TIME SERIES FORECASTING PROJECT

Submitted by,
VIDYA V

PGPDSBA.O.2023.B
11.11.2023

CONTENTS

List of Figures

Rose Dataset

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.

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

	Rose	month	year
1980-01-01	112.0	Jan	1980
1980-02-01	118.0	Feb	1980
1980-03-01	129.0	Mar	1980
1980-04-01	99.0	Apr	1980
1980-05-01	116.0	May	1980

Fig.1.1. Rose Dataset

	Rose	year
count	187.000000	187.000000
mean	89.909091	1987.299465
std	39.244440	4.514749
min	28.000000	1980.000000
25%	62.500000	1983.000000
50%	85.000000	1987.000000
75%	111.000000	1991.000000
max	267.000000	1995.000000

Fig.1.2. Rose Dataset Description

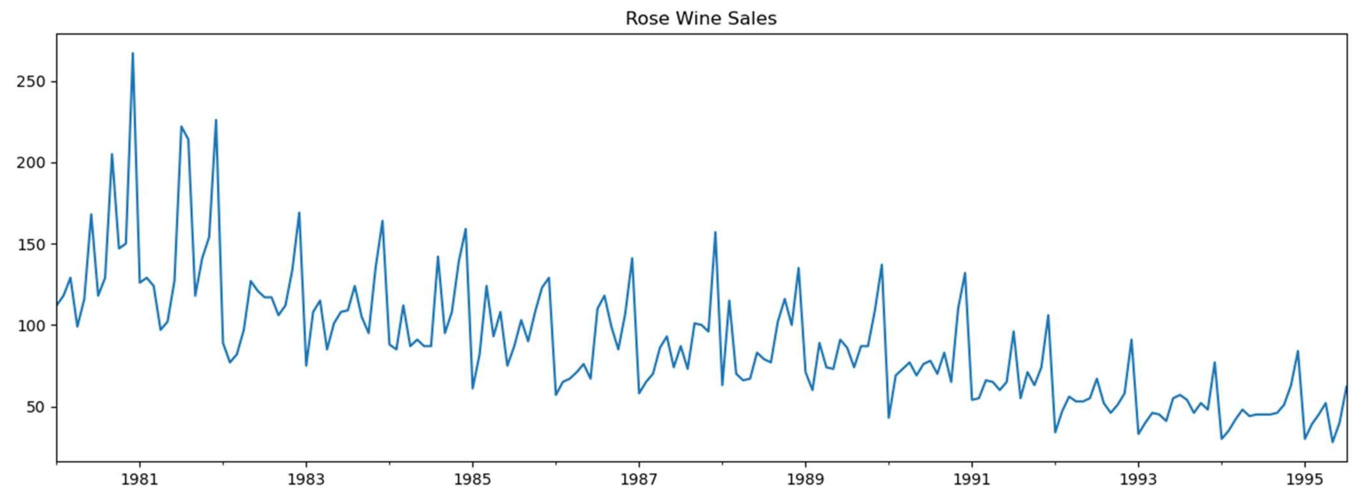

Fig.1.3. Rose wine sales

Observations:

- The plot represents the Rose wine sales from Jan 1980 to July 1995, covering a span of 15.5 years- 187 values
- Two values were missing, and the same was imputed with the last observed value
- There seems to be a declining trend and some seasonality associated with this plot.
- The minimum sales was 28, the maximum sales was 267, with a mean of 89.9

