BUSINESS REPORT

ABSTRACT

ABC Estate Wines, wants to analyze and forecast wine sales in the 20th century. They have provided datasets regarding two of their wine sales, Rose and Sparkling namely.

Yashveer Kothari. A
Post Graduate Programme in Data Science & Business Analytics

LIST OF TABLES

LIST OF FIGURES		
FIGURE#	FIGURE NAME	PAGE#
1	TIME SERIES PLOT	14
2	CHECKING FOR OUTLIERS	17
3	DISTRIBUTION PLOT	18
4	MONTHLY BOX PLOT	19
5	YEARLY BOXPLOT	21
6	DECOMPOSITION ADDITIVE - ROSE	22
7	DECOMPOSITION ADDITIVE MODEL - SPARKLING	23
8	DECOMPOSITION MULTIPLICATIVE MODEL - ROSE	25
9	DECOMPOSITION MULTIPLICATIVE MODEL - SPARKLING	26
10	TRAIN AND TEST SPLIT PLOT -ROSE	30
11	TRAIN AND TEST SPLIT PLOT -SPARKLING	30
12	LINEAR REGRESSION PLOT	31
13	NAÏVE APPROACH PLOT	33
14	SIMPLE AVERAGE PLOT	34
15	MOVING AVERAGE PLOT - ROSE	36
16	MOVING AVERAGE PLOT - SPARKLING	37
17	MODEL COMPARISION	38
18	SINGLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS - ETS (A, N, N)	40
19	DOUBLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS - ETS (A, A, N)	42
20	TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS, ADDITIVE SEASONALITY - ETS (A, A, A)	44
21	TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS, MULTIPLICATIVE SEASONALITY - ETS (A, A, M)	47
22	TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE DAMPED TRENDS, ADDITIVE SEASONALITY - ETS (A, AD, A)	48

ABOUT FORECASTING

- Time series forecasting is the process of analysing time series data using statistics and modelling to make predictions and inform strategic decision-making.
- Forecasting is required to reduce the adverse risk which may occur in the future by the company/organisation.
- Regression, Data mining and Times series are methods of forecasting

ABOUT TIME SERIES

- A Time series is a sequence of measurements on the same variable collected over time. The measurements are made at regular time intervals.
- The regular intervals are Yearly, Quarterly, Monthly, Weekly, Daily and Hourly.
- Data collected on multiple items at the same point of time is not a time series.
- When time periods are not the same.
- In times series the data is not independent and Ordering of the data in the dataset is very important because there is dependency and may change the data structure.
- Time series does not accept missing data therefore such data has to be imputed and the dataset should be continuous.
- There are three components of a time series:
    1. **Trend**: It is a pattern in data that shows the movement of a series to relatively higher or lower values over a long period of time.
    2. **Seasonality**: Seasonality is the relative increase or decrease of sales (demand or consumption) every period (quarter or month) compared to the yearly average
    3. **Irregular component (Error)**: The error or variability associated with the series is the Irregular component, this component is a random component. The part of the series that cannot be explained through Systematic component forms the Irregular Component Other names of this component is Error or White Noise. This component is assumed to have a normal distribution with 0 mean and constant variance $\sigma 2$.

ABOUT DECOMPOSITION

- Decomposition is a forecasting technique that separates or decomposes historical data into different components and uses them to create a forecast that is more accurate than a simple trend line.
- Decomposition is done for the following reasons:

1. To understand revenue generation without the quarterly effects
    i. De-seasonalize the series
    ii. Estimate and adjust by seasonality
2. Compare the long-term movement of the series (Trend) vis-a-vis short-term movement seasonality) to understand which has the higher influence
3. If revenue for multiple sectors is to be compared and if the sectors show non-uniform seasonality, de-seasonalized series needs to be compared
4. Additive model: Observation = Trend + Seasonality + Error

$$Yt = Tt + St + It$$

5. Multiplicative model: Observation = Trend * Seasonality * Error

$$Yt = Tt * St * It$$

Y_t: time series value (actual data) at period t.

S_t: seasonal component (index) at period t.

T_t: trend cycle component at period t.

I_t: irregular (remainder) component at period t.

TYPES OF FORECASTING MODELS

- Naïve forecast: Estimating technique in which the last period's actuals are used as this period's forecast, without adjusting them or attempting to establish causal factors. It is used only for comparison with the forecasts generated by the better (sophisticated) techniques.

- Moving average forecast: The moving average is a statistical method used for forecasting long-term trends. The technique represents taking an average of a set of numbers in a given range while moving the range.

EXPONENTIAL SMOOTHENING FORECAST

- Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component.
- It is a powerful forecasting method that may be used as an alternative to the popular box-Jenkins ARIMA family of methods.
- Here, only the recent observations matter
- Weights decay as the observations get older, one or more parameters control how the weights decay and the parameters have values between 0 and 1.
- If α is closer to 1, forecasts follow the actual observations more closely. If α is closer to 0, forecasts are farther from the actual observations and the line is smooth.

SIMPLE EXPONENTIAL: If the time series neither has a pronounced trend nor seasonality

DOUBLE EXPONENTIAL

- Applicable when data has Trend but no seasonality
- An extension of SES
- Two separate components are considered: Level and Trend
- Level is the local mean
- One smoothing parameter αcorresponds to the level series
- A second smoothing parameter βcorresponds to the trend series
- Also known as Holt.

TRIPLE EXPONENTIAL

- Triple exponential smoothing is used to handle the time series data containing a seasonal component.
- This method is based on three smoothing equations: stationary component, trend, and seasonal.
- Both seasonal and trend can be additive or multiplicative.
- Also known as Holt-Winters Model.

AUTO REGRESSIVE INTEGRATED MOVING AVERAGE

- Auto-regression means regression of a variable on itself.
- One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process.
- When the time series data is not stationary, then we have to convert the non-stationary time-series data to stationary time-series before applying AR models
- **STATIONARITY**: A Time Series is considered to be stationary whose statistical properties such as the variance and (auto) correlation are all constant over time. The properties of a stationary time series do not depend on time. The (auto) correlation observations only depend on how far apart these observations are and not where they are.
- **Augmented Dickey Fuller Test (ADF)**:

1. To check whether the series is stationary, we use the Augmented Dickey Fuller (ADF) test whose null and alternate hypothesis can be simplified to
2. Null Hypothesis H0: Time Series is non-stationary
3. Alternate Hypothesis Ha: Time Series is stationary
4. At our desired level of significance (chosen alpha value), we can test for stationarity using the ADF test.

- We look at the Auto-Correlations of a stationary Time Series to understand the order of a Moving-Average models.
- For a MA model, the ACF (Auto Correlation Function) values cut-off at a certain lag. The ACF values closes to 0 (at appropriate confidence intervals for the ACF plots) beyond that order (or lag).

- Auto-correlation of lag k, $\boldsymbol{\rho\rho}_{qk}$, is the correlation between $\boldsymbol{YY_{tt}}$and $\boldsymbol{YY_{tt-kk}}$. This particular function does not depend on 't' since the Time Series is stationary.
- For building MA models, we look at the ACF plots and determine the order of the MA model.

PARTIAL AUTO CORRELATION

- We look at the Partial Auto-Correlations of a stationary Time Series to understand the order of an Auto-Regressive models
- For an AR model, the PACF (Partial Auto Correlation Function) values cuts-off after a certain lag. The PACF values closes to 0 (at appropriate confidence intervals for the PACF plots) beyond that order (or lag).
- Partial auto-correlation of lag k, $\boldsymbol{\rho\rho ppkk}$, is the correlation between $\boldsymbol{YYtt}$and $\boldsymbol{YYtt-kk}$when the influence of all intermediate values ($\boldsymbol{YYtt-11}$, $\boldsymbol{YYtt-22}$, ⋯, $\boldsymbol{YYtt-kk+11}$) is removed from both $\boldsymbol{YYttaannddYYtt-kk}$
- For building AR models, we look at the PACF plots and determine the order of the AR model.

ARIMA (p, d, q) MODEL

- An ARIMA model consists of the Auto-Regressive (AR) part and the Moving Average (MA) part after we have made the Time Series stationary by taking the correct degree/order of differencing.
- The AR order is selected by looking at where the PACF plot cuts-off (for appropriate confidence interval bands) and the MA order is selected by looking at where the ACF plots cuts-off (for appropriate confidence interval bands)
- The correct degree or order of difference gives us the value of 'd' while the 'p' value is for the order of the AR model and the 'q' value is for the order of the MA model.
- This is the Box-Jenkins methodology for building the ARIMA models.
- ARIMA models can be built keeping the Akaike Information Criterion (AIC) in mind as well. In this case, we choose the 'p' and 'q' values to determine the AR and MA orders respectively which gives us the lowest AIC value. Lower the AIC better is the model.

- Coding languages tries different orders of ‘p’ and ‘q’ to arrive to this conclusion. Remember, even for such a way of choosing the ‘p’ and ‘q’ values, we must make sure that the series is stationary.
- The formula for calculating the AIC is 2k – 2ln(L), where k is the number of parameters to be estimated and L is the likelihood.

SEASONAL ARIMA (p, d, q) (P, D, Q) MODEL

- For a Seasonal Auto-Regressive Integrated Moving Average we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d) and seasonal differencing (D). Here, the ‘F’ parameter indicates the seasonality/seasonal effects over a particular period.
- We can follow the Box-Jenkins method over here as well to decide the ‘p’, ‘q’, ‘P’ and ‘Q’ values
- For deciding the ‘P’ and ‘Q’ values, we need to look at the PACF and the ACF plots respectively at lags which are the multiple of ‘F’ and see where these cut-offs (for appropriate confidence interval bands).
- For the SARIMA models, we can also estimate ‘p’, ‘q’, ‘P’ and ‘Q’ by looking at the lowest AIC values.
- The seasonal parameter ‘F’ can be determined by looking at the ACF plots. The ACF plot is expected to show a spike at multiples of ‘F’ thereby indicating a presence of seasonality.
- Also, for Seasonal models, the ACF and the PACF plots are going to behave a bit different and they will not always continue to decay as the number of lags increase.

SARIMAX AND TVLM MODELS

- For (S)ARIMAX (X stands for exogenous variables) models, we can include Exogenous variables as well. These are variables which affect the target stationary Time Series. These variables can also be included in our models to aid in the decrease of forecast errors.
- There are Time Varying Linear Models as well in which the exogenous variables are also allowed to look at their past observations for forecasting into the future.

INTRODUCTION

The dataset consists sales data regarding two wines Rose and Sparkling namely. The forecasting regarding the future sales of the wines has to be done along with the exploratory data analysis of the same. The main aim of this problem is to forecast the sale using various forecasting methods mentioned above.

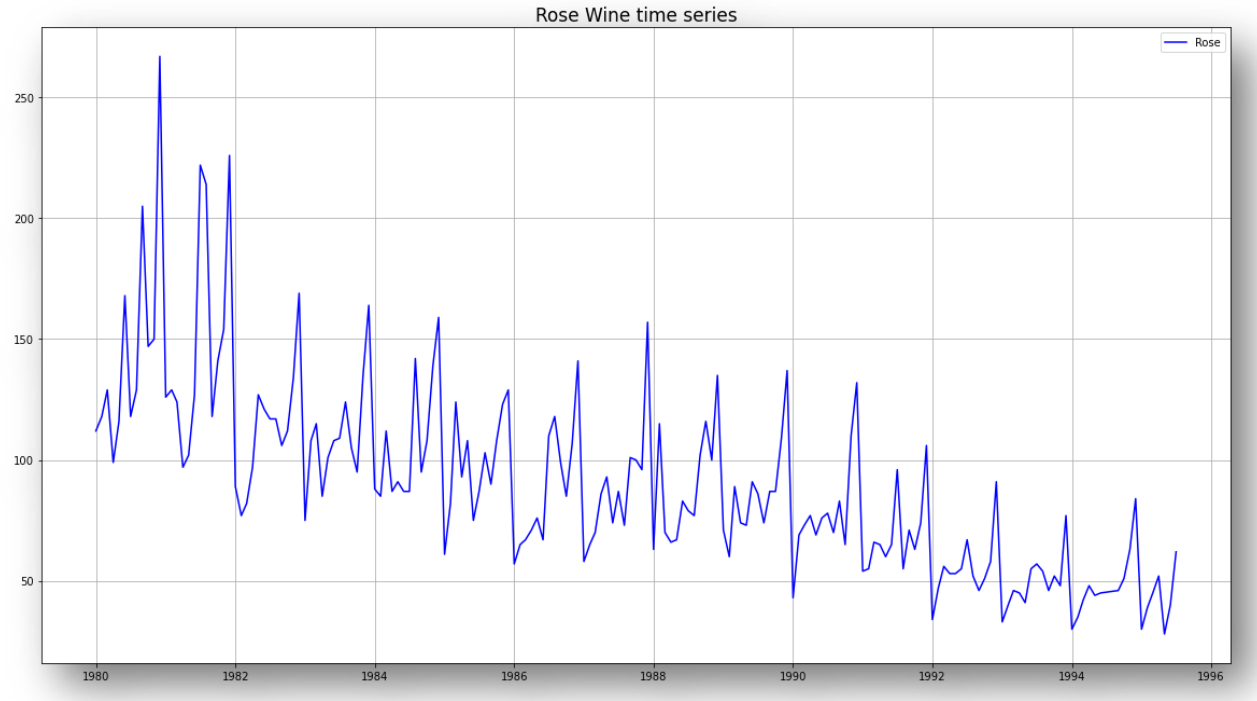
Q.1. Read the data as an appropriate Time Series data and plot the data.

