|  |  |
| --- | --- |
|  |  |
| TIME SERIES FORECASTING PROJECT |  |
|  |  |
|  | Submitted by,VIDYA V |
|  | PGPDSBA.O.2023.B 11.11.2023 |

**CONTENTS**

|  |  |
| --- | --- |
| Rose Wine Sales Dataset | 4 |
| 1. Read the data as an appropriate Time Series data and plot the data. | 4 |
| 1. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. | 5 |
| 1. Split the data into training and test. The test data should start in 1991. | 8 |
| 1. 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, and moving average models should also be built on the training data and check the performance on the test data using RMSE. | 8 |
| 1. 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. | 12 |
| 1. 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. | 14 |
| 1. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data. | 17 |
| 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. | 18 |
| 1. Comment on the model thus built, report your findings, and suggest the measures that the company should be taking for future sales. | 19 |
|  |  |

**List of Figures**

|  |  |
| --- | --- |
| **Name** | **Page No.** |
| Fig.1.1. Rose Dataset | 4 |
| Fig.1.2. Rose Dataset Description | 4 |
| Fig.1.3. Rose wine sales | 5 |
| Fig.1.4 Yearly boxplot- Rose Dataset | 5 |
| Fig.1.5. Monthly Boxplot- Rose wine sales | 6 |
| Fig.1.6. Month-plot- Rose wine sales | 6 |
| Fig.1.7. Monthly sales over the years | 6 |
| Fig.1.8. Yearly sales- Rose Wine | 7 |
| Fig.1.9. Additive Decomposition- Rose sales | 7 |
| Fig.1.10. Multiplicative Decomposition- Rose wine sales | 8 |
| Fig.1.11 Linear Regression model test forecast plot- Rose sales | 8 |
| Fig.1.12. Naive forecast of test data- Rose wine sales | 9 |
| Fig.1.13. Simple Average forecast of test data- Rose wine sales | 9 |
| Fig.1.14.Moving Average forecast of test data- Rose wine sales | 10 |
| Fig.1.15**.** Simple Exponential smoothing forecast of test data- Rose wine sales | 10 |
| Fig.1.16. Simple Exponential smoothing forecast of test data- Rose wine sales optimized for lowest RMSE | 11 |
| Fig.1.17**.** Holt forecast of test data- Rose wine sales | 11 |
| Fig.1.18. Holt forecast of test data- Rose wine sales - optimized for lowest RMSE | 11 |
| Fig.1.19**.** Holt Winters forecast of test data- Rose wine sales | 12 |
| Fig.1.20. Holt Winters smoothing forecast of test data- Rose wine sales - optimized for lowest RMSE | 12 |
| Fig.1.21. Rose dataset -Stationarity test rolling mean and Standard deviation plots | 13 |
| Fig.1.22. Dickey Fuller Test results- Rose Dataset | 13 |
| Fig.1.23. Differenced series -Stationarity test rolling mean and Standard deviation plots | 14 |
| Fig.1.24. Dickey Fuller Test results | 14 |
| Fig.1.25. ARIMA results Summary- Rose Dataset | 15 |
| Fig.1.26. ARIMA model forecast on test data- Rose Dataset | 15 |
| Fig.1.27. SARIMA results Summary- Rose Dataset | 16 |
| Fig.1.28. SARIMA model forecast on test data- Rose Dataset | 16 |
| Fig.1.29. Diagnostics and Correlogram- SARIMA Model | 17 |
| Fig.1.30. Rose Dataset model Results- Test RMSE | 17 |
| Fig.1.31. SARIMA Model forecast for next 12 months- Rose Dataset | 18 |
| Fig.1.32. Holt Winters Model forecast for next 12 months- Rose Dataset | 18 |
| Fig.1.33. Forecast for 12 months- SARIMA and HW | 19 |

