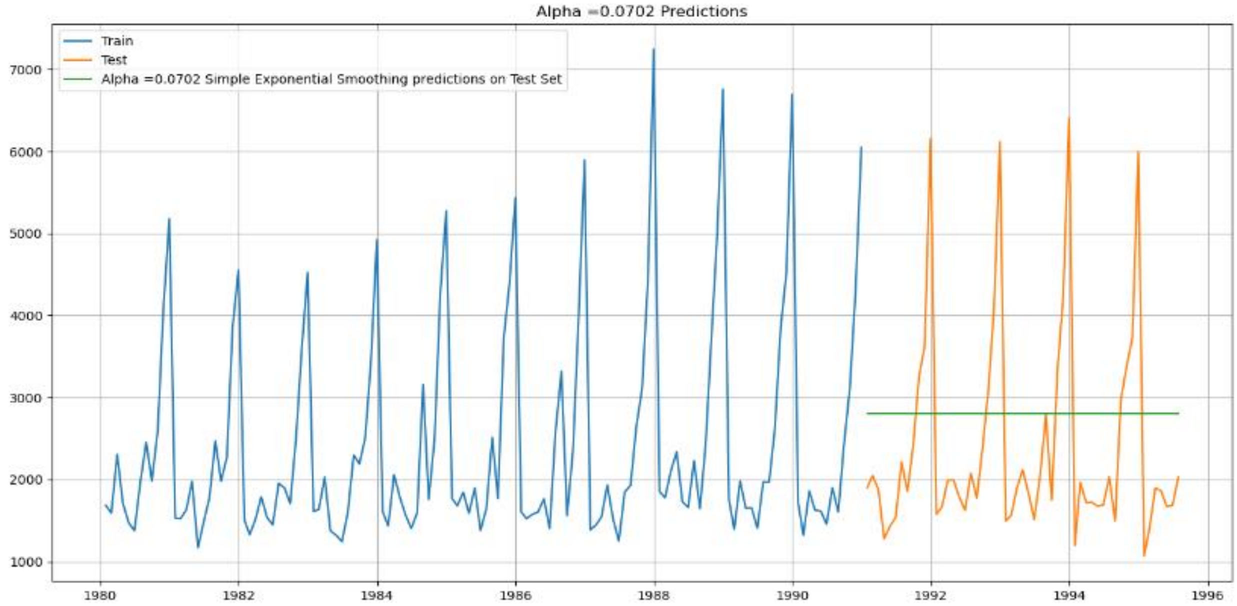
Observations:

- Simple average model does not show any trend, as it uses the mean of training values.
- RMSE on the test data for Simple Average model is 1275.08.

d) Model 4: Simple Exponential Smoothing:

- Simple exponential smoothing model is generally used when the data does not have any trend or seasonality. However, in this case, since the data is having seasonality, SES model may not be effective.
- Let’s Initialize and Fit the SES model on the train set.
- We got a smoothing level of 0.0702 when we fit the model on the data.

Figure 11: SES model on train and test data (Alpha=0.0702)

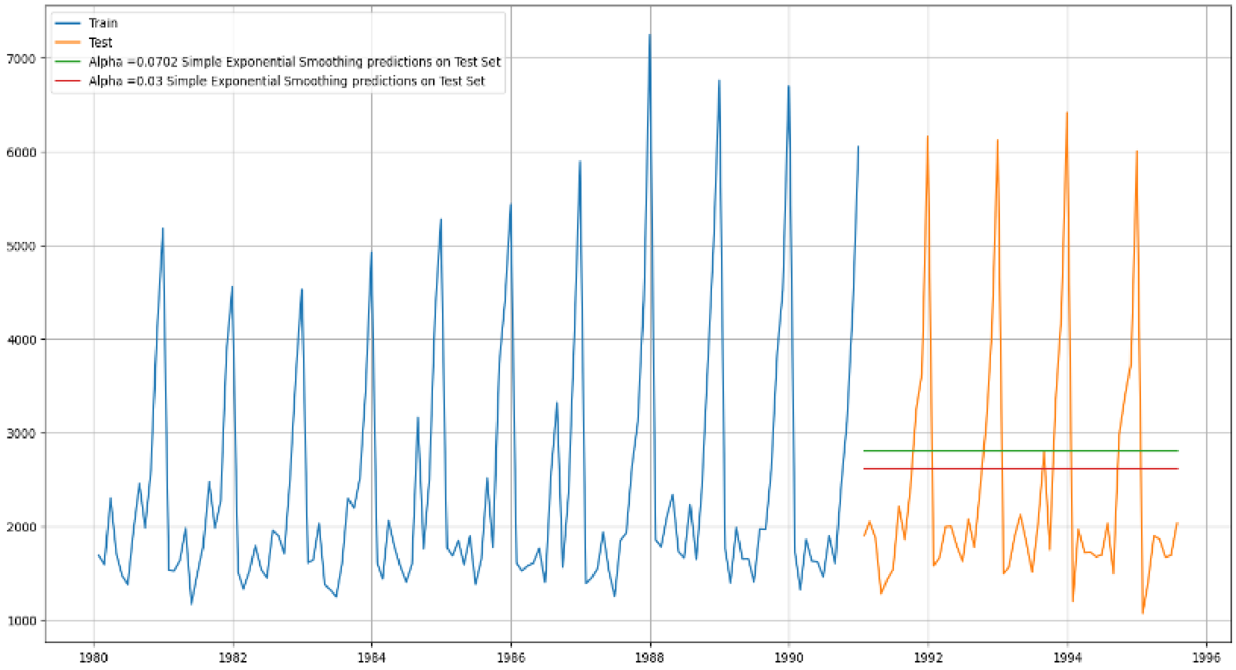


<u>Observations:</u>

- We can observe that there is no presence of trend or seasonality on the predicted sales.
- RMSE on the test data for Simple exponential smoothing model is 1338.01.

We can run a loop in order to find the alpha value that works for best on the test data.

Figure 12: SES model on train and test data (Alpha=0.03)



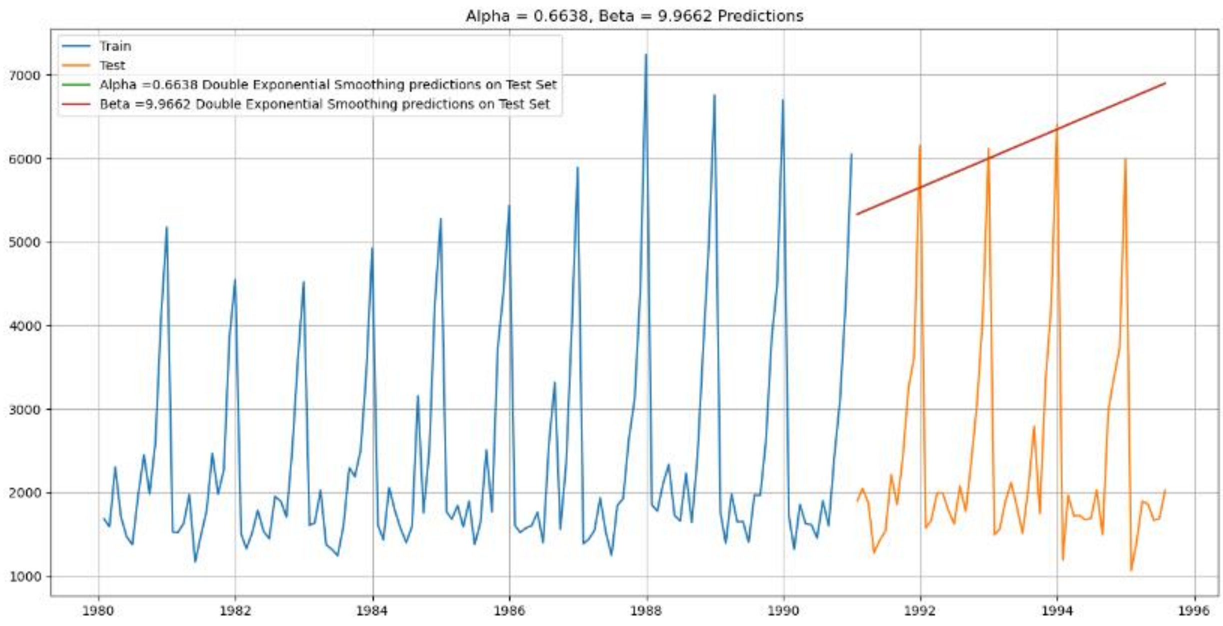
Observation:

- The Alpha value 0.03 works on the test data well, as it has lower RMSE of 1292.57 among other alpha values.

e) Model 5: Double Exponential Smoothing:

- Double exponential smoothing model is used when the data has trend but no seasonality. However, in this case, since the data is having only seasonality, DES model may not be effective.
- Let's Initialize and Fit the DES model on the train set.
- We got a smoothing level of 0.6638 and smoothing trend of 9.9662 when we fit the model on the data.

Figure 13: DES model on train and test data



Observations:

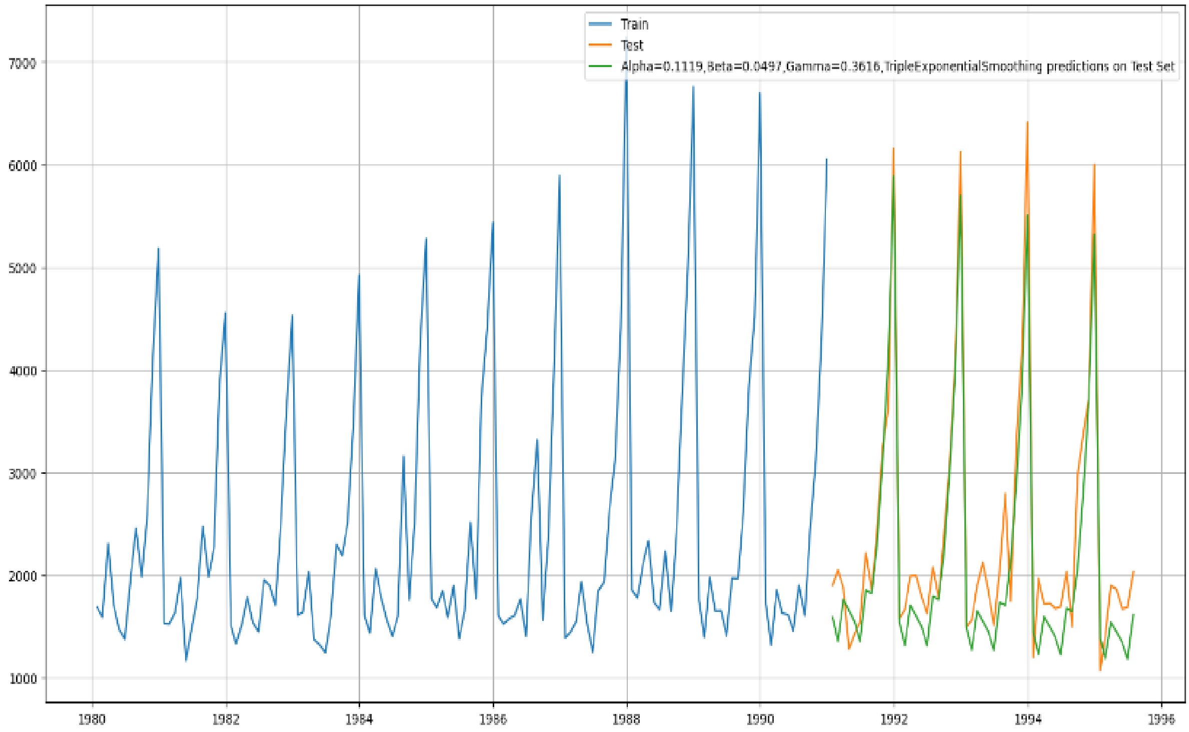
- We can observe that there is a presence of trend but without seasonality on the predicted values.
- RMSE on the test data for Double exponential smoothing model is 3950.

There are no other parameters i.e., Alpha and Beta with lesser RMSE of 3950.

f) Model 6: Triple Exponential Smoothing:

- Triple exponential smoothing model is used when the data consists of trend and seasonality. However, in this case, there is no trend but seasonality so TES model may not be highly effective.
- Let's Initialize and Fit the TES model on the train set.
- We got a smoothing level of 0.1119, smoothing trend of 0.0498 and smoothing seasonal of 0.3617 when we fit the model on the data.

Figure 14: TES model on train and test data (Alpha=0.1119, Beta=0.0498, Gamma=0.3617)

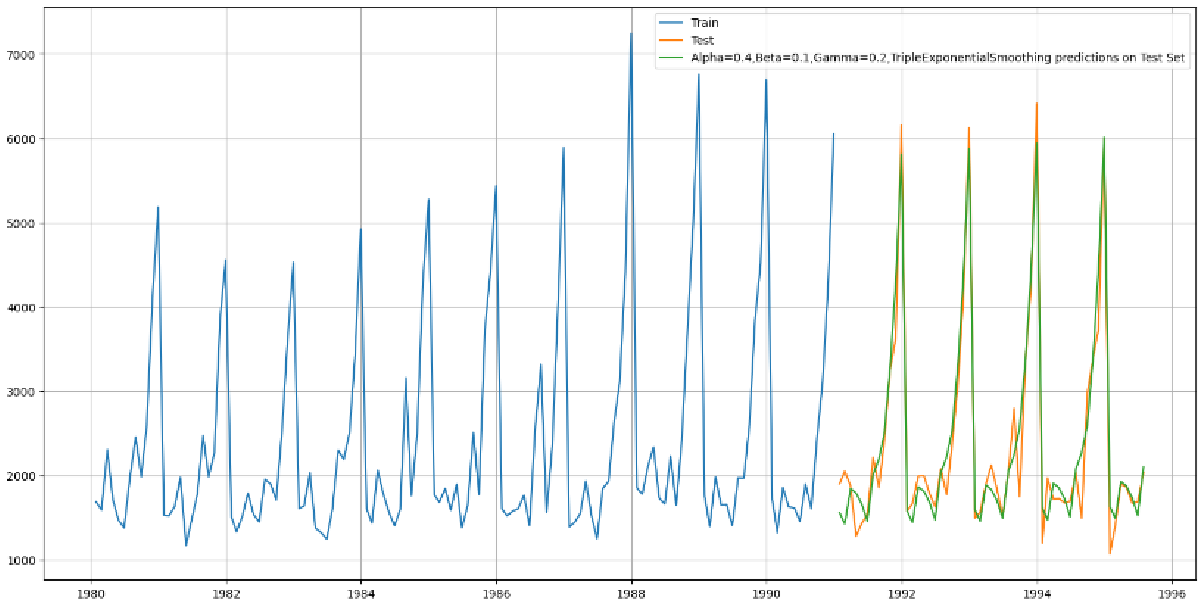


Observations:

- We can observe that there is no presence of trend but seasonality on the predicted sales.
- RMSE on the test data for Triple exponential smoothing model is 406.51.

We can run a loop in order to find the best alpha, beta and gamma values with lower RMSE value.

Figure 15: TES model on train and test data (Alpha=0.4, Beta=0.1, Gamma=0.2)



Observation:

- The Alpha, Beta and Gamma values with 0.4, 0.1 and 0.2 have lower RMSE of 317.43 and works well on the test data.

Conclusion:

Based on lowest RMSE value (317.43), we can opt Triple exponential smoothing model among all other models.

1.5. 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.

Ans:

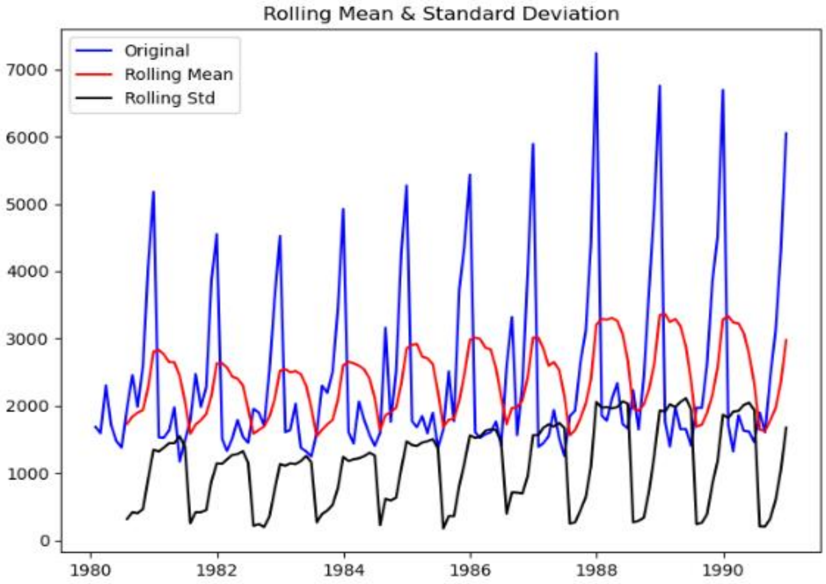
- In order to forecast and build models, time series data is expected to be stationary i.e., variance and correlation to remain constant.
- Augmented Dicky Fueller test is used to check stationarity of data, where null hypothesis states that time series data is not stationary. Null hypothesis will be rejected if the p-value is less than 5%.