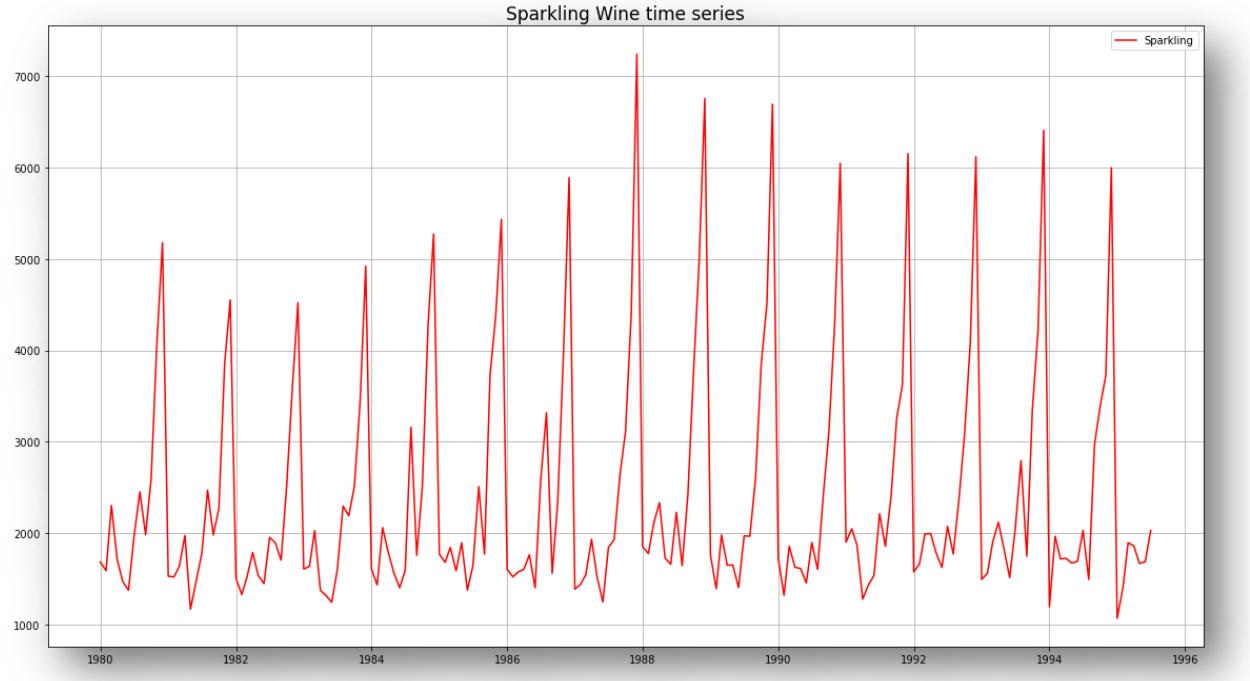
TABLE 1: TOP 5 DATASET SAMPLES

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

FIGURE 1: TIME SERIES PLOT





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

TABLE 2: DATASET INFORMATION

```
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
```

```
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)
```

TABLE 3: MISSING VALUES

```
Rose    2
dtype: int64
```

- There are 2 missing values in the Rose Dataset

```
Sparkling    0
dtype: int64
```

- There are no missing values in the sparkling dataset

TABLE 4: INTERPOLATING THE MISSING VALUES

- **Interpolation** is a technique used to estimate unknown data points between two known data points. Interpolation is mostly used to impute missing values in the data frame or series while pre-processing data.

```
Rose    0
dtype: int64
```

- After, interpolation there are no missing values in the dataset Rose.

TABLE 5: DATASET DESCRIPTION

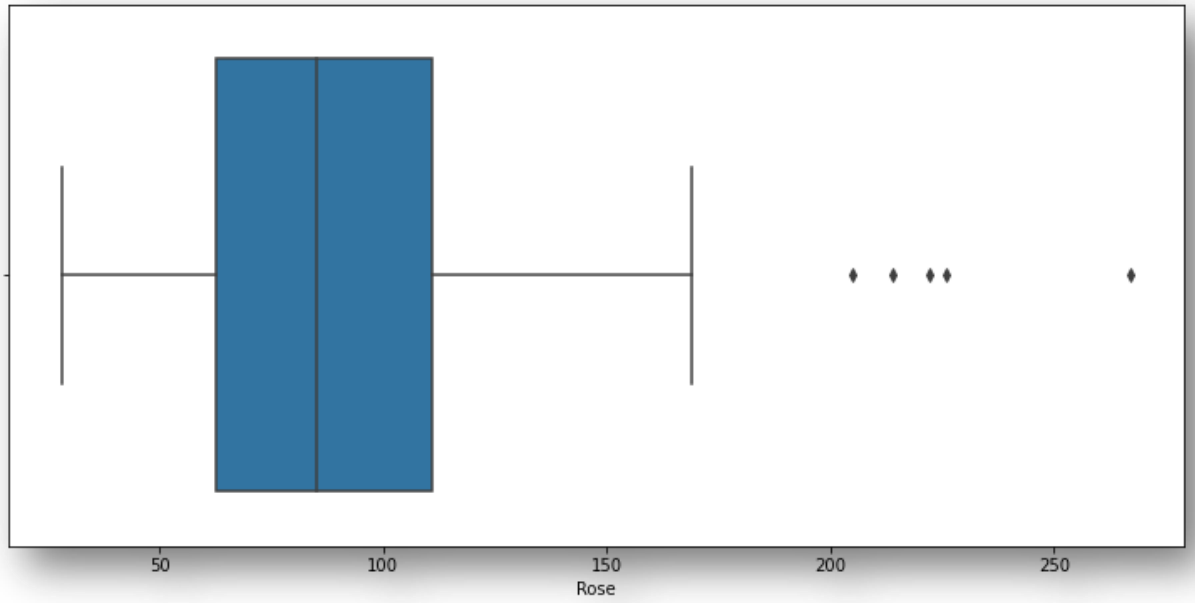
	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.914439	39.238325	28.0	62.5	85.0	111.0	267.0

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

TABLE 6: EXTRACTING DUPLICATES

- There are a total of 89 duplicate values in the dataset Rose.
- There are a total of 11 duplicate values in the dataset Sparkling.

FIGURE 2: CHECKING FOR OUTLIERS



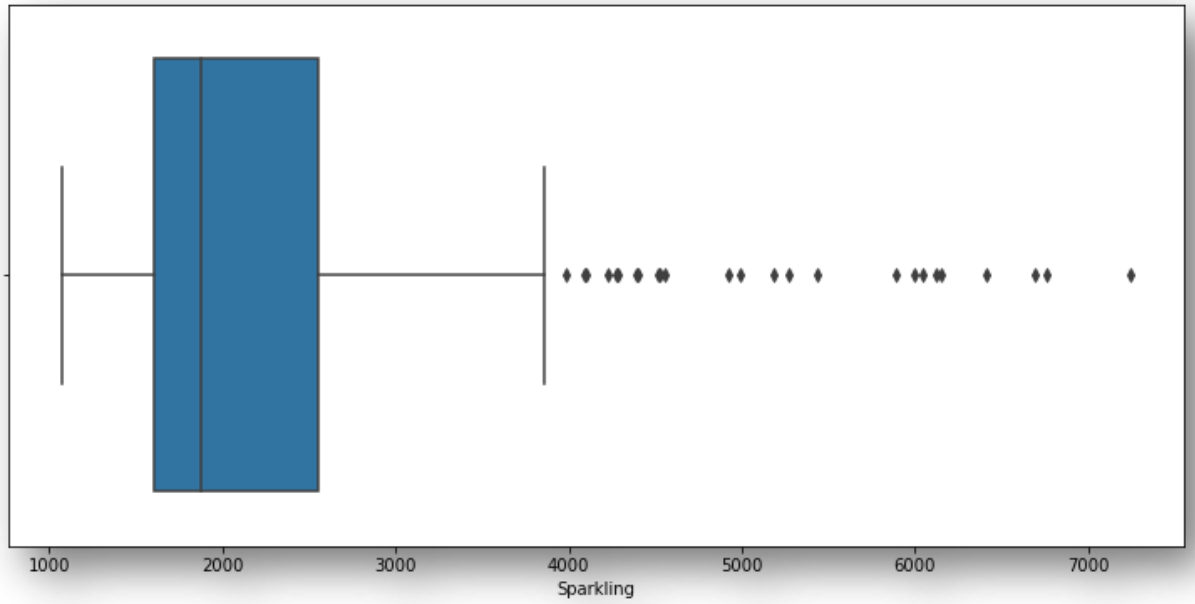
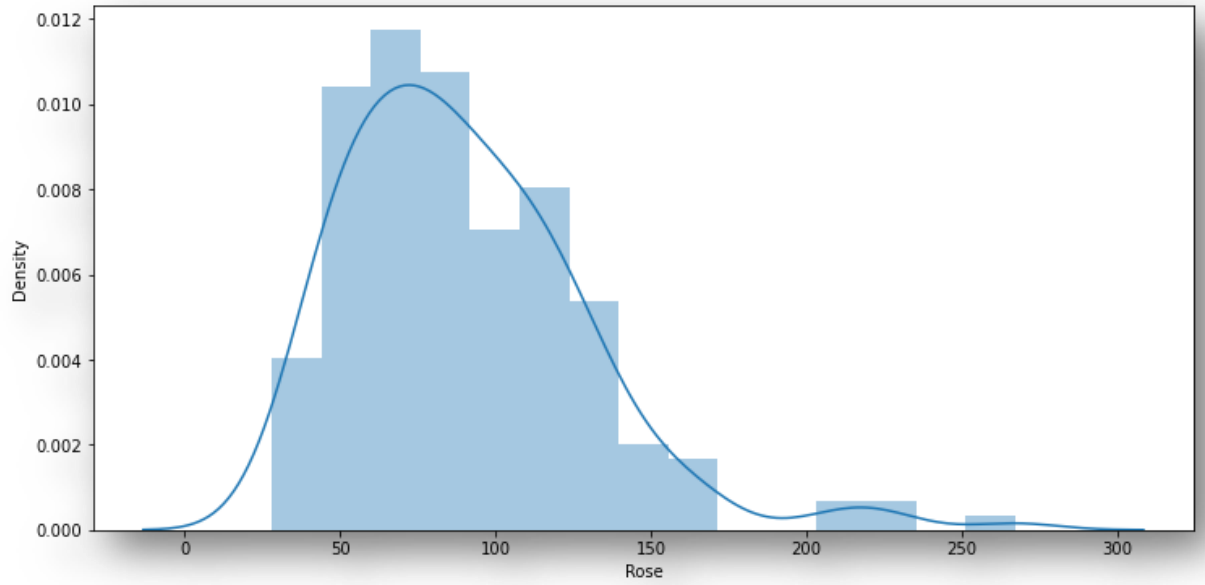
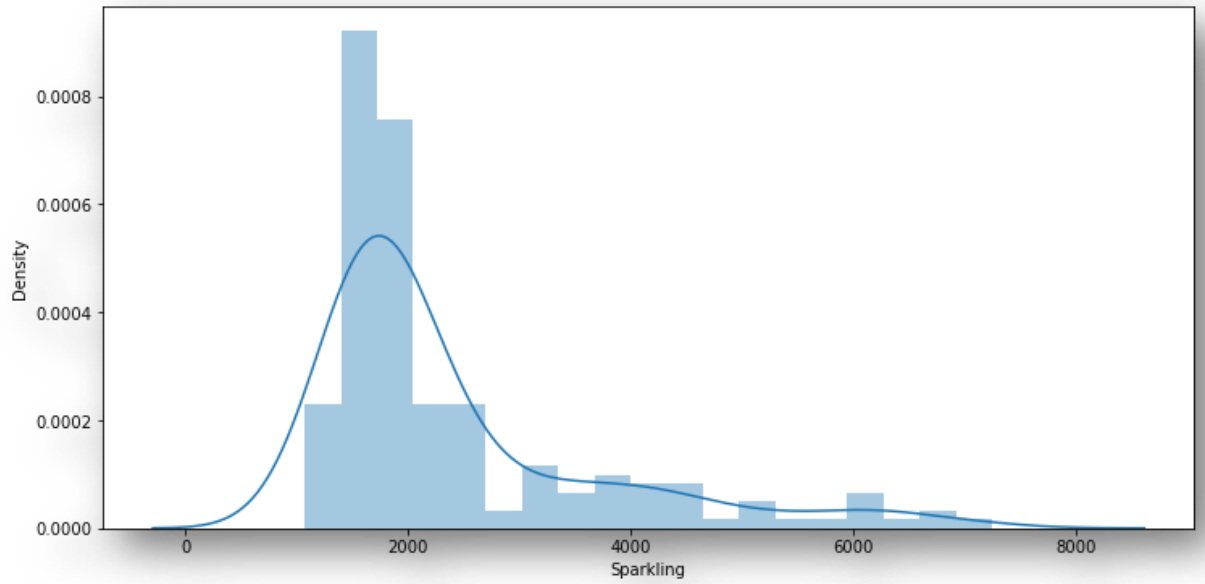