# 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.**

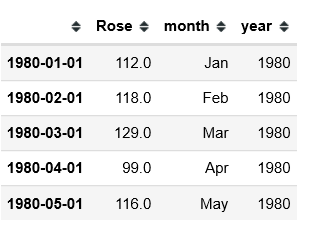


Fig.1.1. Rose Dataset



Fig.1.2. Rose Dataset Description

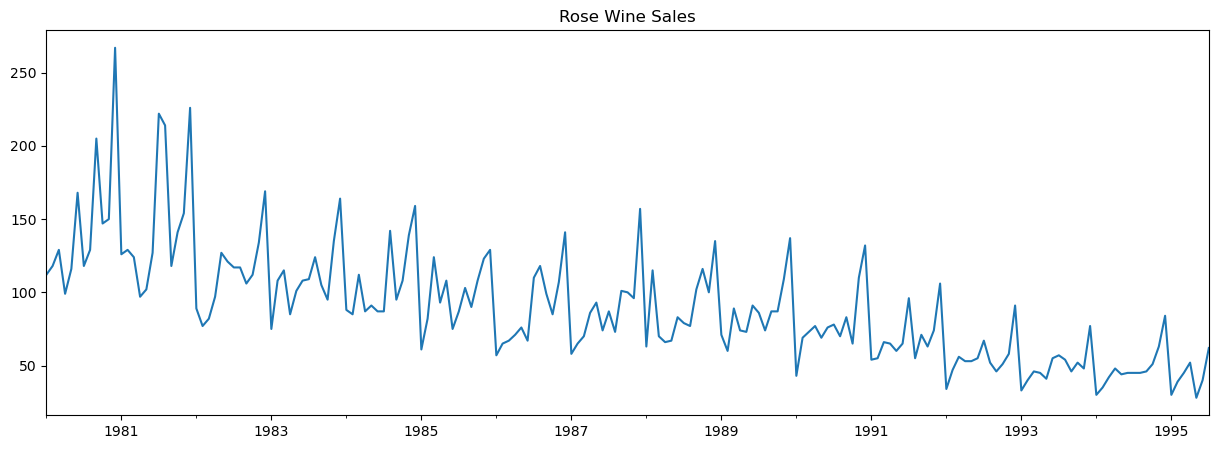


Fig.1.3. Rose wine sales

**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

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

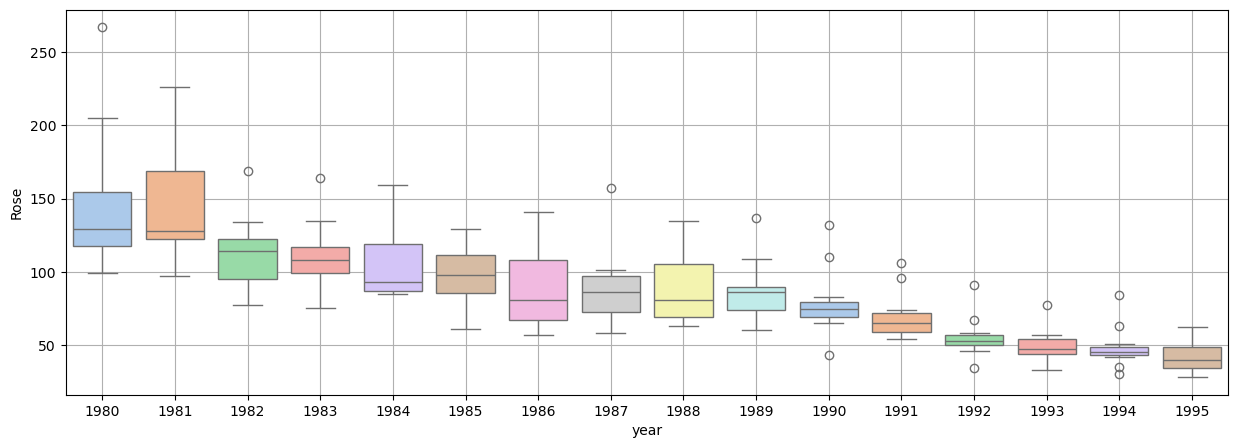


Fig.1.4 Yearly boxplot- Rose Dataset

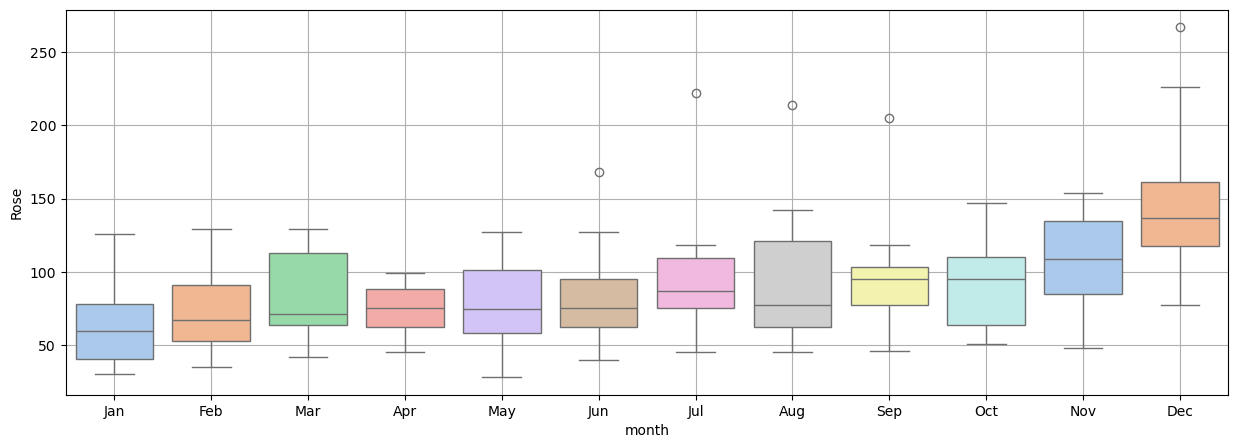


Fig.1.5. Monthly Boxplot- Rose wine sales

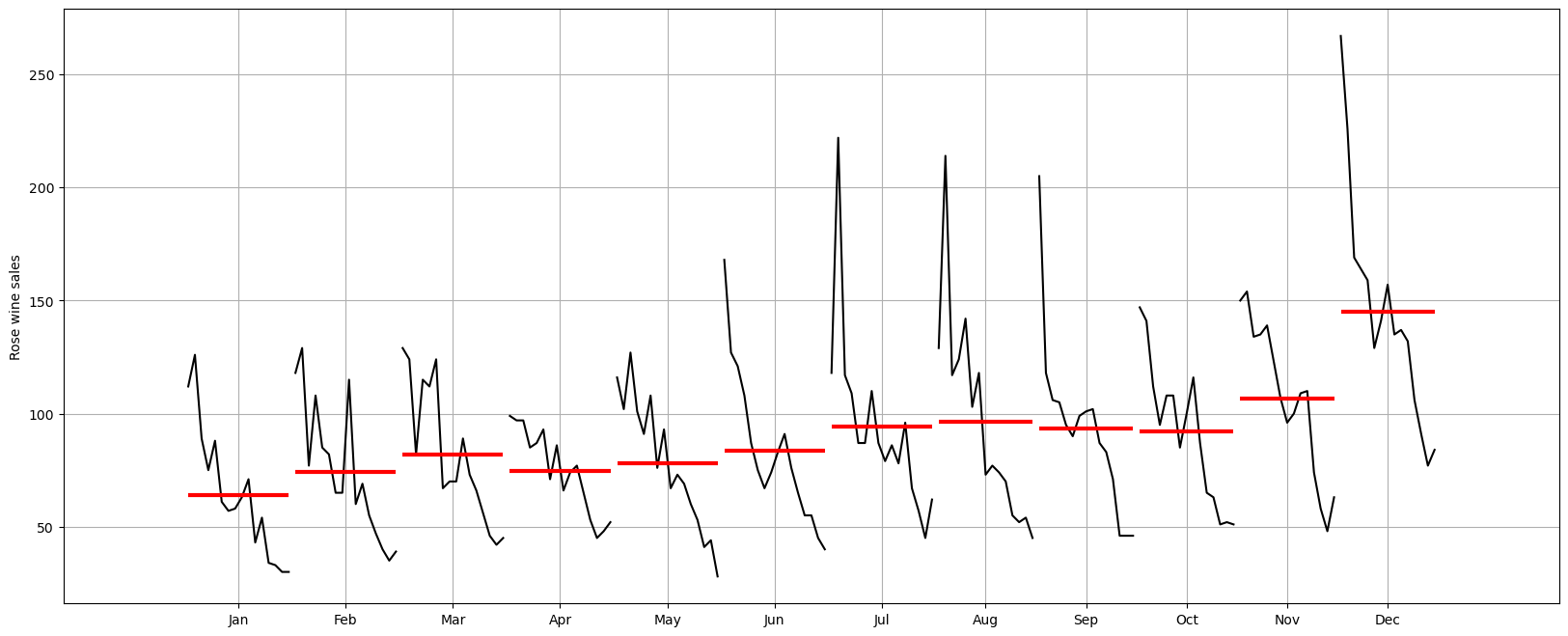


Fig.1.6. Month-plot- Rose wine sales

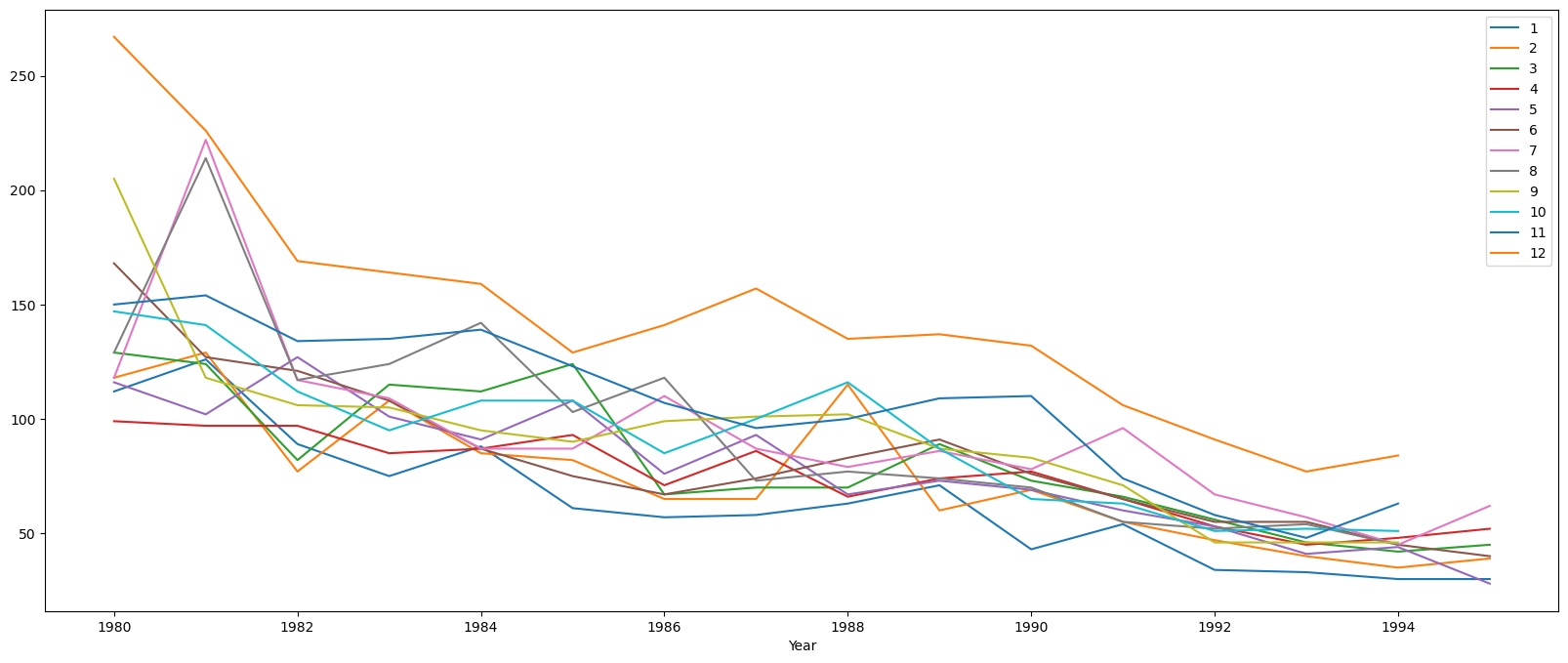


Fig.1.7. Monthly sales over the years

