TIME SERIES FORECASTING PROJECT

Submitted by,
VIDYA V

PGPDSBA.O.2023.B
11.11.2023

CONTENTS

List of Figures

Sparkling 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.**

	Sparkling	month	year
1980-01-01	1686	Jan	1980
1980-02-01	1591	Feb	1980
1980-03-01	2304	Mar	1980
1980-04-01	1712	Apr	1980
1980-05-01	1471	May	1980

Fig.1.1. Sparkling Dataset

	Sparkling	year
count	187.000000	187.000000
mean	2402.417112	1987.299465
std	1295.111540	4.514749
min	1070.000000	1980.000000
25%	1605.000000	1983.000000
50%	1874.000000	1987.000000
75%	2549.000000	1991.000000
max	7242.000000	1995.000000

Fig.1.2. Sparkling Dataset Description

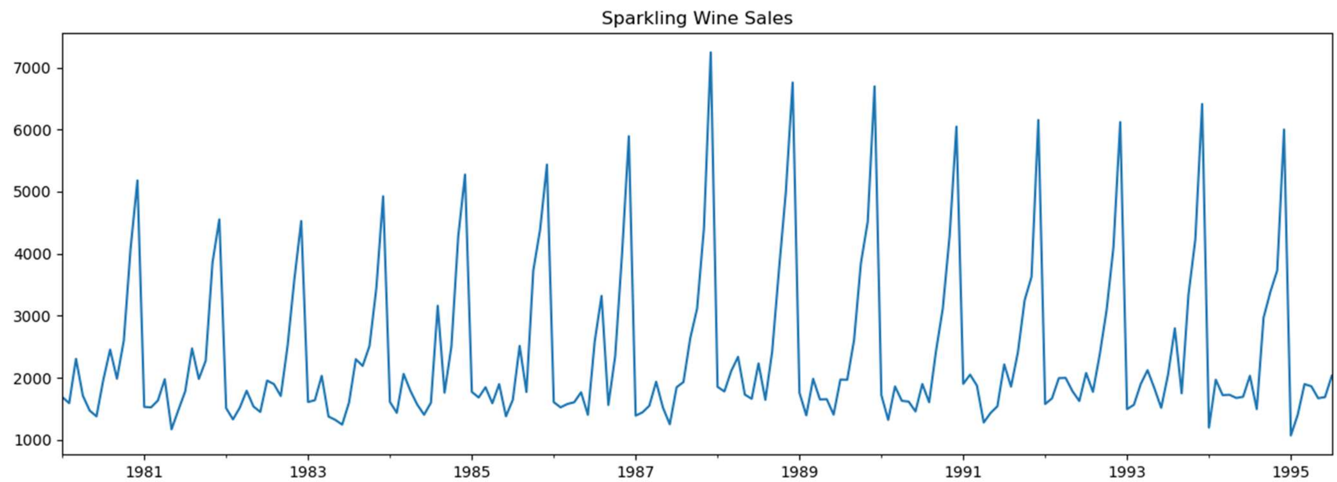

Fig.1.3. Sparkling wine sales

Observations:

- The plot represents the Sparkling wine sales from Jan 1980 to July 1995, covering a span of 15.5 years- 187 values
- There seems to be some seasonality associated with this plot.
- The minimum sales was 1070, the maximum sales was 7242, with a mean of 2402
- There are no null values

