7/31/2022

Project –Time Series Forecasting

VIBHAV JAISWAL

CONTENTS

Problem: 3

1. Read the data as an appropriate Time Series data and plot the data. 3

2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. 5

3. Split the data into training and test. The test data should start in 1991. 15

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. 21

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 40

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. 42

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. 49

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. 61

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. 61

10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. 64

PROBLEM:

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.

Data set for the Problem: Sparkling.csv and Rose.csv

Please do perform the following questions on each of these two data sets separately.

1. READ THE DATA AS AN APPROPRIATE TIME SERIES DATA AND PLOT THE DATA.

Solution:

• Both Datasets are read and stored as Pandas Data Frames for analysis

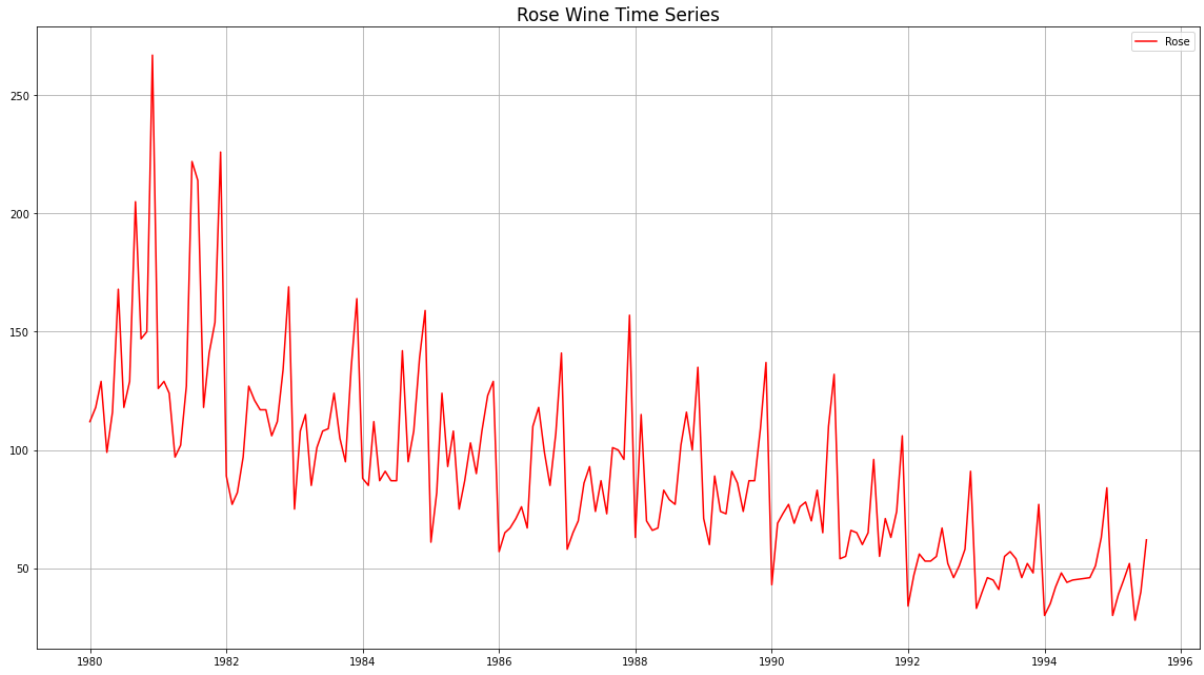
• Datasets are read as Time Series data using parse_dates=True & index_col='YearMonth'

• First 5 rows of both the data are given below

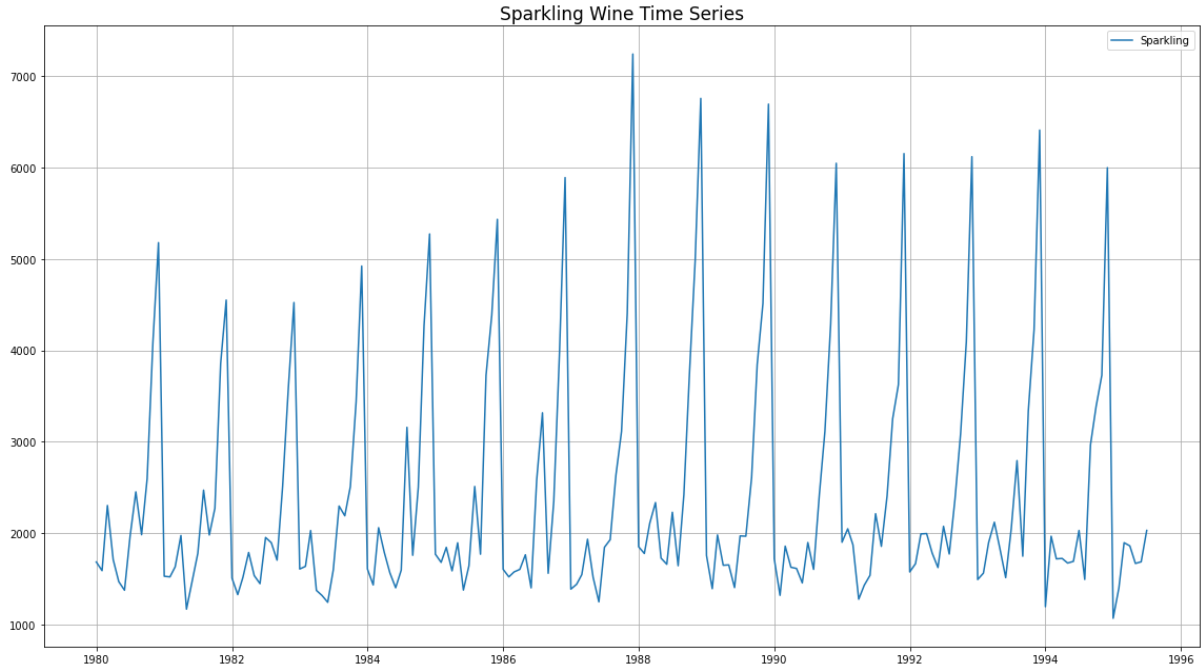
YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

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

• Rose Data plot -



• Sparkling Data plot –



2. PERFORM APPROPRIATE EXPLORATORY DATA ANALYSIS TO UNDERSTAND THE DATA AND ALSO PERFORM DECOMPOSITION.

Solution:

Exploratory Data Analysis-

Descriptive Stats of Rose and Sparkling datasets

	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

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417	1295.112	1070.0	1605.0	1874.0	2549.0	7242.0

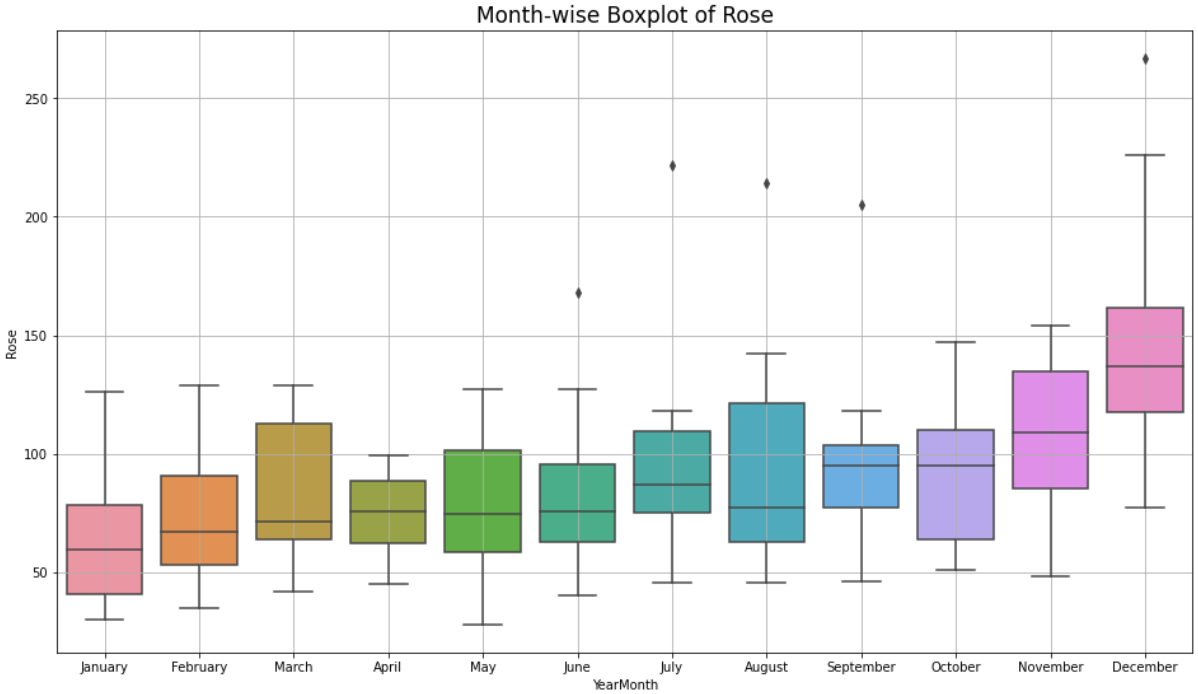
Info - Rose data

```
<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   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

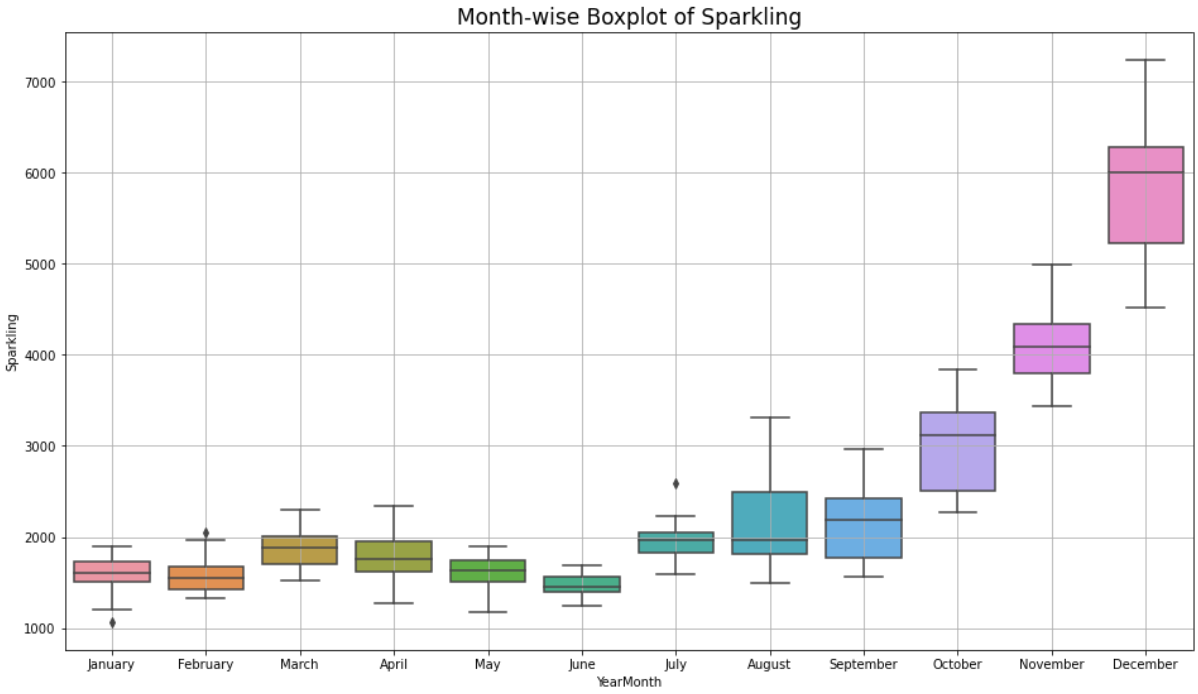
Info - Sparkling data

```
<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
```

• **Month-wise Boxplot of Rose –**



• **Month-wise Boxplot of Sparkling –**

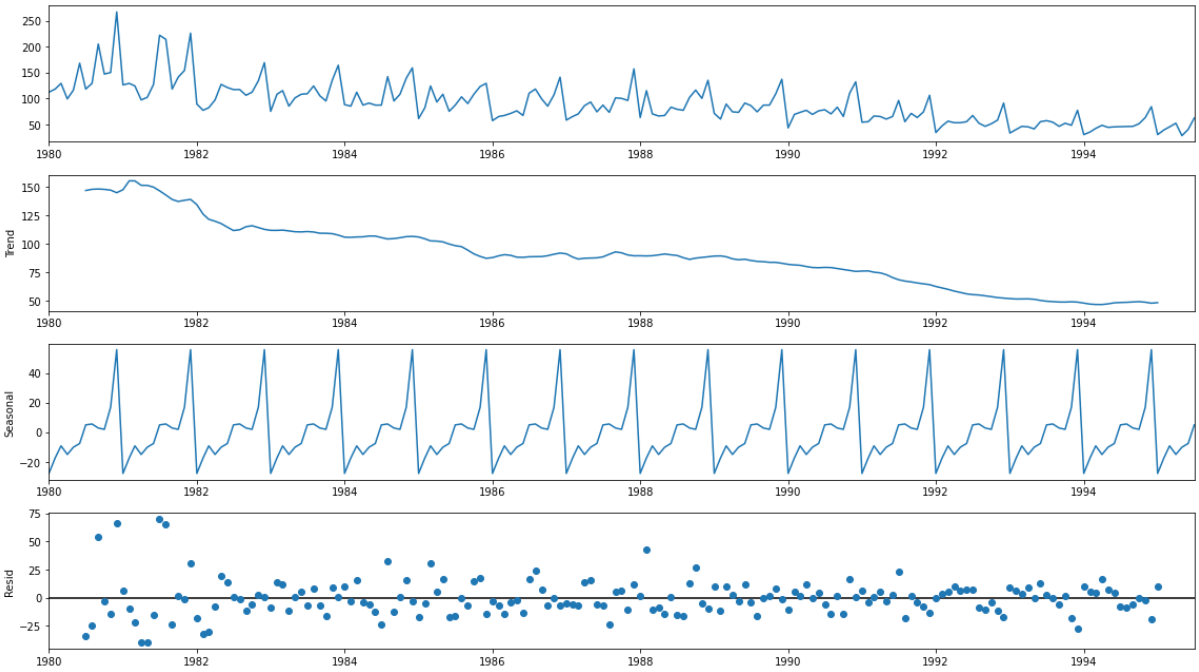


• Sales of both - Rose and Sparkling, show a spike in the last quarter of Oct to Dec

• Spike is much more accentuated in Sparkling sales

• This spike may be due to the Holiday season starting in Oct

Additive Decomposition of Rose –



Rose Trend:

```
Trend
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01    147.083333
1980-08-01    148.125000
1980-09-01    148.375000
1980-10-01    148.083333
1980-11-01    147.416667
1980-12-01    145.125000
Name: trend, dtype: float64
```

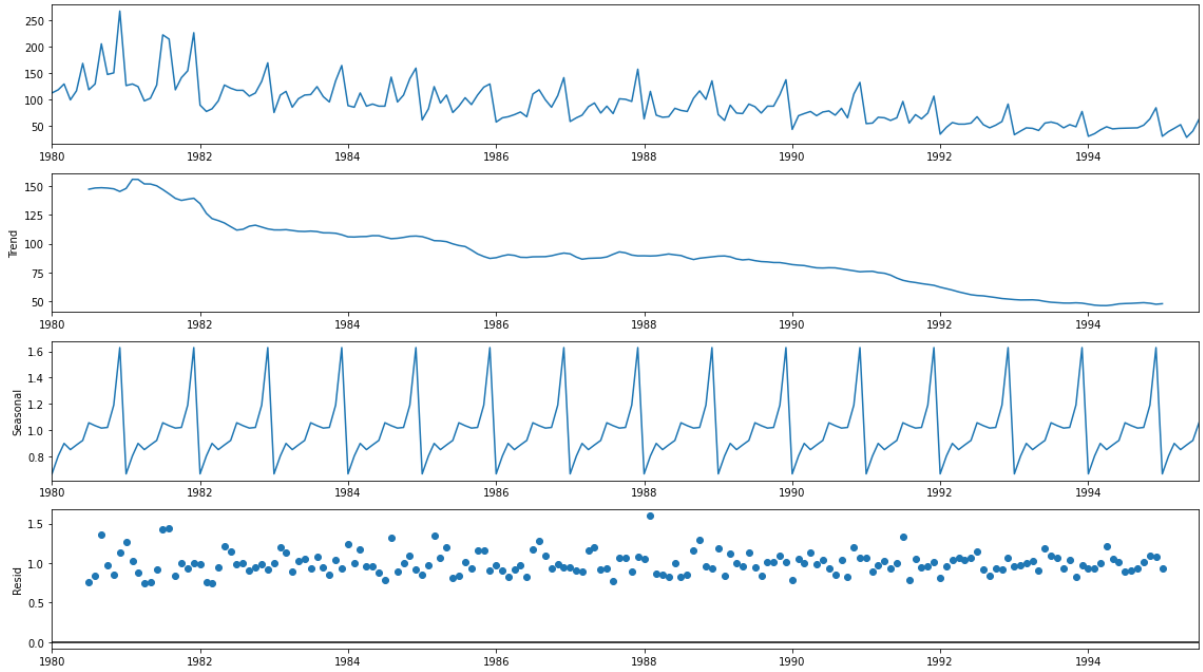
Rose Seasonality:

```
Seasonality
 YearMonth
1980-01-01   -27.908647
1980-02-01   -17.435632
1980-03-01    -9.285830
1980-04-01   -15.098330
1980-05-01   -10.196544
1980-06-01    -7.678687
1980-07-01     4.896908
1980-08-01     5.499686
1980-09-01     2.774686
1980-10-01     1.871908
1980-11-01    16.846908
1980-12-01    55.713575
Name: seasonal, dtype: float64
```

Rose Residual:

```
Residual
 YearMonth
1980-01-01          NaN
1980-02-01          NaN
1980-03-01          NaN
1980-04-01          NaN
1980-05-01          NaN
1980-06-01          NaN
1980-07-01   -33.980241
1980-08-01   -24.624686
1980-09-01    53.850314
1980-10-01    -2.955241
1980-11-01   -14.263575
1980-12-01    66.161425
Name: resid, dtype: float64
```

Multiplicative Decomposition of Rose:



Rose Trend:

```
Trend
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01    147.083333
1980-08-01    148.125000
1980-09-01    148.375000
1980-10-01    148.083333
1980-11-01    147.416667
1980-12-01    145.125000
Name: trend, dtype: float64
```

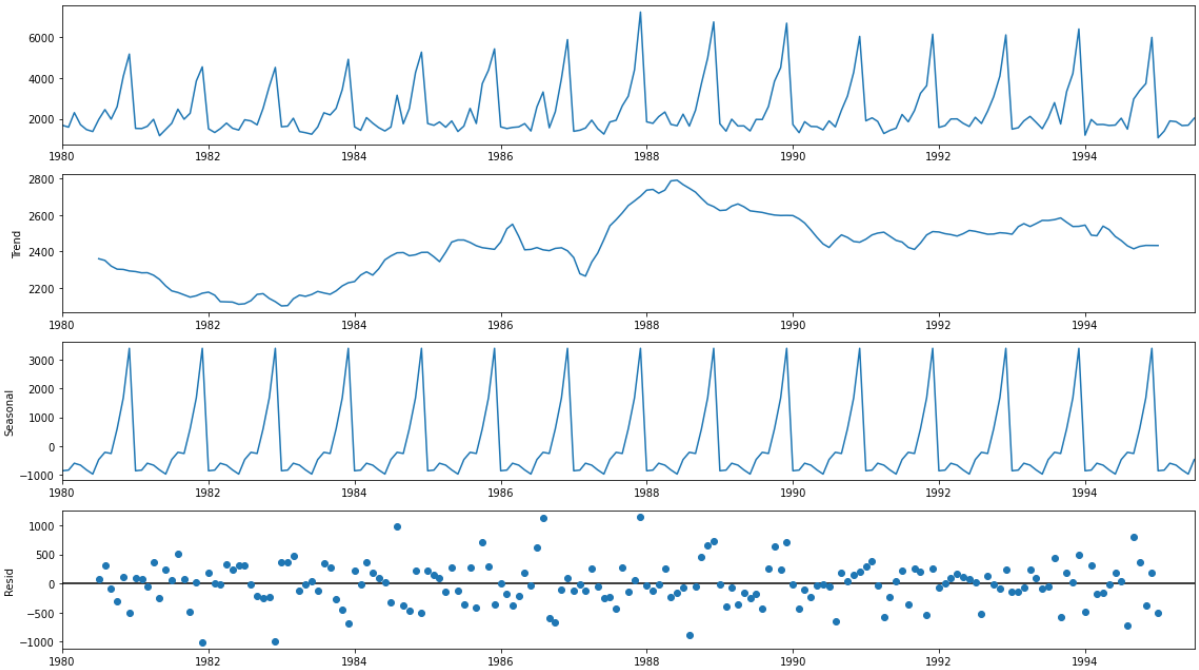
Rose Seasonality:

```
Seasonality
 YearMonth
1980-01-01    0.670111
1980-02-01    0.806163
1980-03-01    0.901164
1980-04-01    0.854024
1980-05-01    0.889415
1980-06-01    0.923985
1980-07-01    1.058038
1980-08-01    1.035881
1980-09-01    1.017648
1980-10-01    1.022573
1980-11-01    1.192349
1980-12-01    1.628646
Name: seasonal, dtype: float64
```

Rose Residual:

```
Residual
 YearMonth
1980-01-01         NaN
1980-02-01         NaN
1980-03-01         NaN
1980-04-01         NaN
1980-05-01         NaN
1980-06-01         NaN
1980-07-01    0.758258
1980-08-01    0.840720
1980-09-01    1.357674
1980-10-01    0.970771
1980-11-01    0.853378
1980-12-01    1.129646
Name: resid, dtype: float64
```

Additive Decomposition of Sparkling –



Sparkling Trend:

```
Trend
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01    2360.666667
1980-08-01    2351.333333
1980-09-01    2320.541667
1980-10-01    2303.583333
1980-11-01    2302.041667
1980-12-01    2293.791667
Name: trend, dtype: float64
```

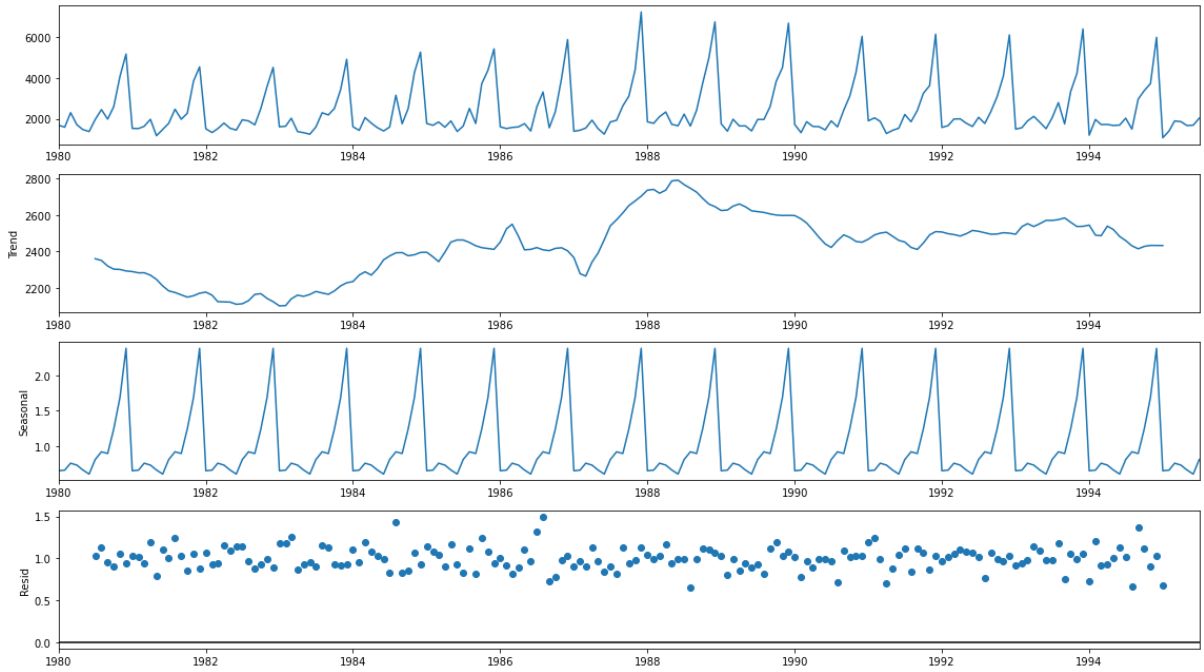
Sparkling Seasonality:

```
Seasonality
 YearMonth
1980-01-01    -854.260599
1980-02-01    -830.350678
1980-03-01    -592.356630
1980-04-01    -658.490559
1980-05-01    -824.416154
1980-06-01    -967.434011
1980-07-01    -465.502265
1980-08-01    -214.332821
1980-09-01    -254.677265
1980-10-01     599.769957
1980-11-01    1675.067179
1980-12-01    3386.983846
Name: seasonal, dtype: float64
```

Sparkling Residual:

```
Residual
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01     70.835599
1980-08-01    315.999487
1980-09-01    -81.864401
1980-10-01   -307.353290
1980-11-01    109.891154
1980-12-01   -501.775513
Name: resid, dtype: float64
```

Multiplicative Decomposition of Sparkling –



Sparkling Trend:

```
Trend
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01    2360.666667
1980-08-01    2351.333333
1980-09-01    2320.541667
1980-10-01    2303.583333
1980-11-01    2302.041667
1980-12-01    2293.791667
Name: trend, dtype: float64
```

Sparkling Seasonality:

```
Seasonality
 YearMonth
1980-01-01    0.649843
1980-02-01    0.659214
1980-03-01    0.757440
1980-04-01    0.730351
1980-05-01    0.660609
1980-06-01    0.603468
1980-07-01    0.809164
1980-08-01    0.918822
1980-09-01    0.894367
1980-10-01    1.241789
1980-11-01    1.690158
1980-12-01    2.384776
Name: seasonal, dtype: float64
```

Sparkling Residual:

```
Residual
 YearMonth
1980-01-01         NaN
1980-02-01         NaN
1980-03-01         NaN
1980-04-01         NaN
1980-05-01         NaN
1980-06-01         NaN
1980-07-01    1.029230
1980-08-01    1.135407
1980-09-01    0.955954
1980-10-01    0.907513
1980-11-01    1.050423
1980-12-01    0.946770
Name: resid, dtype: float64
```

• Additive Models –

- The seasonality is relatively constant over time
- yt = Trend + Seasonality + Residual

• Multiplicative Models –

- The seasonality increases or decreases over time. It is proportionate to the trend
- yt = Trend * Seasonality * Residual

• Here by just observing the Residual patterns of Additive and Multiplicative models of Rose and Sparkling datasets. It seems that –

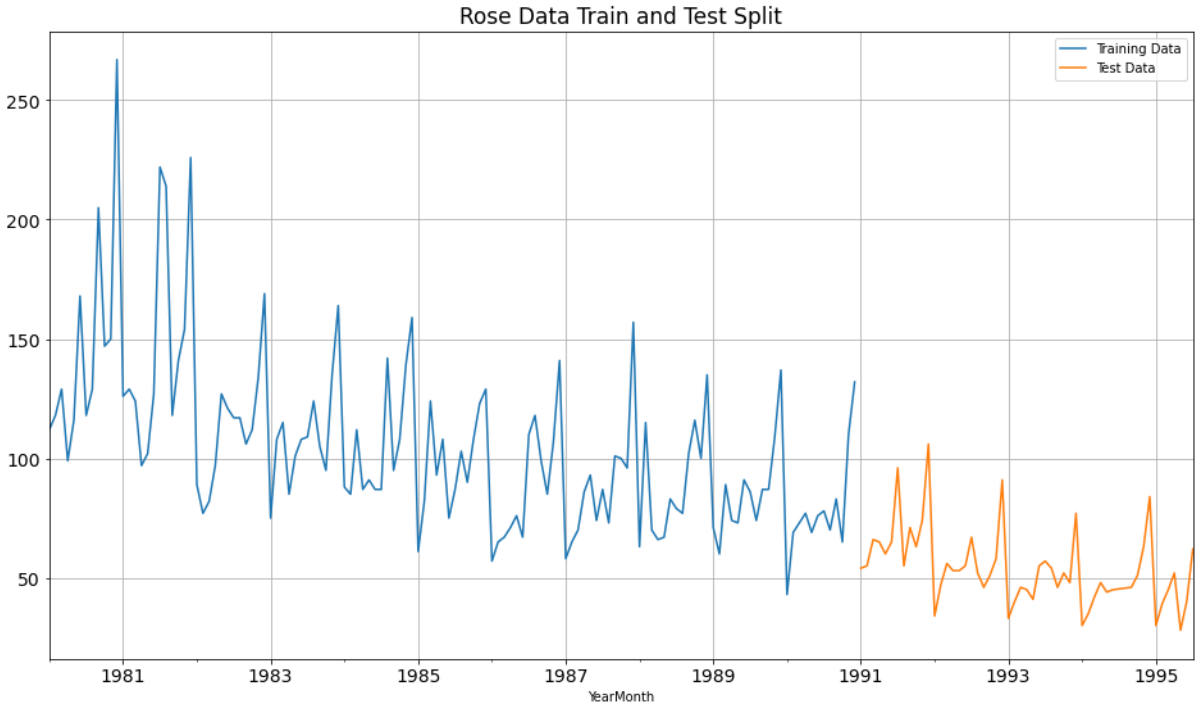
- Rose is Multiplicative
- Sparkling is Additive

3. SPLIT THE DATA INTO TRAINING AND TEST. THE TEST DATA SHOULD START IN 1991.

Solution:

• Both datasets of Rose and Sparkling are split at the year 1991

• Test datasets start at 1991



• **Rose dataset – TRAIN**

```
First few rows of Rose Training Data
```

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