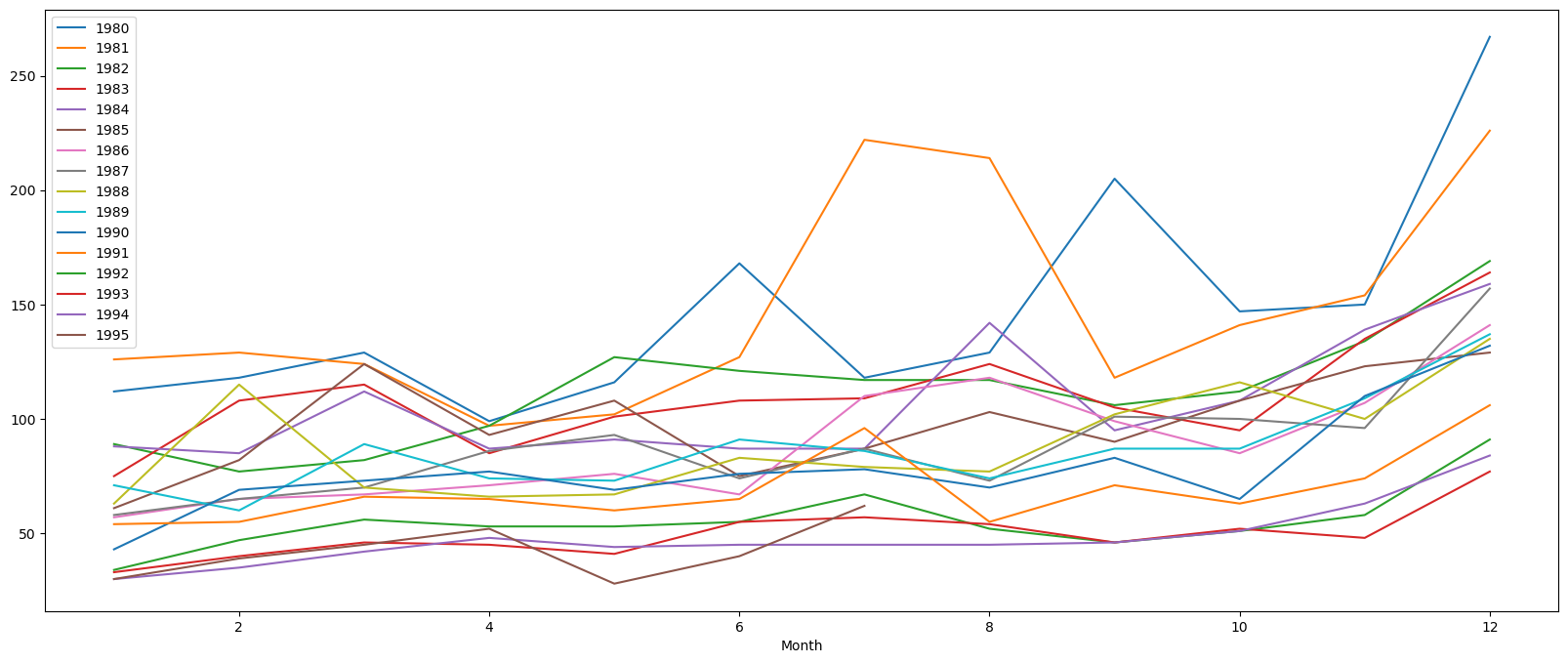


Fig.1.8. Yearly sales- Rose Wine

**Observations:**

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

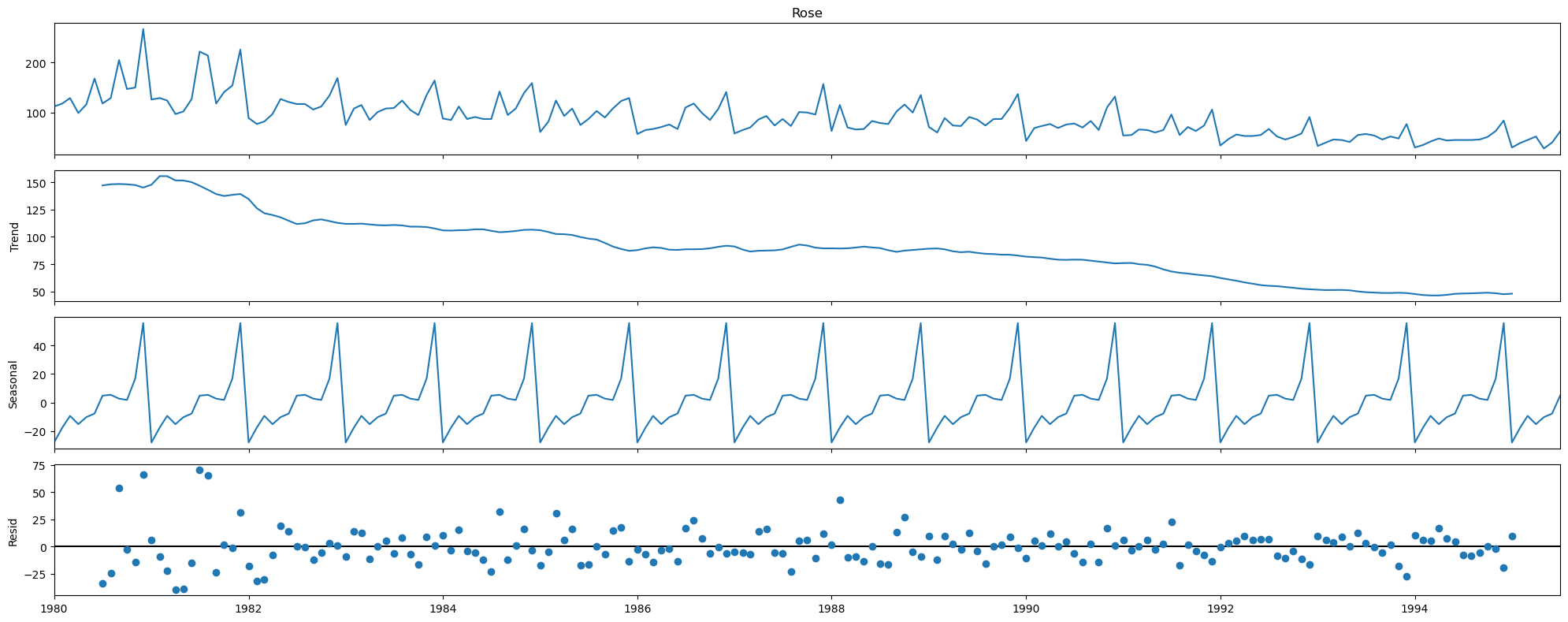
****

Fig.1.9. Additive Decomposition- Rose sales

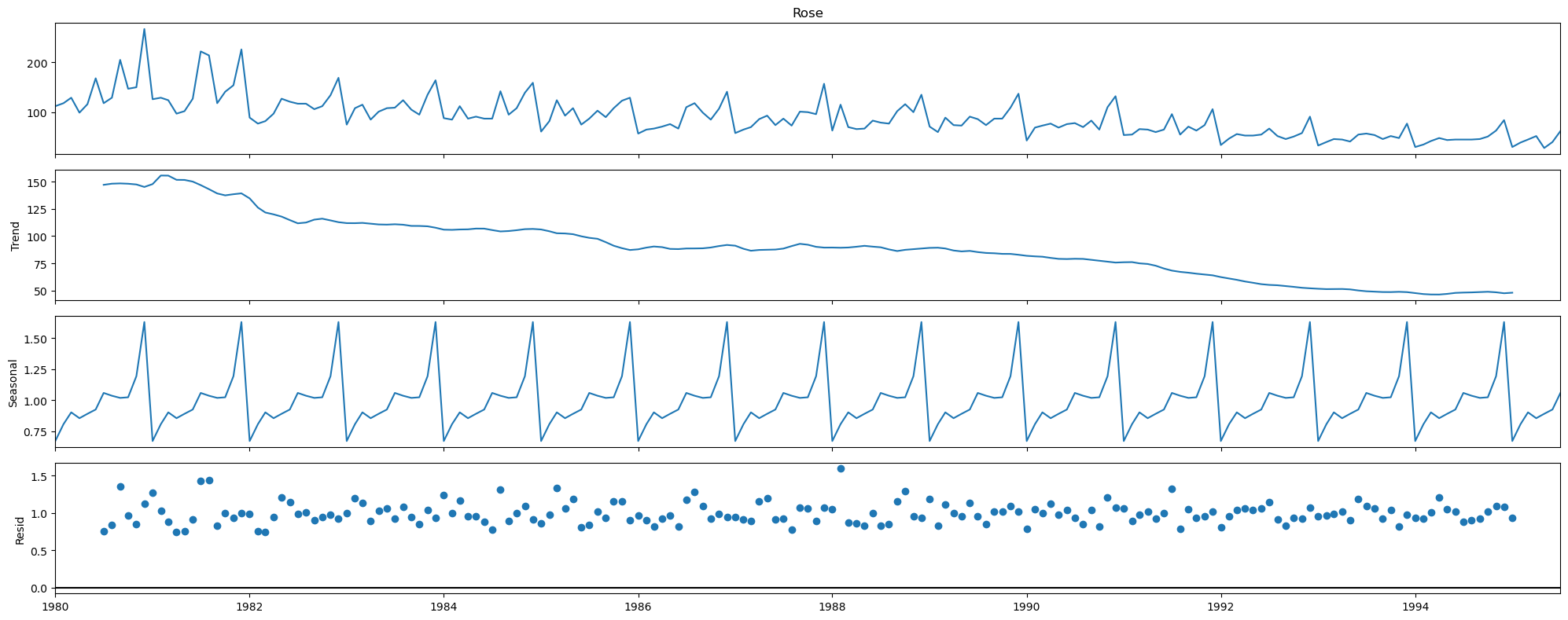


Fig.1.10. Multiplicative Decomposition- Rose wine sales

**Observations:**

- Clear seasonality component observed

- The three conditions for multiplicative seasonality are fulfilled, and hence we can assume multiplicative seasonality

1. **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

1. **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.**
   1. **Simple Models**
      1. **Linear Regression**