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

2.1. EDA

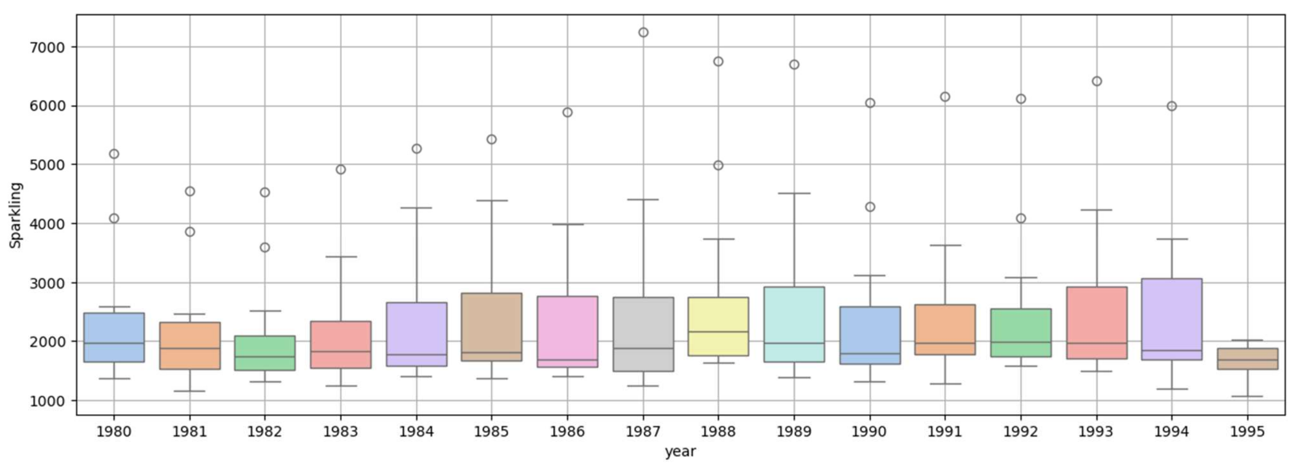

Fig.1.4 Yearly boxplot- Sparkling Dataset

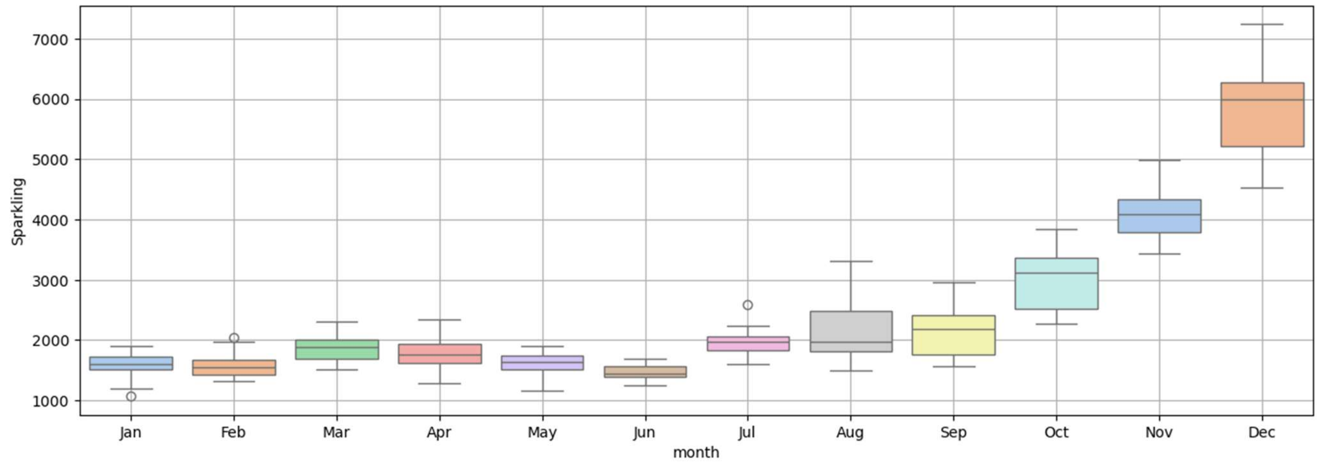

Fig.1.5. Monthly Boxplot- Sparkling wine sales

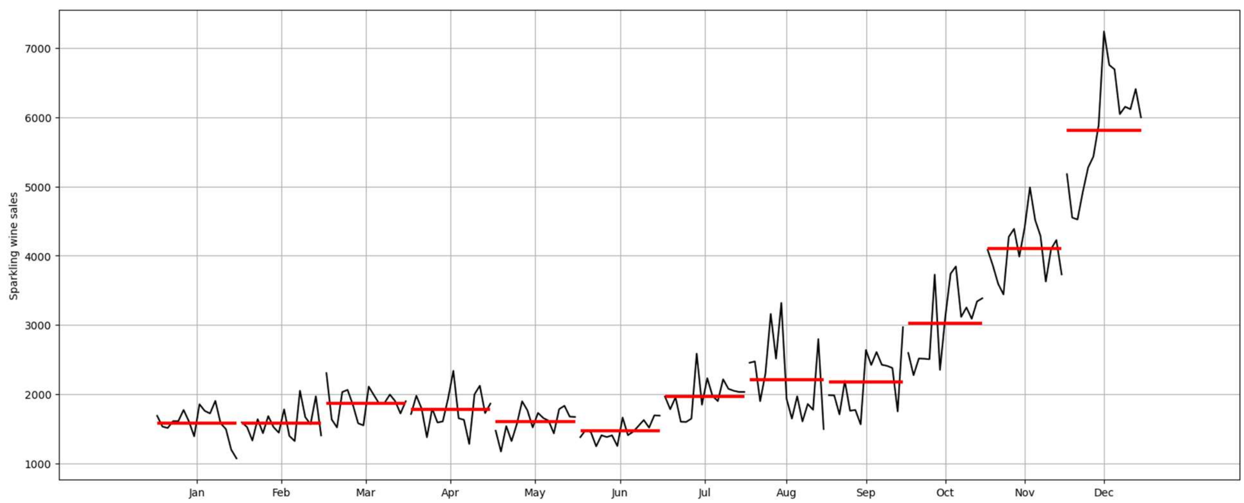

Fig.1.6. Month-plot- Sparkling wine sales

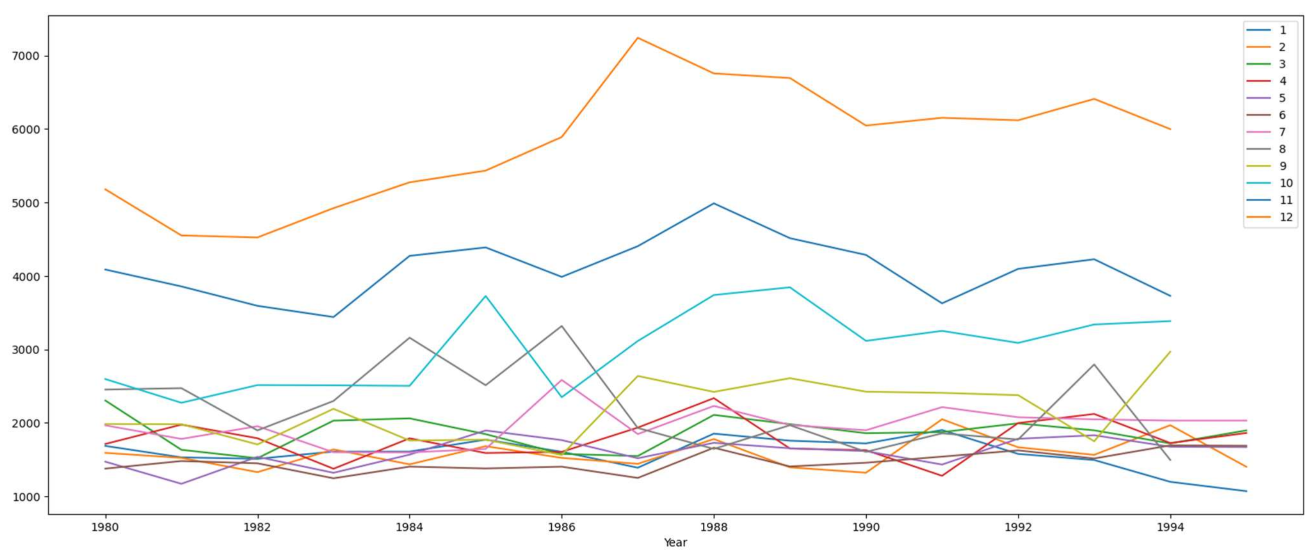

Fig.1.7. Monthly sales over the years

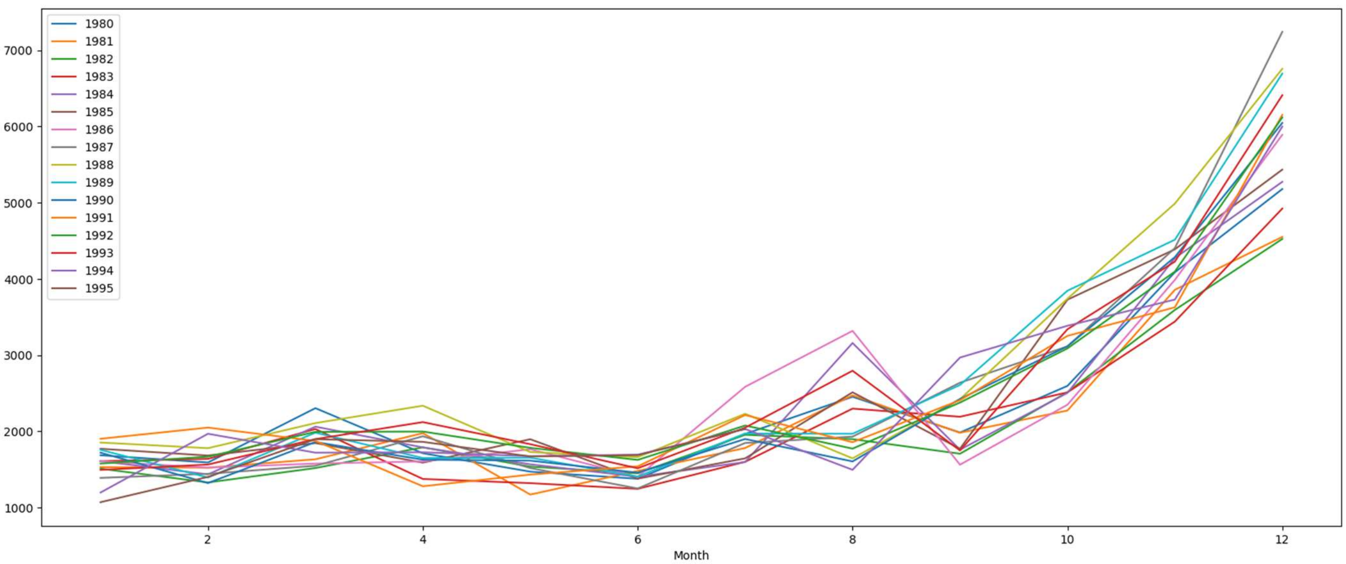

Fig.1.8. Yearly sales- Sparkling Wine

Observations:

- The sales remains low for the first half of the year, and increases in the second half
- Peak sales is in the month of December
- The variability also changes from Jan to Dec. Seasonality is indicated
- No trend can be discerned from the plots