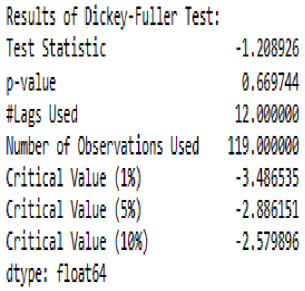
Null Hypothesis Ho: Time Series is non-stationary

Alternate Hypothesis Ha: Time Series is stationary

a) <u>ADF test on train data</u>:

Figure 16: ADF test on train set





```
Results of Dickey-Fuller Test:
Test Statistic                  -1.208926
p-value                          0.669744
#Lags Used                      12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
dtype: float64
```

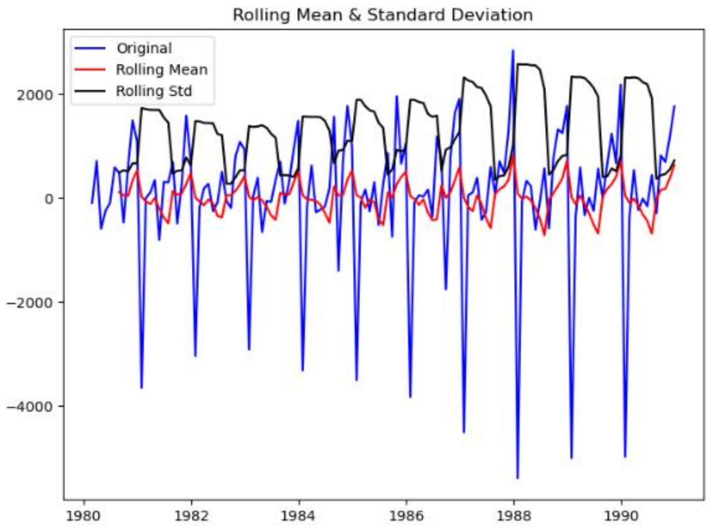
<u>Observations:</u>

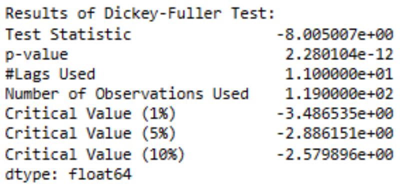
- P value 0.67 > 0.05, hence, we fail to reject Null Hypothesis i.e., time series is not stationary.
- Test statistic value is -1.21.

b) <u>Performing differencing (d=1) on original train data:</u>

- We are performing differencing (d=1) as the data is not stationary.
- It calculates the variations between current and next month.

Figure 17: Differencing (d=1) on train set




```
Results of Dickey-Fuller Test:
Test Statistic                  -8.005007e+00
p-value                          2.280104e-12
#Lags Used                       1.100000e+01
Number of Observations Used      1.190000e+02
Critical Value (1%)             -3.486535e+00
Critical Value (5%)             -2.886151e+00
Critical Value (10%)            -2.579896e+00
dtype: float64
```

Observations:

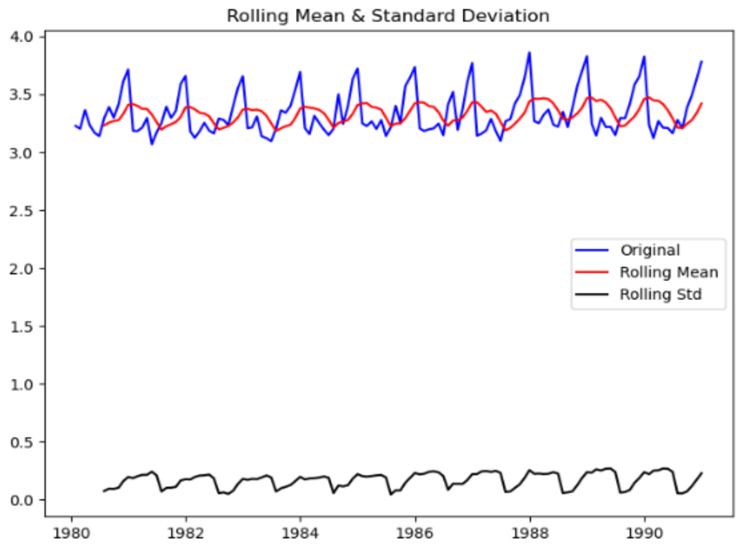
- Here, the p value is lesser than 5% significant level, hence, we can reject Null Hypothesis i.e., time series is stationary.
- Test statistic value is -8.005007e+00.
- We can observe the variation between months, of which, some are positive and negative values.

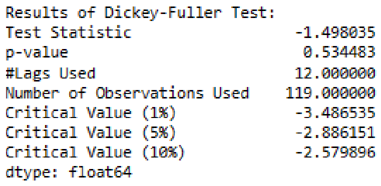
- We can also use **log method** to make time series stationary.

c) Log transformation of the data:

- Log method converts the multiplicative effect to additive.

Figure 18: Log transformation on train set



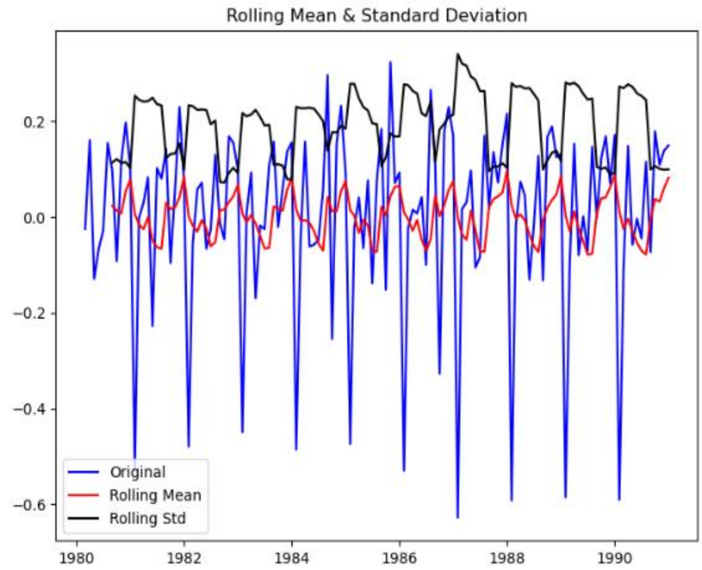

```
Results of Dickey-Fuller Test:
Test Statistic                   -1.498035
p-value                           0.534483
#Lags Used                       12.000000
Number of Observations Used     119.000000
Critical Value (1%)              -3.486535
Critical Value (5%)              -2.886151
Critical Value (10%)             -2.579896
dtype: float64
```

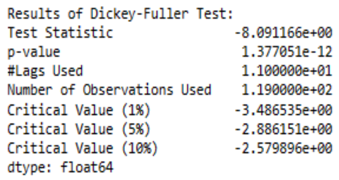
Observations:

- P value 0.53 > 0.05, hence, we fail to reject Null Hypothesis.
- Test statistic value is -1.50.
- Also, we can observe that log function to the base 10 has almost solved the variance issue.

d) Performing differencing (d=1) on log transformed data:

Figure 19: Differencing (d=1) on log transformed data





```
Results of Dickey-Fuller Test:
Test Statistic                 -8.091166e+00
p-value                         1.377051e-12
#Lags Used                      1.100000e+01
Number of Observations Used     1.190000e+02
Critical Value (1%)            -3.486535e+00
Critical Value (5%)            -2.886151e+00
Critical Value (10%)           -2.579896e+00
dtype: float64
```

Observations:

- Here, the p value is lesser than 5% significant level, hence, we can reject Null Hypothesis.
- Test statistic value is -8.091166e+00.
- We can observe very less variability and no trend.

Conclusion:

Time series data which was non-stationary has been made stationary by applying differencing (d=1) method on the data. The transformed data does not have any trend or variability, and is in a good form to take it further to build different models.

1.6. 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.

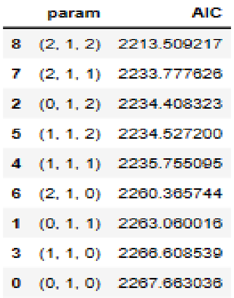
Ans:

We are choosing **ARIMA model** for the evaluation.

a) <u>ARIMA automated version (AIC):</u>

- For p and q values, we are choosing a range between (0,3)
- We are opting Differencing (d=1) method in order to make the data stationary.
- AIC values of the opted parameter combinations and sorting the AIC values in ascending order.

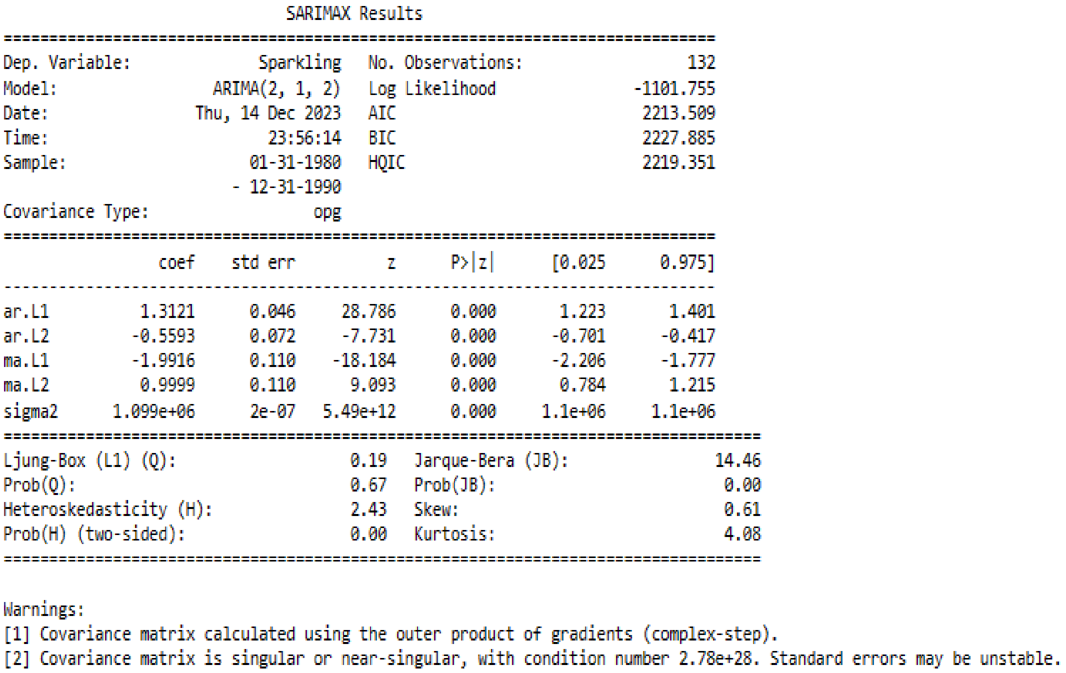
Table 4: AIC values



	param	AIC
8	(2, 1, 2)	2213.509217
7	(2, 1, 1)	2233.777626
2	(0, 1, 2)	2234.408323
5	(1, 1, 2)	2234.527200
4	(1, 1, 1)	2235.755095
6	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
3	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

- Based on the minimum AIC value, we choose order (2,1,2) as parameter for the model.

Figure 20: SARIMAX results for ARIMA automated version (AIC)



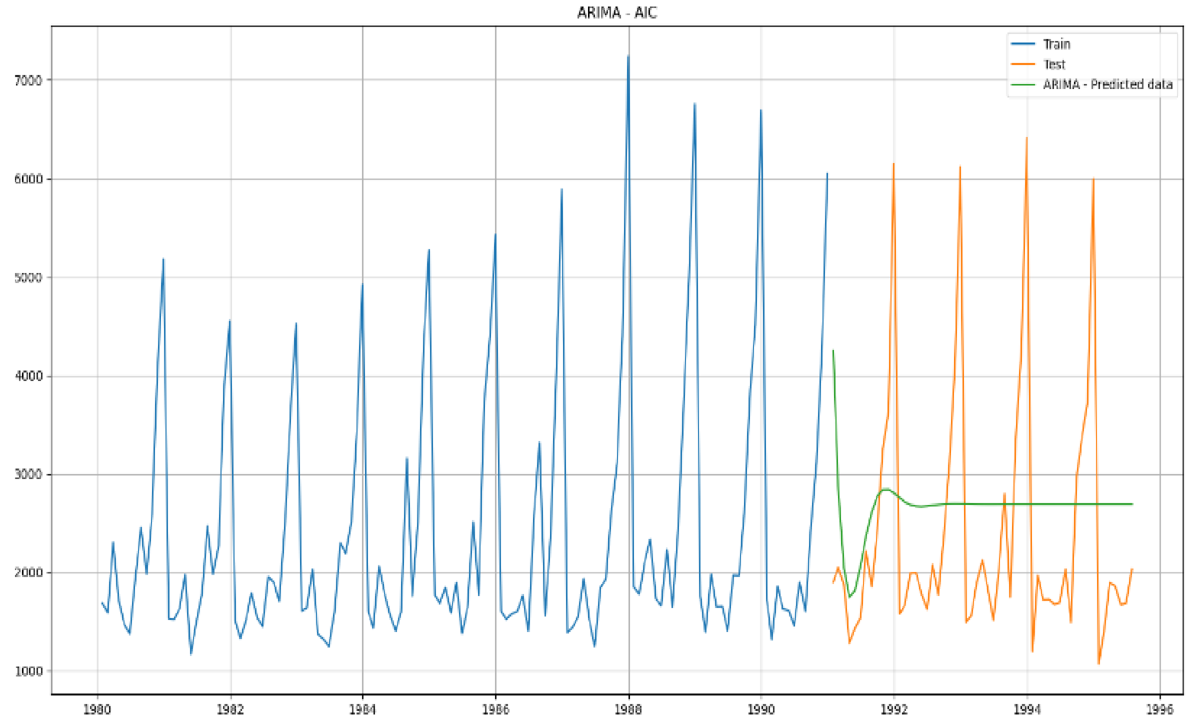
```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Thu, 14 Dec 2023   AIC                           2213.509
Time:                        23:56:14   BIC                           2227.885
Sample:                    01-31-1980   HQIC                          2219.351
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.786      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.731      0.000      -0.701      -0.417
ma.L1         -1.9916      0.110    -18.184      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.093      0.000       0.784       1.215
sigma2      1.099e+06      2e-07   5.49e+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
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 2.78e+28. Standard errors may be unstable.
```

Observations:

- All the coefficients are significant as p values are less than 0.05.
- Prob (JB) is 0.00, so we can say that the data is not normal.
- The data is positively skewed (0.61).

- Now we can do the prediction on test data.

Figure 21: Predictions on test data ARIMA automated version (AIC)



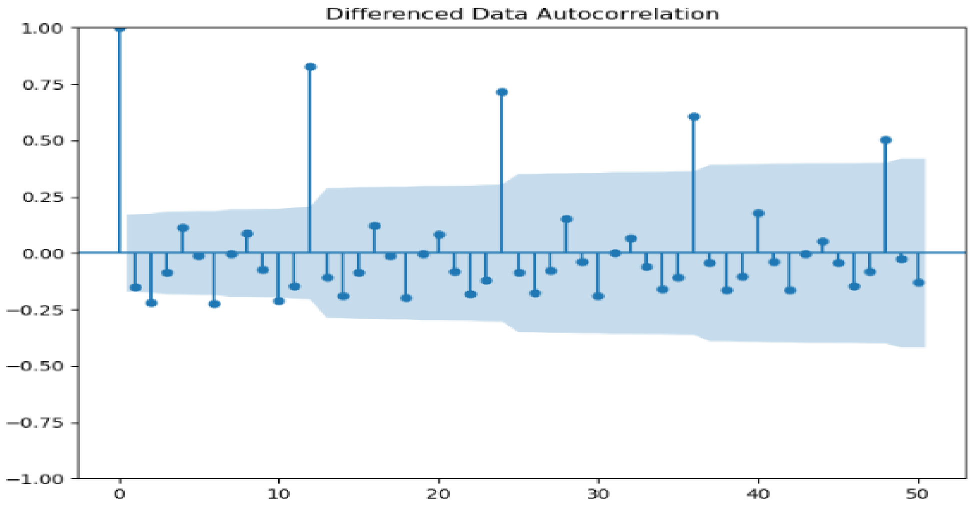
Observations:

- We can say that the model has not predicted the values correctly, as the seasonality is not considered in ARIMA model.
- RMSE on the test data for ARIMA automated version (AIC) model is 1299.98

b) ARIMA manual version (ACF and PACF):

- In this method, ‘p’ and ‘q’ parameters are determined based on the significant lag from the ACF and PACF plots.

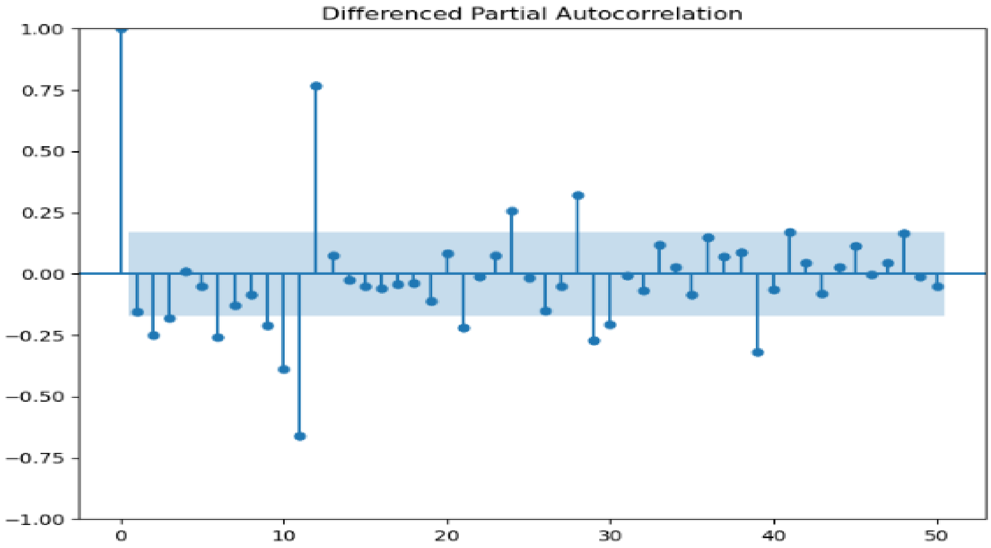
Figure 22: ACF



<u>Insights</u>**:**

- We have used differencing (d=1) method to make the data stationary.
- Moving Average (q) is considered from the ACF plot based on the significant lag.
- We can observe from the above plot that significant lag cuts-off at 0, i.e., q = 0.