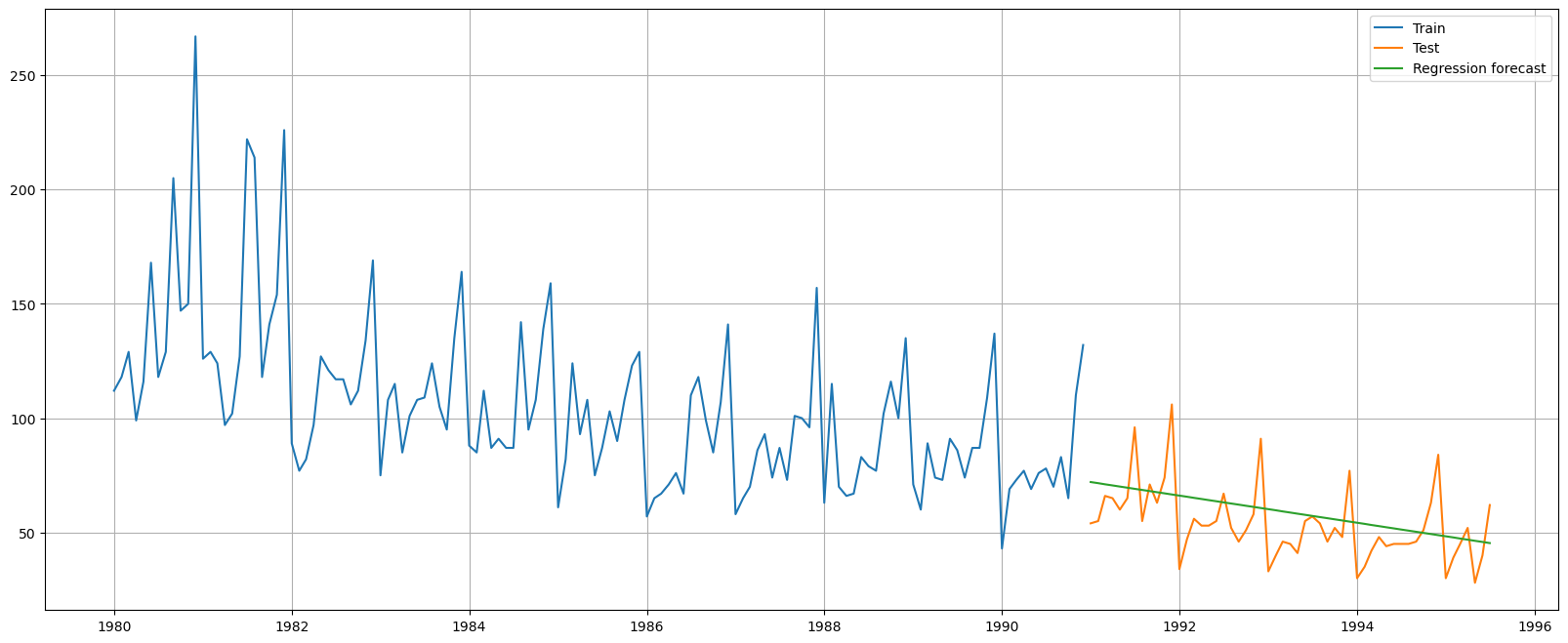
****

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

* + 1. **Naïve Forecast**

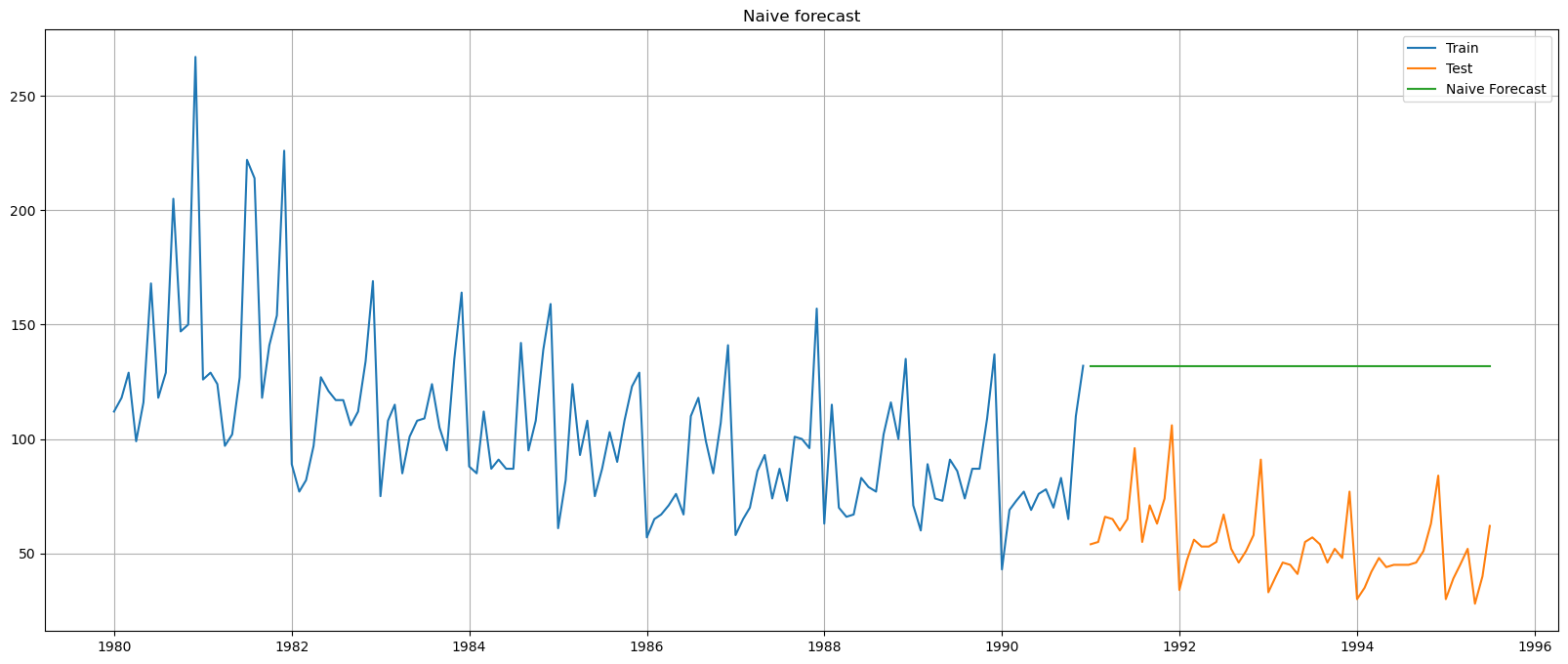
****

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

* + 1. **Simple Average**

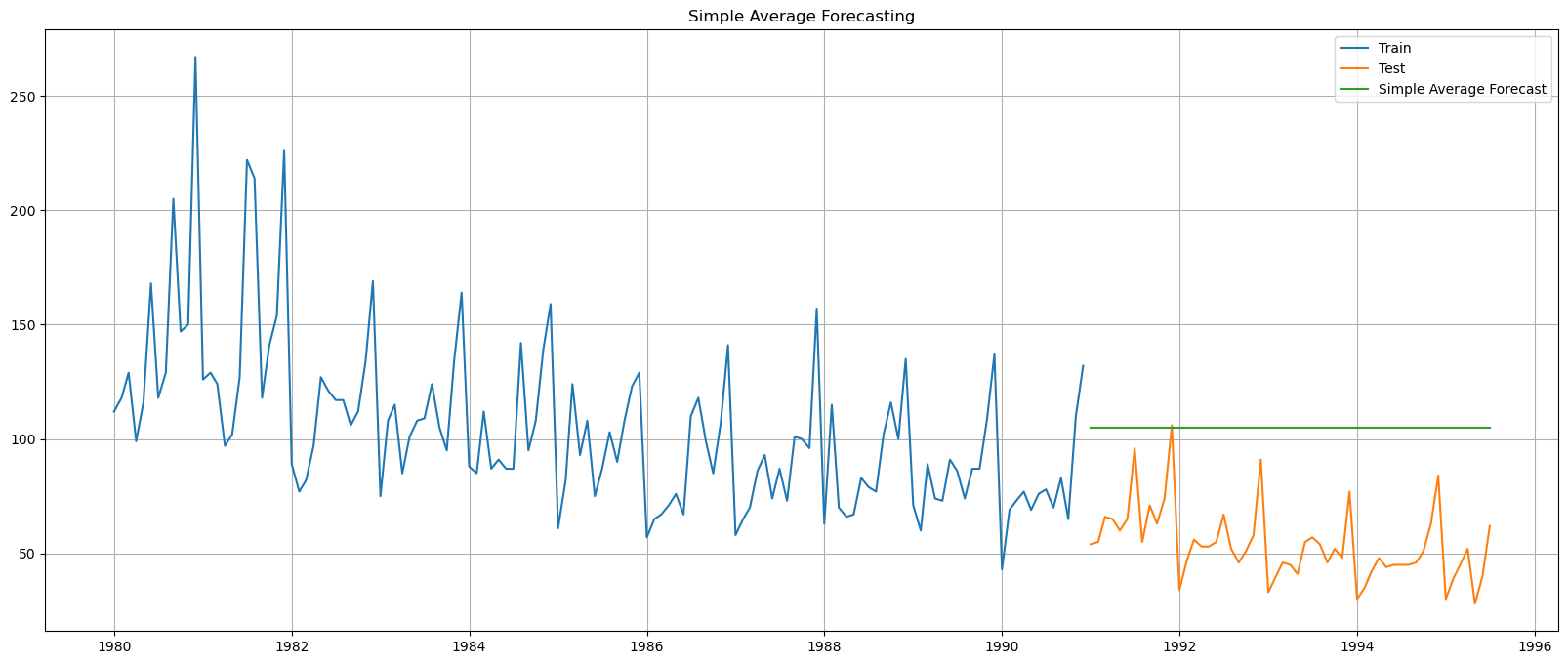
****

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

* + 1. **Moving Average**

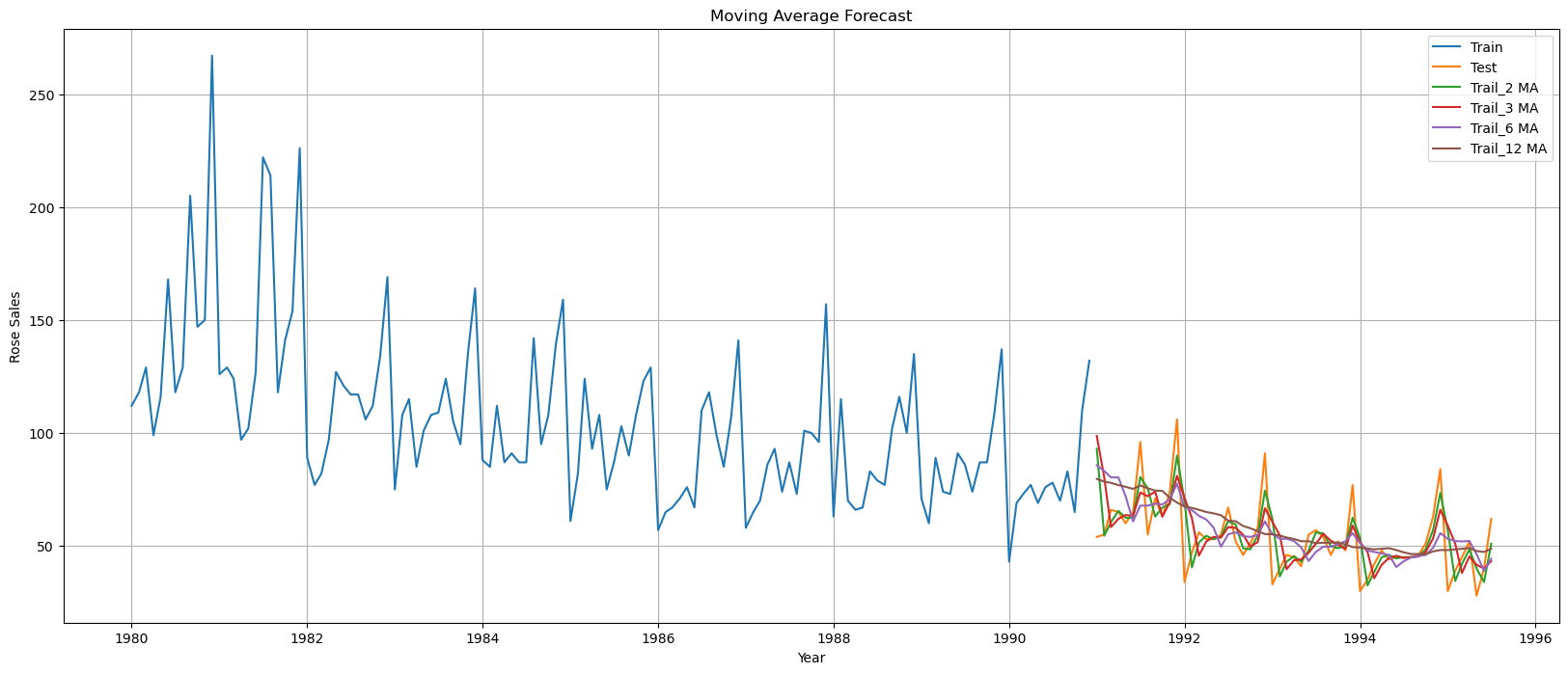
****

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

Observations:

* + - Best fit occurs in MA trail 2 model
  1. **Exponential Smoothing Models**
     1. **Simple Exponential Smoothing**