2.2. Decomposition

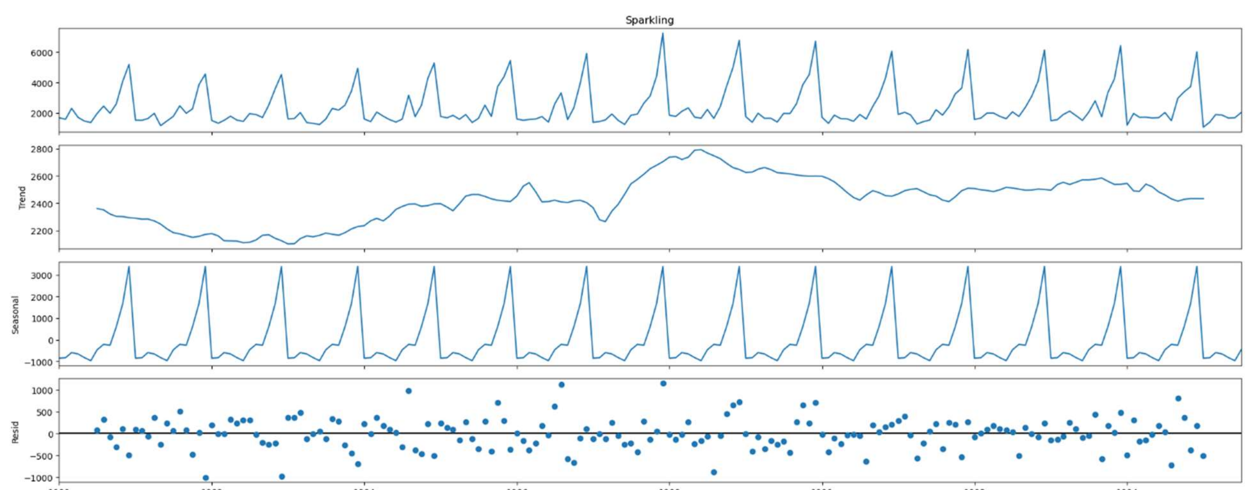

Fig.1.9. Additive Decomposition- Sparkling sales

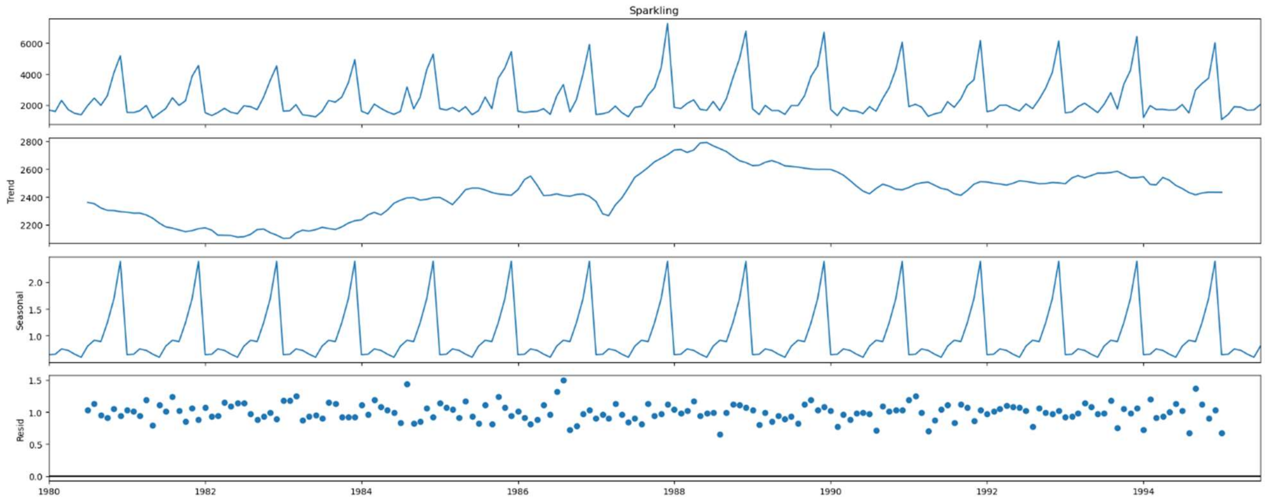

Fig.1.10. Multiplicative Decomposition- Sparkling wine sales

Observations:

- Clear seasonality component observed in both types of decomposition
- The residual plots of both the decompositions look similar. Hence, we can adopt the simpler of the two- additive 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