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

2.1. EDA

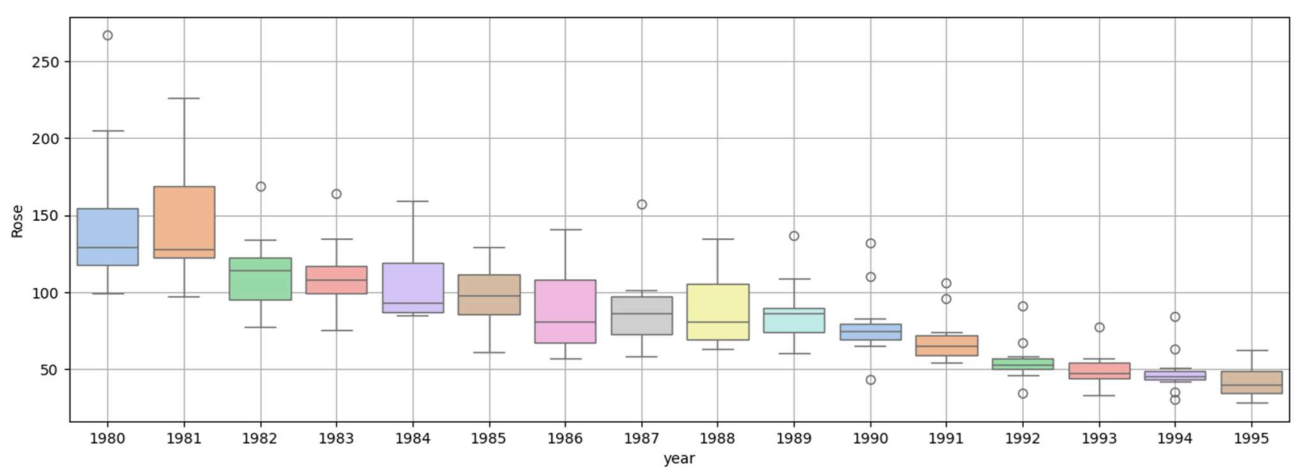

Fig.1.4 Yearly boxplot- Rose Dataset

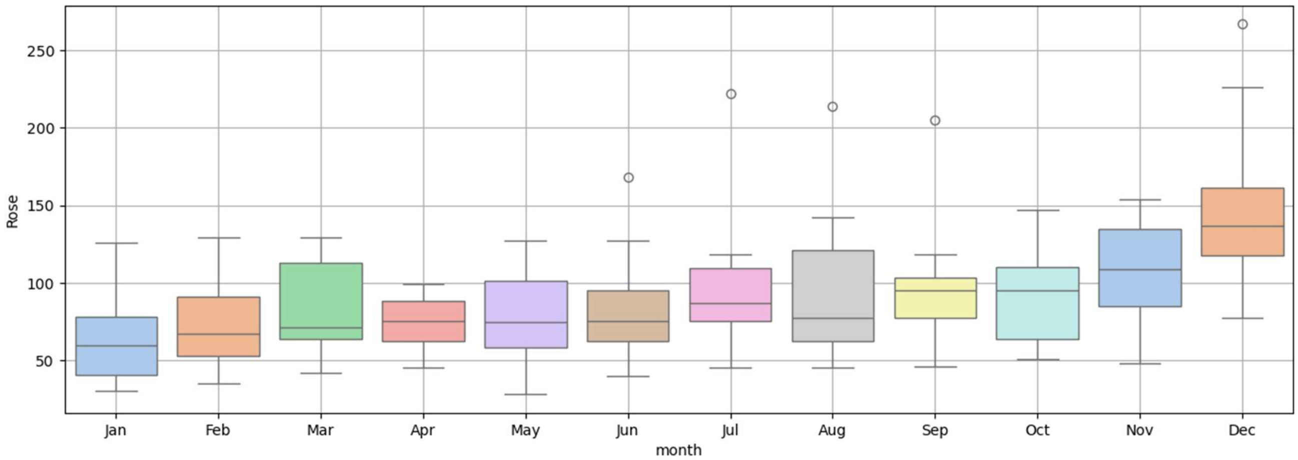

Fig.1.5. Monthly Boxplot- Rose wine sales

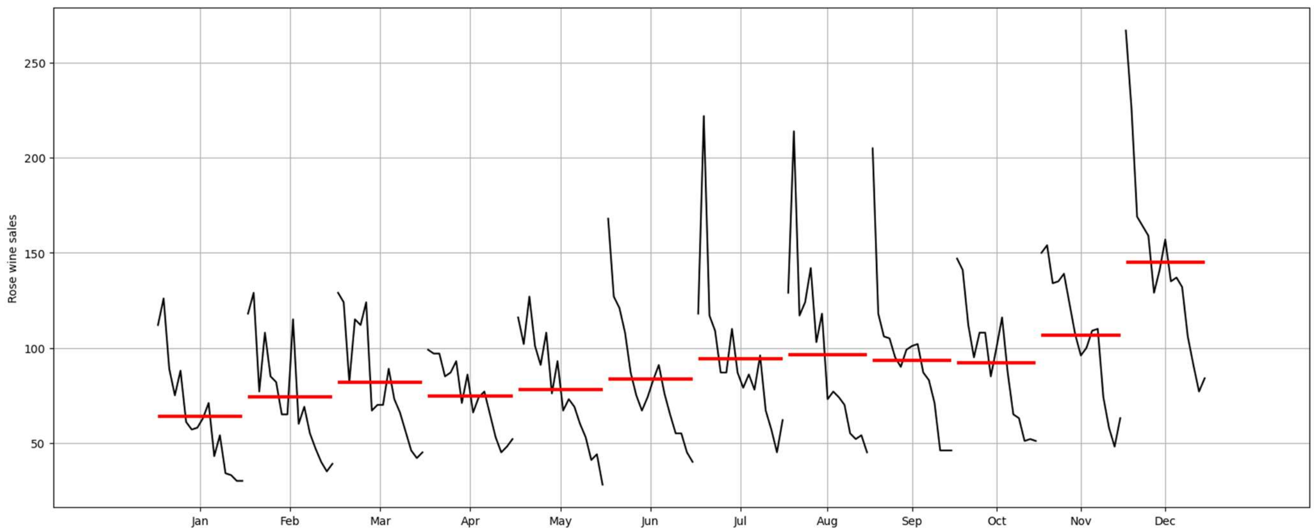

Fig.1.6. Month-plot- Rose wine sales

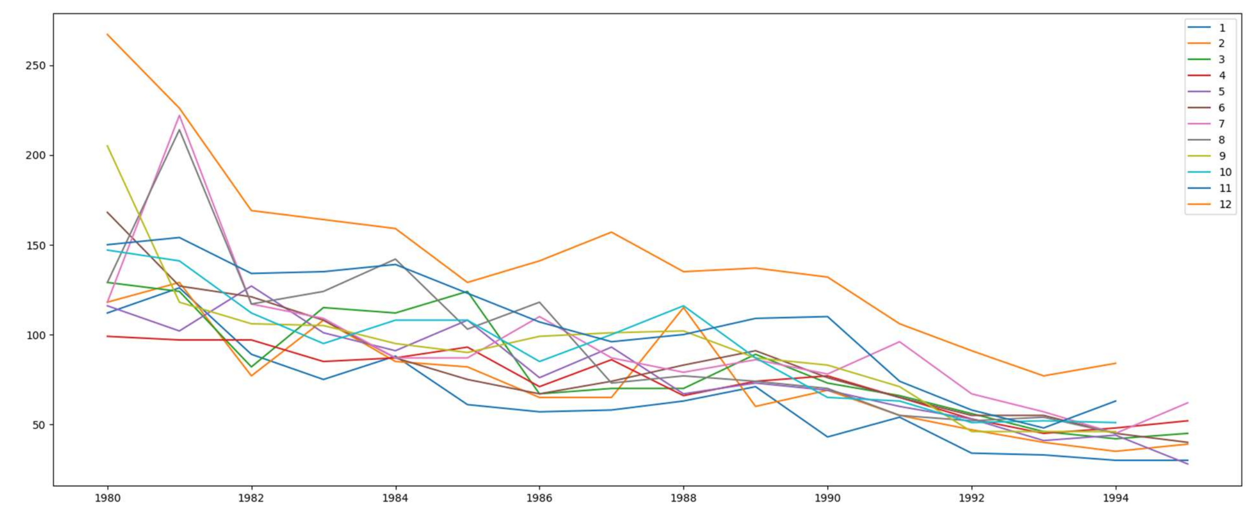

Fig.1.7. Monthly sales over the years

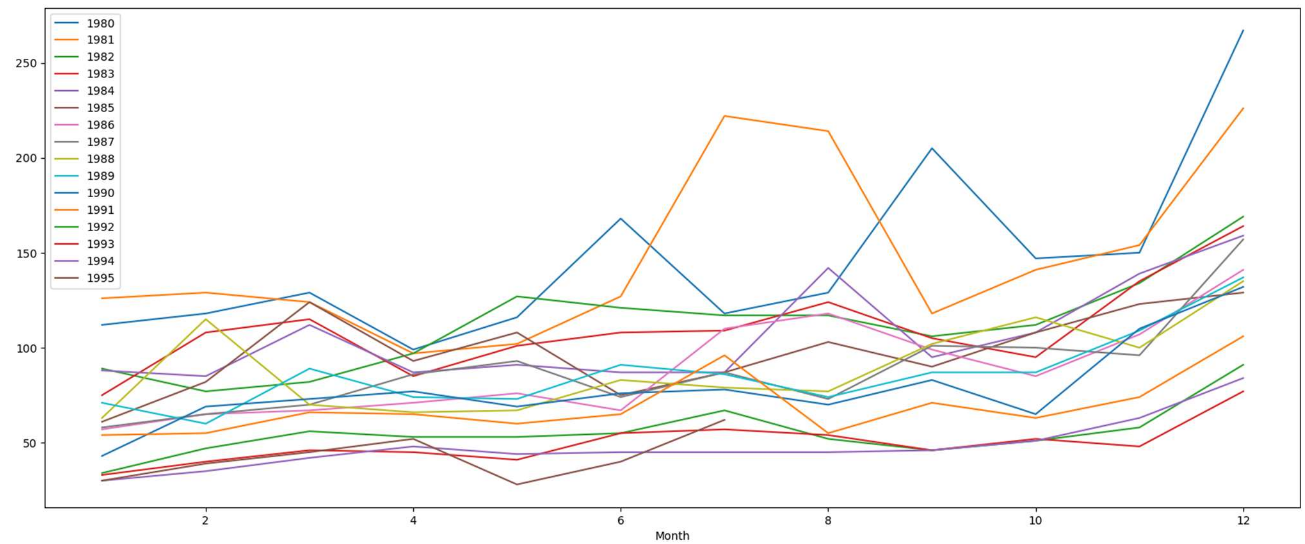

Fig.1.8. Yearly sales- Rose Wine

Observations:

- Clear declining trend observed
- The sales remain flat for the first 9 months of the year, and then rise, peaking in December

