CHETAN
DUDHANE

28-MAR-2021

Project on

TIME SERIES

PGP DSBA - JULY B GROUP 2
chetan.dudhane@gmail.com

INTRODUCTION

This report consists of Time Series analysis and forecasting of 2 datasets -

- DATASET 1 - Sales data of Rose Wine
- DATASET 2 - Sales data of Sparkling Wine

Please find the Jupyter Code Notebook here. Analysis code is in Python.

PROBLEM - Wine Sales

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.

Datasets used - Sales of Rose Wine and Sparkling Wine

SYNOPSIS

1. Total No. Of Rose Data Entries = 187

   Total No. Of Sparkling Data Entries = 187

2. No. Of Missing Values in Rose data = 2

   - We use INTERPOLATION to impute these missing values

   No. Of Missing Values in Sparkling data = 0

3. No. Of Duplicate entries in Rose data = 0

   No. Of Duplicate entries in Sparkling data = 0

4. Both datasets are split in Train : Test at year 1991 - Test data starts at 1991

5. Various forecasting models applied are -

   1. Linear Regression
   2. Naive Bayes
   3. Simple Average
   4. 2-pt Moving Average

5. 4-pt Moving Average
6. 6-pt Moving Average
7. 9-pt Moving Average
8. Single Exponential Smoothing
9. Double Exponential Smoothing (Holt's Model)
10. Triple Exponential Smoothing (Holt-Winter Model)
11. ARIMA / SARIMA (Auto fitted)
12. ARIMA / SARIMA (Manually fitted)

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

- Both Datasets are read and stored as Pandas Data Frames for analysis
- Datasets are read as Time Series data using `parse_dates=True & index_col='YearMonth'`
- First 5 rows of both the data are given below -

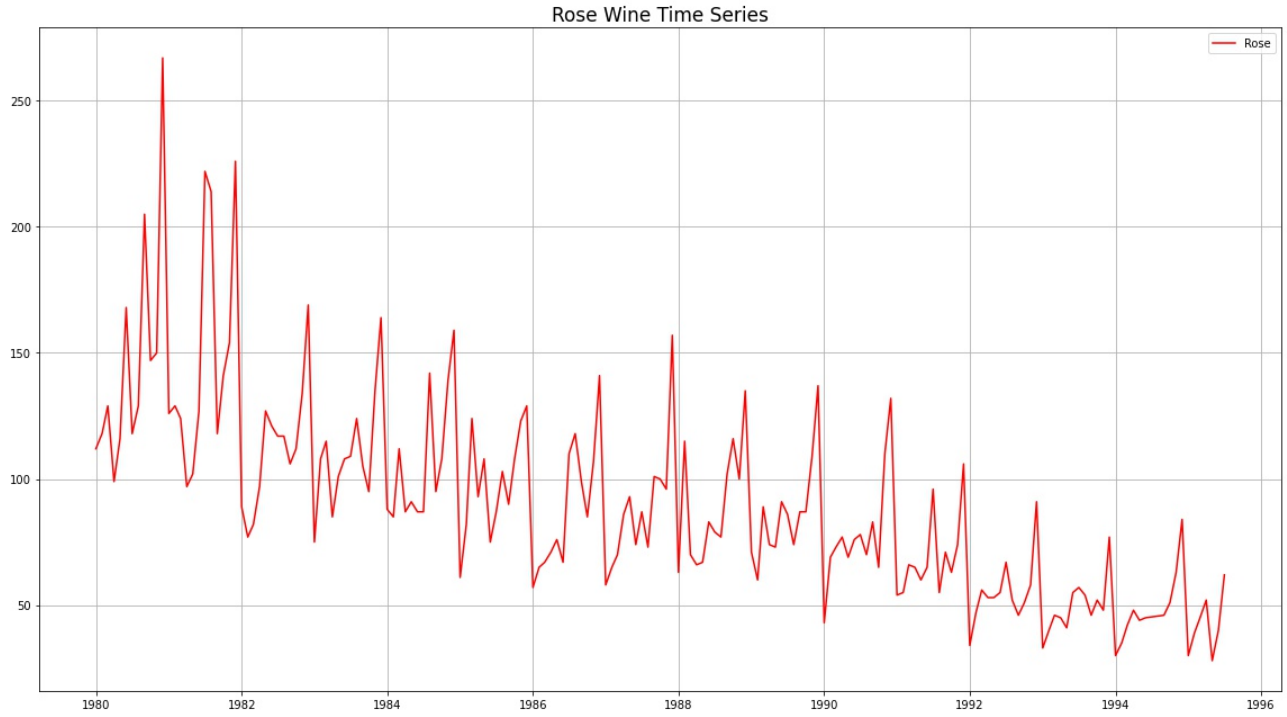
YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Rose Data Header

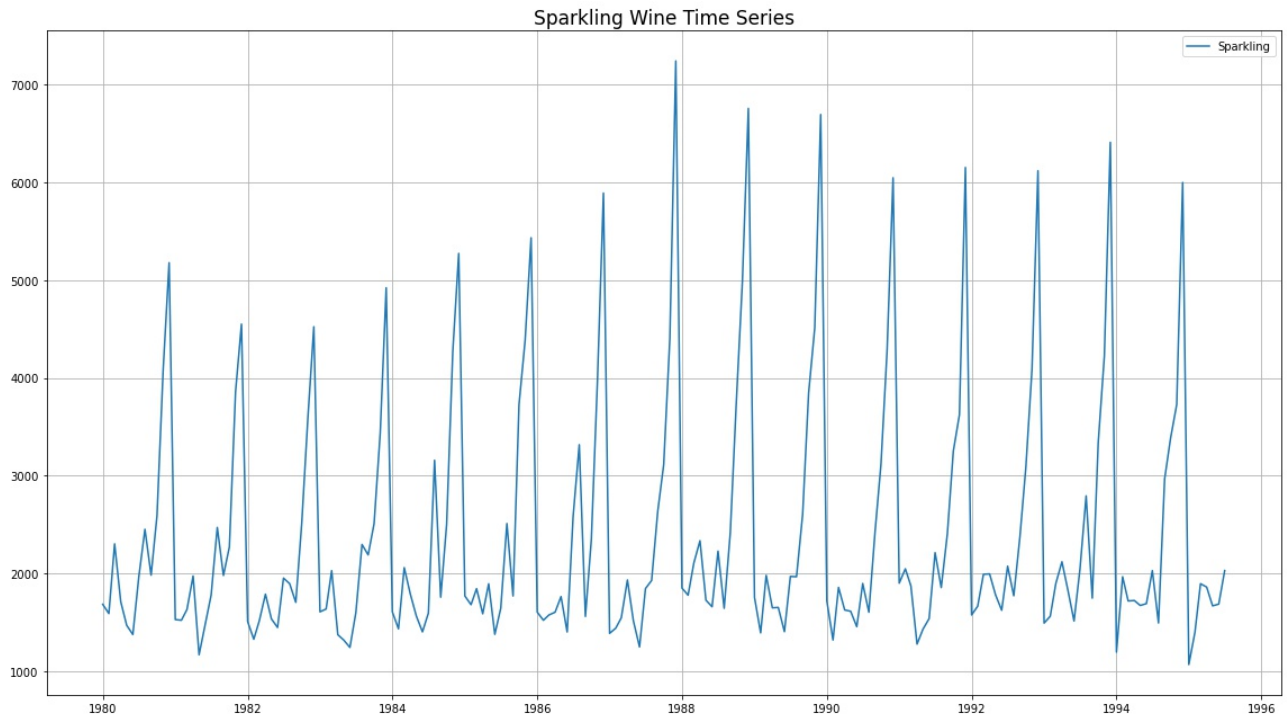
YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Sparkling Data Header

- Rose Data plot -



- Sparkling Data plot -



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

✦ Exploratory Data Analysis -

	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.914	39.238	28.0	62.5	85.0	111.0	267.0
Sparkling	187.0	2402.417	1295.112	1070.0	1605.0	1874.0	2549.0	7242.0

Descriptive Stats of Rose and Sparkling datasets

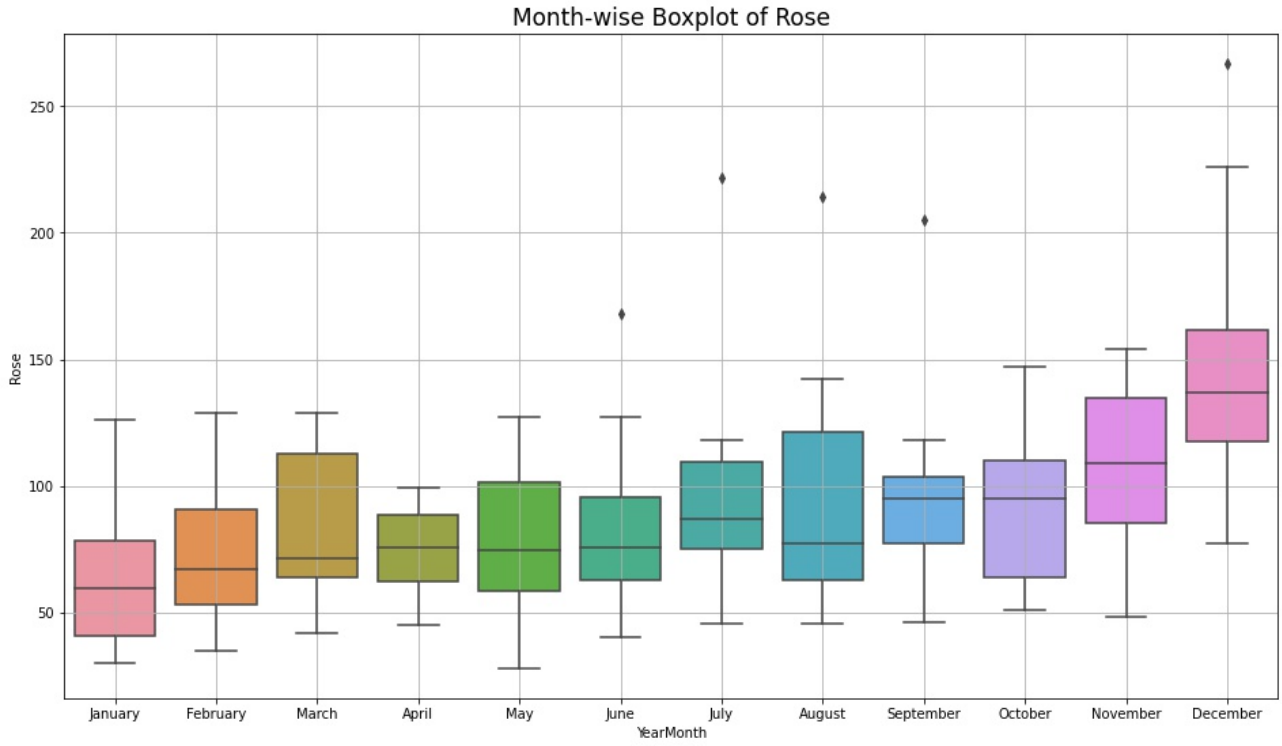
```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries,
1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    187 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Info - Rose data

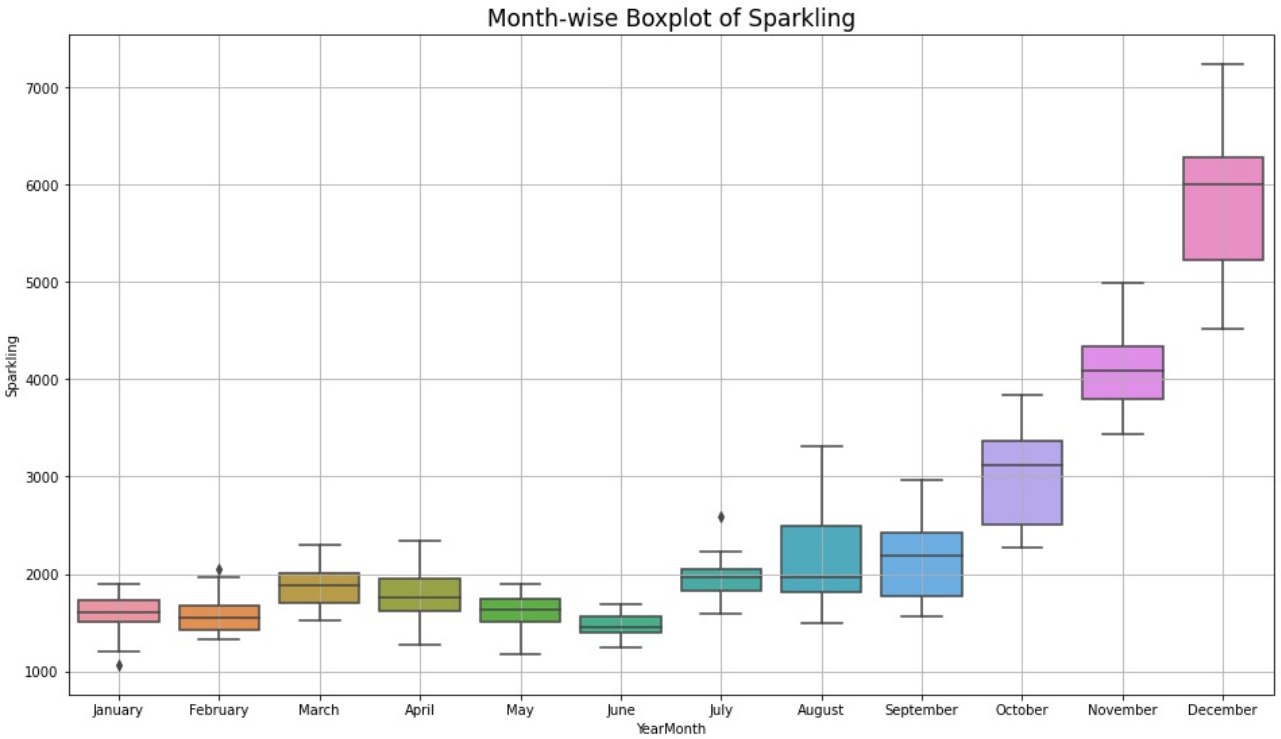
```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries,
1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Info - Sparkling data

- Month-wise Boxplot of Rose -

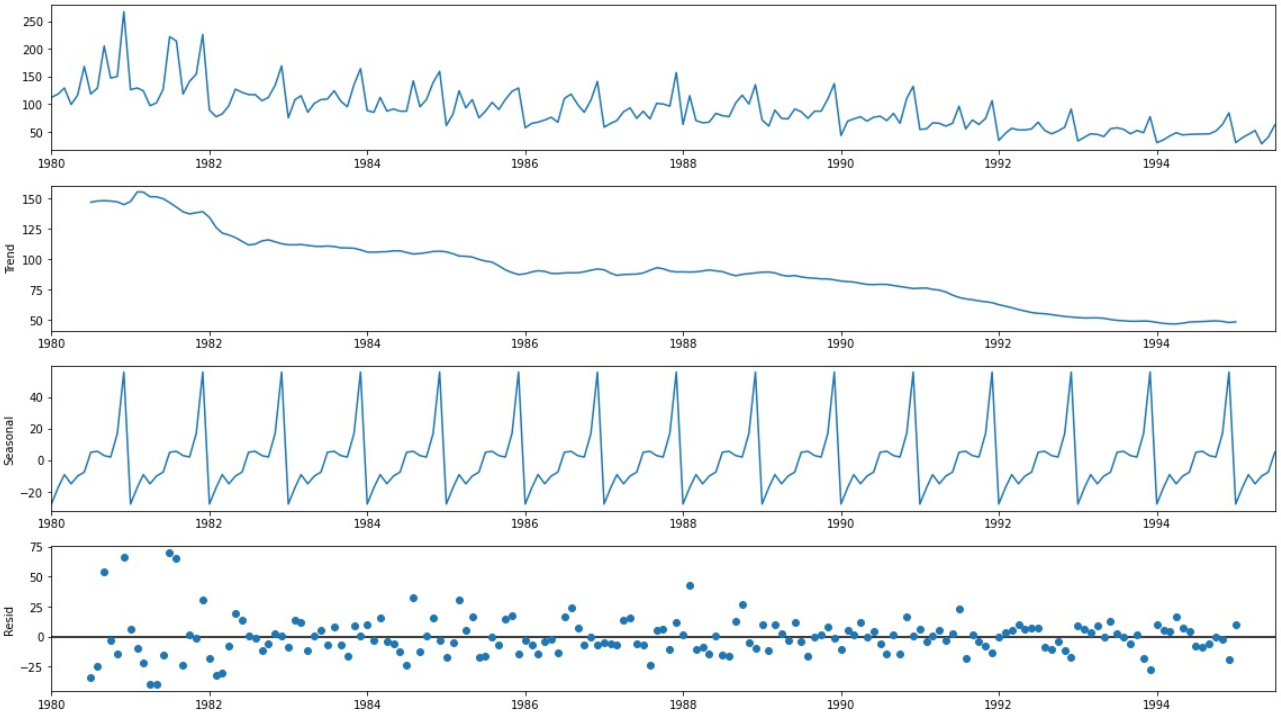


- Month-wise Boxplot of Sparkling -



  - Sales of both - Rose and Sparkling, show a spike in the last quarter of Oct to Dec
  - Spike is much more accentuated in Sparkling sales
  - This spike may be due to the Holiday season starting in Oct

✦ Additive Decomposition of Rose -



YearMonth	trend
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	147.08
1980-08-01	148.13
1980-09-01	148.38
1980-10-01	148.08
1980-11-01	147.42
1980-12-01	145.13

Rose Trend

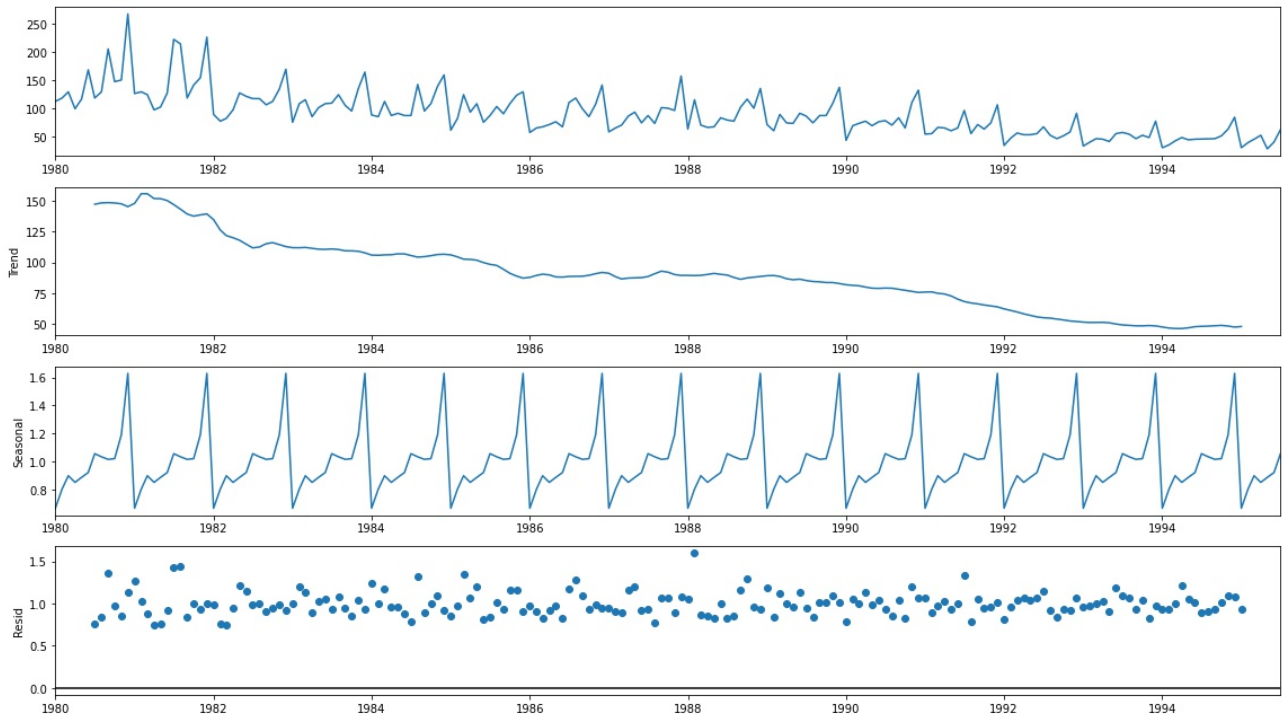
YearMonth	seasonal
1980-01-01	-27.91
1980-02-01	-17.44
1980-03-01	-9.29
1980-04-01	-15.10
1980-05-01	-10.20
1980-06-01	-7.68
1980-07-01	4.90
1980-08-01	5.50
1980-09-01	2.77
1980-10-01	1.87
1980-11-01	16.85
1980-12-01	55.71

Rose Seasonality

YearMonth	resid
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	-33.98
1980-08-01	-24.62
1980-09-01	53.85
1980-10-01	-2.96
1980-11-01	-14.26
1980-12-01	66.16

Rose Residual

✦ Multiplicative Decomposition of Rose -



YearMonth	trend
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	147.08
1980-08-01	148.13
1980-09-01	148.38
1980-10-01	148.08
1980-11-01	147.42
1980-12-01	145.13

Rose Trend

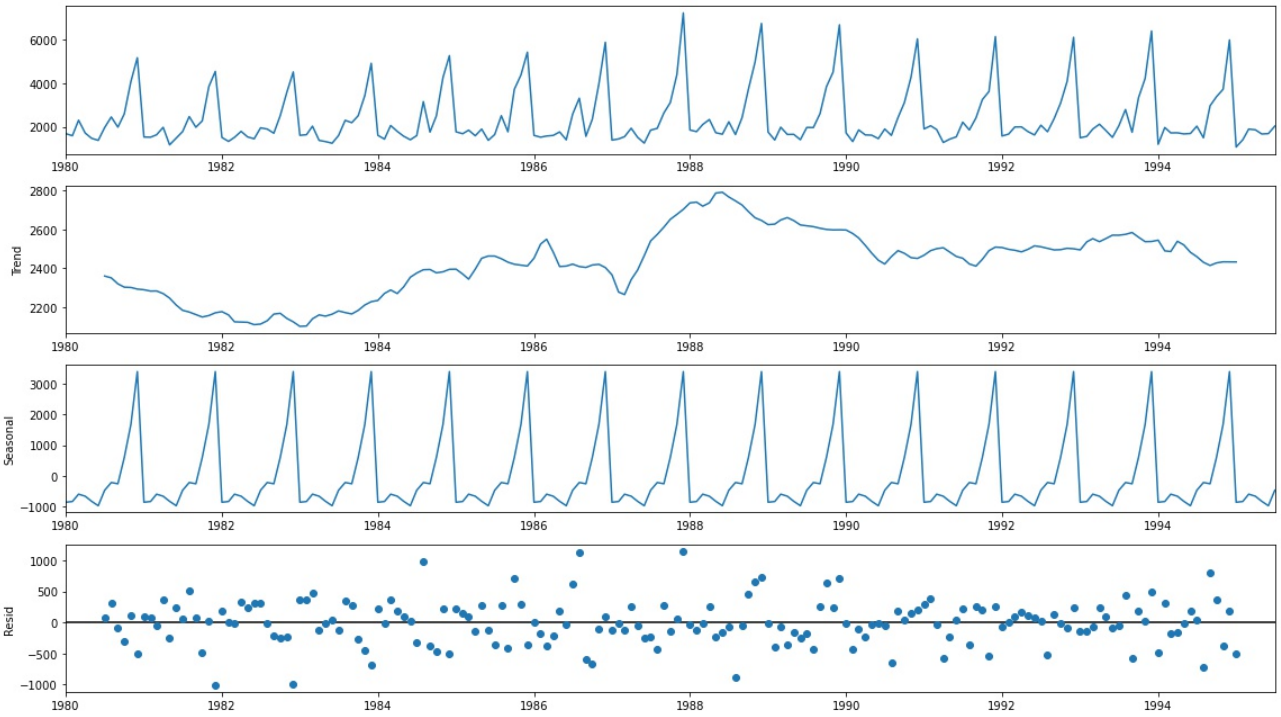
YearMonth	seasonal
1980-01-01	0.67
1980-02-01	0.81
1980-03-01	0.90
1980-04-01	0.85
1980-05-01	0.89
1980-06-01	0.92
1980-07-01	1.06
1980-08-01	1.04
1980-09-01	1.02
1980-10-01	1.02
1980-11-01	1.19
1980-12-01	1.63

Rose Seasonality

YearMonth	resid
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	0.76
1980-08-01	0.84
1980-09-01	1.36
1980-10-01	0.97
1980-11-01	0.85
1980-12-01	1.13

Rose Residual

✦ Additive Decomposition of Sparkling -



YearMonth	trend
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	2360.67
1980-08-01	2351.33
1980-09-01	2320.54
1980-10-01	2303.58
1980-11-01	2302.04
1980-12-01	2293.79

Sparkling Trend

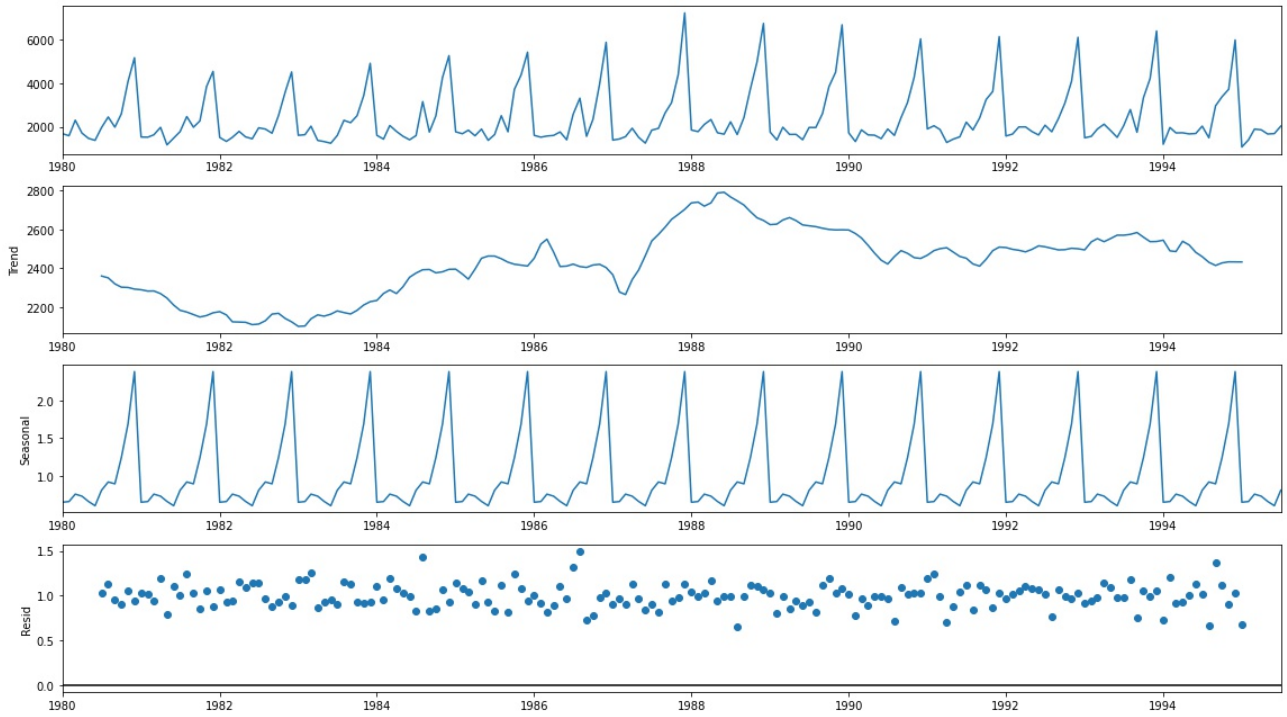
YearMonth	seasonal
1980-01-01	-854.26
1980-02-01	-830.35
1980-03-01	-592.36
1980-04-01	-658.49
1980-05-01	-824.42
1980-06-01	-967.43
1980-07-01	-465.50
1980-08-01	-214.33
1980-09-01	-254.68
1980-10-01	599.77
1980-11-01	1675.07
1980-12-01	3386.98

Sparkling Seasonality

YearMonth	resid
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	70.84
1980-08-01	316.00
1980-09-01	-81.86
1980-10-01	-307.35
1980-11-01	109.89
1980-12-01	-501.78

Sparkling Residual

✦ Multiplicative Decomposition of Sparkling -



YearMonth	trend
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	2360.67
1980-08-01	2351.33
1980-09-01	2320.54
1980-10-01	2303.58
1980-11-01	2302.04
1980-12-01	2293.79

Sparkling Trend

YearMonth	seasonal
1980-01-01	0.65
1980-02-01	0.66
1980-03-01	0.76
1980-04-01	0.73
1980-05-01	0.66
1980-06-01	0.60
1980-07-01	0.81
1980-08-01	0.92
1980-09-01	0.89
1980-10-01	1.24
1980-11-01	1.69
1980-12-01	2.38

Sparkling Seasonality

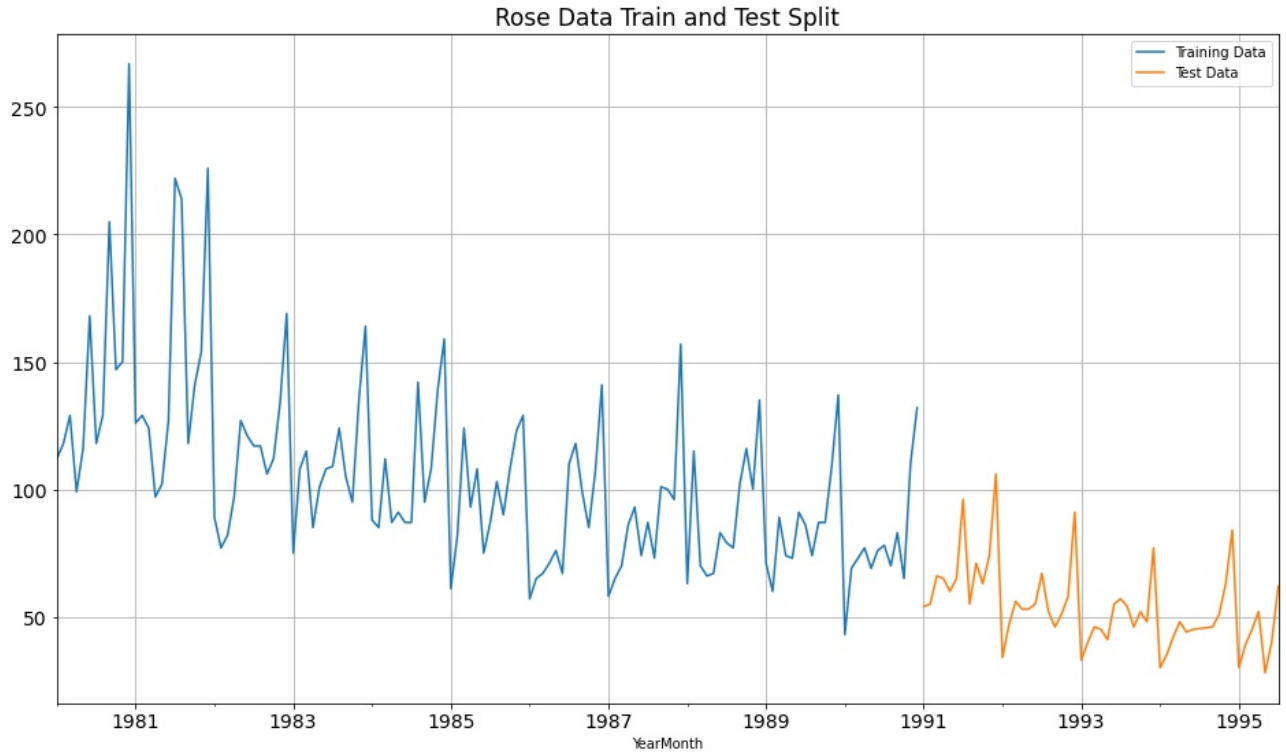
YearMonth	resid
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	1.03
1980-08-01	1.14
1980-09-01	0.96
1980-10-01	0.91
1980-11-01	1.05
1980-12-01	0.95