```
Last few rows of Rose Training Data
```

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

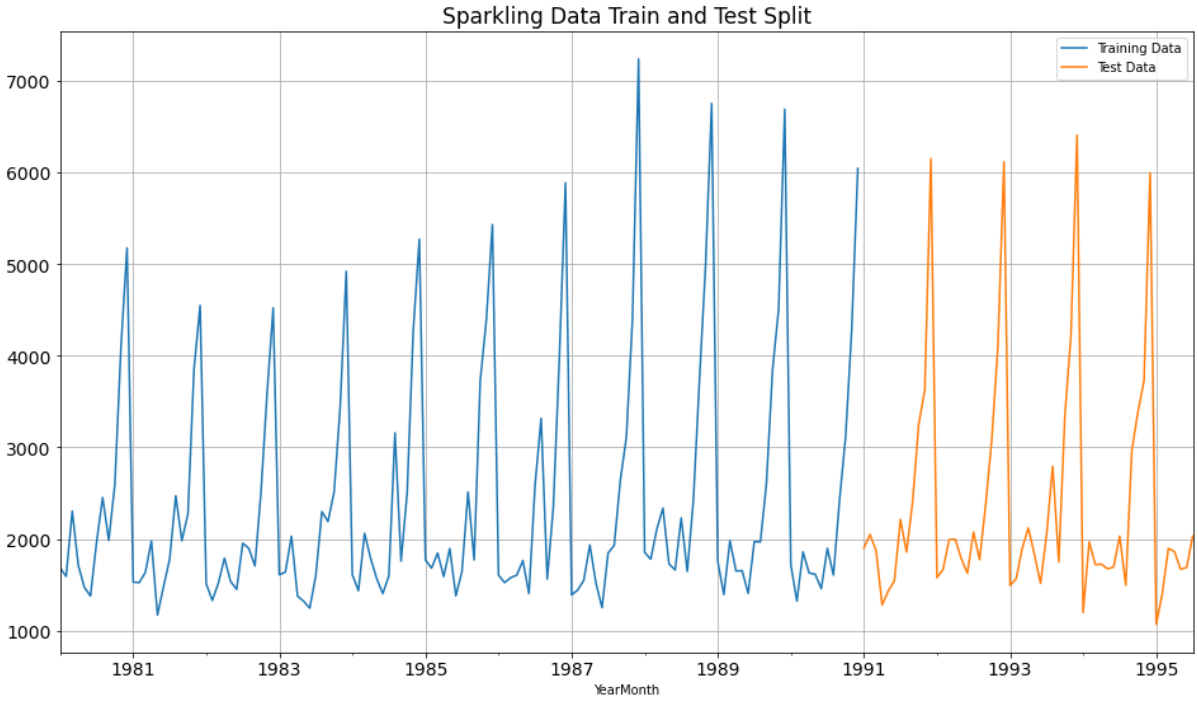
• **Rose dataset – TEST**

First few rows of Rose Test Data

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last few rows of Rose Test Data

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0



- **Sparkling dataset – TRAIN**

First few rows of Sparkling Training Data

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

Last few rows of Sparkling Training Data

YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

• **Sparkling dataset – TEST**

First few rows of Sparkling Test Data

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

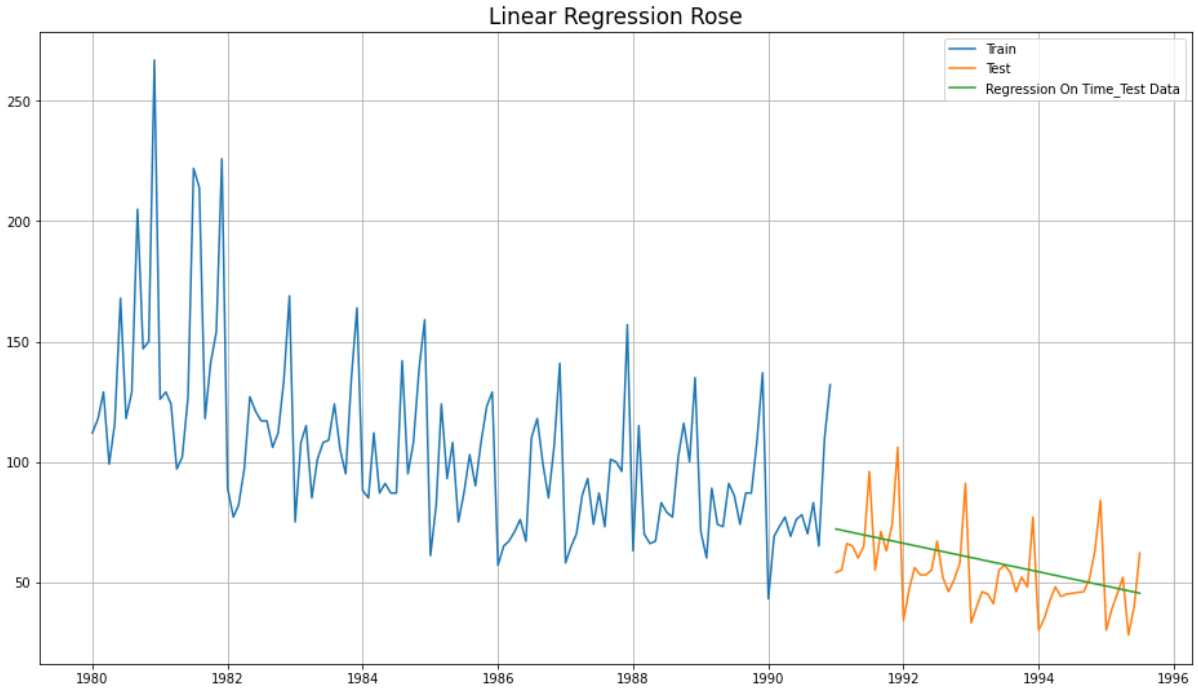
Last few rows of Sparkling Test Data

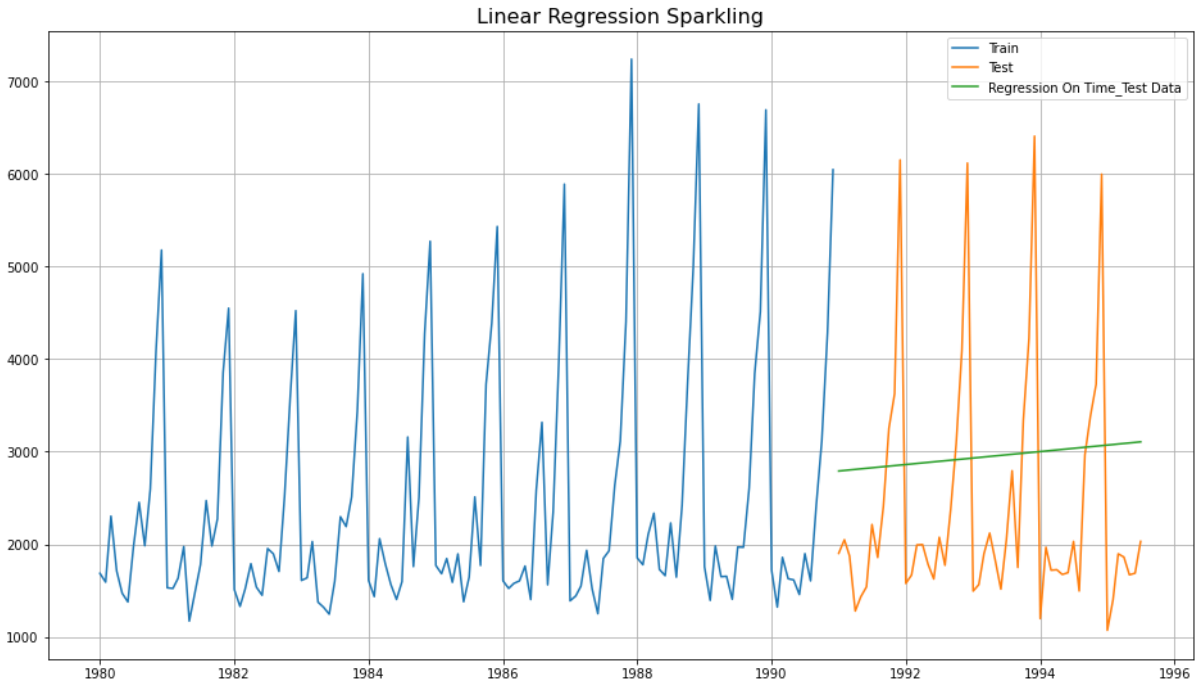
YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

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.

Solution:

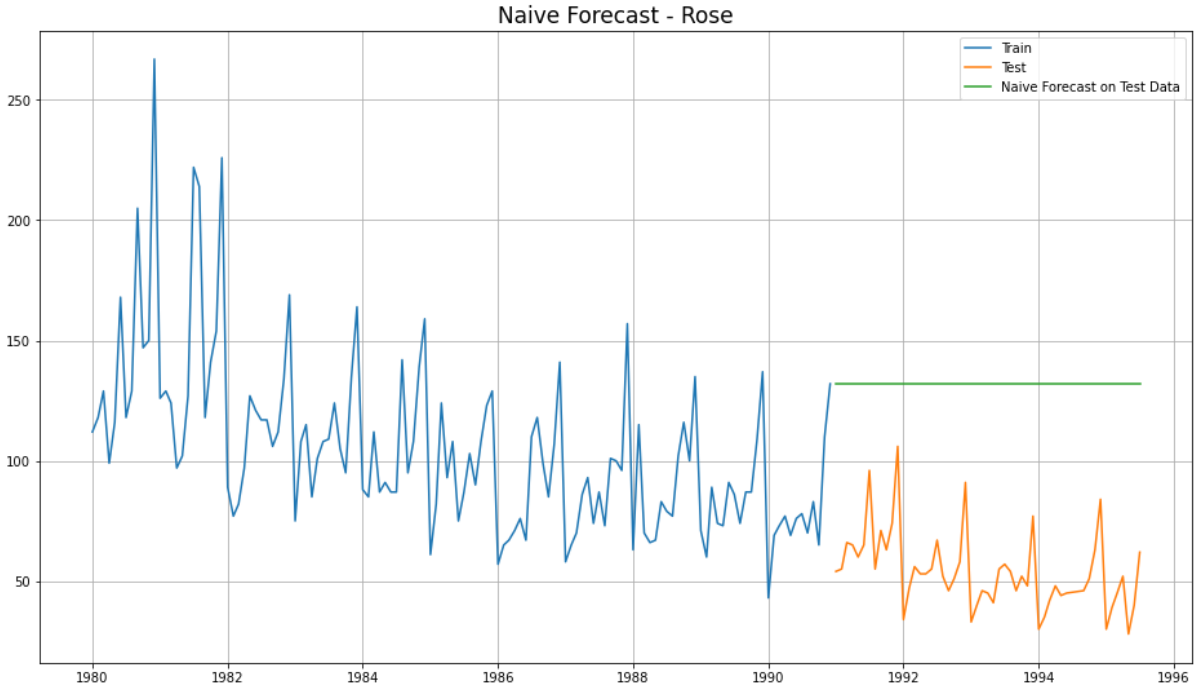
Model 1 - Linear Regression

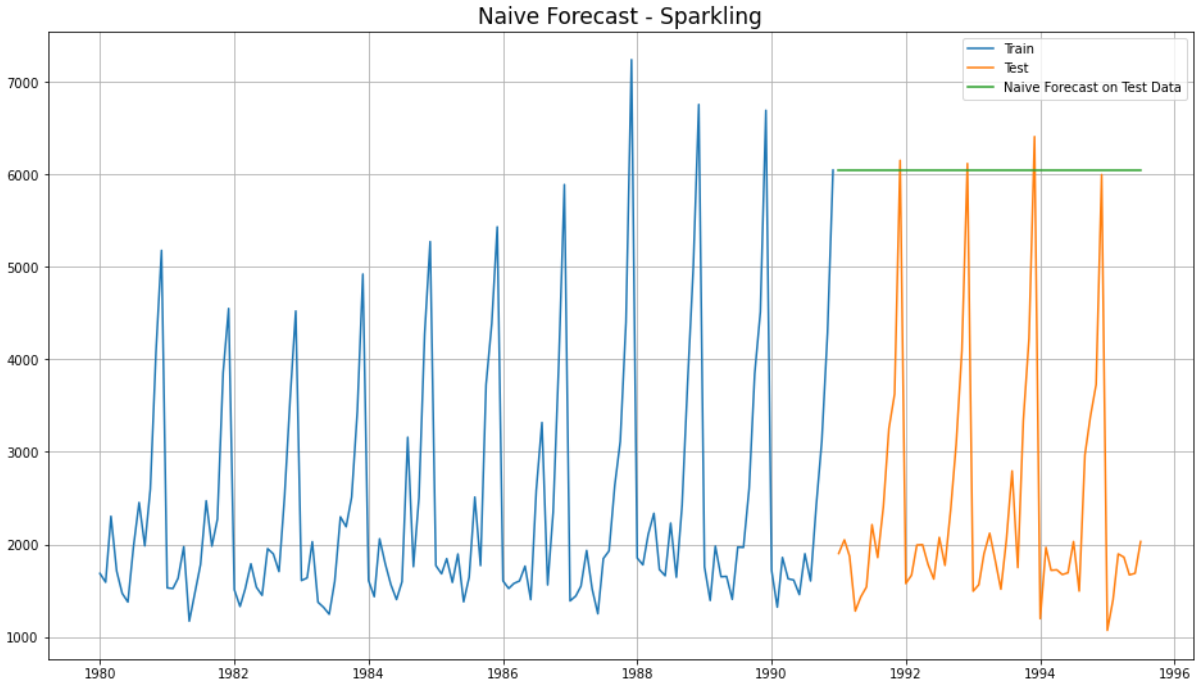




	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175

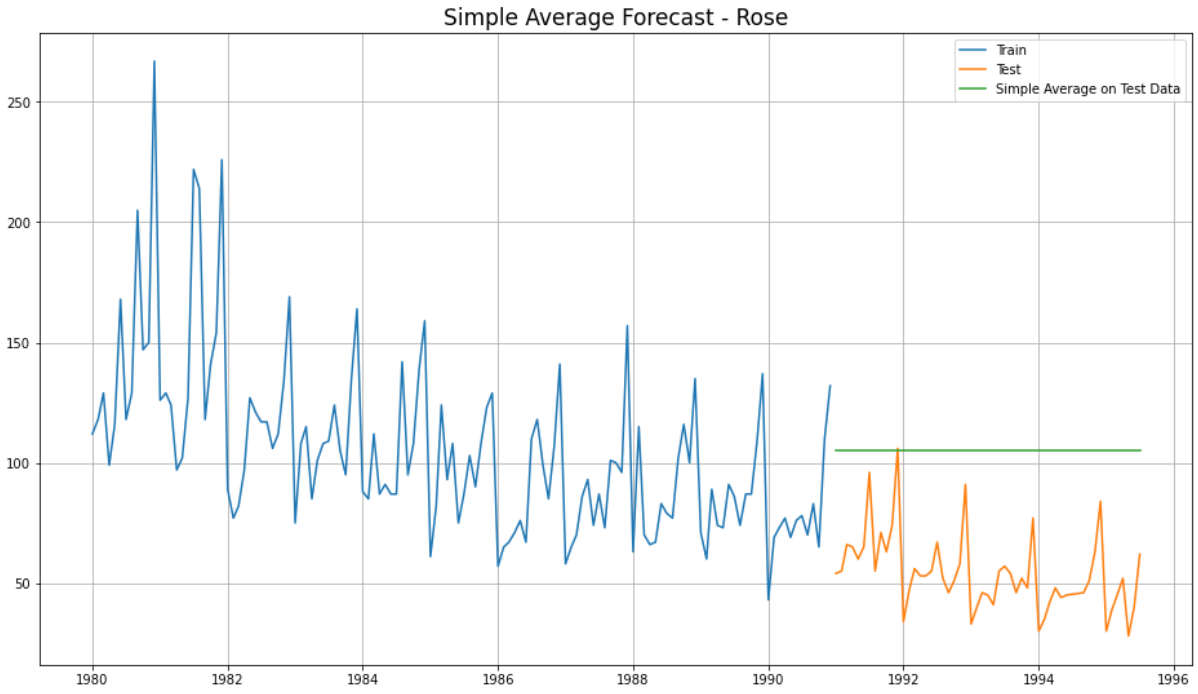
Model 2 - Naive Bayes

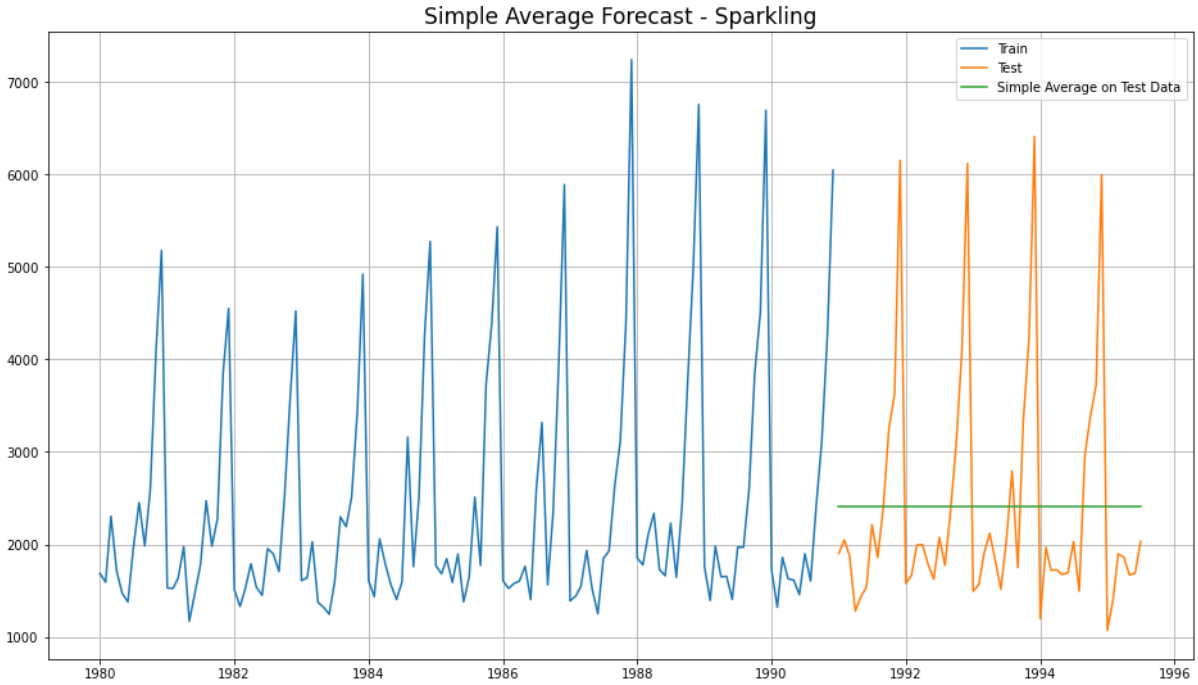




	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352

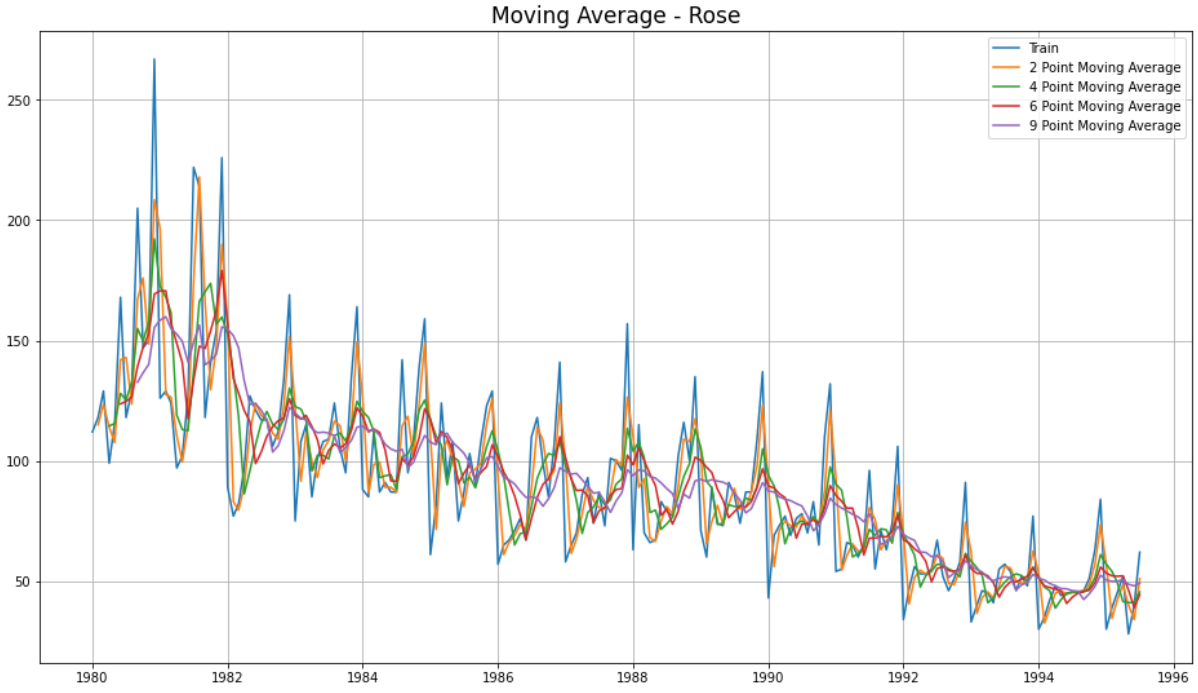
Model 3 - Simple Average





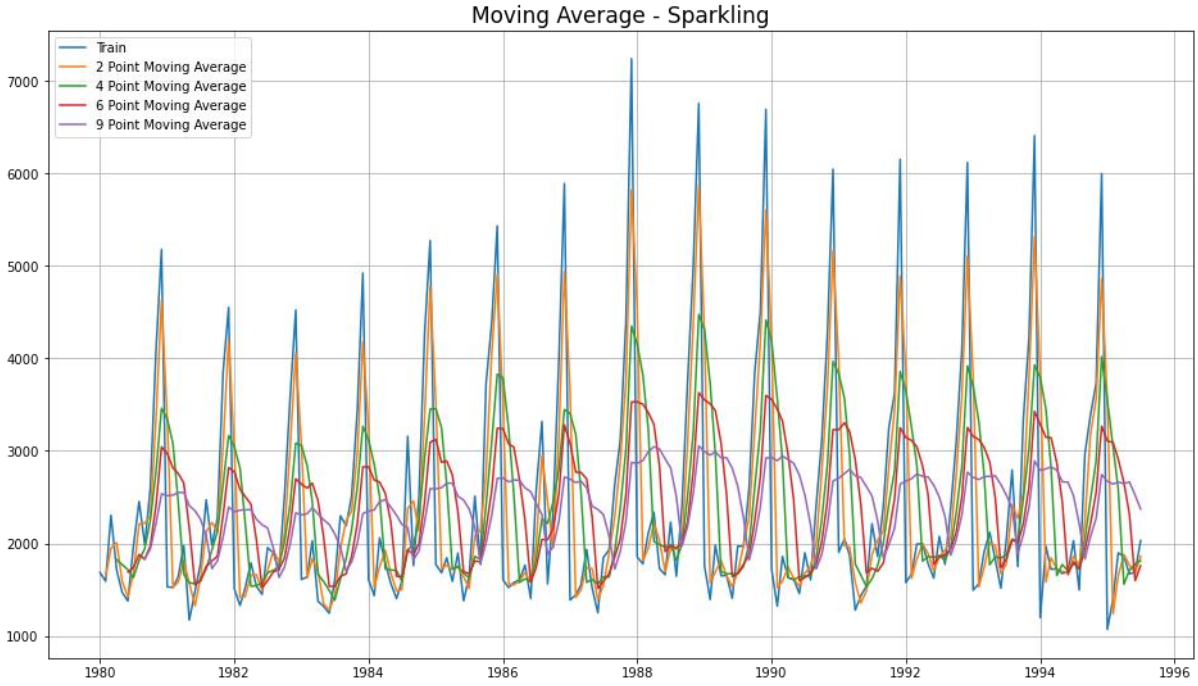
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804

Model 4: Moving Average (MA) – Rose



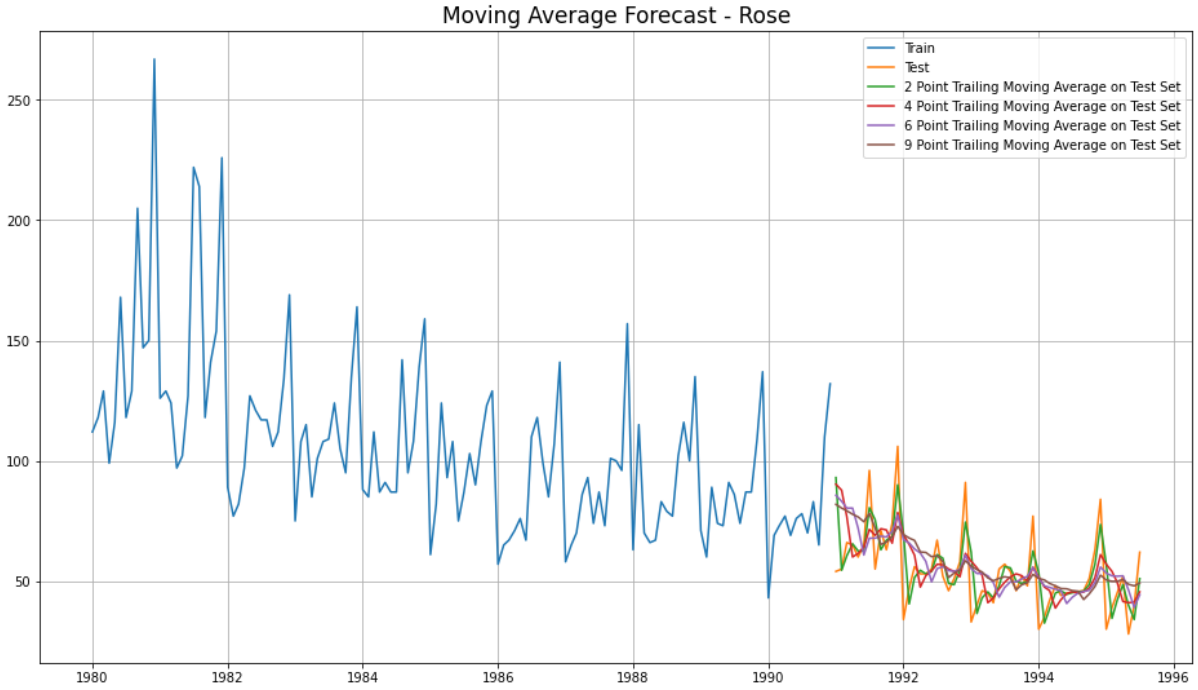
	Test RMSE Rose
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630

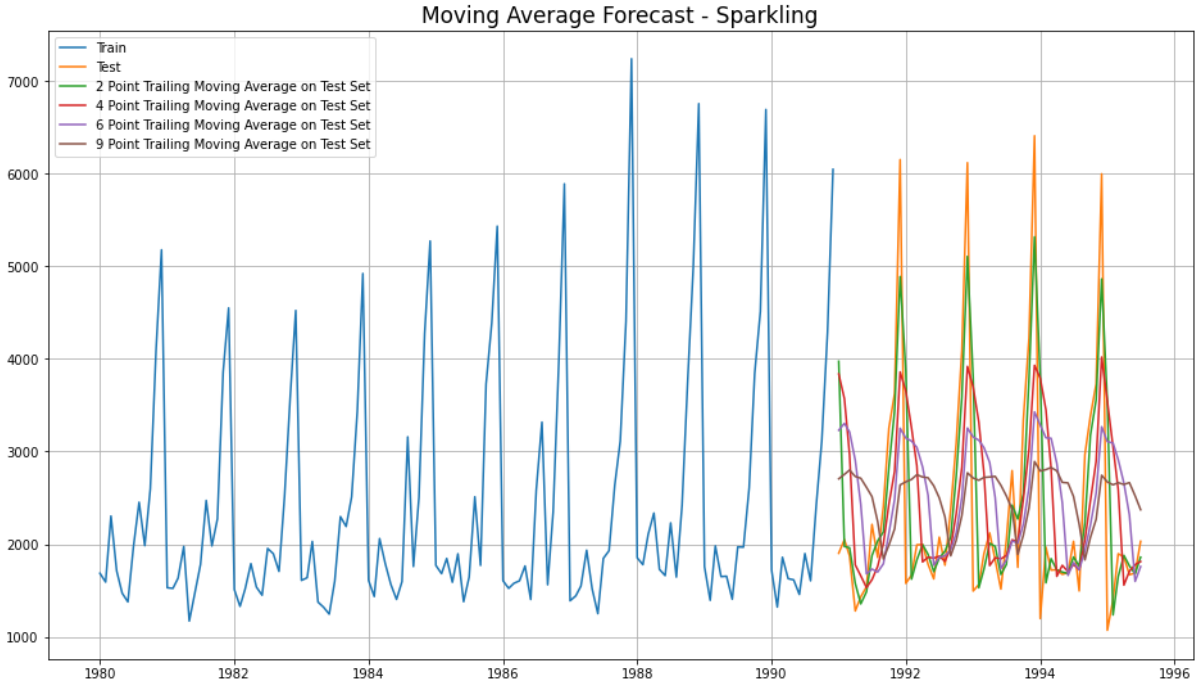
Model 4 - Moving Average (Sparkling)



	Test RMSE Sparkling
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

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

- 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 ( )

- 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 ( )

- 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

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

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315

• 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

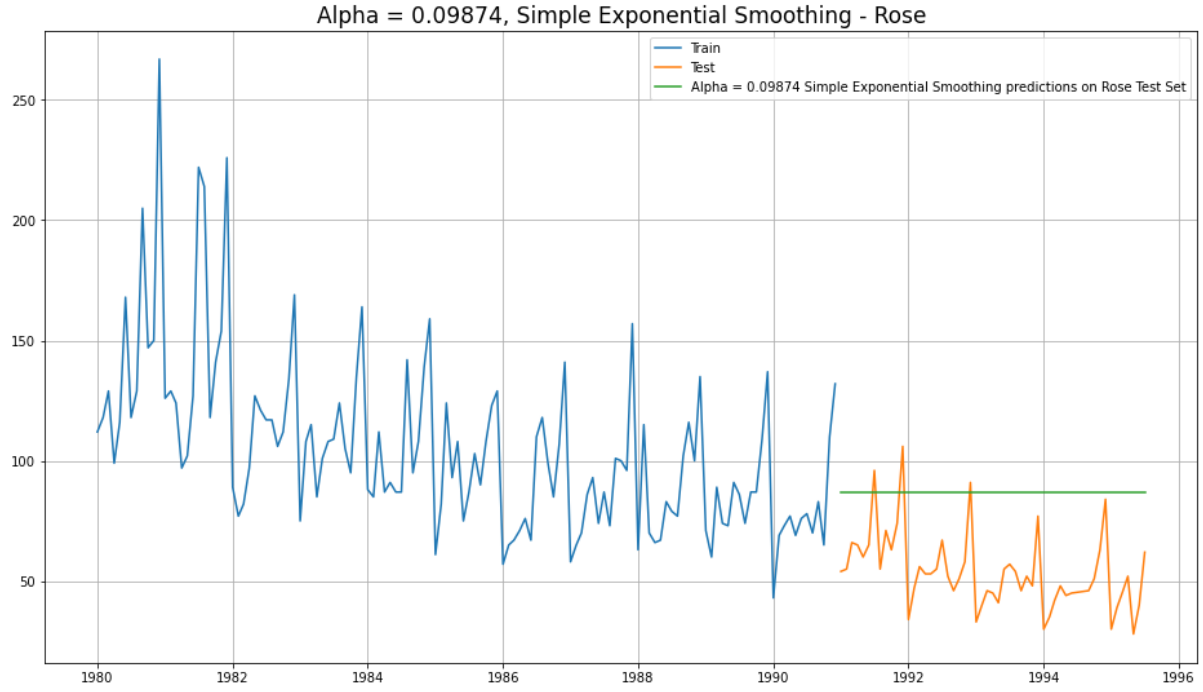
• Exponential smoothing methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality) where each can be None(N), Additive (N), Additive damped (Ad), Multiplicative (M) or Multiplicative damped (Md)

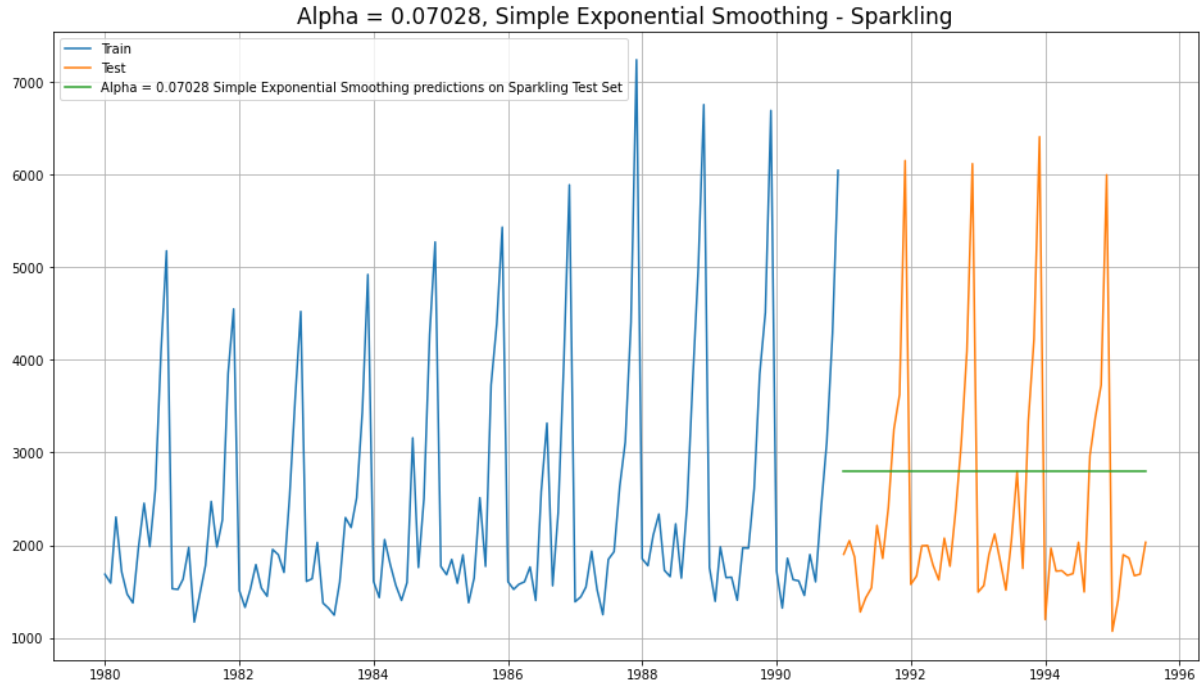