2.2. Decomposition

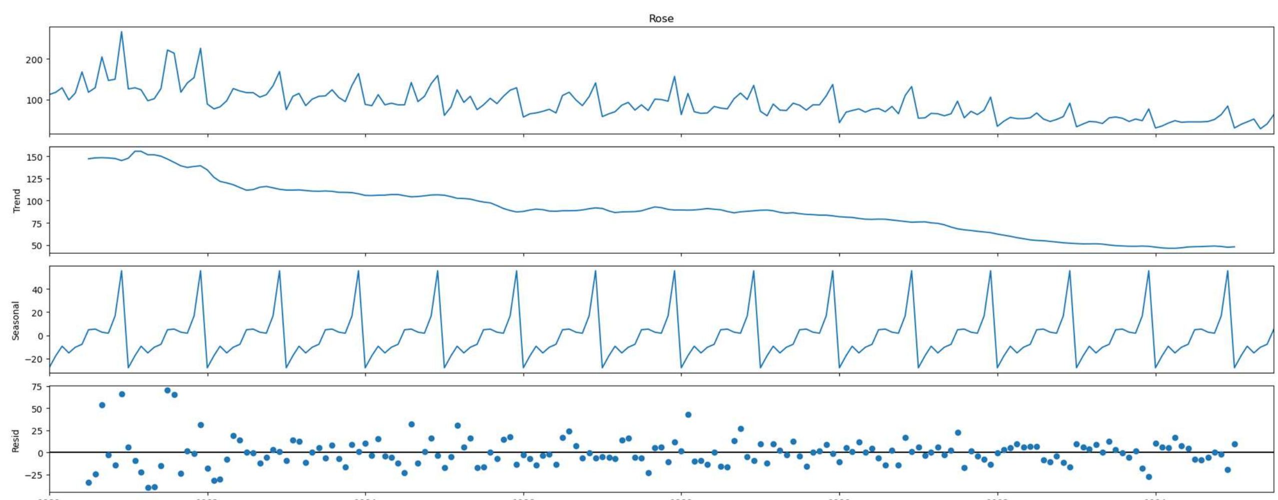

Fig.1.9. Additive Decomposition- Rose sales

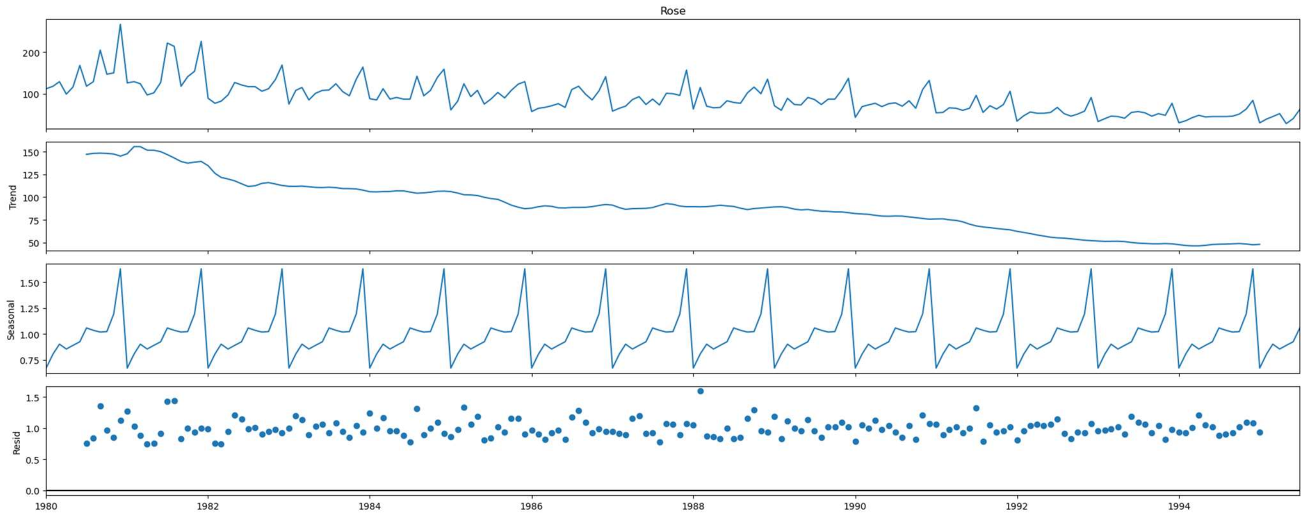

Fig.1.10. Multiplicative Decomposition- Rose wine sales

Observations:
- Clear seasonality component observed
- The three conditions for multiplicative seasonality are fulfilled, and hence we can assume multiplicative seasonality

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

Observations:
- After the split, the train dataset contains 132 values
- The test dataset contains 55 values, starting from Jan 1991