Figure 23: PACF

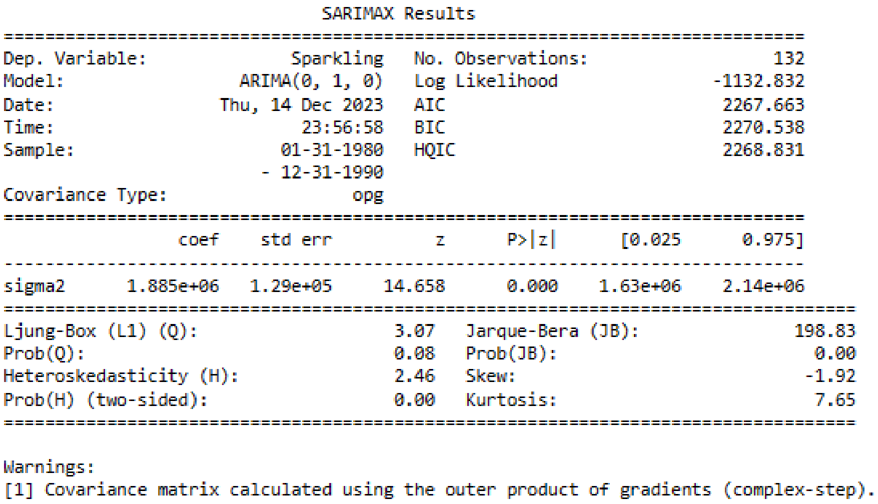


<u>Insights</u>**:**

- We have used differencing (d=1) method to make the data stationary.
- Auto Regression (p) is considered from the PACF plot based on the significant lag.
- We can observe from the above plot that significant lag cuts-off at 0, i.e., p = 0.

- Based on the above plots, we determine parameter of (0,1,0) for the ARIMA model.

Figure 24: SARIMAX results for ARIMA manual version (ACF and PACF)