• One or more parameters control how fast the weights decay. The values of the parameters lie between 0 and 1

• 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)



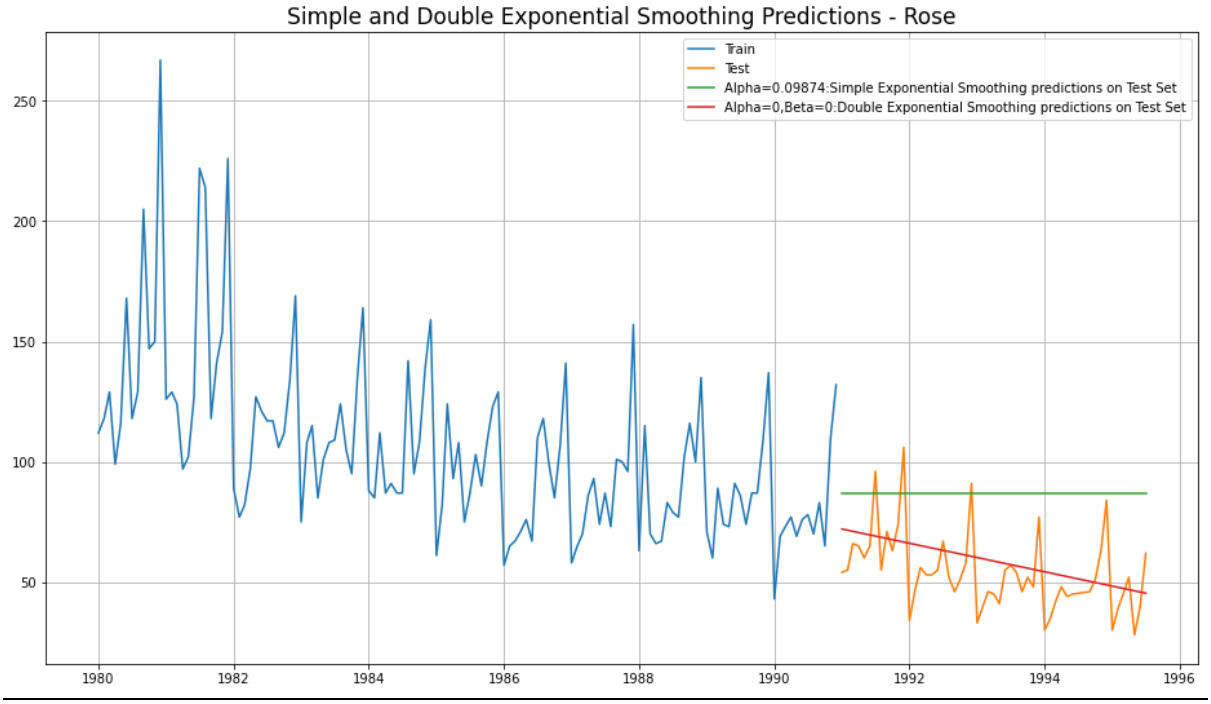


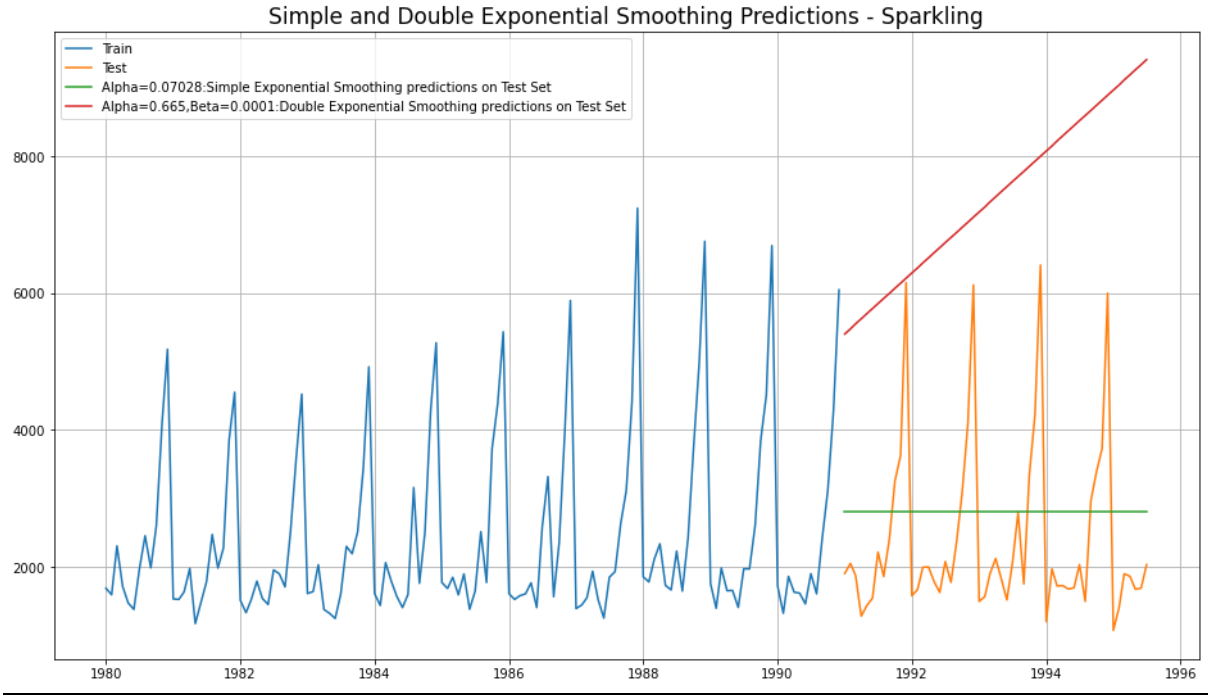
- For Rose - Level Parameter, Alpha = 0.09874
- For Sparkling - Level Parameter, Alpha = 0.07028

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315
Simple Exponential Smoothing	36.796241	1338.008384

• 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

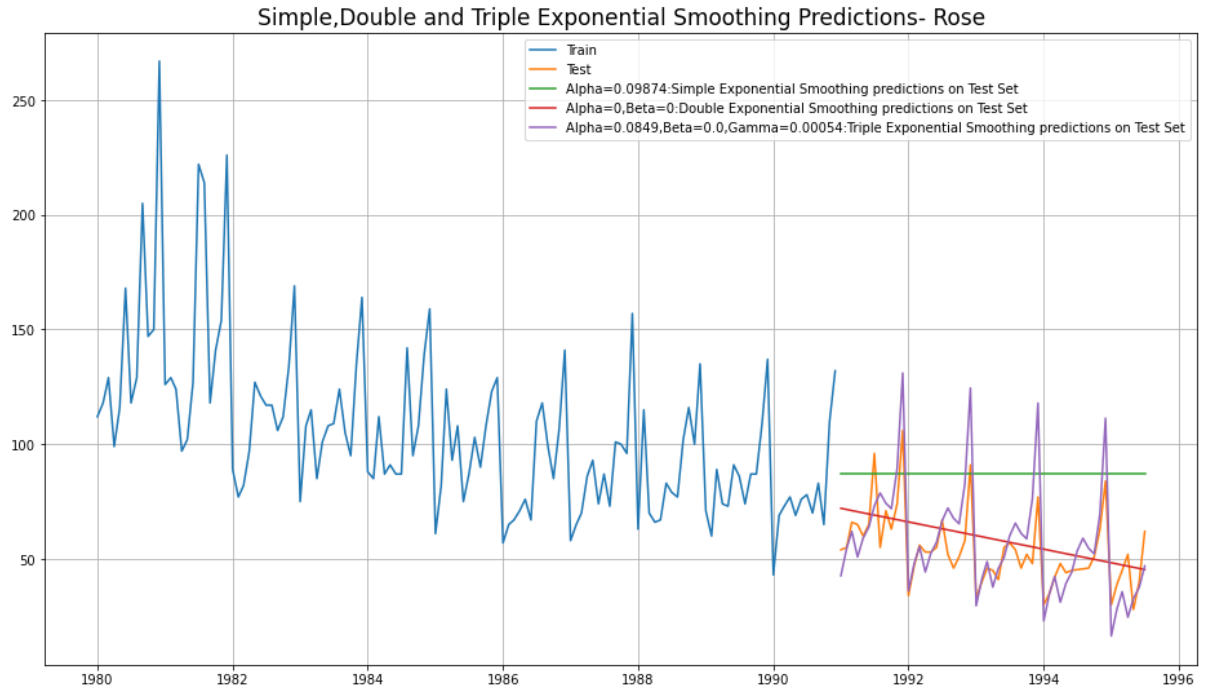
• Sparkling - Level parameter, Alpha = 0.665

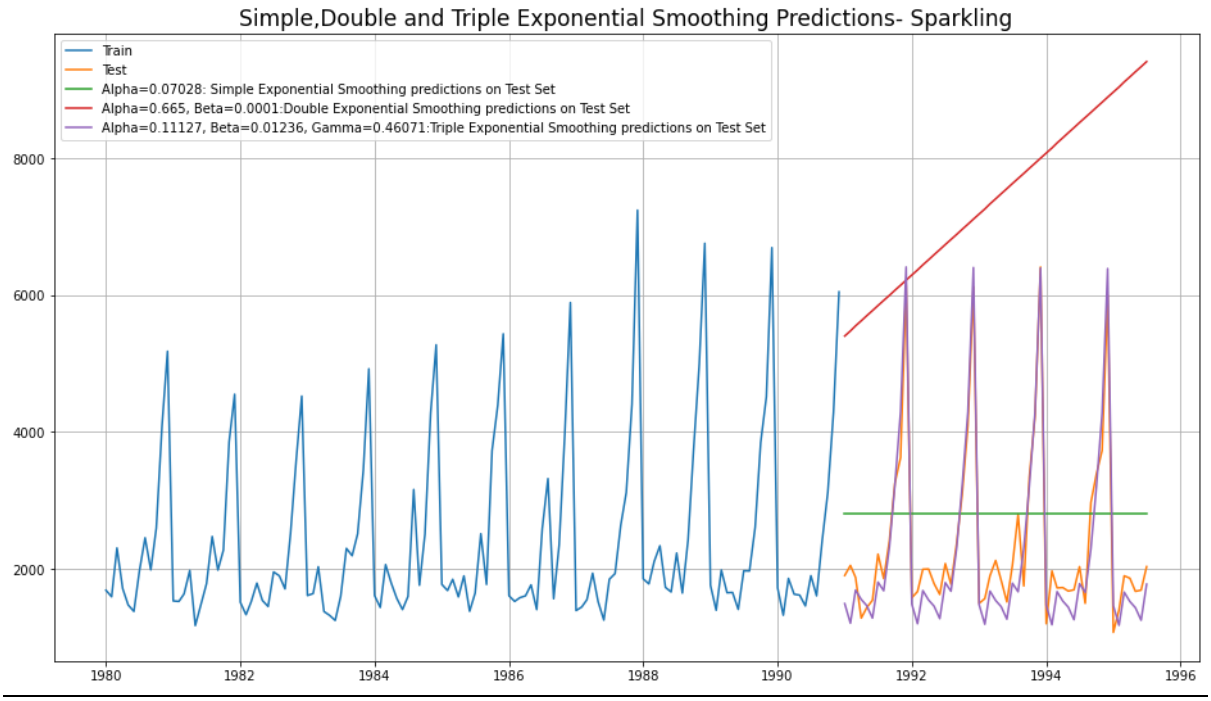
Trend parameter, Beta = 0.0001

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315
Simple Exponential Smoothing	36.796241	1338.008384
Double Exponential Smoothing	15.268944	5291.879833

• 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

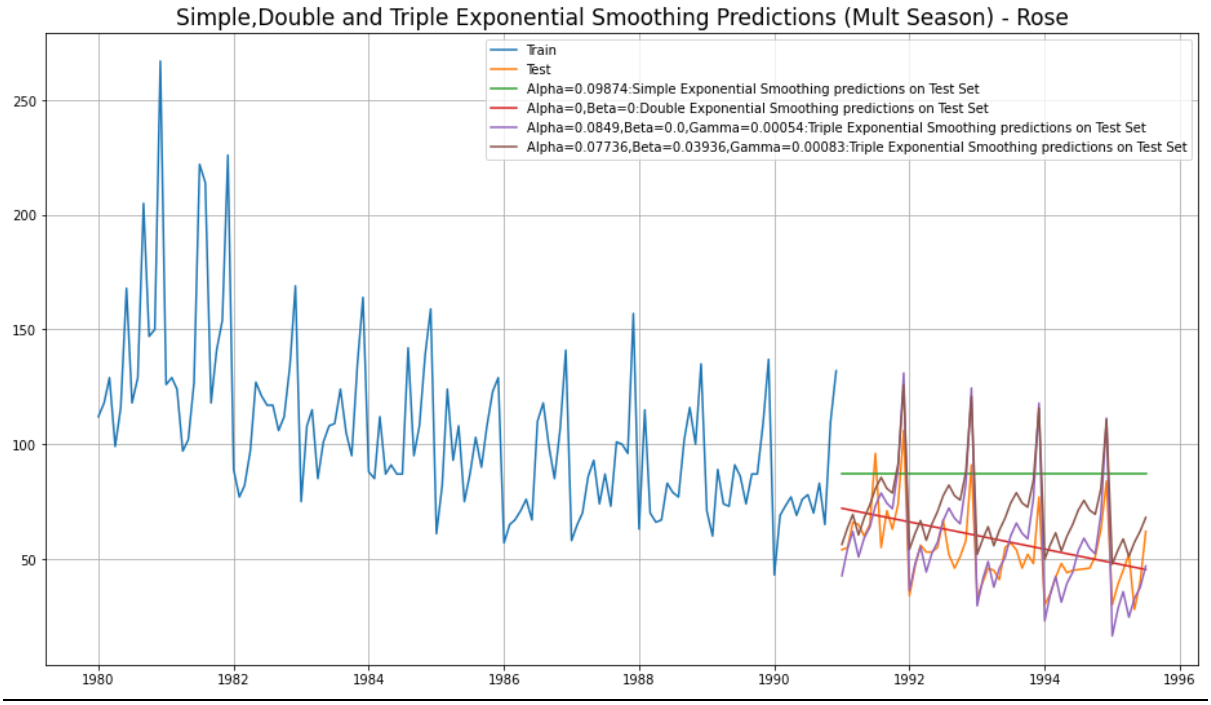
Seasonality parameter, Gamma = 0.46071

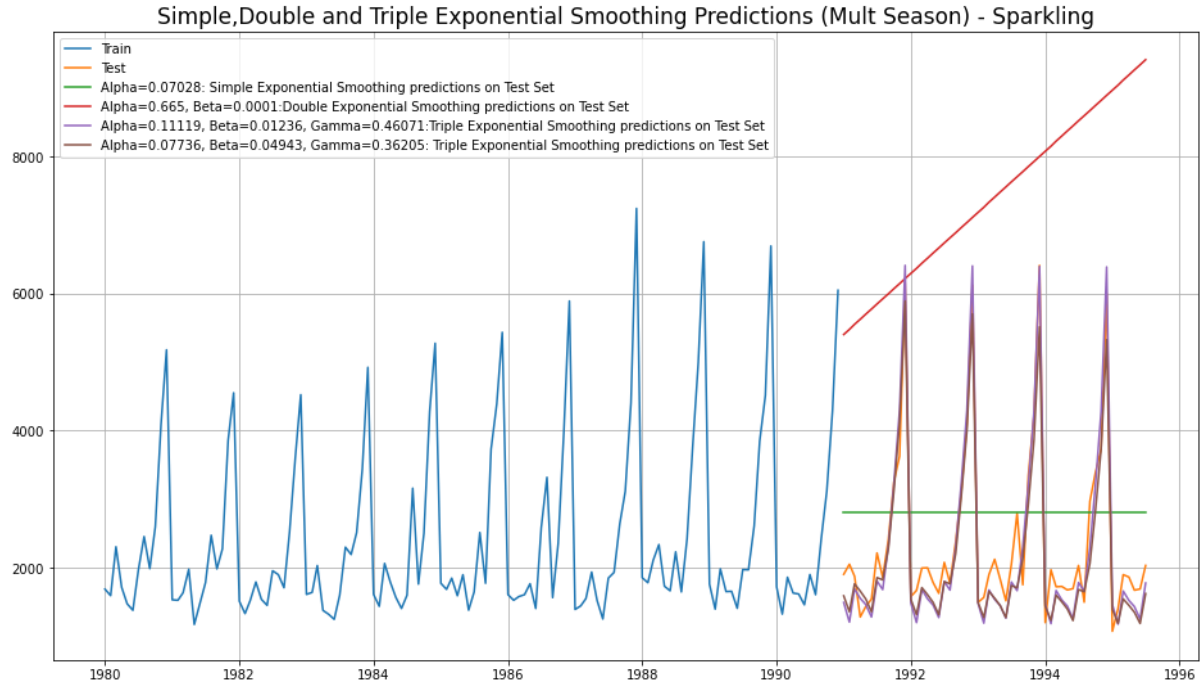
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315
Simple Exponential Smoothing	36.796241	1338.008384
Double Exponential Smoothing	15.268944	5291.879833
Triple Exponential Smoothing (Additive Season)	14.249661	378.951023

• 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

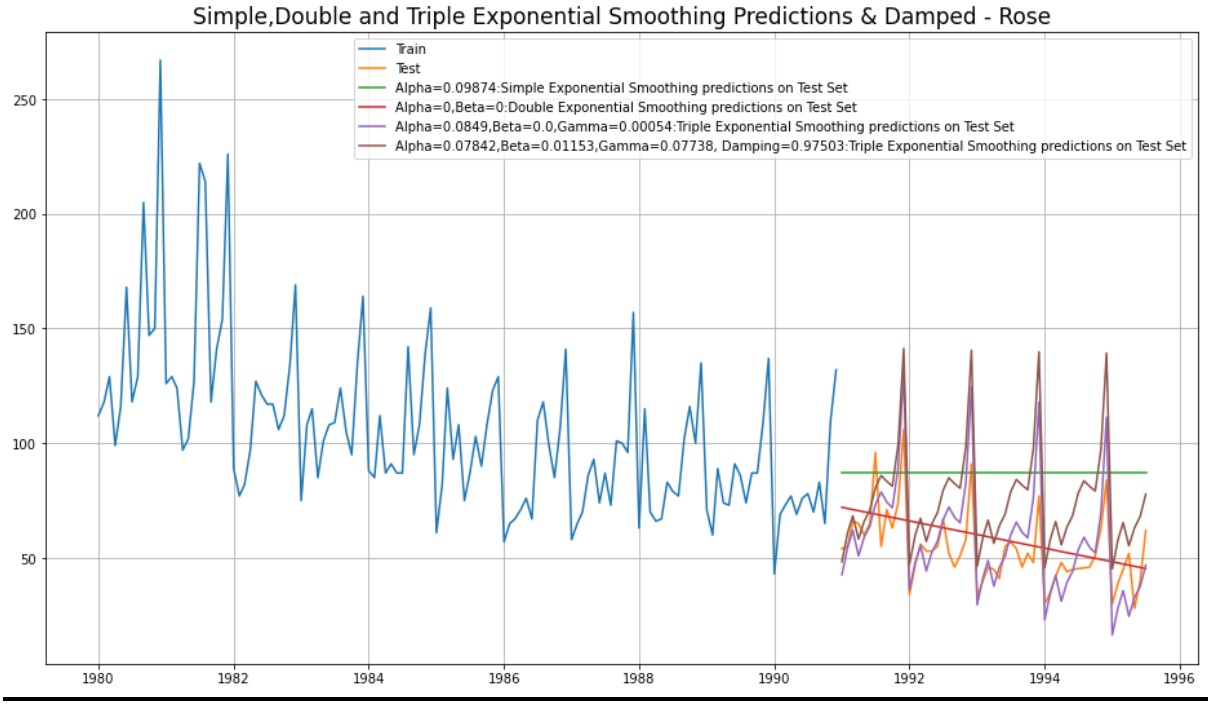
Seasonality parameter, Gamma = 0.36205

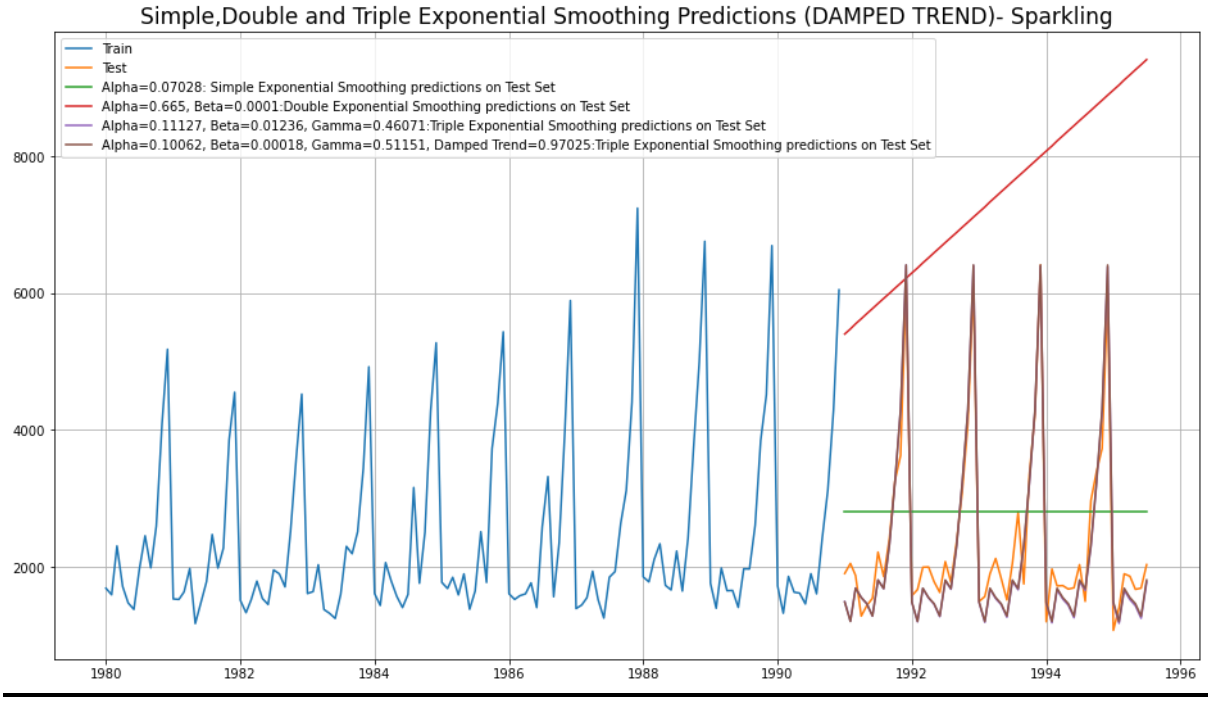
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315
Simple Exponential Smoothing	36.796241	1338.008384
Double Exponential Smoothing	15.268944	5291.879833
Triple Exponential Smoothing (Additive Season)	14.249661	378.951023
Triple Exponential Smoothing (Multiplicative Season)	20.156763	404.286809

• 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

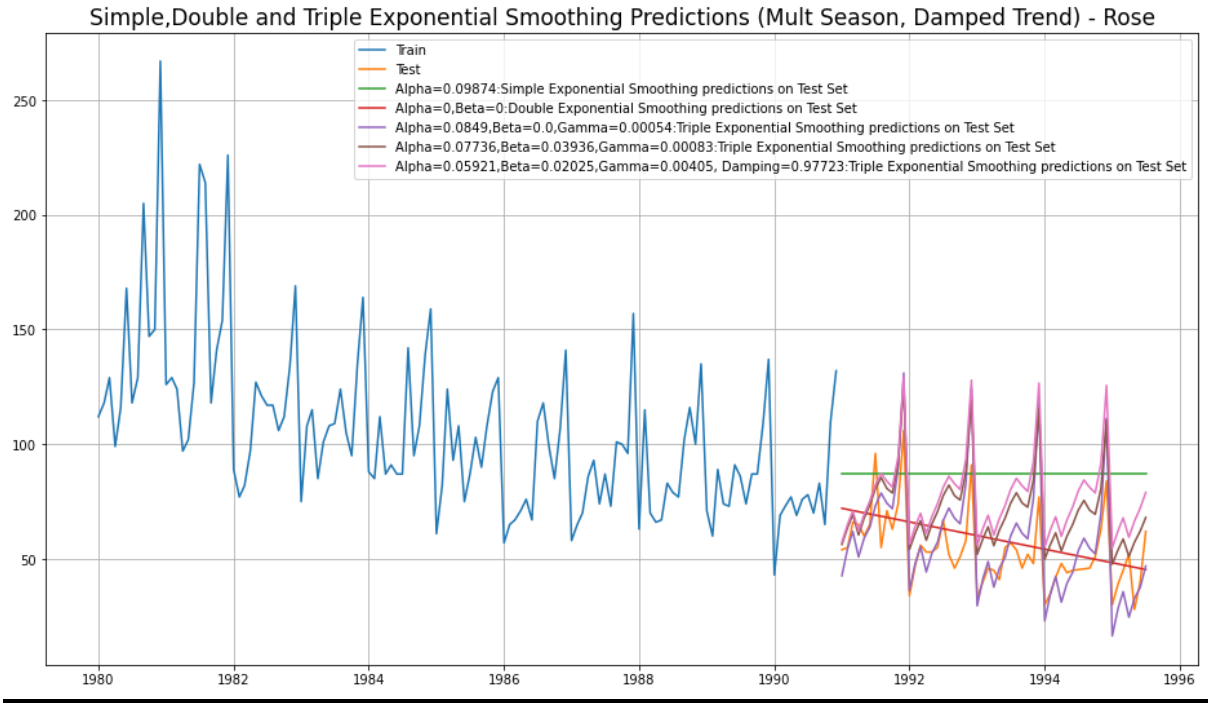
Damping factor = 0.97025

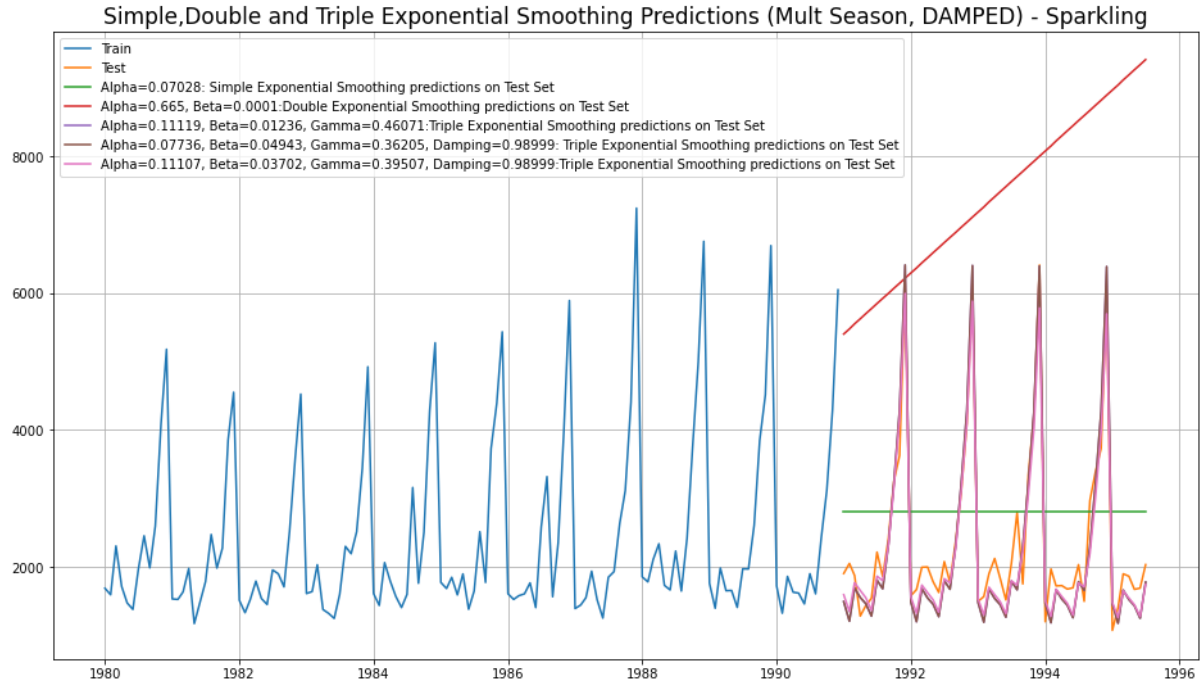
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315
Simple Exponential Smoothing	36.796241	1338.008384
Double Exponential Smoothing	15.268944	5291.879833
Triple Exponential Smoothing (Additive Season)	14.249661	378.951023
Triple Exponential Smoothing (Multiplicative Season)	20.156763	404.286809
Triple Exponential Smoothing (Additive Season, Damped Trend)	26.360083	378.951023

• 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

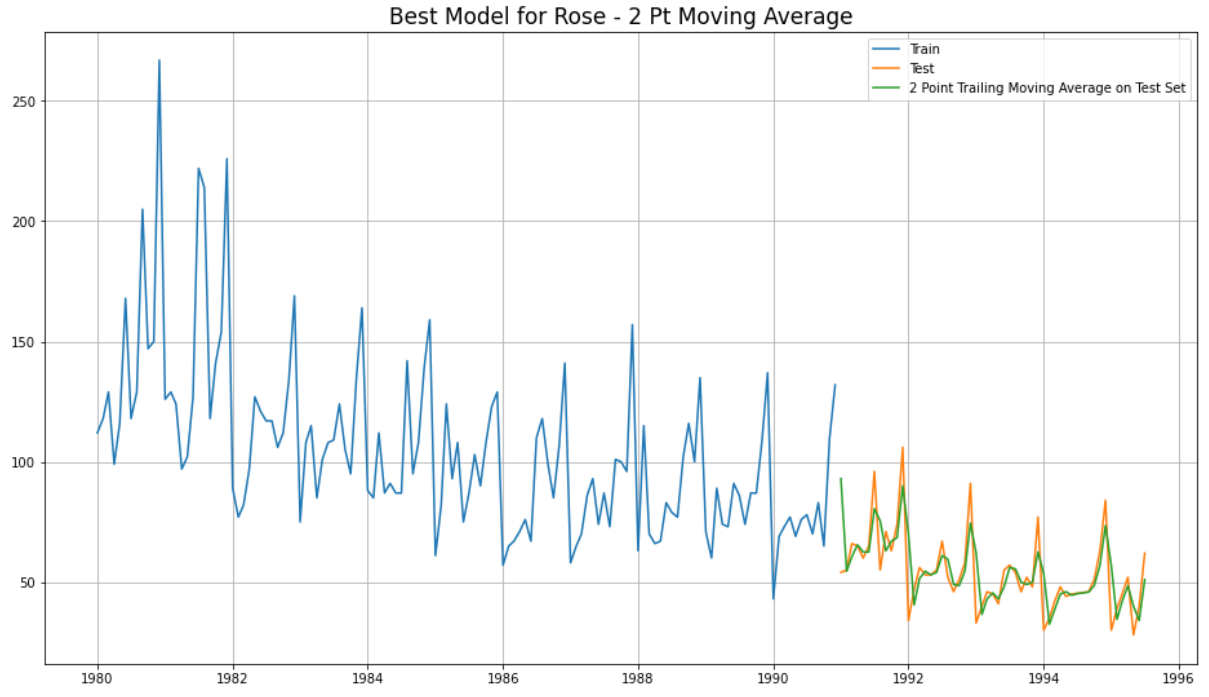
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315
Simple Exponential Smoothing	36.796241	1338.008384