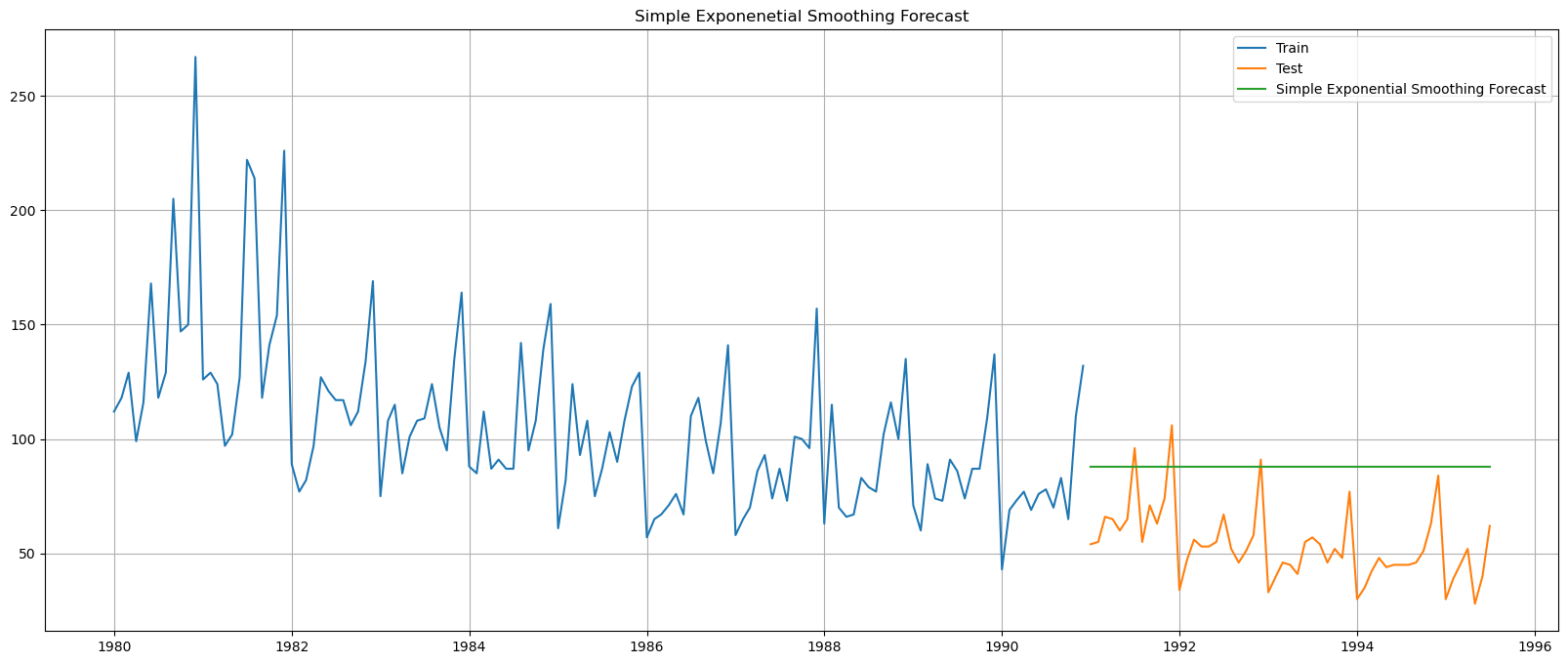
****

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

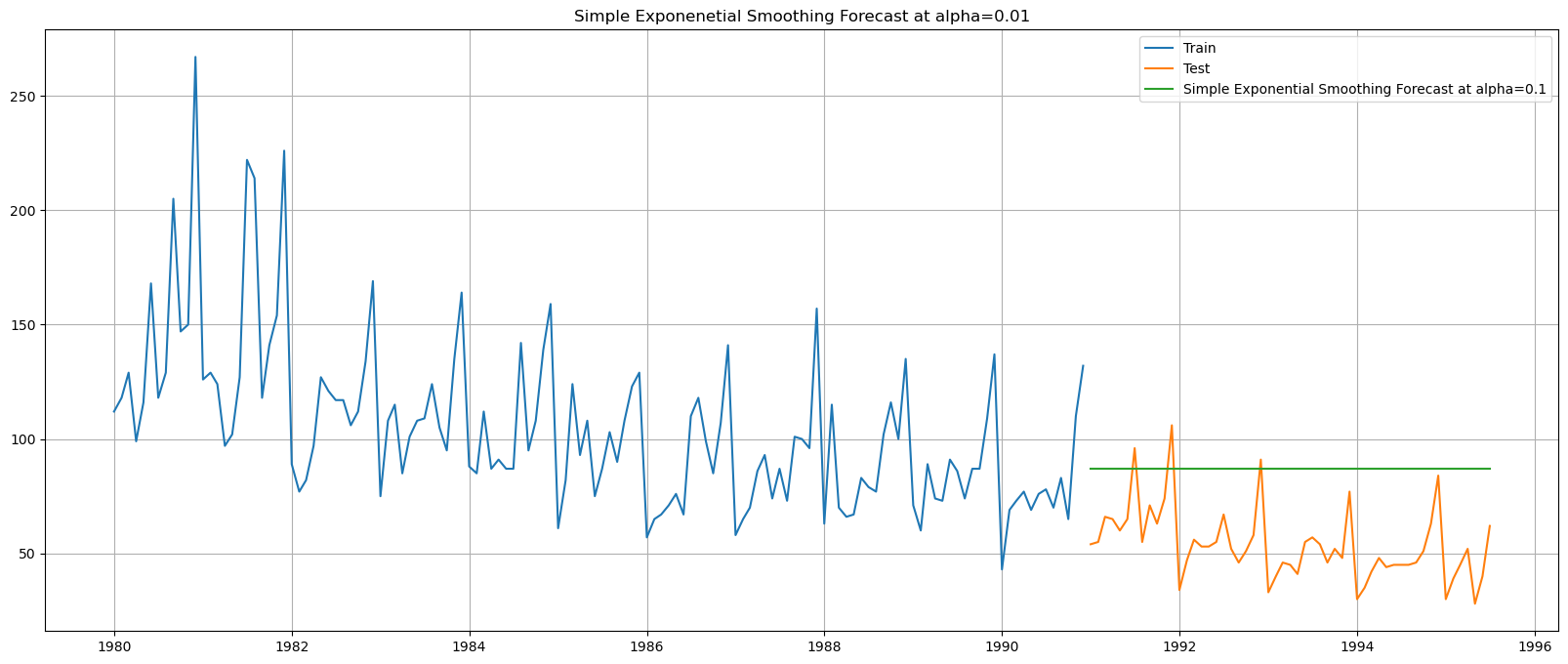
****

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
    1. **Holt Double Exponential Smoothing**

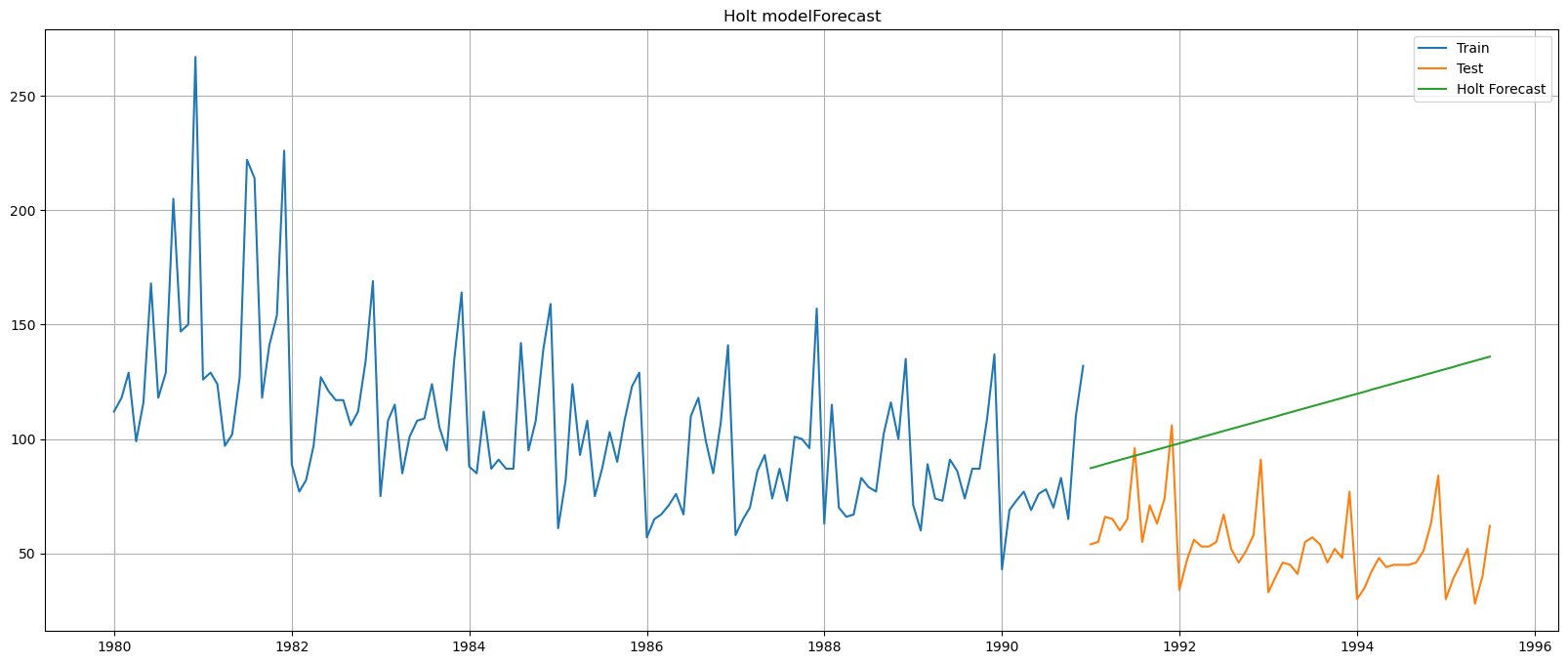
****

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

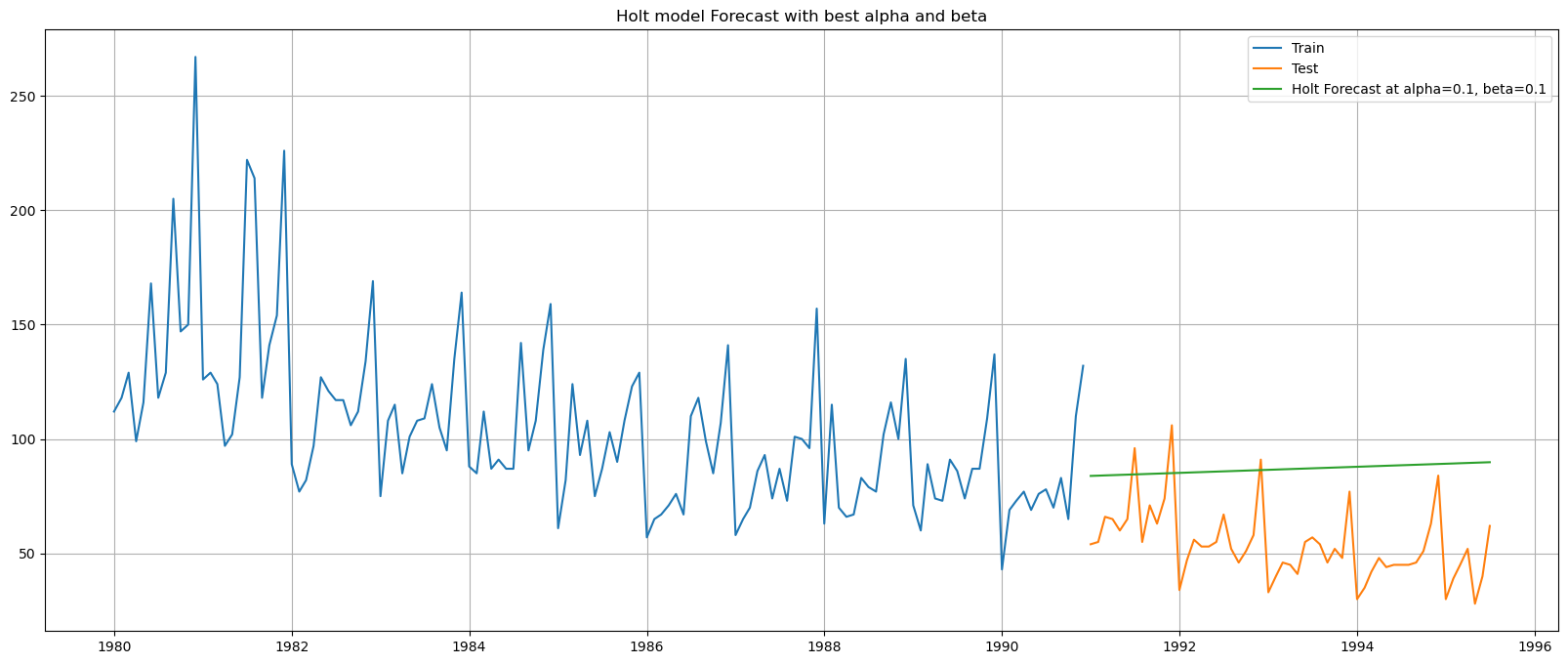
****

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
    1. **Holt-Winters Triple Exponential Smoothing**