FIGURE 3: DISTRIBUTION PLOT

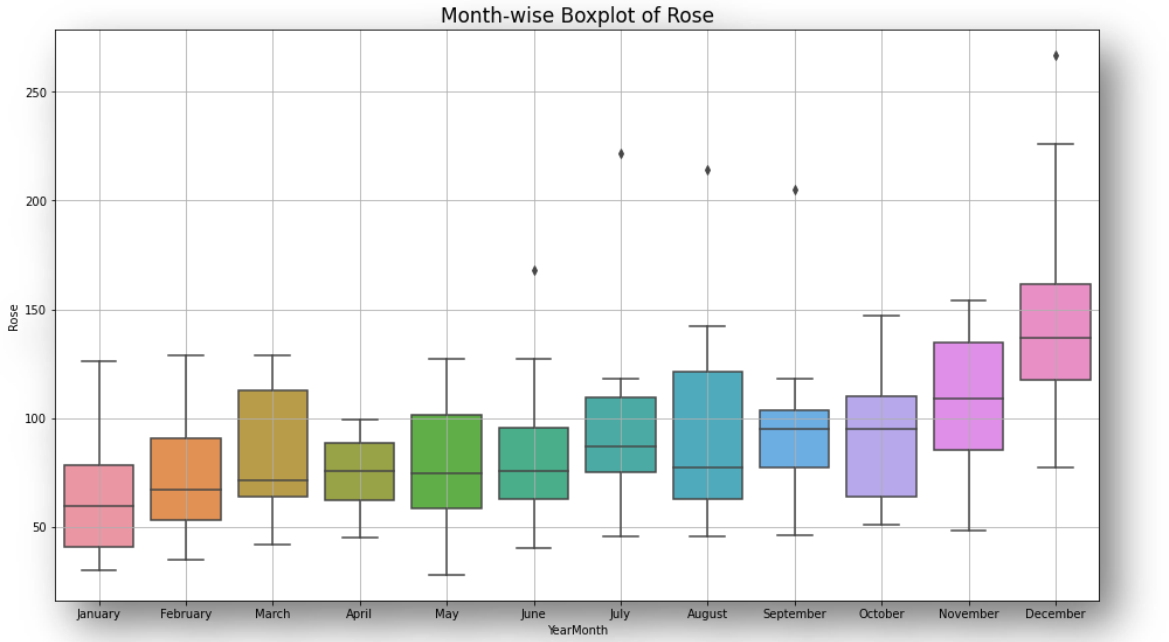


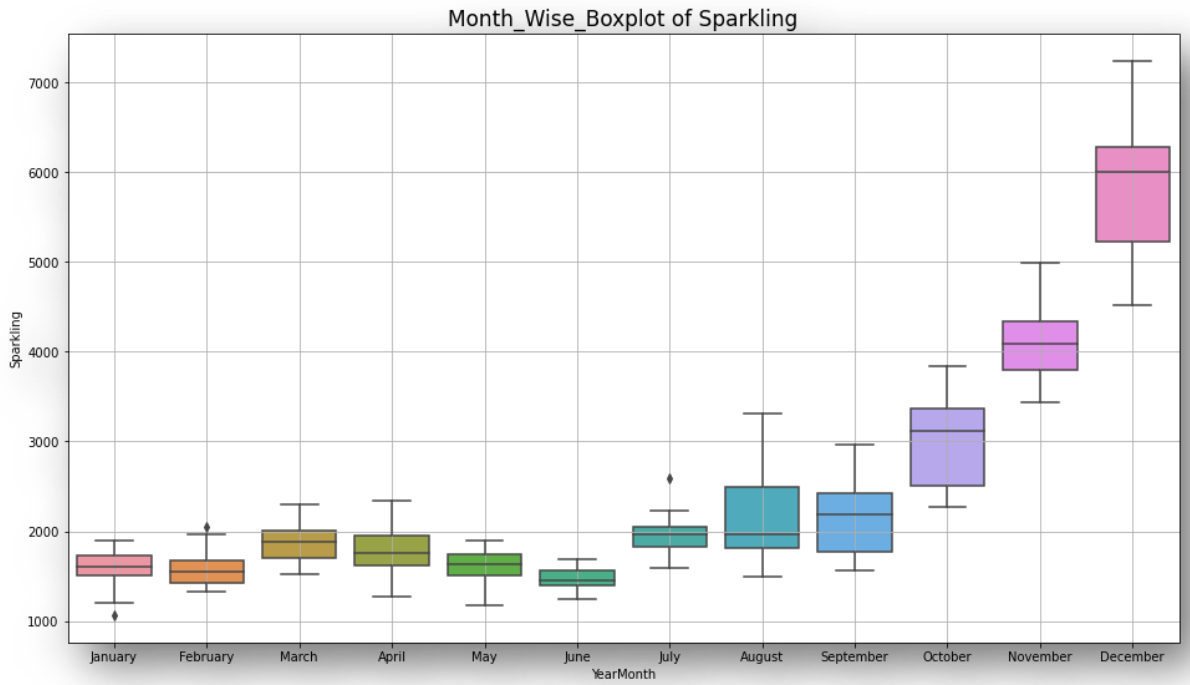


- Till here, we can see that there are significant outliers present in both datasets, though it is not clear in which month there are outliers particularly, we shall explore that further below.
- The dataset is almost distributed normally and right skewed for both the datasets.
- There are 185 rows and 1 column in dataset Rose and 187 rows and 1 column in dataset Sparkling. (Table 2).

- The description dataset shows that in the rose dataset the mean is 90 which is lower than the sparkling mean which is 2402.41, hence can be told that the sales of sparkling comparatively higher than rose.
- The standard deviation is also lower than the mean which says that both the datasets are normally distributed.

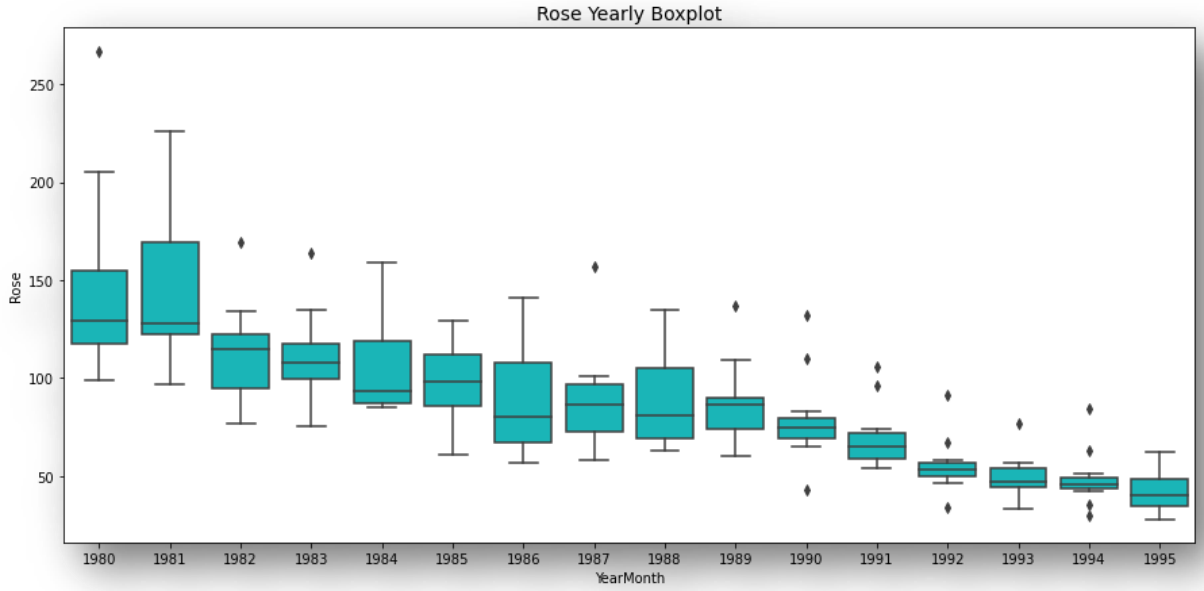
FIGURE 4: MONTHLY BOXPLOT

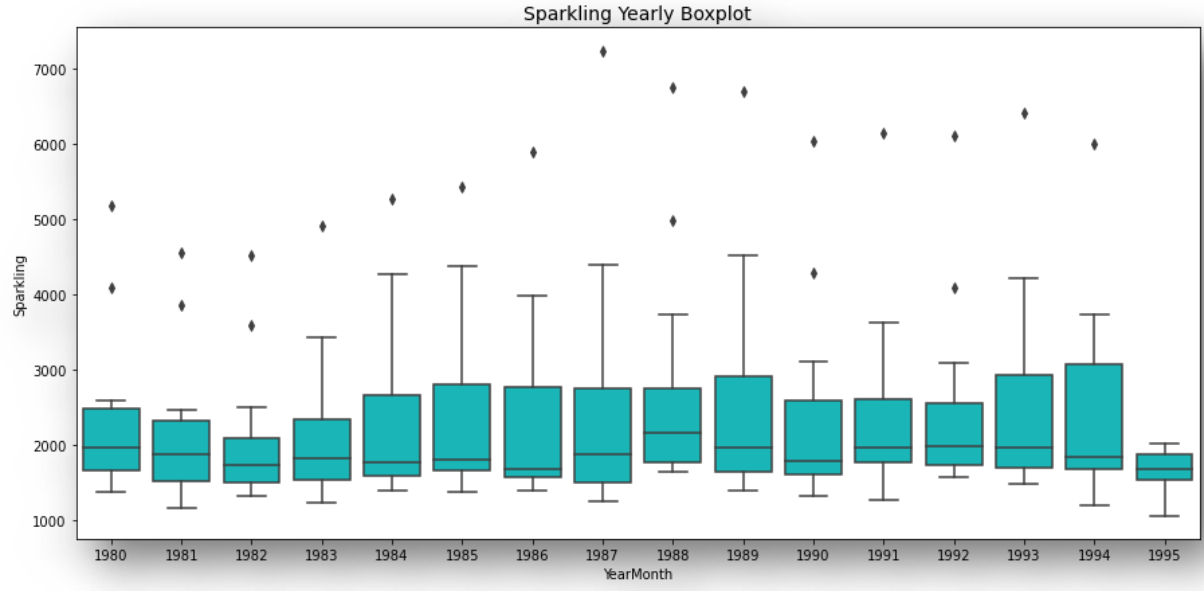




- From Figure 4, we can analyse that there is a spike in the sales of both the wines.
- The reason behind the spike can be the fall in the temperature at the end of the year and also the beginning of holidays from October.
- When compared between sparkling and Rose, Sales of Sparkling is higher for the last three months than rose
- The lowest sale happened in the month of June.

FIGURE 5: YEARLY BOXPLOT





- In Rose the highest sales happened in the year 1981 and from then on the sales have gradually fallen, with 1994-95 having the lowest sales.
- In Sparkling, the highest sale was in the year 1985 and 1987, the lowest in 1995. Sparkling sales is not falling when compared to Rose's Sales.