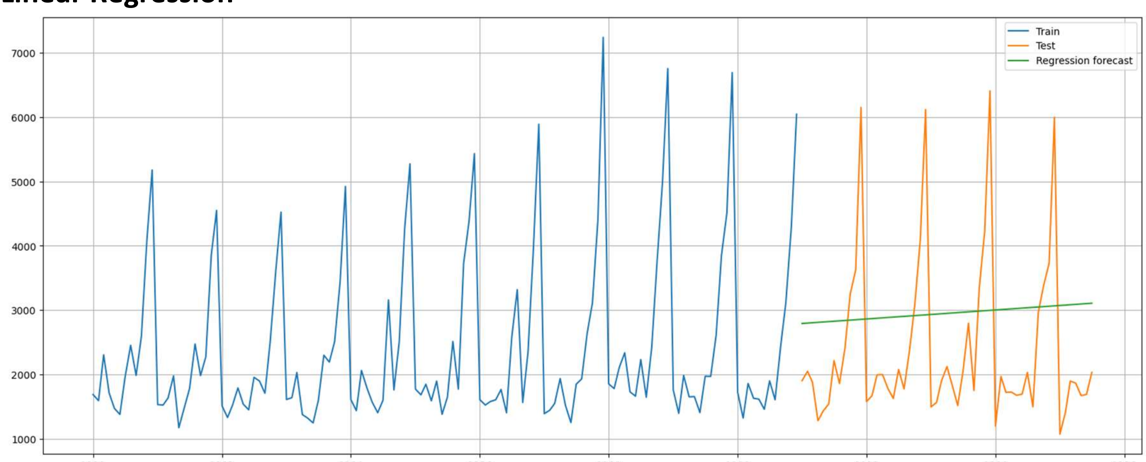

Fig.1.11 Linear Regression model test forecast plot- Sparkling sales

4.1.2. Naïve Forecast

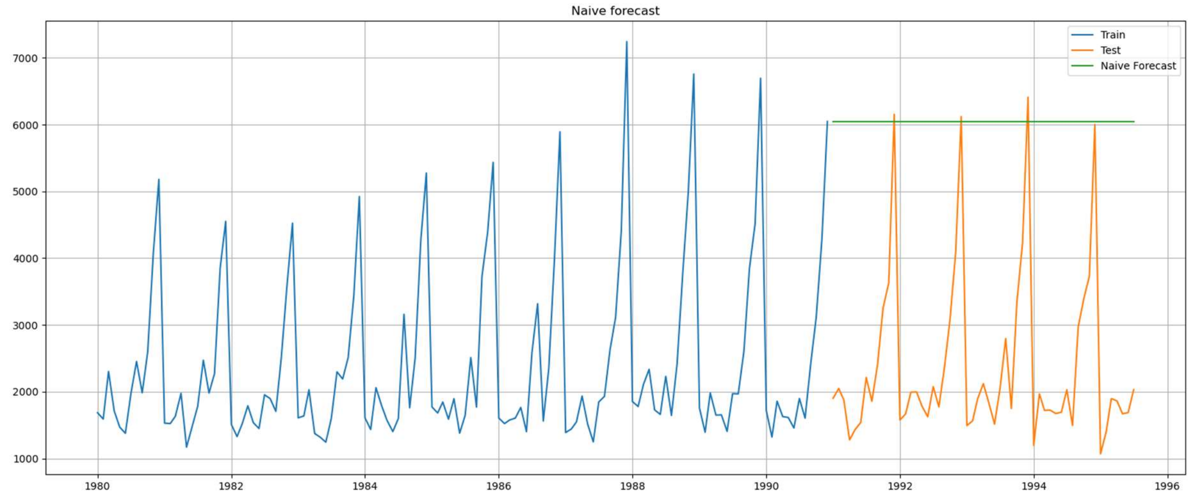

Fig.1.12. Naive forecast of test data- Sparkling wine sales

4.1.3. Simple Average

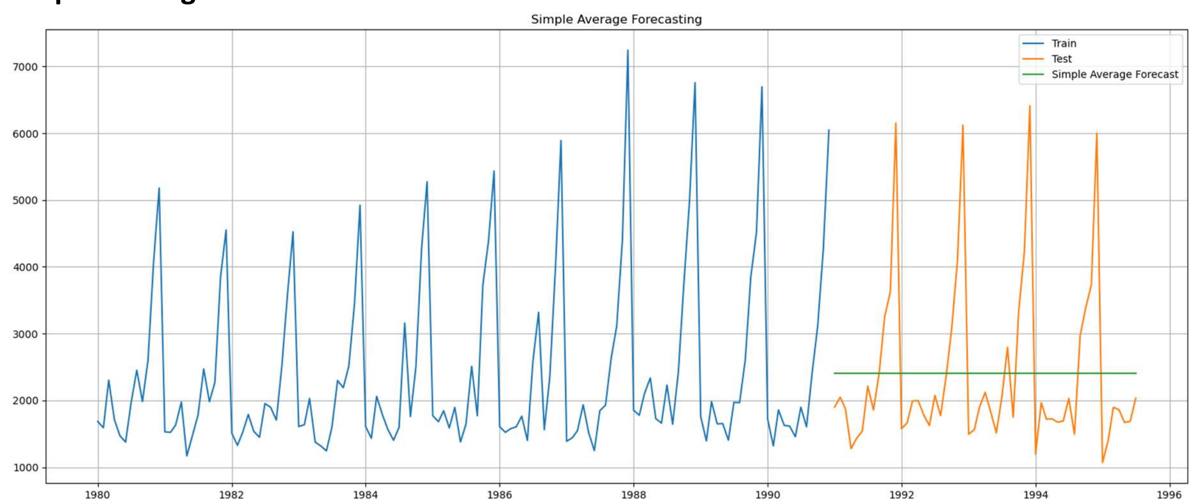

Fig.1.13. Simple Average forecast of test data- Sparkling wine sales

4.1.4. Moving Average

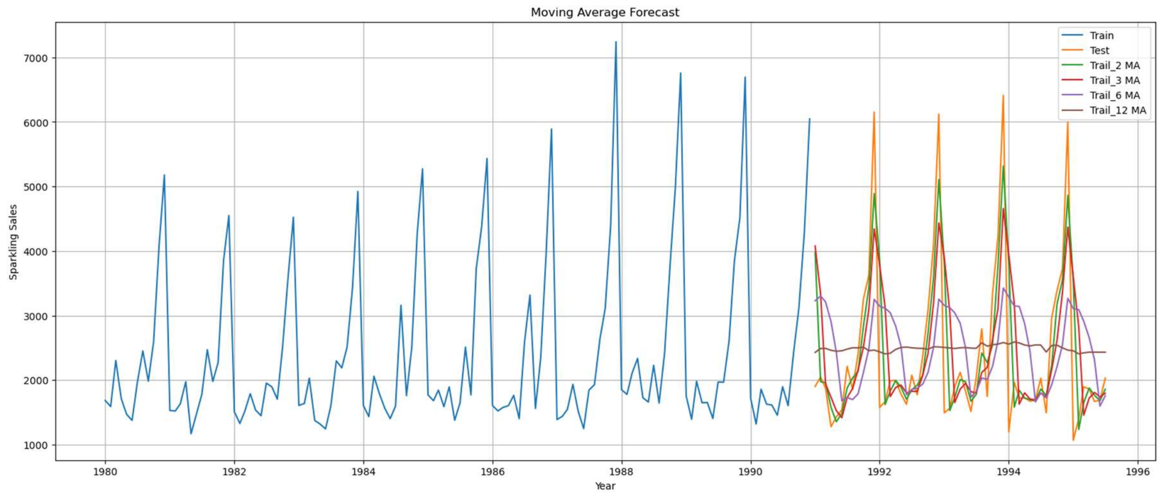

Fig.1.14. Moving Average forecast of test data- Sparkling wine sales

Observations:

- Best fit occurs in MA trail 2 model

4.2. Exponential Smoothing Models

4.2.1. Simple Exponential Smoothing

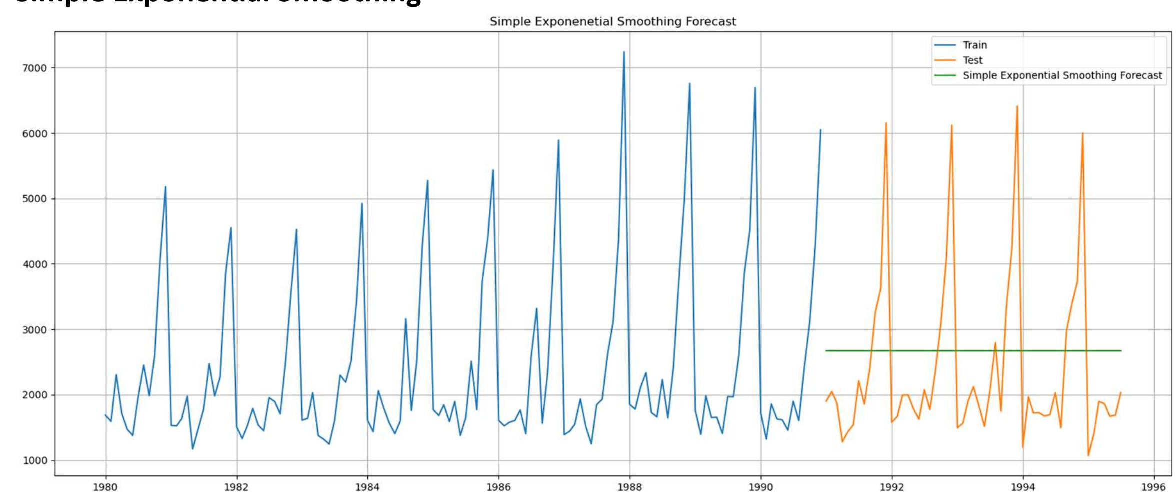

Fig.1.15. Simple Exponential smoothing forecast of test data- Sparkling wine sales

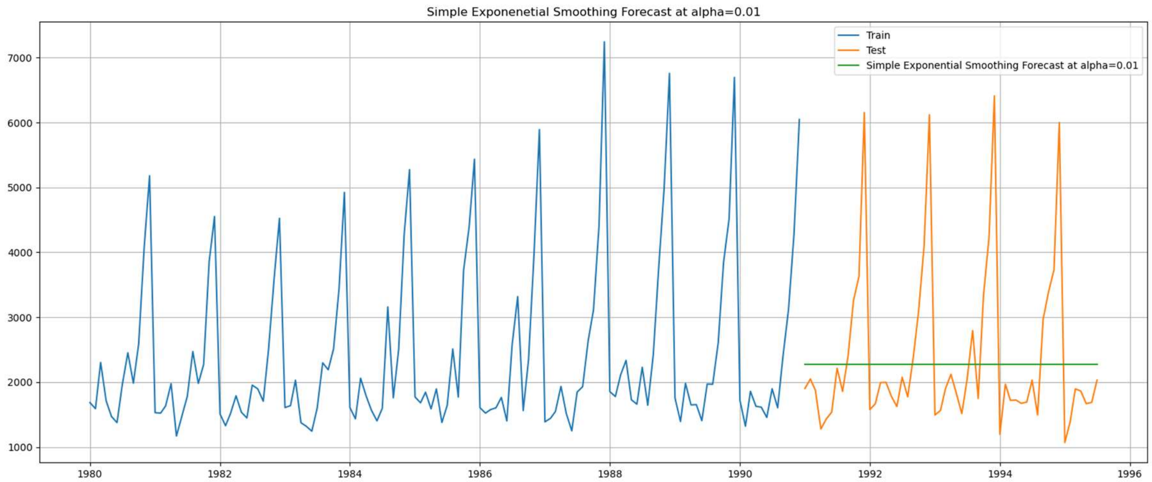

Fig.1.16. Simple Exponential smoothing forecast of test data- Sparkling wine sales optimized for lowest RMSE

Observations:

- RMSE is the lowest for alpha=0.01

4.2.2. Holt Double Exponential Smoothing

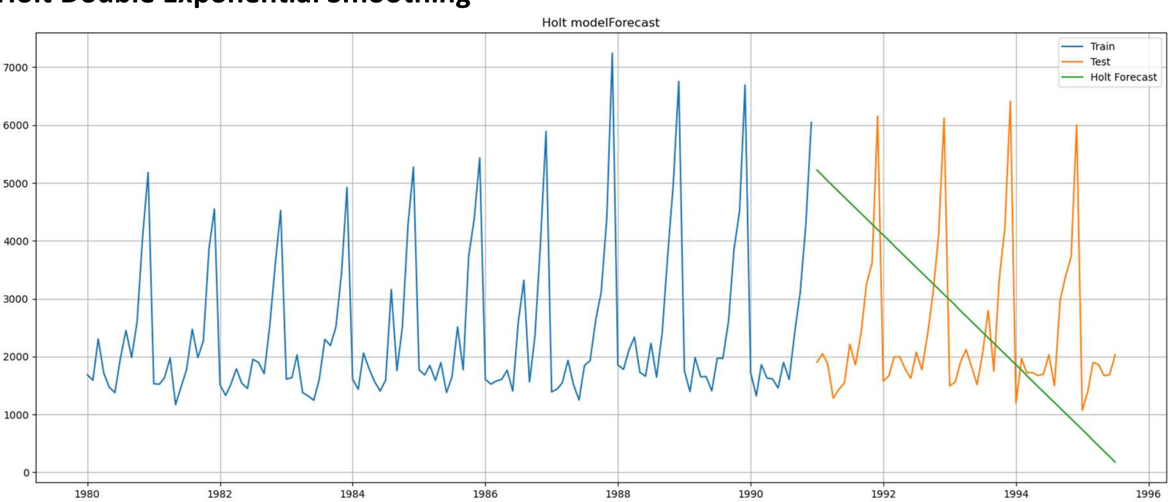

Fig.1.17**.** Holt forecast of test data- Sparkling wine sales

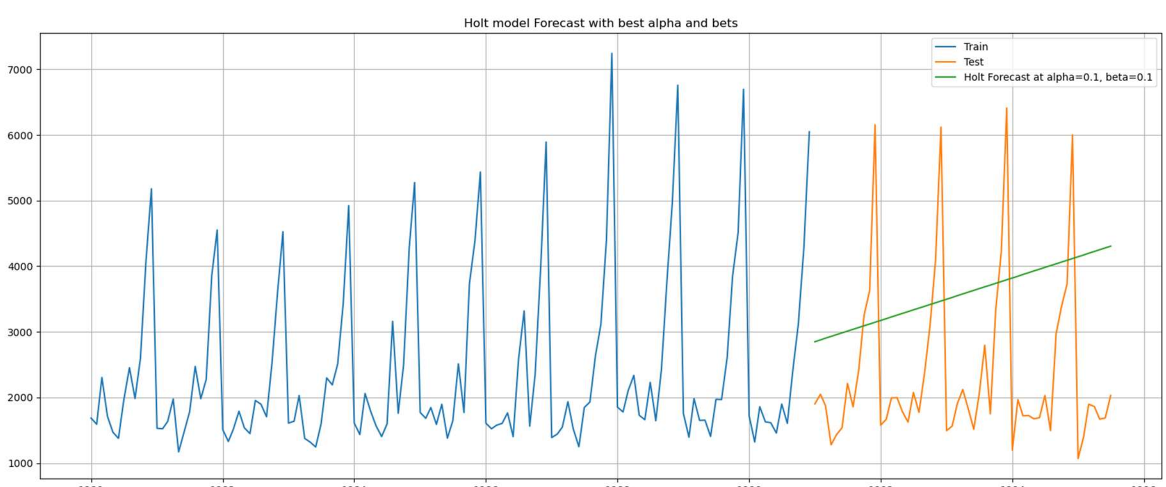

Fig.1.18. Holt forecast of test data- Sparkling wine sales - optimized for lowest RMSE

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

4.2.3. Holt-Winters Triple Exponential Smoothing

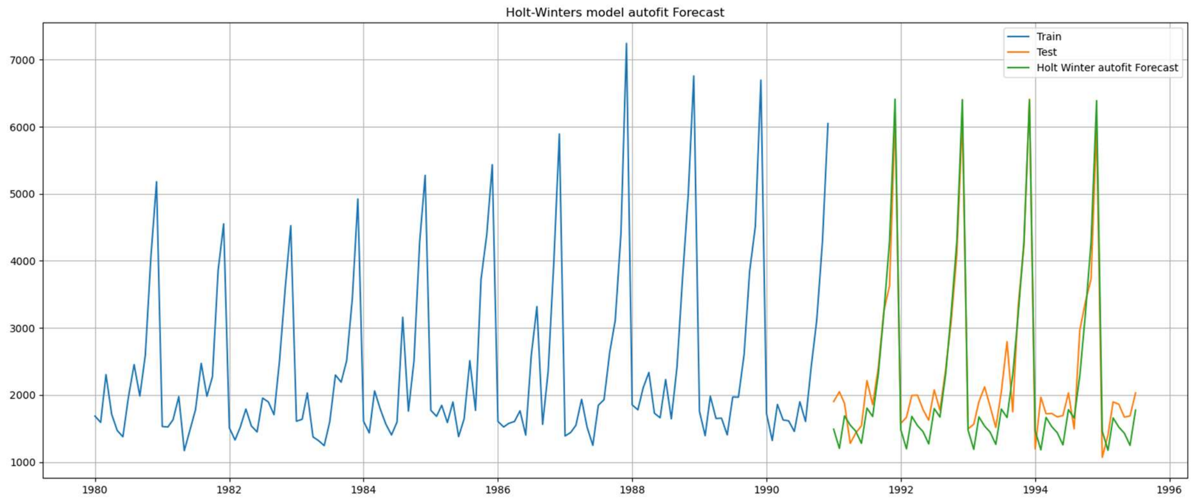

Fig.1.19**.** Holt Winters forecast of test data- Sparkling wine sales

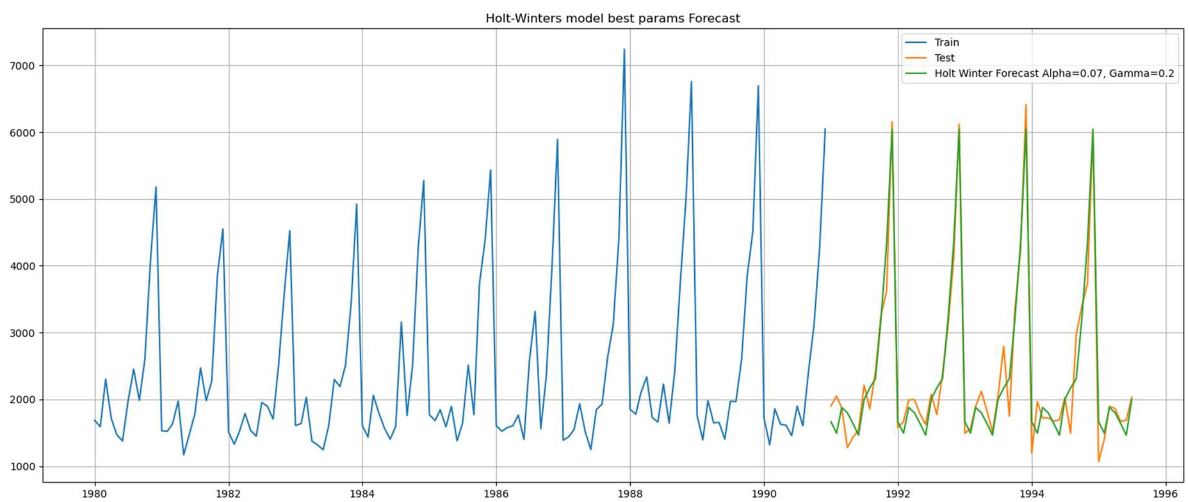

Fig.1.20. Holt Winters smoothing forecast of test data- Sparkling wine sales - optimized for lowest RMSE

Observations:
- RMSE is the lowest for alpha=0.07, and gamma=0.2, irrespective of beta value
- We noticed that the given series has no trend, and hence the results obtained are consistent with the assumptions