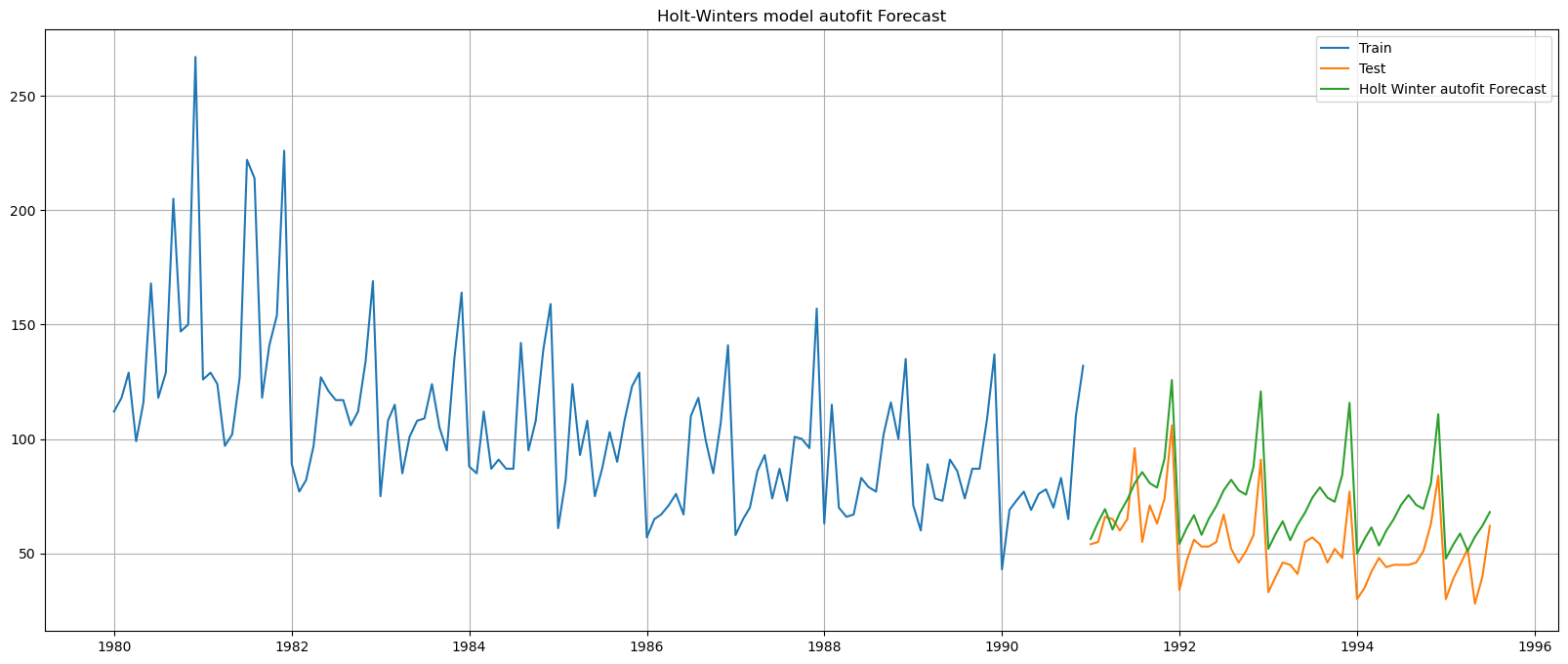
****

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

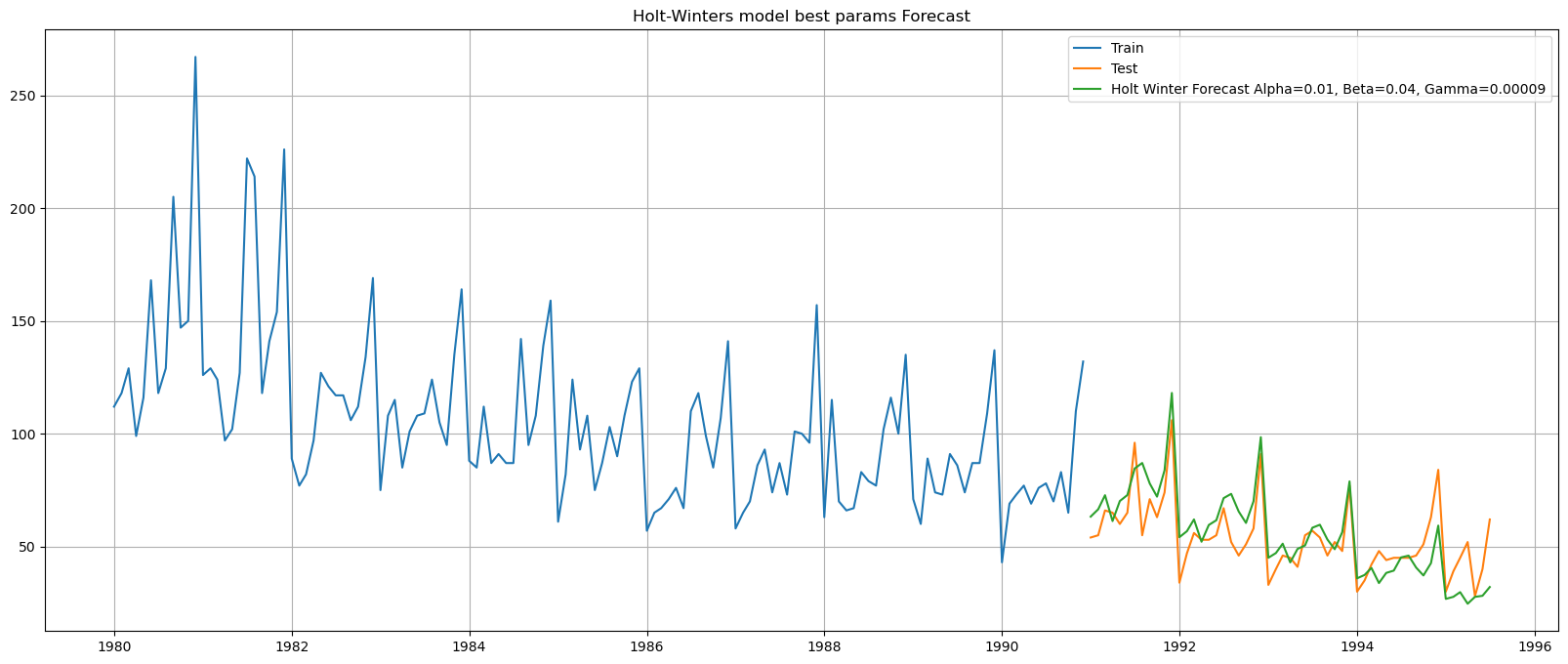


Fig.1.20. Holt Winters smoothing forecast of test data- Rose wine sales - optimized for lowest RMSE