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.**

4.1. Simple Models

4.1.1. Linear Regression

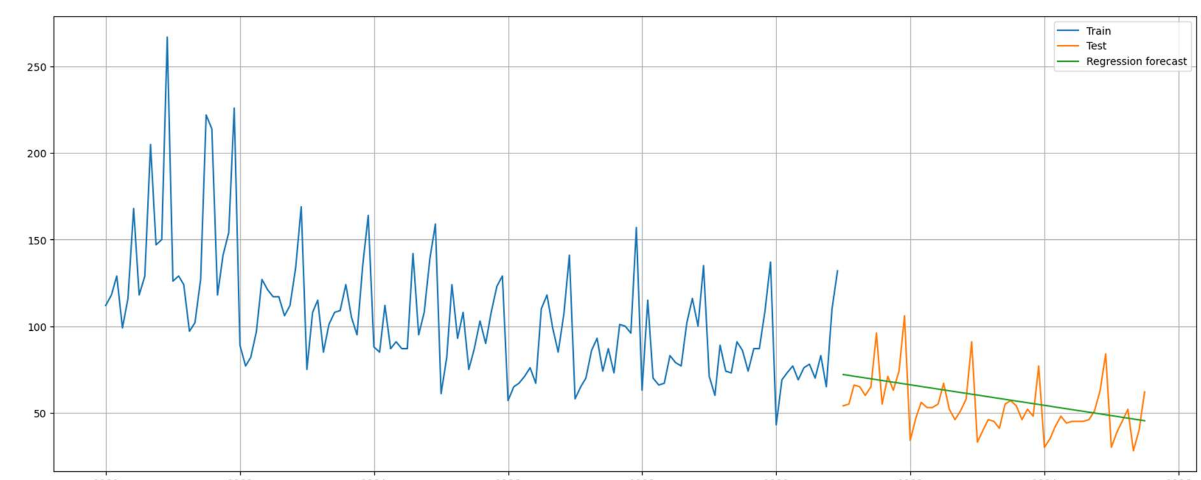

Fig.1.11 Linear Regression model test forecast plot- Rose sales

4.1.2. Naïve Forecast

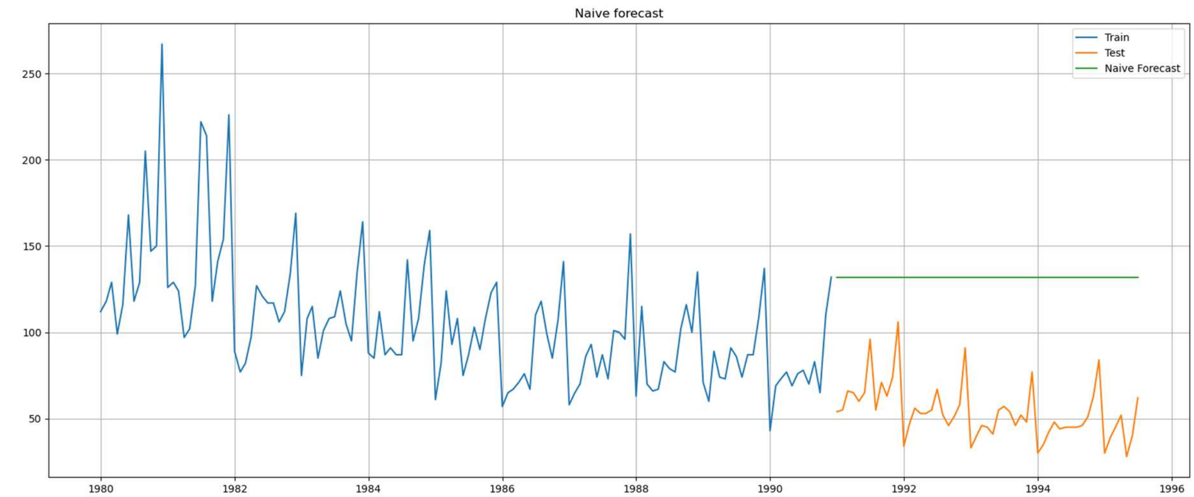

Fig.1.12. Naive forecast of test data- Rose wine sales

4.1.3. Simple Average

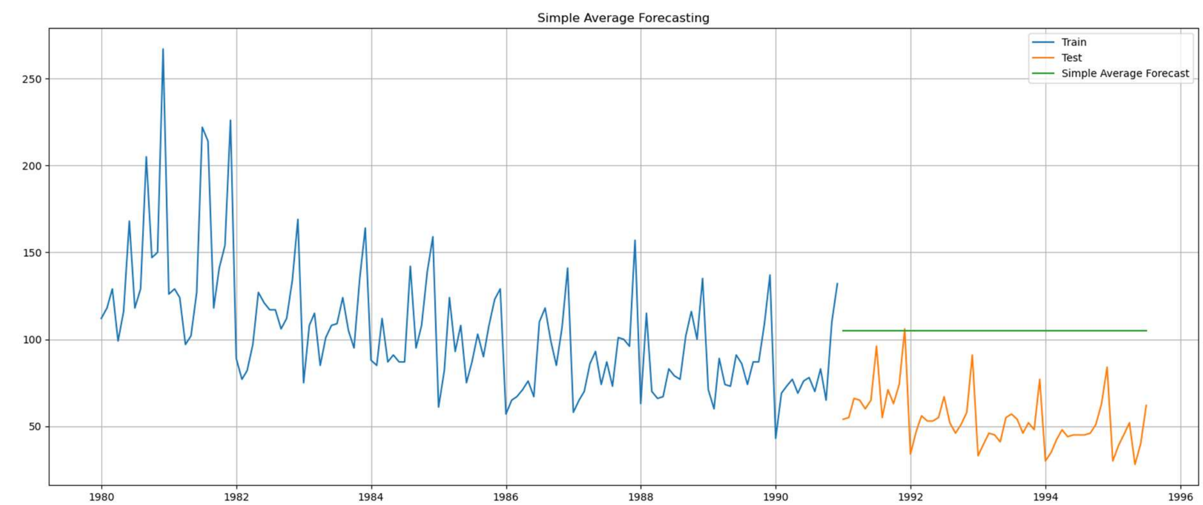

Fig.1.13. Simple Average forecast of test data- Rose wine sales

4.1.4. Moving Average

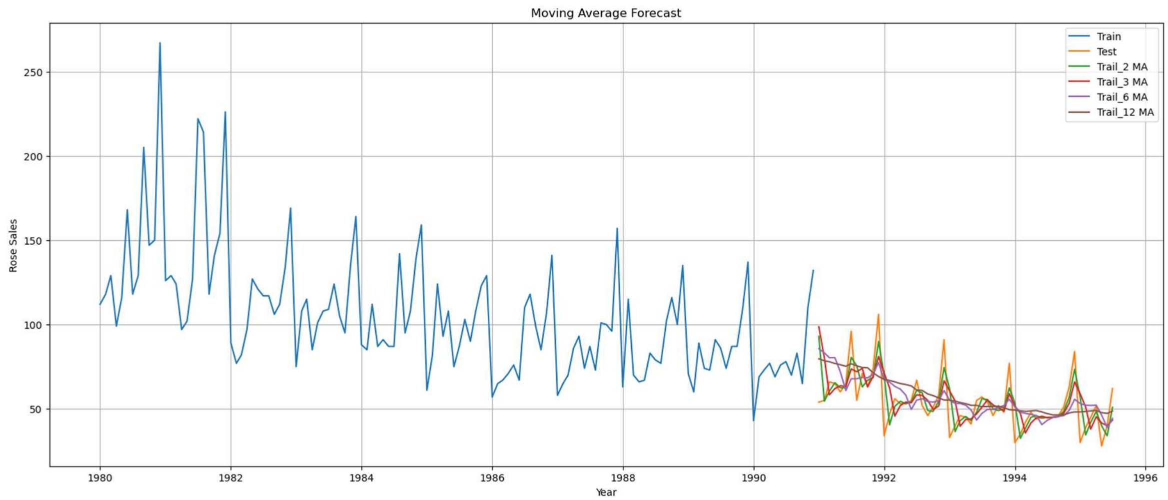

Fig.1.14. Moving Average forecast of test data- Rose wine sales

Observations:

- Best fit occurs in MA trail 2 model

4.2. Exponential Smoothing Models

4.2.1. Simple Exponential Smoothing

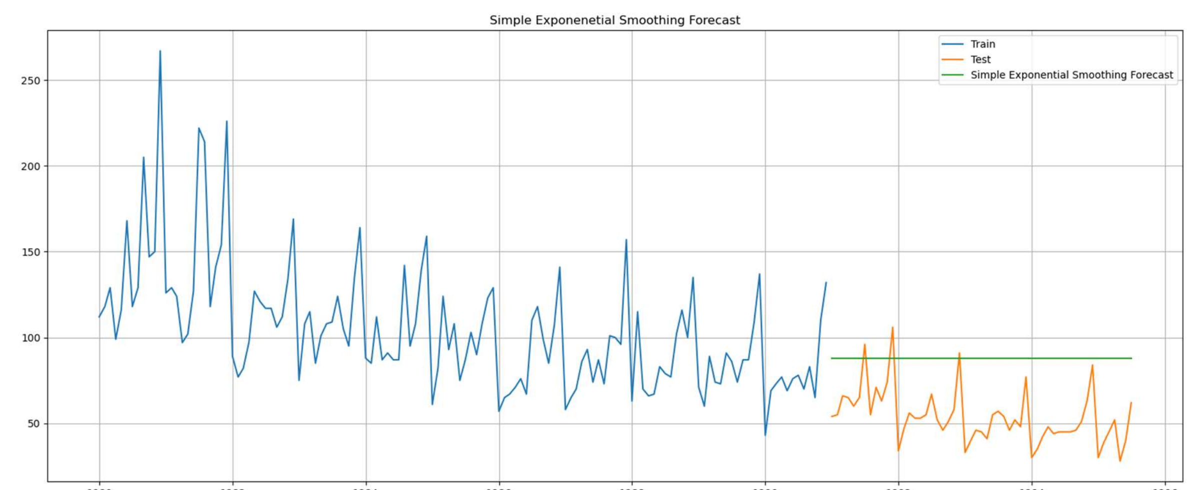

Fig.1.15. Simple Exponential smoothing forecast of test data- Rose wine sales

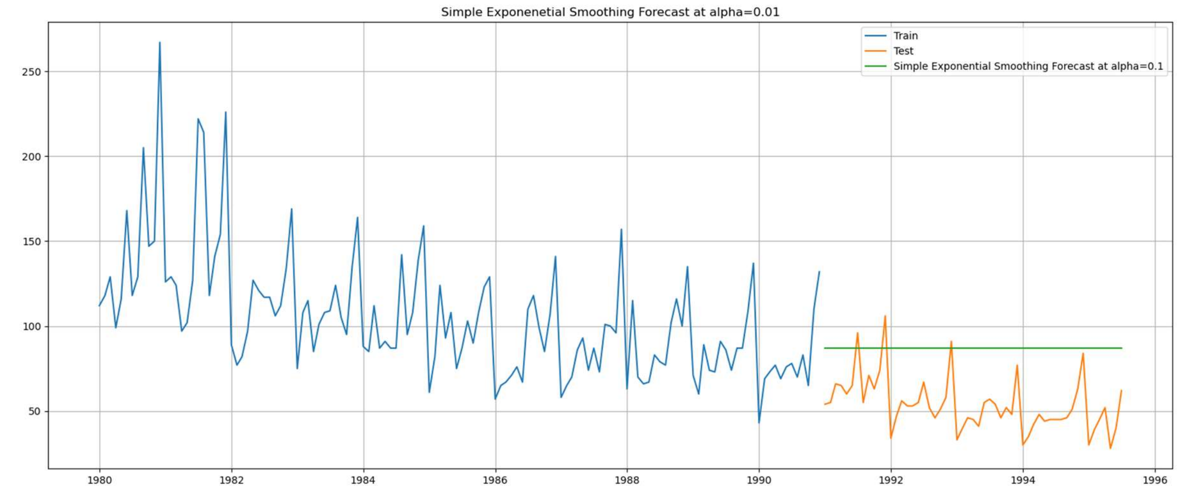

Fig.1.16. Simple Exponential smoothing forecast of test data- Rose wine sales optimized for lowest RMSE

Observations:

- RMSE is the lowest for alpha=0.1

4.2.2. Holt Double Exponential Smoothing

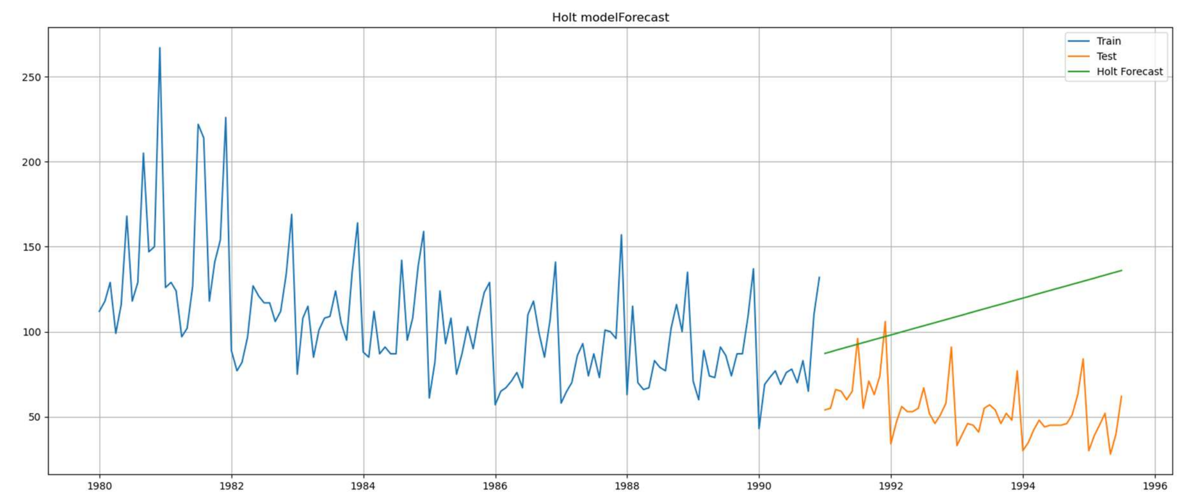

Fig.1.17**.** Holt forecast of test data- Rose wine sales

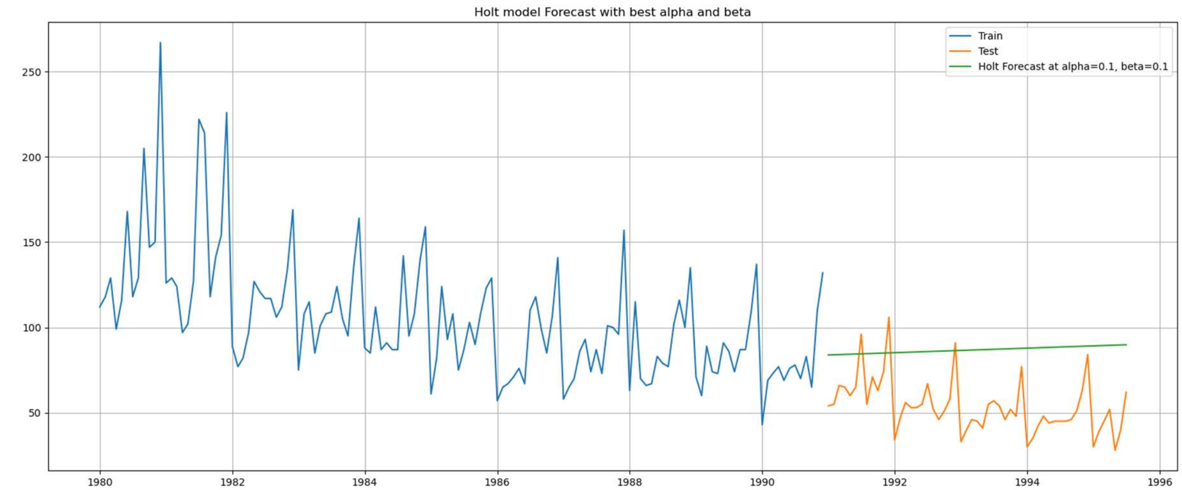

Fig.1.18. Holt forecast of test data- Rose wine sales - optimized for lowest RMSE

Observations:

- RMSE is the lowest for alpha=0.1, beta=0.1

4.2.3. Holt-Winters Triple Exponential Smoothing

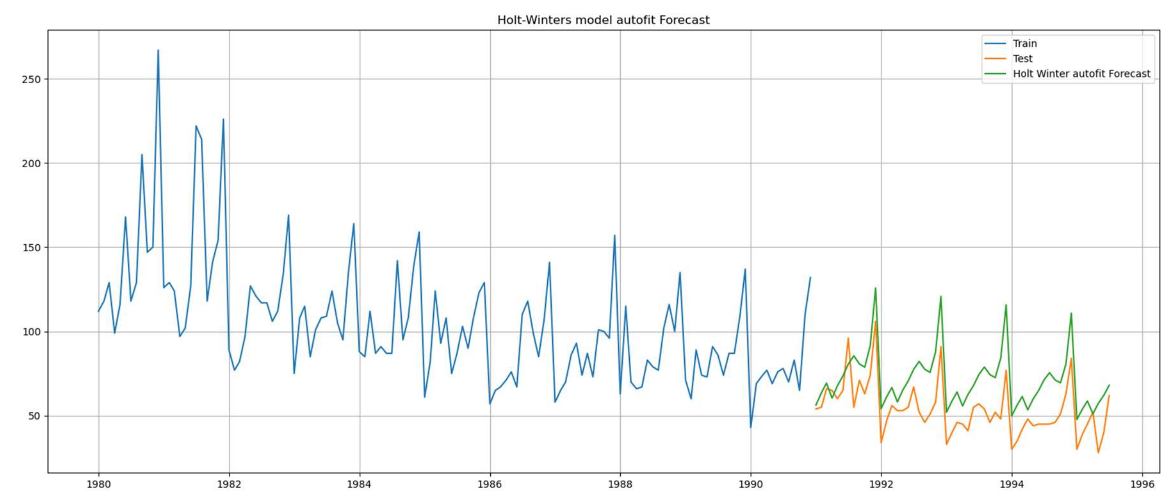

Fig.1.19**.** Holt Winters forecast of test data- Rose wine sales

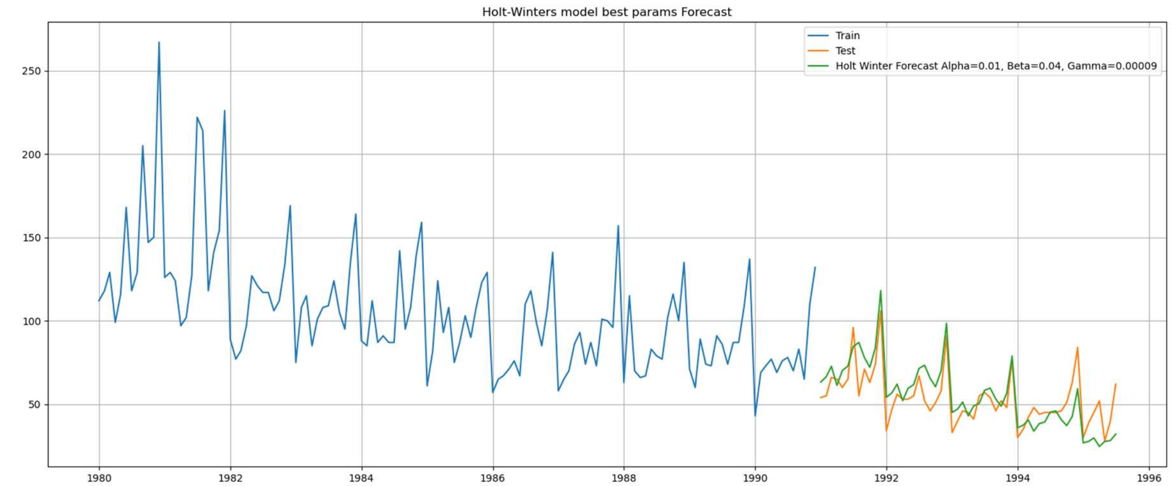

Fig.1.20. Holt Winters smoothing forecast of test data- Rose wine sales - optimized for lowest RMSE

Observations:

- RMSE is the lowest for alpha=0.01, beta=0.04 and gamma=0.00009

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.**

   The stationarity of the data can be ascertained by the Dickey-Fuller test. The Null and alternate hypothesis are as follows:

- **H0: The series is non-stationary**
- **Ha: The series is stationary**

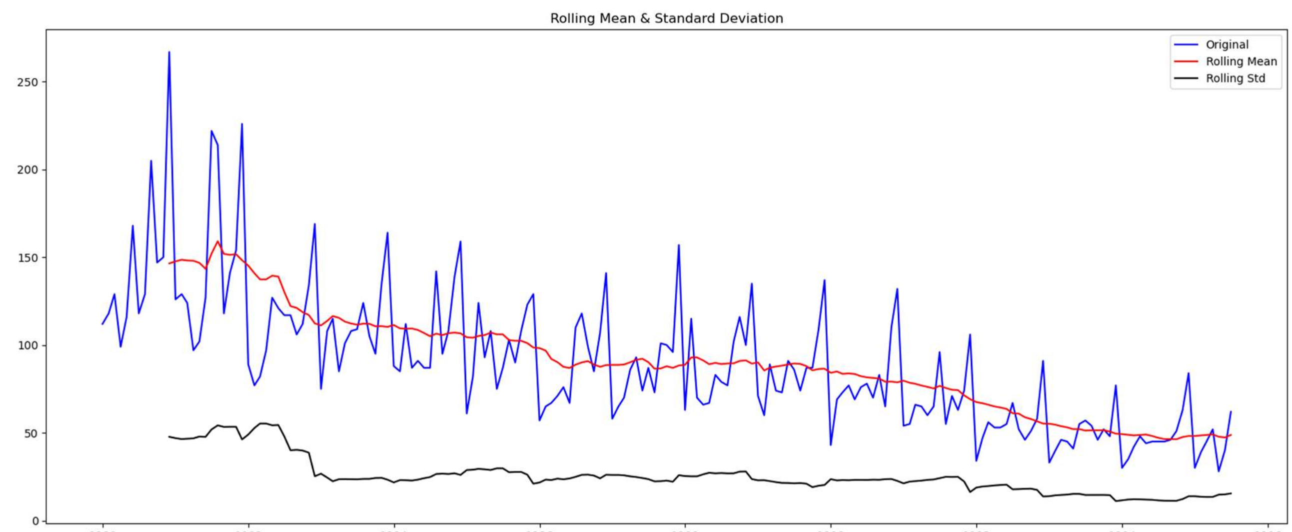

Fig.1.21. Rose dataset -Stationarity test rolling mean and Standard deviation plots

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.874856
p-value                          0.343981
#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
```

Fig.1.22. Dickey Fuller Test results- Rose Dataset

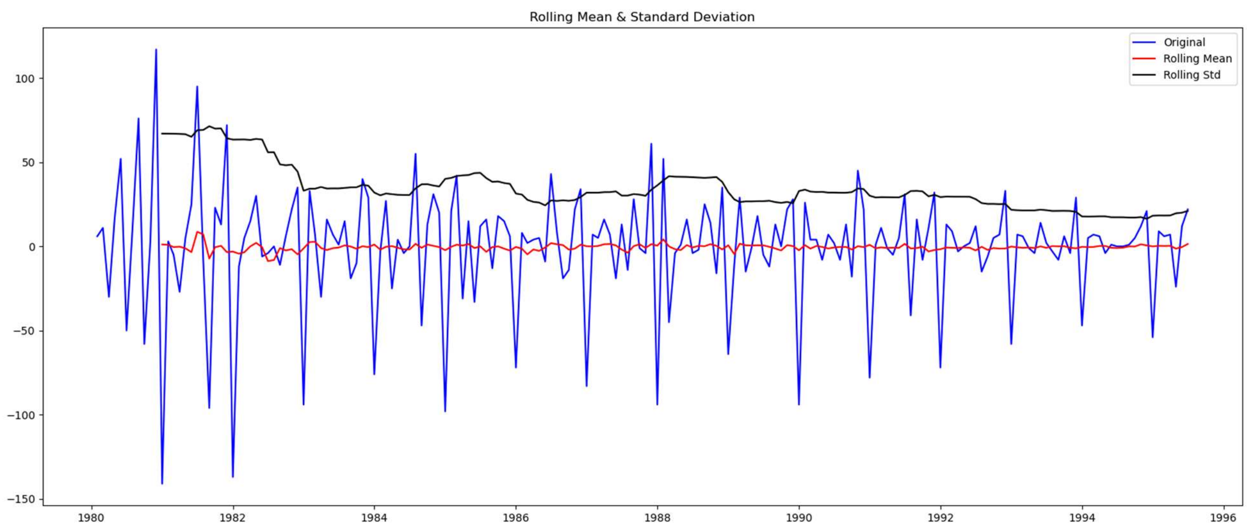

Fig.1.23. Differenced series -Stationarity test rolling mean and Standard deviation plots

```
Results of Dickey-Fuller Test:
Test Statistic                -8.044139e+00
p-value                        1.813580e-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
```

Fig.1.24. Dickey Fuller Test results- Differenced series

Observations:

- The given series was originally non- stationary, as evidenced by the Dickey Fuller test, with resulted in a p-value of 0.3
- After performing a first order differencing, stationarity was established. The Dickey fuller test on the differenced series resulted in a p-value of 0.0, which is less than the critical value of 0.05.

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.**

6.1. ARIMA model

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -636.836
Date:                Sat, 11 Nov 2023   AIC                           1279.672
Time:                        11:52:12   BIC                           1288.297
Sample:                    01-01-1980   HQIC                          1283.176
                         - 12-01-1990
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
===================================================================================
```

Fig.1.25. ARIMA results Summary- Rose Dataset

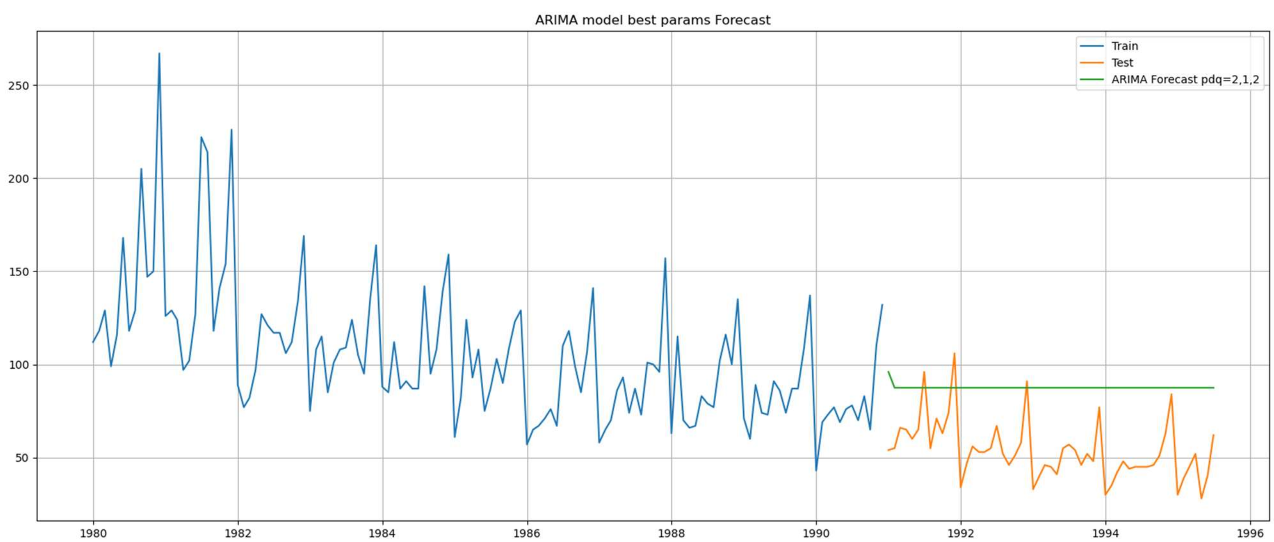

Fig.1.26. ARIMA model forecast on test data- Rose Dataset

Observations:

- Lowest AIC obtained for (p,d,q)=(0,1,2)
- This is consistent with the d=1 obtained during stationarity check