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**

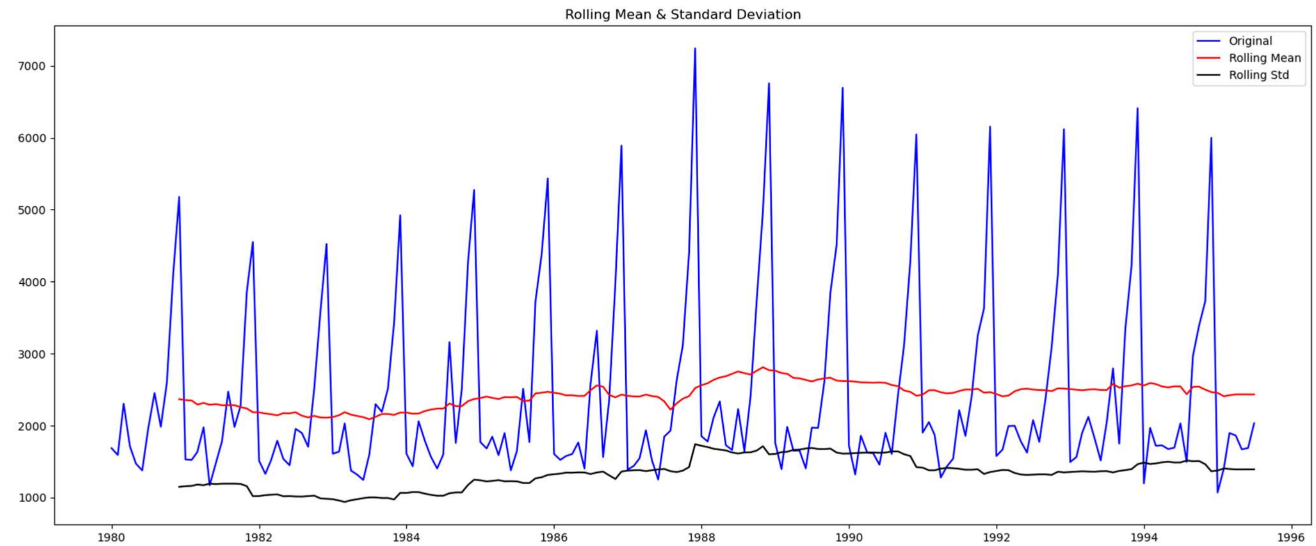

Fig.1.21. Sparkling dataset -Stationarity test rolling mean and Standard deviation plots

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

Fig.1.22. Dickey Fuller Test results- Sparkling Dataset

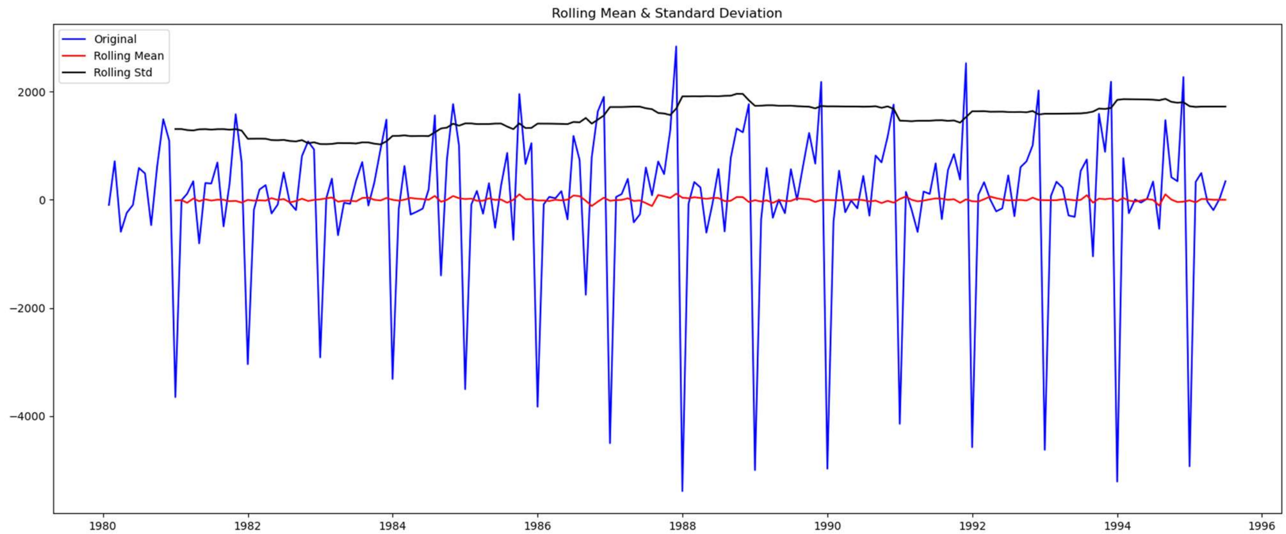

Fig.1.23. Differenced series -Stationarity test rolling mean and Standard deviation plots

```
Results of Dickey-Fuller Test:
Test Statistic                 -45.050301
p-value                          0.000000
#Lags Used                      10.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
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 0.6
- 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:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Fri, 10 Nov 2023   AIC                           2213.509
Time:                        22:09:17   BIC                           2227.885
Sample:                    01-01-1980   HQIC                          2219.351
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.781      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.741      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.218      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                14.46
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================
```

Fig.1.25. ARIMA results Summary- Sparkling Dataset

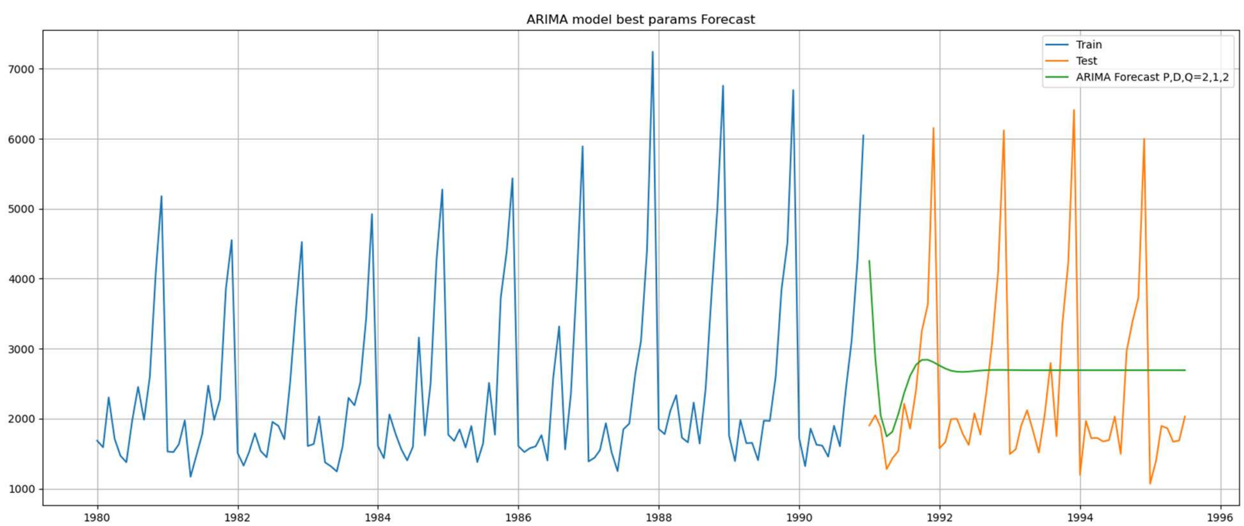

Fig.1.26. ARIMA model forecast on test data- Sparkling Dataset

Observations:

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

6.2. SARIMA model

```
                                    SARIMAX Results
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(0, 1, 2, 12)   Log Likelihood                -685.174
Date:                            Sat, 11 Nov 2023   AIC                           1382.348
Time:                                    07:29:54   BIC                           1397.479
Sample:                                01-01-1980   HQIC                          1388.455
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5507      0.287     -1.922      0.055      -1.112       0.011
ma.L1         -0.1612      0.235     -0.687      0.492      -0.621       0.299
ma.L2         -0.7218      0.175     -4.132      0.000      -1.064      -0.379
ma.S.L12      -0.4062      0.092     -4.401      0.000      -0.587      -0.225
ma.S.L24      -0.0274      0.138     -0.198      0.843      -0.298       0.243
sigma2      1.705e+05   2.45e+04      6.956      0.000    1.22e+05    2.19e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                13.48
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               0.89   Skew:                             0.60
Prob(H) (two-sided):                  0.75   Kurtosis:                         4.44
===================================================================================
```