**Observations:**

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

1. **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**

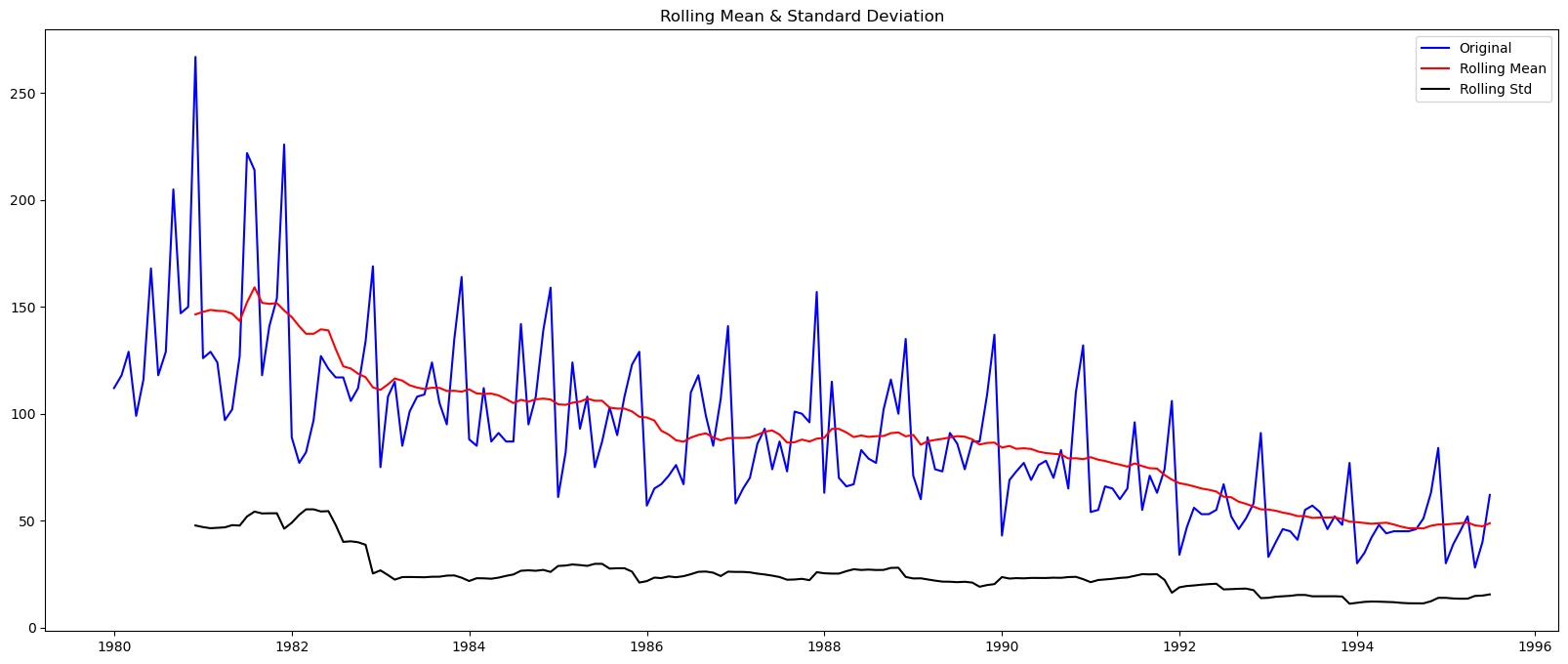
****

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

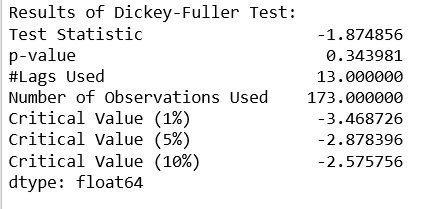


Fig.1.22. Dickey Fuller Test results- Rose Dataset

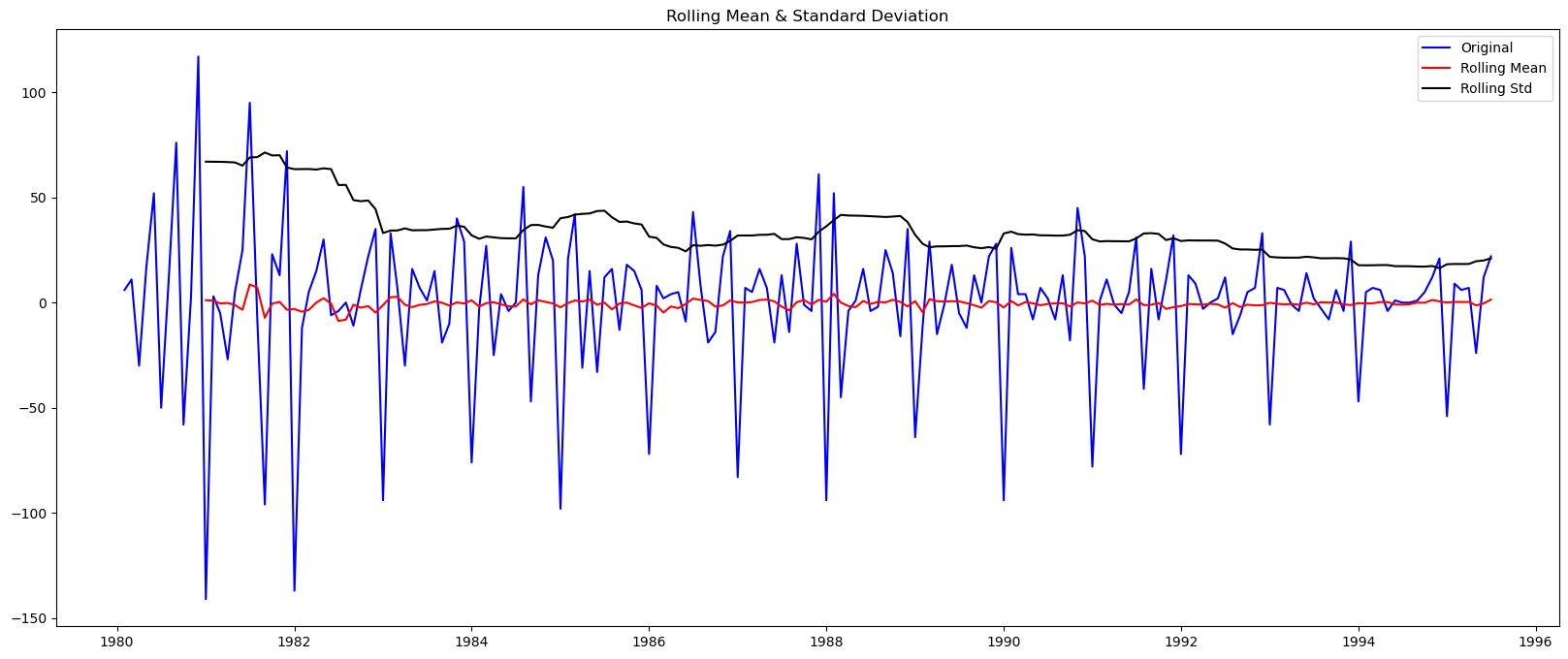
****

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

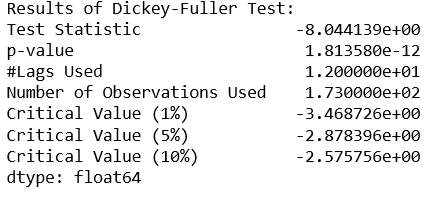


Fig.1.24. Dickey Fuller Test results- Differenced series

**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.

1. **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.**
   1. **ARIMA model**

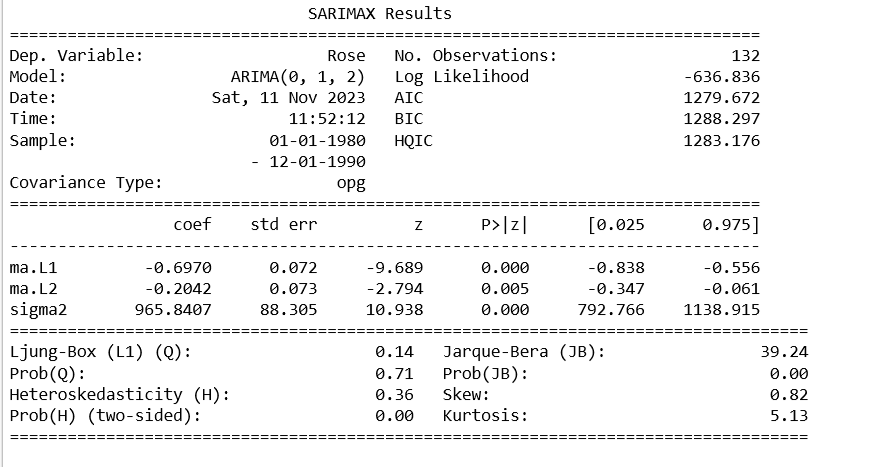


Fig.1.25. ARIMA results Summary- Rose Dataset

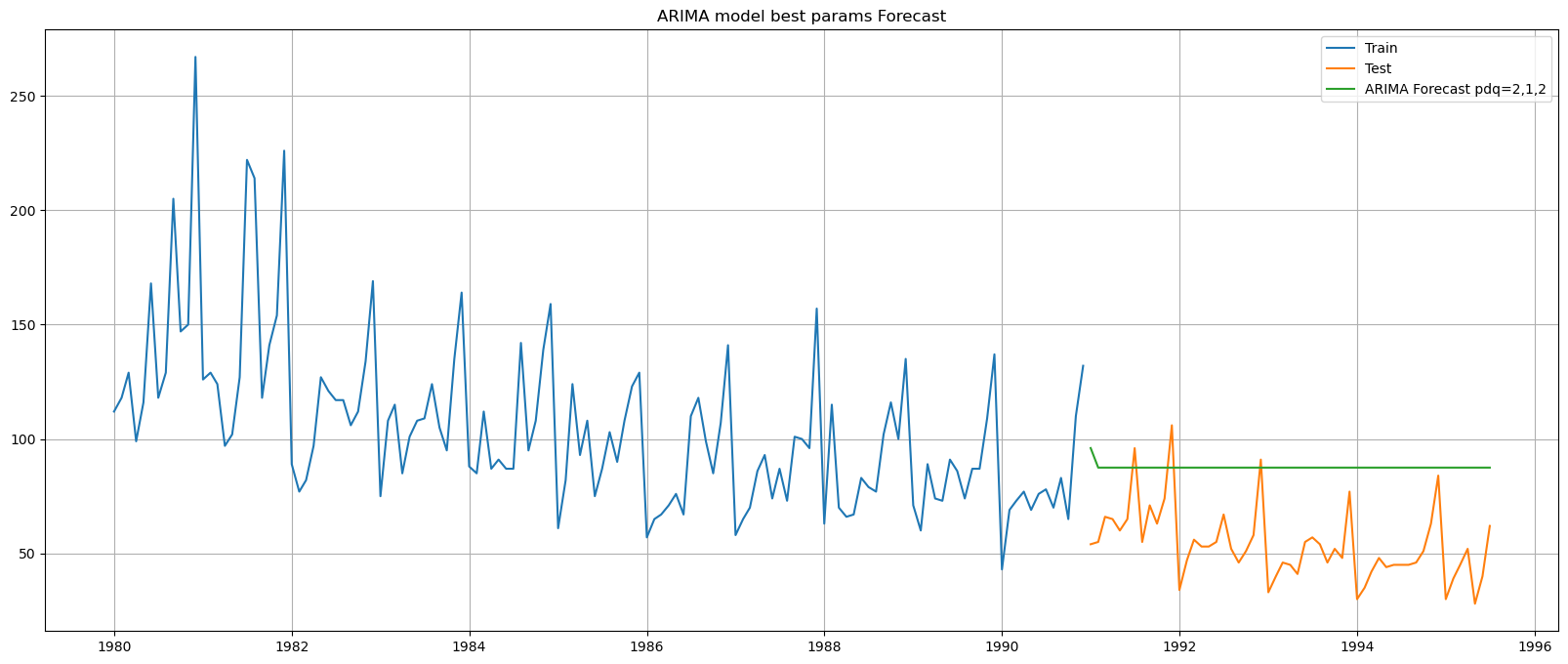


Fig.1.26. ARIMA model forecast on test data- Rose Dataset

**Observations:**

* + Lowest AIC obtained for (p,d,q)=(0,1,2)
  + This is consistent with the d=1 obtained during stationarity check
  1. **SARIMA model**