Double Exponential Smoothing	15.268944	5291.879833
Triple Exponential Smoothing (Additive Season)	14.249661	378.951023
Triple Exponential Smoothing (Multiplicative Season)	20.156763	404.286809
Triple Exponential Smoothing (Additive Season, Damped Trend)	26.360083	378.951023
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)	25.955974	352.439828

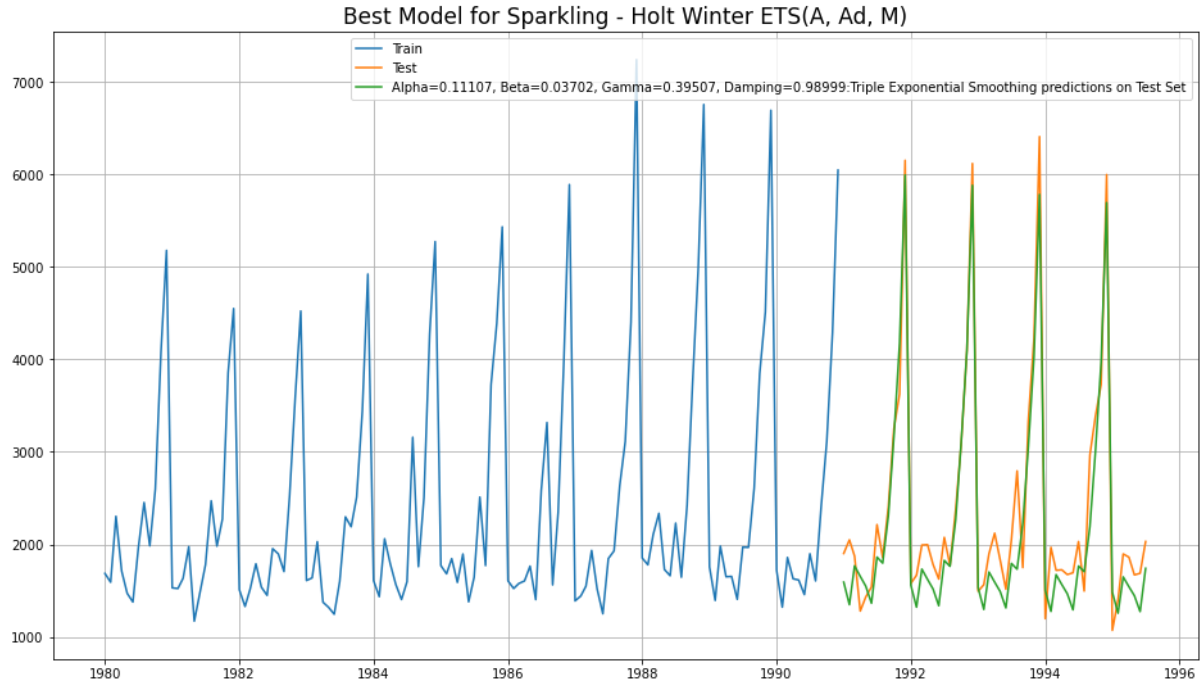
• We conclude that models with least RMSE,

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 –





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

Solution:

<u>To Check Stationarity of Data –</u>

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

• Hypotheses of ADF Test :

H_0 = Time Series is not Stationary

H_a = 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

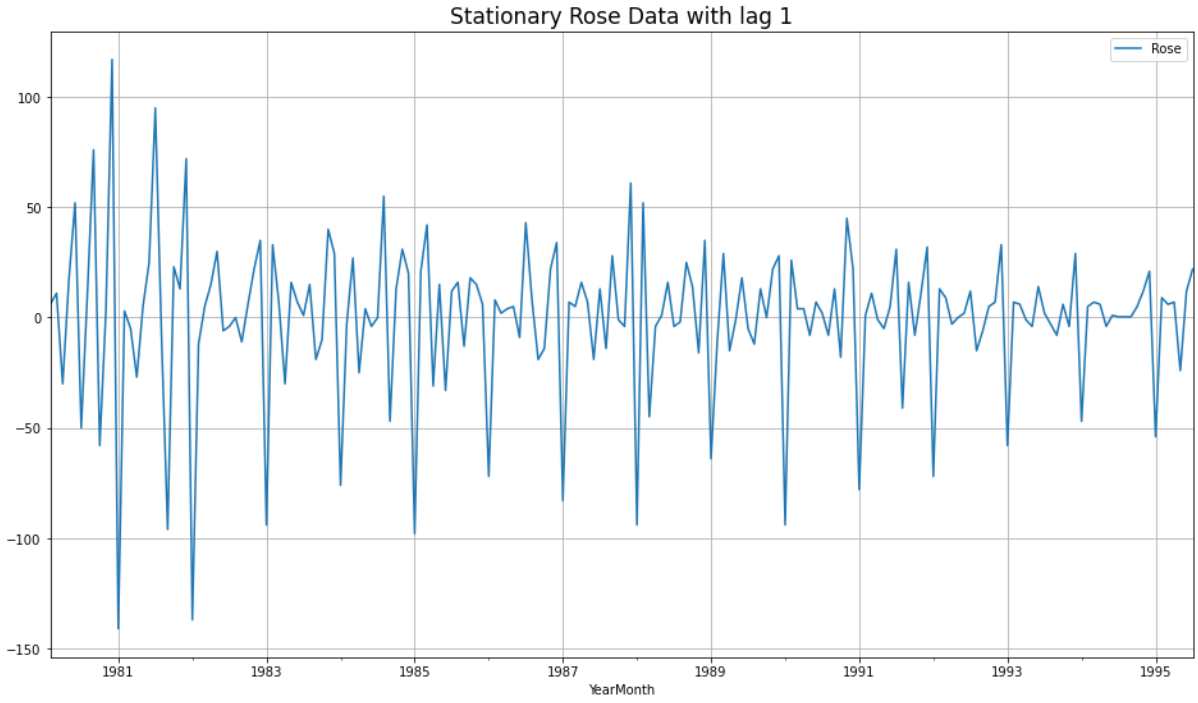
• We fail to reject the Null Hypothesis and hence conclude that Rose Wine Time Series is Not Stationary

• 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

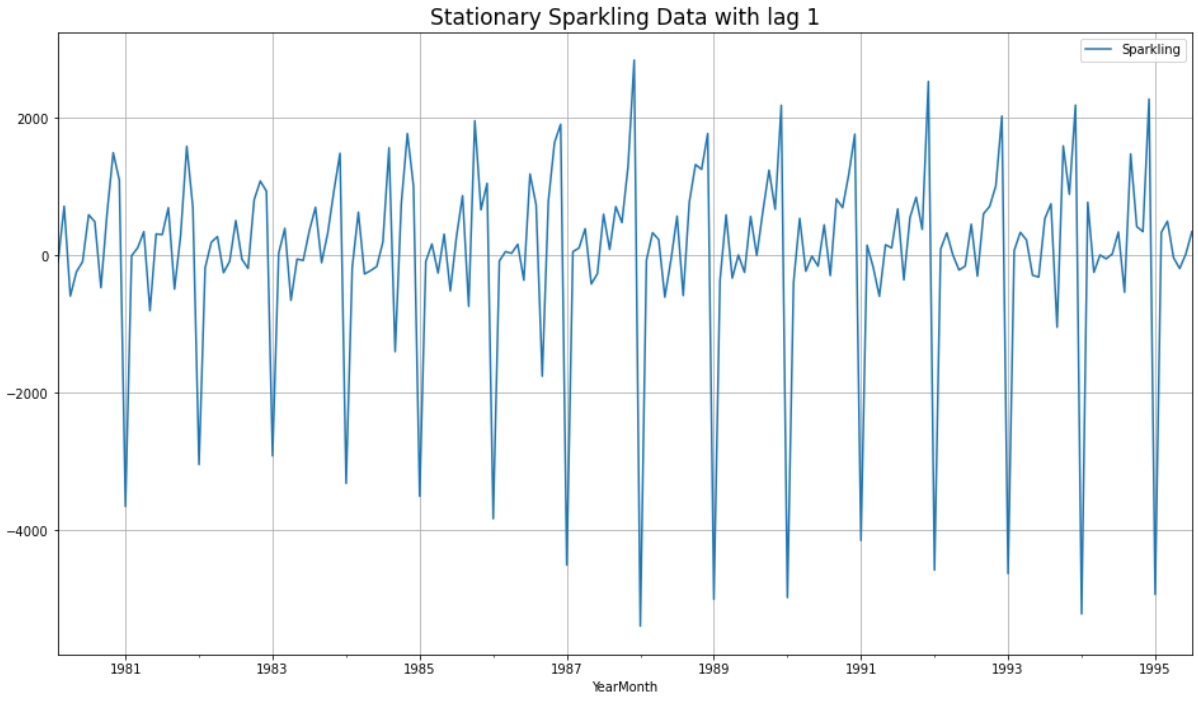
• We fail to reject the Null Hypothesis and hence conclude that Sparkling Wine Time Series is Not Stationary

• 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



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.

Solution:

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)

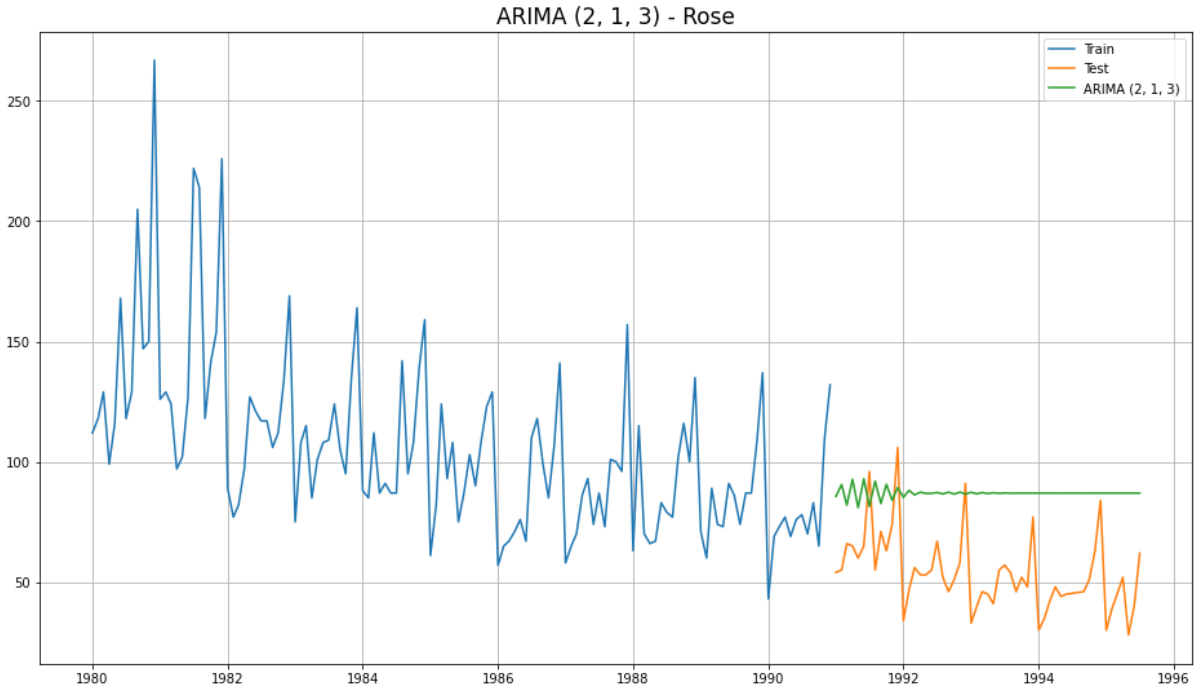
• We fit ARIMA models to each of these combinations for both datasets

• We choose the combination with the least Akaike Information Criteria (AIC)

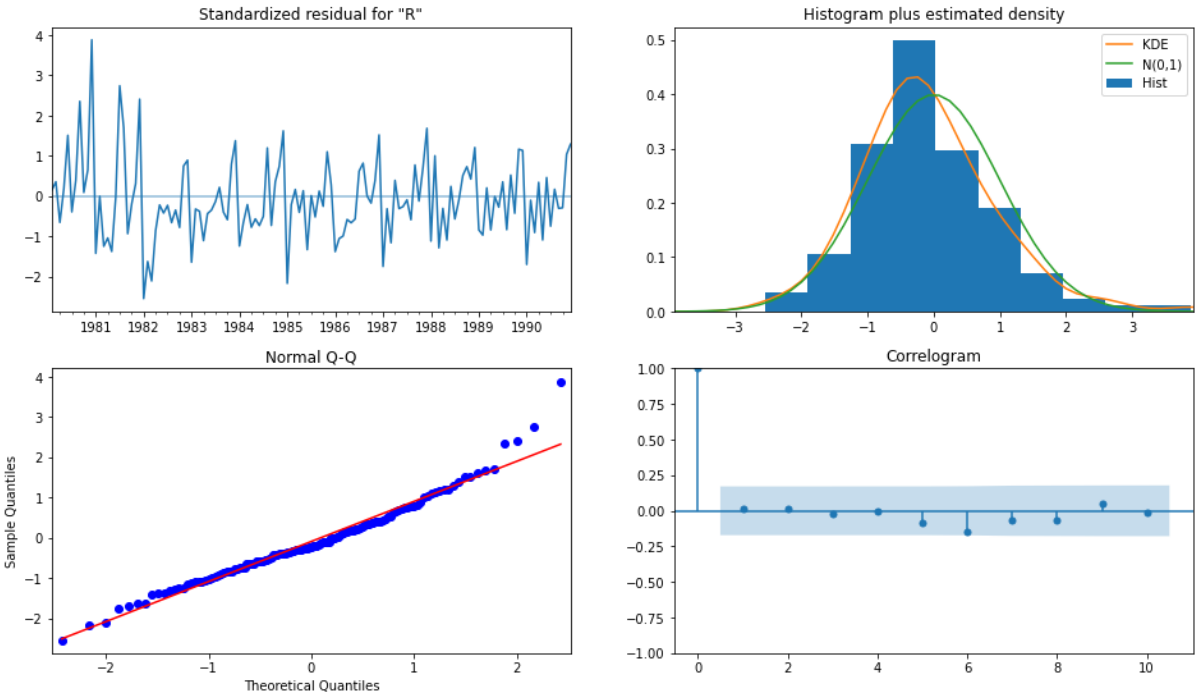
• We fit ARIMA to this combination of (p, d, q) to the Train set and forecast on the Test set

• 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)**

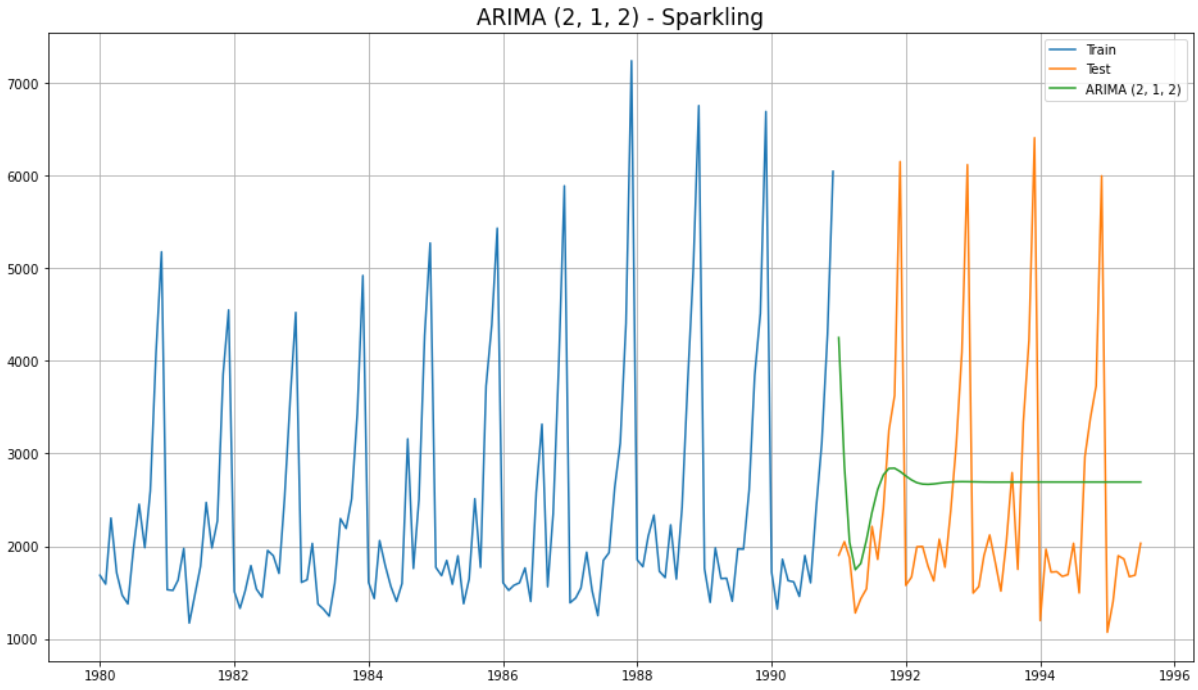


ARIMA (2, 1, 3) Diagnostic Plot – Rose

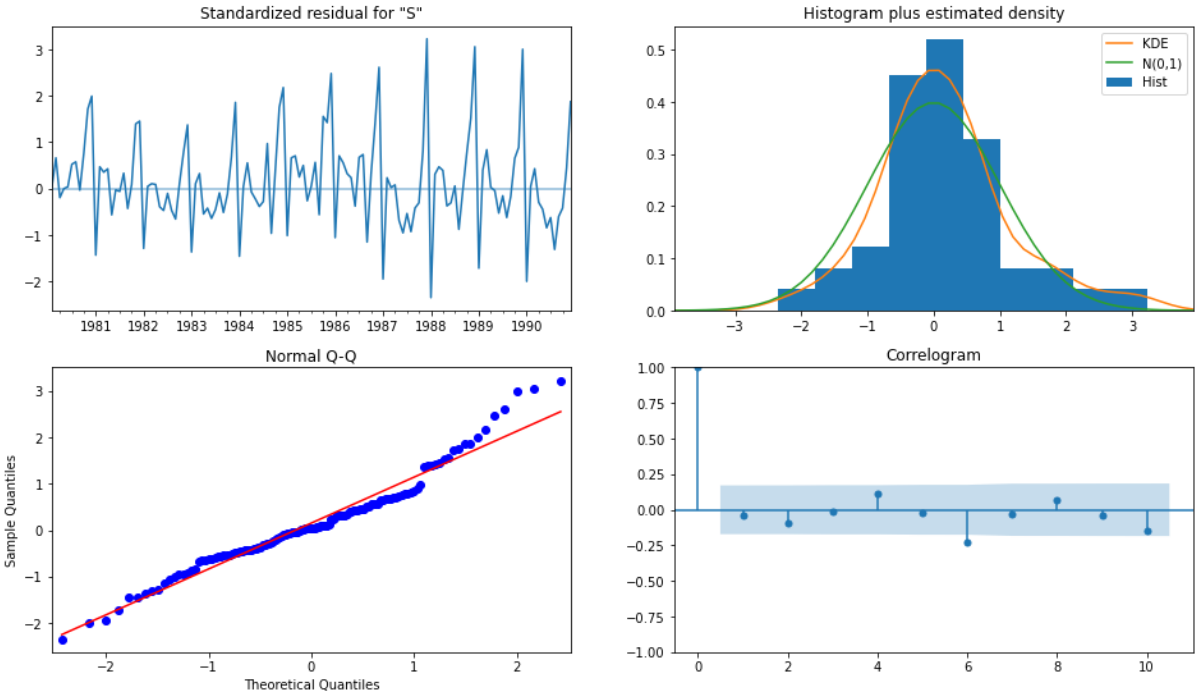


	Test RMSE Rose	Test MAPE Rose
ARIMA(2,1,3)	36.812984	75.838991

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



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



	RMSE	MAPE
ARIMA(2,1,2)	1299.97964	47.099986

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)

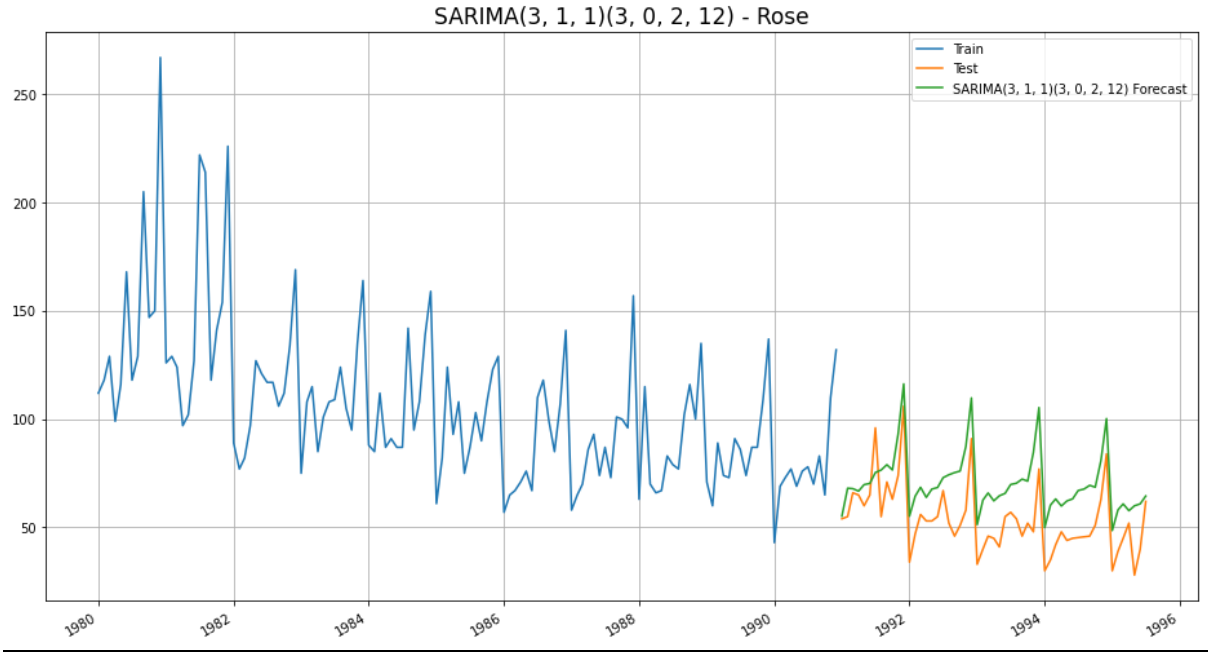
Model: (2, 1, 3)(2, 0, 3, 12)

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

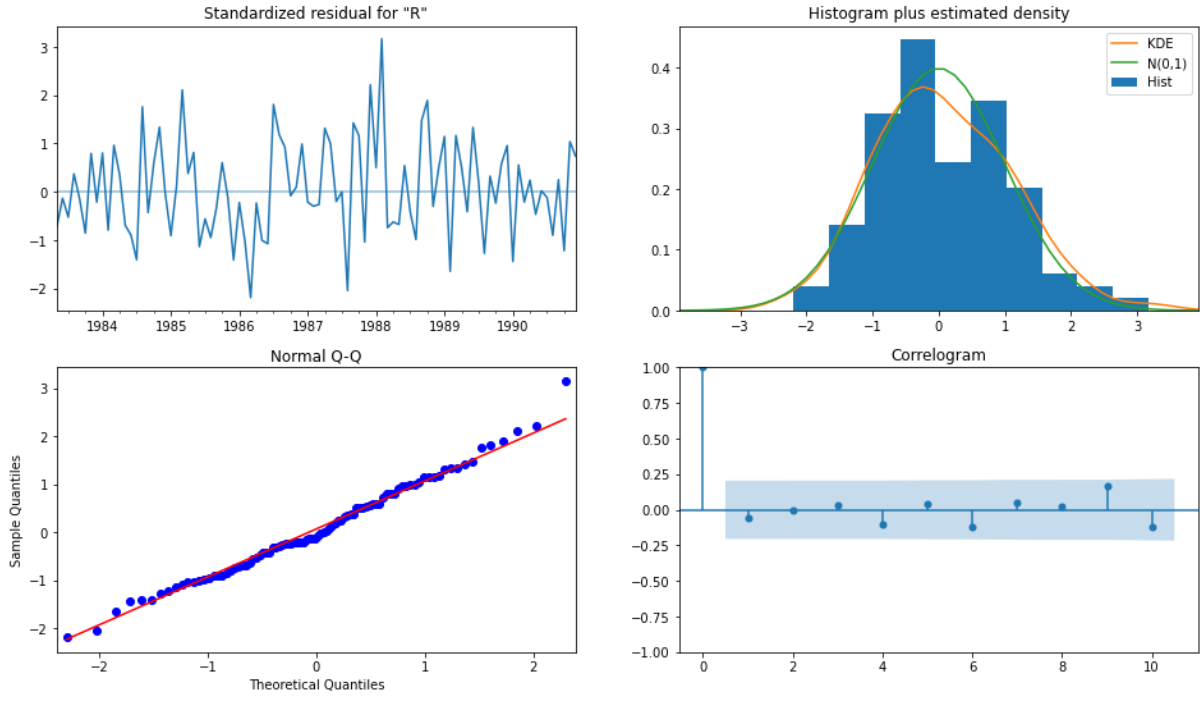
• We fit SARIMA models to each of these combinations and select with least AIC

• 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)**

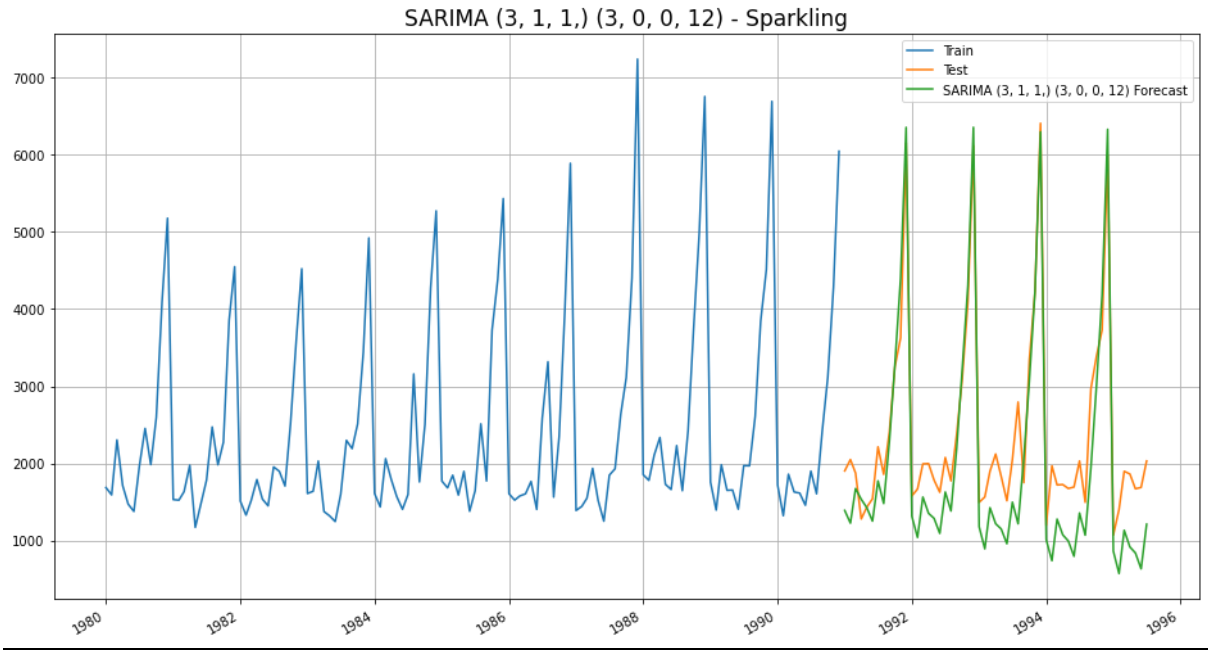