Fig.1.27. SARIMA results Summary- Sparkling Dataset

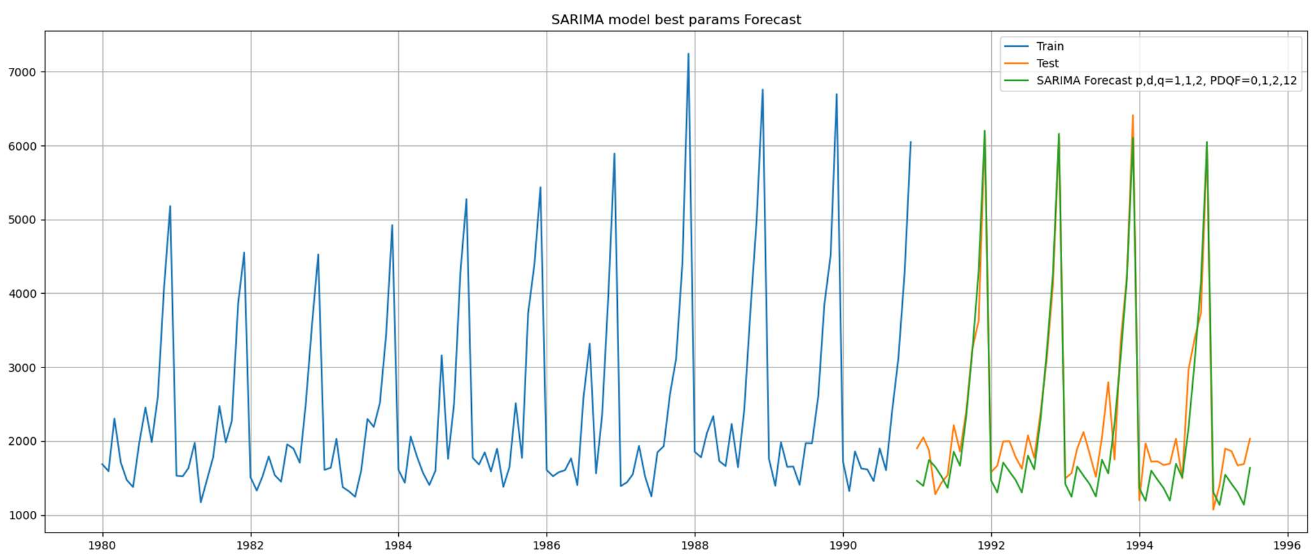

Fig.1.28. SARIMA model forecast on test data- Sparkling Dataset

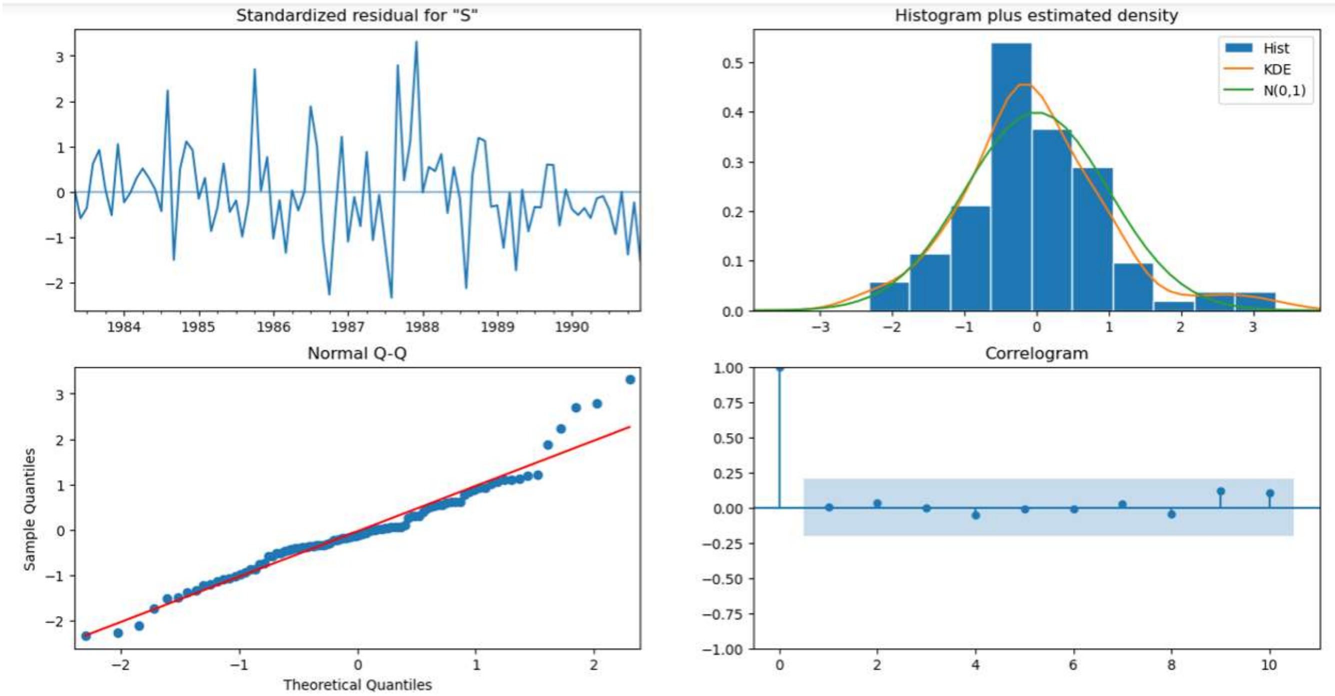

Fig.1.29. Diagnostics and Correlogram- SARIMA Model

Observations:

- Lowest AIC obtained for (p,d,q)x(P,D,Q,F)=(1,1,2)x(0,1,2,12)
- 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
HoltWintersalpha0.07gamma0.2	302.41
HoltWintersAutofit	378.95
SARIMA	382.58
MA_trail2	813.40
MA_trail3	1028.61
MA_trail12	1267.93
SimpleAvgForecast	1275.08
SimpExpSmoothingAlpha0.01	1281.03
MA_trail6	1283.93
ARIMA	1299.98
SimpleExpSmoothing	1304.93
LinearRegression	1389.14
HoltBestAlphaBeta	1778.56
HoltAutofit	2007.24
NaiveForecast	3864.28

Fig.1.30. Sparkling Dataset model Results- Test RMSE

Observations:

- Lowest RMSE is obtained for Holt winters model having alpha=0.07 and gamma=0.2
- And SARIMA model for the above said parameters

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.