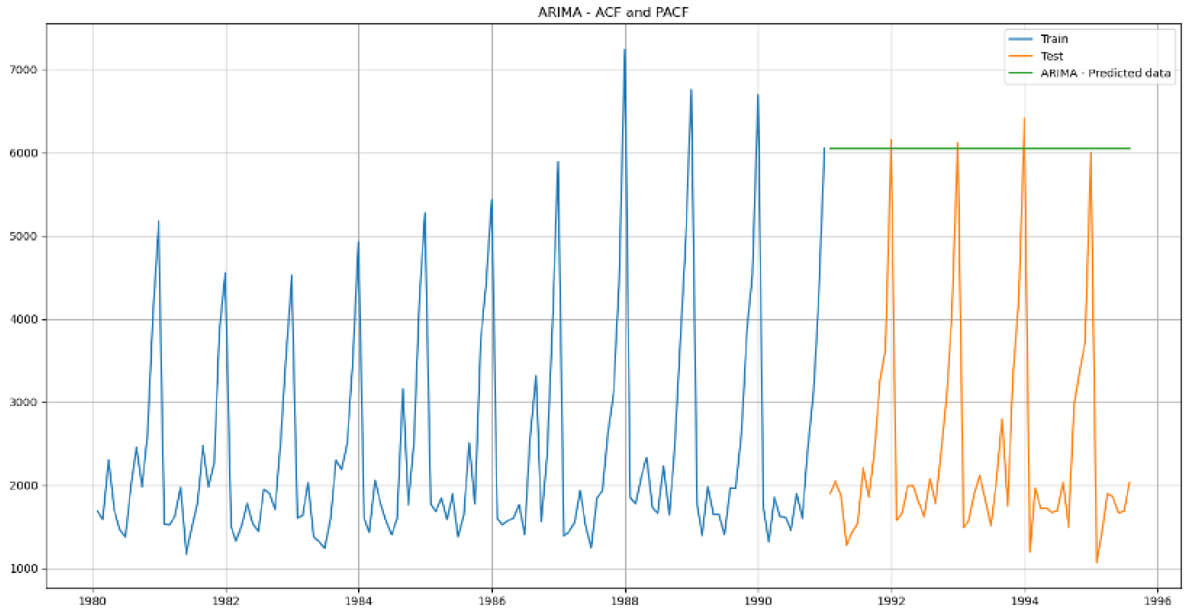
```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.832
Date:                Thu, 14 Dec 2023   AIC                           2267.663
Time:                        23:56:58   BIC                           2270.538
Sample:                    01-31-1980   HQIC                          2268.831
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      1.885e+06   1.29e+05     14.658      0.000    1.63e+06    2.14e+06
===================================================================================
Ljung-Box (L1) (Q):                   3.07   Jarque-Bera (JB):               198.83
Prob(Q):                              0.08   Prob(JB):                         0.00
Heteroskedasticity (H):               2.46   Skew:                            -1.92
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.65
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Observations:

- Coefficient is significant as p value is less than 0.05.
- Data is not normal as Prob (JB) is 0.00.
- The data is negatively skewed (-1.92).
- The model has log likelihood of -1132.832.

- Now we can do the prediction on test data.

Figure 25: Predictions on test data ARIMA manual version (ACF and PACF)



Observations:

- We can say that the model has not predicted the values correctly, as the seasonality is not considered in this model.
- RMSE on the test data for ARIMA (ACF and PACF) model is 3864.28

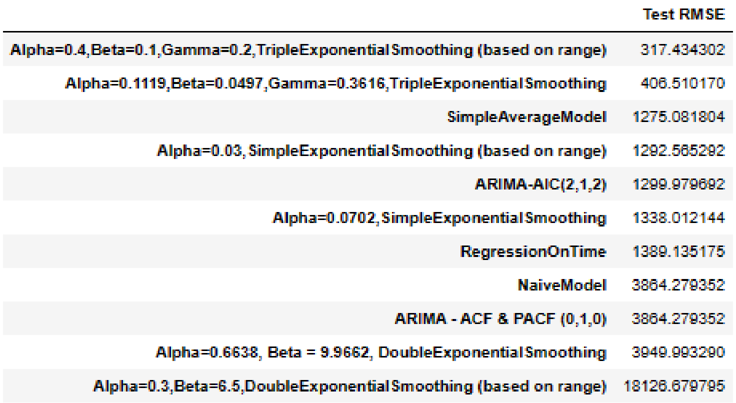
Conclusion:

Based on RMSE, automated ARIMA (AIC) model has lower RMSE 1299.98

1.7. 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.

Ans:

Table 5: Data frame showing built models with RMSE values



	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing (based on range)	317.434302
Alpha=0.1119,Beta=0.0497,Gamma=0.3616,TripleExponentialSmoothing	406.510170
SimpleAverageModel	1275.081804
Alpha=0.03, SimpleExponentialSmoothing (based on range)	1292.565292
ARIMA-AIC(2,1,2)	1299.979692
Alpha=0.0702, SimpleExponentialSmoothing	1338.012144
RegressionOnTime	1389.135175
NaiveModel	3864.279352
ARIMA - ACF & PACF (0,1,0)	3864.279352
Alpha=0.6638, Beta = 9.9662, DoubleExponentialSmoothing	3949.993290
Alpha=0.3,Beta=6.5,DoubleExponentialSmoothing (based on range)	18126.679795

Observations:

- The lowest RMSE value of all these models is 317.43 and for Triple exponential smoothing (Alpha = 0.4, Beta = 0.1, Gamma = 0.2) model, for which the parameters were derived based on the range.
- The highest RMSE value of all these models is 18126.68 and for Double exponential smoothing (Alpha = 0.3, Beta = 6.5) model, for which the parameters were derived based on the range.

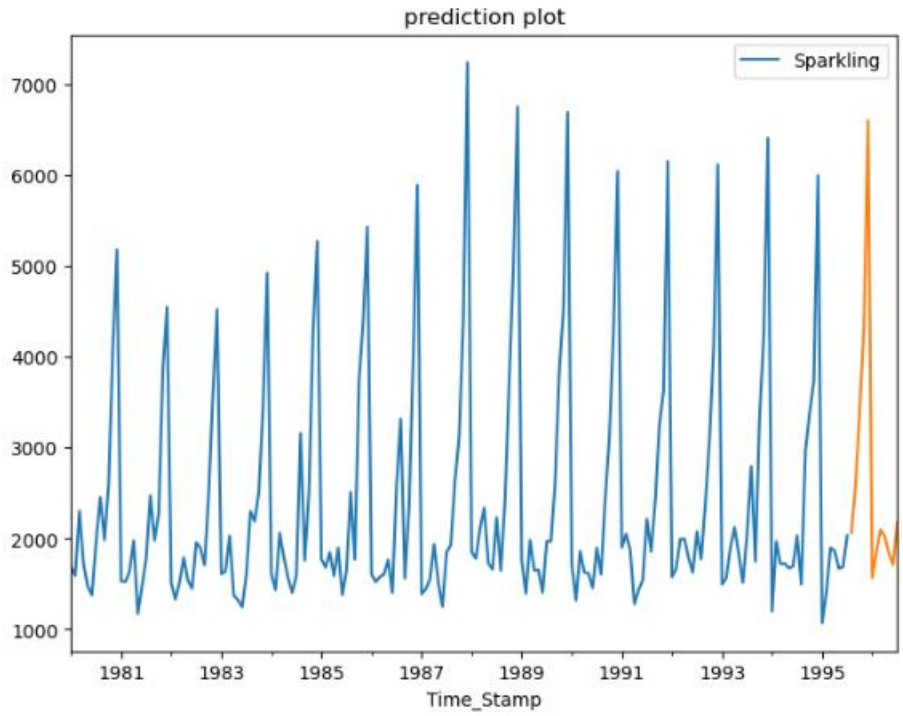
1.8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Ans:

- Based on the RMSE value, we opt Triple exponential smoothing iterative model with Alpha = 0.4, Beta = 0.1 and Gamma = 0.2.
- Initialising and fitting the model on the complete data.

- Forecasting the data for 12 months.
- RMSE value of the complete data is 376.77
- Prediction on the test data, i.e., for future 12 months.

Figure 26: Predictions on test data (future 12 months)

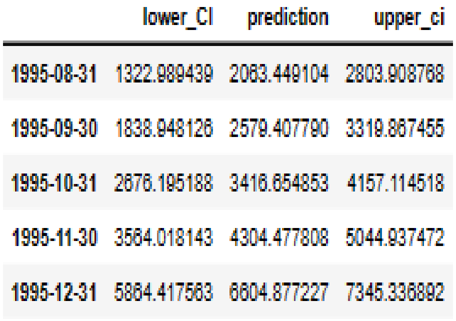


<u>Observation:</u>

- We can observe a slight increase in seasonality on the above plot.

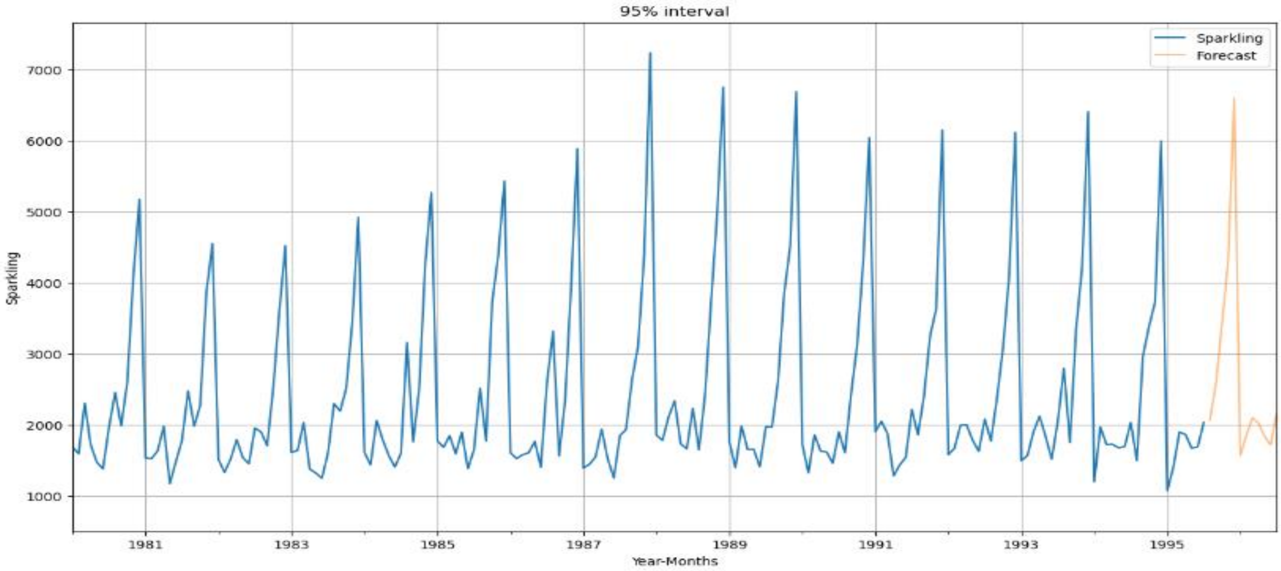
- We can calculate the lower and upper confidence bands at 95% confidence interval.

Table 6: 95% confidence interval



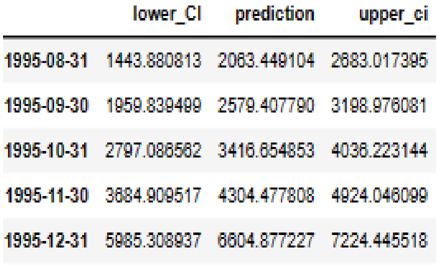
	lower_CI	prediction	upper_ci
1995-08-31	1322.989439	2063.449104	2803.908768
1995-09-30	1838.948126	2579.407790	3319.867455
1995-10-31	2676.195188	3416.654853	4157.114518
1995-11-30	3564.018143	4304.477808	5044.937472
1995-12-31	5864.417563	6604.877227	7345.336892

Figure 27: Plot showing 95% confidence interval



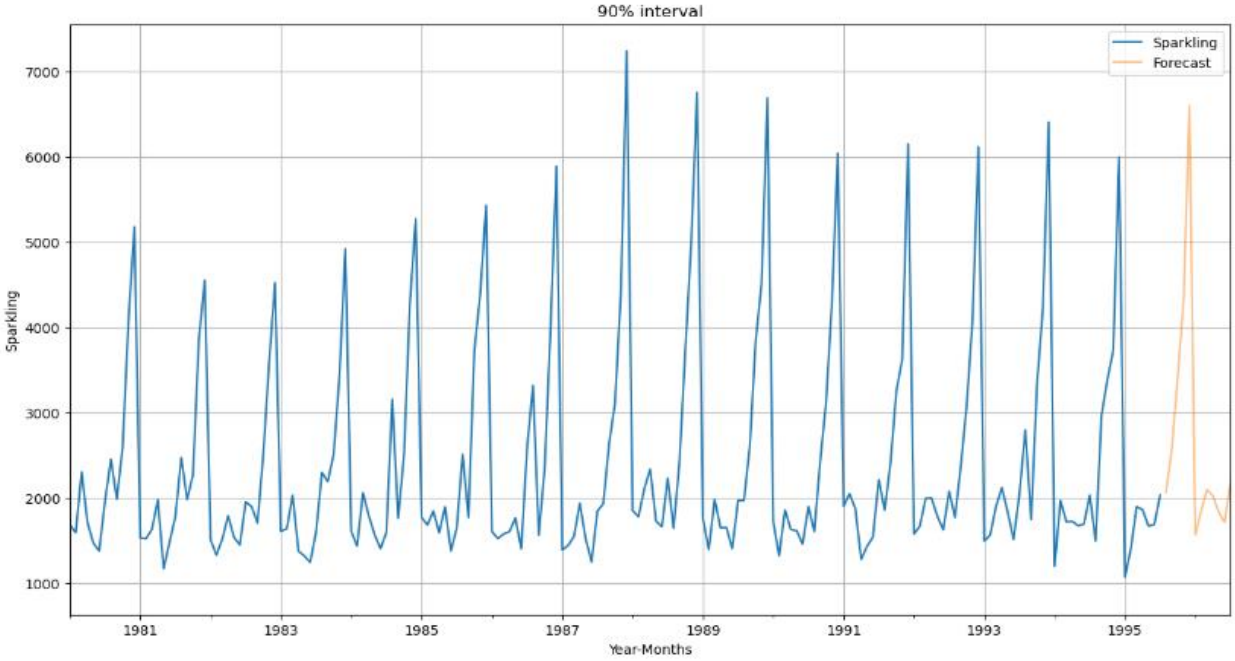
- We can calculate the lower and upper confidence bands at 90% confidence interval.

Table 7: 90% confidence interval



	lower_CI	prediction	upper_ci
1995-08-31	1443.880813	2063.449104	2683.017395
1995-09-30	1959.839499	2579.407790	3198.976081
1995-10-31	2797.086562	3416.654853	4036.223144
1995-11-30	3684.909517	4304.477808	4924.046099
1995-12-31	5985.308937	6604.877227	7224.445518

Figure 28: Plot showing 90% confidence interval



1.9. 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.

Ans:

Insights and Findings:

- We have chosen Triple exponential smoothing iterative model as an optimum model based on the lower RMSE value.
- Mean sales for the predicted year is 2685.65, which was 2402.42 in previous years.
- Median sales have increased from 1874 (previous years) to 2081.16 (predicted year).
- Minimum sale in a month has increased from 1070 to 1564.54 in the predicted year.
- We can observe an increase in sales in the last quarter of a year. We can expect a maximum sale of 6605 in December and there would be a sudden drop in sales in January.
- We can observe a slight increase in seasonality and there is no presence of trend on the predicted year.

Suggestions and Recommendations:

- One advantage for the company is that there is no decrease in trend as it is stable, but they need to adopt suitable measures in order to increase the trend over the coming years.
- Company is expected to have almost 10% more sales in next December, than last December, so it should be well prepared to match the demand of the customers.
- Company can introduce discounts, give away vouchers, gift coupons in order to attract more customers and boost sales for the period (Jan to July), as the sales are very low during this period.
- Company should have a proper mission and vision for the upcoming years.
- From the past years data, we can observe that there is no presence of trend, so company should improve promotional strategies so as to create demand for the product.

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.

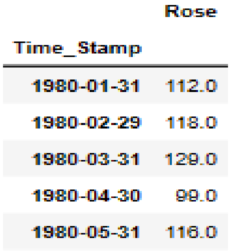
Rose data:

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

Ans:

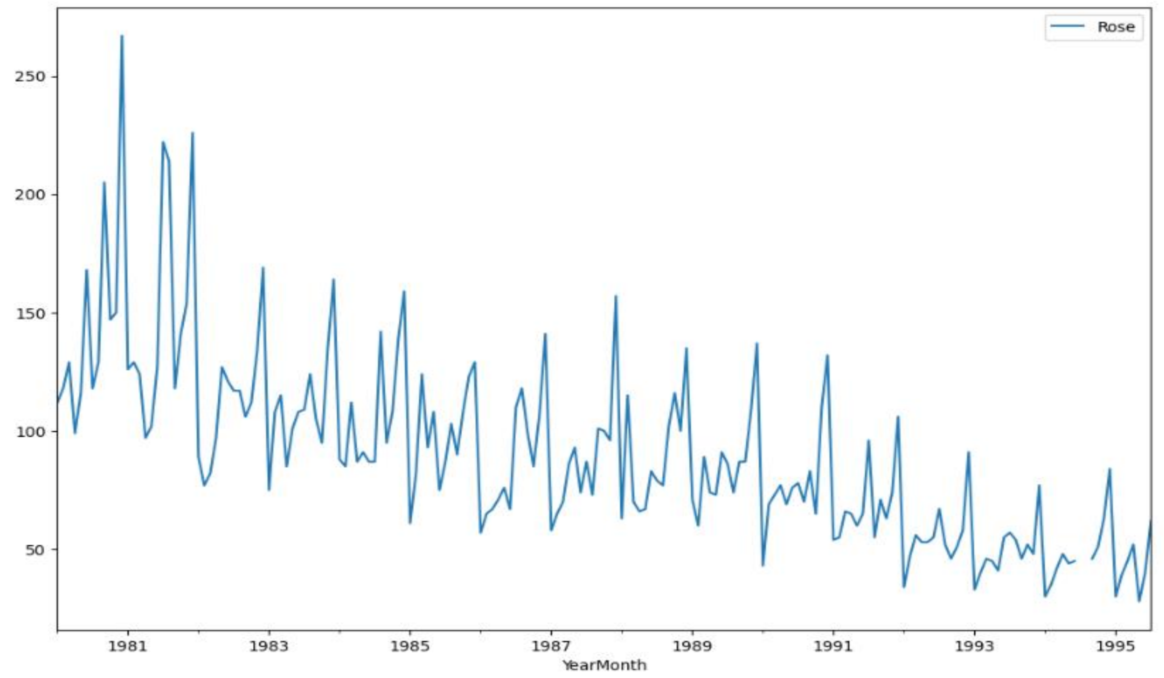
- There are 187 rows and 2 columns in the given dataset.
- Creating a timestamp column and adding it to the dataframe.
- Removing 'YearMonth' column and updating the timestamp as index.

Table 8: Top 5 rows of the dataset



	Rose
Time_Stamp	
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Figure 29: Plotting the time series data



Insights:

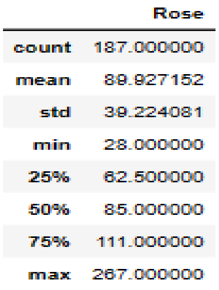
- The above plot has a downward trend.

- There is no presence of seasonality in the above time series data.

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

Ans:

Table 9: Statistical summary of the dataset

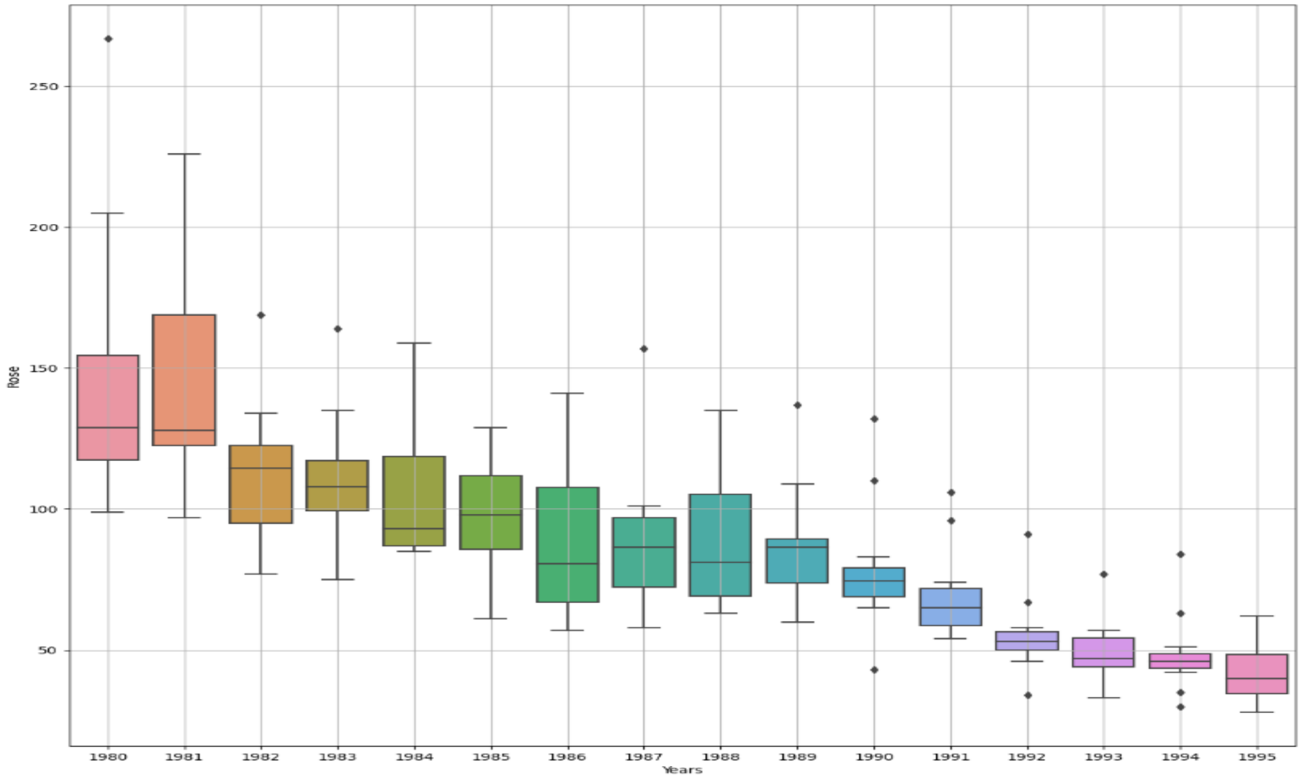


	Rose
count	187.000000
mean	89.927152
std	39.224081
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Observations:

- Mean sales of Rose wine is 90.
- Minimum and maximum sale is 28 and 267.

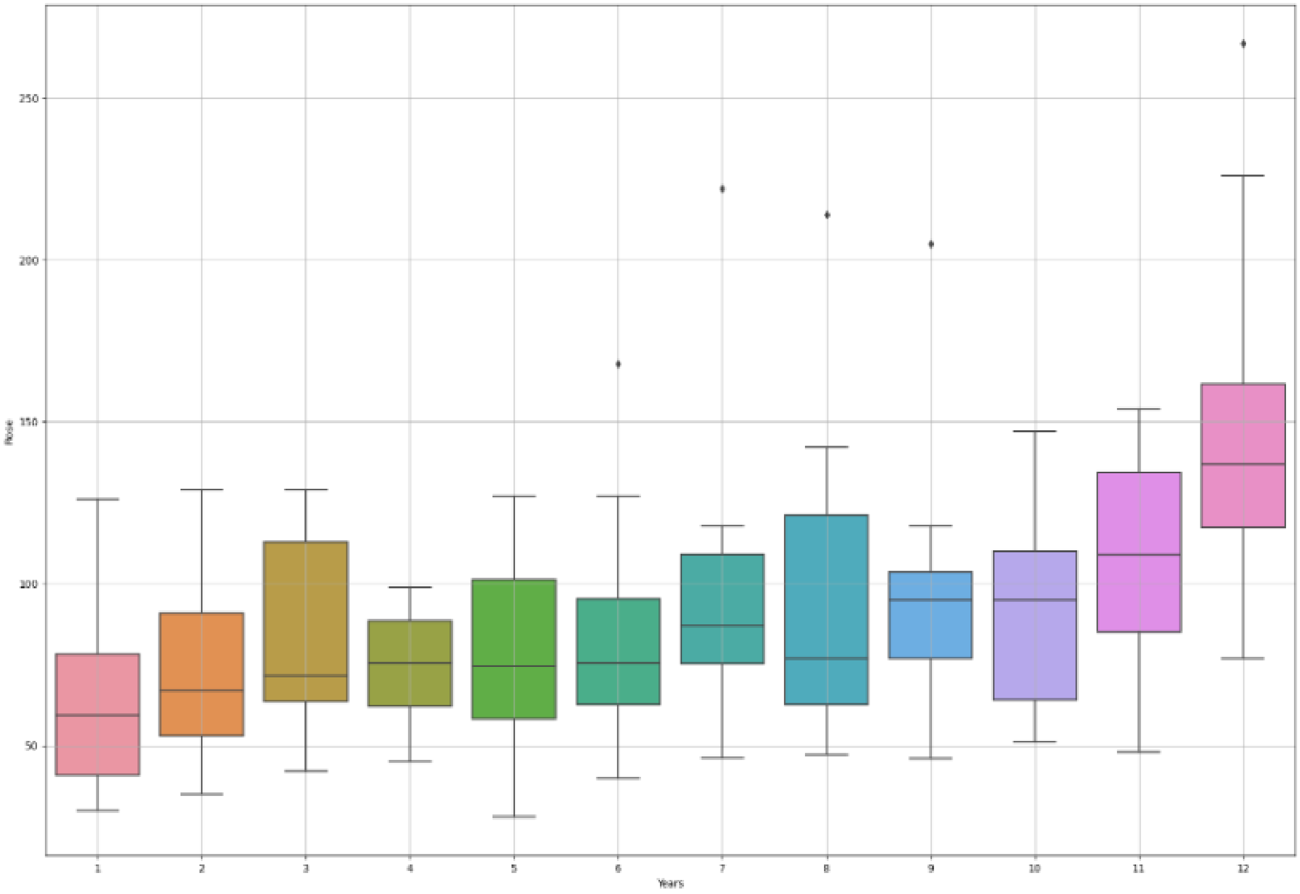
Figure 30: Yearly BoxPlot



Observations:

- We can observe outliers in the given dataset.