Sparkling Residual

- Additive Models -
    - The seasonality is relatively constant over time
    - $y_t = Trend + Seasonality + Residual$
- Multiplicative Models -
    - The seasonality increases or decreases over time. It is proportionate to the trend
    - $y_t = Trend * Seasonality * Residual$
- Here by just observing the Residual patterns of Additive and Multiplicative models of Rose and Sparkling datasets. It seems that -
    - Rose is Multiplicative
    - Sparkling is Additive

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

- Both datasets of Rose and Sparkling are split at the year 1991
- Test datasets start at 1991



- Rose dataset - TRAIN

YearMonth	Rose
1980-01-01	112.00
1980-02-01	118.00
1980-03-01	129.00
1980-04-01	99.00
1980-05-01	116.00

Rose Train - First 5 rows

YearMonth	Rose
1990-08-01	70.00
1990-09-01	83.00
1990-10-01	65.00
1990-11-01	110.00
1990-12-01	132.00

Rose Train - Last 5 rows

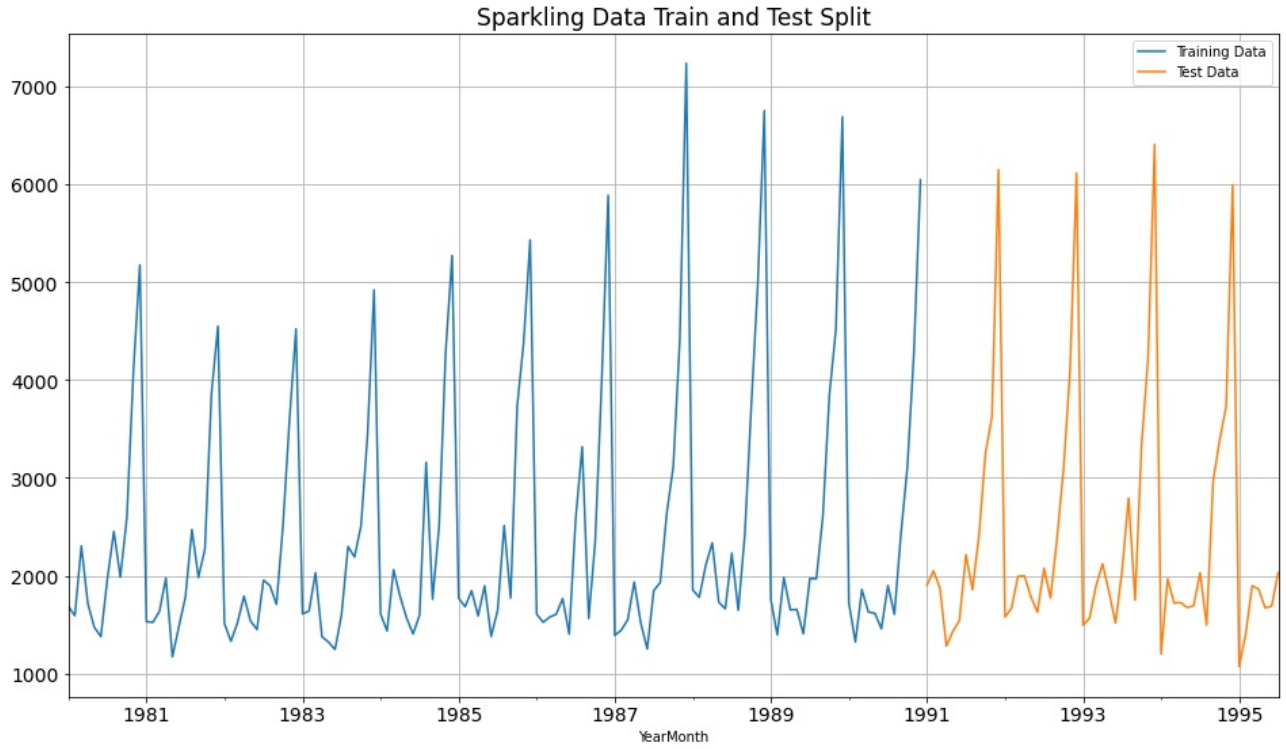
- Rose dataset - TEST

YearMonth	Rose
1991-01-01	54.00
1991-02-01	55.00
1991-03-01	66.00
1991-04-01	65.00
1991-05-01	60.00

Rose Test - First 5 rows

YearMonth	Rose
1995-03-01	45.00
1995-04-01	52.00
1995-05-01	28.00
1995-06-01	40.00
1995-07-01	62.00

Rose Test - Last 5 rows



- Sparkling dataset - TRAIN

YearMonth	Sparkling
1980-01-01	1686.00
1980-02-01	1591.00
1980-03-01	2304.00
1980-04-01	1712.00
1980-05-01	1471.00

Sparkling Train - First 5

YearMonth	Sparkling
1990-08-01	1605.00
1990-09-01	2424.00
1990-10-01	3116.00
1990-11-01	4286.00
1990-12-01	6047.00

Sparkling Train - Last 5

- Sparkling dataset - TEST

YearMonth	Sparkling
1991-01-01	1902.00
1991-02-01	2049.00
1991-03-01	1874.00
1991-04-01	1279.00
1991-05-01	1432.00

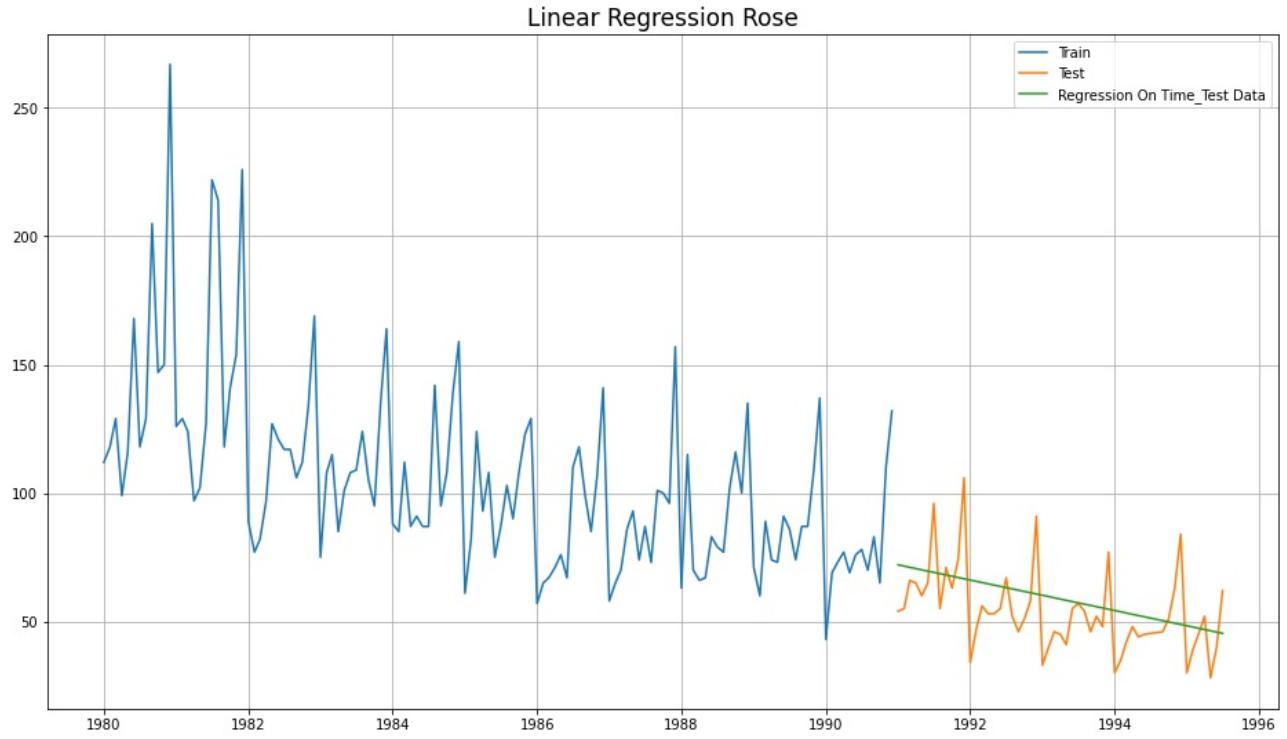
Sparkling Test - First 5

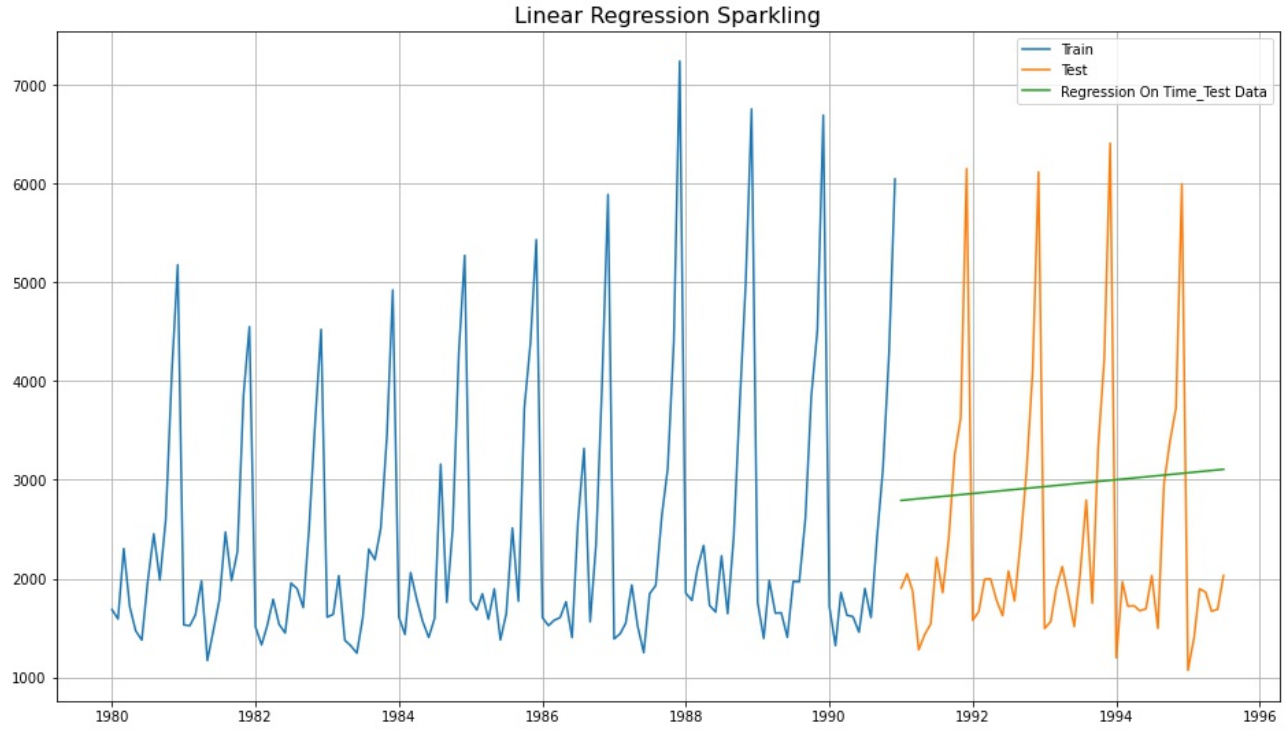
YearMonth	Sparkling
1995-03-01	1897.00
1995-04-01	1862.00
1995-05-01	1670.00
1995-06-01	1688.00
1995-07-01	2031.00

Sparkling Test - Last 5

[Q 4] Build various 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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

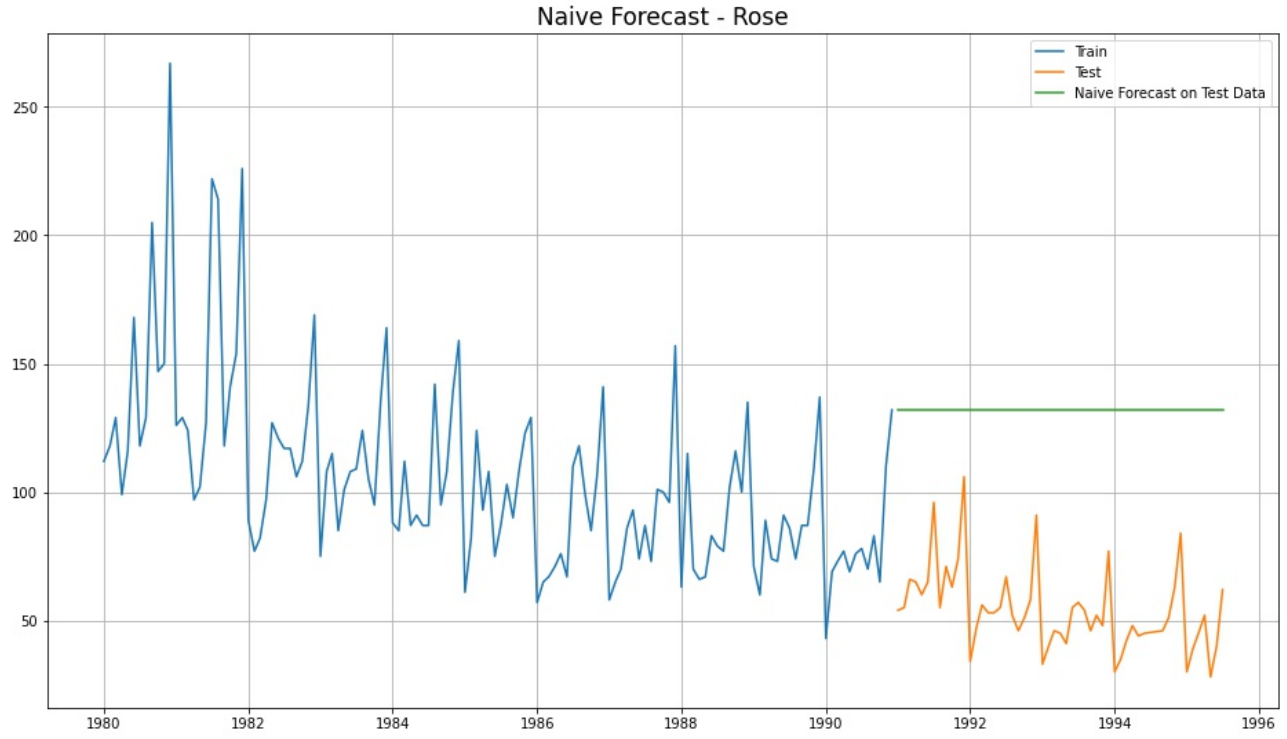
✦ Model 1 - Linear Regression

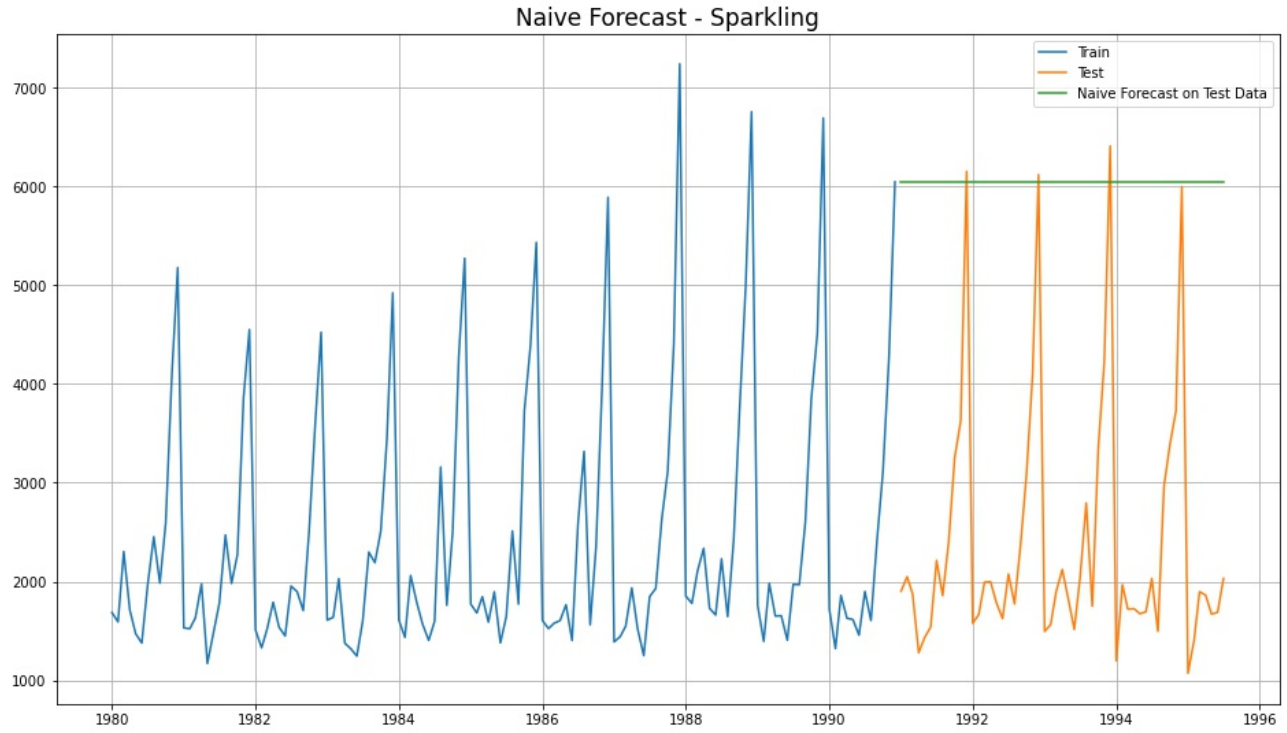




	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14

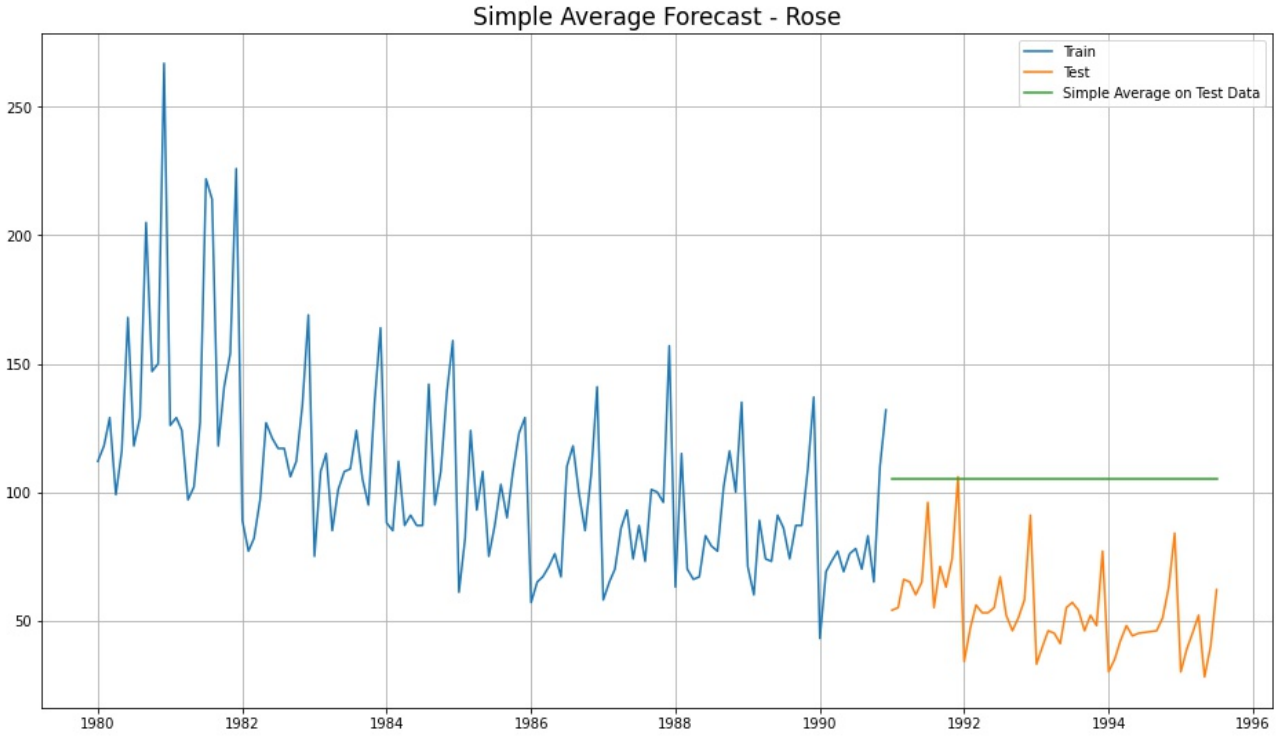
✦ Model 2 - Naive Bayes

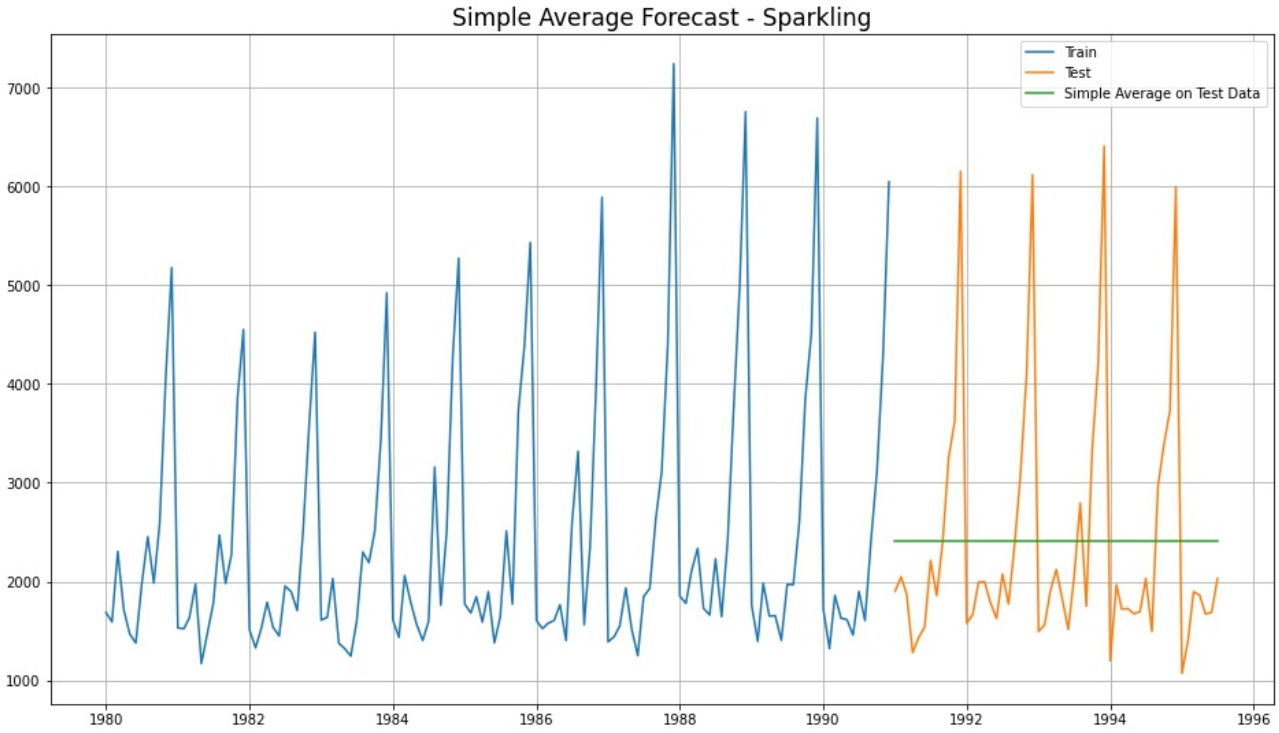




	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28

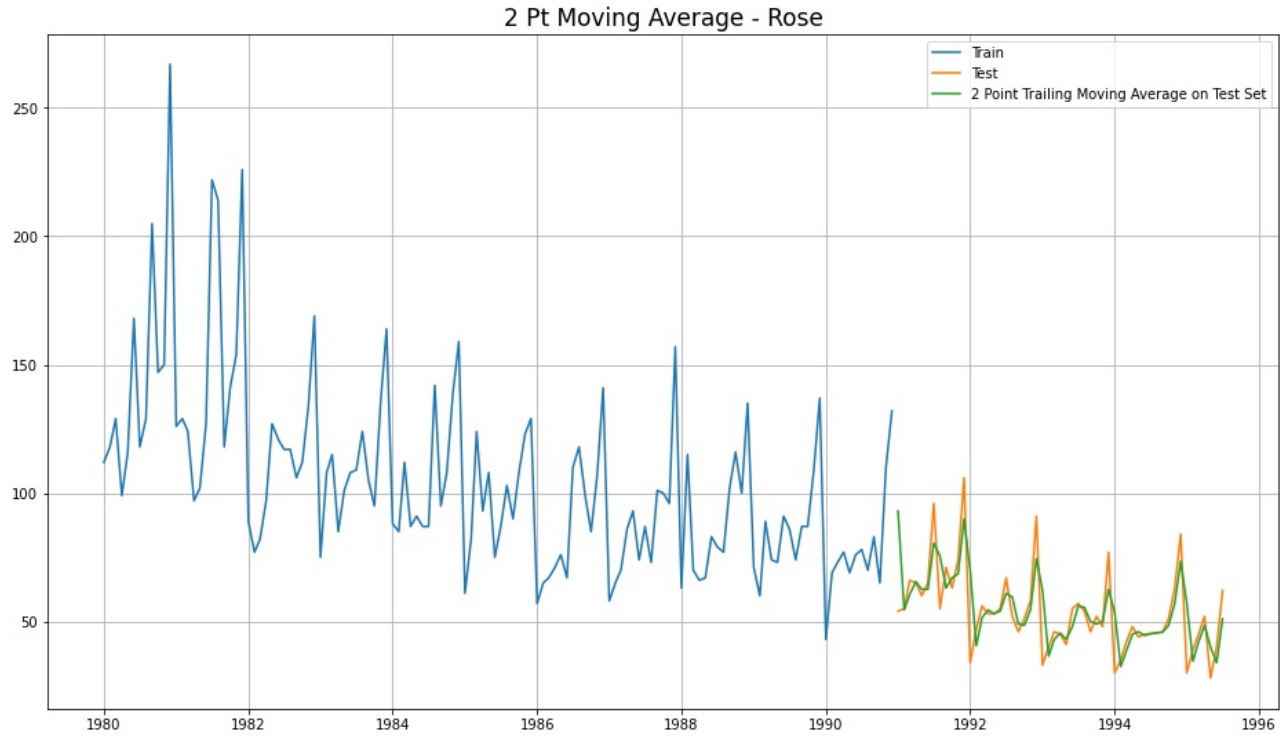
✦ Model 3 - Simple Average

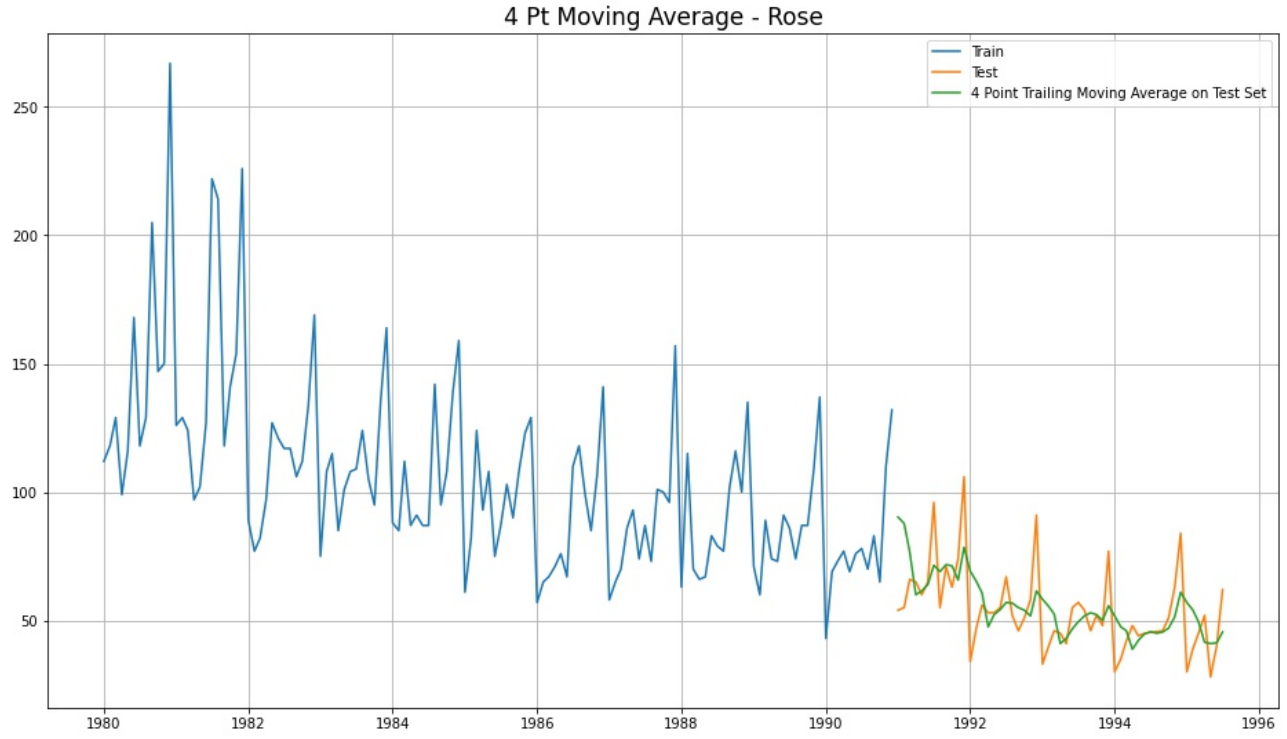


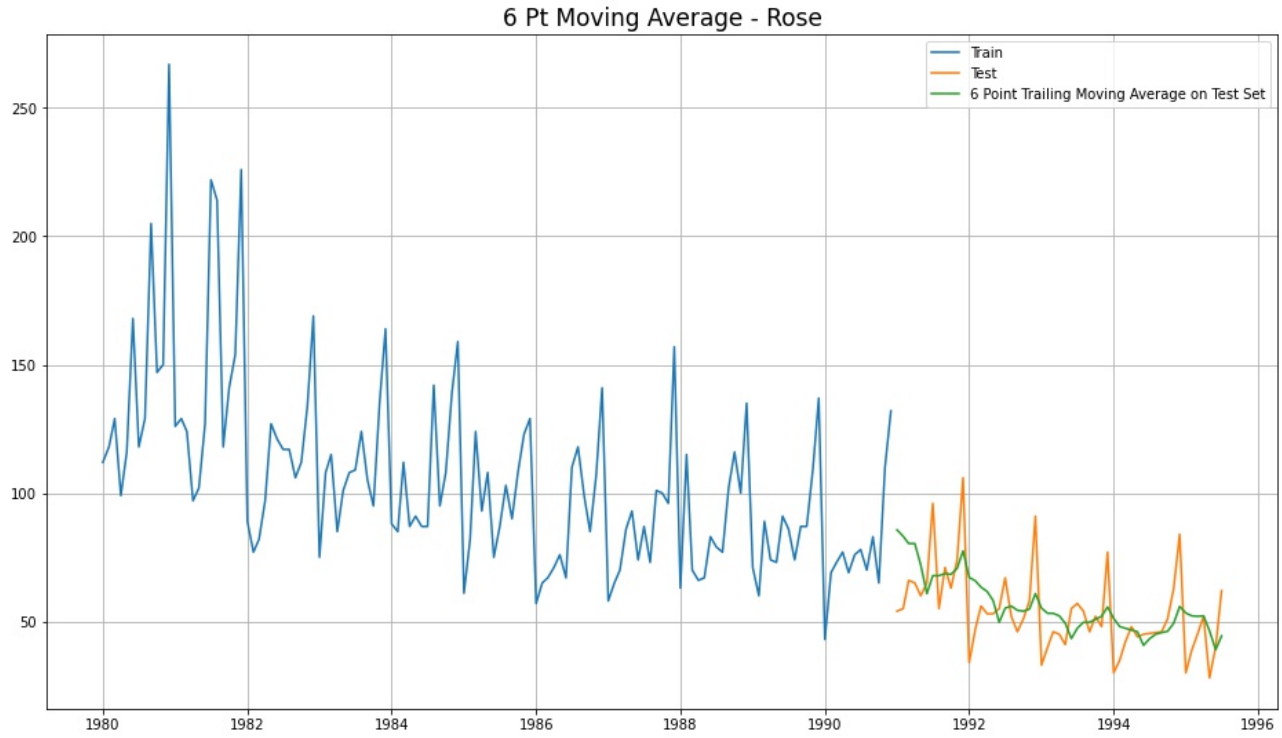


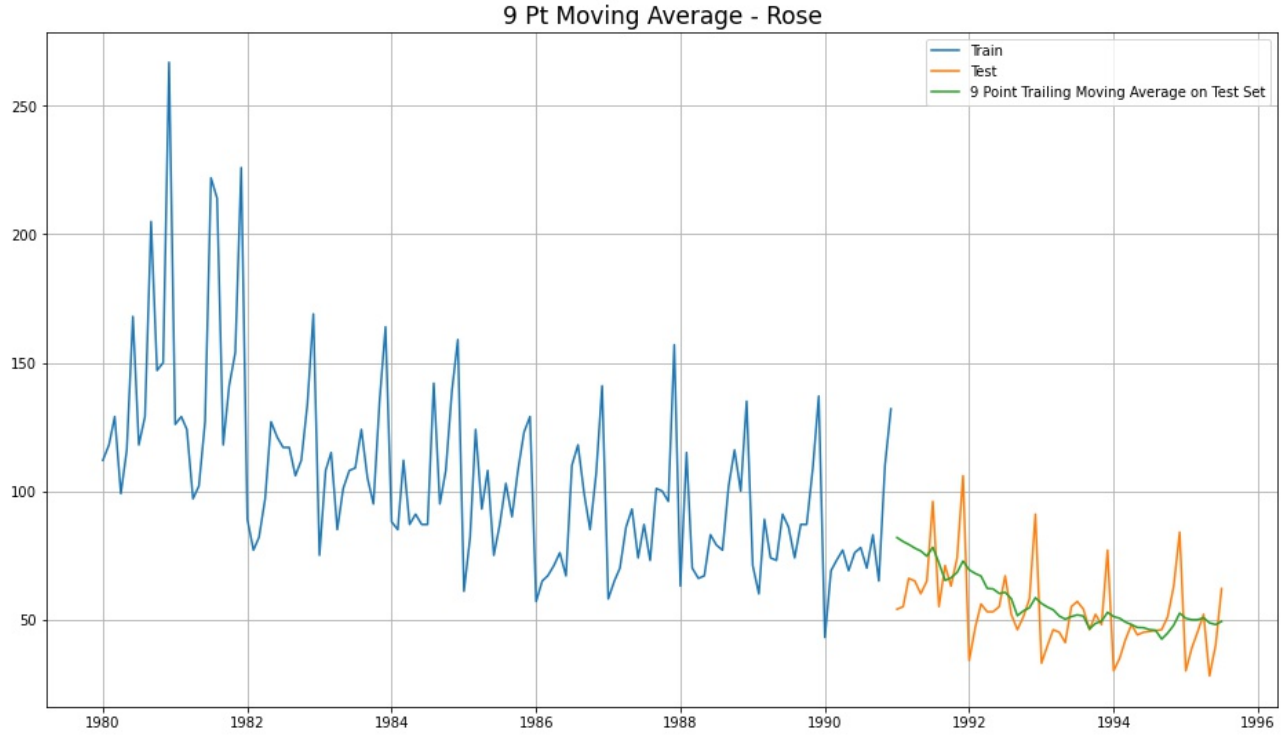
	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28
SimpleAverageModel	53.46	1275.08