SARIMA (3, 1, 1) (3, 0, 2, 12) Diagnostic Plot - ROSE

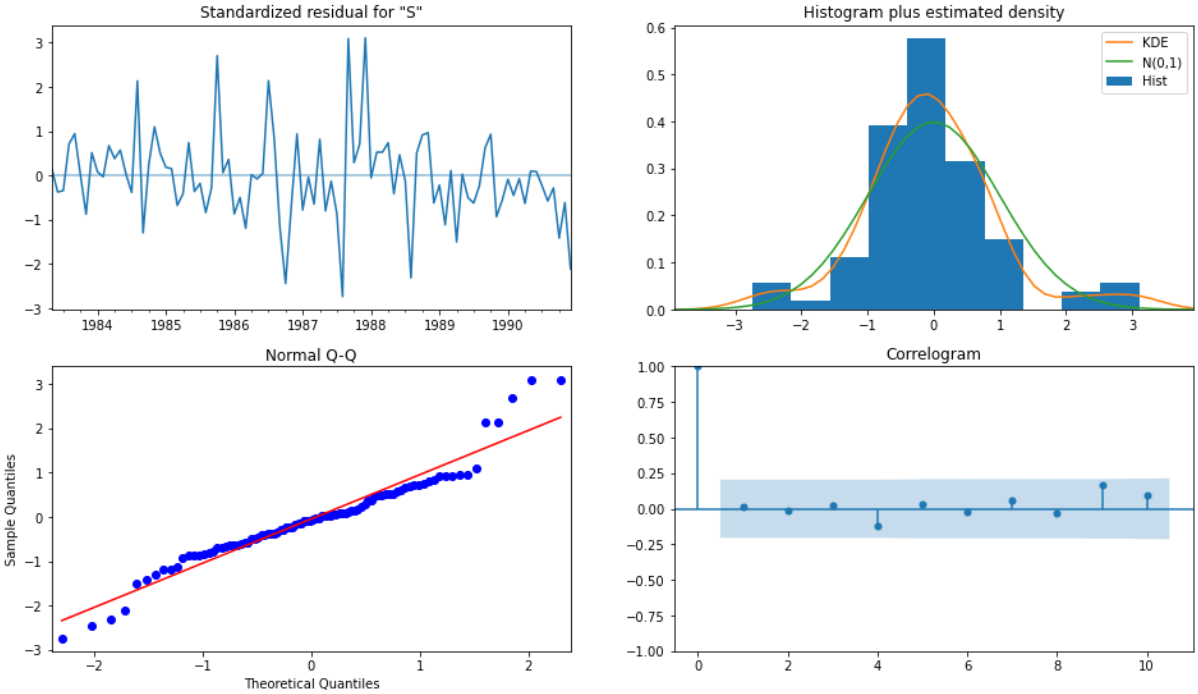


	Test RMSE Rose	Test MAPE Rose
ARIMA(2,1,3)	36.812984	75.838991
ARIMA(2,1,2)	36.871197	76.056213
SARIMA(3, 1, 1)(3, 0, 2, 12)	18.881910	36.375428

• 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



	RMSE	MAPE
ARIMA(2,1,2)	1299.979640	47.099986
ARIMA(0,1,0)	3864.279352	201.327650
SARIMA(3,1,1)(3,0,2,12)	601.140492	25.865406

• 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)

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.

Solution:

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

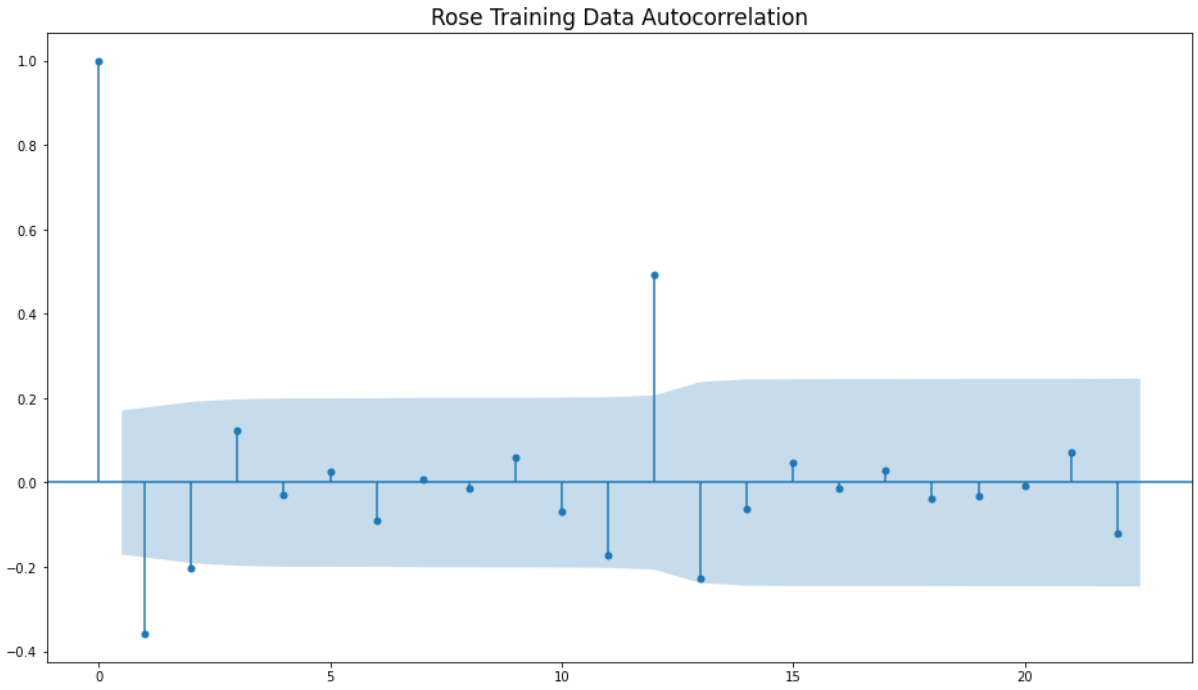
• PACF indicates the value of 'p' - which is the Auto-Regressive parameter in ARIMA / SARIMA models

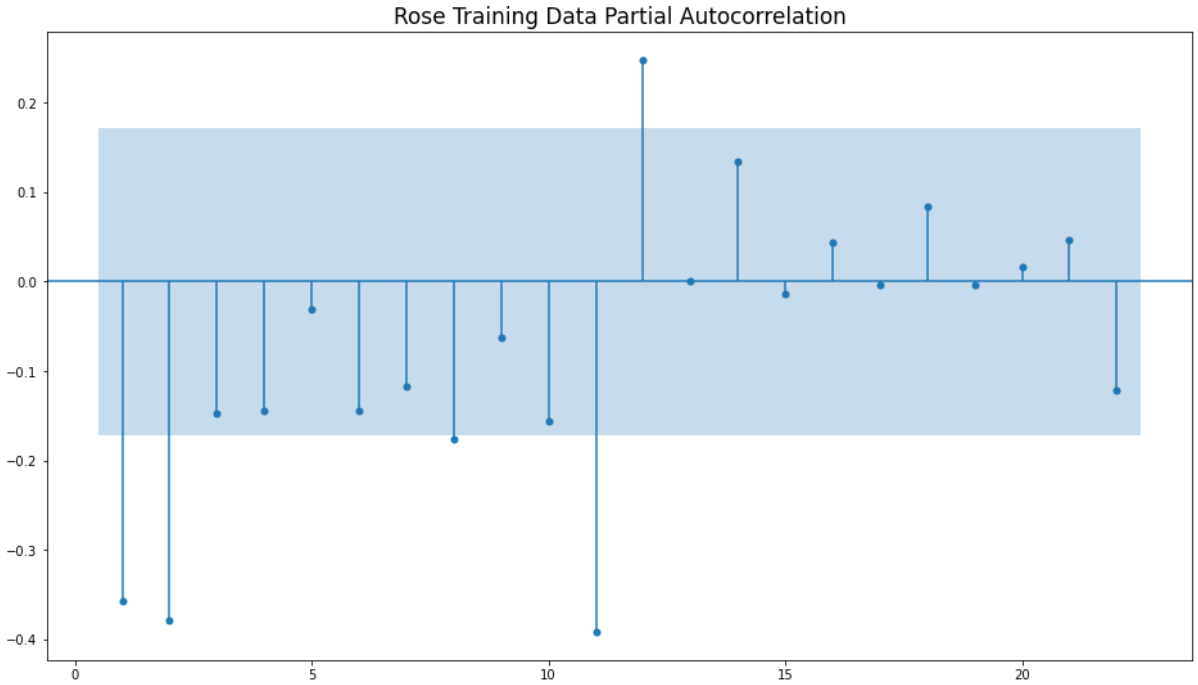
ACF & PACF of Rose –

• 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



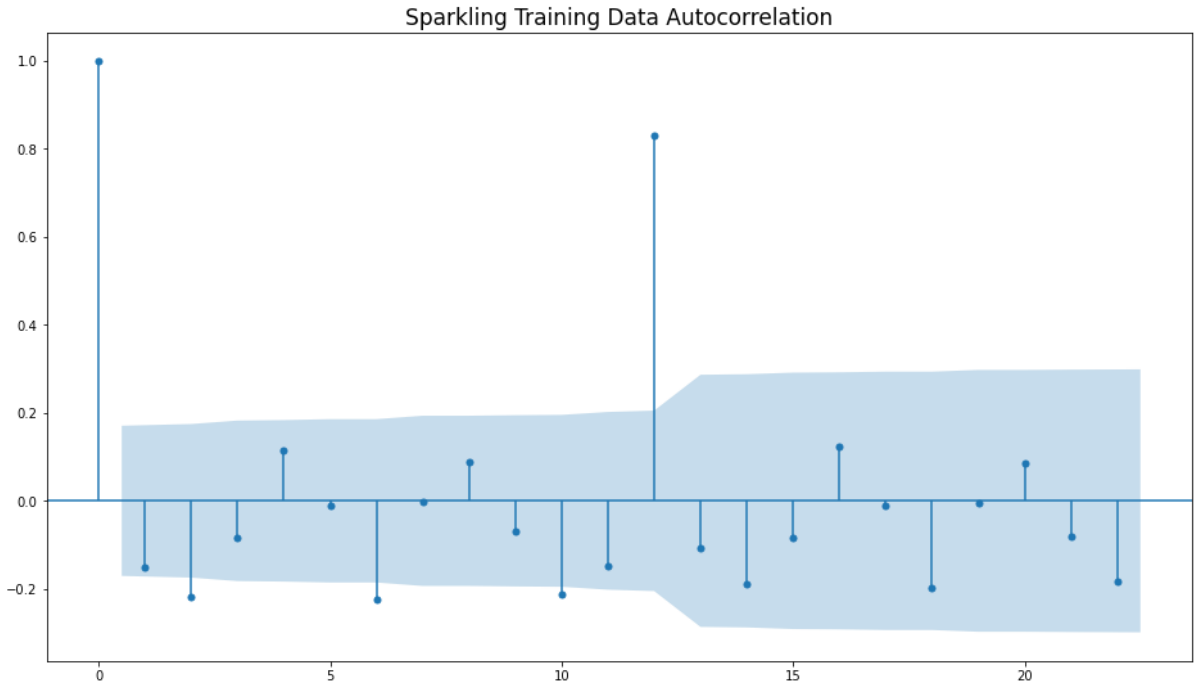


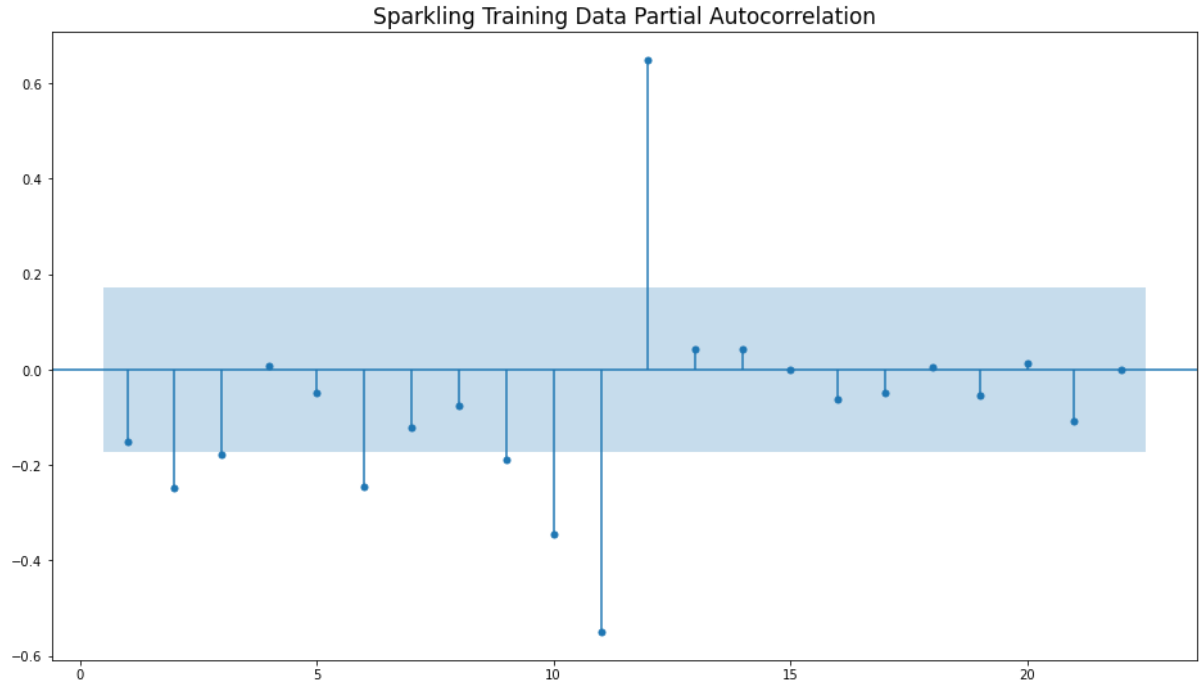
ACF & PACF of Sparkling –

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

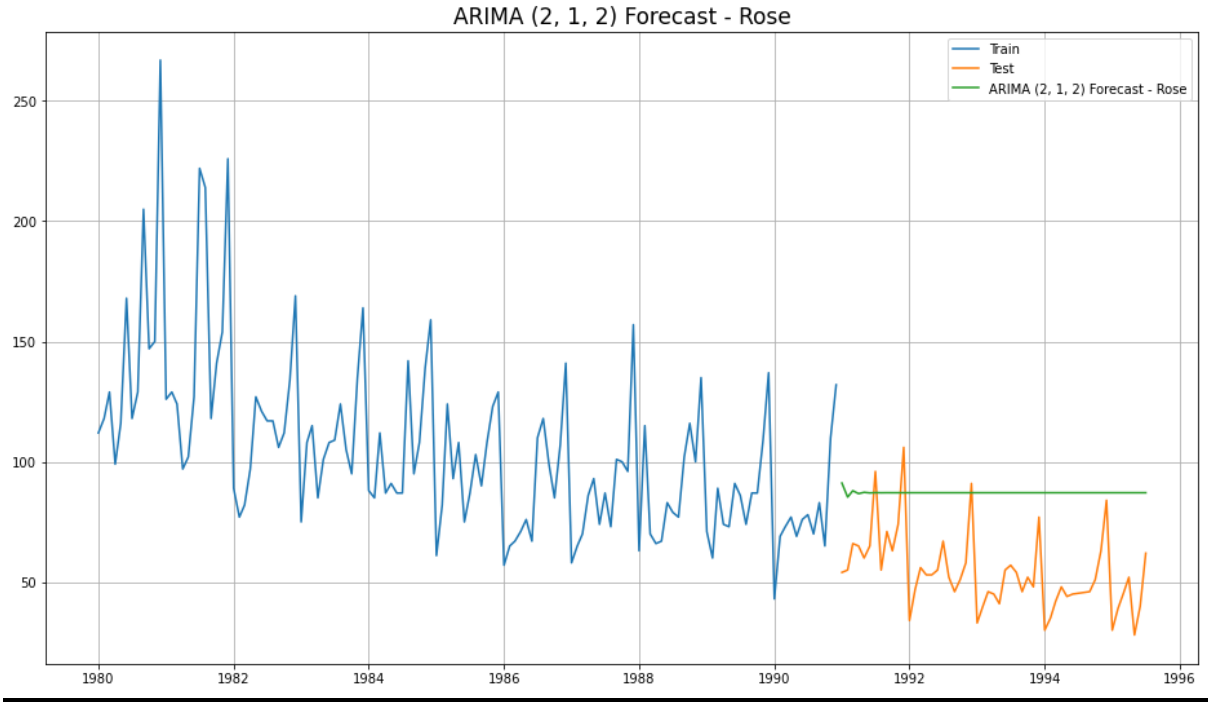
FOR ARIMA —-> p = 0, q = 0 and difference d = 1

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

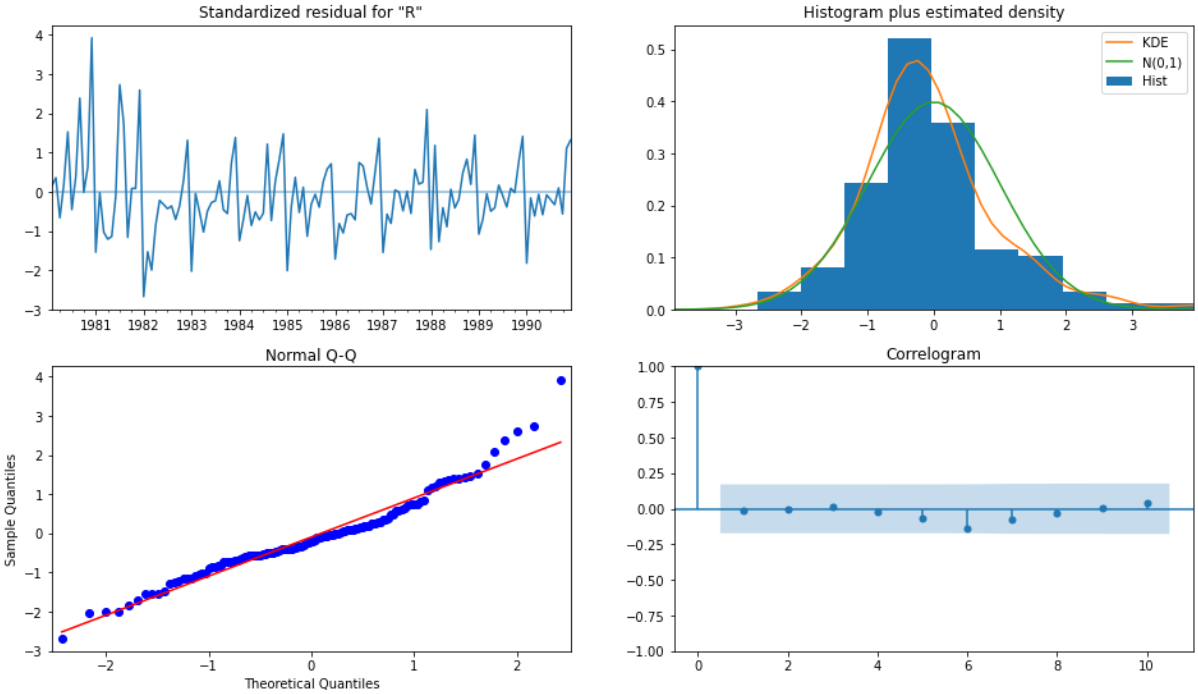




1. ARIMA Manual - Rose - (2, 1, 2)

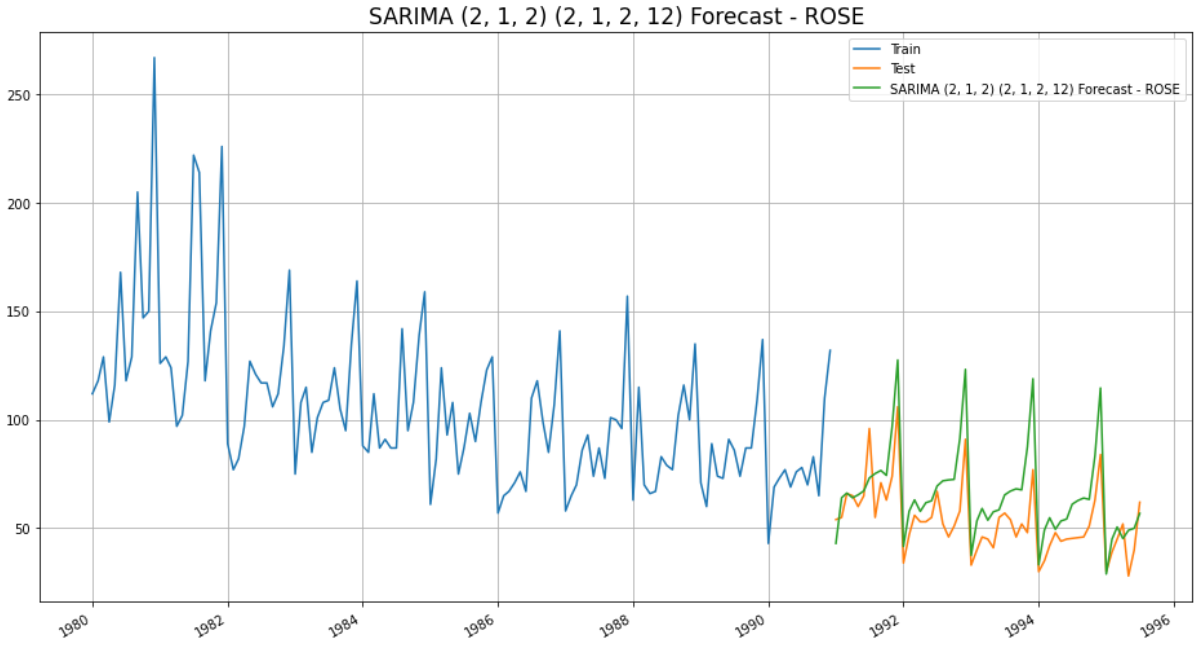


ARIMA (2, 1, 2) Diagnostic Plot – ROSE

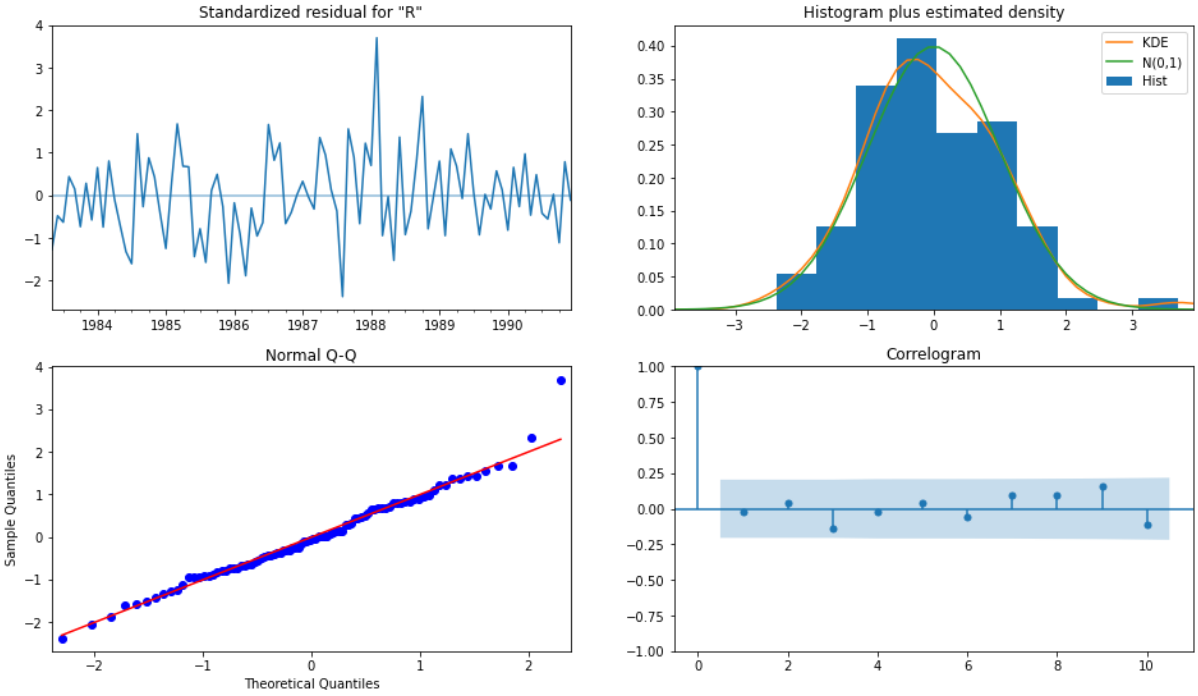


	Test RMSE Rose	Test MAPE Rose
ARIMA(2,1,3)	36.812984	75.838991
ARIMA(2,1,2)	36.871197	76.056213

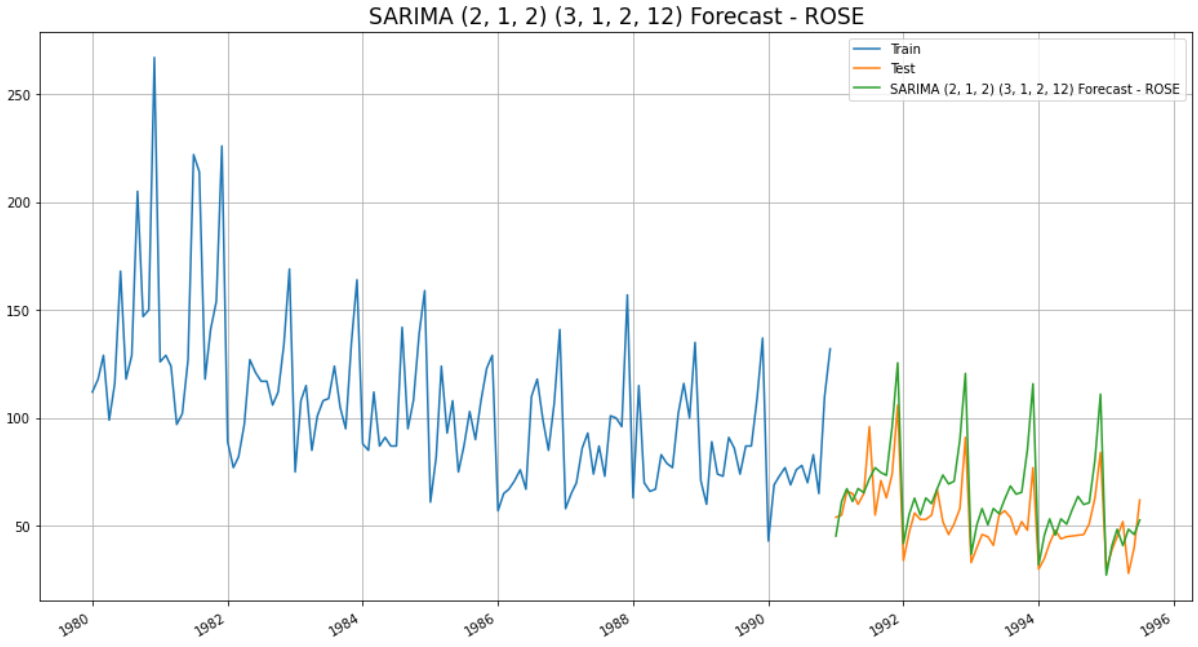
2. SARIMA Manual - Rose - (2, 1, 2) (2, 1, 2, 12)



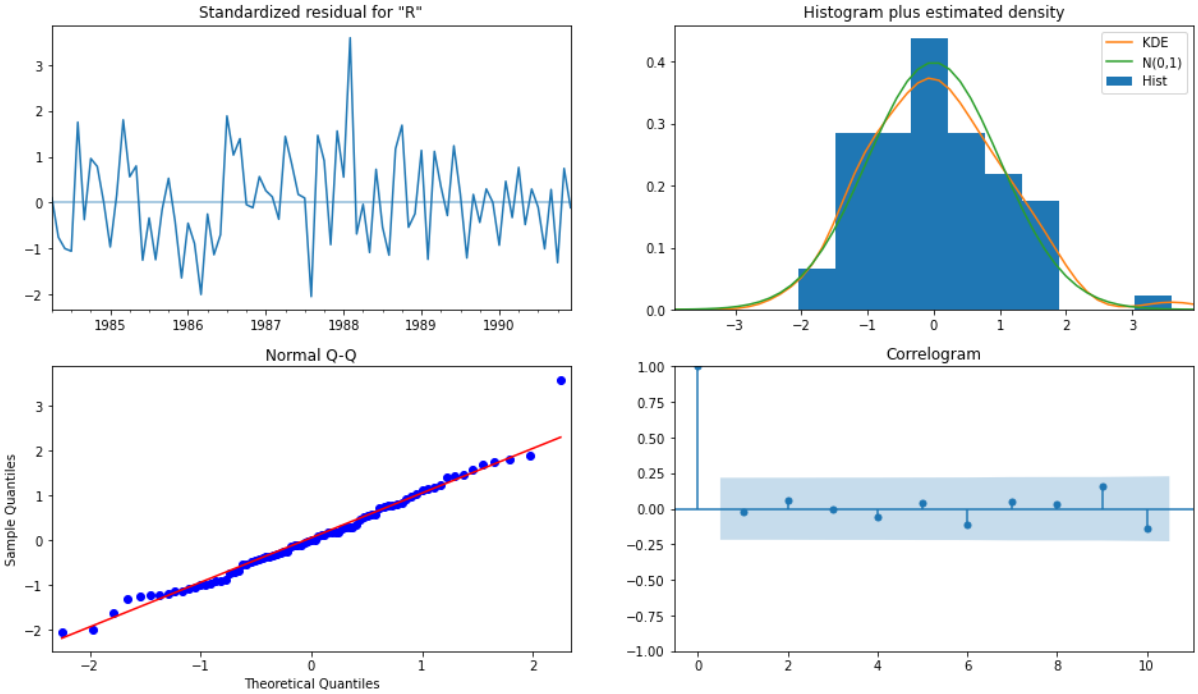
SARIMA (2, 1, 2) (2, 1, 2, 12) Diagnostic Plot – ROSE



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

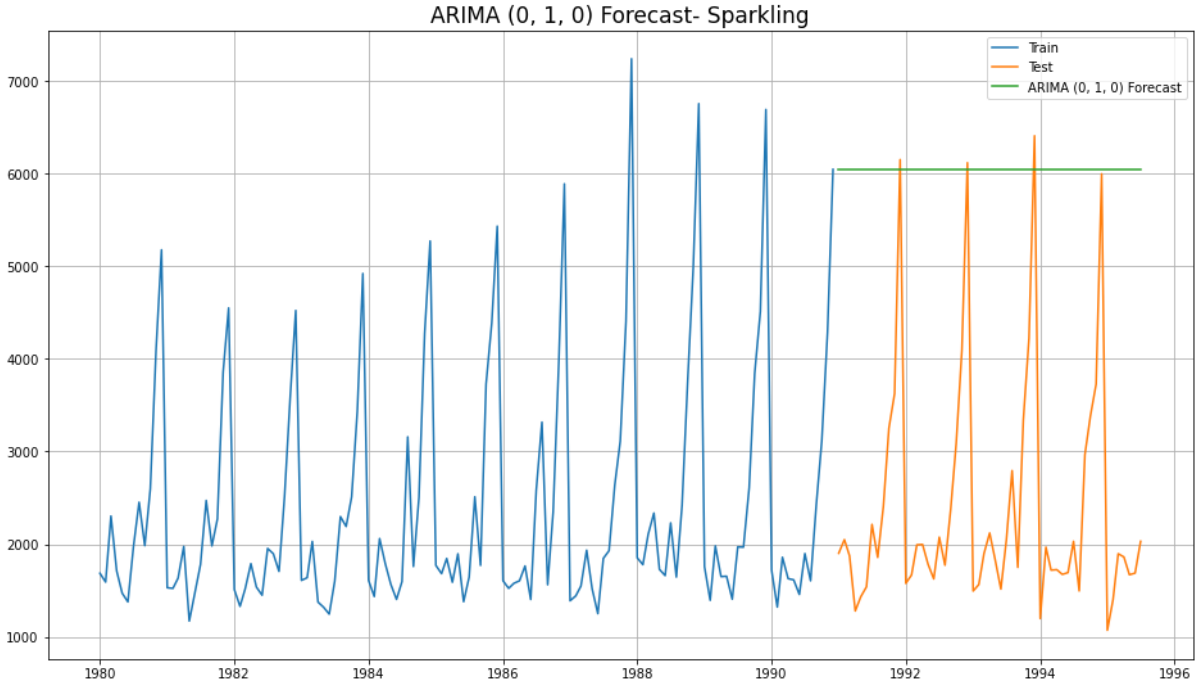


SARIMA (2, 1, 2) (3, 1, 2, 12) Diagnostic Plot – ROSE

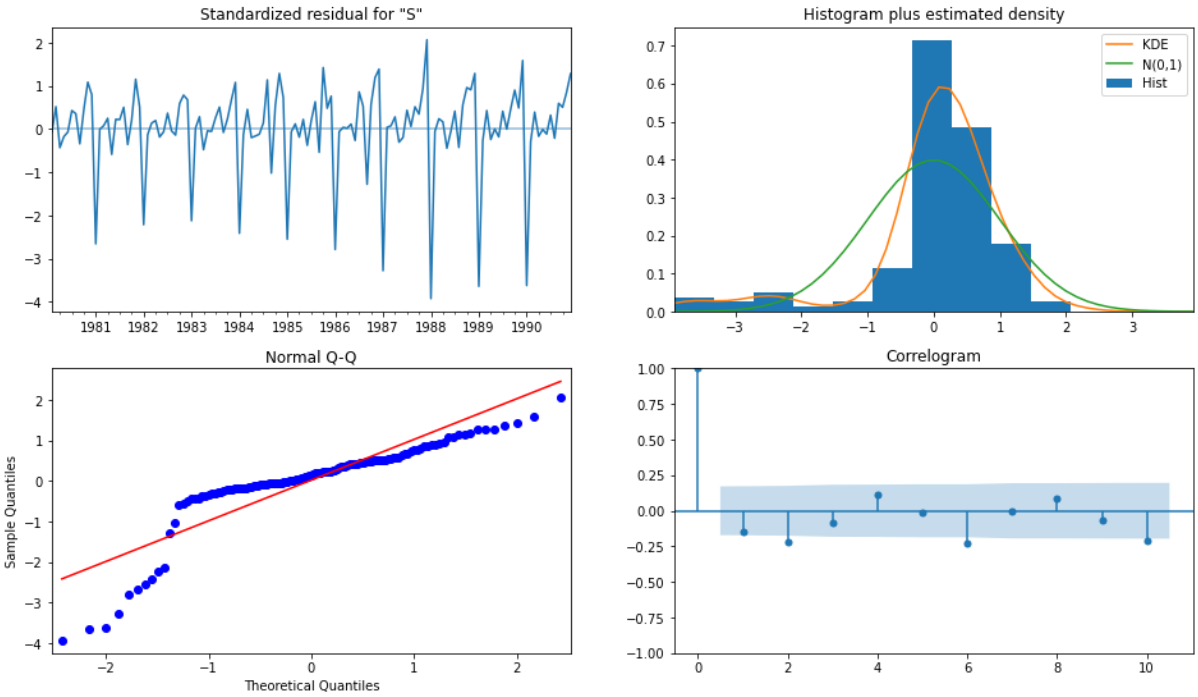


	Test RMSE Rose	Test MAPE Rose
ARIMA(2,1,3)	36.812984	75.838991
ARIMA(2,1,2)	36.871197	76.056213
SARIMA(3, 1, 1)(3, 0, 2, 12)	18.881910	36.375428
SARIMA(2,1,2)(3,1,2,12)	15.356802	22.954786

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

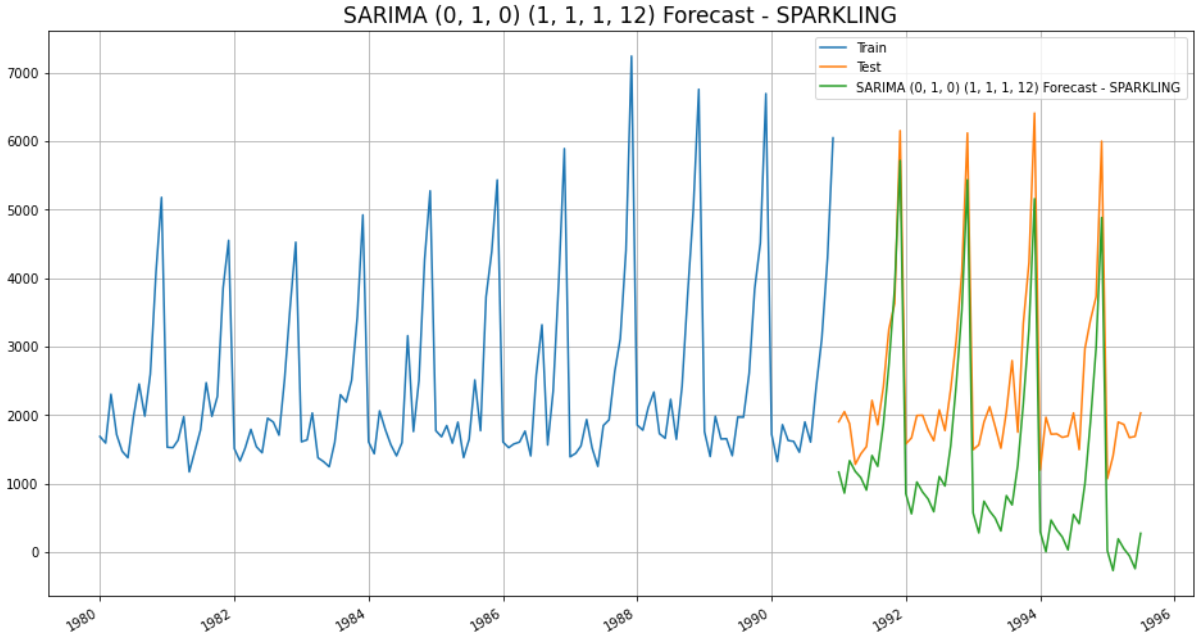


ARIMA (0, 1, 0) Diagnostic Plot – SPARKLING

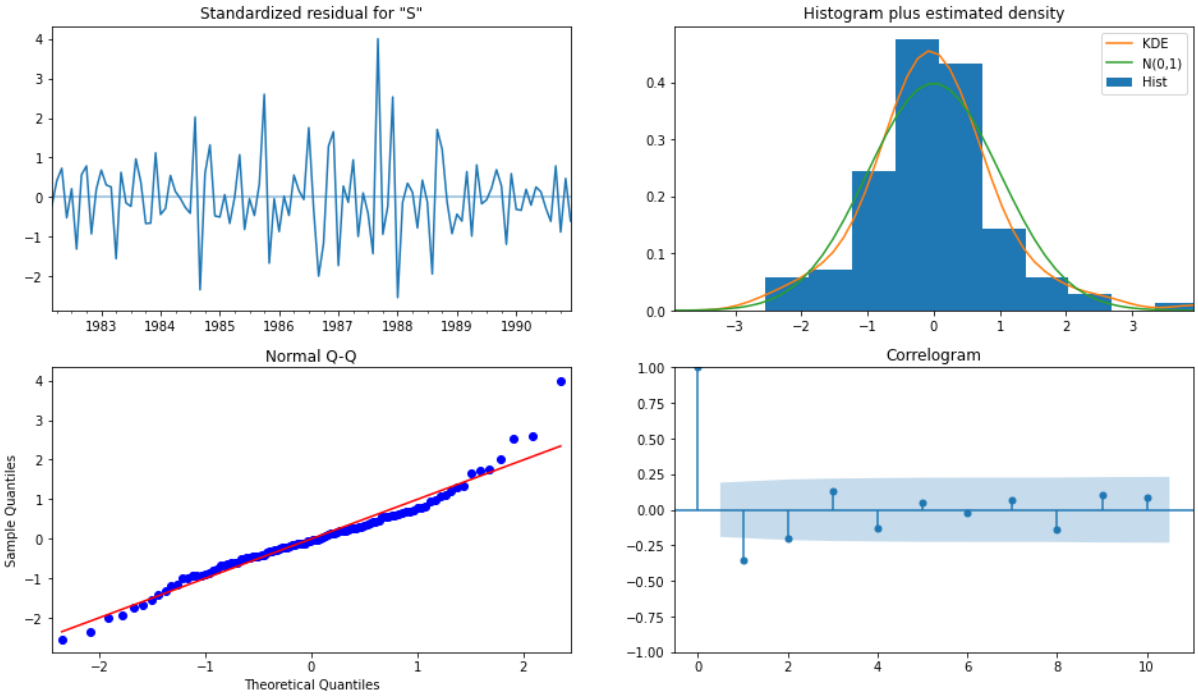


	RMSE	MAPE
ARIMA(2,1,2)	1299.979640	47.099986
ARIMA(0,1,0)	3864.279352	201.327650

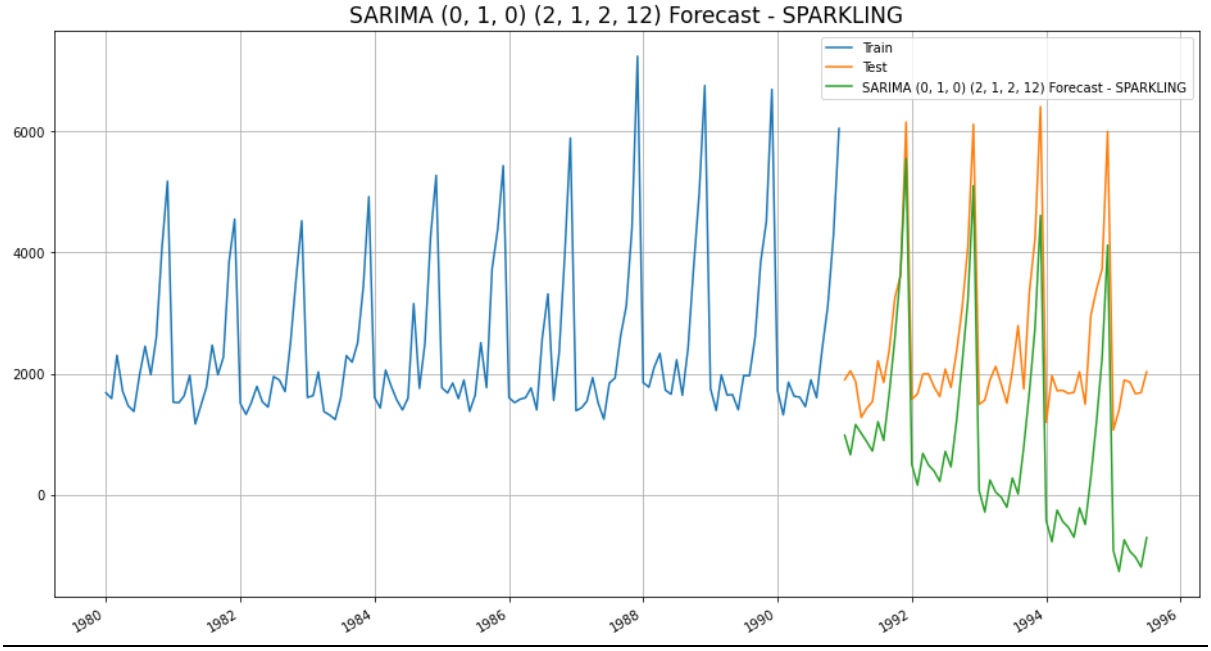
5. SARIMA Manual - Sparkling - (0, 1, 0) (1, 1, 1, 12)



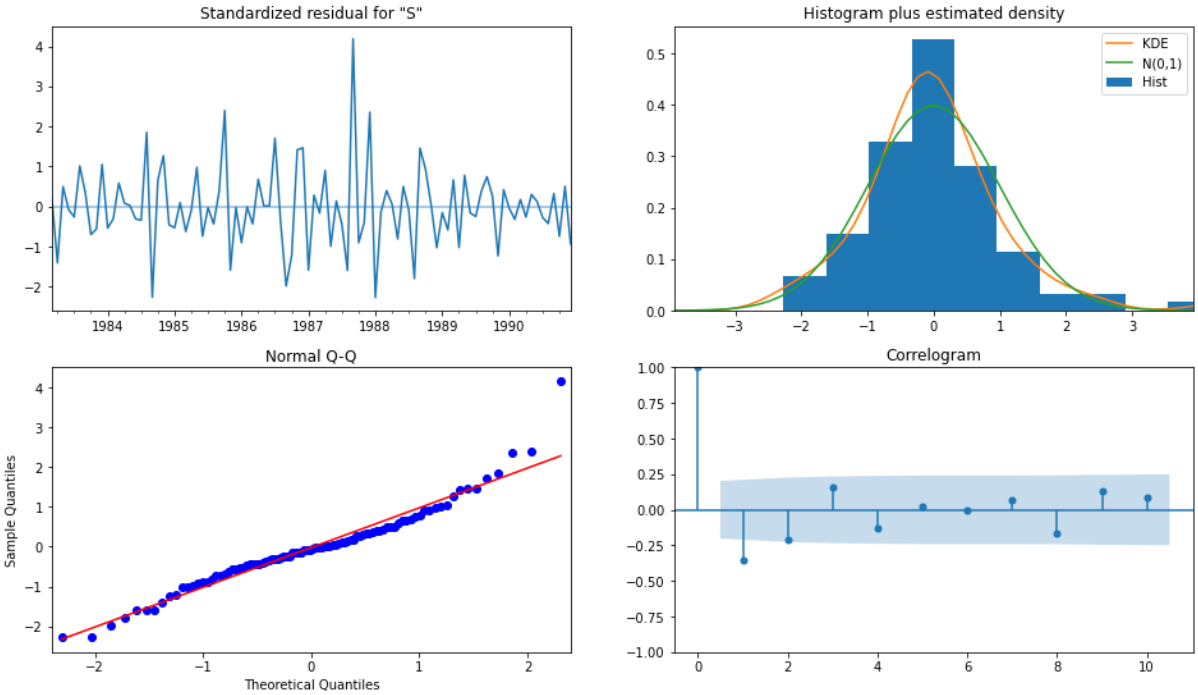