6.2. SARIMA model

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -380.485
Date:                              Sat, 11 Nov 2023   AIC                            774.969
Time:                                      12:00:48   BIC                            792.622
Sample:                                  01-01-1980   HQIC                           782.094
                                       - 12-01-1990
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9524      0.184     -5.166      0.000      -1.314      -0.591
ma.L2         -0.0764      0.126     -0.605      0.545      -0.324       0.171
ar.S.L12       0.0480      0.177      0.271      0.786      -0.299       0.395
ar.S.L24      -0.0419      0.028     -1.513      0.130      -0.096       0.012
ma.S.L12      -0.7526      0.301     -2.503      0.012      -1.342      -0.163
ma.S.L24      -0.0721      0.204     -0.354      0.723      -0.472       0.327
sigma2       187.8646     45.275      4.149      0.000      99.127     276.602
===================================================================================
Ljung-Box (L1) (Q):                   0.06   Jarque-Bera (JB):                 4.86
Prob(Q):                              0.81   Prob(JB):                         0.09
Heteroskedasticity (H):               0.91   Skew:                             0.41
Prob(H) (two-sided):                  0.79   Kurtosis:                         3.77
===================================================================================
```

Fig.1.27. SARIMA results Summary- Rose Dataset

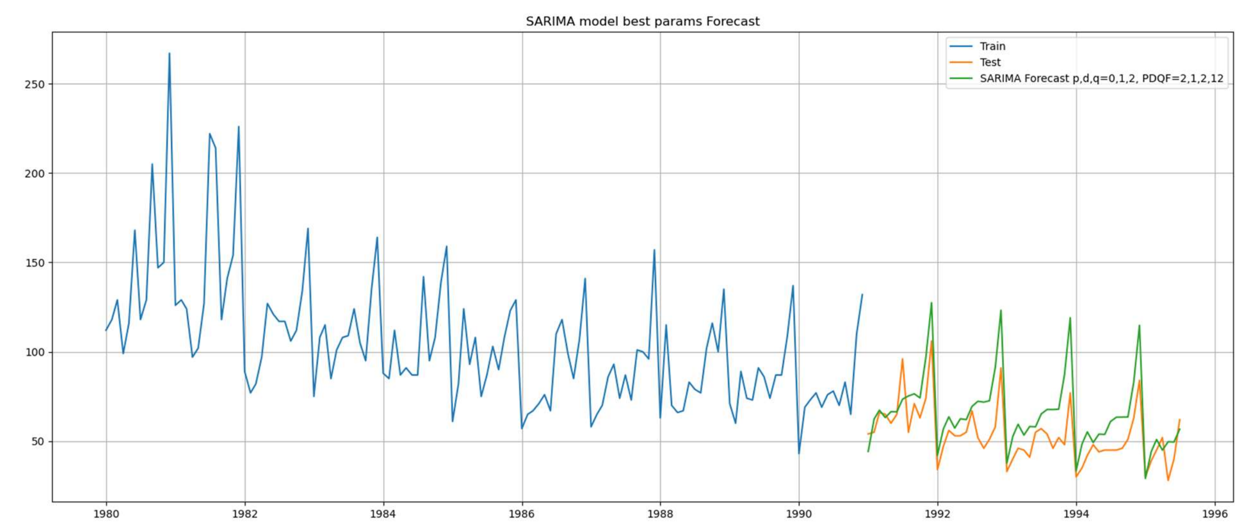

Fig.1.28. SARIMA model forecast on test data- Rose Dataset

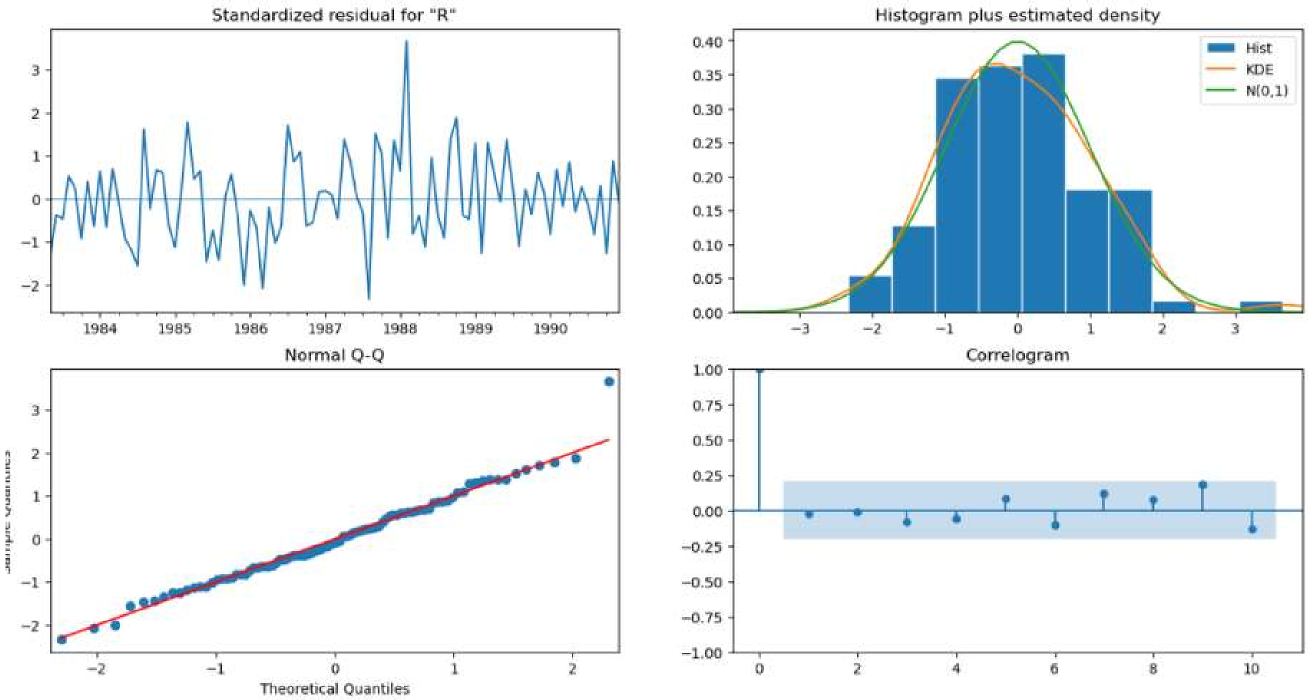

Fig.1.29. Diagnostics and Correlogram- SARIMA Model

Observations:

- Lowest AIC obtained for (p,d,q)x(P,D,Q,F)=(0,1,2)x(2,1,2,12)