- There is downward in trend across years.

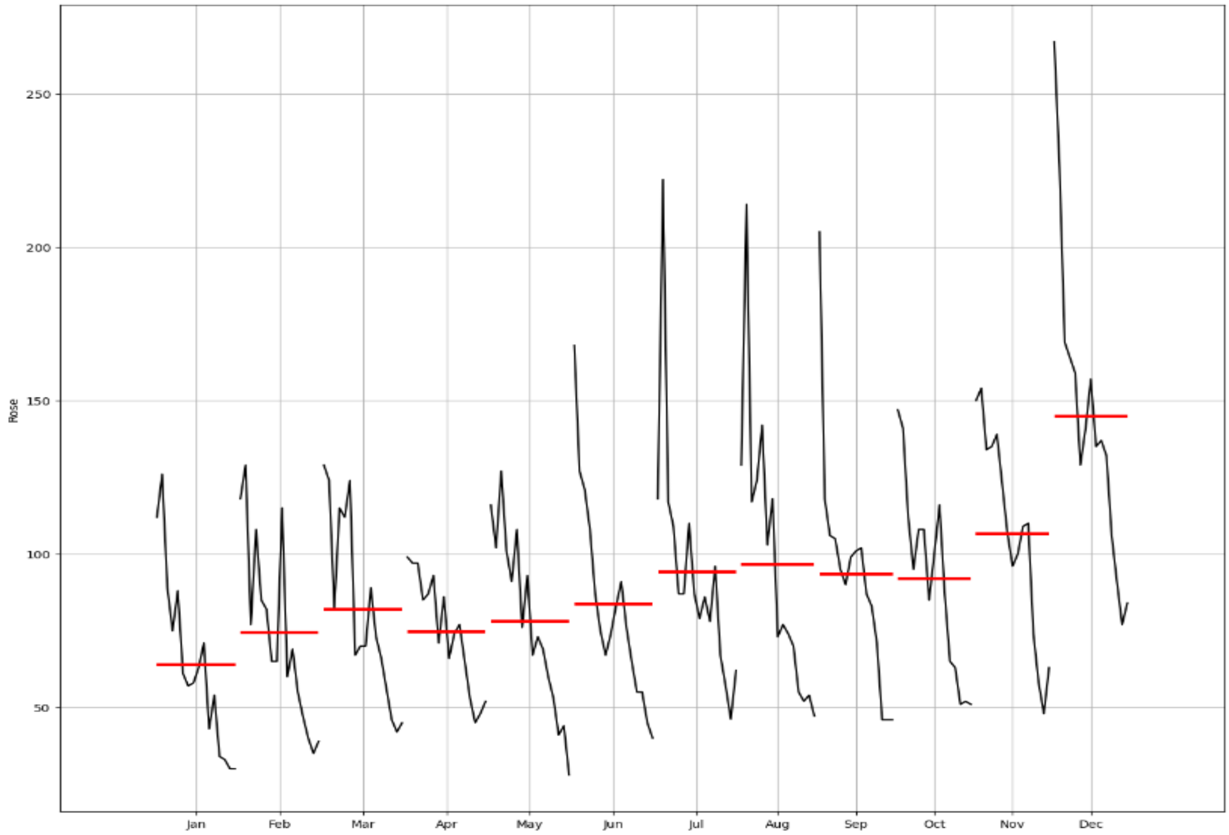
Figure 31: Monthly BoxPlot



Observation:

- Maximum sale is observed in the month of December, followed by November.

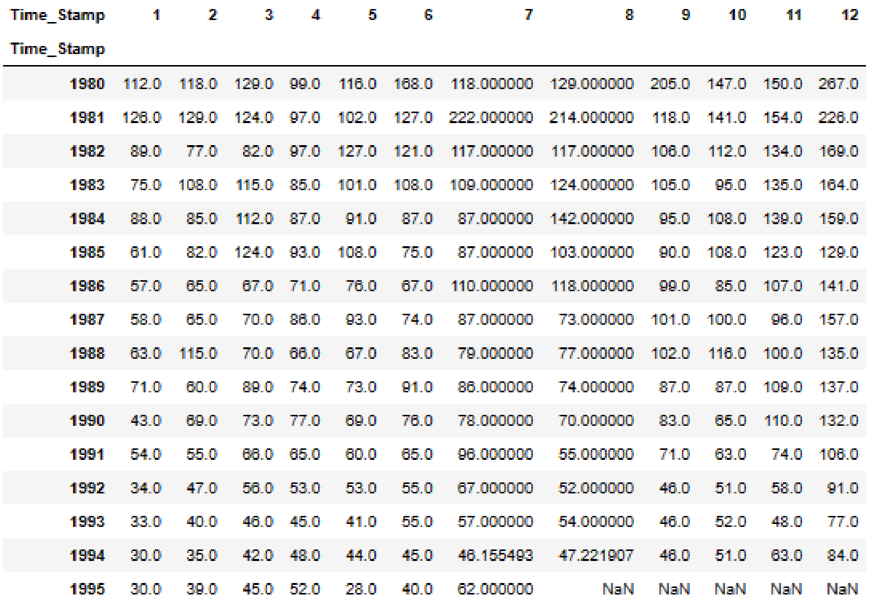
Figure 32: MonthPlot



Observation:

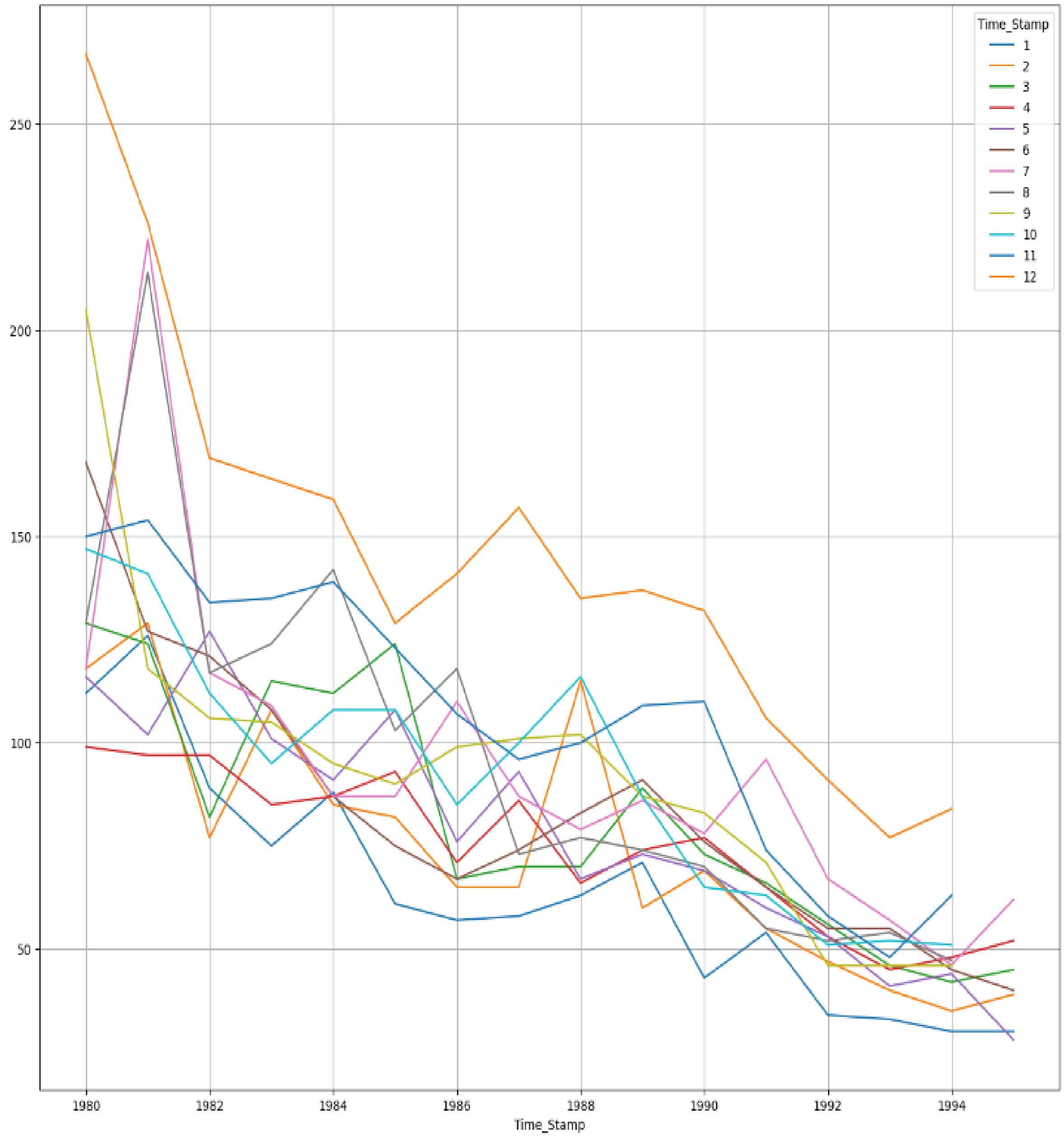
- January sees lowest sales across years and peak sales can be observed in the last quarter.

Table 10: Monthly wine sales across years



Time_Stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_Stamp												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.000000	124.000000	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.000000	142.000000	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.000000	103.000000	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.000000	118.000000	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.000000	73.000000	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.000000	77.000000	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.000000	74.000000	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.000000	70.000000	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.000000	55.000000	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.000000	52.000000	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.000000	54.000000	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	46.155493	47.221907	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.000000	NaN	NaN	NaN	NaN	NaN

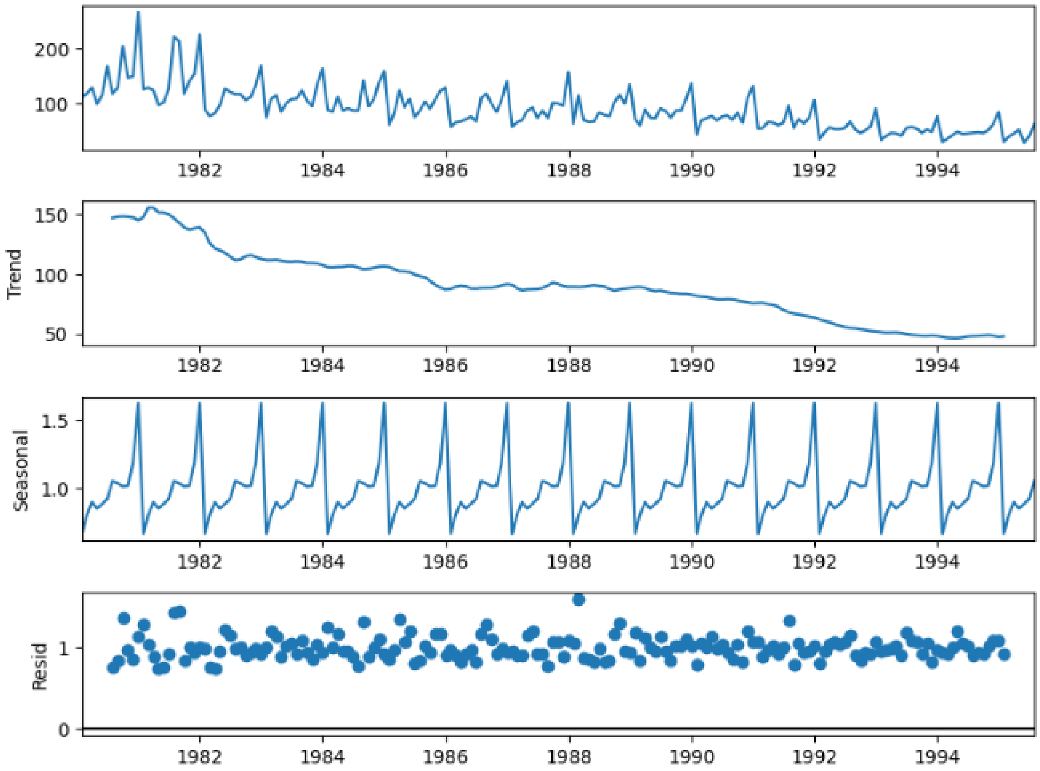
Figure 33: Plot showing monthly wine sales across years



Observations:

- December month shows the maximum sales across years.
- 1980 December has the highest and 1995 May has the lowest sales in Rose wine.

Figure 34: Decomposition with multiplicative seasonality



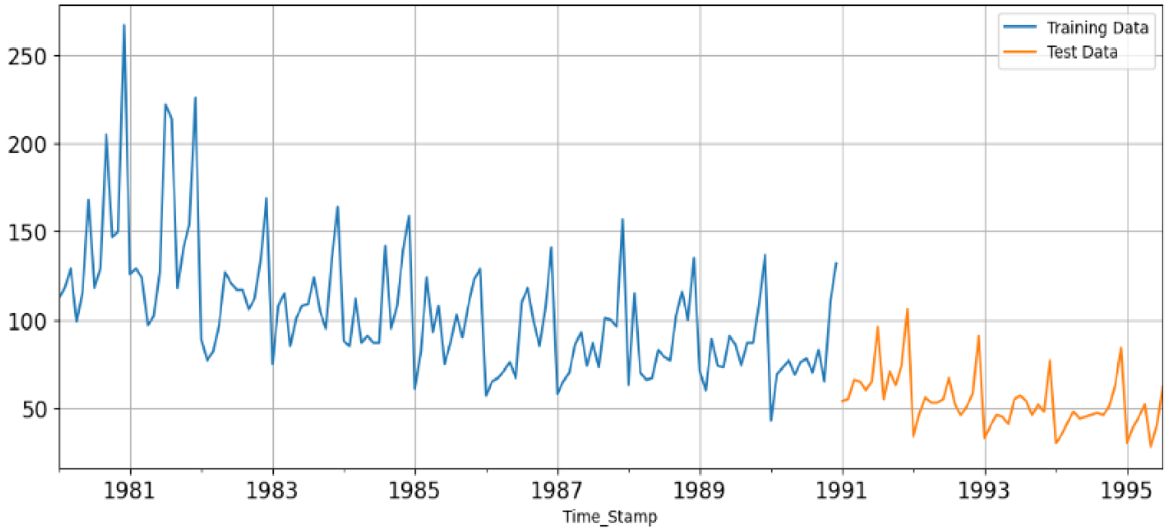
Observations:

- There is no particular pattern of residuals in multiplicative seasonality.
- There is both trend and seasonality in the given data set, i.e., there is a downward trend in the sales across years.

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

Ans:

Figure 35: Train and Test data



Observations:

- Test data is considered from 1991 onwards.
- There are 132 and 55 values in the train and test set.

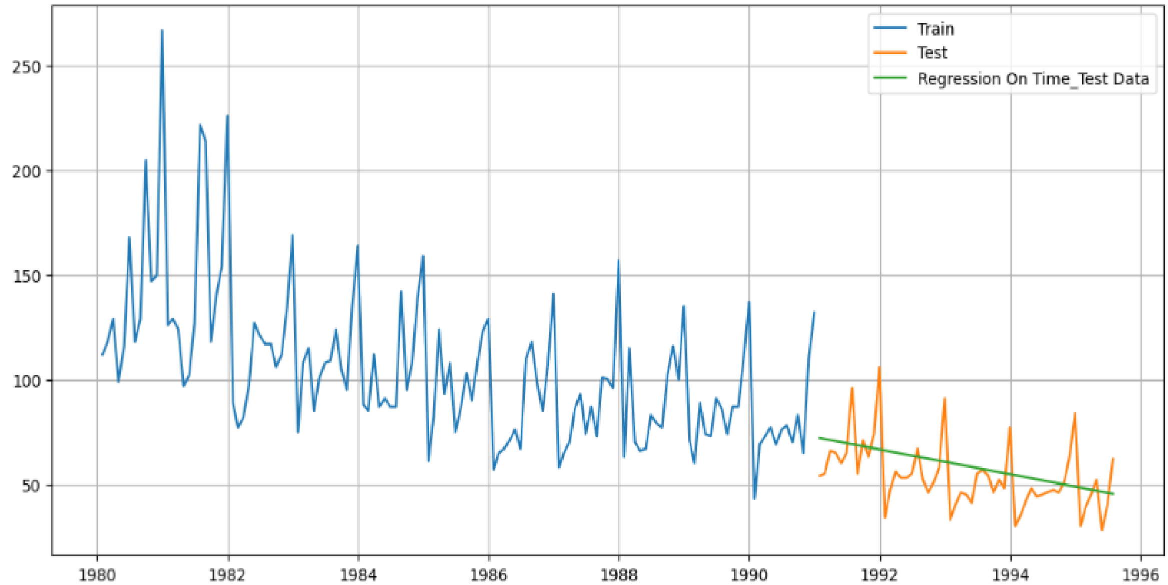
2.4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Ans:

a) Model 1: Linear Regression:

- We are generating the numerical time instance and adding the same to the train and test set.
- Initializing and Fitting the Linear regression model to the train set.

Figure 36: Linear regression model on train and test data



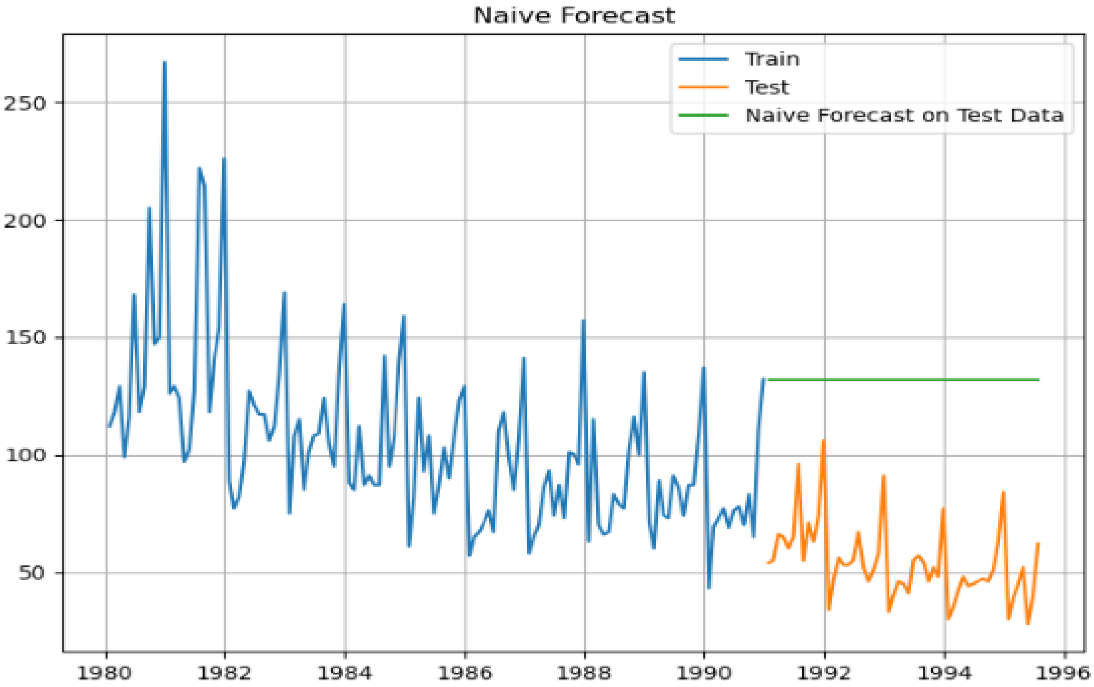
Observations:

- The above plot shows a decrease in trend on the predicted values.
- RMSE on the test data for Linear Regression is 15.26.

b) Model 2: Naïve Forecast:

- Naïve model focuses on the latest value to predict future sales.

Figure 37: Naïve forecast model on train and test data



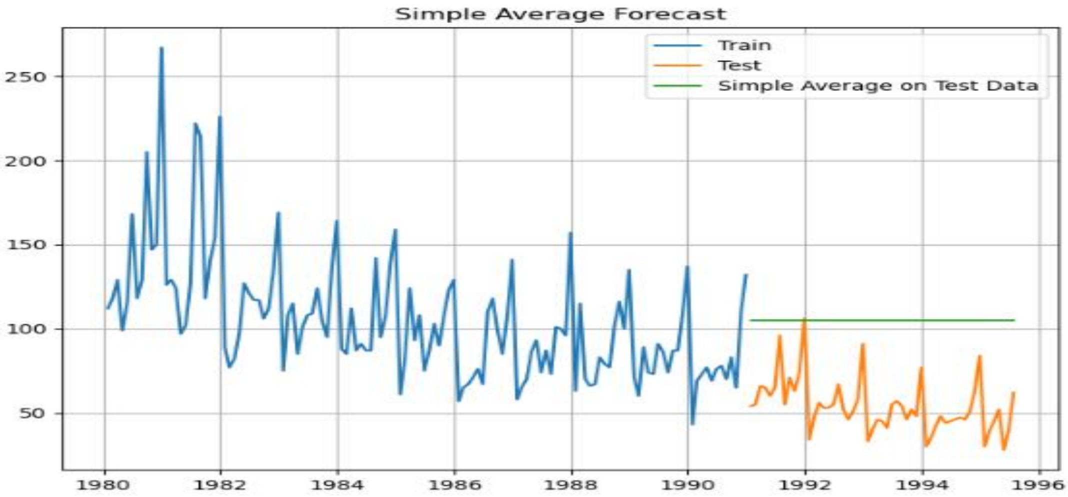
Observations:

- The above plot shows a flat trend, as it predicts values based on the latest value.
- RMSE on the test data for Naïve forecast model is 79.67.

c) Model 3: Simple Average:

- Simple Average method uses the mean of the training values to predict future sales.

Figure 38: Simple average model on train and test data



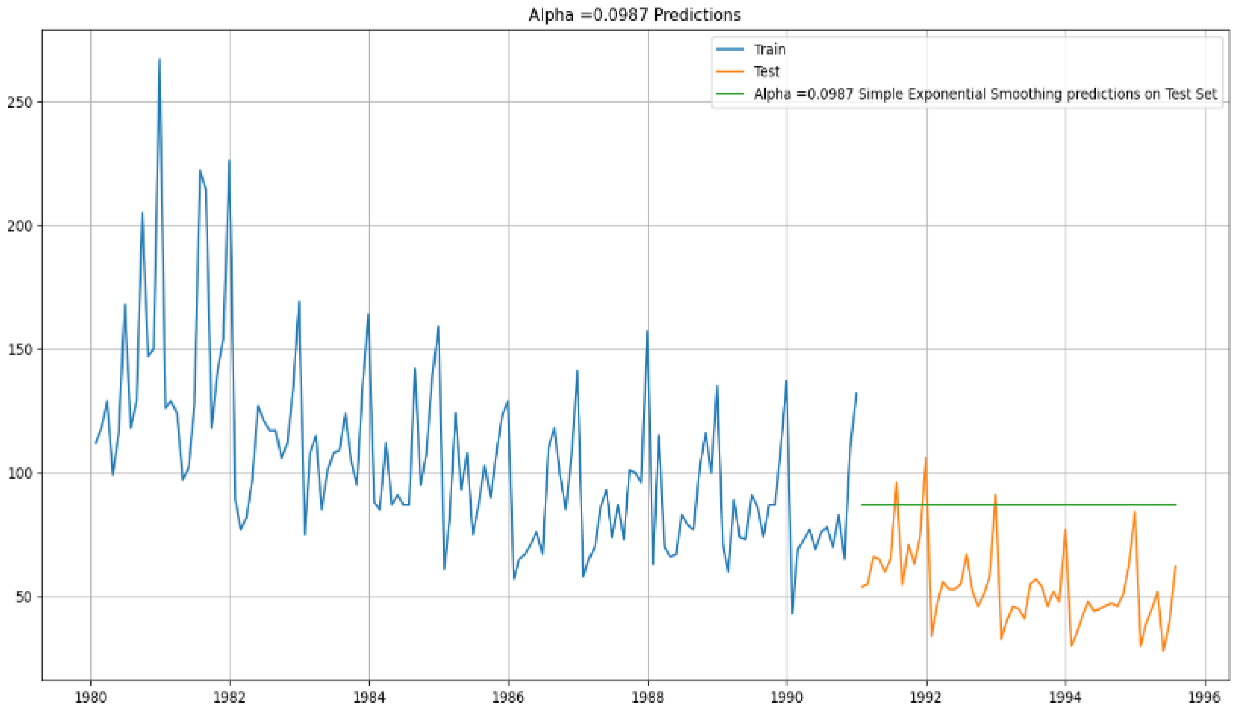
Observations:

- Simple average model does not show any trend, as it uses the mean of training values.
- RMSE on the test data for Simple Average model is 53.41.

d) Model 4: Simple Exponential Smoothing:

- Simple exponential smoothing model is generally used when the data does not have any trend or seasonality. However, in this case, there are both trend and seasonality, so SES model may not be effective.
- Let's Initialize and Fit the SES model on the train set.
- We got a smoothing level of 0.0987 when we fit the model on the data.

Figure 39: SES model on train and test data (Alpha=0.0987)

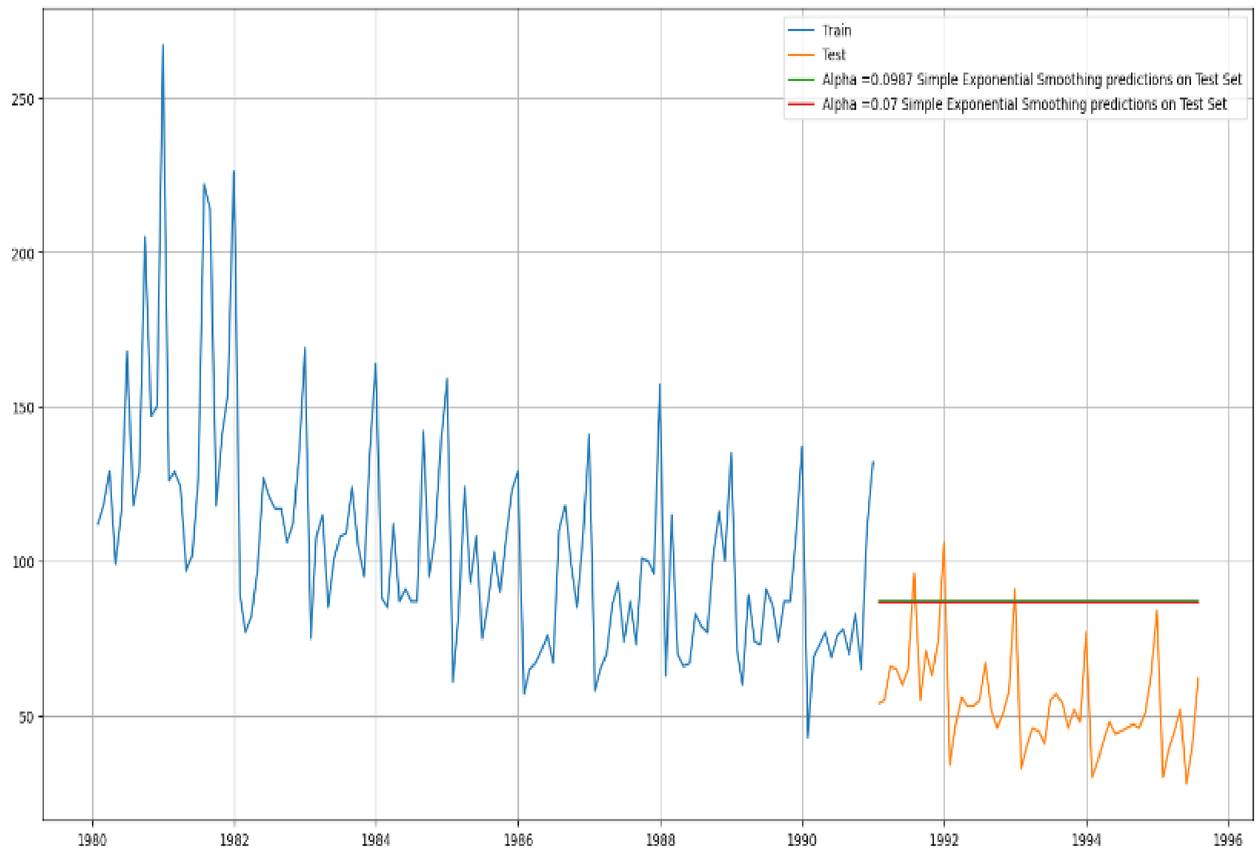


Observations:

- We can observe that there is no presence of trend or seasonality on the predicted sales.
- RMSE on the test data for Simple exponential smoothing model is 36.75.

- We can run a loop in order to find the alpha value that works for best on the test data.

Figure 40: SES model on train and test data (Alpha=0.07)



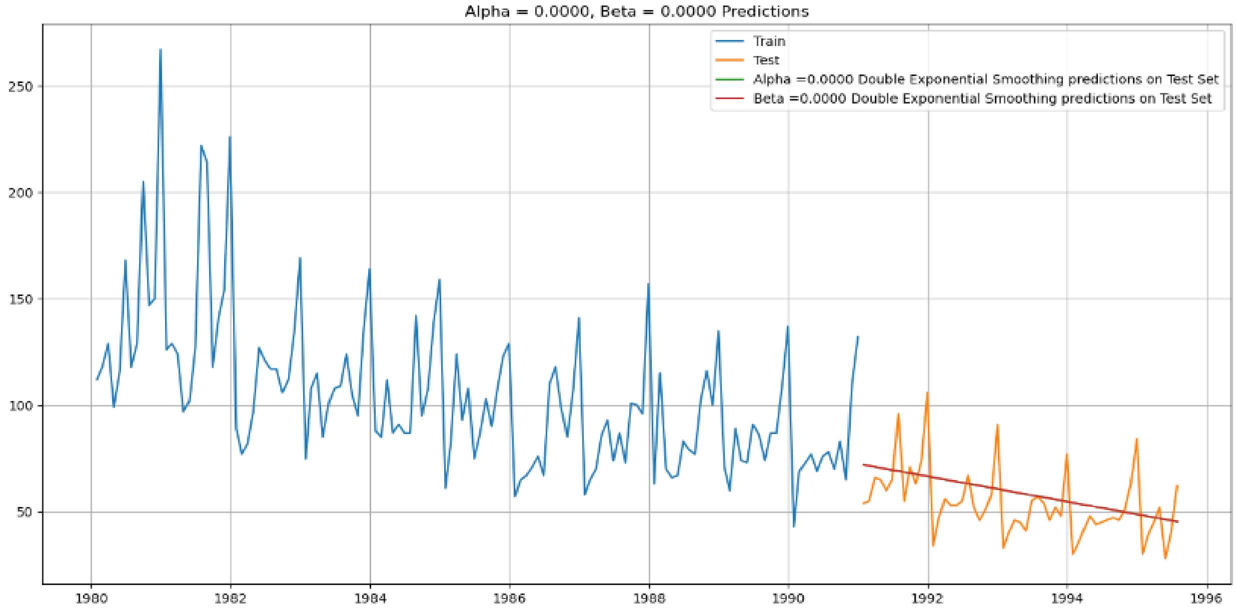
Observation:

- The Alpha value 0.07 works on the test data well, as it has lower RMSE of 36.38 among other alpha values. However, there is just a decimal difference between the RMSE of alpha values.

e) Model 5: Double Exponential Smoothing:

- Double exponential smoothing model is used when the data has trend but no seasonality. However, in this case, since the data have both trend and seasonality, thus, DES model may not be effective.
- Let's Initialize and Fit the DES model on the train set.
- We got a smoothing level of 0.0000 and smoothing trend of 0.0000 when we fit the model on the data.

Figure 41: DES model on train and test data



<u>Observations:</u>

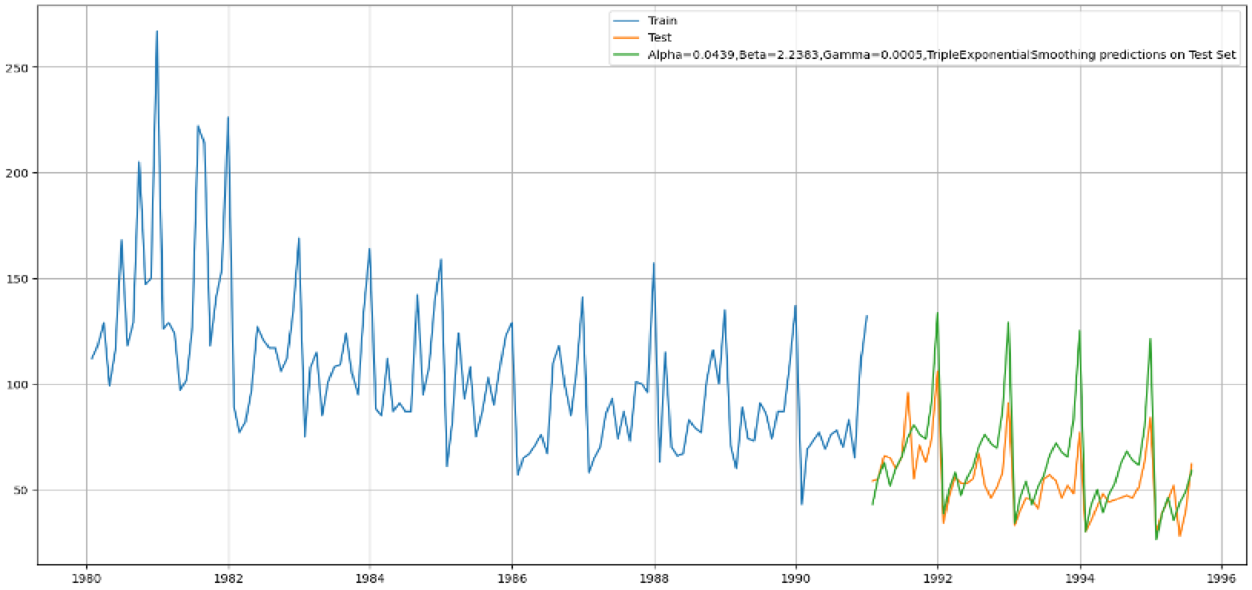
- We can observe that there is a presence of trend but without seasonality on the predicted values.
- RMSE on the test data for Double exponential smoothing model is 15.26.

There are no other parameters i.e., Alpha and Beta with lesser RMSE of 15.26.

f) <u>Model 6: Triple Exponential Smoothing:</u>

- Triple exponential smoothing model is used when the data consists of trend and seasonality. However, in this case, there are both trend and seasonality.
- Let’s Initialize and Fit the TES model on the train set.
- We got a smoothing level of 0.0439, smoothing trend of 2.2383 and smoothing seasonal of 0.0005 when we fit the model on the data.

Figure 42: TES model on train and test data (Alpha=0.0439, Beta=2.2383, Gamma=0.0005)



Observations:

- We can observe that there is presence of trend and seasonality on the predicted sales.
- RMSE on the test data for Triple exponential smoothing model is 16.63.

We can run a loop in order to find the best alpha, beta and gamma values with lower RMSE value. However, there are no other parameters i.e., Alpha and Beta with lesser RMSE of 16.63.

Conclusion:

Of all these models, based on the given criteria i.e., trend and seasonality, the best model we can opt is Triple exponential smoothing model with lowest RMSE of 16.63.

2.5. 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.

Ans:

- In order to forecast and build models, time series data is expected to be stationary i.e., variance and correlation to remain constant.

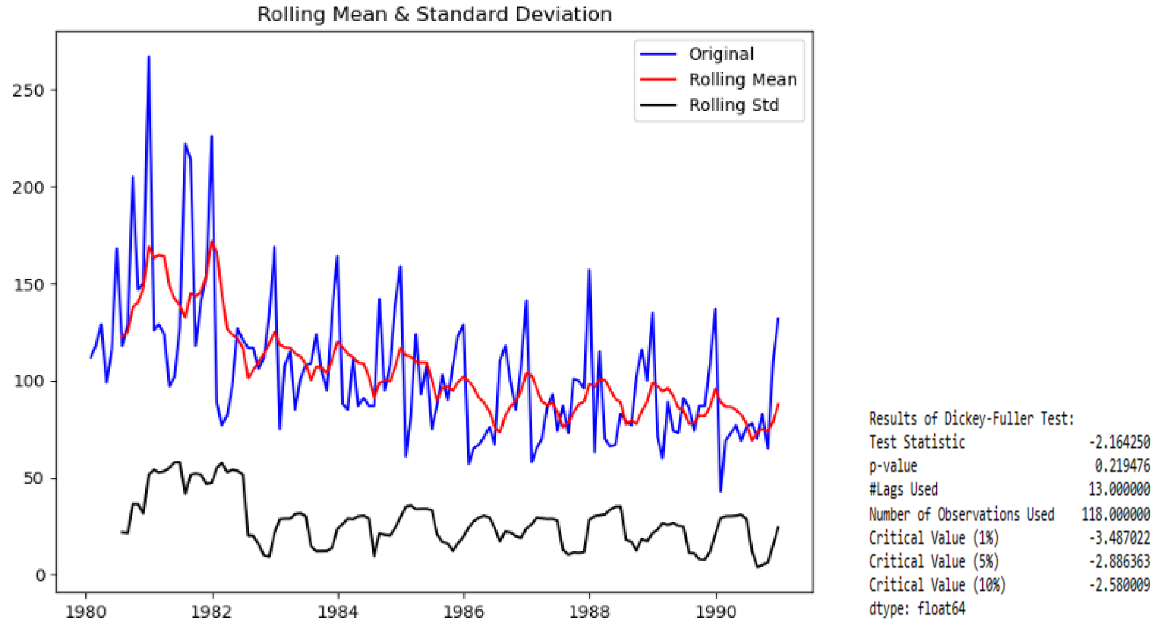
- Augmented Dicky Fueller test is used to check stationarity of data, where null hypothesis states that time series data is not stationary. Null hypothesis will be rejected if the p-value is less than 5%.

Null Hypothesis Ho: Time Series is non-stationary

Alternate Hypothesis Ha: Time Series is stationary

a) <u>ADF test on train data</u>:

Figure 43: ADF test on train set