FIGURE 6: DECOMPOSITION ADDITIVE MODEL - ROSE

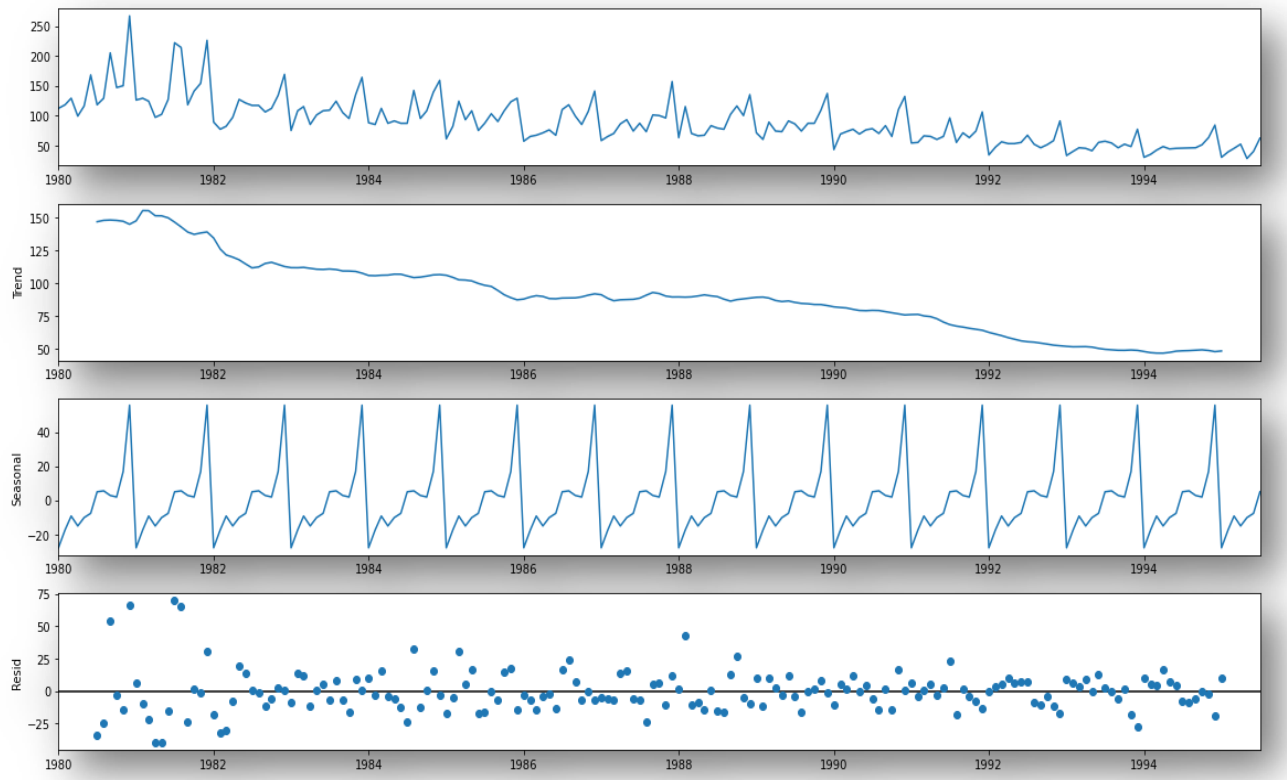


TABLE 7: DECOMPOSITION ADDITIVE MODEL - ROSE

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

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

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

FIGURE 7: DECOMPOSITION ADDITIVE MODEL - SPARKLING

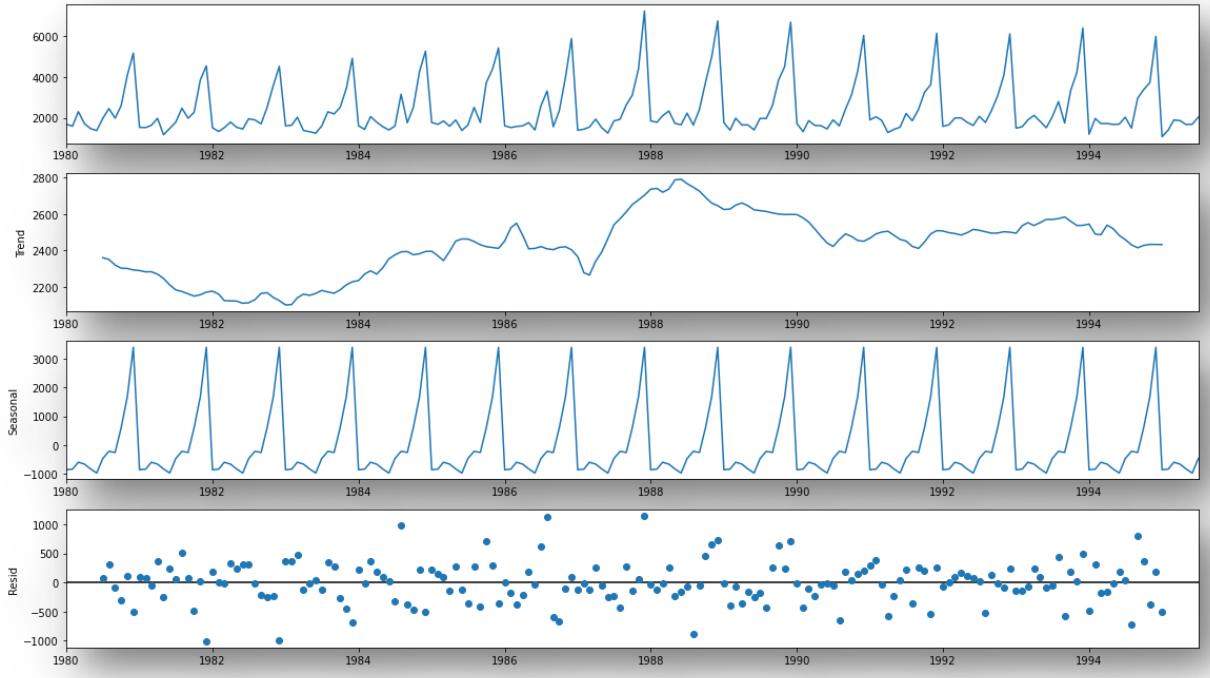


TABLE 8: DECOMPOSITION ADDITIVE MODEL - SPARKLING

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

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

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

FIGURE 8: DECOMPOSITION MULTIPLICATIVE MODEL - ROSE

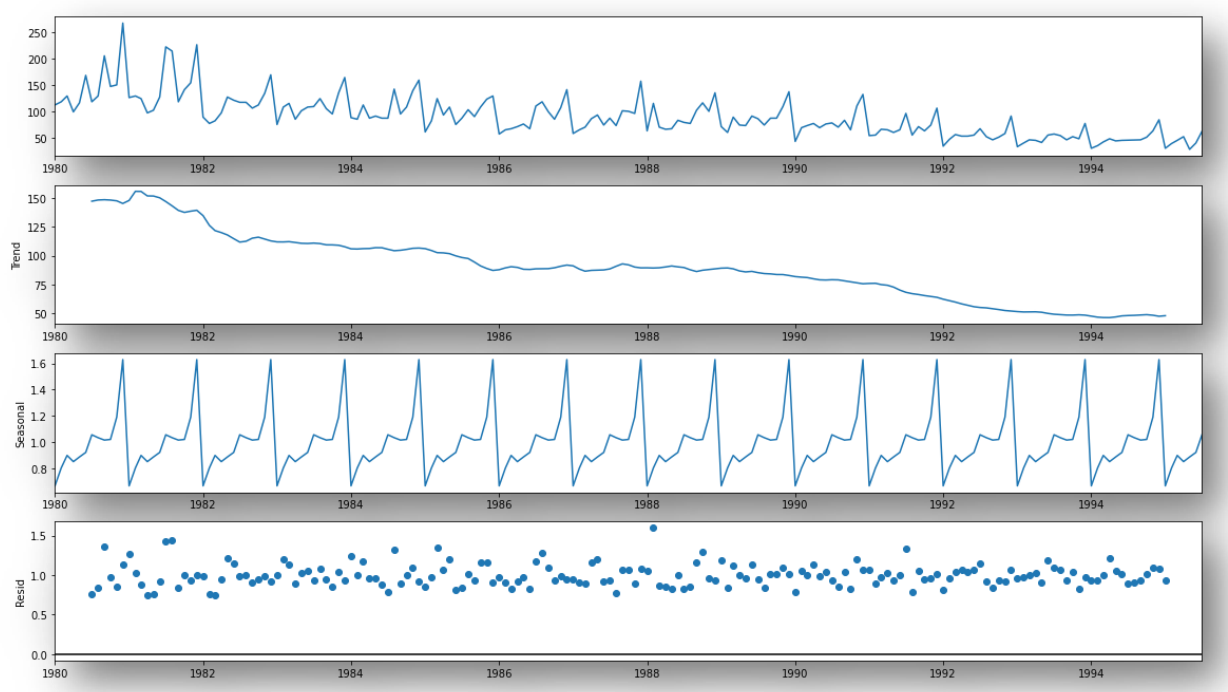


TABLE 9: DECOMPOSITION MULTIPLICATIVE MODEL - ROSE

YearMonth	Trend
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

YearMonth	Seasonality
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

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

FIGURE 9: DECOMPOSITION MULTIPLICATIVE MODEL - SPARKLING

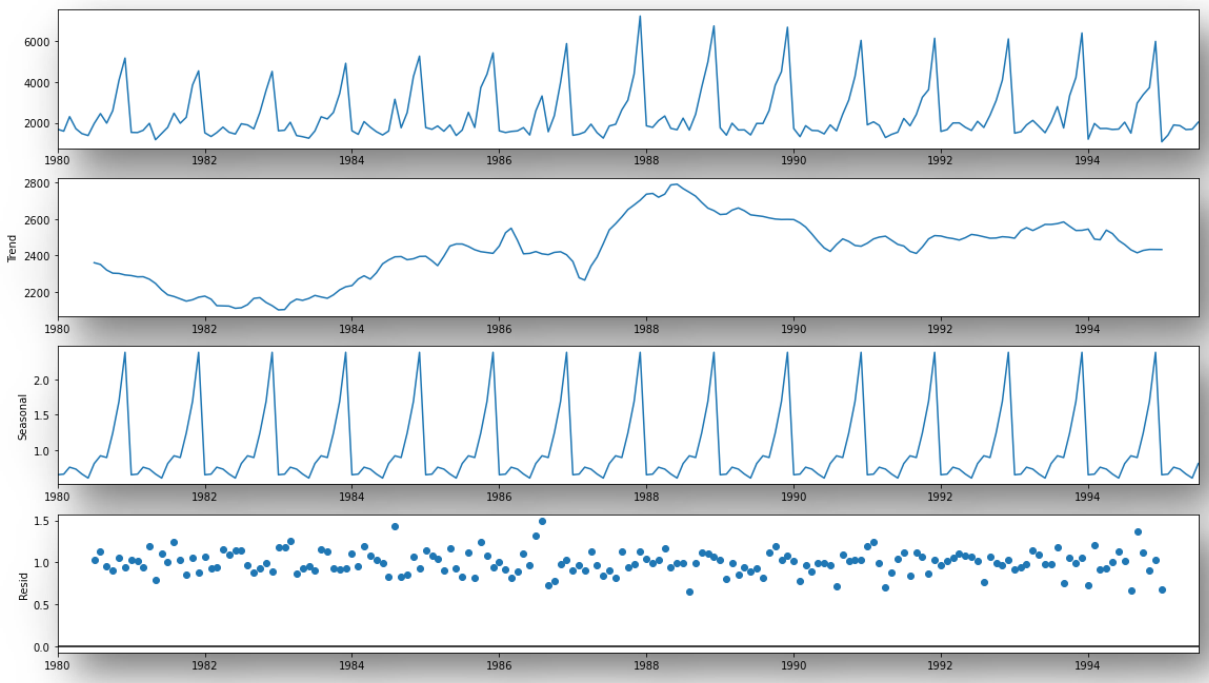


TABLE 10: DECOMPOSITION MULTIPLICATIVE MODEL - SPARKLING

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

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

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

- From the above decomposition of both the datasets into additive and multiplicative separately, we can infer that from the residual plot we can confirm that the Rose dataset is Multiplicative and Sparkling is Additive.

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

TABLE 11: SHAPE OF THE SPLIT DATASETS

```
(132, 1)
(55, 1)

(132, 1)
(55, 1)
```

The shape of the both dataset Rose and Sparkling after the split is 132 rows and 1 column for train data and 55 rows and 1 column for test data.