- This is consistent with the d=1 obtained during stationarity check

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.

	Test RMSE
MA_trail2	11.53
HoltWintersalpha0.01beta0.04gamma0.00009	12.01
MA_trail3	14.13
MA_trail6	14.57
MA_trail12	15.24
LinearRegression	15.28
SARIMA	16.52
HoltWintersAutofit	20.18
SimpExpSmoothingAlpha0.1	36.85
HoltBestAlphaBeta	36.94
ARIMA	37.33
SimpleExpSmoothing	37.61
SimpleAvgForecast	53.48
HoltAutofit	63.07
NaiveForecast	79.74

Fig.1.30. Rose Dataset model Results- Test RMSE

Observations:

- From the above table, we can observe that the best model for the given time series is Holt Winters and SARIMA with appropriate params
- Eventhough models like MA and Linear Regression have better RMSEs the seasonality component is not incorporated in the, and hence cannot be considered

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.

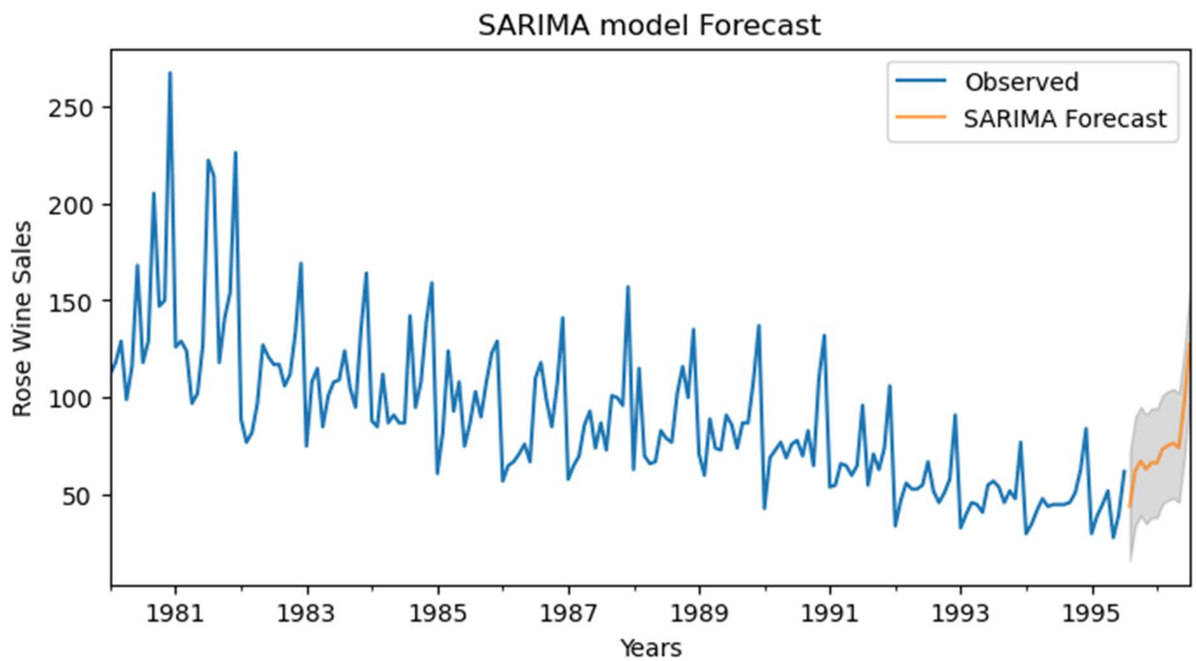

Fig.1.31. SARIMA Model forecast for next 12 months- Rose Dataset

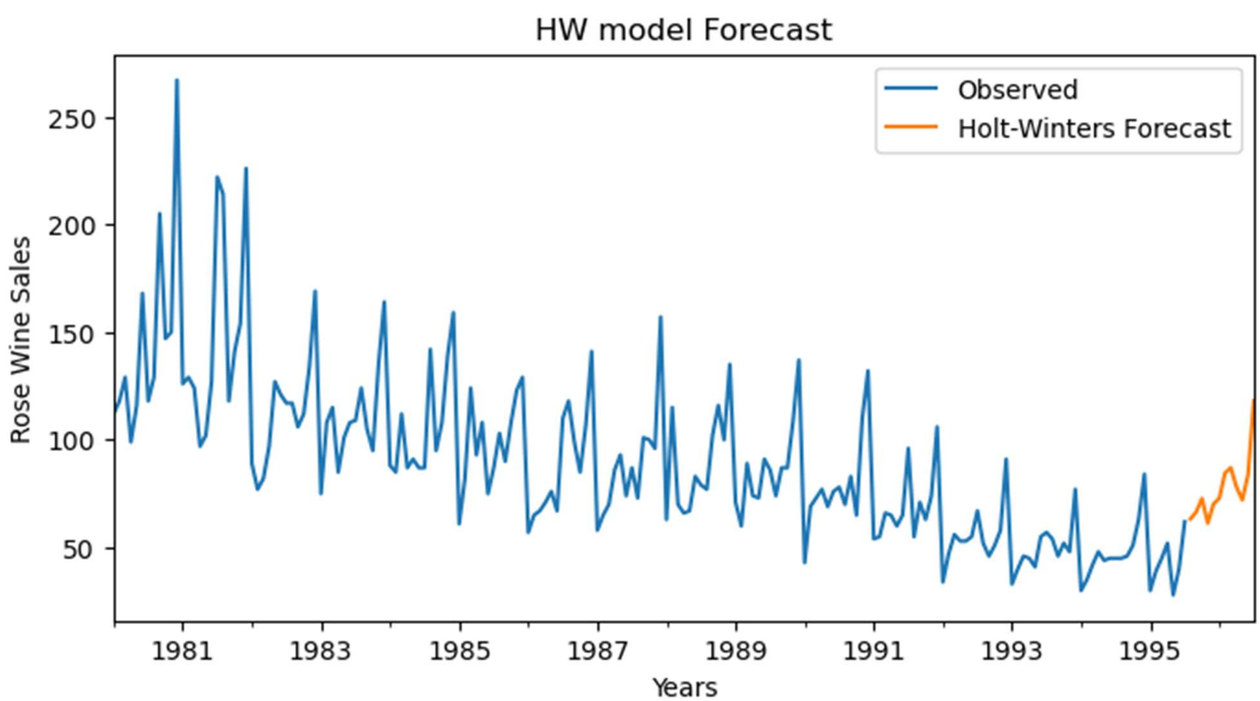

Fig.1.32. Holt Winters Model forecast for next 12 months- Rose Dataset

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

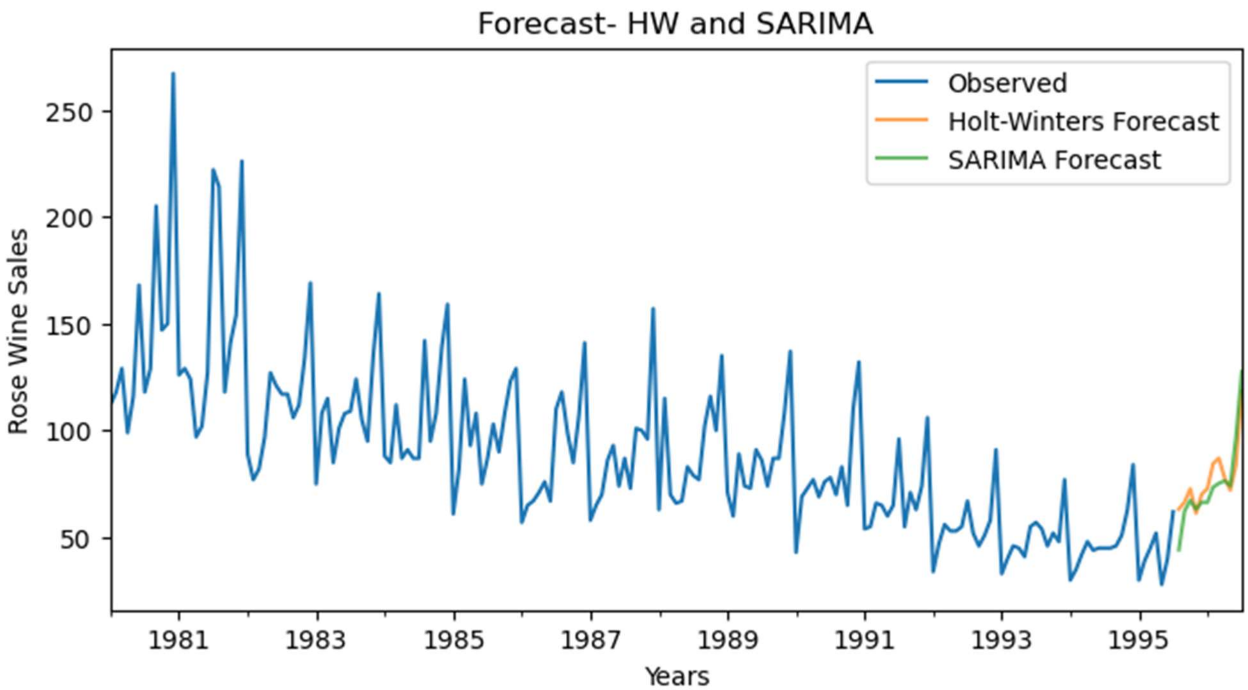

Fig.1.33. Forecast for 12 months- SARIMA and HW

Observations:

- The wine sales peaks during the months of November and december, probably due to the holiday season.
- The sales data exhibits declining trend
- For the Years 1994 and 1995, a slight improvement in sales, especially during the peak seasons is observed
- This is replicated in the forecast

Insights:

- The seasonality component of sales can be capitalized, and can try to push sales in the peak months
- The trend component needs immediate addressing. The reasons for the declining trend need to be investigated and the sales has to be improved.