SARIMA (0, 1, 0) (1, 1, 1, 12 Diagnostic Plot – SPARKLING



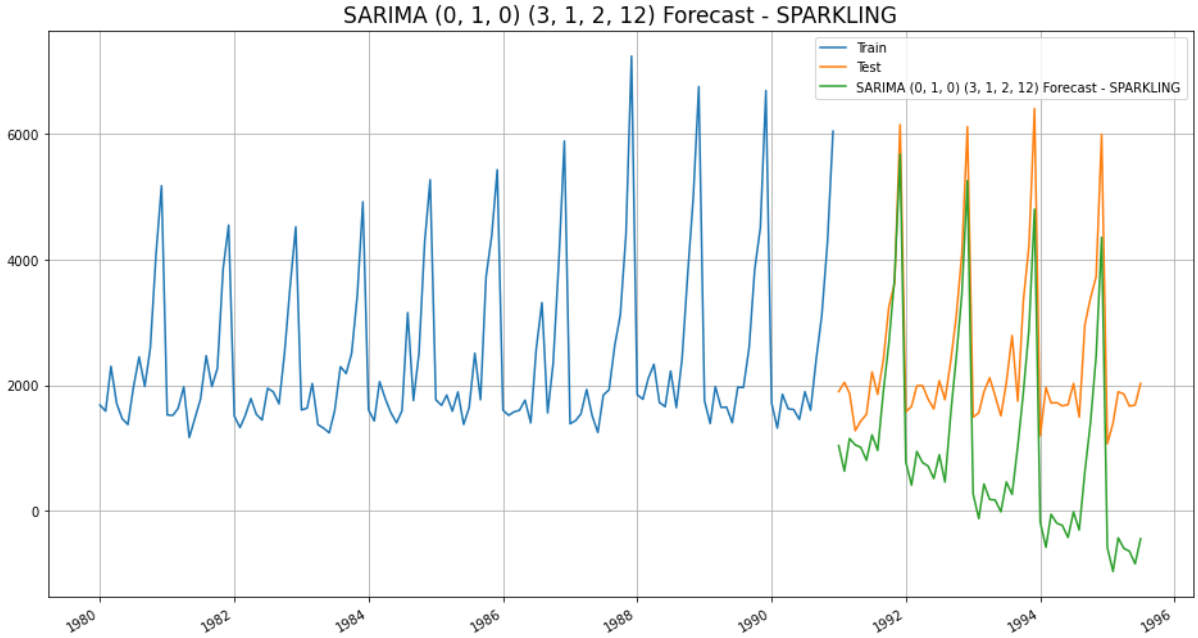
6. SARIMA Manual - Sparkling - (0, 1, 0) (2, 1, 2, 12)



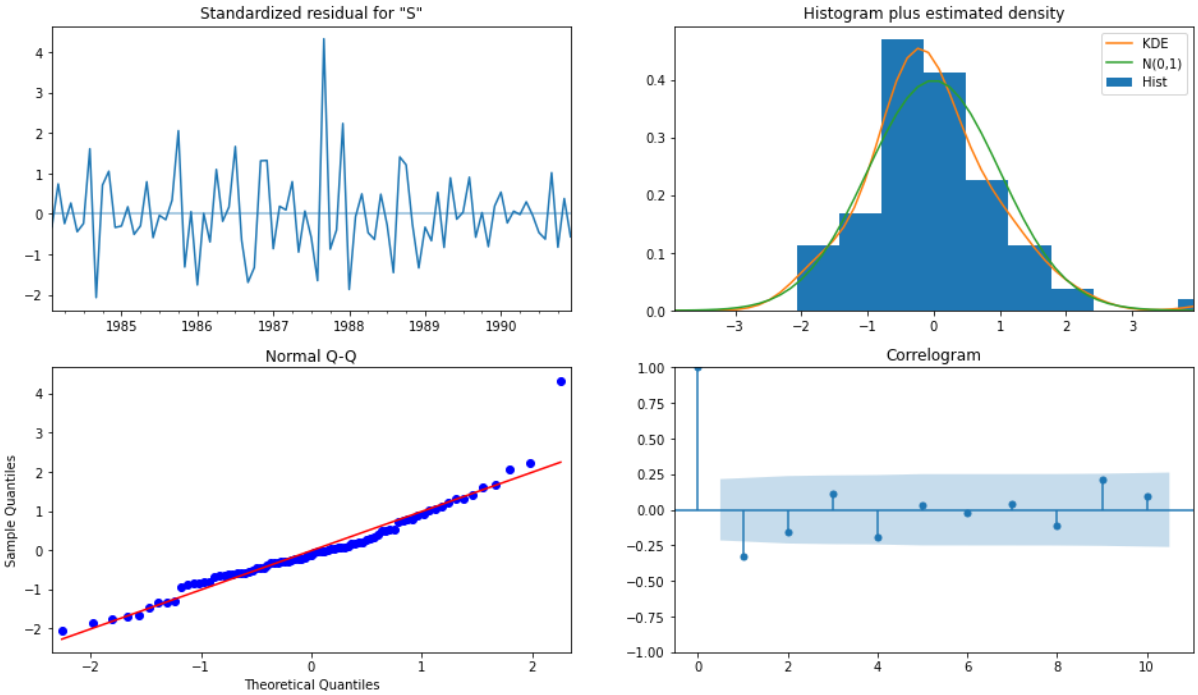
SARIMA (0, 1, 0) (2, 1, 2, 12 Diagnostic Plot – SPARKLING



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



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



	RMSE	MAPE
ARIMA(2,1,2)	1299.979640	47.099986
ARIMA(0,1,0)	3864.279352	201.327650
SARIMA(3,1,1)(3,0,2,12)	601.140492	25.865406
SARIMA(0,1,0)(3,1,2,12)	1189.835782	54.872536
SARIMA(0,1,0)(2,1,2,12)	1757.727286	81.785254
SARIMA(0,1,0)(3,1,2,12)	1551.645942	71.566181

• 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

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.

Solution:

Below is the required data frame

	Test RMSE Rose	Test RMSE Sparkling	RMSE	MAPE
RegressionOnTime	15.268955	1389.135175	NaN	NaN
NaiveModel	79.718773	3864.279352	NaN	NaN
SimpleAverageModel	53.460570	1275.081804	NaN	NaN
2pointTrailingMovingAverage	11.529278	813.400684	NaN	NaN
4pointTrailingMovingAverage	14.451403	1156.589694	NaN	NaN
6pointTrailingMovingAverage	14.566327	1283.927428	NaN	NaN
9pointTrailingMovingAverage	14.727630	1346.278315	NaN	NaN
Simple Exponential Smoothing	36.796241	1338.008384	NaN	NaN
Double Exponential Smoothing	15.268944	5291.879833	NaN	NaN
Triple Exponential Smoothing (Additive Season)	14.249661	378.951023	NaN	NaN
Triple Exponential Smoothing (Multiplicative Season)	20.156763	404.286809	NaN	NaN
Triple Exponential Smoothing (Additive Season, Damped Trend)	26.360083	378.951023	NaN	NaN
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)	25.955974	352.439828	NaN	NaN
ARIMA(2,1,2)	NaN	NaN	1299.979640	47.099986
ARIMA(0,1,0)	NaN	NaN	3864.279352	201.327650
SARIMA(3,1,1)(3,0,2,12)	NaN	NaN	601.140492	25.865406
SARIMA(0,1,0)(3,1,2,12)	NaN	NaN	1189.835782	54.872536
SARIMA(0,1,0)(2,1,2,12)	NaN	NaN	1757.727286	81.785254
SARIMA(0,1,0)(3,1,2,12)	NaN	NaN	1551.645942	71.566181

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.

Solution:

• 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 ——>

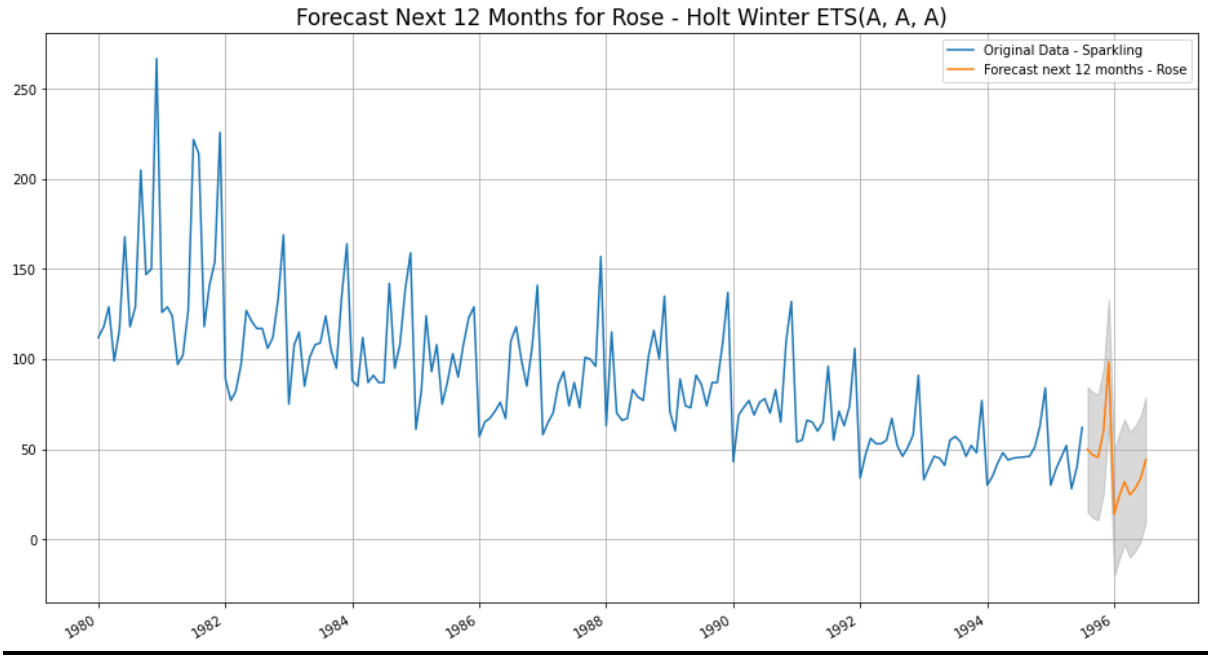
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)

Rose Forecast Next 12 months – 2 Pt Moving Average

• 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

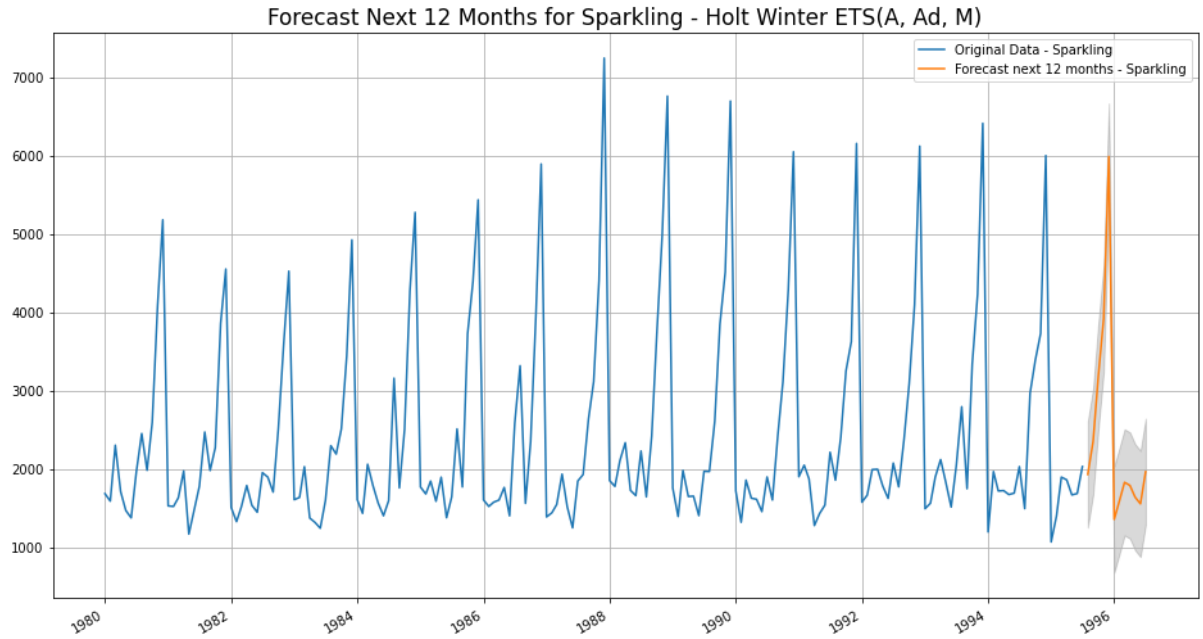
Rose Forecast Next 12 months - Triple Exponential Smoothing ETS (A, A, A)



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