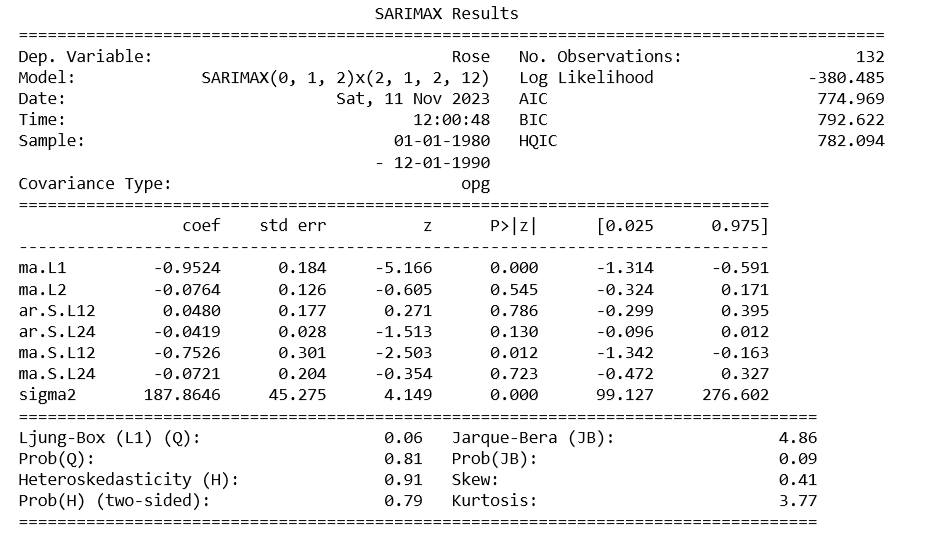


Fig.1.27. SARIMA results Summary- Rose Dataset

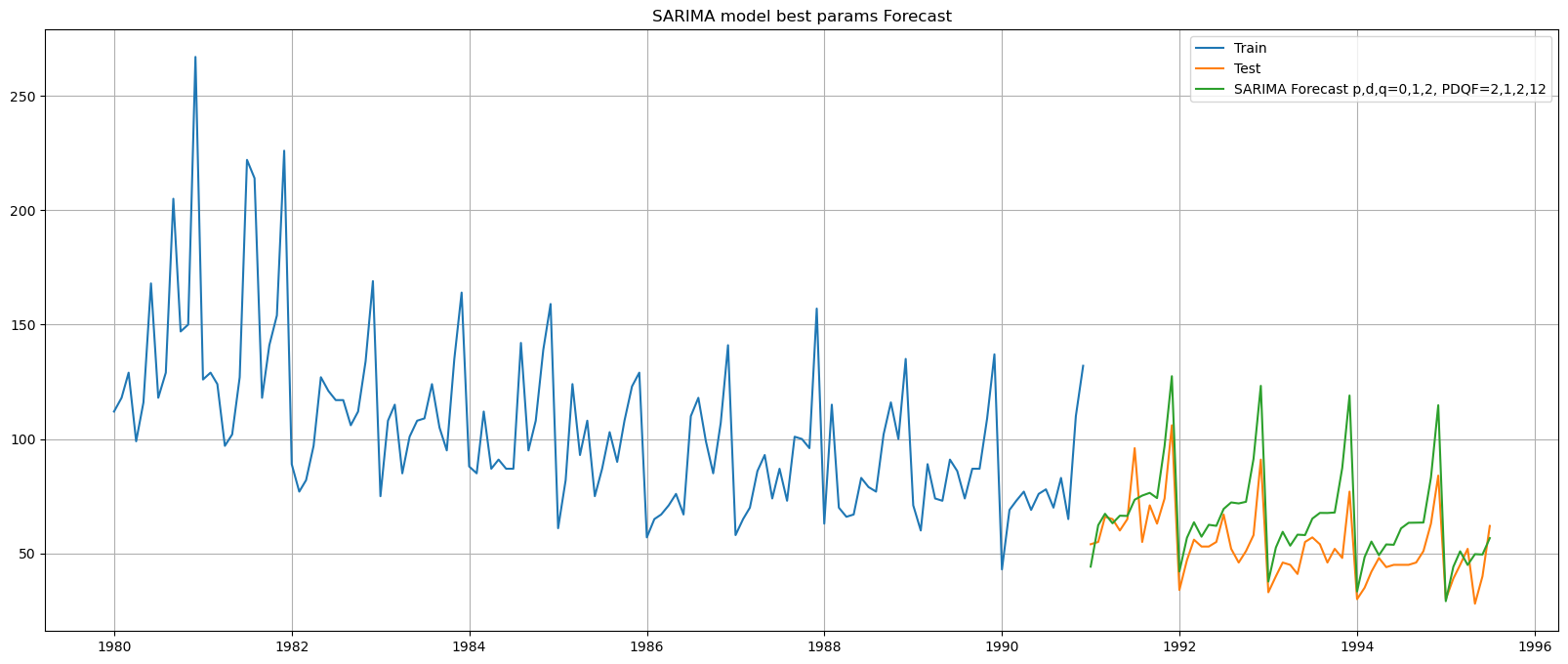
****

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

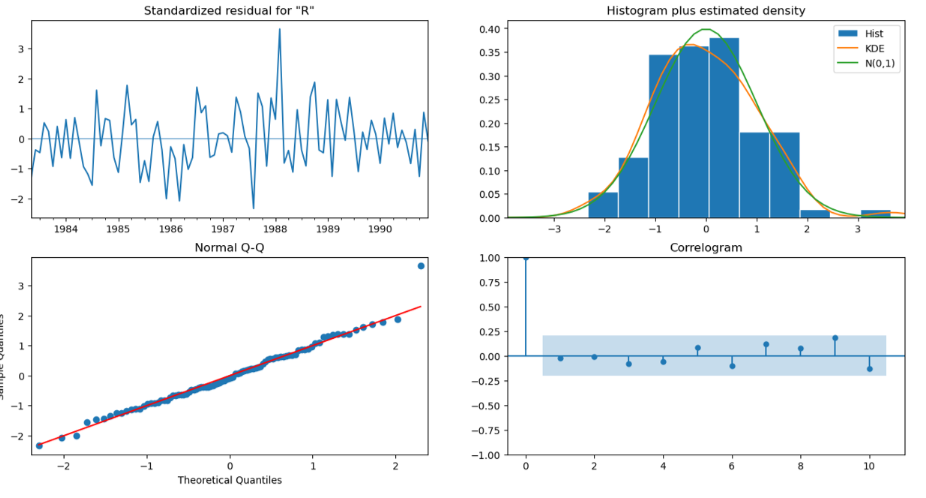


Fig.1.29. Diagnostics and Correlogram- SARIMA Model

**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

1. **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.**

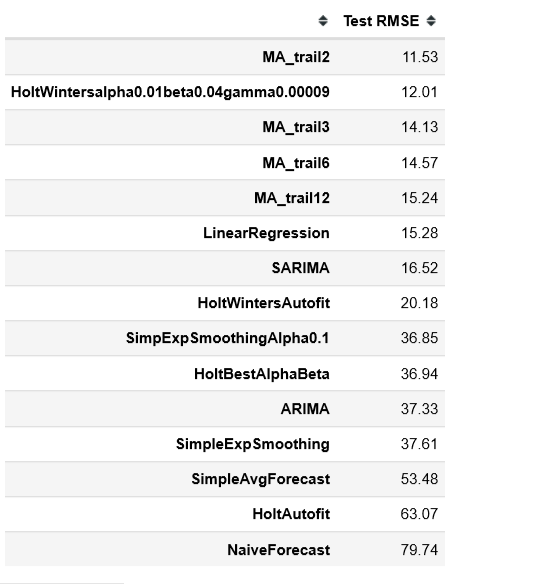


Fig.1.30. Rose Dataset model Results- Test RMSE

**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

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.**

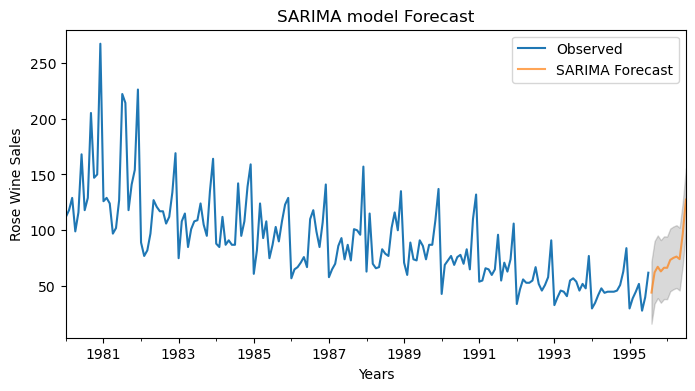


Fig.1.31. SARIMA Model forecast for next 12 months- Rose Dataset

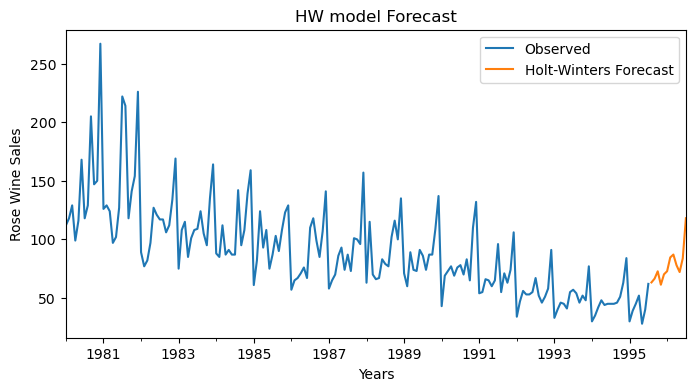


Fig.1.32. Holt Winters Model forecast for next 12 months- Rose Dataset

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

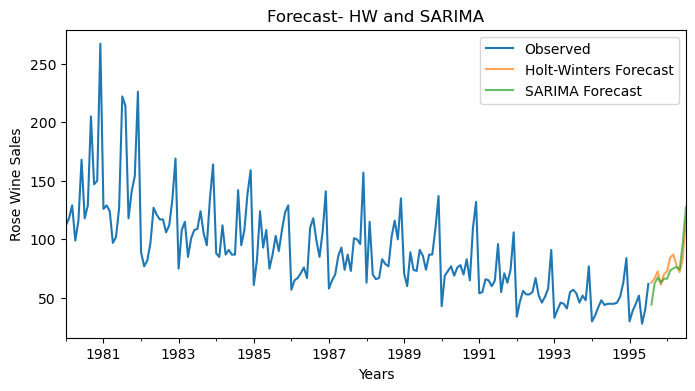


Fig.1.33. Forecast for 12 months- SARIMA and HW

**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.