```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

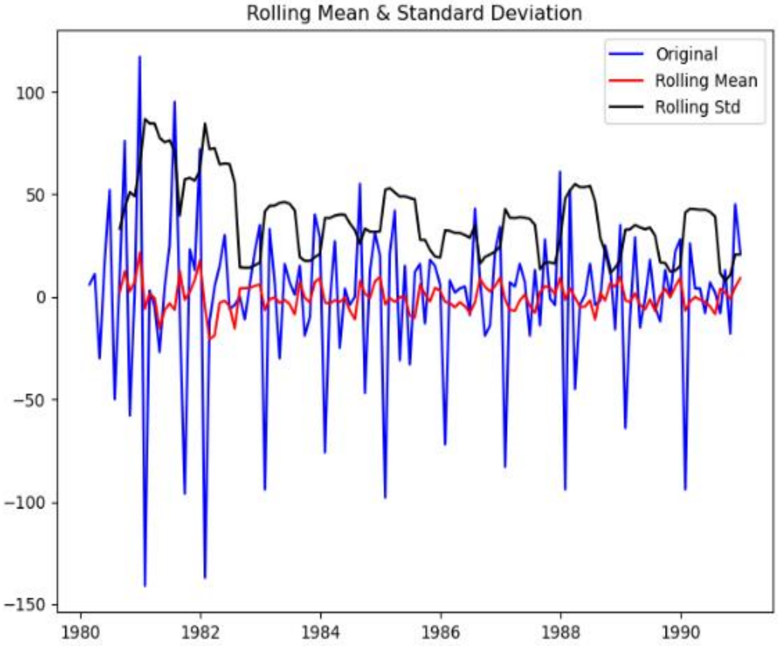
<u>Observations:</u>

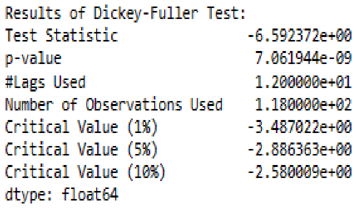
- P value 0.22 > 0.05, hence, we fail to reject Null Hypothesis i.e., time series is not stationary.
- Test statistic value is -2.16.
- We can observe a downward trend on given data.

b) <u>Performing differencing (d=1) on original train data:</u>

- We are performing differencing (d=1) as the data is not stationary.
- It calculates the variations between current and next month.