✦ Model 4.A - Moving Average (Rose)



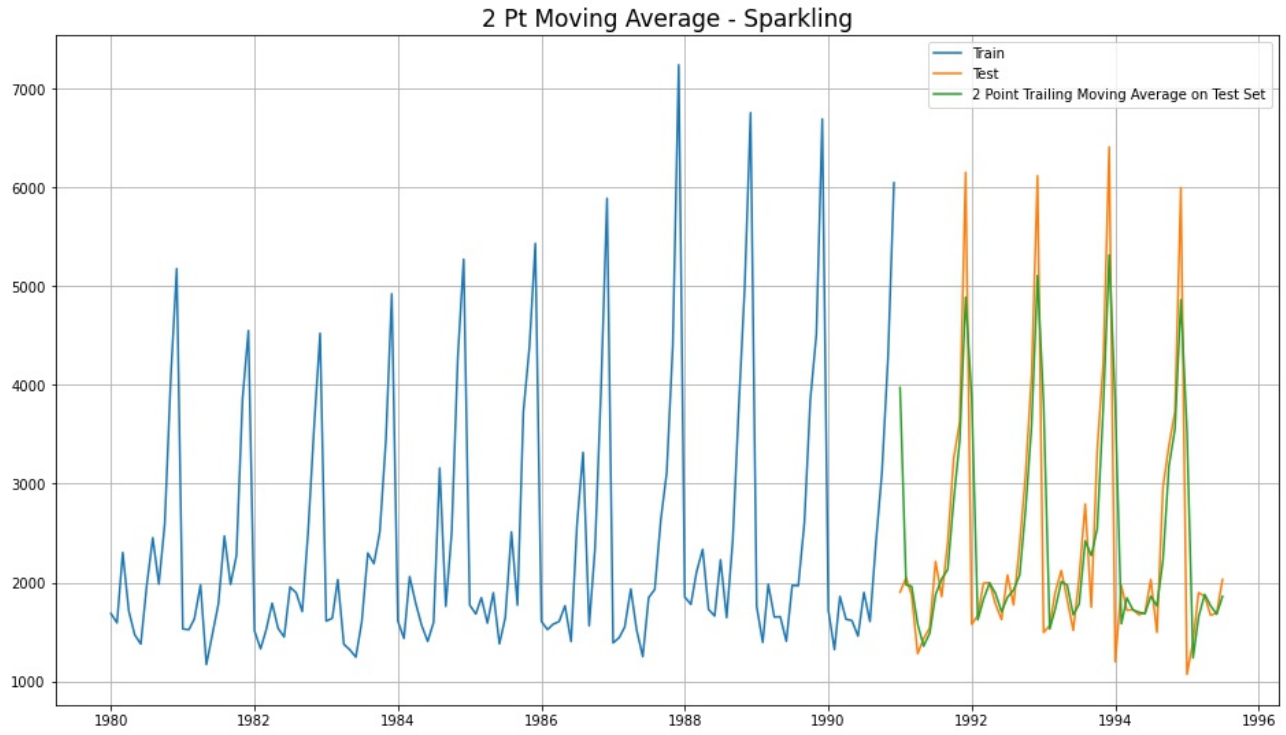


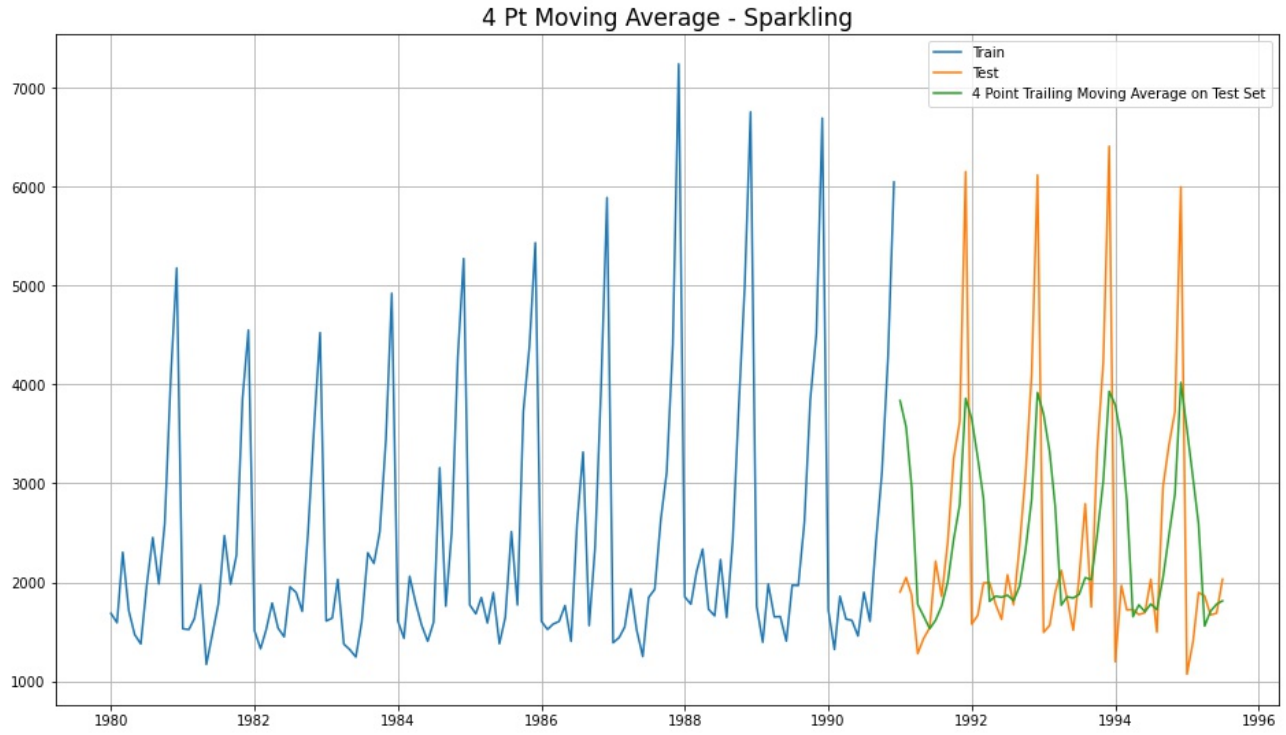


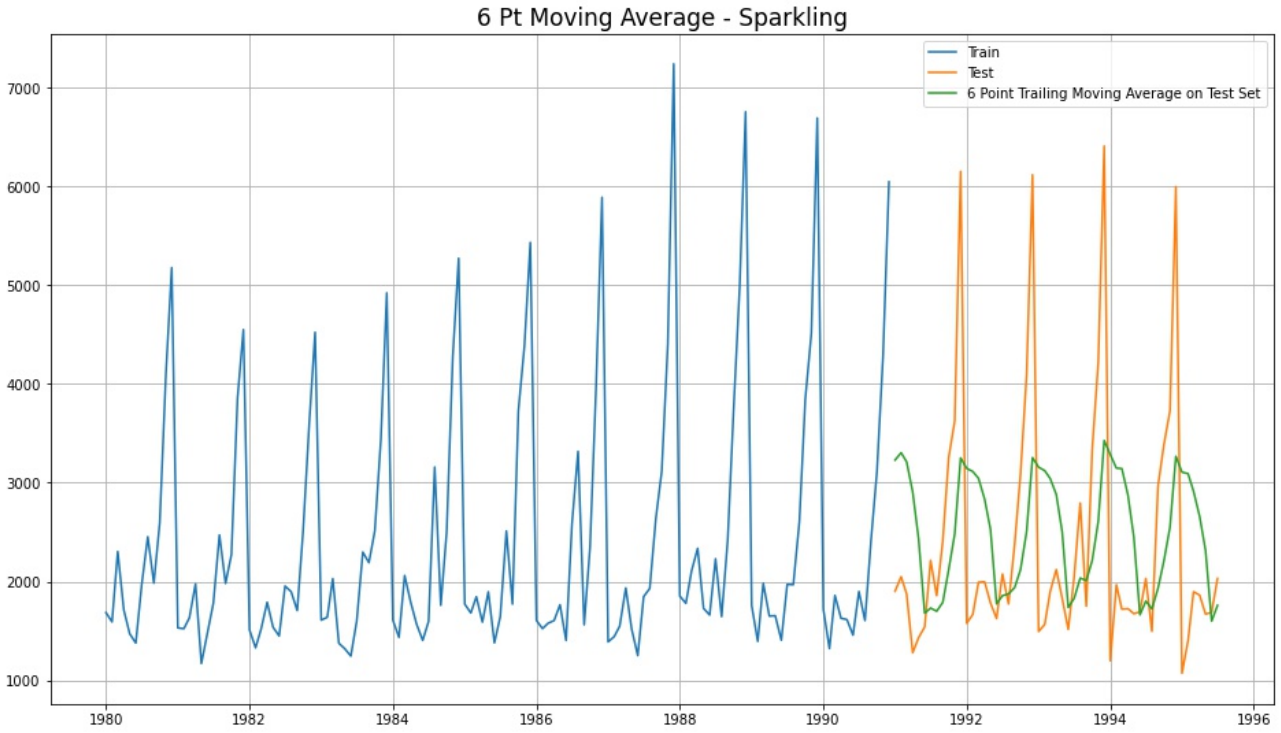


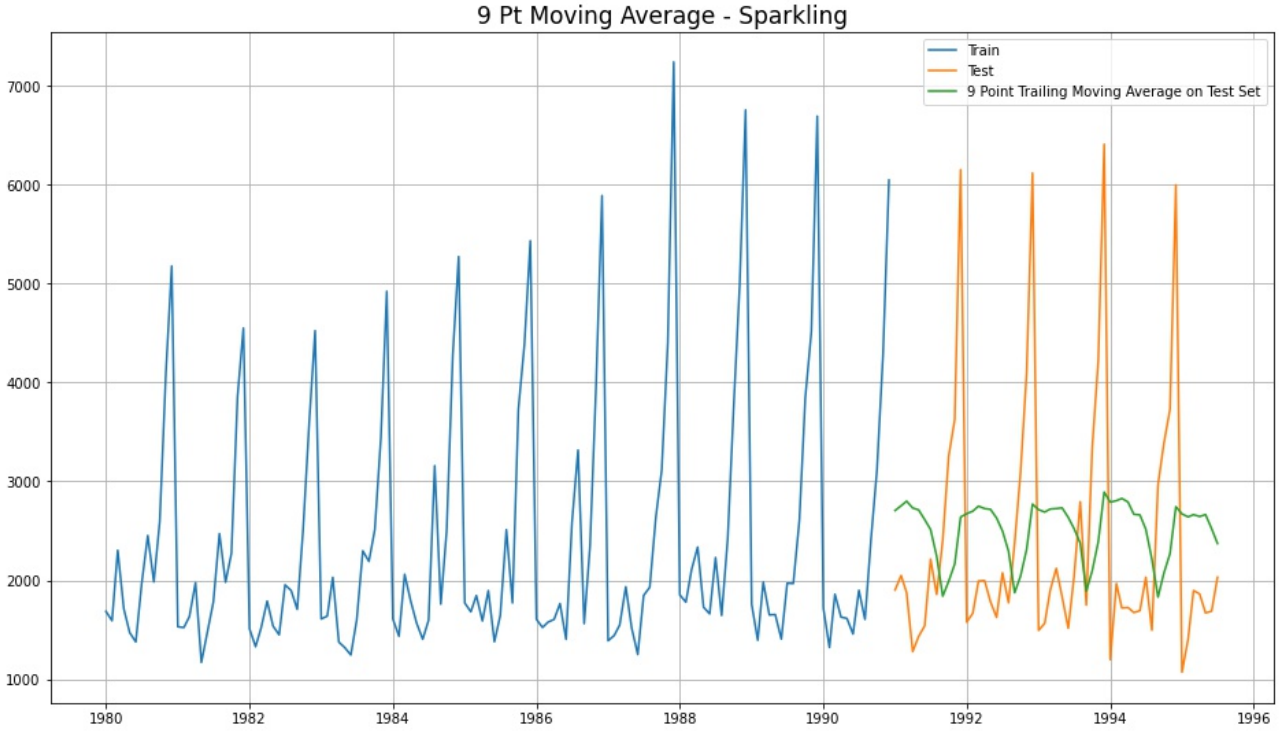
	Test RMSE Rose
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73

✦ Model 4.B - Moving Average (Sparkling)



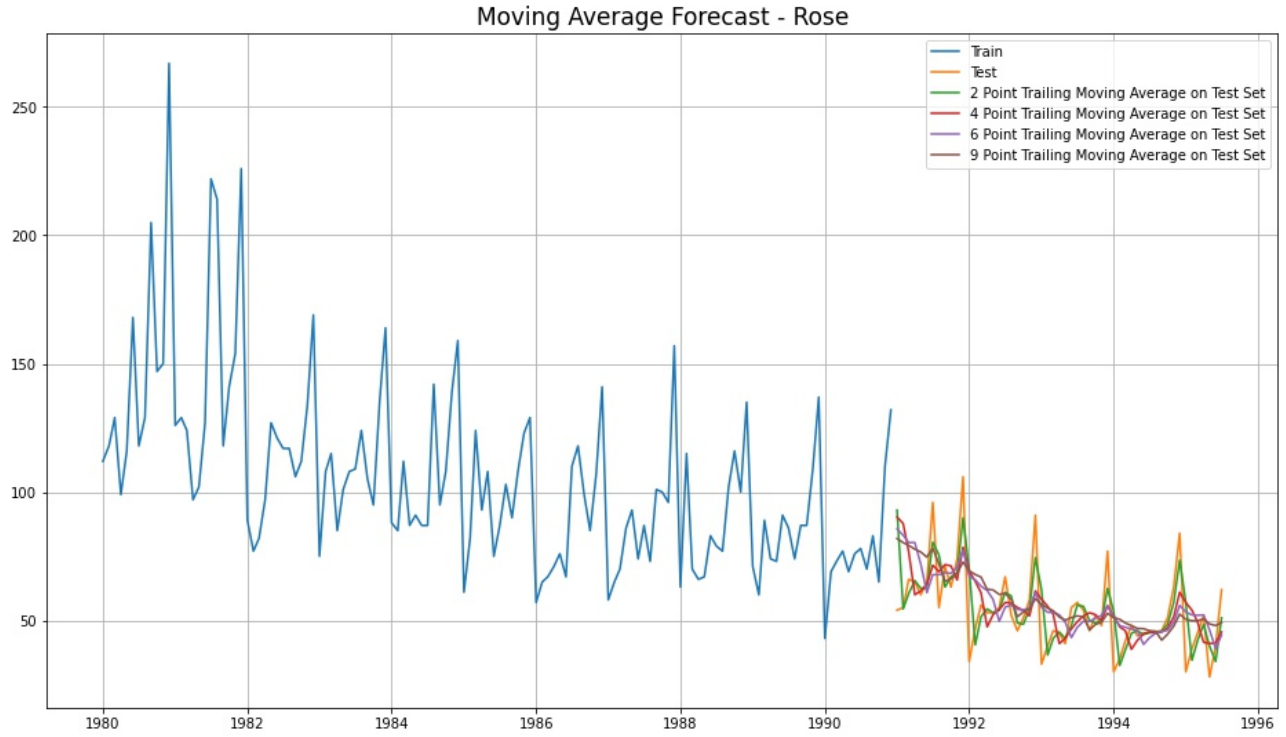


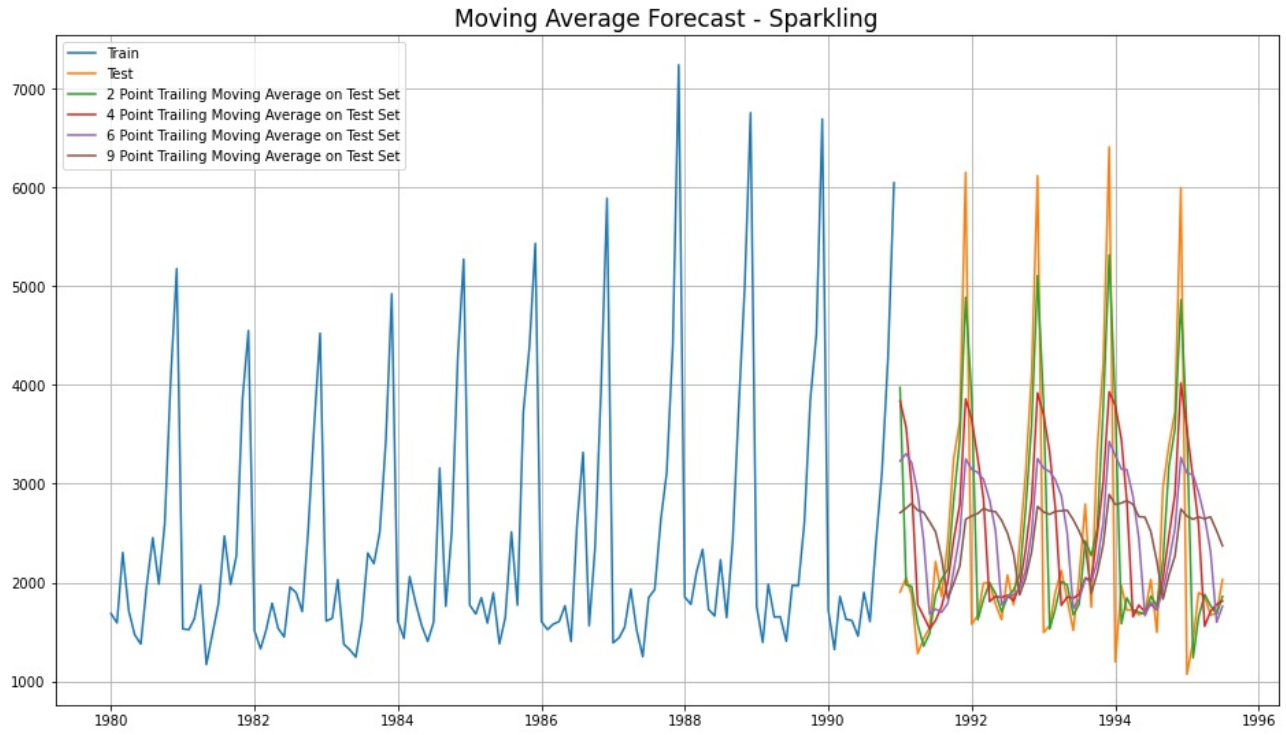




	Test RMSE Sparkling
2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28

✦ Consolidated Moving Average Forecasts (Rose & Sparkling)





✦ **NOTE -**

- We have built 4 models till now for both Rose and Sparkling Wine datasets
- We fitted various models to the Train split and Tested it on Test split. Accuracy metrics used is Root Mean Squared Error (RMSE) on Test data
- Model 1 - Linear Regression ($y_t = \beta_0 + \beta_1 X_t + \epsilon_t$)
  - We regressed variables 'Rose' and 'Sparkling' against their individual time instances
  - We modified the datasets and tagged individual sales to their time instances
  - TEST RMSE ROSE = 15.27 | TEST RMSE SPARKLING = 1389.14
- Model 2 - Naive Approach ($\hat{y}_{t+1} = y_t$)
  - Naive approach says that prediction for tomorrow is same as today
  - And, prediction for day-after is same as tomorrow
  - So, effectively all future predictions are going to be same as today
  - TEST RMSE ROSE = 79.72 | TEST RMSE SPARKLING = 3864.28
- Model 3 - Simple Average ($\hat{y}_{t+1} = \hat{y}_{t+2} = \ldots = \hat{y}_{t+n} = Mean(y_1, y_2, \ldots, y_t)$)
  - All future predictions are the same as the simple average of all data till today
  - TEST RMSE ROSE = 53.46 | TEST RMSE SPARKLING = 1275.08
- Model 4 - Moving Average (MA)
  - We calculate rolling means (Moving averages) over different intervals for the whole train data
  - 2 Pt MA ====> means, we find average of 1st and 2nd to predict 3rd
    similarly, average of 2nd and 3rd to predict 4th and so on
  - 4 Pt MA ====> means, we find average of 1st, 2nd, 3rd & 4th to predict 5th
    also, average of 2nd, 3rd, 4th & 5th to predict 6th and so on

- 2 PT MA ====>

  TEST RMSE ROSE = 11.53 | TEST RMSE SPARKLING = 813.40

- 4 PT MA ====>

  TEST RMSE ROSE = 14.45 | TEST RMSE SPARKLING = 1156.59

- 6 PT MA ====>

  TEST RMSE ROSE = 14.57 | TEST RMSE SPARKLING = 1283.93

- 9 PT MA ====>

  TEST RMSE ROSE = 14.73 | TEST RMSE SPARKLING = 1346.28

	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28
SimpleAverageModel	53.46	1275.08
2pointTrailingMovingAverage	**11.53**	**813.40**
4pointTrailingMovingAverage	14.45	1156.59
6pointTrailingMovingAverage	14.57	1283.93
9pointTrailingMovingAverage	14.73	1346.28

Consolidated Scores of Regression, Naive, Simple Average & Moving Average

- **Till now, Best Model which gives lowest RMSE score for both Rose and Sparkling is ——> 2 Pt Moving Average Model**

- We'll continue to forecast using Exponential Smoothing Models for both datasets of Rose and Sparkling Wine Sales

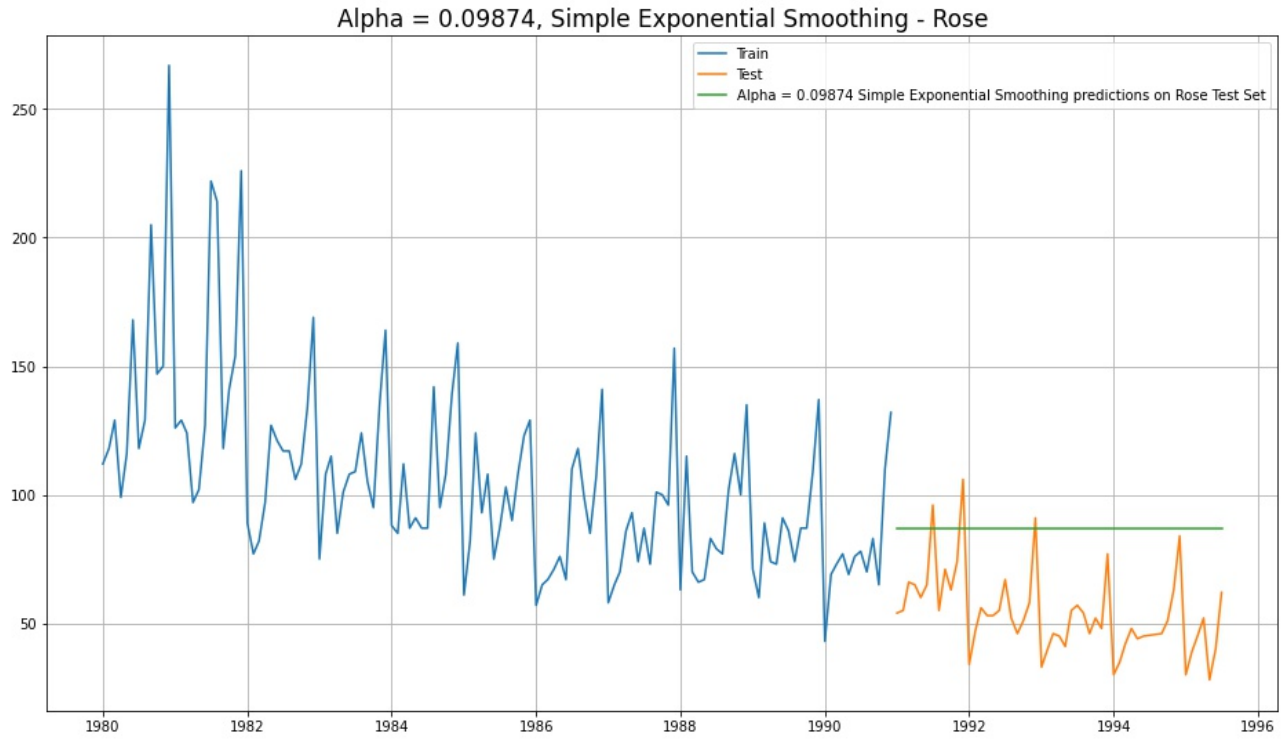
- Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous periods data with exponentially declining influence on the older observations

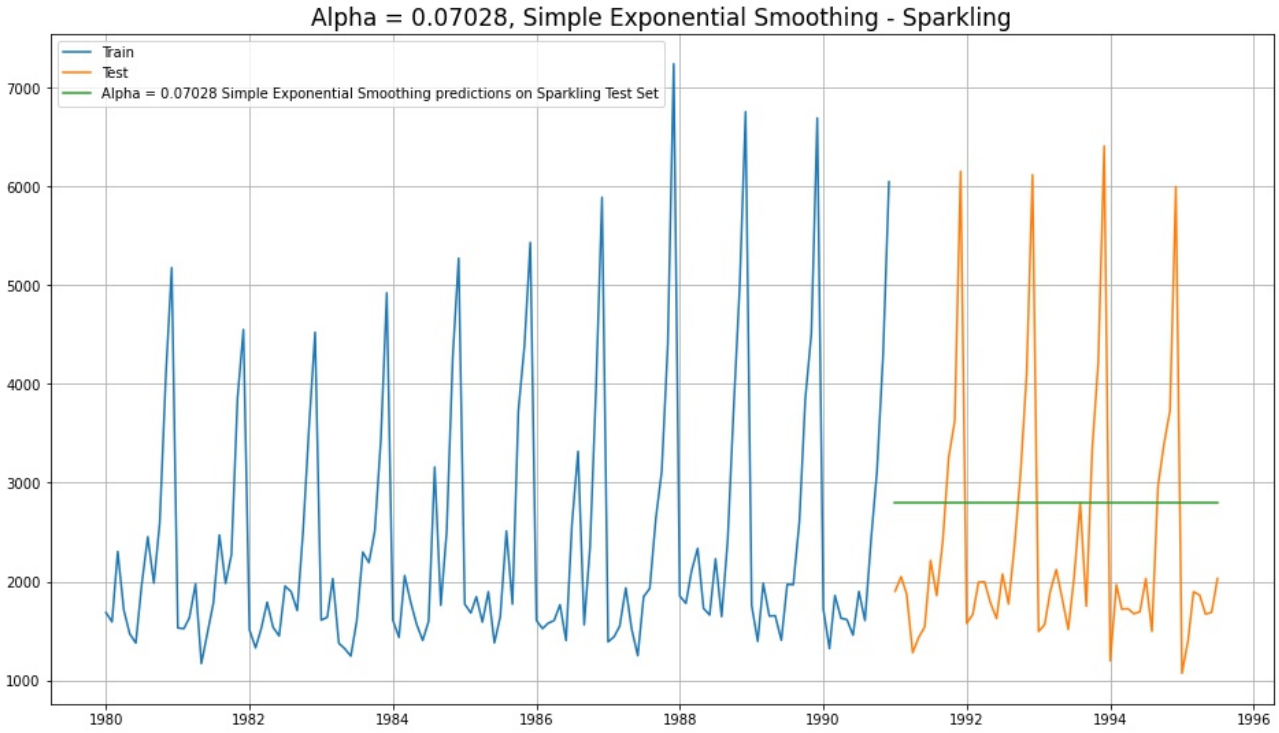
- Exponential smoothing methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality) where each can be None(N), Additive (N), Additive damped (Ad), Multiplicative (M) or Multiplicative damped (Md)

- One or more parameters control how fast the weights decay. The values of the parameters lie between 0 and 1

- We'll build following Exponential Smoothing Models -
    - Single Exponential Smoothing with Additive Errors - **ETS(A, N, N)**
    - Double Exponential Smoothing with Additive Errors, Additive Trends - **ETS(A, A, N)**
    - Triple Exponential Smoothing with Additive Errors, Additive Trends, Additive Seasonality - **ETS(A, A, A)**
    - Triple Exponential Smoothing with Additive Errors, Additive Trends, Multiplicative Seasonality - **ETS(A, A, M)**
    - Triple Exponential Smoothing with Additive Errors, Additive **DAMPED** Trends, Additive Seasonality - **ETS(A, Ad, A)**
    - Triple Exponential Smoothing with Additive Errors, Additive **DAMPED** Trends, Multiplicative Seasonality - **ETS(A, Ad, M)**

✦ Single Exponential Smoothing with Additive Errors - ETS(A, N, N)