TABLE 12: DATASET SAMPLES AFTER SPLIT -ROSE

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

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

TABLE 13: DATASET SAMPLES AFTER SPLIT - SPARKLING

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

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

FIGURE 10: TRAIN AND TEST SPLIT PLOT -ROSE

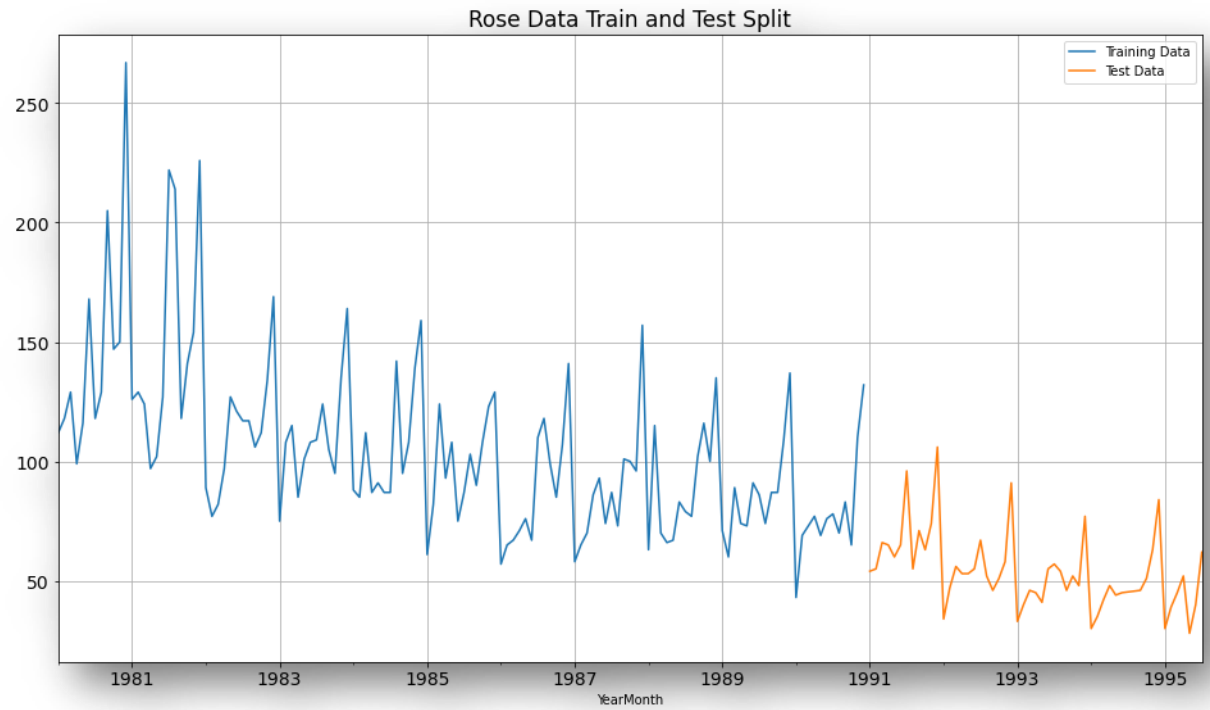
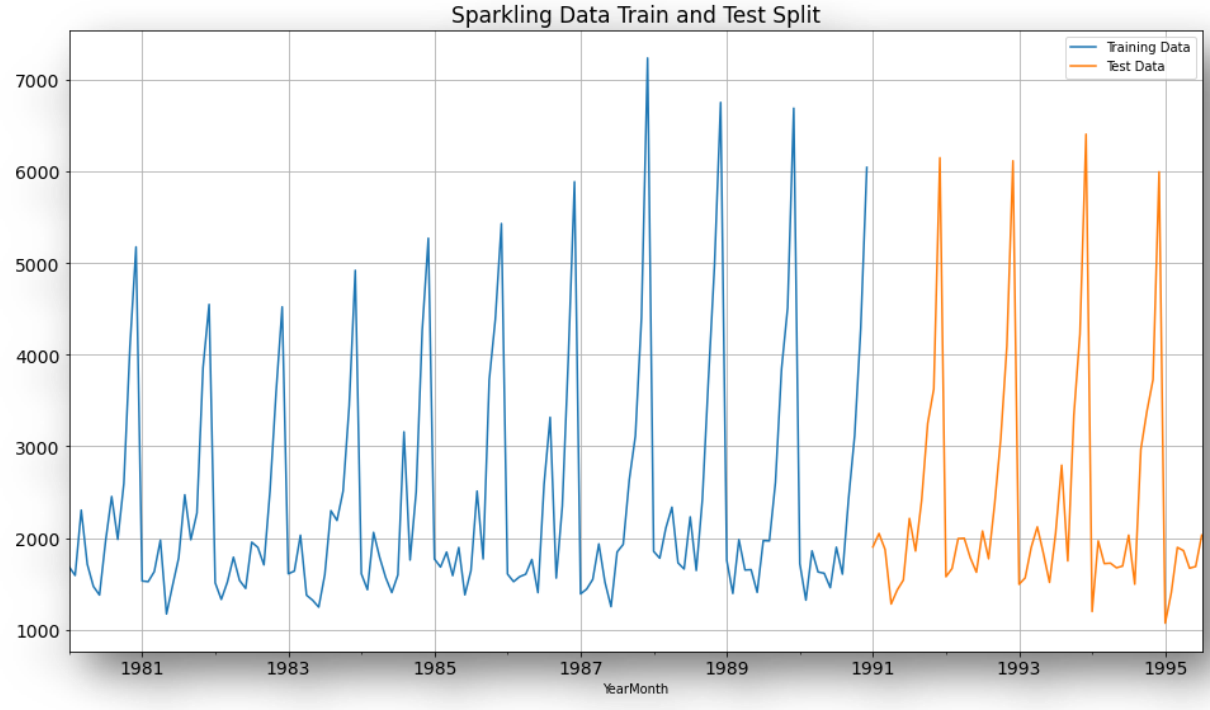
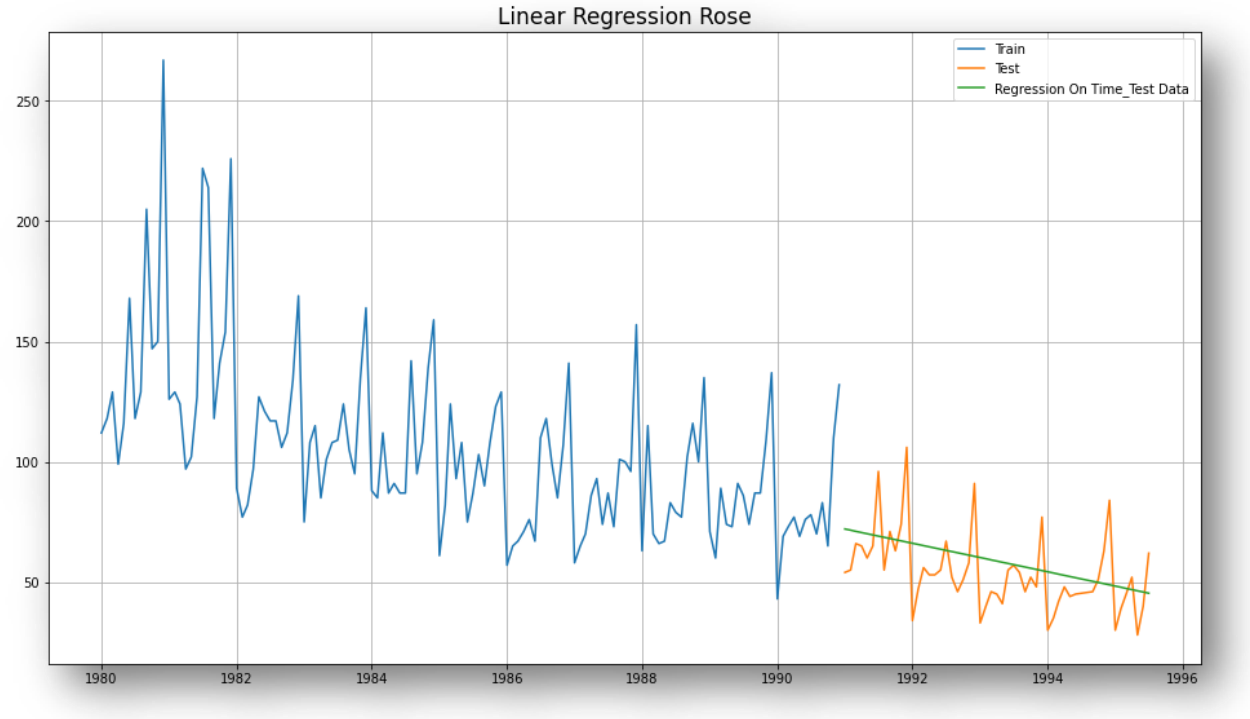


FIGURE 11: TRAIN AND TEST SPLIT PLOT -SPARKLING



Q.4.Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

FIGURE 12: LINEAR REGRESSION PLOT



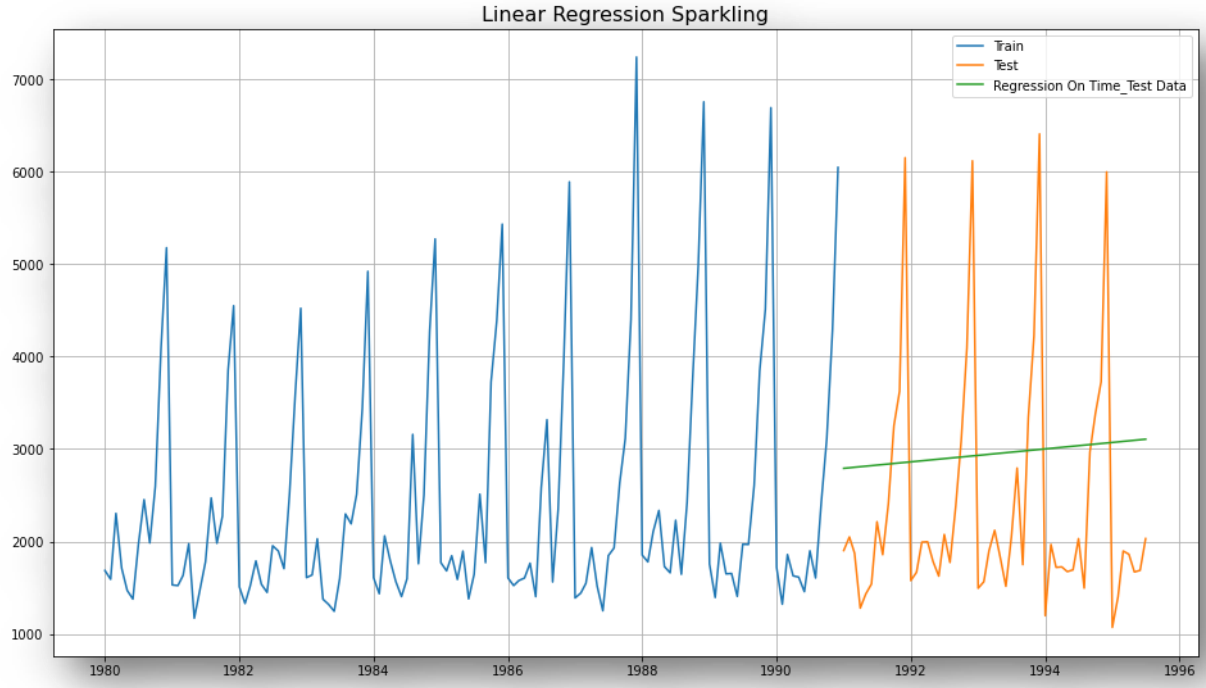
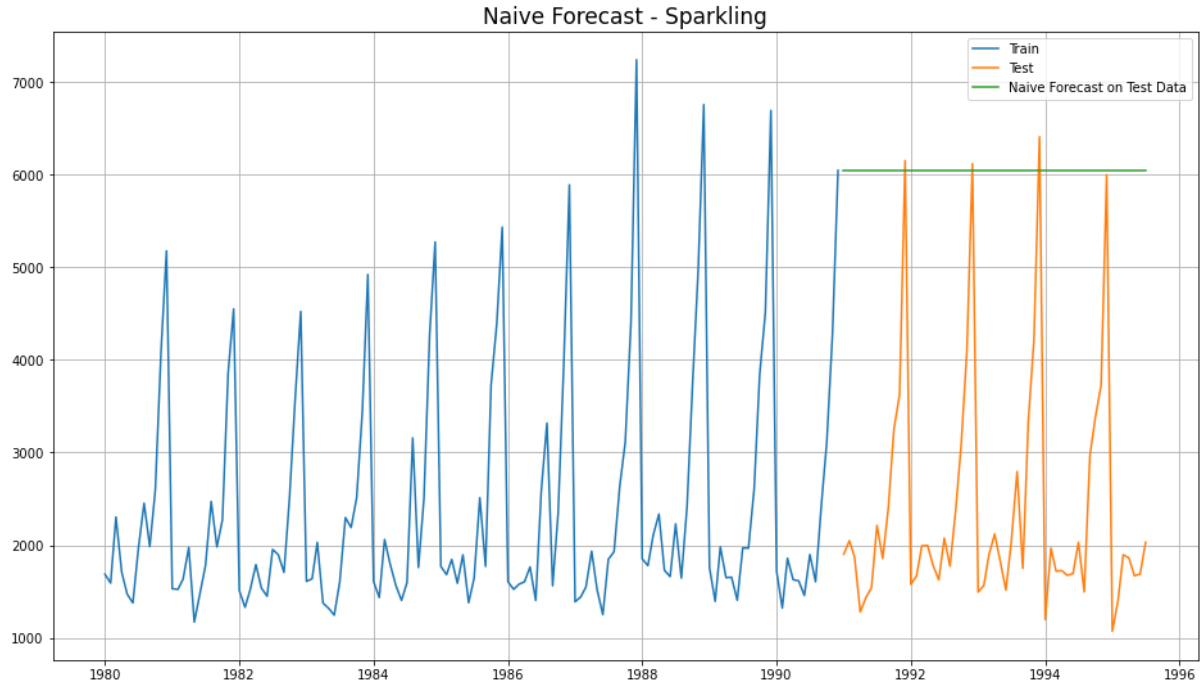