Figure 44: Differencing (d=1) on train set





```
Results of Dickey-Fuller Test:
Test Statistic                 -6.592372e+00
p-value                         7.061944e-09
#Lags Used                      1.200000e+01
Number of Observations Used     1.180000e+02
Critical Value (1%)            -3.487022e+00
Critical Value (5%)            -2.886363e+00
Critical Value (10%)           -2.580009e+00
dtype: float64
```

Observations:

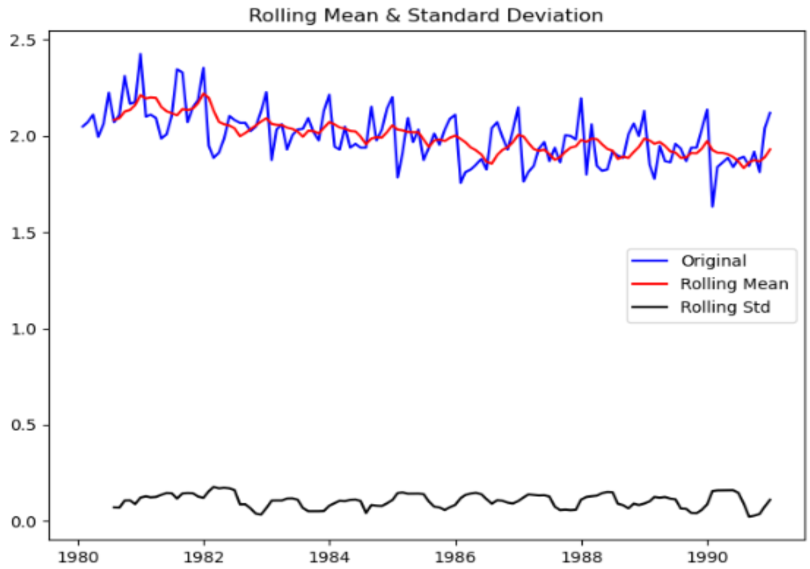
- Here, the p value is lesser than 5% significant level, hence, we can reject Null Hypothesis i.e., time series is stationary.
- Test statistic value is -6.592372e+00.
- We can observe the variation between months, of which, some are positive and negative values.

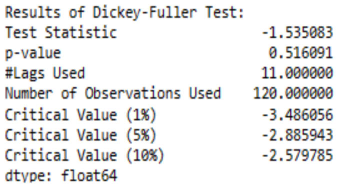
- We can also use **log method** to make time series stationary.

c) Log transformation of the data:

- Log method converts the multiplicative effect to additive.

Figure 45: Log transformation on train set





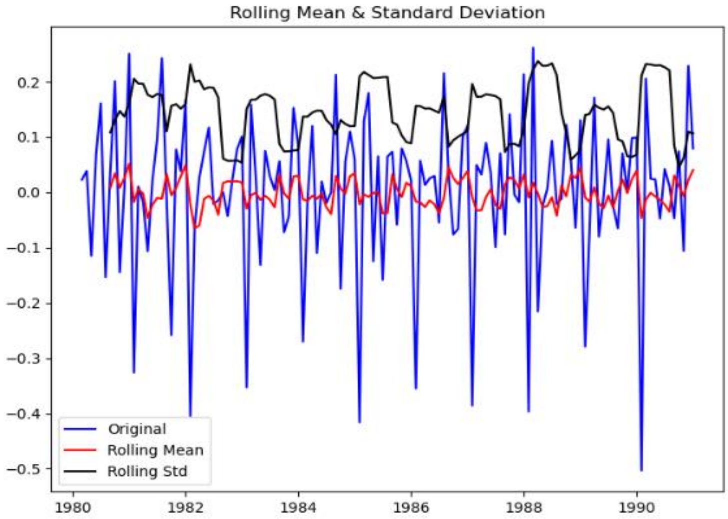
```
Results of Dickey-Fuller Test:
Test Statistic                 -1.535083
p-value                         0.516091
#Lags Used                     11.000000
Number of Observations Used   120.000000
Critical Value (1%)            -3.486056
Critical Value (5%)            -2.885943
Critical Value (10%)           -2.579785
dtype: float64
```

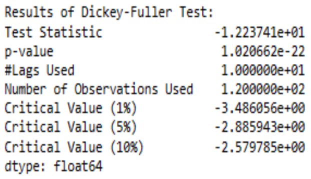
Observation:

- P value 0.52 > 0.05, hence, we fail to reject Null Hypothesis.
- Test statistic value is -1.54.
- Also, we can observe that log function to the base 10 has almost solved the variance issue.

d) Performing differencing (d=1) on log transformed data:

Figure 46: Differencing (d=1) on log transformed data





```
Results of Dickey-Fuller Test:
Test Statistic                -1.223741e+01
p-value                        1.020662e-22
#Lags Used                     1.000000e+01
Number of Observations Used    1.200000e+02
Critical Value (1%)           -3.486056e+00
Critical Value (5%)           -2.885943e+00
Critical Value (10%)          -2.579785e+00
dtype: float64
```

Observations:

- Here, the p value is lesser than 5% significant level, hence, we can reject Null Hypothesis.

- Test statistic value is -1.223741e+01.
- We can observe very less variability and no presence of trend.

<u>Conclusion</u>:

Time series data which was non-stationary has been made stationary by applying differencing (d=1) method on the data. The transformed data does not have any trend or variability, and is in a good form to take it further to build different models.

2.6. 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.

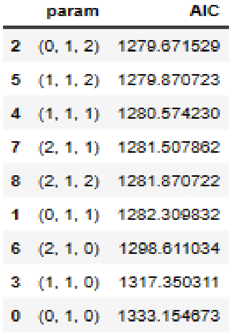
Ans:

We are choosing **ARIMA model** for the evaluation.

a) <u>ARIMA automated version (AIC):</u>

- For p and q values, we are choosing a range between (0,3)
- We are opting Differencing (d=1) method in order to make the data stationary.
- AIC values of the opted parameter combinations and sorting the AIC values in ascending order.

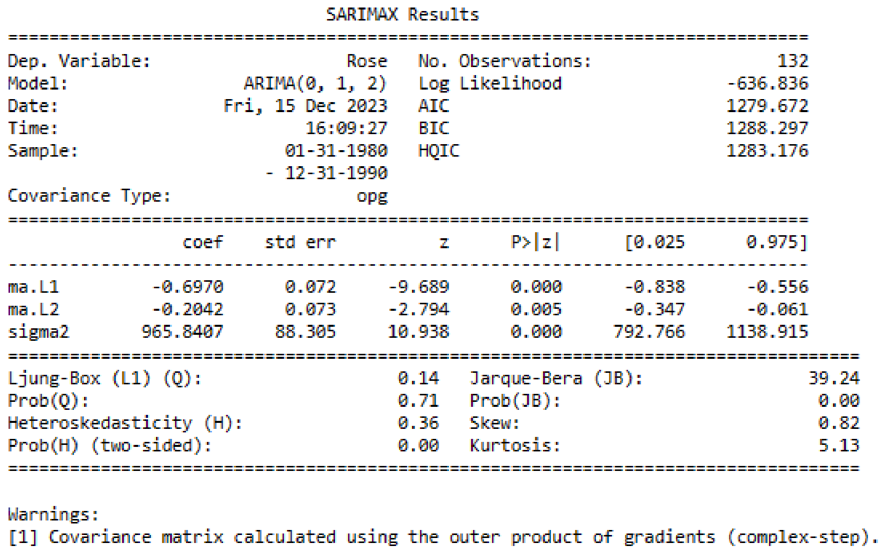
Table11: AIC values



	param	AIC
2	(0, 1, 2)	1279.671529
5	(1, 1, 2)	1279.870723
4	(1, 1, 1)	1280.574230
7	(2, 1, 1)	1281.507862
8	(2, 1, 2)	1281.870722
1	(0, 1, 1)	1282.309832
6	(2, 1, 0)	1298.611034
3	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

- Based on the minimum AIC value, we choose order (0,1,2) as parameter for the model.

Figure 47: SARIMAX results for ARIMA automated version (AIC)



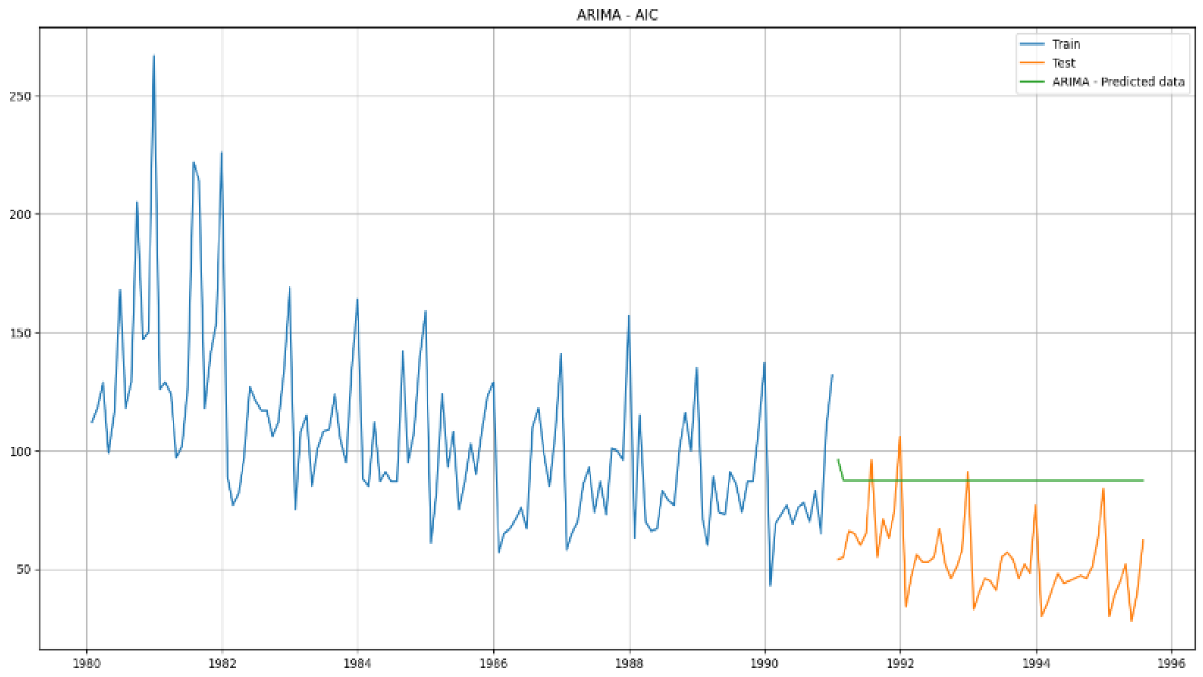
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -636.836
Date:                Fri, 15 Dec 2023   AIC                           1279.672
Time:                        16:09:27   BIC                           1288.297
Sample:                    01-31-1980   HQIC                          1283.176
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.6970      0.072     -9.689      0.000      -0.838      -0.556
ma.L2         -0.2042      0.073     -2.794      0.005      -0.347      -0.061
sigma2       965.8407     88.305     10.938      0.000     792.766    1138.915
===================================================================================
Ljung-Box (L1) (Q):                   0.14   Jarque-Bera (JB):                39.24
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               0.36   Skew:                             0.82
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.13
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

<u>Observations:</u>

- All the coefficients are significant as p values are less than 0.05.
- Prob (JB) is 0.00, so we can say that the data is not normal.
- The data is positively skewed (0.82).

- Now we can do the prediction on test data.

Figure 48: Predictions on test data ARIMA automated version (AIC)



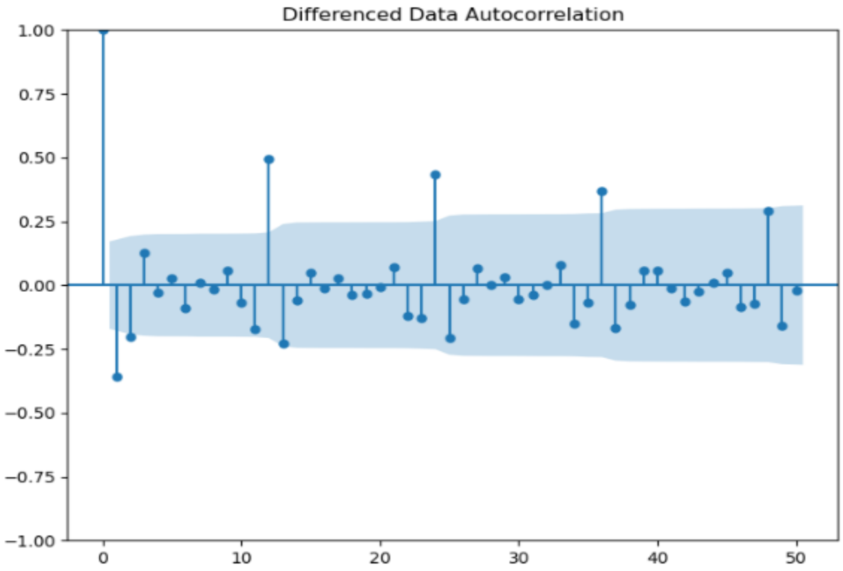
Observations:

- We can say that the model has not predicted the values correctly, as the seasonality is not considered in this ARIMA model.
- RMSE on the test data for ARIMA automated version (AIC) model is 37.26

b) **ARIMA manual version (ACF and PACF):**

- In this method, ‘p’ and ‘q’ parameters are determined based on the significant lag from the ACF and PACF plots.

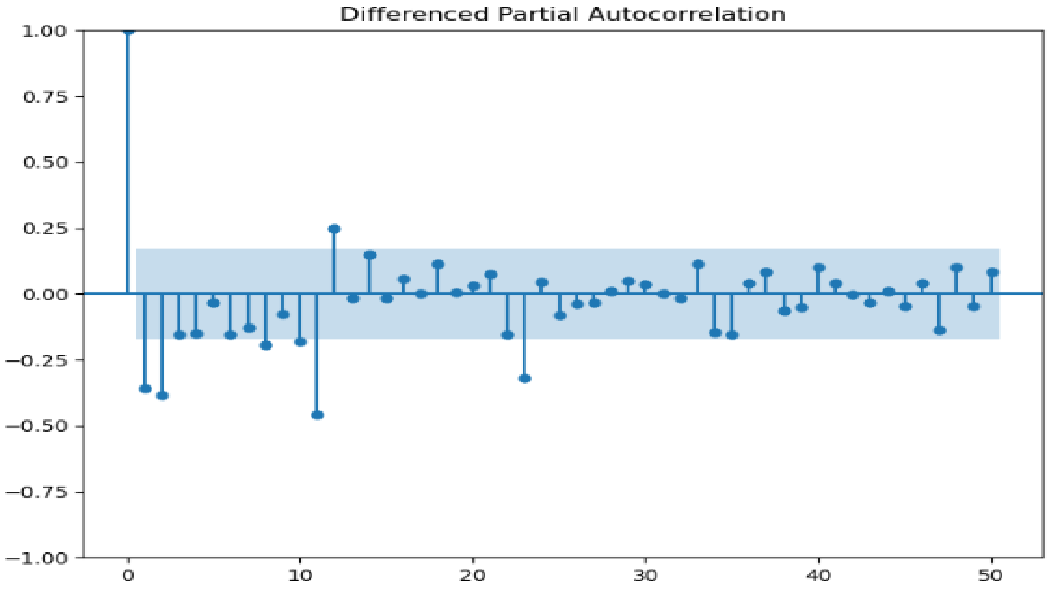
Figure 49: ACF



Insights**:**

- We have used differencing (d=1) method to make the data stationary.
- Moving Average (q) is considered from the ACF plot based on the significant lag.
- We can observe from the above plot that significant lag cuts-off at 2, i.e., q = 2.

Figure 50: PACF

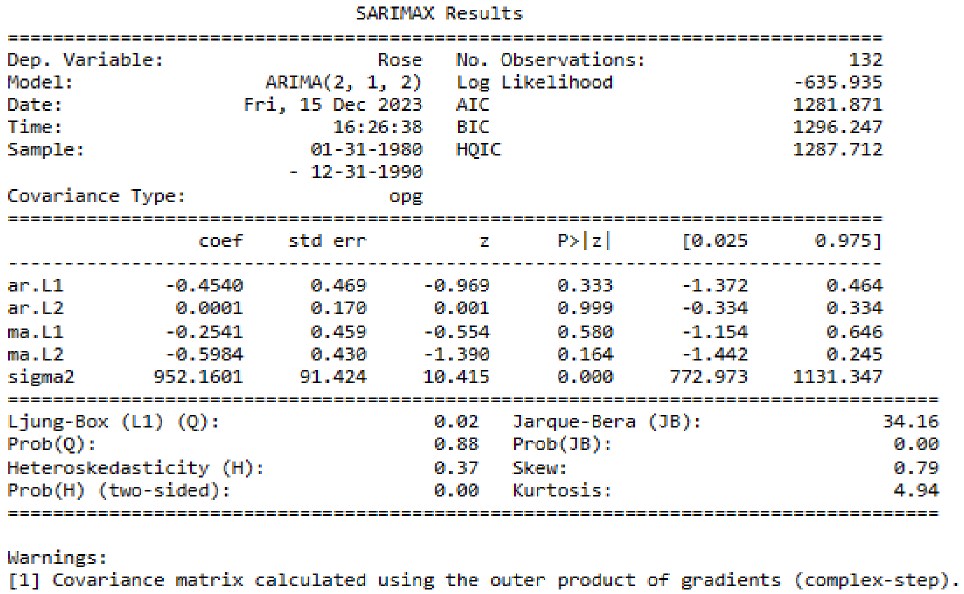


Insights:

- We have used differencing (d=1) method to make the data stationary.
- Auto Regression (p) is considered from the PACF plot based on the significant lag.
- We can observe from the above plot that significant lag cuts-off at 2, i.e., p = 2.

- Based on the above plots, we determine parameter of (2,1,2) for the ARIMA model.