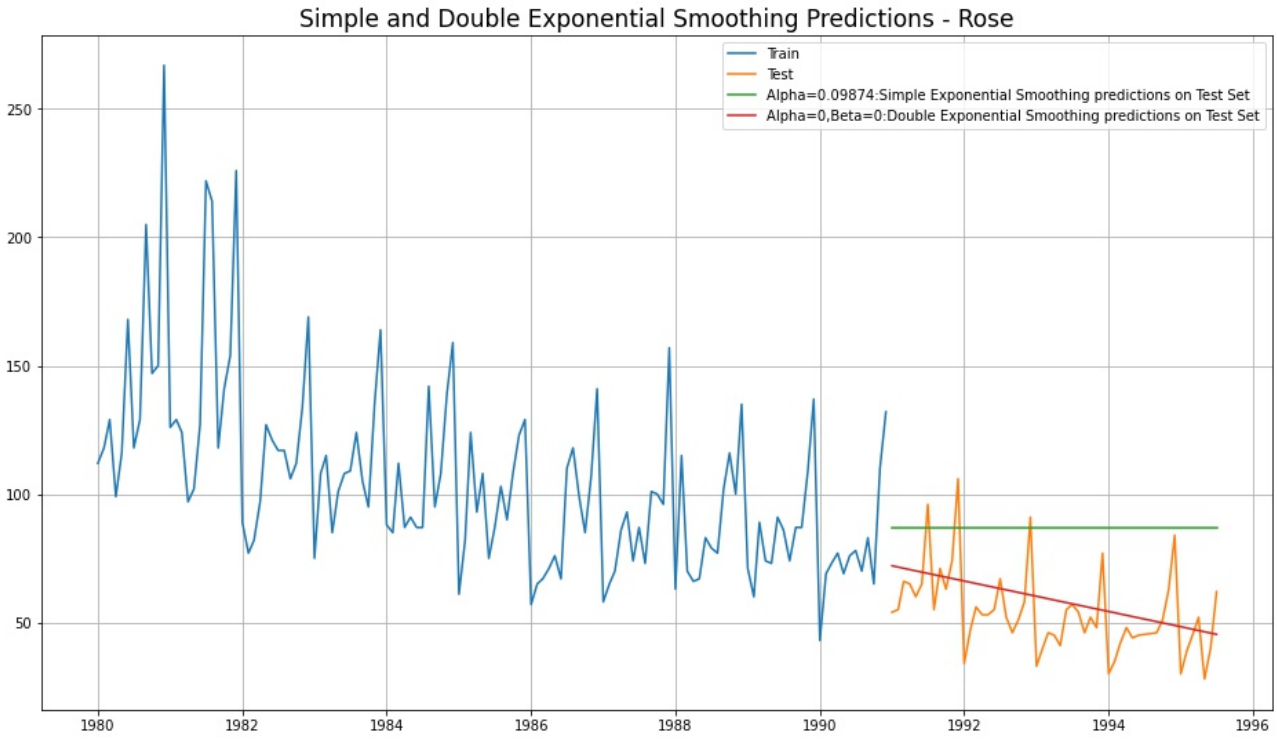


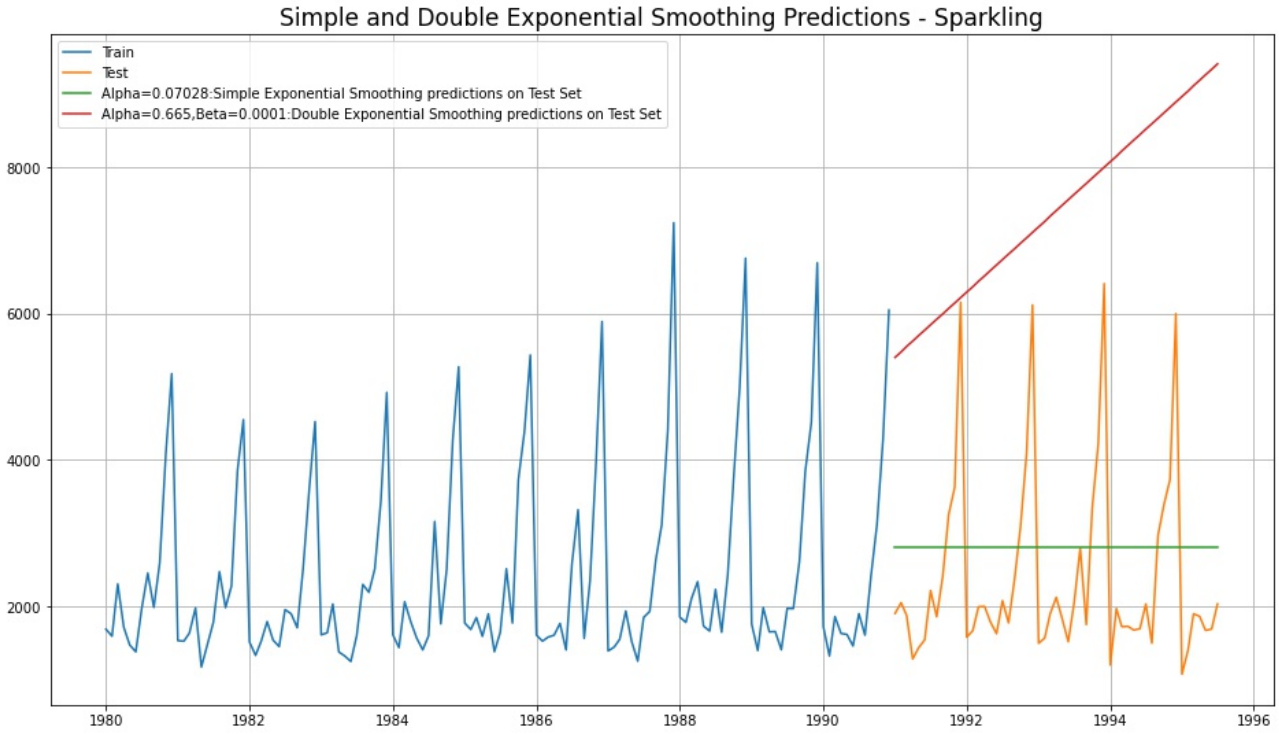
- For Rose - Level Parameter, Alpha = 0.09874
- For Sparkling - Level Parameter, Alpha = 0.07028

	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28
SimpleAverageModel	53.46	1275.08
2pointTrailingMovingAverage	**11.53**	**813.40**
4pointTrailingMovingAverage	14.45	1156.59
6pointTrailingMovingAverage	14.57	1283.93
9pointTrailingMovingAverage	14.73	1346.28
Simple Exponential Smoothing	36.80	1338.00

- Best Model till now for Rose and Sparkling ——- > 2 Pt Moving Average Model

✦ Double Exponential Smoothing with Additive Errors, Additive Trends - ETS(A, A, N)





- In Rose - DES has picked up the trend well. DES seems to perform better than SES here

- In Sparkling - DES shows a non-existent trend. DES does not perform well here

- Rose - Level parameter, Alpha = 0

  Trend parameter, Beta = 0

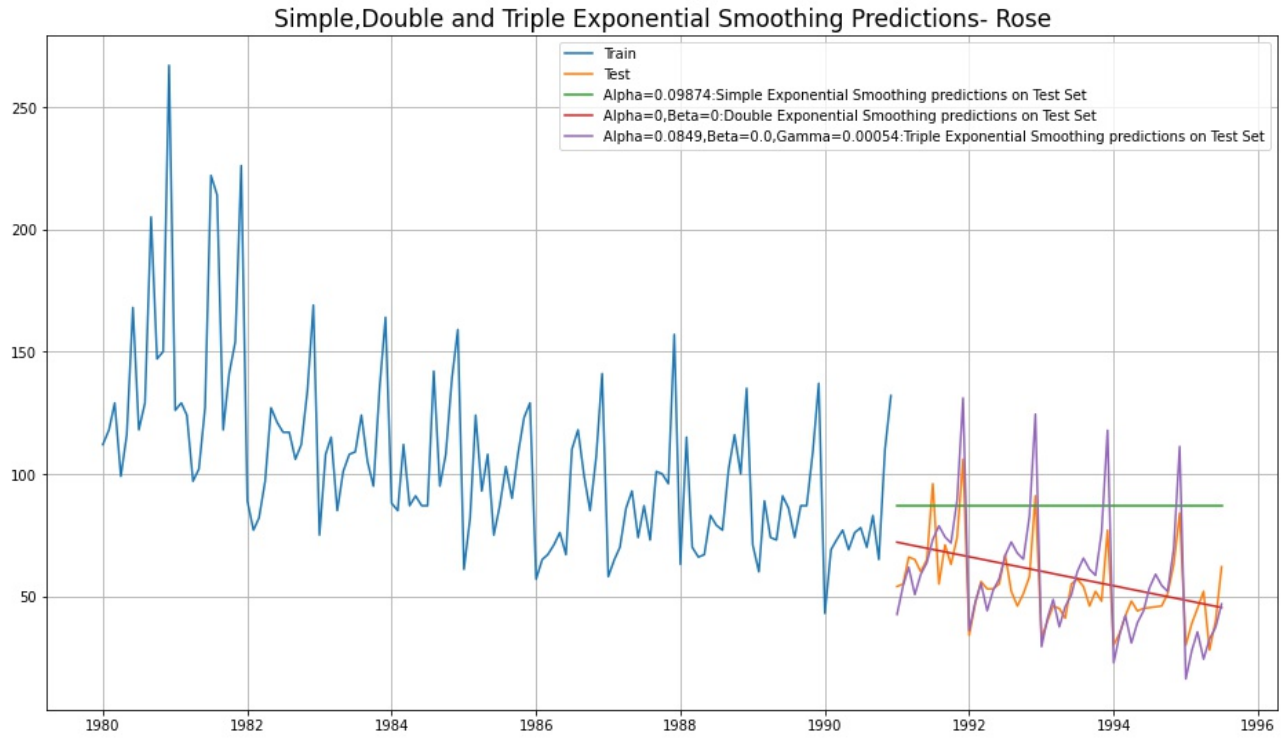
- Sparkling - Level parameter, Alpha = 0.665

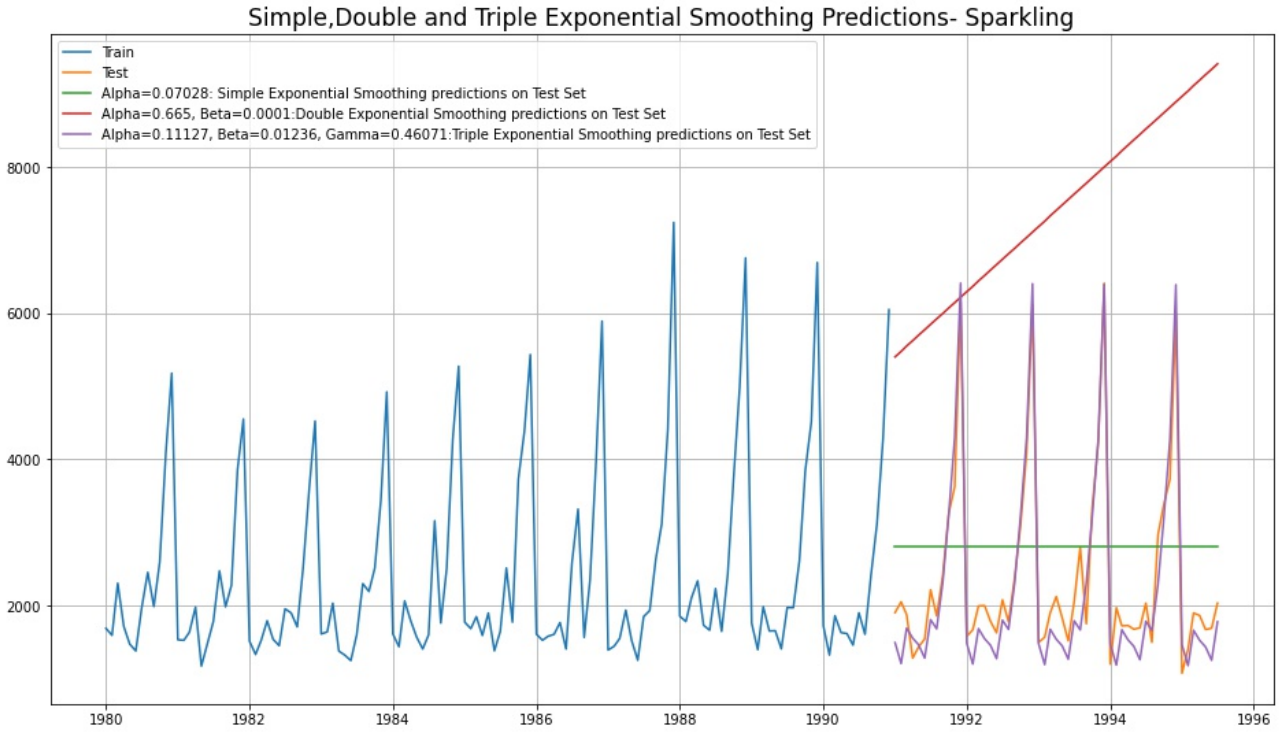
  Trend parameter, Beta = 0.0001

	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28
SimpleAverageModel	53.46	1275.08
2pointTrailingMovingAverage	**11.53**	**813.40**
4pointTrailingMovingAverage	14.45	1156.59
6pointTrailingMovingAverage	14.57	1283.93
9pointTrailingMovingAverage	14.73	1346.28
Simple Exponential Smoothing	36.80	1338.00
Double Exponential Smoothing	15.27	5291.88

- Best Model till now for Rose and Sparkling ——- > 2 Pt Moving Average Model

✦ Triple Exponential Smoothing with Additive Errors, Additive Trends, Additive Seasonality - ETS(A, A, A)





- In Rose & Sparkling - TES has picked up the trend and seasonality very well
- Rose - Level parameter, Alpha = 0.0849

  Trend parameter, Beta = 0.0

  Seasonality parameter, Gamma = 0.00054
- Sparkling - Level parameter, Alpha = 0.11127

  Trend parameter, Beta = 0.01236

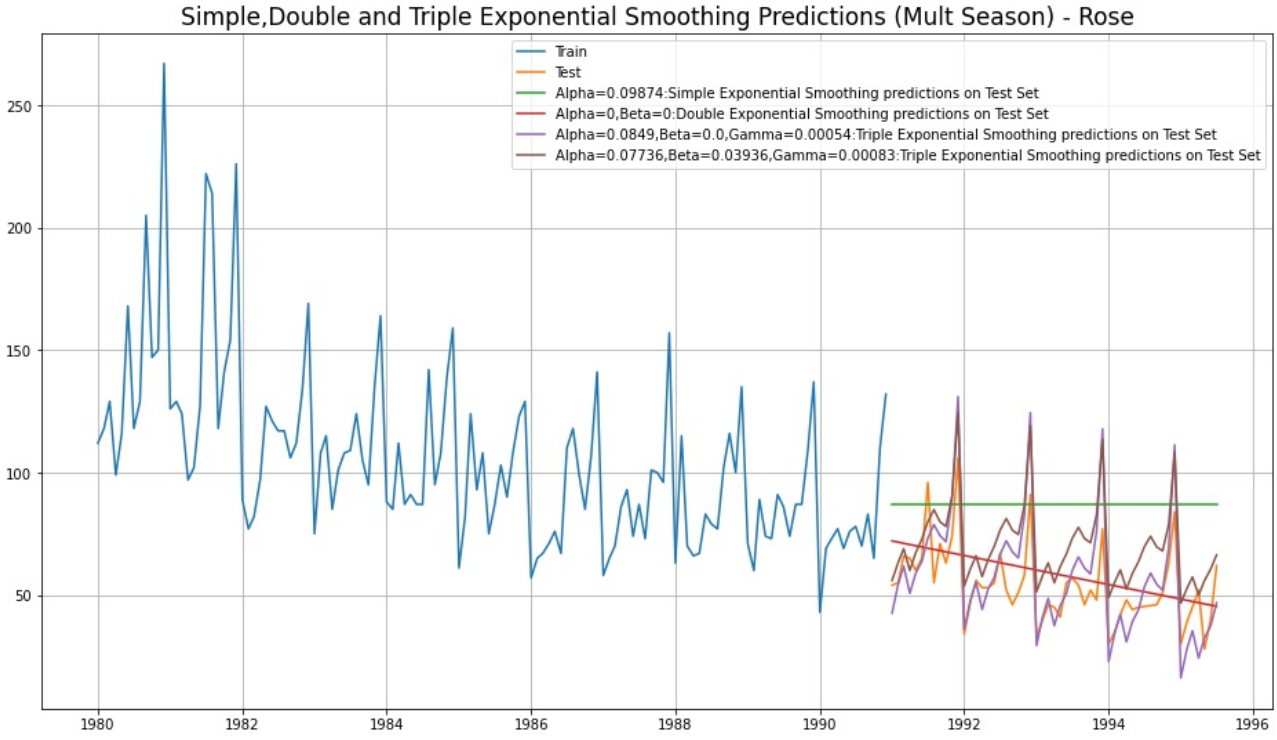
  Seasonality parameter, Gamma = 0.46071

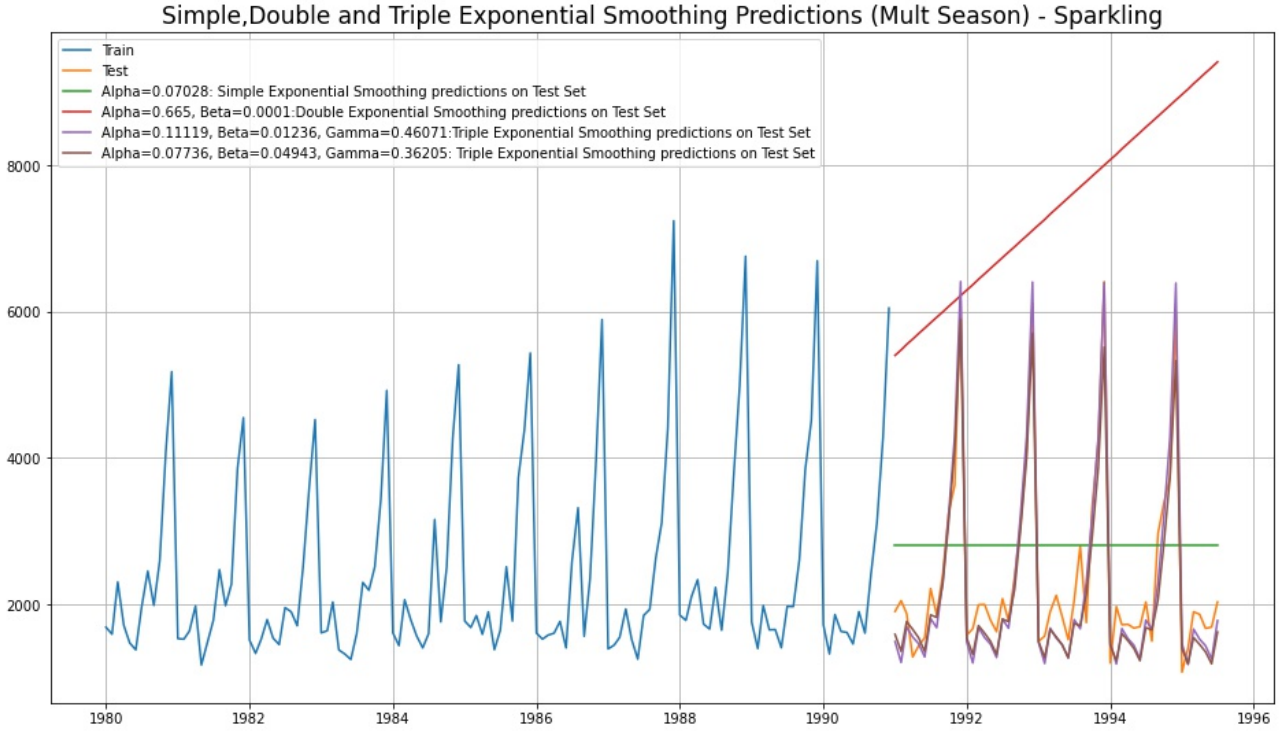
	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28
SimpleAverageModel	53.46	1275.08
2pointTrailingMovingAverage	**11.53**	813.40
4pointTrailingMovingAverage	14.45	1156.59
6pointTrailingMovingAverage	14.57	1283.93
9pointTrailingMovingAverage	14.73	1346.28
Simple Exponential Smoothing	36.80	1338.00
Double Exponential Smoothing	15.27	5291.88
Triple Exponential Smoothing (Additive Season)	14.24	**378.63**

- Till now, Best Model for Rose ——> 2 Pt Moving Average

  Best Model for Sparkling ——> Holt-Winter - ETS (A, A, A)

✦ Triple Exponential Smoothing with Additive Errors, Additive Trends, Multiplicative Seasonality - ETS(A, A, M)





- Rose - Level parameter, Alpha = 0.07736

  Trend parameter, Beta = 0.03936

  Seasonality parameter, Gamma = 0.00083

- Sparkling - Level parameter, Alpha = 0.07736

  Trend parameter, Beta = 0.04943

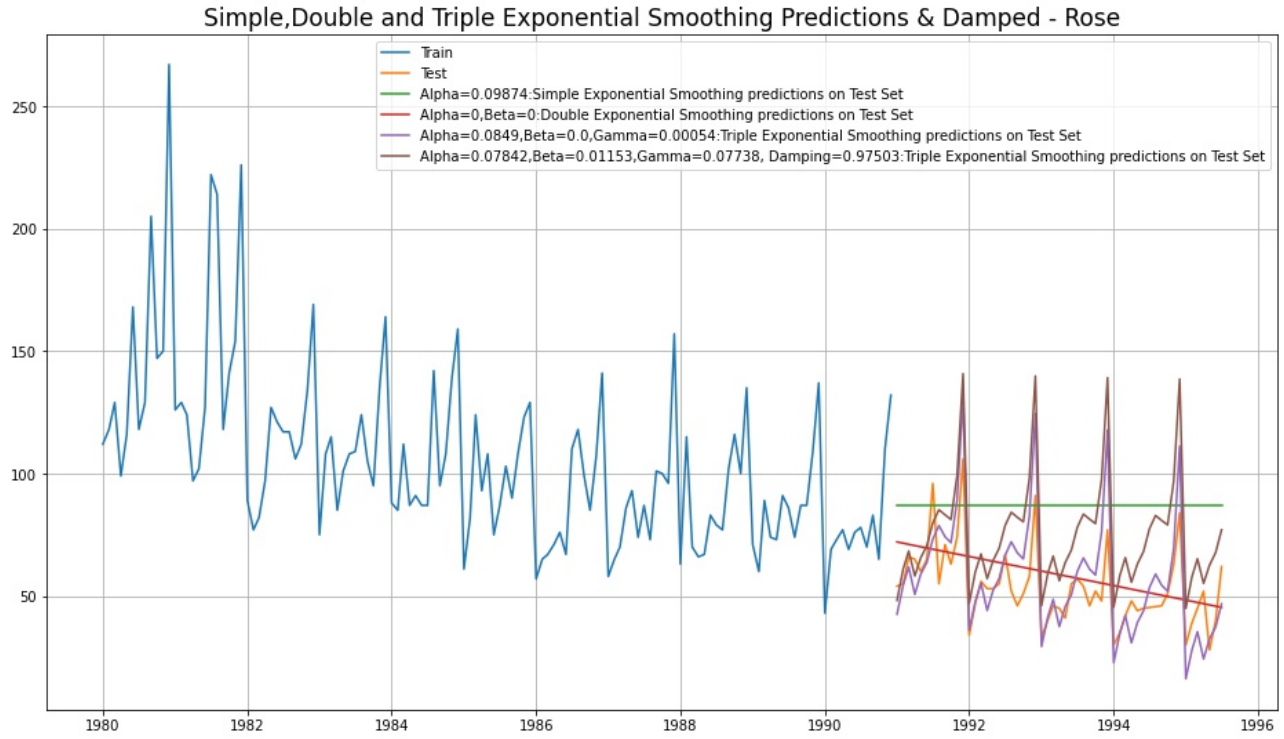
  Seasonality parameter, Gamma = 0.36205

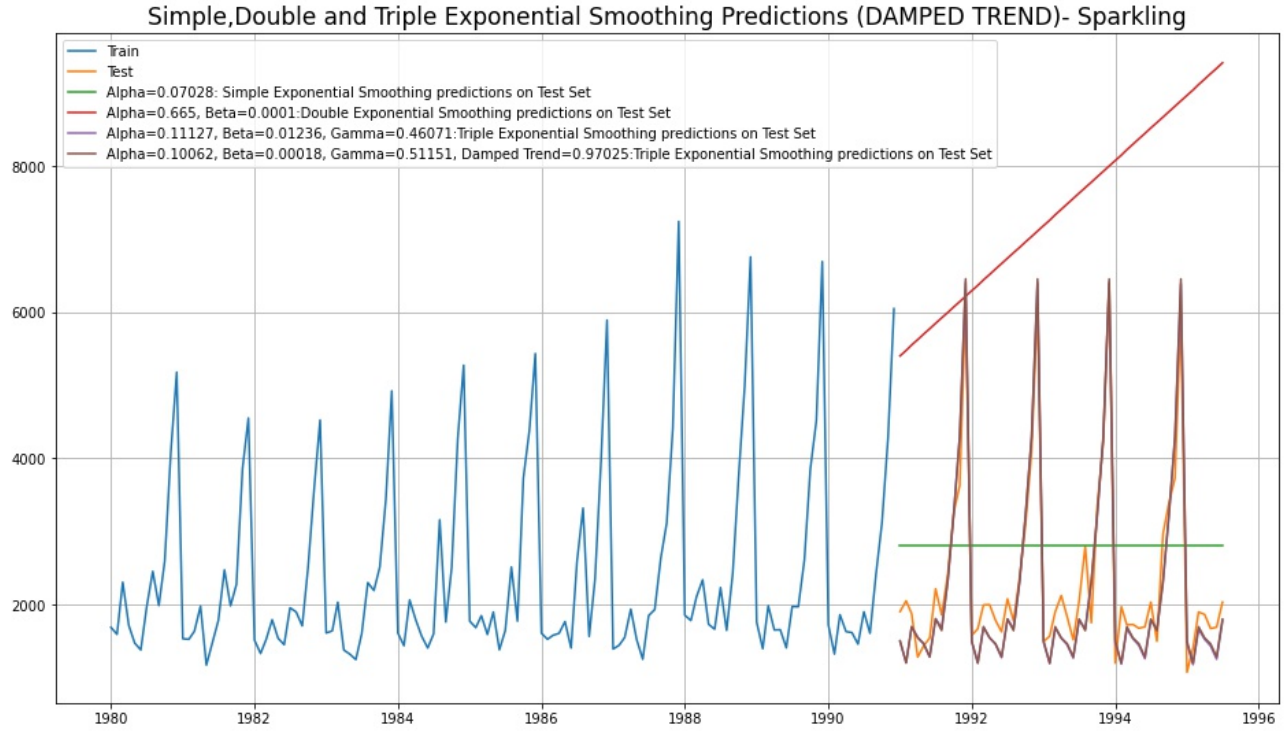
	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28
SimpleAverageModel	53.46	1275.08
2pointTrailingMovingAverage	**11.53**	813.40
4pointTrailingMovingAverage	14.45	1156.59
6pointTrailingMovingAverage	14.57	1283.93
9pointTrailingMovingAverage	14.73	1346.28
Simple Exponential Smoothing	36.80	1338.00
Double Exponential Smoothing	15.27	5291.88
Triple Exponential Smoothing (Additive Season)	14.24	**378.63**
Triple Exponential Smoothing (Multiplicative Season)	19.11	403.71

- Till now, Best Model for Rose ——> 2 Pt Moving Average

  Best Model for Sparkling ——> Holt-Winter - ETS (A, A, A)

✦ Triple Exponential Smoothing with Additive Errors, Additive DAMPED Trends, Additive Seasonality - ETS(A, Ad, A)





- Rose - Level parameter, Alpha = 0.07842

  Trend parameter, Beta = 0.01153

  Seasonality parameter, Gamma = 0.07738

  Damping factor = 0.97503

- Sparkling - Level parameter, Alpha = 0.10062

  Trend parameter, Beta = 0.00018

  Seasonality parameter, Gamma = 0.97025

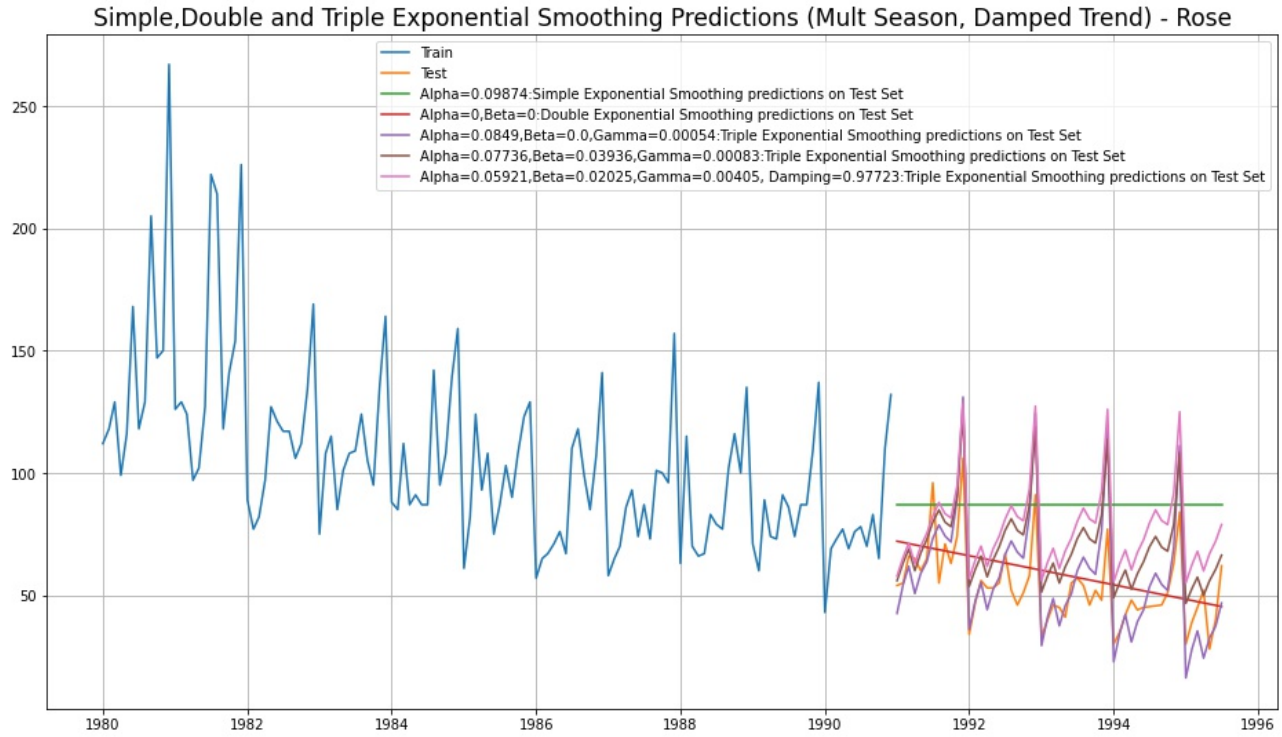
  Damping factor = 0.97025

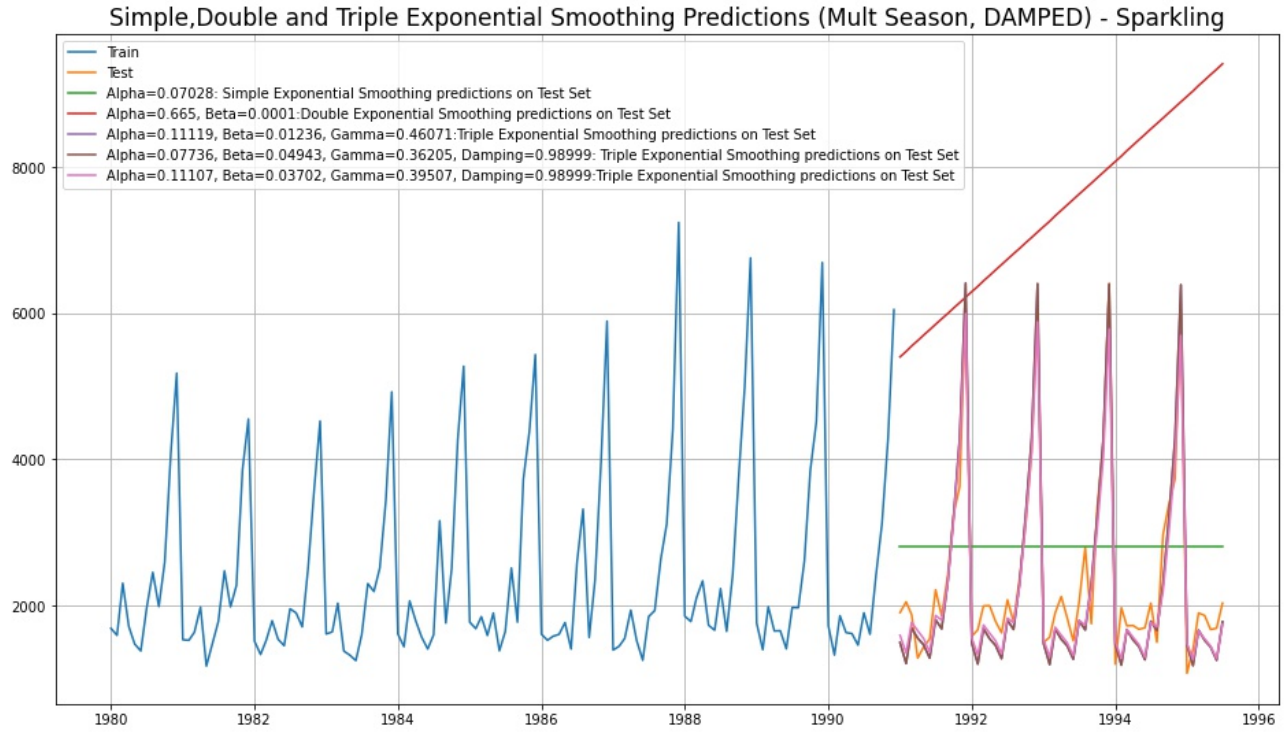
	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28
SimpleAverageModel	53.46	1275.08
2pointTrailingMovingAverage	**11.53**	813.40
4pointTrailingMovingAverage	14.45	1156.59
6pointTrailingMovingAverage	14.57	1283.93
9pointTrailingMovingAverage	14.73	1346.28
Simple Exponential Smoothing	36.80	1338.00
Double Exponential Smoothing	15.27	5291.88
Triple Exponential Smoothing (Additive Season)	14.24	**378.63**
Triple Exponential Smoothing (Multiplicative Season)	19.11	403.71
Triple Exponential Smoothing (Additive Season, Damped Trend)	26.04	378.63