TABLE 14: LINEAR REGRESSION RMSE

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175

FIGURE 12: NAÏVE APPROACH PLOT



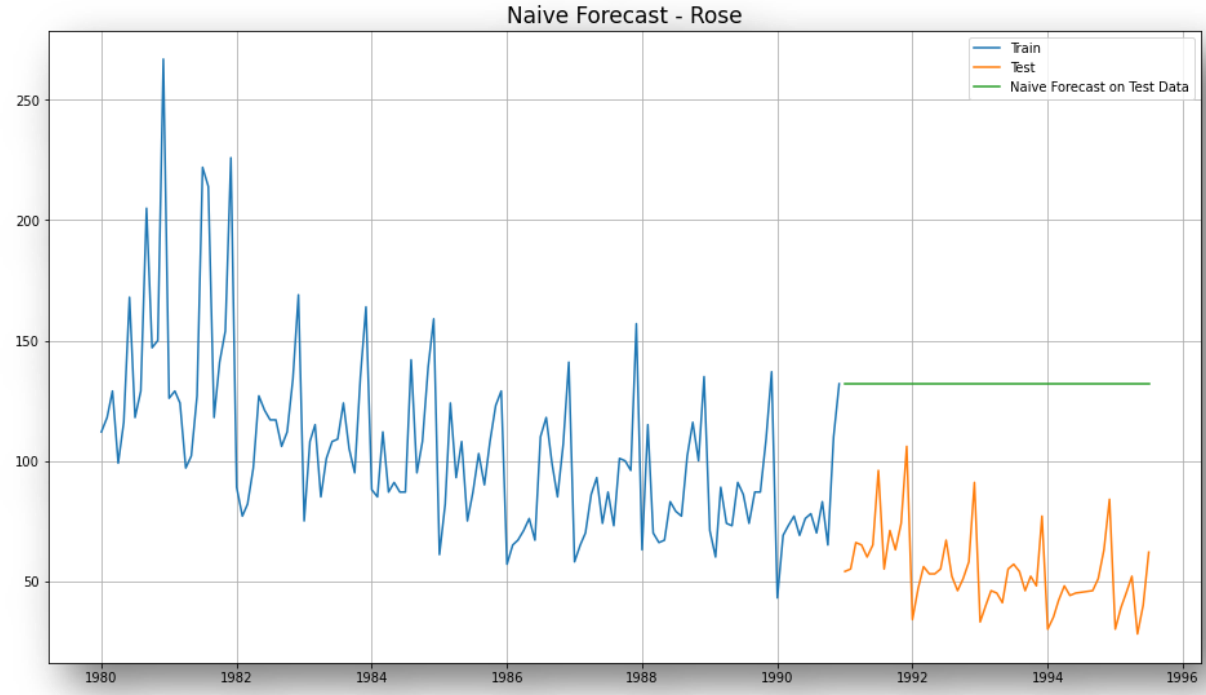
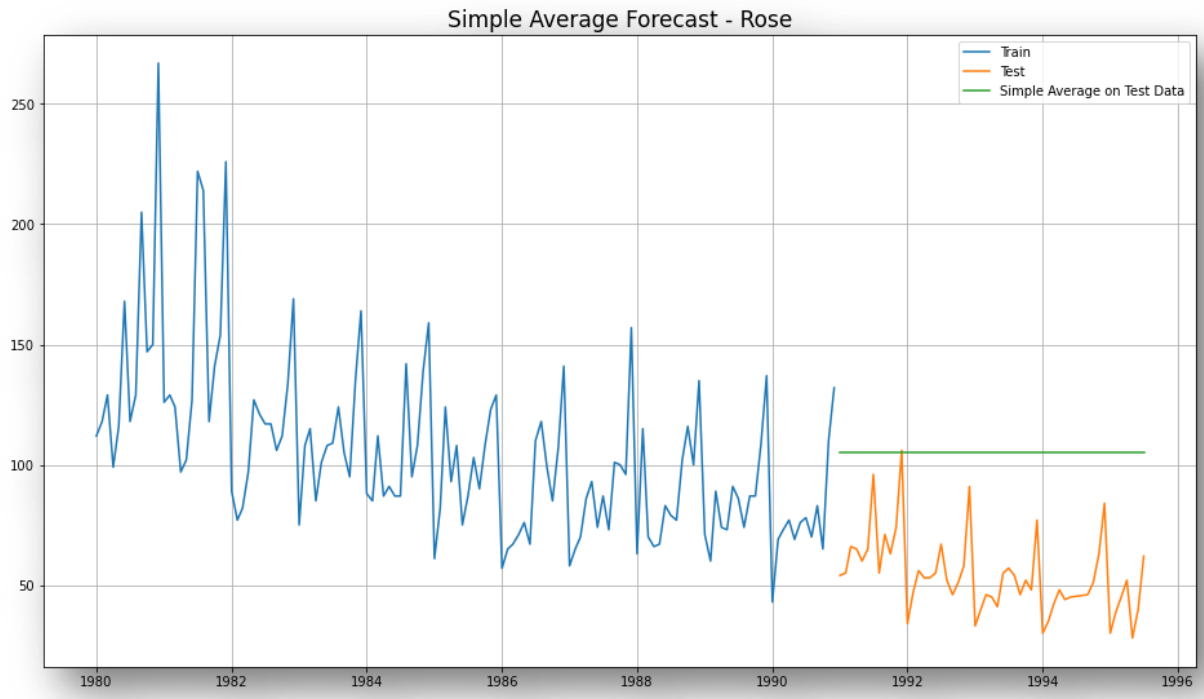