```
1995-08-01    49.772739
1995-09-01    46.634326
1995-10-01    45.406017
1995-11-01    60.028624
1995-12-01    98.360619
1996-01-01    13.816110
1996-02-01    24.258995
1996-03-01    31.889828
1996-04-01    24.606456
1996-05-01    27.978689
1996-06-01    33.570506
1996-07-01    44.099788
Freq: MS, dtype: float64
```

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



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

```
1995-08-01    1931.548659
1995-09-01    2352.035451
1995-10-01    3179.715287
1995-11-01    3918.609828
1995-12-01    5986.825950
1996-01-01    1357.492648
1996-02-01    1599.117790
1996-03-01    1830.306505
1996-04-01    1790.955356
1996-05-01    1641.907867
1996-06-01    1556.364789
1996-07-01    1965.888644
Freq: MS, dtype: float64
```

10. COMMENT ON THE MODEL THUS BUILT AND REPORT YOUR FINDINGS AND SUGGEST THE MEASURES THAT THE COMPANY SHOULD BE TAKING FOR FUTURE SALES.

Solution:

• 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

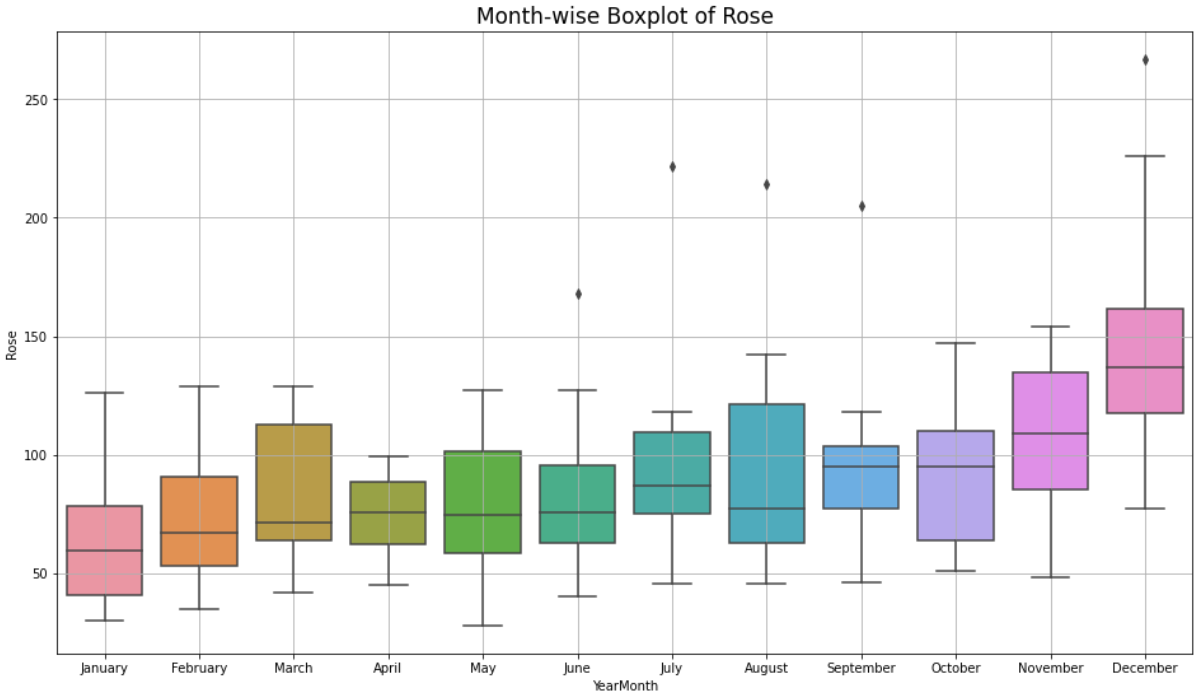
- This might be due to the Holiday season in this period

- 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

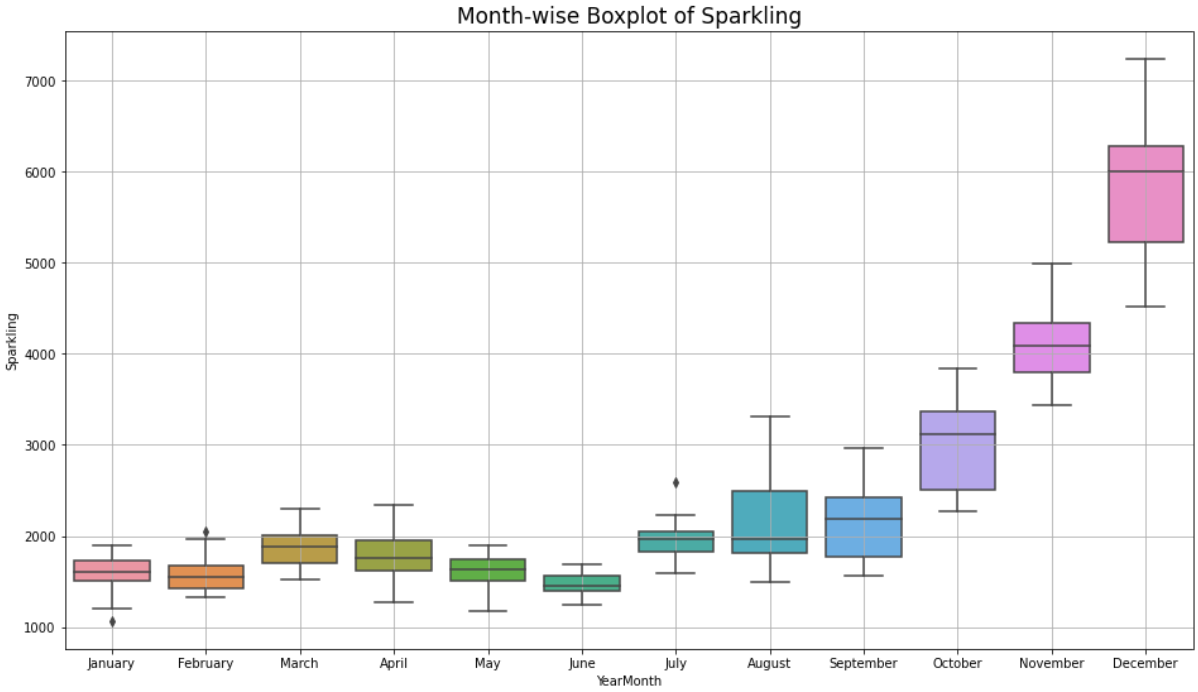
- This shows flat sales over long term range

• 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