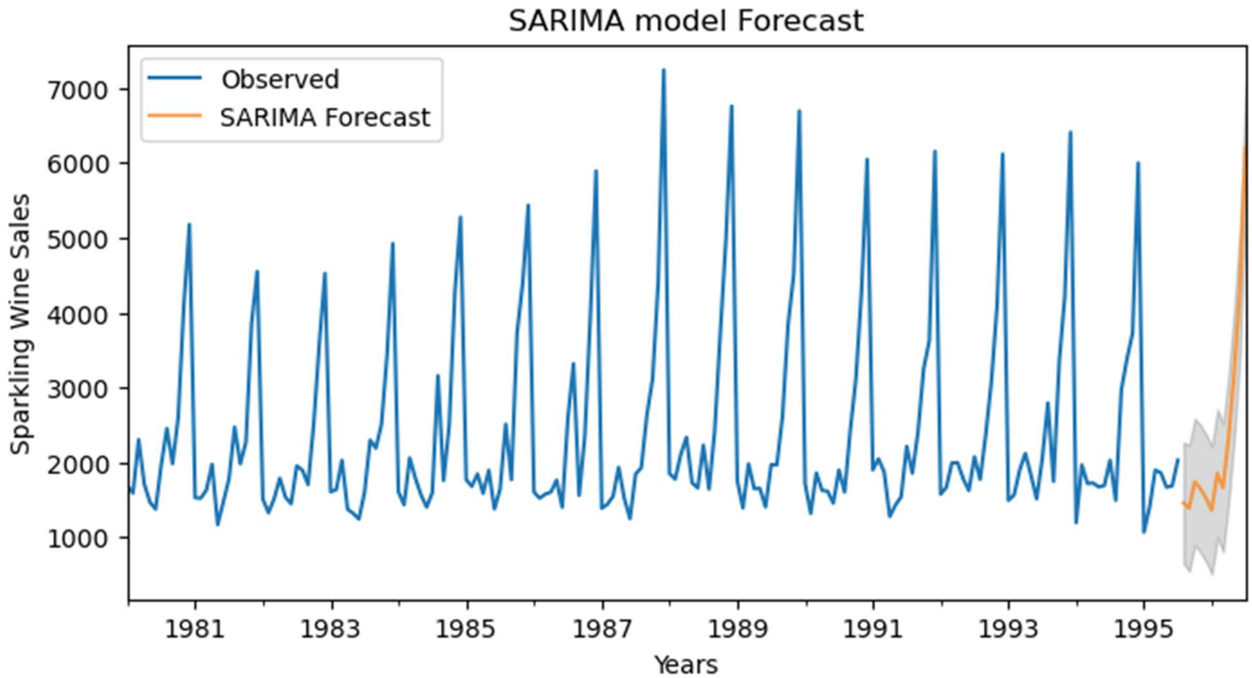

Fig.1.31. SARIMA Model forecast for next 12 months- Sparkling Dataset

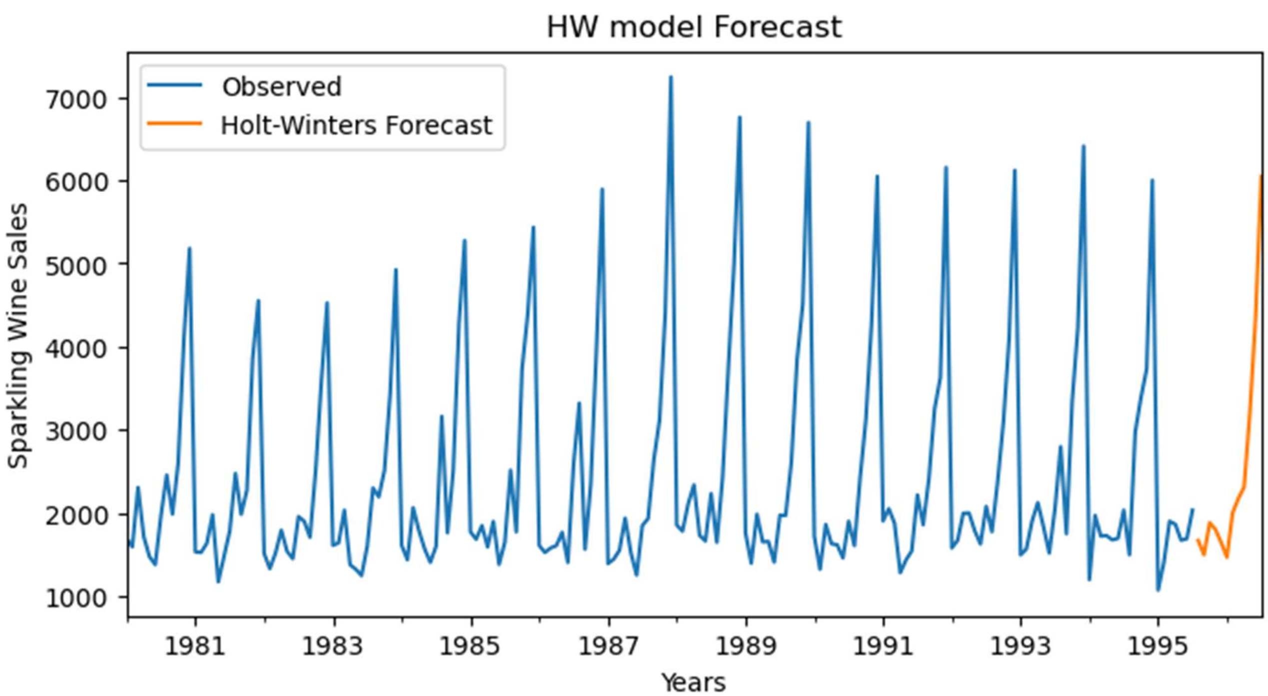

Fig.1.32. Holt Winters Model forecast for next 12 months- Sparkling Dataset

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

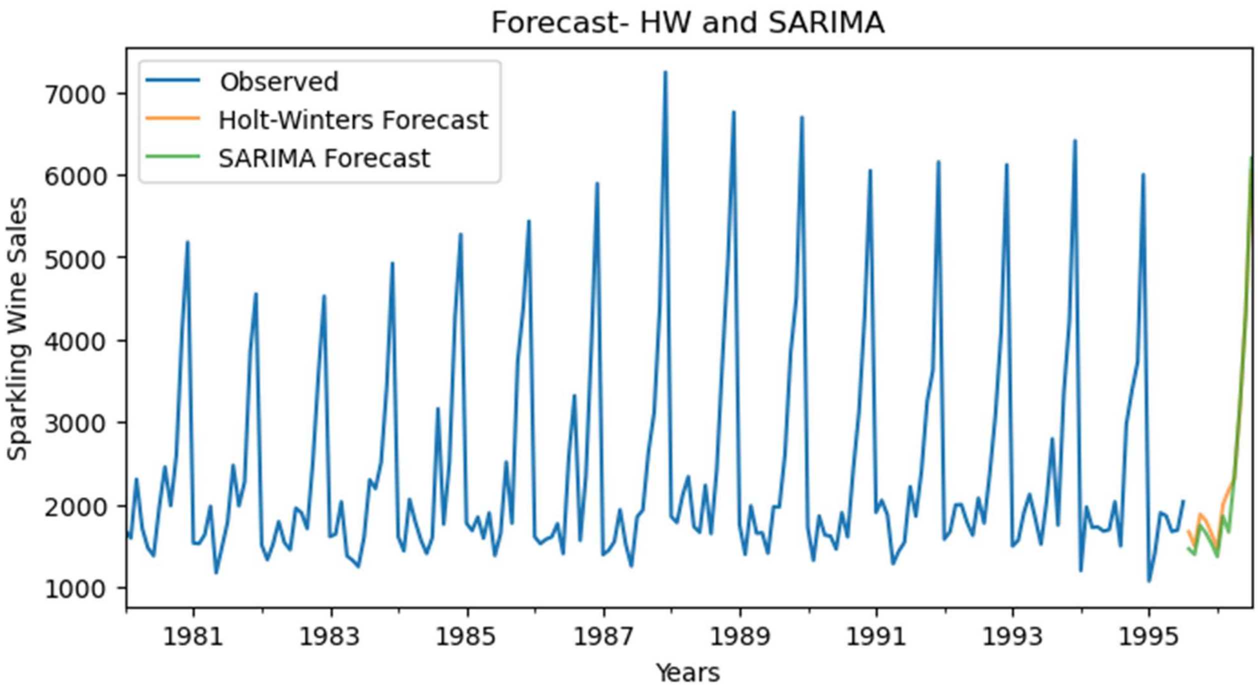

Fig.1.33. Forecast for 12 months- SARIMA and HW

Observations:

- The wine sales peaks during the months of November and december, probably due to the holiday season.
- The sales data does not exhibit any trend
- The forecast replicates the existing seasonality

Insights:

- The seasonality component of sales can be capitalized, and can try to push sales in the peak months
- The trend component needs improvement. The company can adopt different marketing strategies by customer segmentation in order to increase the overall trend