- Till now, Best Model for Rose ——> 2 Pt Moving Average

  Best Model for Sparkling ——> Holt-Winter - ETS (A, A, A)

✦ Triple Exponential Smoothing with Additive Errors, Additive DAMPED Trends, Multiplicative Seasonality - ETS(A, Ad, M)





- Rose - Level parameter, Alpha = 0.05921

  Trend parameter, Beta = 0.02025

  Seasonality parameter, Gamma = 0.00405

  Damping factor = 0.97723

- Sparkling - Level parameter, Alpha = 0.11107

  Trend parameter, Beta = 0.03702

  Seasonality parameter, Gamma = 0.39507

  Damping factor = 0.98999

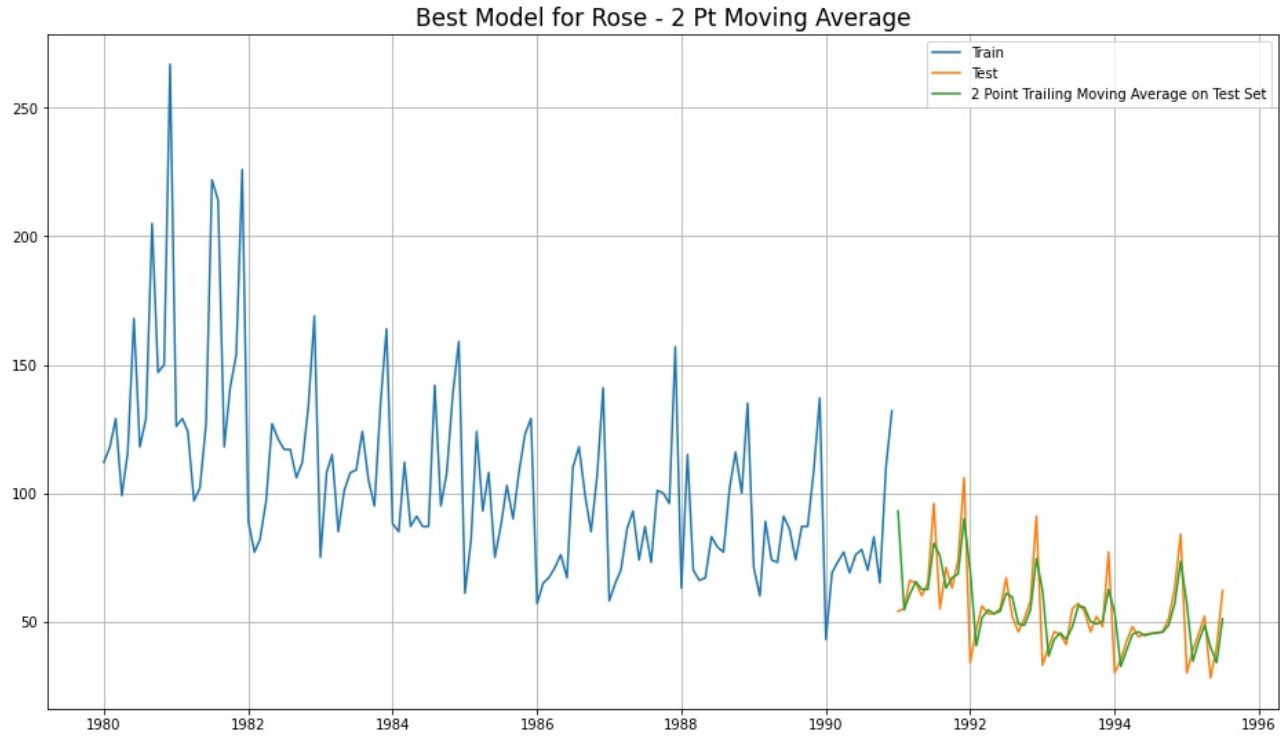
	Test RMSE Rose	**Test RMSE Sparkling**
RegressionOnTime	15.27	1389.14
NaiveModel	79.72	3864.28
SimpleAverageModel	53.46	1275.08
2pointTrailingMovingAverage	**11.53**	813.40
4pointTrailingMovingAverage	14.45	1156.59
6pointTrailingMovingAverage	14.57	1283.93
9pointTrailingMovingAverage	14.73	1346.28
Simple Exponential Smoothing	36.80	1338.00
Double Exponential Smoothing	15.27	5291.88
Triple Exponential Smoothing (Additive Season)	14.24	378.63
Triple Exponential Smoothing (Multiplicative Season)	19.11	403.71
Triple Exponential Smoothing (Additive Season, Damped Trend)	26.04	378.63
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)	25.99	**352.45**

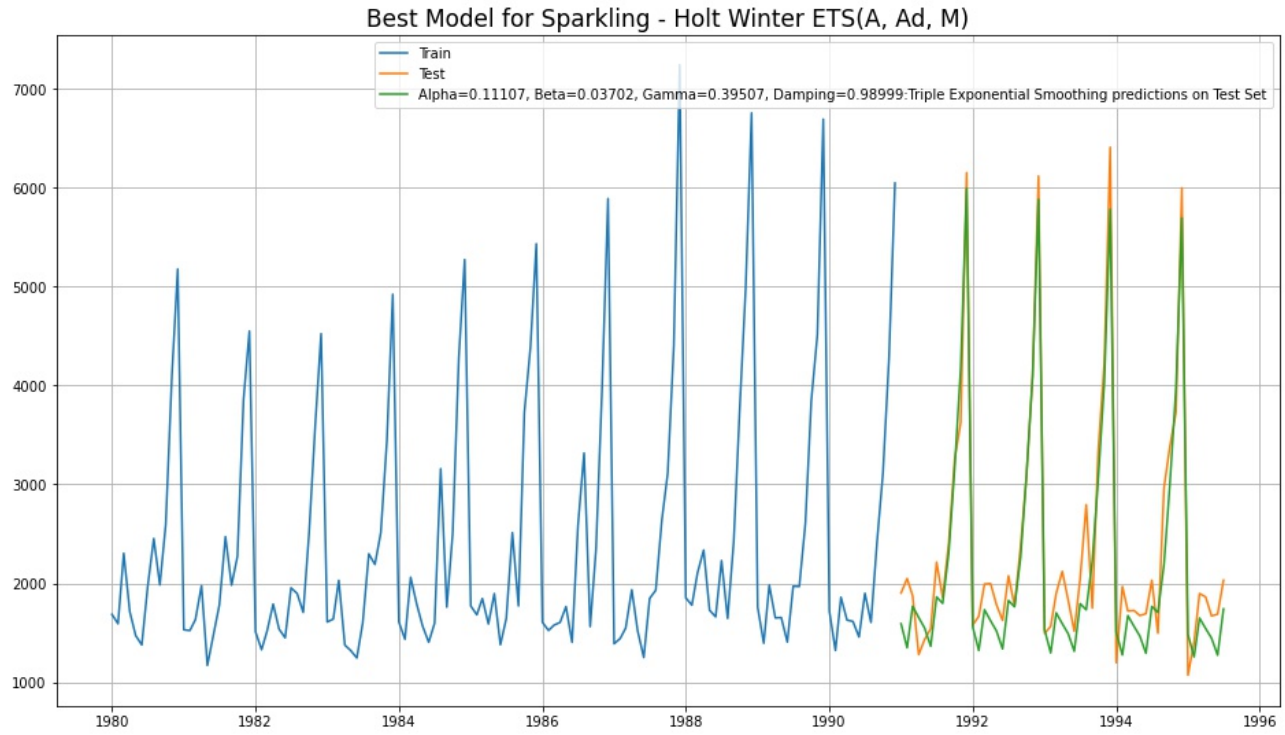
- We conclude that models with least RMSE,

Best Model for Rose ——> 2 Pt Moving Average

Best Model for Sparkling ——> Holt-Winter Damped Trend
ETS (A, Ad, M)

✦ Best Models for Rose and Sparkling -





[Q 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.

✦ **To Check Stationarity of Data** -

- We use Augmented Dicky - Fuller (ADF) Test to check the Stationarity of Data
- Hypotheses of ADF Test :

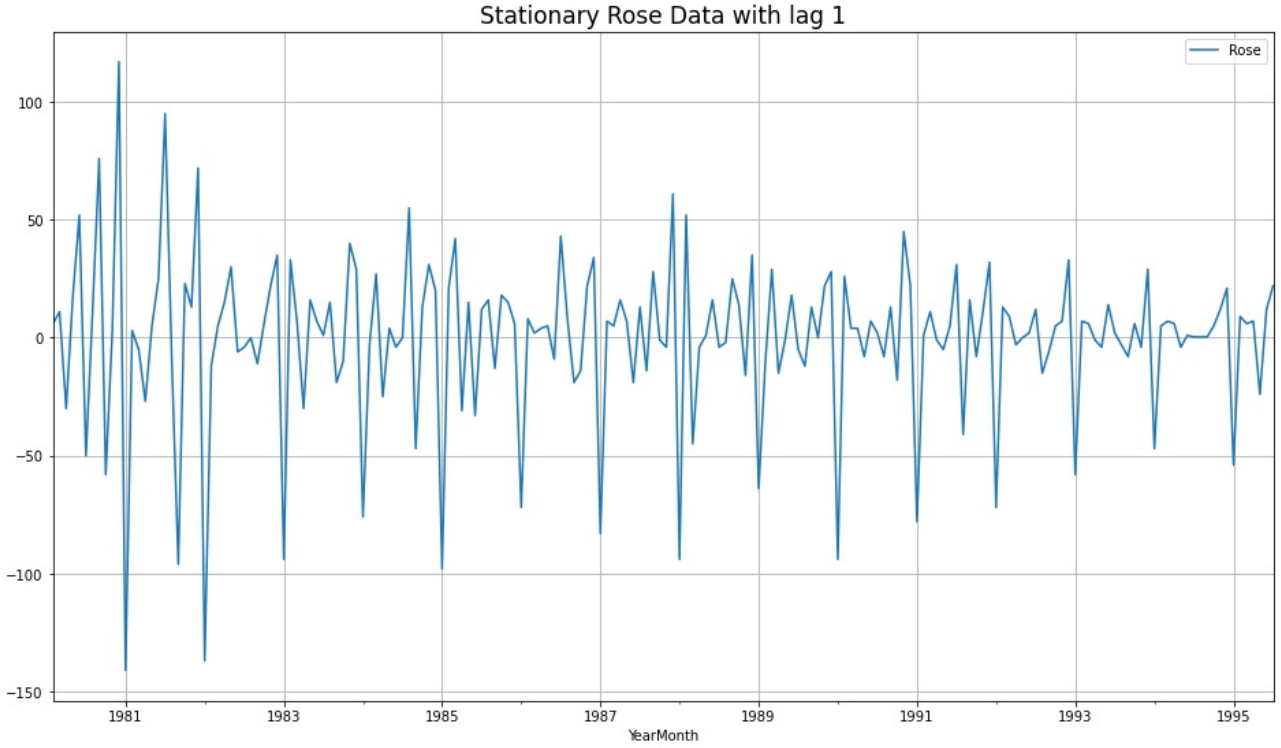
$$H_0 : \textit{Time Series is not Stationary}$$

$$H_a : \textit{Time Series is Stationary}$$

- So for Industry standard (also given for this problem), the Confidence Interval is 95%
- Hence, alpha = 0.05
- So in ADF Test, if p-value < alpha ===> We reject the Null Hypothesis and hence conclude that given Time Series is Stationary
- So in ADF Test, if p-value > alpha ===> We fail to reject the Null Hypothesis and hence conclude that given Time Series is Not Stationary
- If Time Series is not Stationary then we apply one level of differencing and check for Stationarity again.
- Again, if the Time Series is still not Stationary, we apply one more level of differencing and check for Stationarity again
- Generally, with max 2 levels of differencing, Time Series becomes Stationary
- Once the Time Series is Stationary then we are ready to apply ARIMA / SARIMA models

✦ Stationarity of Rose Wine Dataset -

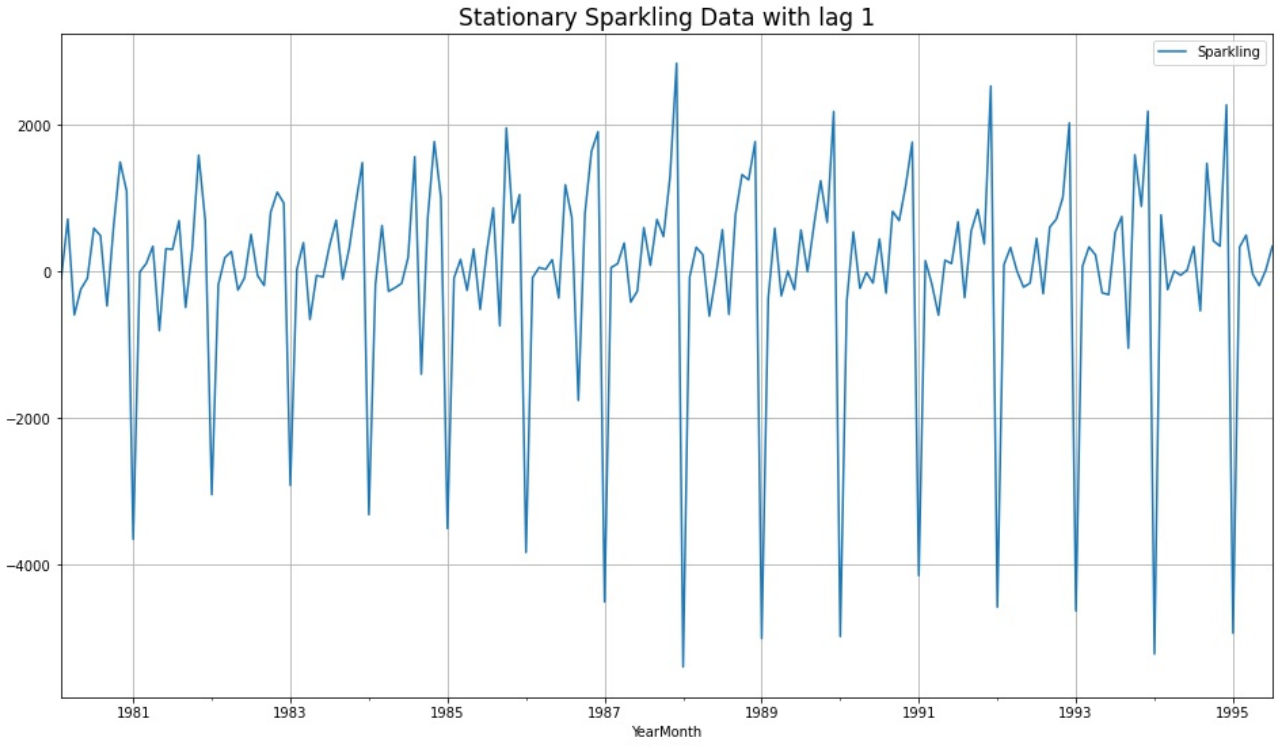
- Augmented Dicky-Fuller Test was applied to the whole Rose dataset
- We found, p-value = 0.4671
- Here, p-value > alpha=0.05
- We fail to reject the Null Hypothesis and hence conclude that Rose Wine Time Series is Not Stationary
- We take 1 level of differencing and check again for Stationarity
- Now, p-value = 3.0159e-11 ≈ 0.00
- Now, p-value < alpha=0.05
- Now, we reject the Null Hypothesis and conclude that Rose Time Series is Stationary with a lag of 1



✦ Stationarity of Sparkling Wine Dataset -

- Augmented Dicky-Fuller Test was applied to the whole Sparkling dataset
- We found, p-value = 0.70559
- Here, p-value > alpha=0.05

- We fail to reject the Null Hypothesis and hence conclude that Sparkling Wine Time Series is Not Stationary
- We take 1 level of differencing and check again for Stationarity
- Now, p-value = 0.00
- Now, p-value < alpha=0.05
- Now, we reject the Null Hypothesis and conclude that Sparkling Time Series is Stationary with a lag of 1



[Q 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

✦ ARIMA / SARIMA Models -

- ARIMA is an acronym for Auto-Regressive Integrated Moving Average
- SARIMA stands for Seasonal ARIMA, when the TS has seasonality

- ARIMA / SARIMA are forecasting models on Stationary Time Series

✦ ARIMA / SARIMA Modelling on Train Rose & Sparkling Data -

- We check for stationarity of Train Rose & Sparkling data by using Augmented Dicky Fuller Test
- We take a difference of 1 and make both these datasets Stationary
- We apply the following iterations to both these datasets -
    1. ARIMA Automated
    2. SARIMA Automated

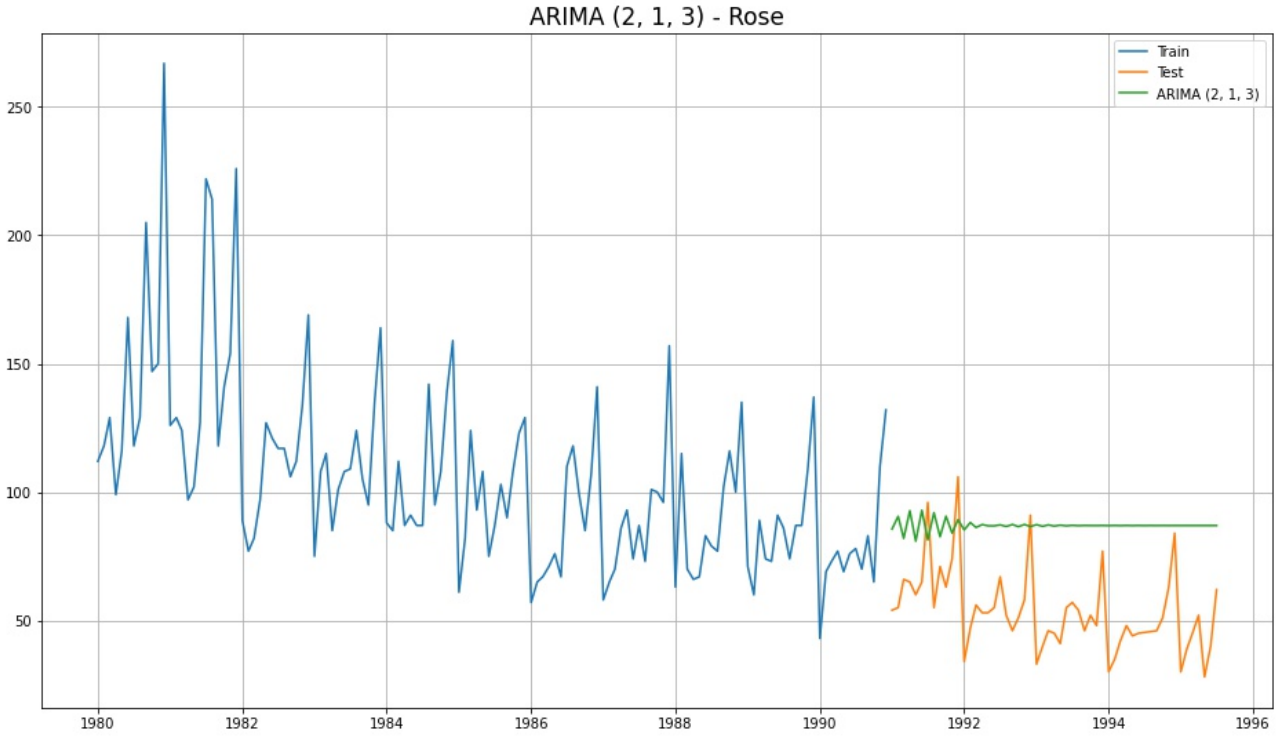
1. ARIMA Automated -

- We create a grid of all possible combinations of (p, d, q)
- Range of p = Range of q = 0 to 3, Constant d = 1
- Few Examples of the grid -

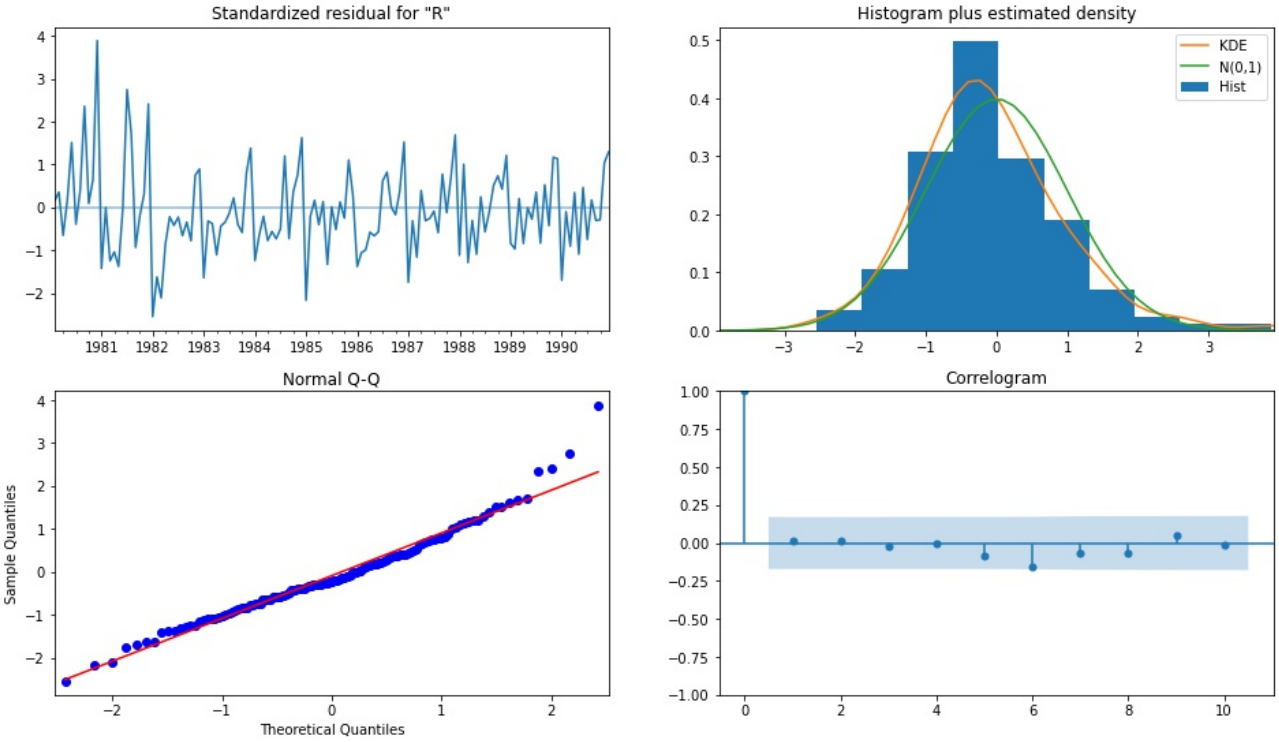
    Model: (0, 1, 2)
    Model: (0, 1, 3)
    Model: (1, 1, 0)
    Model: (1, 1, 1)
    Model: (1, 1, 2)
    Model: (1, 1, 3)
    Model: (2, 1, 0)
    Model: (2, 1, 1)
    Model: (2, 1, 2)
    Model: (2, 1, 3)
    Model: (3, 1, 0)
    Model: (3, 1, 1)

- We fit ARIMA models to each of these combinations for both datasets
- We choose the combination with the least Akaike Information Criteria (AIC)
- We fit ARIMA to this combination of (p, d, q) to the Train set and forecast on the Test set
- Finally, we check the accuracy of this model by checking RMSE of Test set

- For **Rose**, Best Combination with **Least AIC** is - **(p, d, q) —-> (2, 1, 3)**

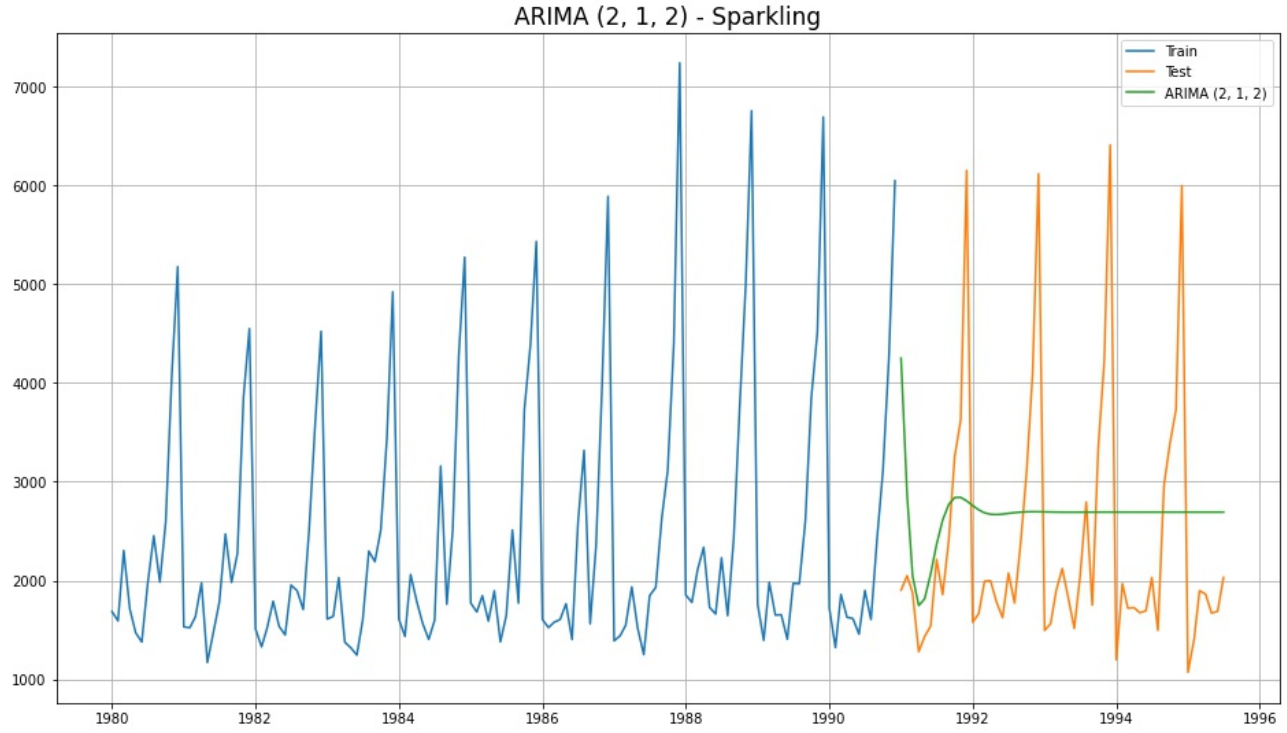


ARIMA (2, 1, 3) Diagnostic Plot - Rose

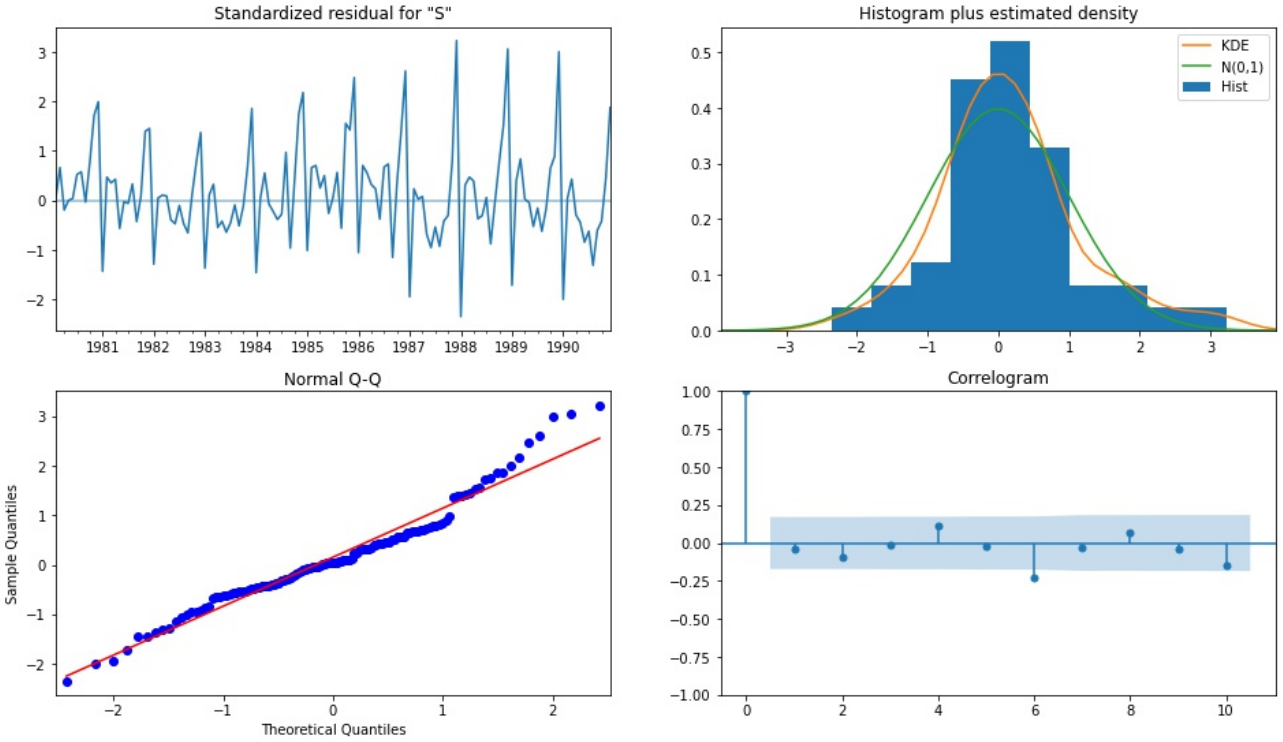


	Test RMSE Rose	**Test MAPE Rose**
ARIMA(2,1,3)	36.81	75.84

- For **Sparkling**, Best Combination with **Least AIC** is - **(p, d, q) —-> (2, 1, 2)**