Figure 51: SARIMAX results for ARIMA manual version (ACF and PACF)



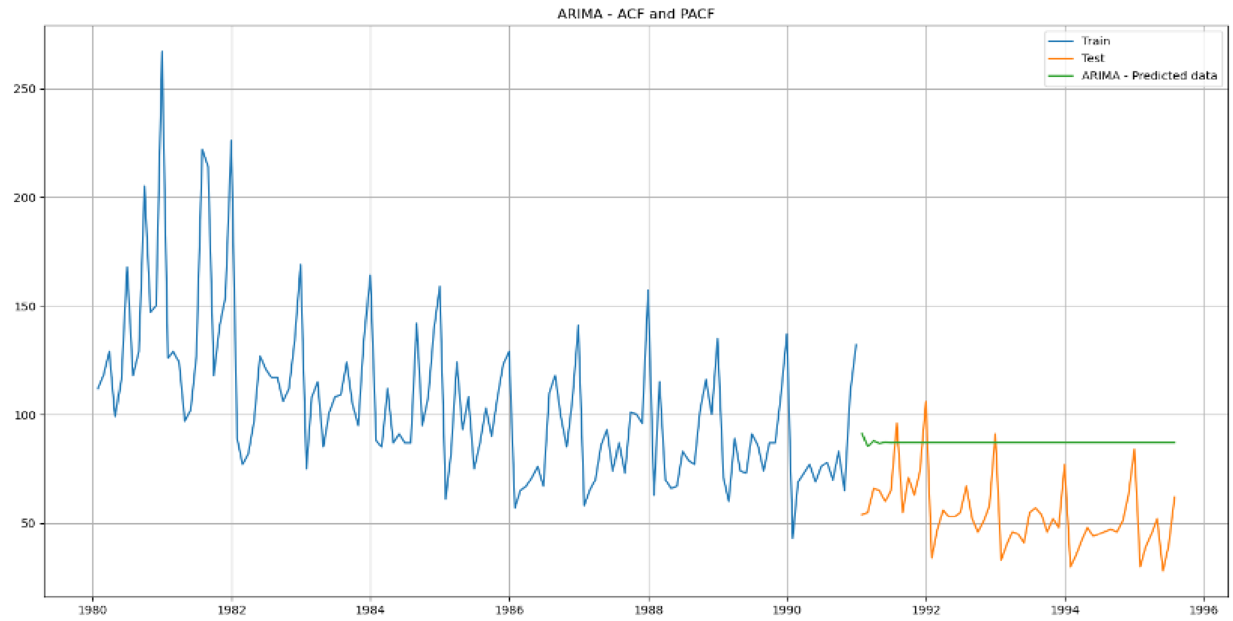
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -635.935
Date:                Fri, 15 Dec 2023   AIC                           1281.871
Time:                        16:26:38   BIC                           1296.247
Sample:                    01-31-1980   HQIC                          1287.712
                         - 12-31-1990
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
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Observations:

- Coefficients are insignificant as p values are more than 0.05, except sigma2 (variance of residuals).
- Data is not normal as Prob (JB) is 0.00.
- The data is positively skewed (0.79).
- The model has log likelihood of -635.935

- Now we can do the prediction on test data.

Figure 52: Predictions on test data ARIMA manual version (ACF and PACF)



Observations:

- We can say that the model has not predicted the values correctly, as the seasonality is not considered completely in this model.
- RMSE on the test data for ARIMA (ACF and PACF) model is 36.82

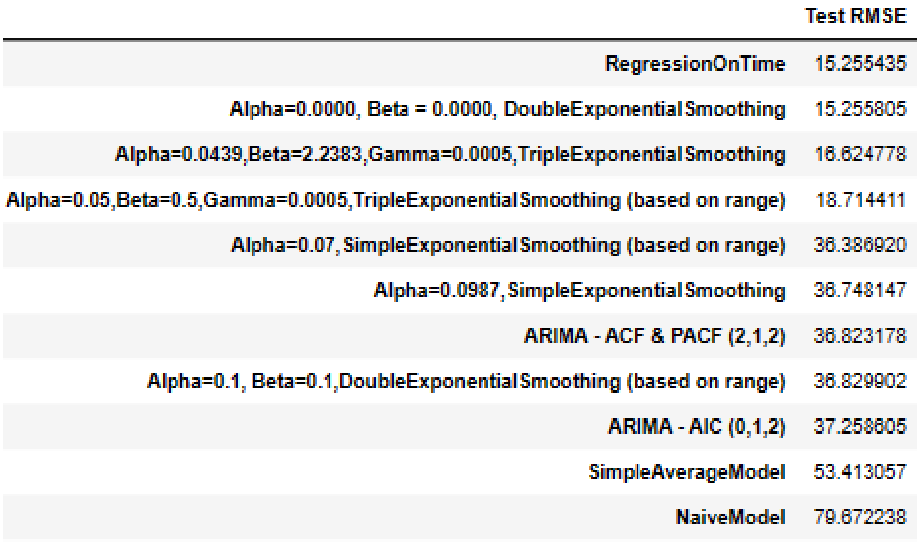
Conclusion:

Based on RMSE, ARIMA (ACF and PACF) model has lower RMSE 36.82.

2.7. 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.

Ans:

Table 12: Data frame showing built models with RMSE values



	Test RMSE
RegressionOnTime	15.255435
Alpha=0.0000, Beta = 0.0000, DoubleExponentialSmoothing	15.255805
Alpha=0.0439,Beta=2.2383,Gamma=0.0005,TripleExponentialSmoothing	16.624778
Alpha=0.05,Beta=0.5,Gamma=0.0005,TripleExponentialSmoothing (based on range)	18.714411
Alpha=0.07, SimpleExponentialSmoothing (based on range)	36.386920
Alpha=0.0987, SimpleExponentialSmoothing	36.748147
ARIMA - ACF & PACF (2,1,2)	36.823178
Alpha=0.1, Beta=0.1,DoubleExponentialSmoothing (based on range)	36.829902
ARIMA - AIC (0,1,2)	37.258605
SimpleAverageModel	53.413057
NaiveModel	79.672238

Observations:

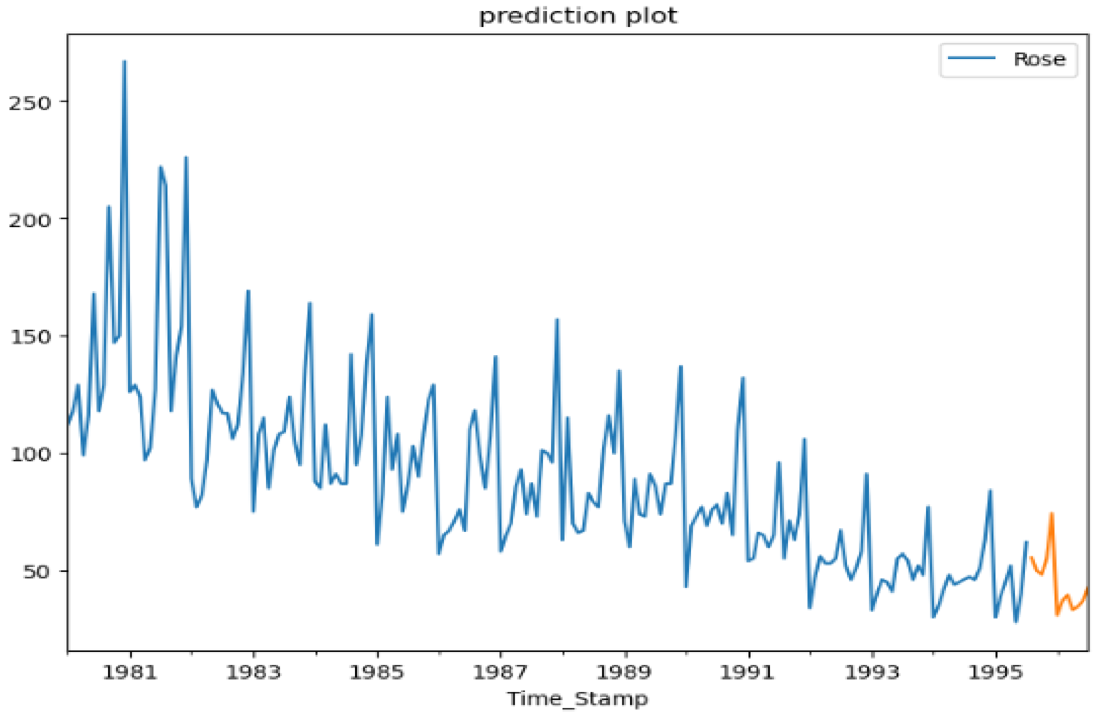
- The lowest RMSE value of all these models is 15.26 and for Linear regression model.
- The highest RMSE value of all these models is 79.67 and for Naïve model.

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.

Ans:

- Based on the RMSE value and seasonality, we opt Triple exponential smoothing model with Alpha = 0.0439, Beta = 2.2383 and Gamma = 0.0005.
- Initialising and fitting the model on the complete data.
- Forecasting the data for 12 months.
- RMSE value of the complete data is 18.48.
- Prediction on the test data, i.e., for future 12 months.

Figure 53: Predictions on test data (future 12 months)

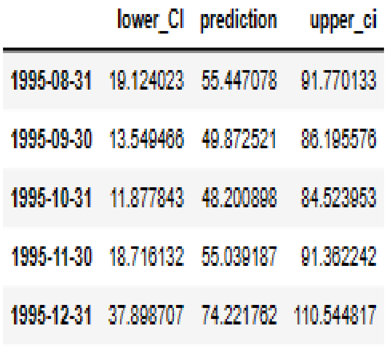


<u>Observation:</u>

- We can observe a downward or decreasing trend and seasonality on the above plot.

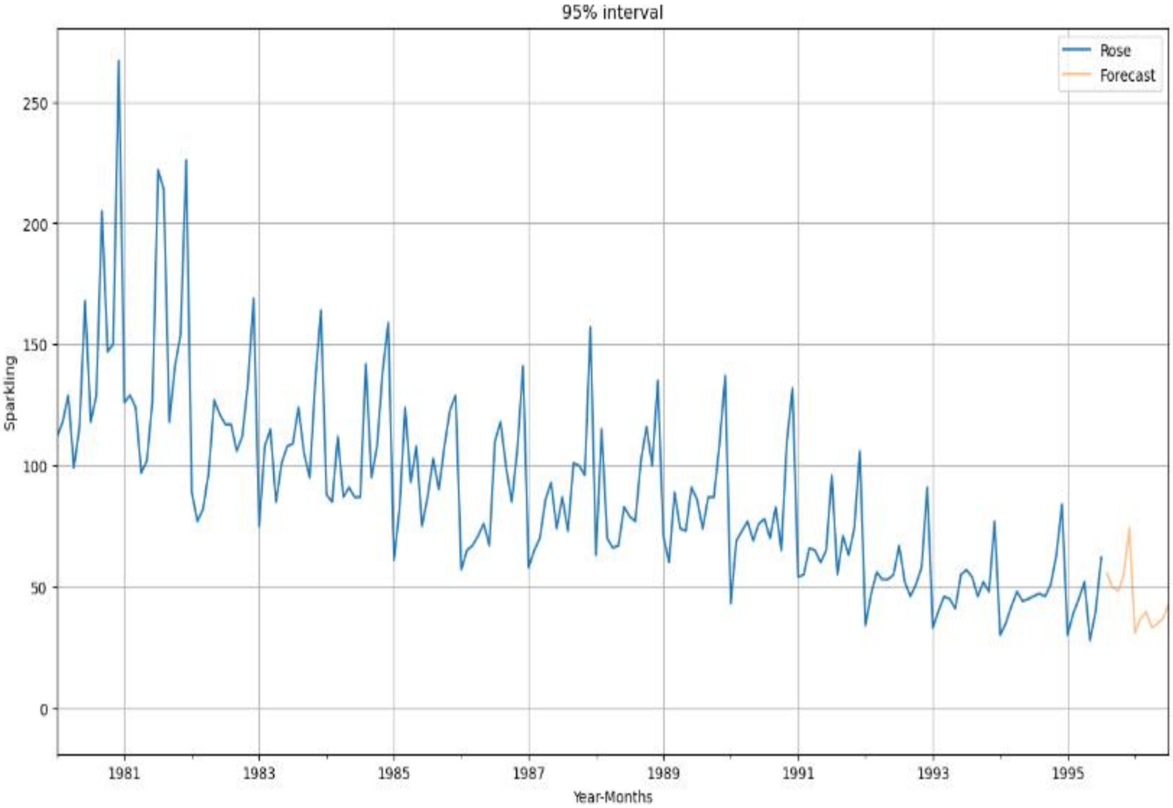
- We can calculate the lower and upper confidence bands at 95% confidence interval.

Table 13: 95% confidence interval



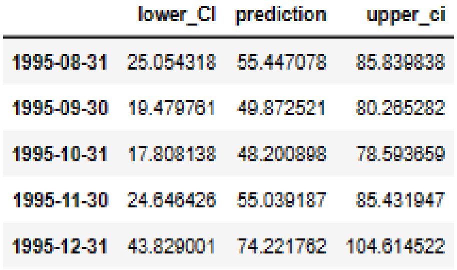
	lower_CI	prediction	upper_ci
1995-08-31	19.124023	55.447078	91.770133
1995-09-30	13.549466	49.872521	86.195576
1995-10-31	11.877843	48.200898	84.523953
1995-11-30	18.716132	55.039187	91.362242
1995-12-31	37.898707	74.221762	110.544817

Figure 54: Plot showing 95% confidence interval



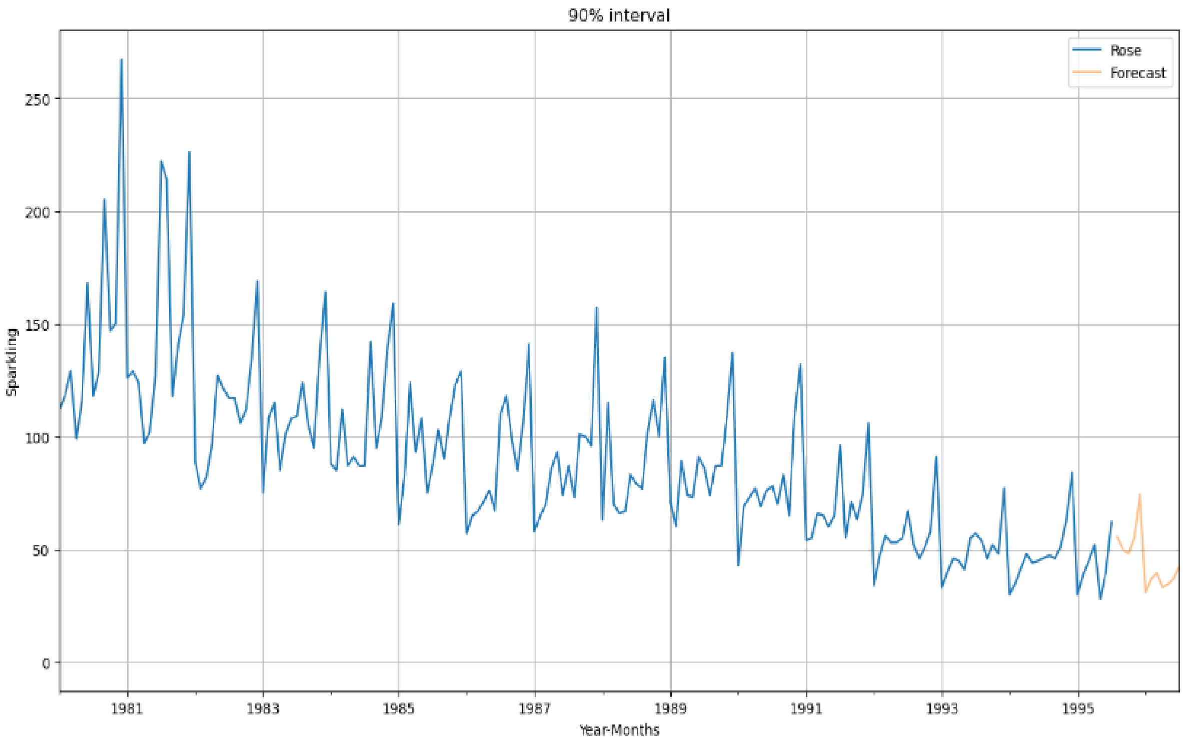
- We can calculate the lower and upper confidence bands at 90% confidence interval.

Table 14: 90% confidence interval



	lower_CI	prediction	upper_ci
1995-08-31	25.054318	55.447078	85.839838
1995-09-30	19.479761	49.872521	80.265282
1995-10-31	17.808138	48.200898	78.593659
1995-11-30	24.646426	55.039187	85.431947
1995-12-31	43.829001	74.221762	104.614522

Figure 55: Plot showing 90% confidence interval



2.9. 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.

Ans:

Insights and Findings:

- We have chosen Triple exponential smoothing model as an optimum model based on the seasonality of the data and RMSE value.
- Mean sales for the predicted year is 44.80, which was 89.93 in previous years.
- Median sales have decreased 85 (previous years) to 41.06 (predicted year).
- We can expect a maximum sale of 74.22 in next December and there would be a sudden drop in sales in January onwards.
- There is seasonality and decreasing trend for the Rose wine sales.

Suggestions and Recommendations:

- Company should focus on attracting new customers and retaining existing customers, as the sales has fallen drastically over the years.

- Company can introduce discounts, give away vouchers, gift coupons in order to attract more customers and boost sales.
- Company should plan the production accordingly, as the demand is expected to go down in the upcoming year. This helps in reducing overproduction.
- Company should have a proper mission and vision for the upcoming years.
- Company should improve promotional strategies so as to create demand for the product. Company can also introduce loyalty points to retain the existing customers.

---THE END--