ARIMA (2, 1, 2) Diagnostic Plot - Sparkling



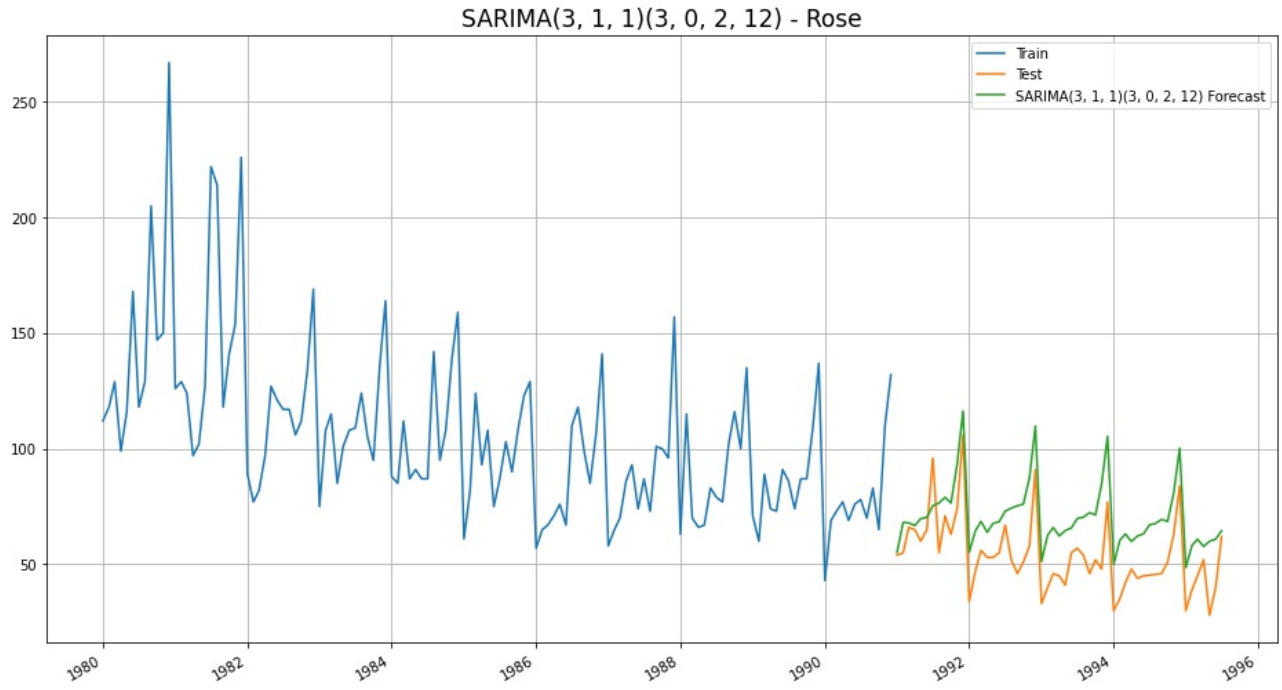
	Test RMSE Sparkling	**Test MAPE Sparkling**
ARIMA(2,1,2)	1299.98	47.10

2. SARIMA Automated -

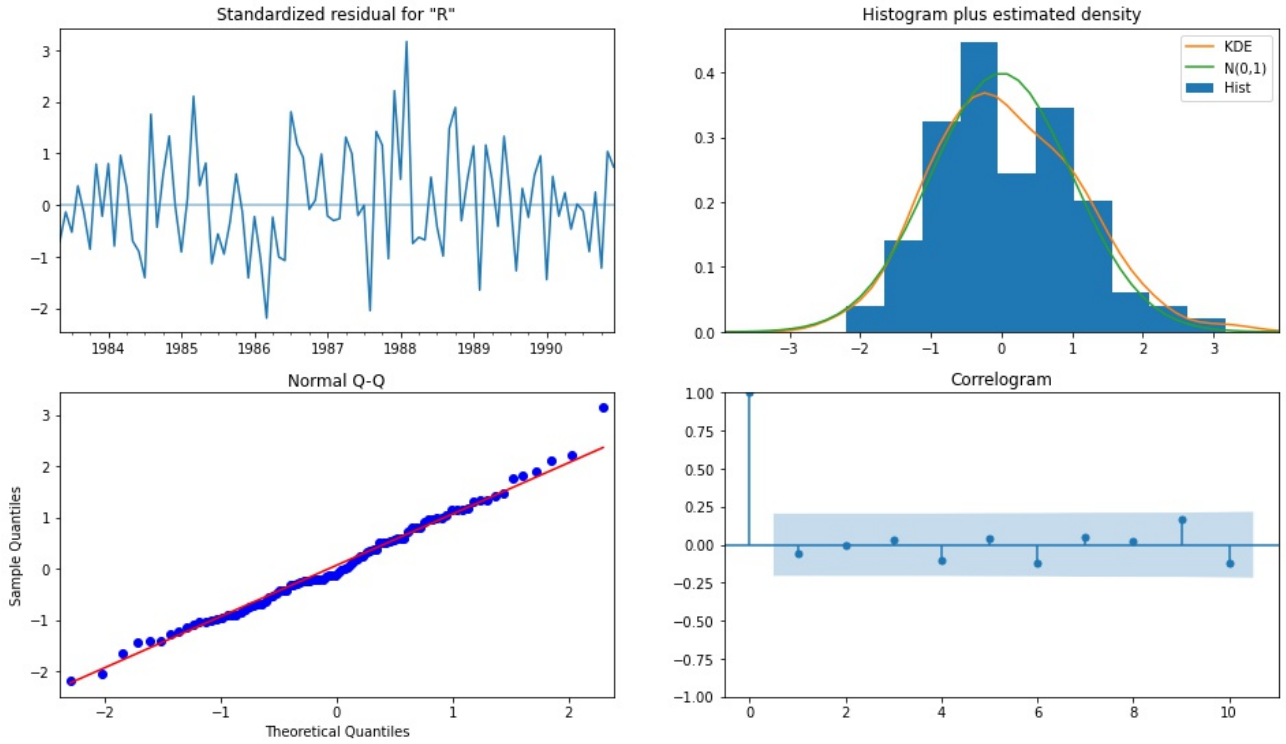
- We create a grid of all possible combinations of (p, d, q) along with Seasonal (P, D, Q) & Seasonality of 12 (for both datasets)
- Range of p = Range of q = 0 to 3, Constant d = 1
- Range of Seasonal P = Range of Seasonal Q = 0 to 3, Constant D = 1, Seasonality m = 12
- Few Examples of the grid (p, d, q) (P, D, Q, m) -

  Model: (0, 1, 2)(0, 0, 2, 12)
  Model: (0, 1, 3)(0, 0, 3, 12)
  Model: (1, 1, 0)(1, 0, 0, 12)
  Model: (1, 1, 1)(1, 0, 1, 12)
  Model: (1, 1, 2)(1, 0, 2, 12)
  Model: (1, 1, 3)(1, 0, 3, 12)
  Model: (2, 1, 0)(2, 0, 0, 12)
  Model: (2, 1, 1)(2, 0, 1, 12)
  Model: (2, 1, 2)(2, 0, 2, 12)
  Model: (2, 1, 3)(2, 0, 3, 12)
  Model: (3, 1, 0)(3, 0, 0, 12)

- We fit SARIMA models to each of these combinations and select with least AIC
- We fit SARIMA to this best combination of (p, d, q) (P, D, Q, m) to the Train set and forecast on the Test set. Then, we check accuracy using RMSE on Test set
- For **Rose**, Best Combination with **Least AIC** is - **(3, 1, 1) (3, 0, 2, 12)**

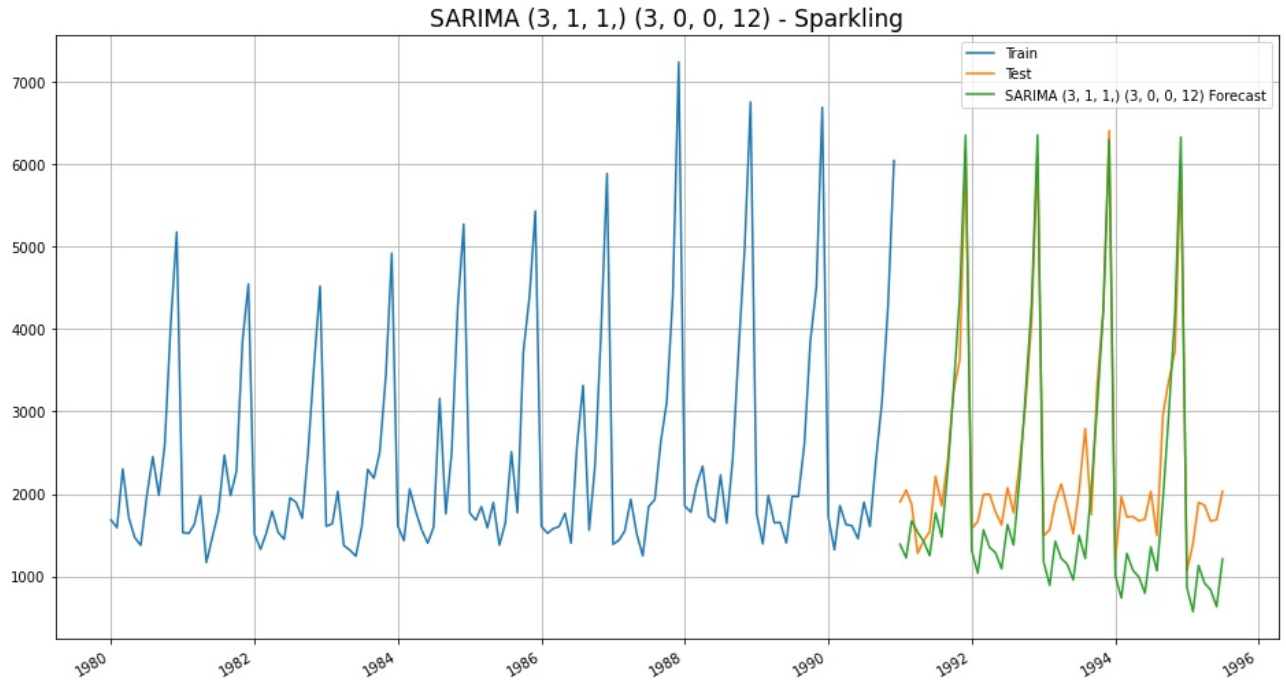


SARIMA (3, 1, 1) (3, 0, 2, 12) Diagnostic Plot - ROSE

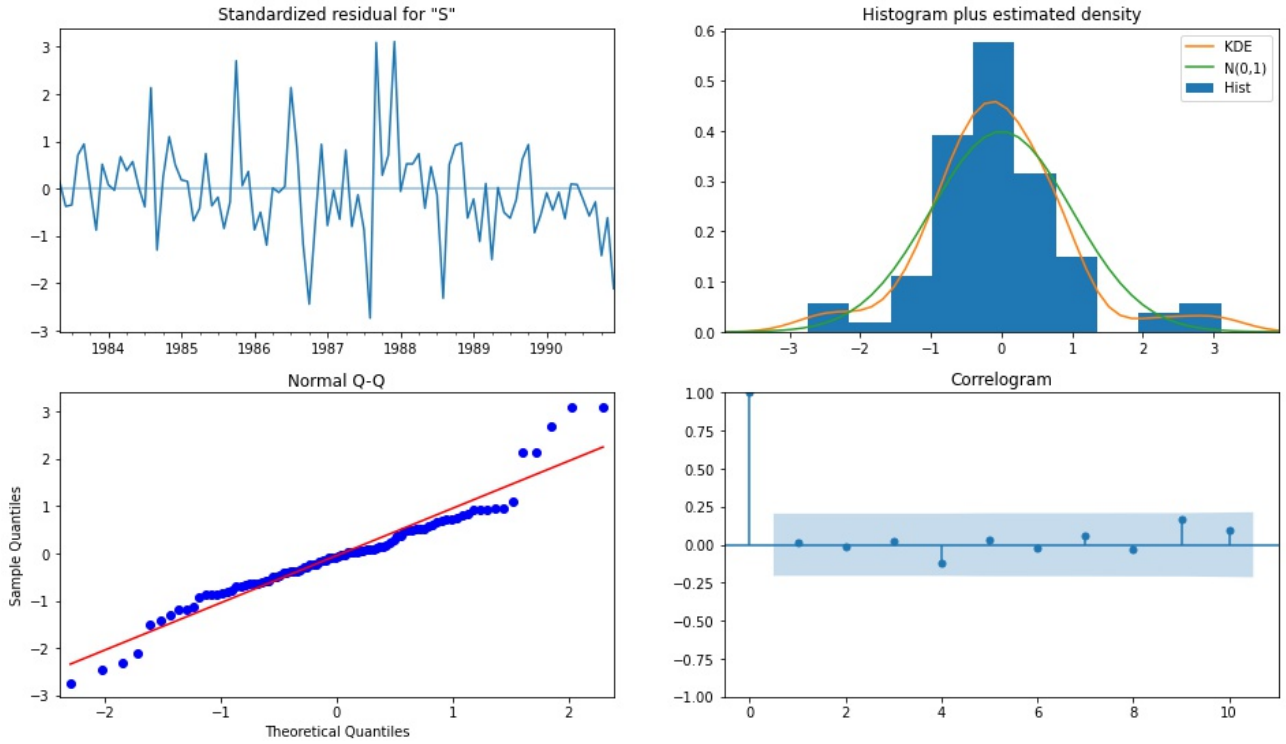


	Test RMSE Rose	**Test MAPE Rose**
ARIMA(2,1,3)	36.81	75.84
SARIMA (3, 1, 1) (3, 0, 2, 12)	18.88	36.38

- For **Sparkling**, Best Combination with **low AIC** and low Test RMSE is - **(3, 1, 1) (3, 0, 0, 12)**



SARIMA (3, 1, 1) (3, 0, 0, 12) Diagnostic Plot - SPARKLING



	Test RMSE Sparkling	**Test MAPE Sparkling**
ARIMA(2,1,2)	1299.98	47.10
SARIMA (3, 1, 1) (3, 0, 0, 12)	601.24	25.87

- **Till Now, Best Model for Rose with Least RMSE —-> SARIMA (3, 1, 1) (3, 0, 2, 12)**
- **Till Now, Best Model for Sparkling with Least RMSE ->SARIMA (3, 1, 1) (3, 0, 0, 12)**

[Q 7] Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

✦ Auto-Correlation Function (ACF) -

- Autocorrelation refers to how correlated a time series is with its past values. e.g. y_t with y_{t-1} also y_{t+1} with y_t and so on.
- 'Auto' part of Autocorrelation refers to Correlation of any time instance with its previous time instance in the SAME Time Series
- ACF is the plot used to see the correlation between the points, up to and including the lag unit

- ACF indicates the value of 'q' - which is the Moving Average parameter in ARIMA / SARIMA models

✦ Partial Auto-Correlation Function (PACF) -

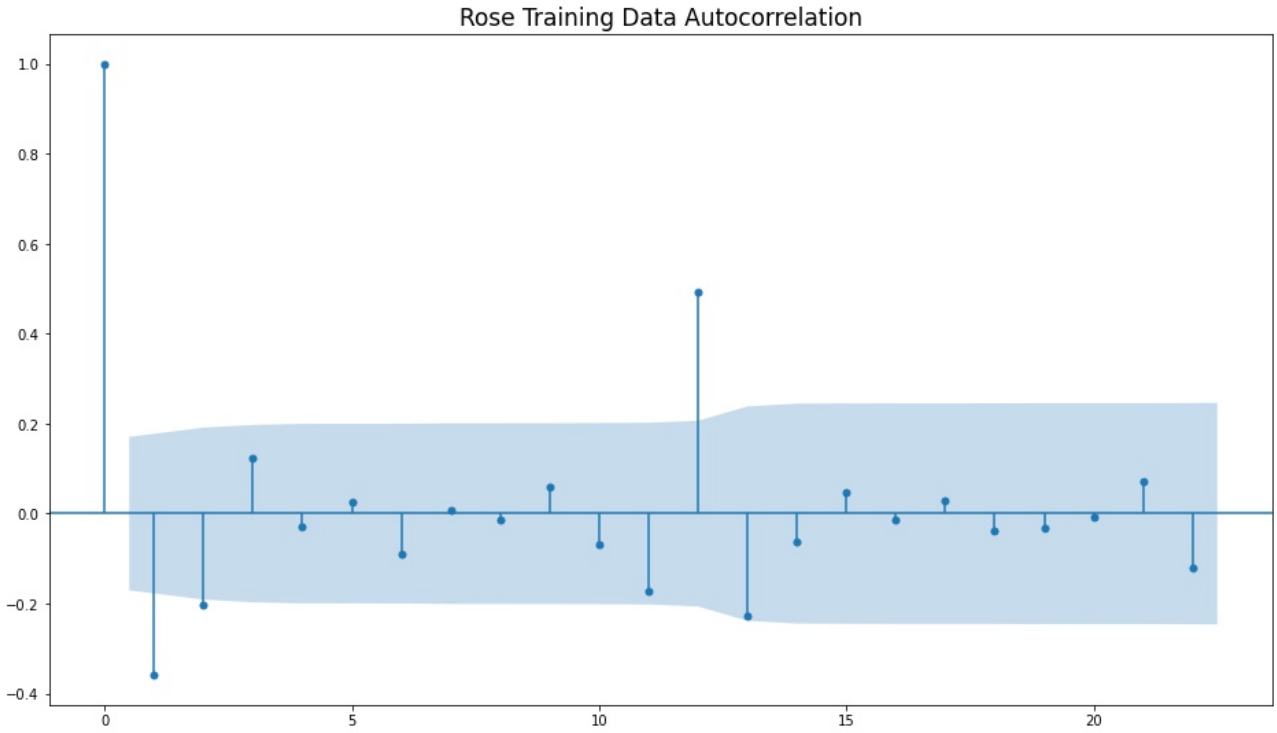
- Partial Autocorrelation refers to how correlated a time series is with its past lag values.
- For example, let lag=k, then Partial Autocorrelation is Correlation of y_t with y_{t-k}, ignoring the effects of all the instances between y_t and y_{t-k}
- PACF is the plot used to see the correlation between the lag points
- PACF indicates the value of 'p' - which is the Auto-Regressive parameter in ARIMA / SARIMA models

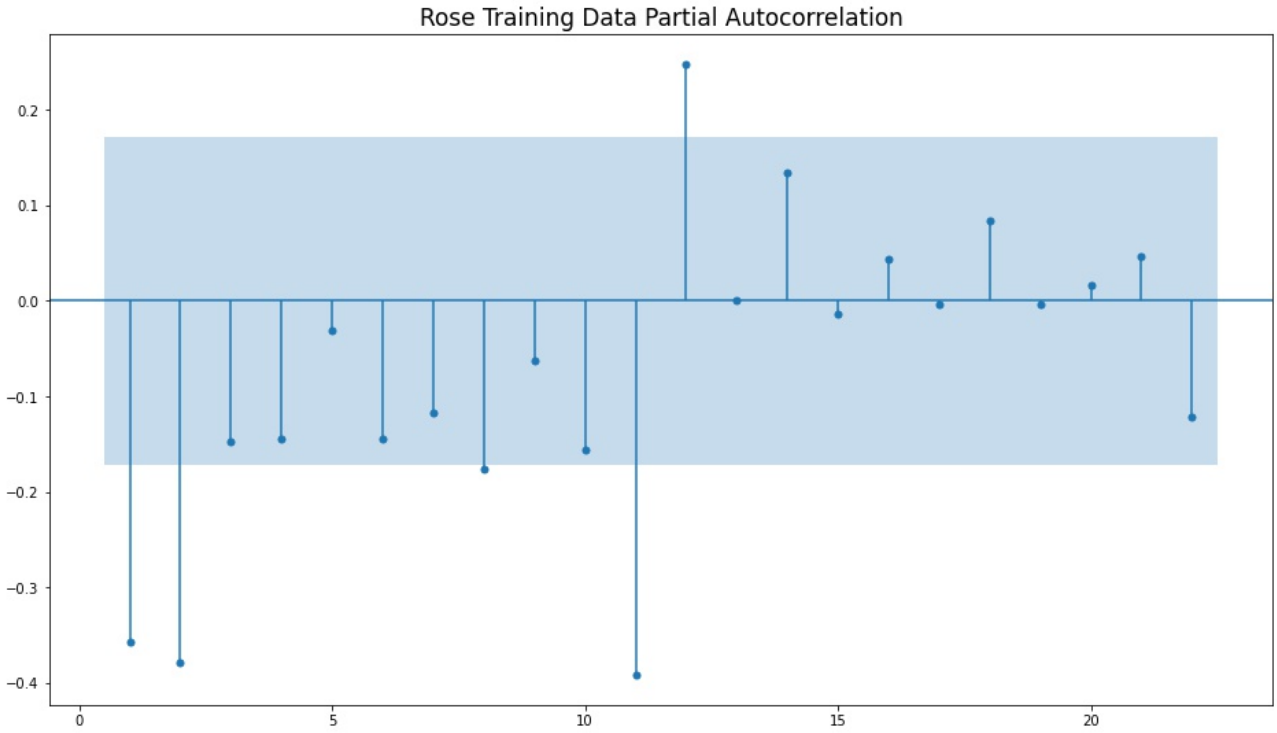
✦ ACF & PACF of Rose -

- Observing the cutoffs in ACF and PACF plots for Rose dataset, we get -

  FOR ARIMA —-> p = 2, q = 2 and difference d = 1

  FOR SARIMA —-> p = 2, q = 2, d = 1 and P = 2, D = 1, Q = 2, Seasonality=12



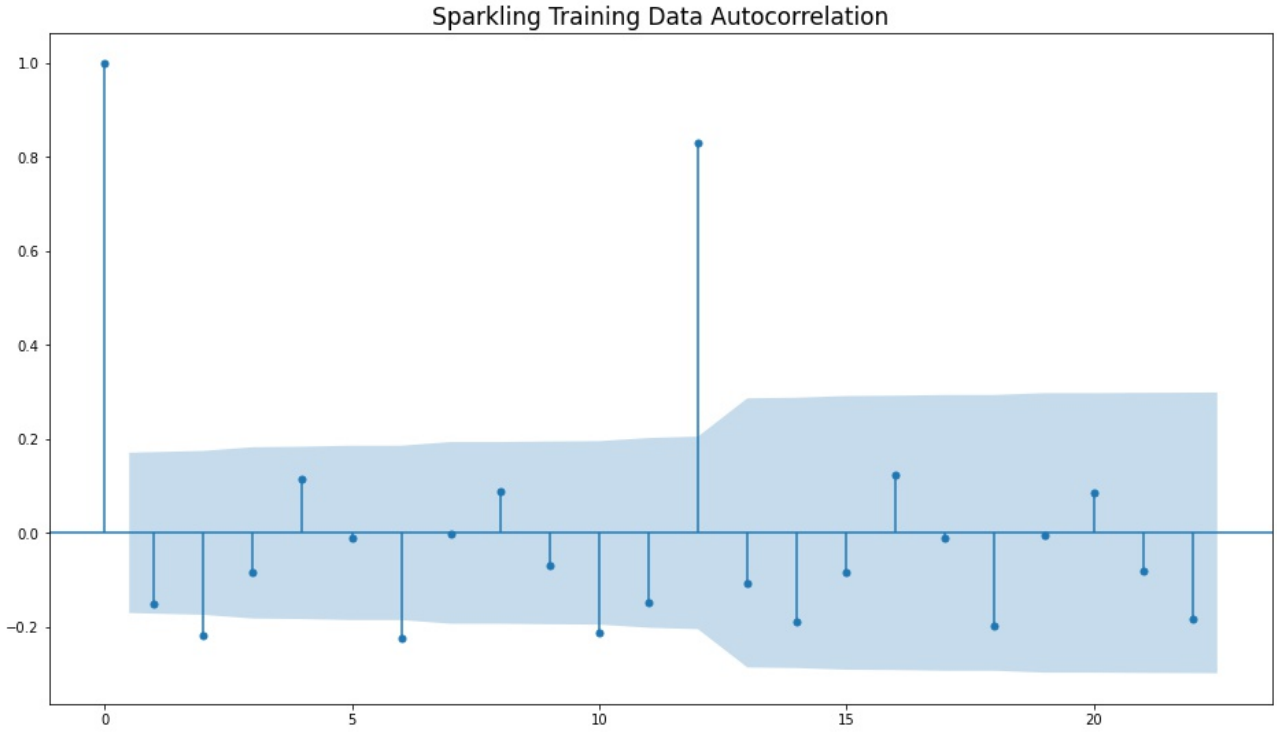


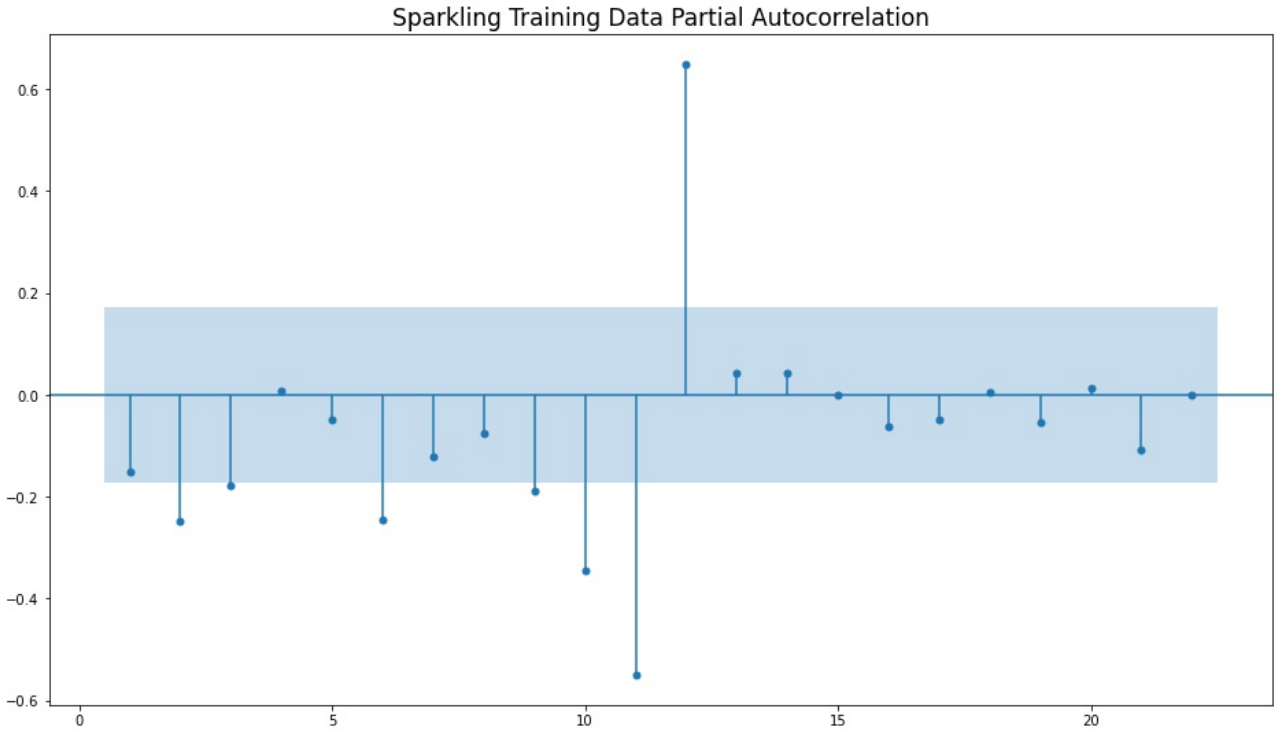
✦ ACF & PACF of Sparkling -

- Observing the cutoffs in ACF and PACF plots for Sparkling dataset, we get -

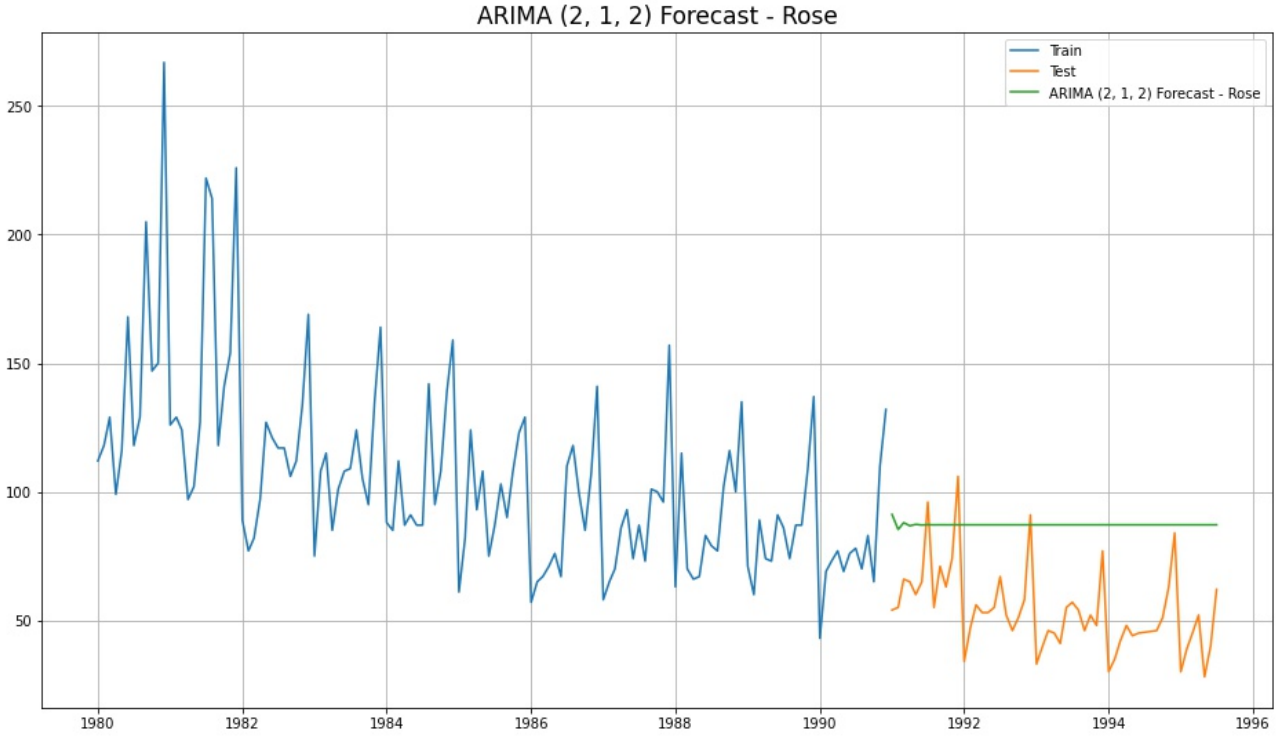
  FOR ARIMA —-> p = 0, q = 0 and difference d = 1

  FOR SARIMA —-> p = 0, q = 0, d = 1 and P = 0, 1, 2, 3 | D = 0, Q = 1, 2, 3

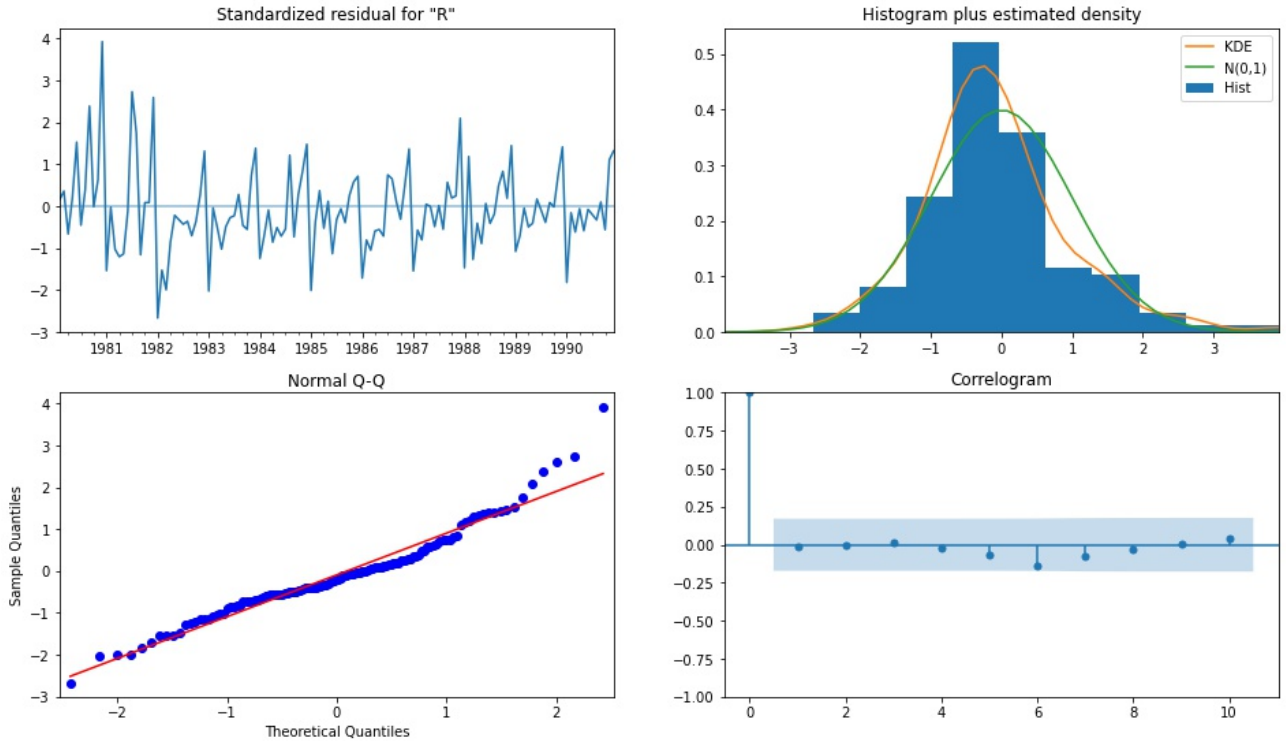




1. ARIMA Manual - Rose - (2, 1, 2)

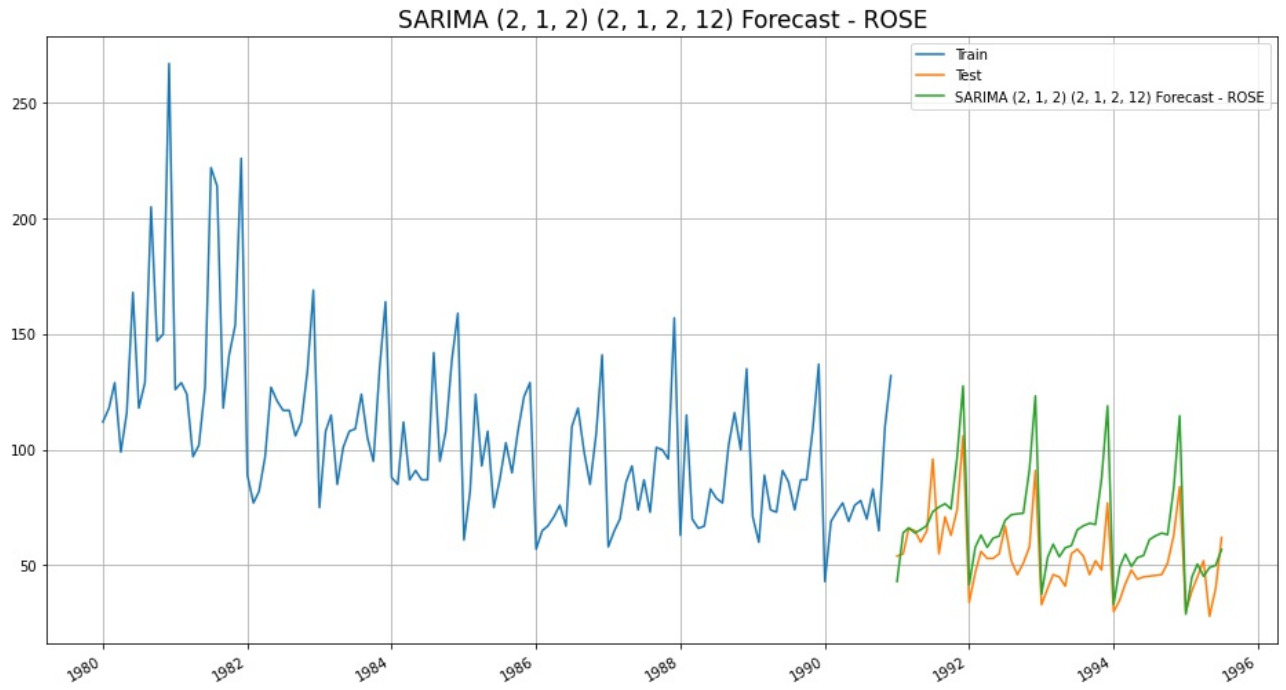


ARIMA (2, 1, 2) Diagnostic Plot - ROSE

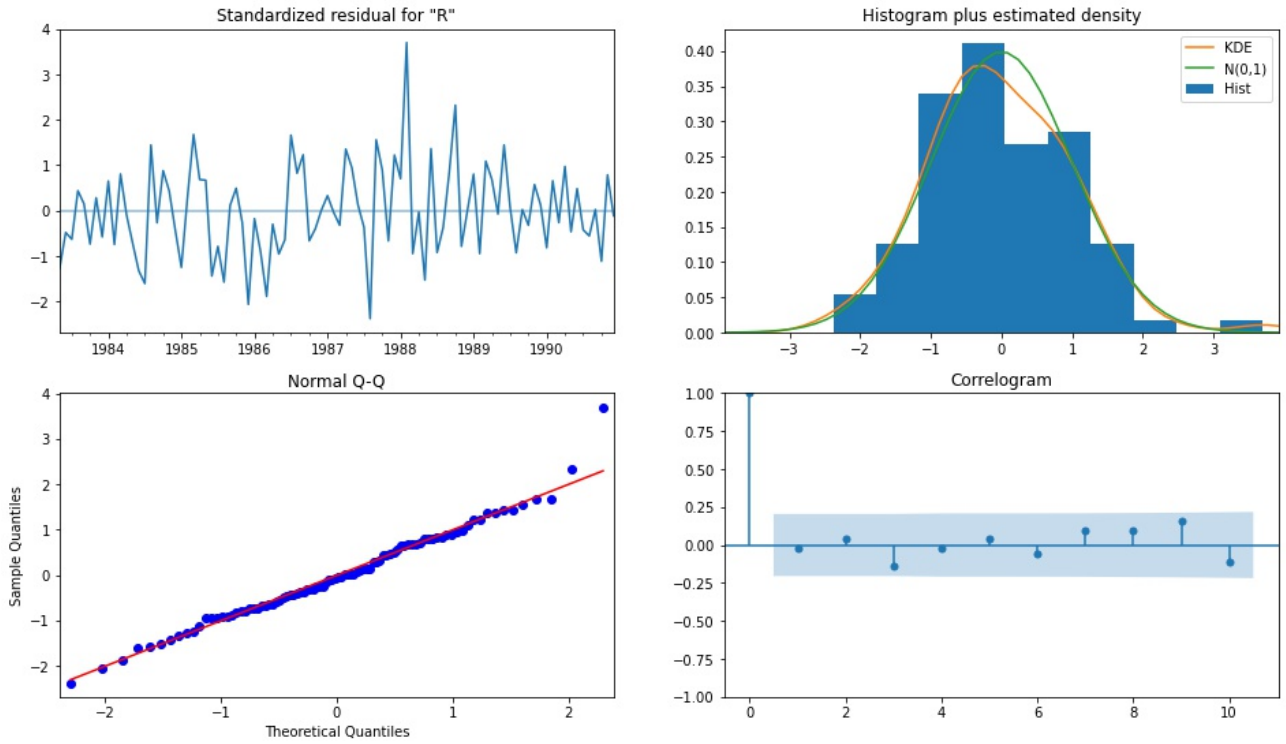


	Test RMSE Rose	**Test MAPE Rose**
ARIMA(2, 1, 2)	36.87	76.06

2. SARIMA Manual - Rose - (2, 1, 2) (2, 1, 2, 12)

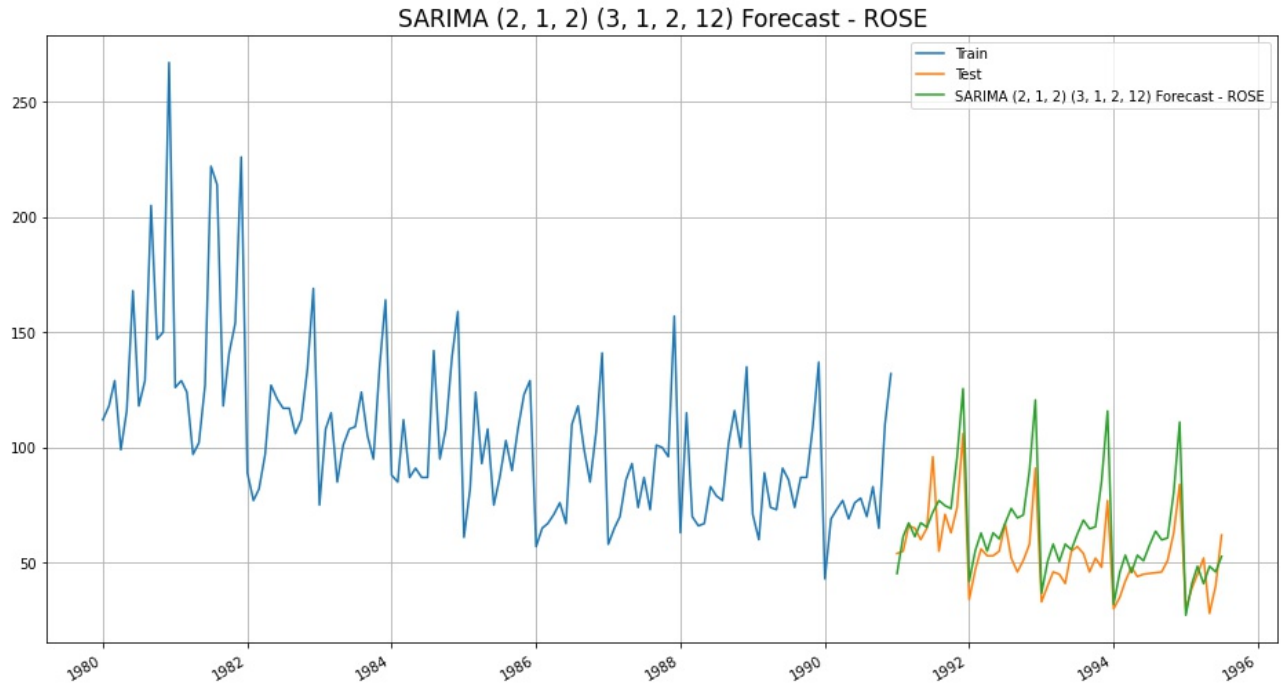


SARIMA (2, 1, 2) (2, 1, 2, 12) Diagnostic Plot - ROSE

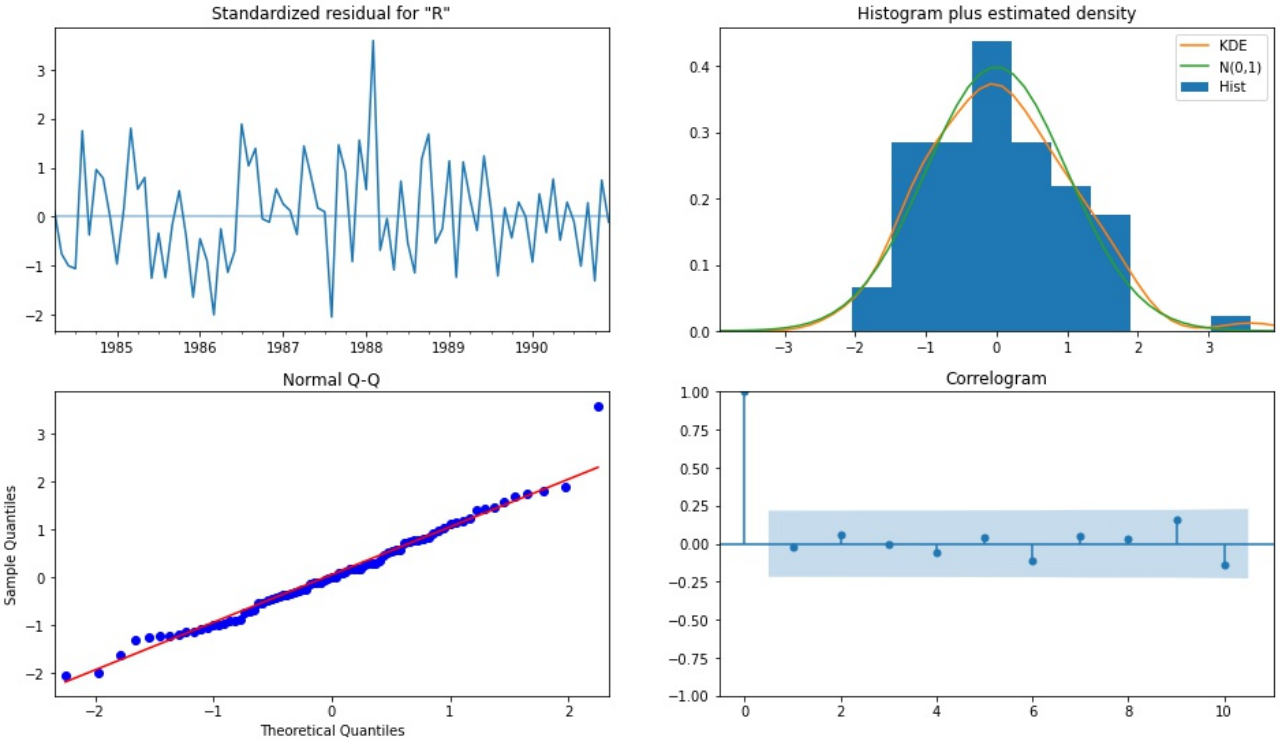


	Test RMSE Rose	**Test MAPE Rose**
ARIMA(2, 1, 2)	36.87	76.06
SARIMA (2, 1, 2) (2, 1, 2, 12)	16.55	25.48

3. SARIMA Manual - Rose - (2, 1, 2) (3, 1, 2, 12)

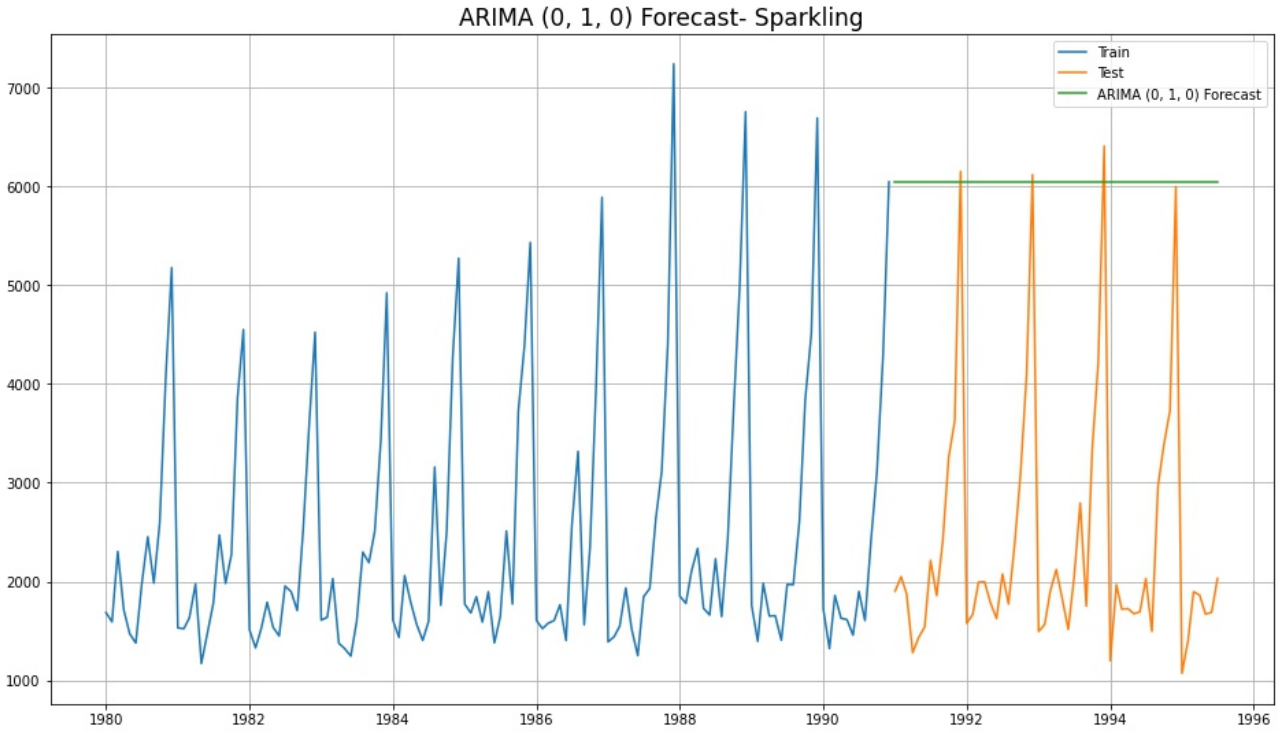


SARIMA (2, 1, 2) (3, 1, 2, 12) Diagnostic Plot - ROSE

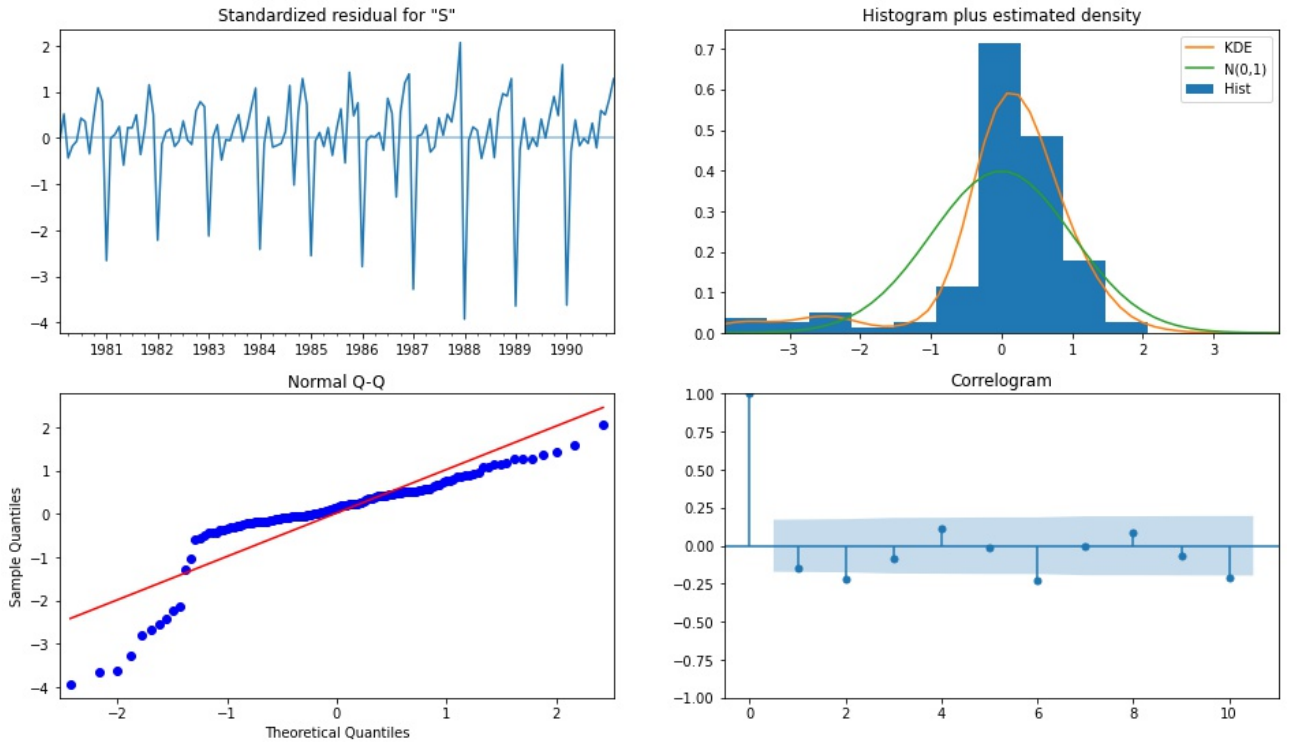


	Test RMSE Rose	**Test MAPE Rose**
ARIMA(2, 1, 2)	36.87	76.06
SARIMA (2, 1, 2) (2, 1, 2, 12)	16.55	25.48
SARIMA (2, 1, 2) (3, 1, 2, 12)	15.36	22.96

4. ARIMA Manual - Sparkling - (0, 1, 0)

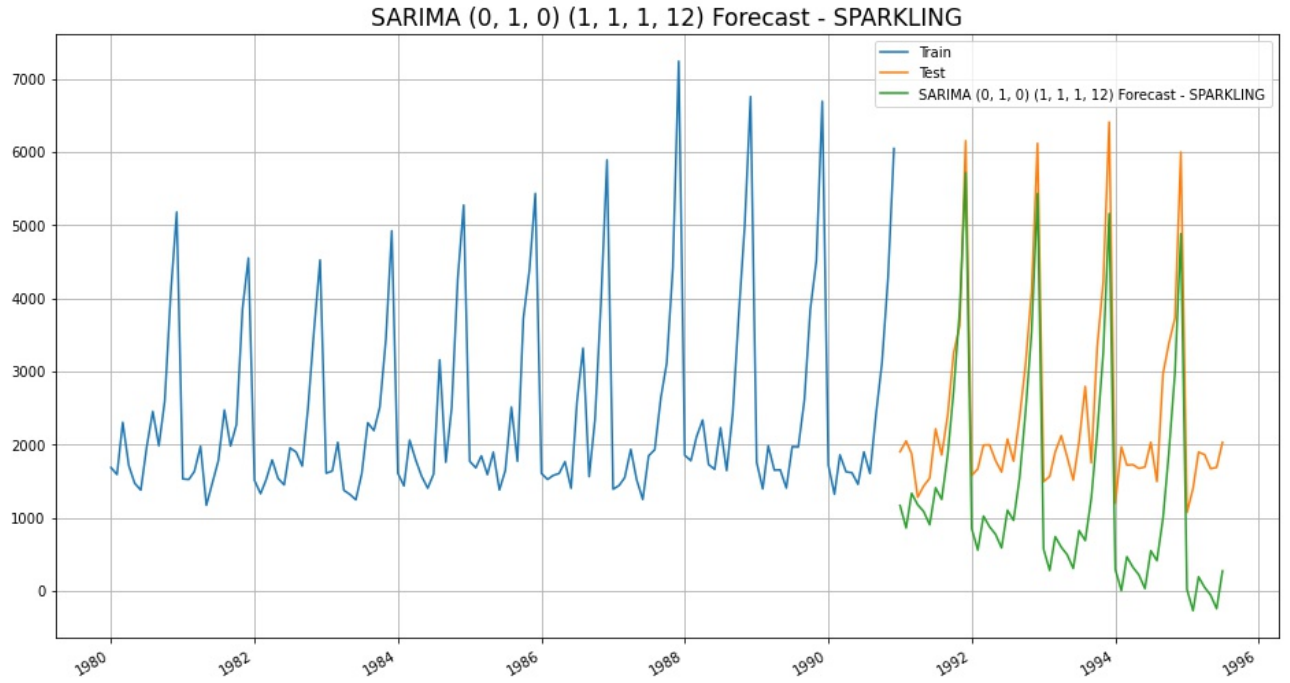


ARIMA (0, 1, 0) Diagnostic Plot - SPARKLING

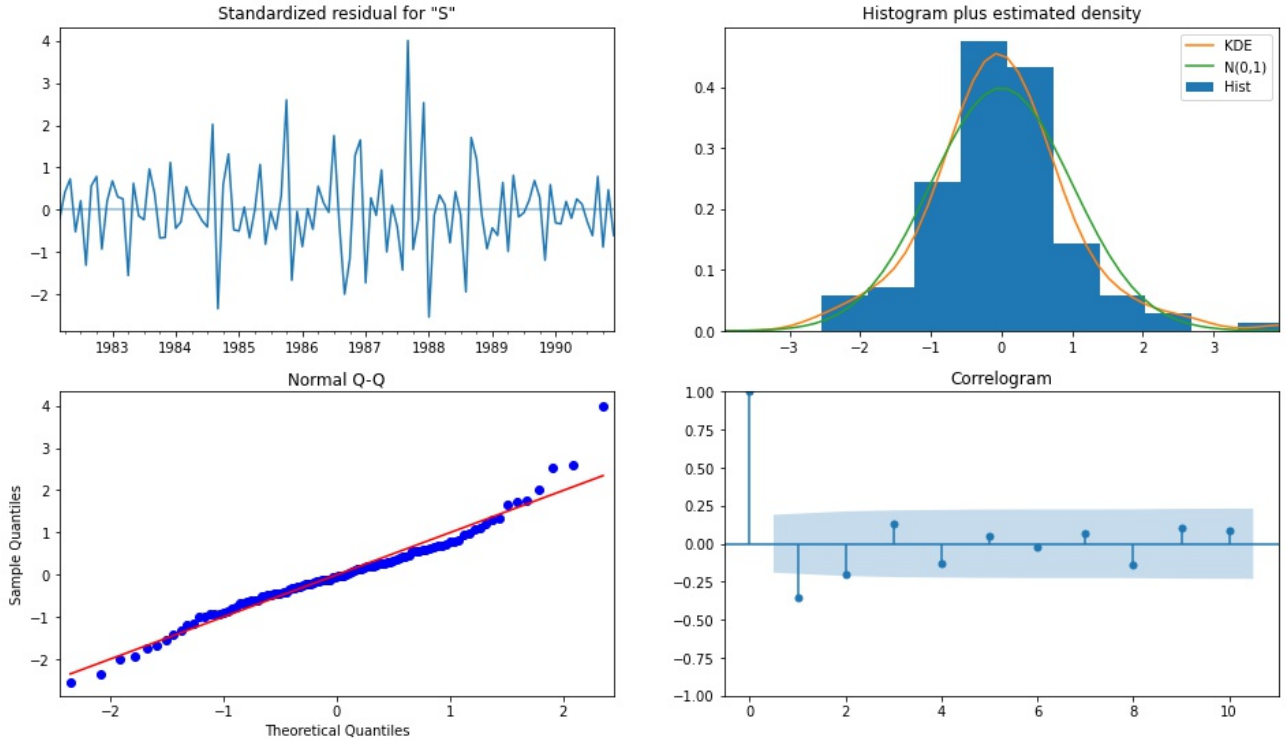


	Test RMSE Sparkling	**Test MAPE Sparkling**
ARIMA(0, 1, 0)	3864.28	201.33

5. SARIMA Manual - Sparkling - (0, 1, 0) (1, 1, 1, 12)

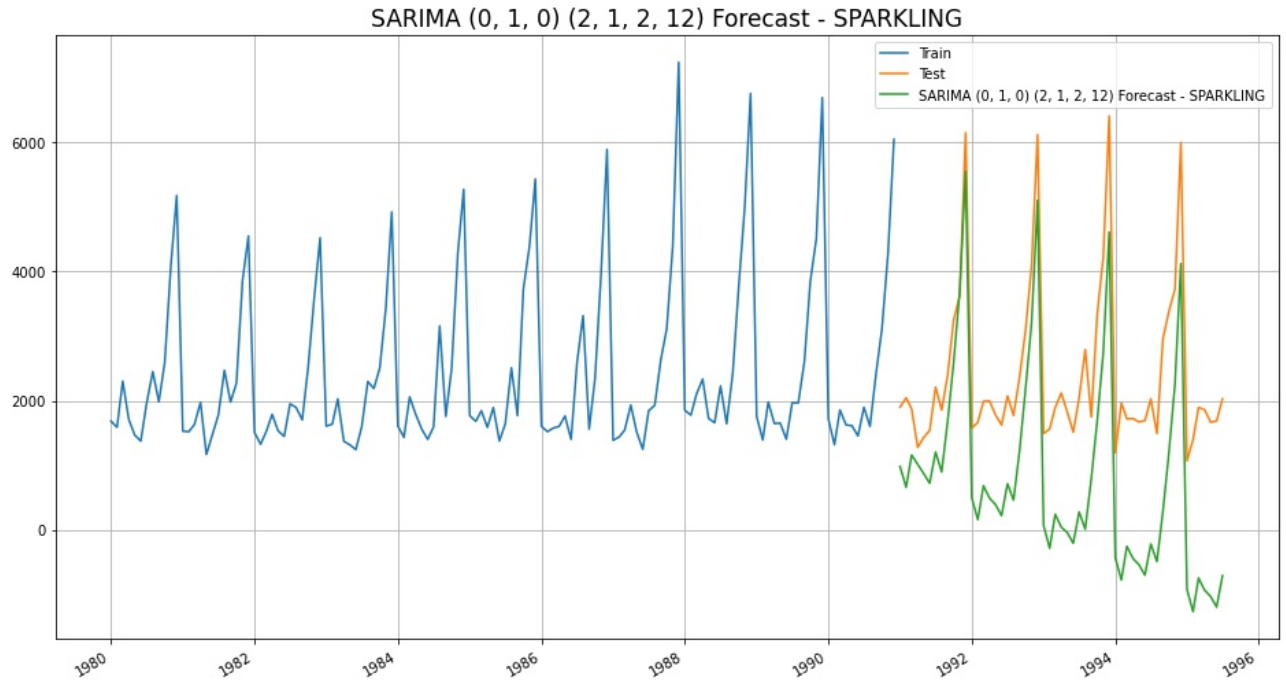


SARIMA (0, 1, 0) (1, 1, 1, 12 Diagnostic Plot - SPARKLING

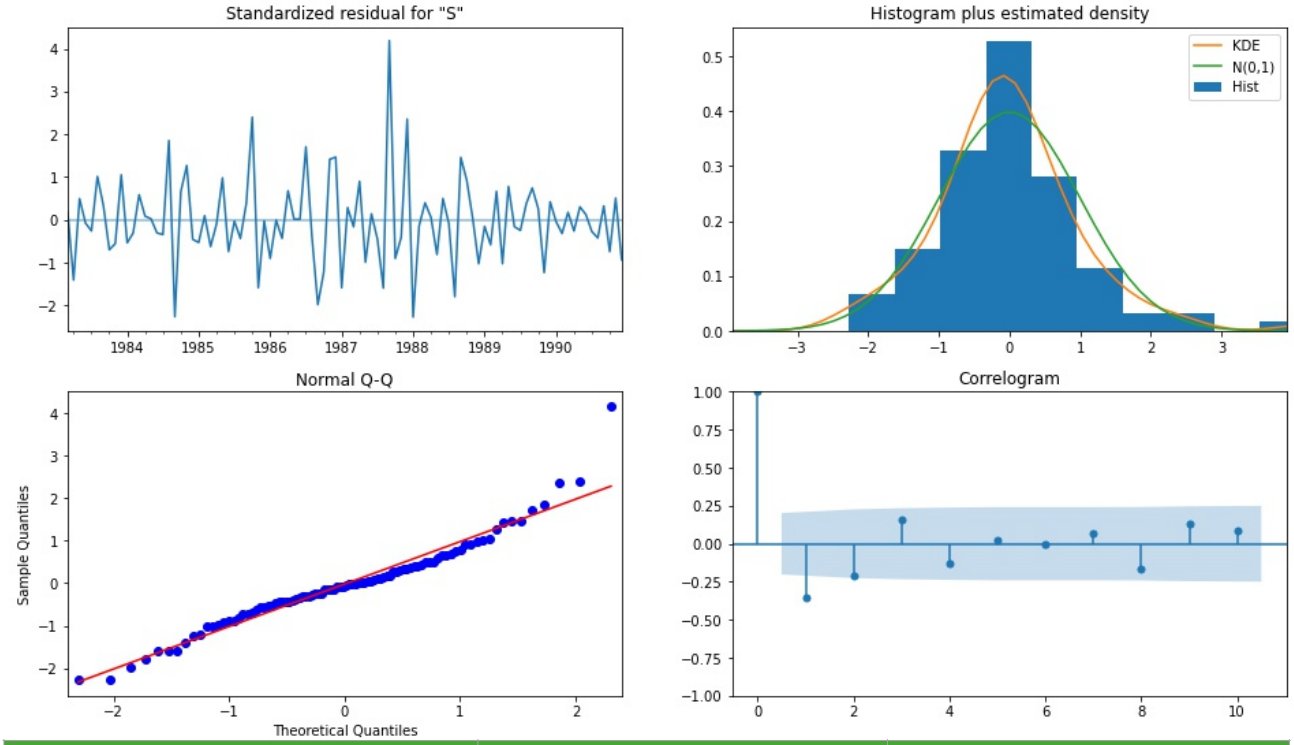


	Test RMSE Sparkling	**Test MAPE Sparkling**
ARIMA(0, 1, 0)	3864.28	201.33
SARIMA (0, 1, 0) (1, 1, 1, 12)	1189.84	54.87

6. SARIMA Manual - Sparkling - (0, 1, 0) (2, 1, 2, 12)

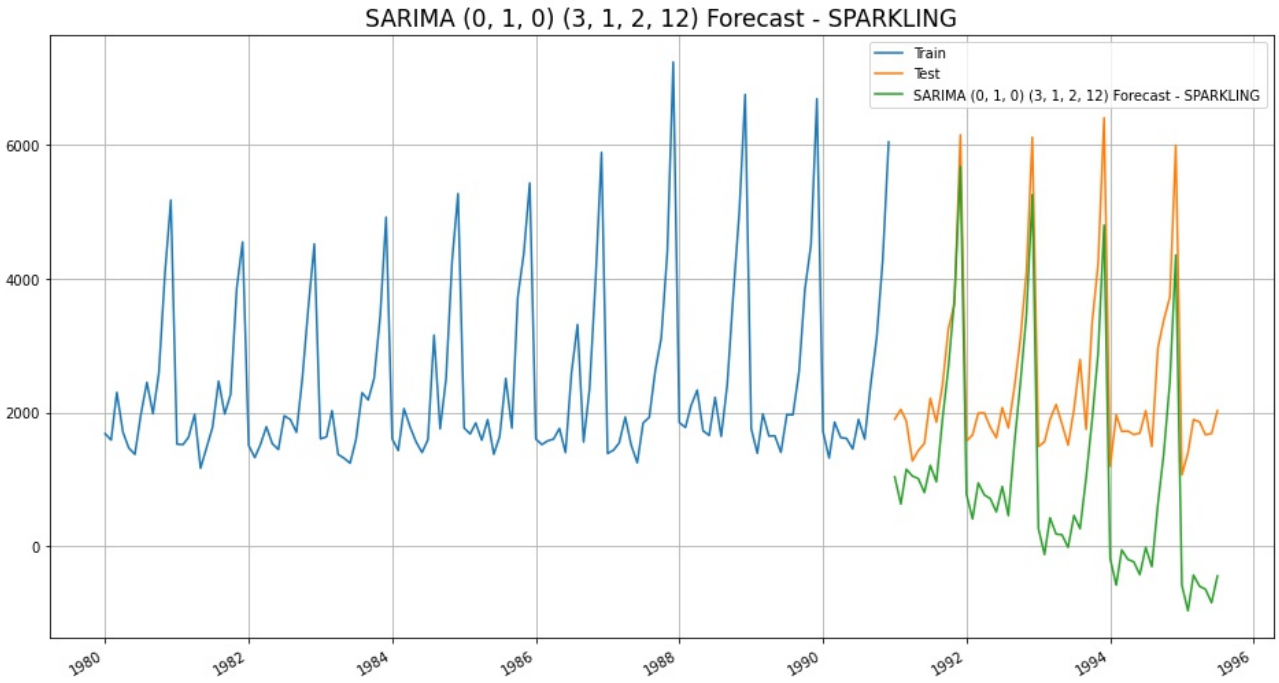


SARIMA (0, 1, 0) (2, 1, 2, 12 Diagnostic Plot - SPARKLING

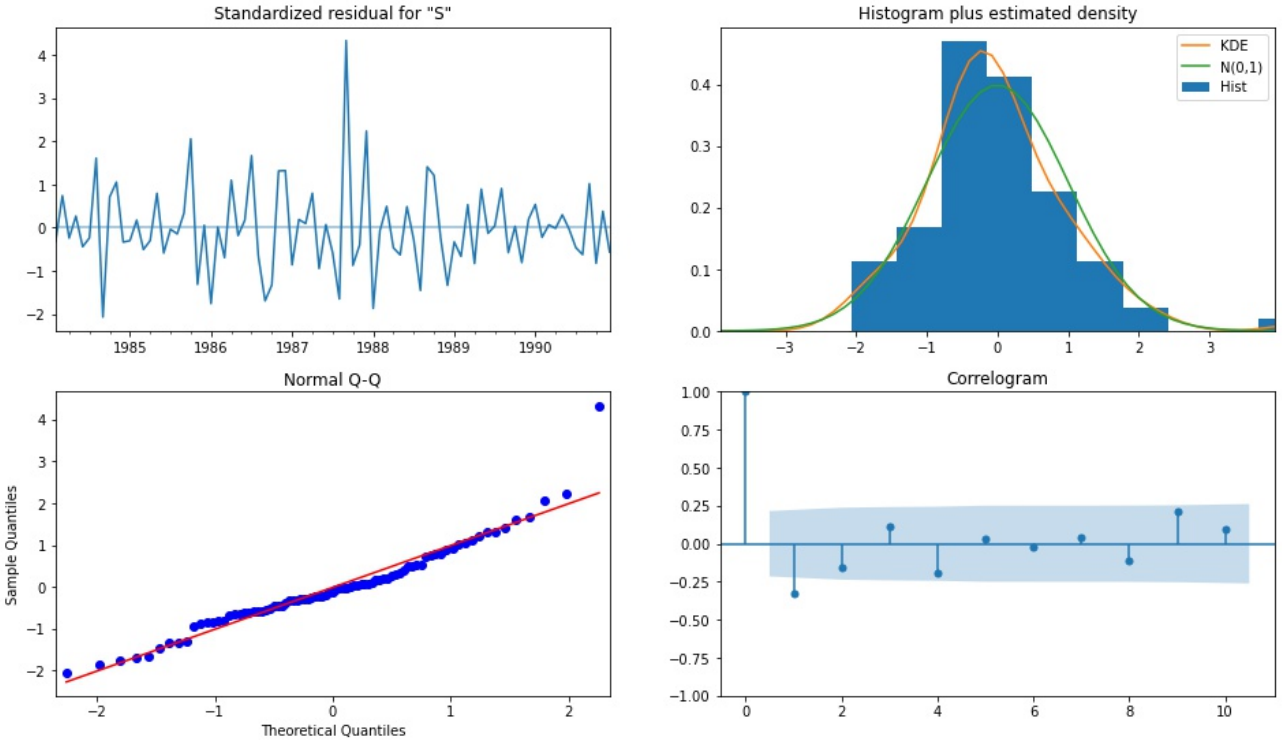


	Test RMSE Sparkling	**Test MAPE Sparkling**
ARIMA(0, 1, 0)	3864.28	201.33
SARIMA (0, 1, 0) (1, 1, 1, 12)	1189.84	54.87
SARIMA (0, 1, 0) (2, 1, 2, 12)	1757.73	81.79

7. SARIMA Manual - Sparkling - (0, 1, 0) (3, 1, 2, 12)



SARIMA (0, 1, 0) (3, 1, 2, 12 Diagnostic Plot - SPARKLING



	Test RMSE Sparkling	**Test MAPE Sparkling**
ARIMA(0, 1, 0)	3864.28	201.33
SARIMA (0, 1, 0) (1, 1, 1, 12)	1189.84	54.87
SARIMA (0, 1, 0) (2, 1, 2, 12)	1757.73	81.79
SARIMA (0, 1, 0) (3, 1, 2, 12)	1551.65	71.57

- **In all Manual methods, Best Model for Rose with Least RMSE**

  —-> SARIMA (2, 1, 2) (3, 1, 2, 12)

- **In all Manual methods, Best Model for Sparkling with Least RMSE**

  —-> SARIMA (0, 1, 0) (1, 1, 1, 12)

- Seasonal P and Q - it was difficult to gauge the correct values here as the data was not enough and cutoffs were not visible
- Hence, we tried multiple combinations of Seasonal P and Q as given above

[Q 8] 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.

✦ All Models built with ROSE (sorted by RMSE) -

Model	Parameters	Test RMSE Rose
2pointTrailingMovingAverage		11.53
Triple Exponential Smoothing (Additive Season)	$\alpha = 0.0849$ $\beta = 5.52e^{-6} \approx 0.00$ $\gamma = 0.00054$	14.24
4pointTrailingMovingAverage		14.45
6pointTrailingMovingAverage		14.57
9pointTrailingMovingAverage		14.73
RegressionOnTime		15.27
Double Exponential Smoothing	$\alpha = 1.49e^{-8} \approx 0.00$ $\beta = 5.488e^{-9} \approx 0.00$	15.27
SARIMA(2,1,2)(3,1,2,12)		15.36
SARIMA(3, 1, 1)(3, 0, 2, 12)		18.88
Triple Exponential Smoothing (Multiplicative Season)	$\alpha = 0.07736$ $\beta = 0.03936$ $\gamma = 0.00083$	19.11
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)	$\alpha = 0.05921$ $\beta = 0.0205$ $\gamma = 0.00405$	25.99
Triple Exponential Smoothing (Additive Season, Damped Trend)	$\alpha = 0.07842$ $\beta = 0.01153$ $\gamma = 0.07738$	26.04
Simple Exponential Smoothing	$\alpha = 0.09874$	36.80
ARIMA(2,1,3)		36.81
ARIMA(2,1,2)		36.87
SimpleAverageModel		53.46
NaiveModel		79.72

✦ All Models built with SPARKLING (sorted by RMSE) -

Model	Parameters	Test RMSE Sparkling
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)	$\alpha = 0.11107$ $\beta = 0.03702$ $\gamma = 0.39507$	352.45
Triple Exponential Smoothing (Additive Season)	$\alpha = 0.1112$ $\beta = 0.01236$ $\gamma = 0.46071$	378.63
Triple Exponential Smoothing (Additive Season, Damped Trend)	$\alpha = 0.10062$ $\beta = 0.00018$ $\gamma = 0.51151$	378.63
Triple Exponential Smoothing (Multiplicative Season)	$\alpha = 0.11119$ $\beta = 0.04943$ $\gamma = 0.36205$	403.71
SARIMA(3,1,1)(3,0,2,12)		601.24
2pointTrailingMovingAverage		813.40
4pointTrailingMovingAverage		1156.59
SARIMA(0,1,0)(3,1,2,12)		1189.84
SimpleAverageModel		1275.08
6pointTrailingMovingAverage		1283.93
ARIMA(2,1,2)		1299.98
Simple Exponential Smoothing	$\alpha = 0.07028$	1338.00
9pointTrailingMovingAverage		1346.28
RegressionOnTime		1389.14
SARIMA(0,1,0)(3,1,2,12)		1551.65
SARIMA(0,1,0)(2,1,2,12)		1757.73
NaiveModel		3864.28
ARIMA(0,1,0)		3864.28
Double Exponential Smoothing	$\alpha = 0.665$ $\beta = 0.0001$	5291.88

[Q 9] 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.

- Best Models as per the Least RMSE on ROSE Test set ——>

  - 2 Pt Trailing Moving Average

  - Triple Exponential Smoothing (Additive Seasonality)

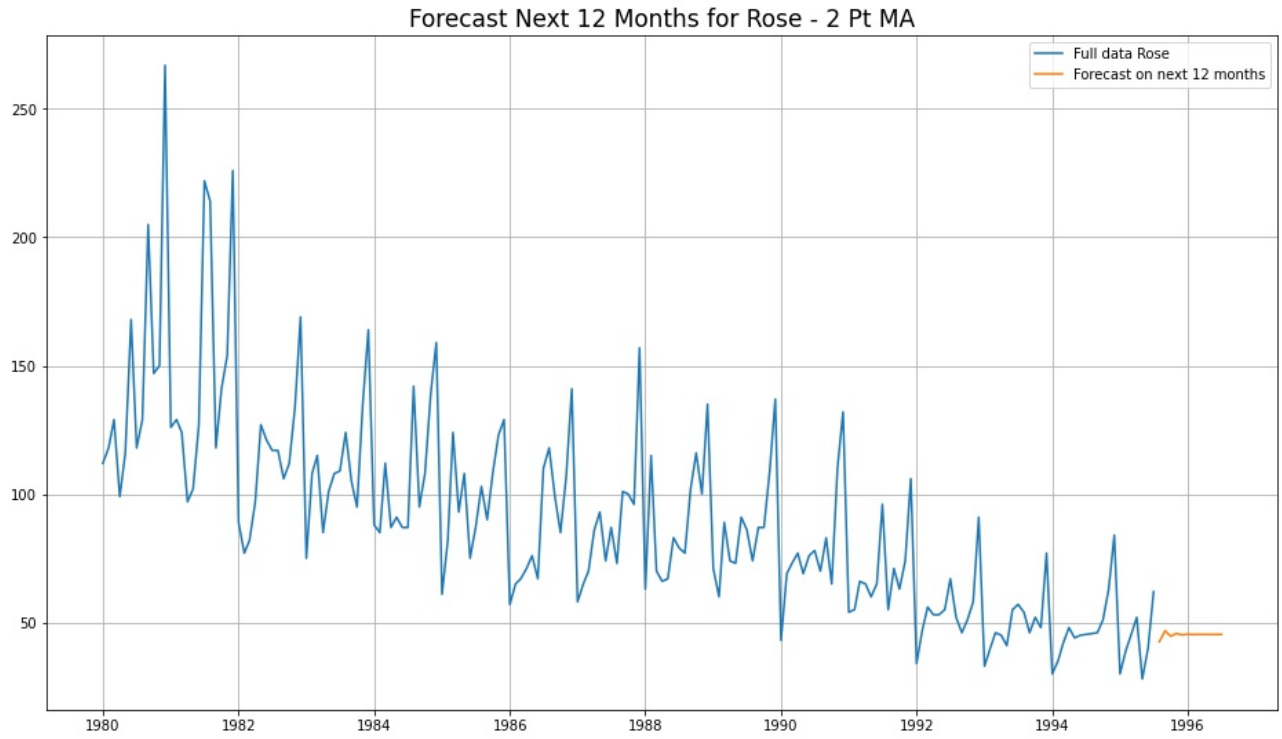
- Best Model as per the Least RMSE on SPARKLING Test set ——>

  Triple Exponential Smoothing

  (Multiplicative Season, Damped Trend)

  $\alpha = 0.11107, \quad \beta = 0.03702, \quad \gamma = 0.39507$

✦ <u>Rose Forecast Next 12 months - 2 Pt Moving Average</u>

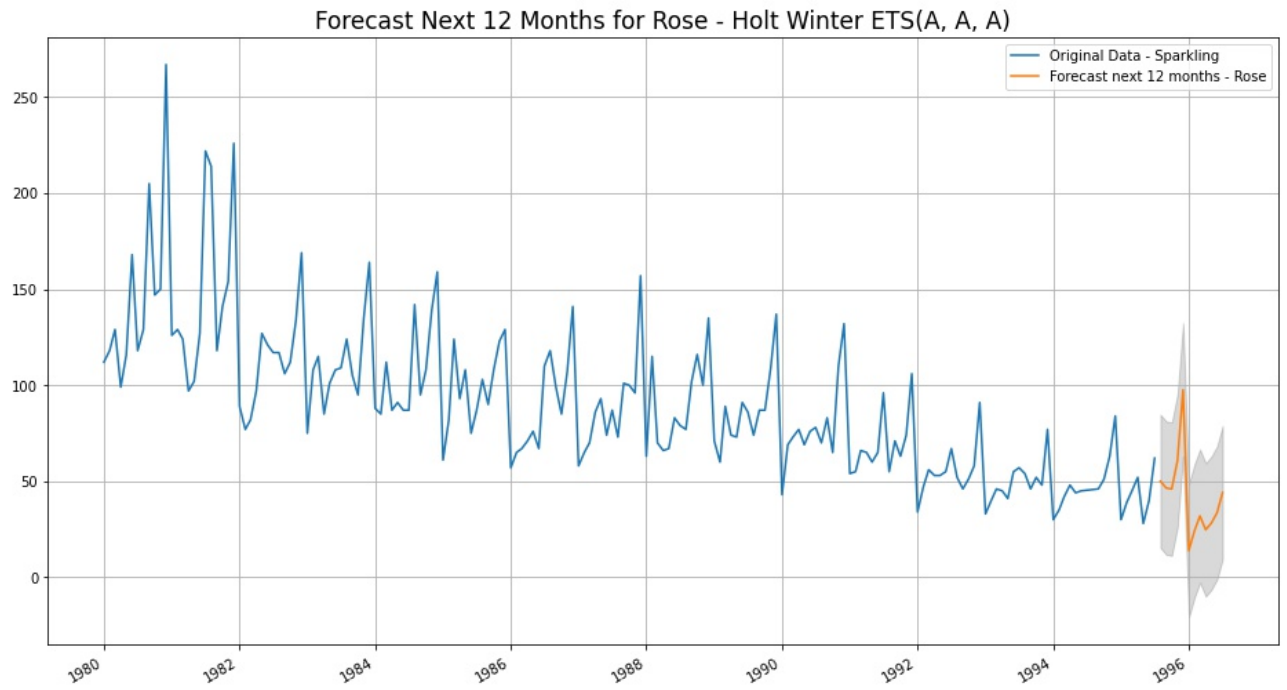


- This model doesn't seem to be predicting very well
- Hence, forecasting on the second best model - Triple Exponential Smoothing ETS(A, A, A) - Additive Seasonality

	Forecast
1995-08-01	42.50
1995-09-01	46.75
1995-10-01	44.63
1995-11-01	45.69
1995-12-01	45.16
1996-01-01	45.42
1996-02-01	45.29
1996-03-01	45.36
1996-04-01	45.32
1996-05-01	45.34
1996-06-01	45.33
1996-07-01	45.33

Rose Forecast Values - Next 12 Months - 2 Pt Moving Average

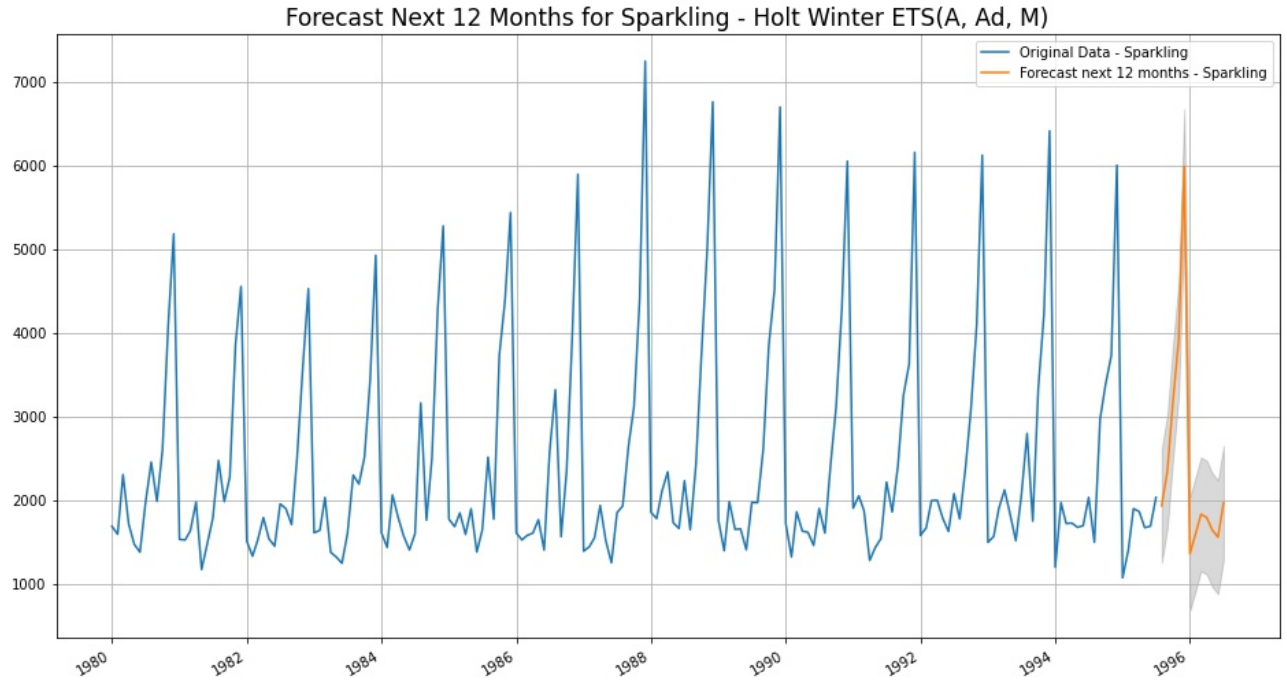
✦ Rose Forecast Next 12 months - Triple Exponential Smoothing ETS (A, A, A)



	Forecast
1995-08-01	50.02
1995-09-01	46.50
1995-10-01	45.94
1995-11-01	60.72
1995-12-01	97.66
1996-01-01	13.89
1996-02-01	24.29
1996-03-01	31.91
1996-04-01	24.73
1996-05-01	28.12
1996-06-01	33.56
1996-07-01	44.09

Rose Forecast Values - Next 12 Months - TES - ETS(A, A, A)

✦ **Sparkling Forecast Next 12 months - Triple Exponential Smoothing ETS (A, Ad, M) - Damped Trend, Multiplicative Seasonality**



	Forecast
1995-08-01	1931.44
1995-09-01	2351.98
1995-10-01	3179.46
1995-11-01	3918.07
1995-12-01	5985.90
1996-01-01	1357.57
1996-02-01	1599.15
1996-03-01	1830.31
1996-04-01	1791.02
1996-05-01	1641.84
1996-06-01	1556.37
1996-07-01	1966.00

Sparkling Forecast Values - Next 12 Months -
TES - ETS(A, Ad, M)

[Q 10] Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

✦ **Rose Wine Sales - Comments :**

- Rose wine shows a clear trend of declining sales since 1980

  - This shows decline in popularity of this variant of wine

- Also, there is a clear spike in sales seen in the last quarter of every year from Oct to Dec

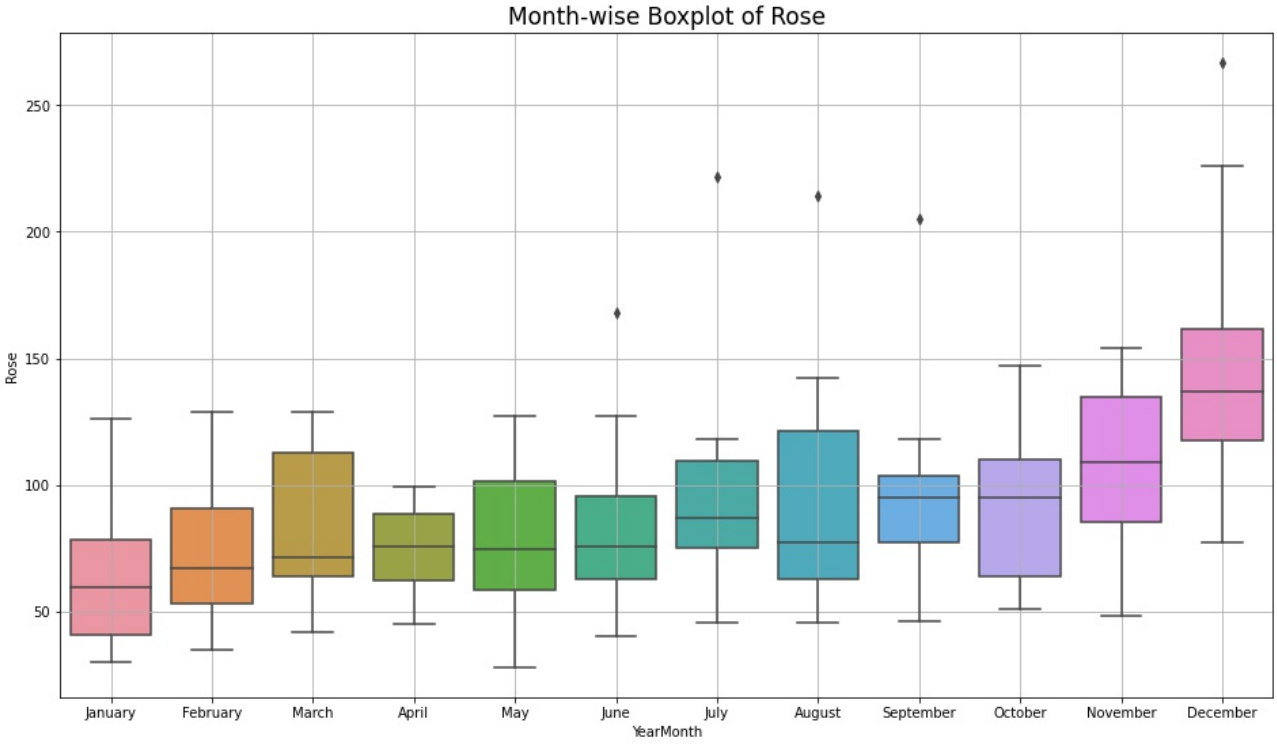
  - This might be due to the Holiday season in this period

  - Highest peak in sales is seen in Dec every year

- There is also an instant crashing slump in sales in the first quarter of every year from Jan

  - This might be due to the after effect or hangover of Holidays

- Sales slowly pick up only after May-June



✦ Rose Wine Sales - Forecast Models :

- Top 2 best models as per lowest Test RMSE were found to be - 2 Pt Moving Average and Holt-Winters - Additive Seasonality & Trend
- 2 Pt Moving Average model, when used for forecasting do not seem to give good predictions. Forecast values level out after a few iterations
- Holt-Winters seems to give a consistent forecast with respect to the data
- Hence, for **final forecast of Rose Wine Sales - we choose Holt-Winters**

✦ Rose Wine Sales - Suggestions :

- Firstly, Holiday season is around the corner and forecast shows increasing sales and sharp peak in Dec. Hence, Company should stock up
- But Declining sales of Rose Wine over the long period should be investigated with more data crunching
- Company can rebrand its Rose variant along-with a new Wine-master

- Company should take advantage of the oncoming spike from Aug-Oct by introducing aggressive offers and Ad campaigns.
- This will entice first time Wine drinkers and fence sitters (who don't have specific loyalties to any particular brand)
- Still if there is no significant upward trend in sales by this Dec, then Company has 2 options - invest in R&D or think of discontinuing this variant and come up with something completely new

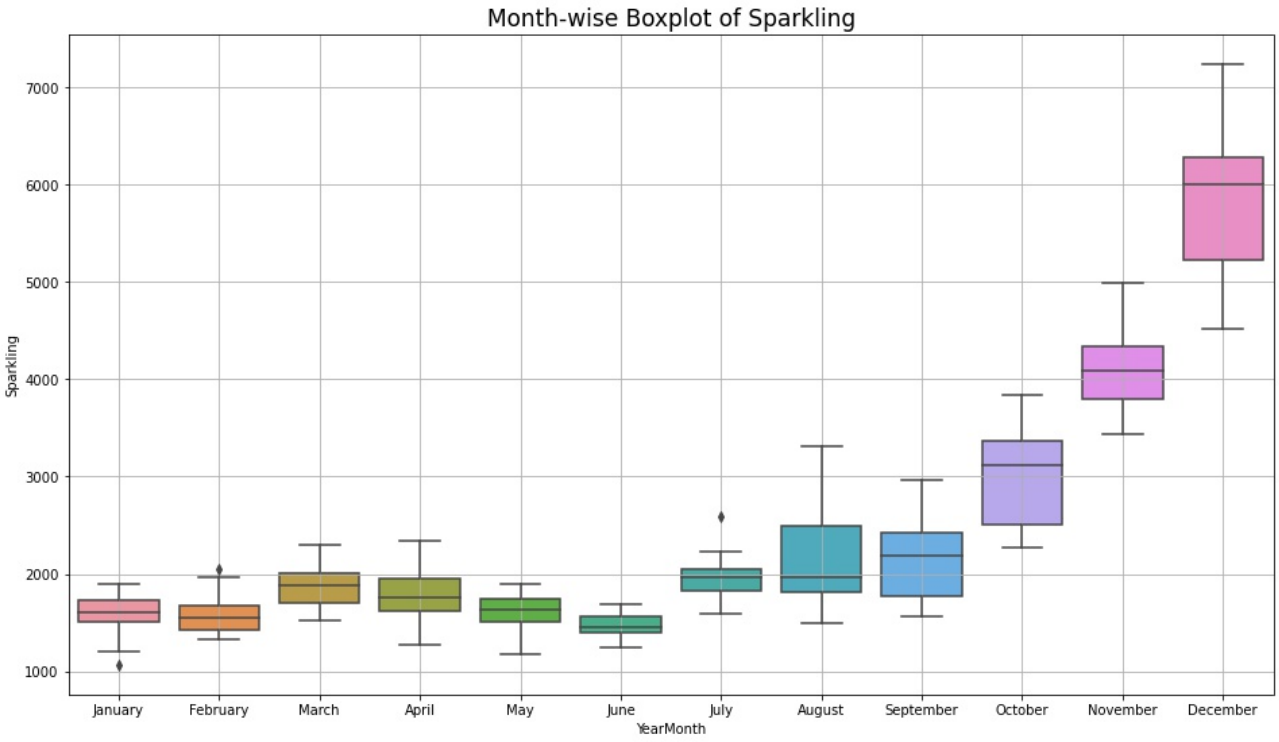
✦ Sparkling Wine Sales - Comments :

- Sparkling wine sales don't show any upward or downward trend

  - This shows flat sales over long term range

- Also, there is very high spike in sales seen in the last quarter of every year from Oct to Dec

  - This might be due to the Holiday season in this period
  - Highest peak in sales is seen in Dec every year
  - Dec sales are almost 3 times of Sep sales



- Similar to Rose Wine, even in Sparkling sales, an instant crashing slump is seen in the first quarter of every year from Jan

  - This might be due to the after effect or hangover of Holidays

- Sales slowly pick up only from Jul-Aug

✦ Sparkling Wine Sales - Forecast Models :

- Triple Exponential Smoothing - Holt-Winters Models perform the best on Sparkling datasets, considering the least RMSE on Test data
- There has been incremental improvements in Test RMSE with each tuning of parameters
- Finally, **for forecast of Sparkling Wine Sales - we choose Holt-Winters with Multiplicative Seasonality and Additive Damped Trend**

✦ Sparkling Wine Sales - Suggestions :

- Even for Sparkling, Holiday season is around the corner and forecast shows increasing sales and sharp peak in Dec. Hence, Company should stock up
- Sparkling wine has great holiday sales, so this shows popularity.
- So no need to introduce any offers here but hammering Ads are suggested in these times of Oct-Dec. This will drive sales even further.
- Sparkling wines are generally associated with celebrations and mainly to burst open.
- A special designer bottle can be introduced at a cheaper price just for bursting. This will maximise profits
- Year on Year sales do not show any significant increase or decrease
- Though, Holiday spikes are extreme, but general Year on Year sales need to be investigated more. Early period from Jan should be used to do this deep dive

——————— END OF PROJECT ———————