TABLE 15: NAÏVE APPROACH RMSE

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

FIGURE 13: SIMPLE AVERAGE PLOT



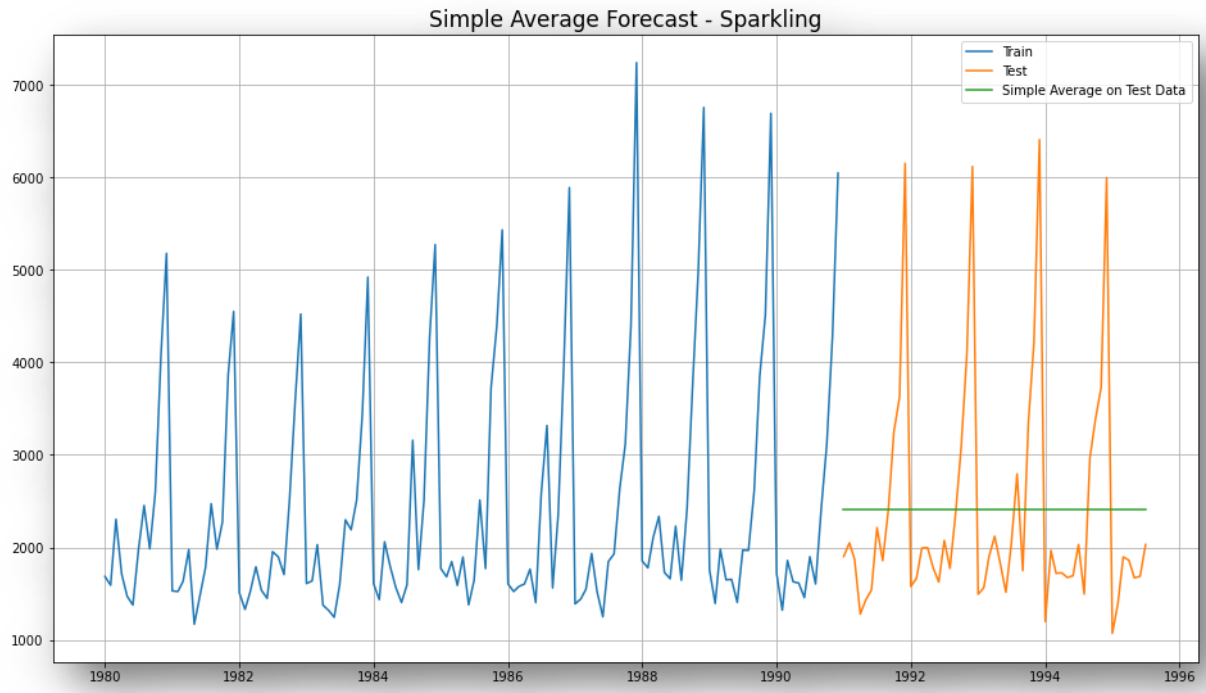


TABLE 16: SIMPLE AVERAGE APPROACH RMSE

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

FIGURE 15: MOVING AVERAGE PLOT - ROSE

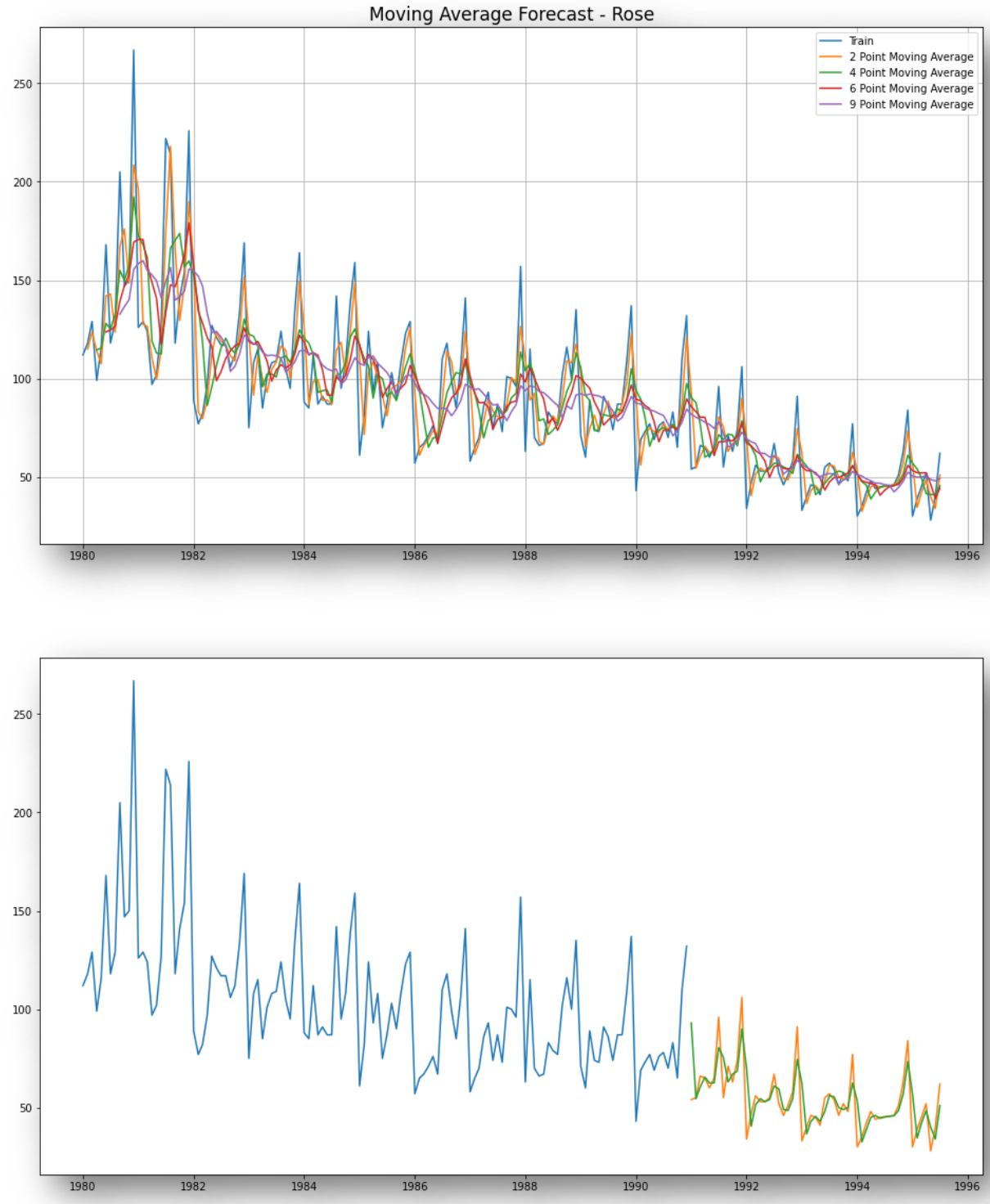
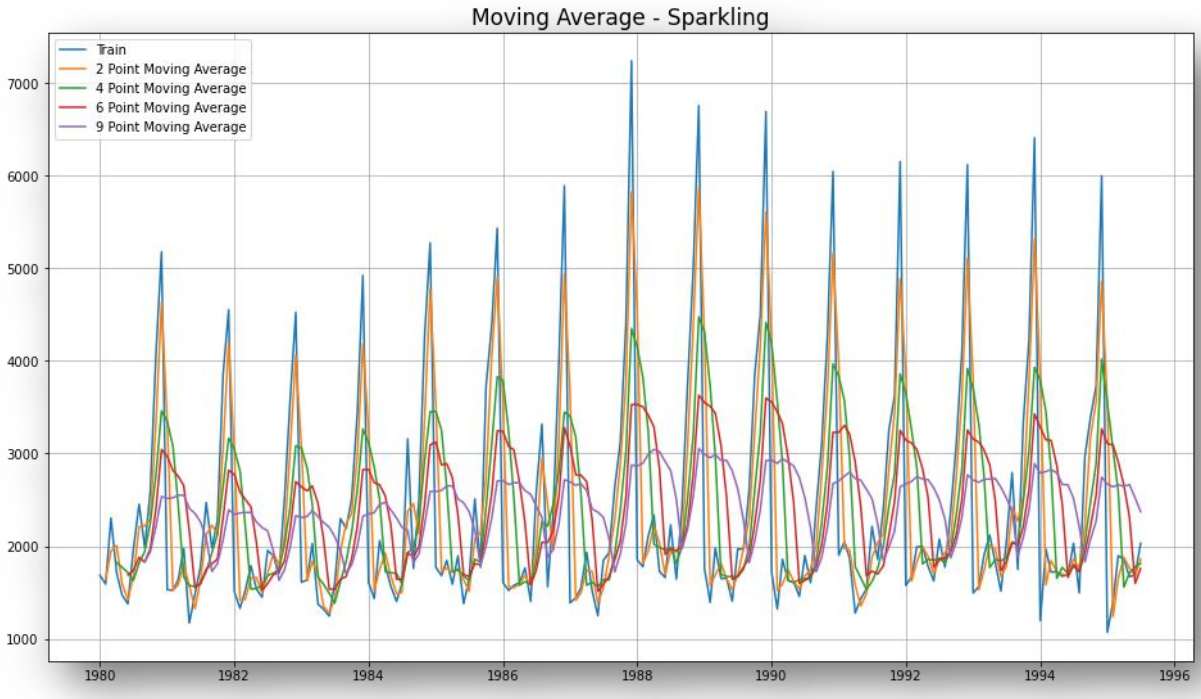


FIGURE 16: MOVING AVERAGE PLOT - SPARKLING



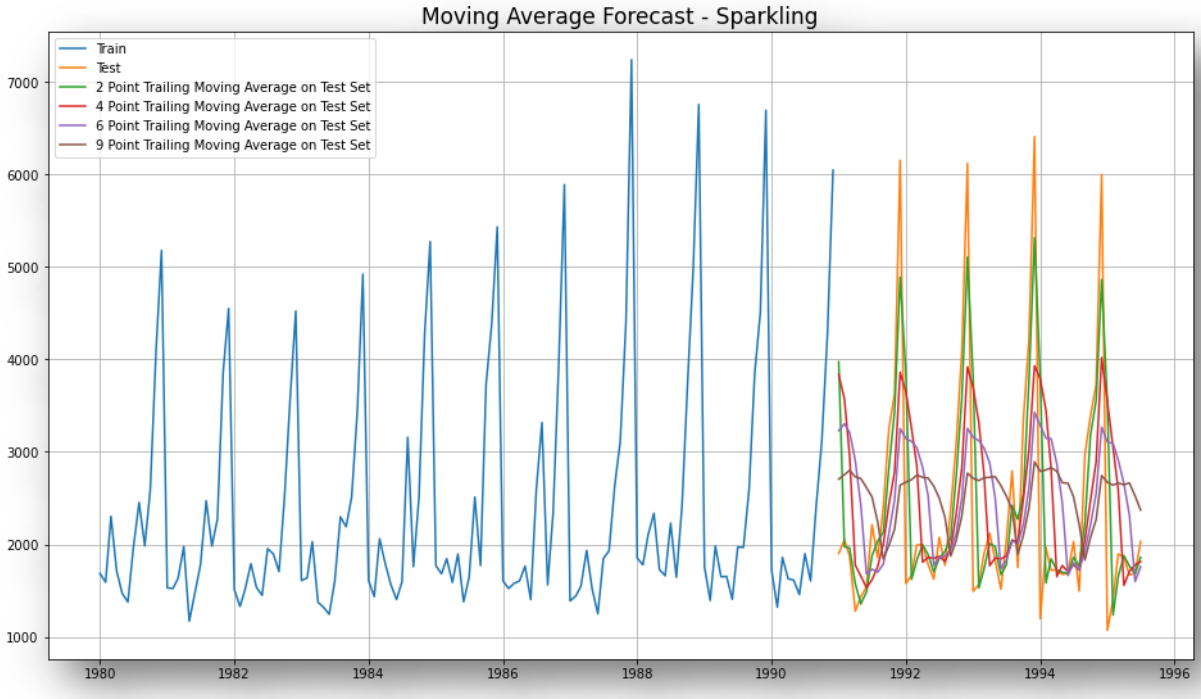
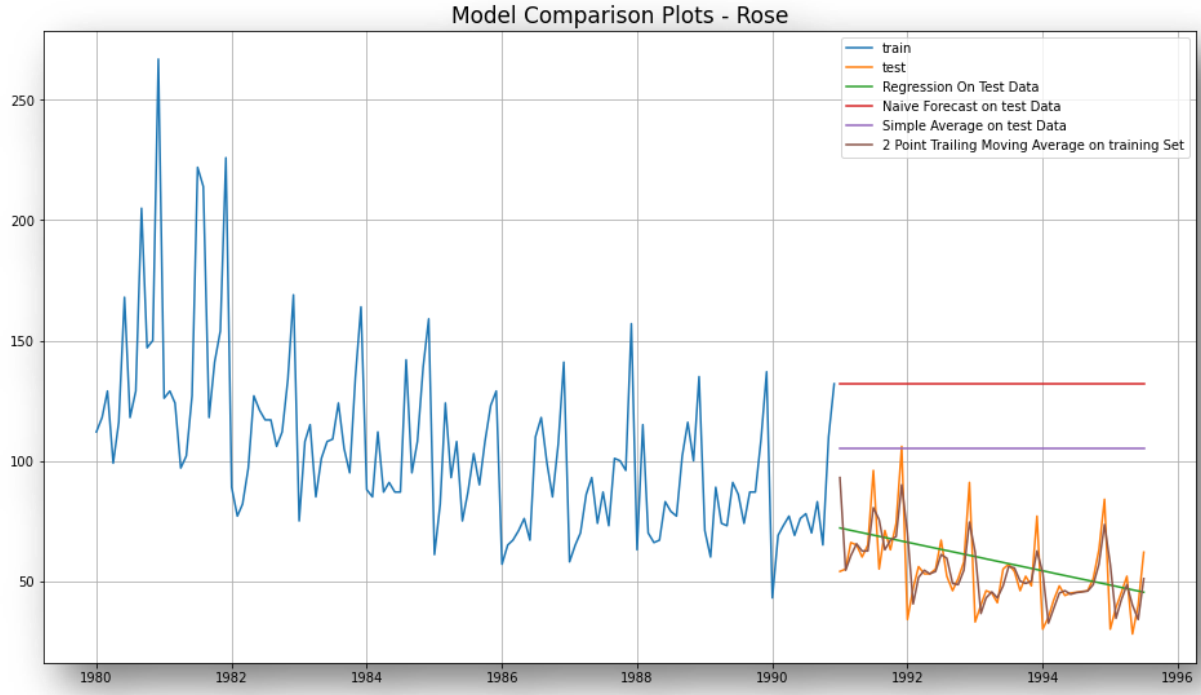
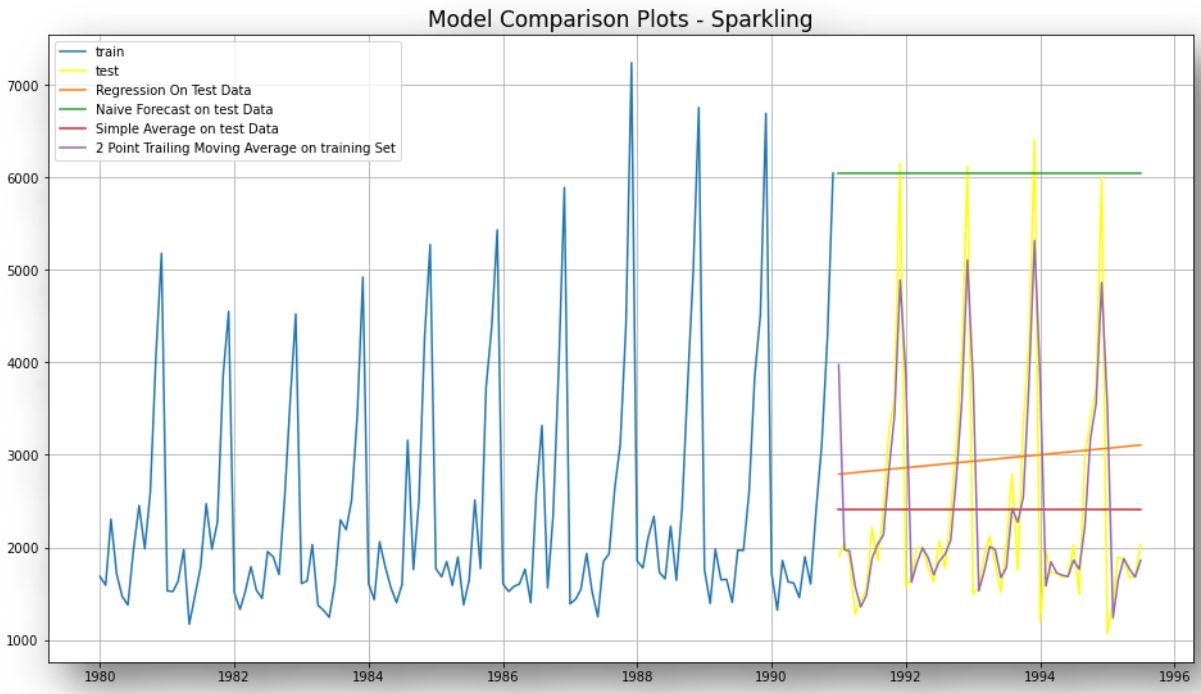


TABLE 17: MOVING AVERAGE APPROACH RMSE

	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

FIGURE 17: MODEL COMPARISION





- From the above models for both Rose and Sparkling Wine datasets we can infer the following:

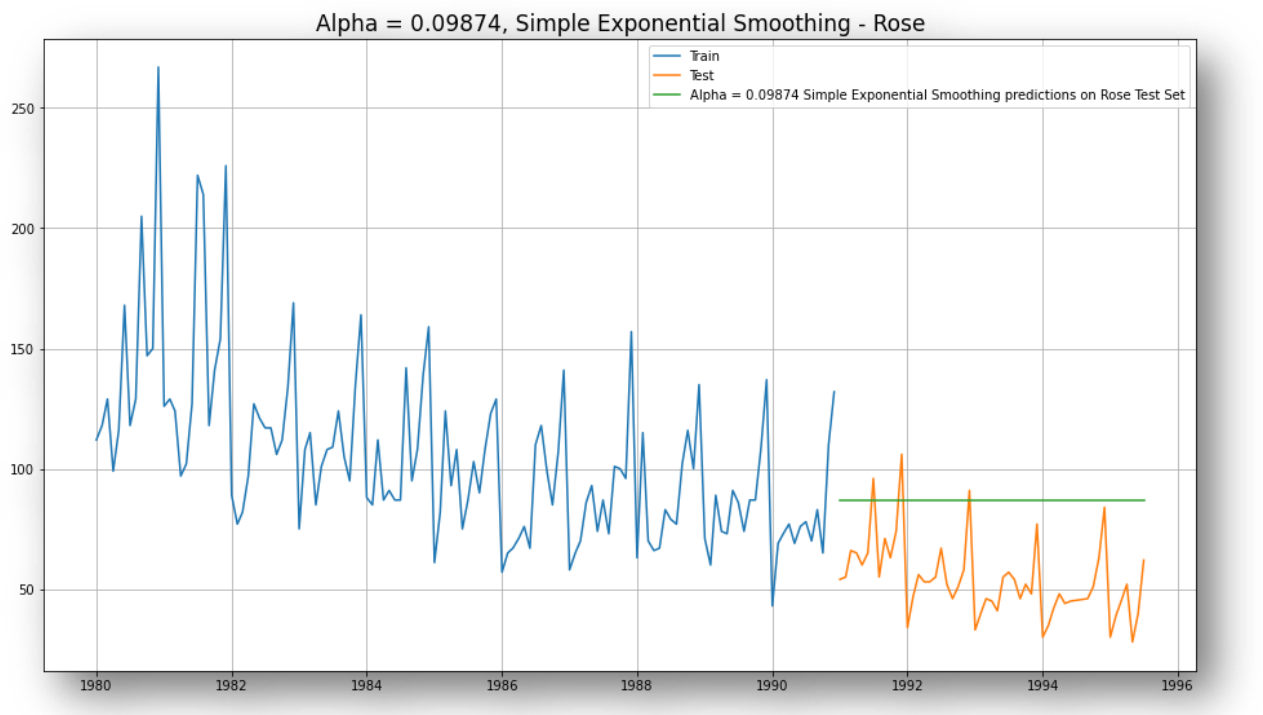
1. Linear Regression: Test RMSE Rose = 15.27, Test RMSE Sparkling = 1389.14
2. Naïve approach: All the future predictions will be same as today. Test RMSE Rose = 79.72, Test RMSE Sparkling = 3864.27
3. Simple Average Model: Test RMSE Rose = 53.46, Test RMSE Sparkling = 1275.08
4. 2 Point Trailing Moving Average: Test RMSE Rose = 11.52, Test RMSE Sparkling = 813.40
5. 4 Point Trailing Moving Average: Test RMSE Rose = 14.45, Test RMSE Sparkling = 1156.58
6. 6 Point Trailing Moving Average: Test RMSE Rose = 14.56, Test RMSE Sparkling = 1283.92
7. 9 Point Trailing Moving Average: Test RMSE Rose = 14.72, Test RMSE Sparkling = 1346.27.

- 2 Point Moving average means, we find average of 1st and 2nd to predict 3rd similarly, average of 2nd and 3rd to predict 4th and so on
- 4 Point Moving Average 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 $y_t = \beta_0 + \beta_1 X_t + \epsilon_t$ $\hat{y}_{t+1} = y_t$ $\hat{y}_{t+1} = \hat{y}_{t+2} = \ldots = \hat{y}_{t+n} = \text{Mean}(y_1, y_2, \ldots, y_t)$

- From the above we can conclude that, the best model which gives the lowest RMSE for both the datasets is the **2 Point Moving Average Model.**
- We'll build following Exponential Smoothing Models –

1. Single Exponential Smoothing with Additive Errors - ETS (A, N, N)
2. Double Exponential Smoothing with Additive Errors, Additive Trends - ETS (A, A, N)
3. Triple Exponential Smoothing with Additive Errors, Additive Trends, Additive Seasonality - ETS (A, A, A)
4. Triple Exponential Smoothing with Additive Errors, Additive Trends, Multiplicative Seasonality - ETS (A, A, M)
5. Triple Exponential Smoothing with Additive Errors, Additive DAMPED Trends, Additive Seasonality - ETS (A, Ad, A)
6. Triple Exponential Smoothing with Additive Errors, Additive DAMPED Trends, Multiplicative Seasonality - ETS (A, Ad, M).

FIGURE 18: SINGLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS - ETS (A, N, N)



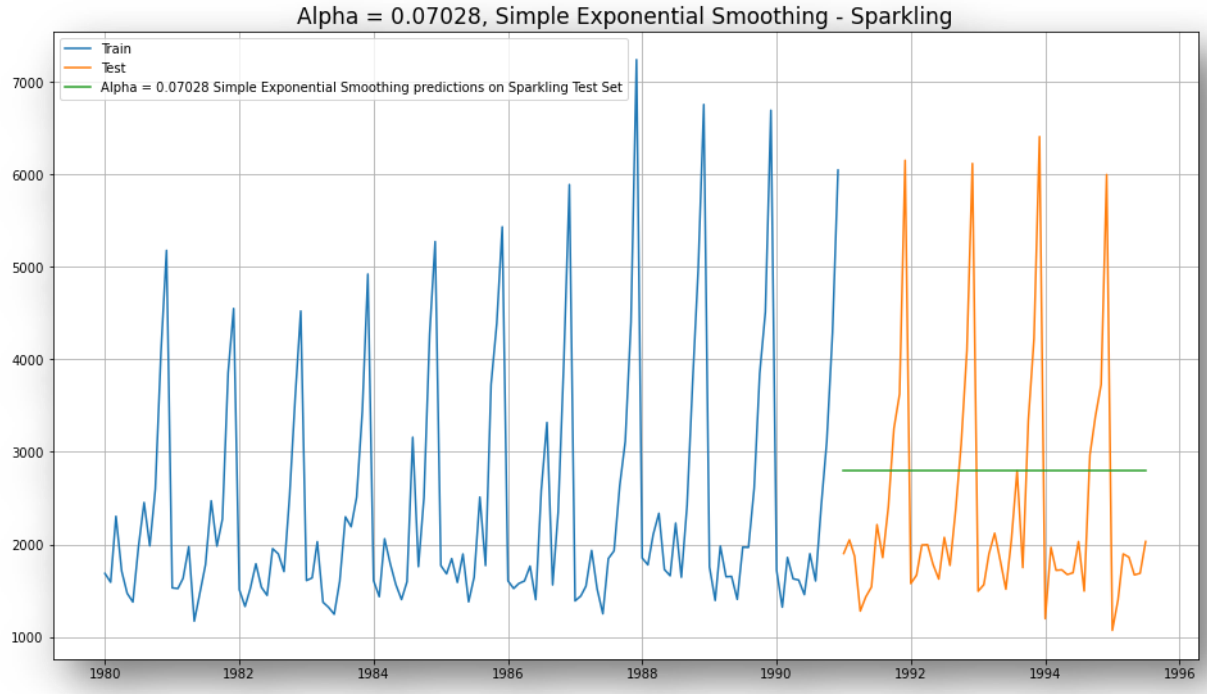
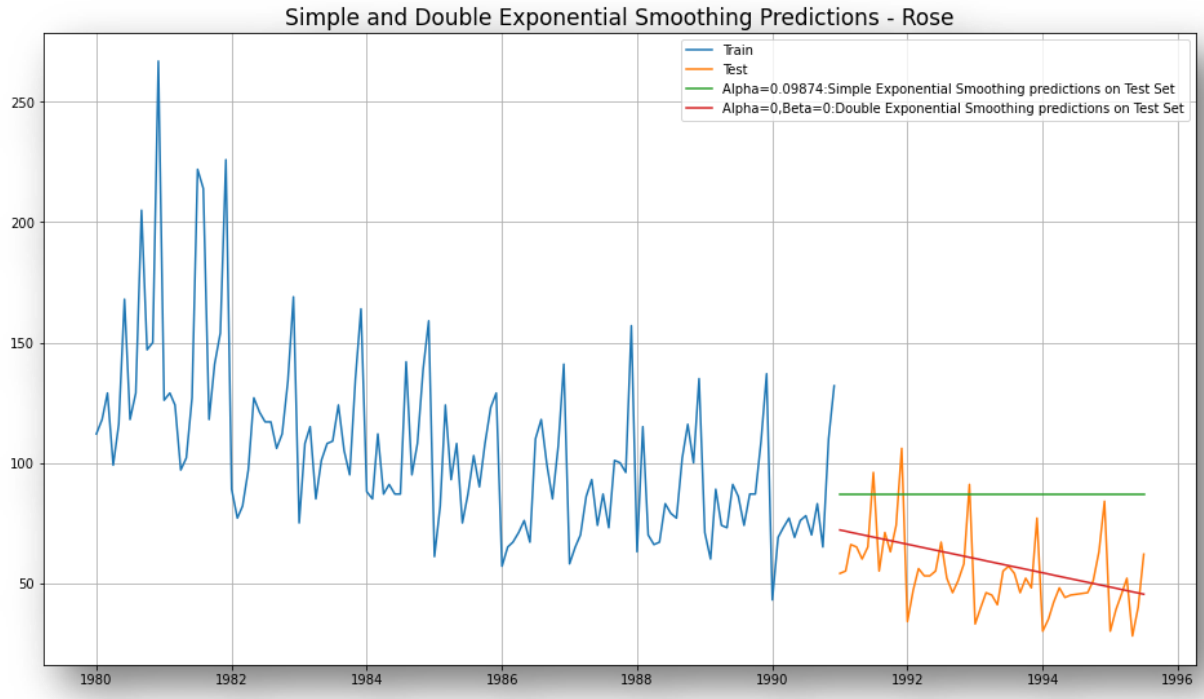


TABLE 18: SINGLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS - ETS (A, N, N) RMSE

	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	36.796228	1338.012144

FIGURE 19: DOUBLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS - ETS (A, A, N)



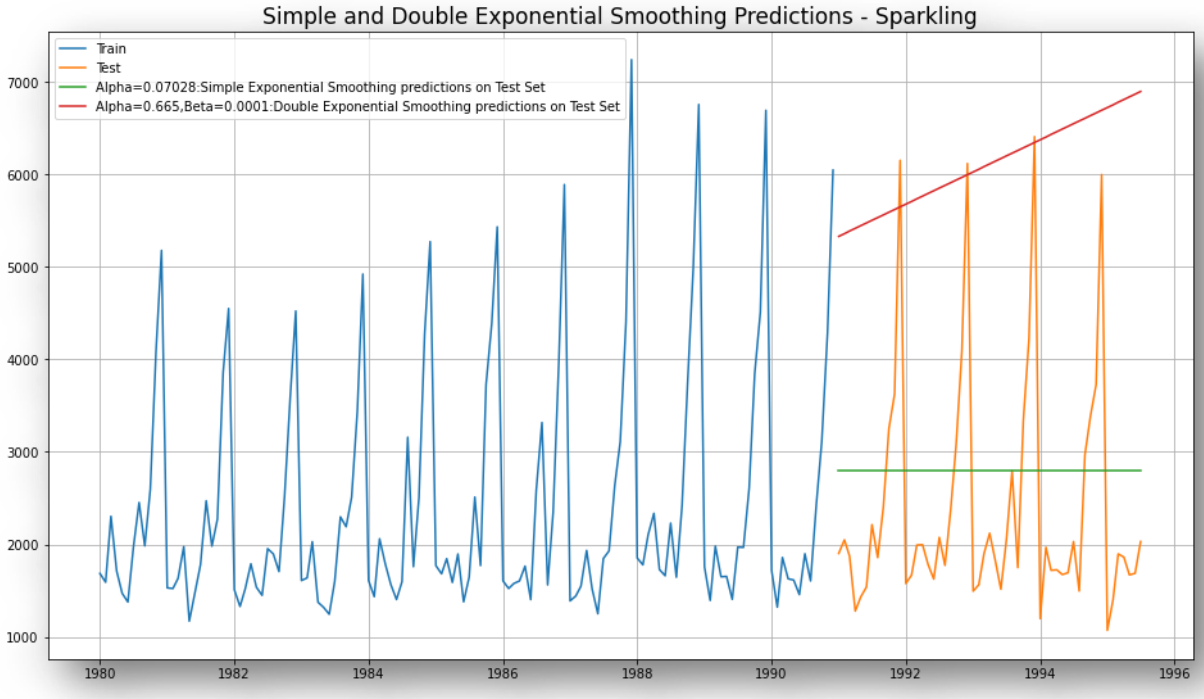


TABLE 19: DOUBLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS - ETS (A, A, N) RMSE

	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	36.796228	1338.012144
Double Exponential Smoothing	15.269328	3949.993290

Here, we see that the Double Exponential Smoothing has actually done well when compared to the Simple Exponential Smoothing. This is because of the fact that the Double Exponential Smoothing model has picked up the trend component as well.

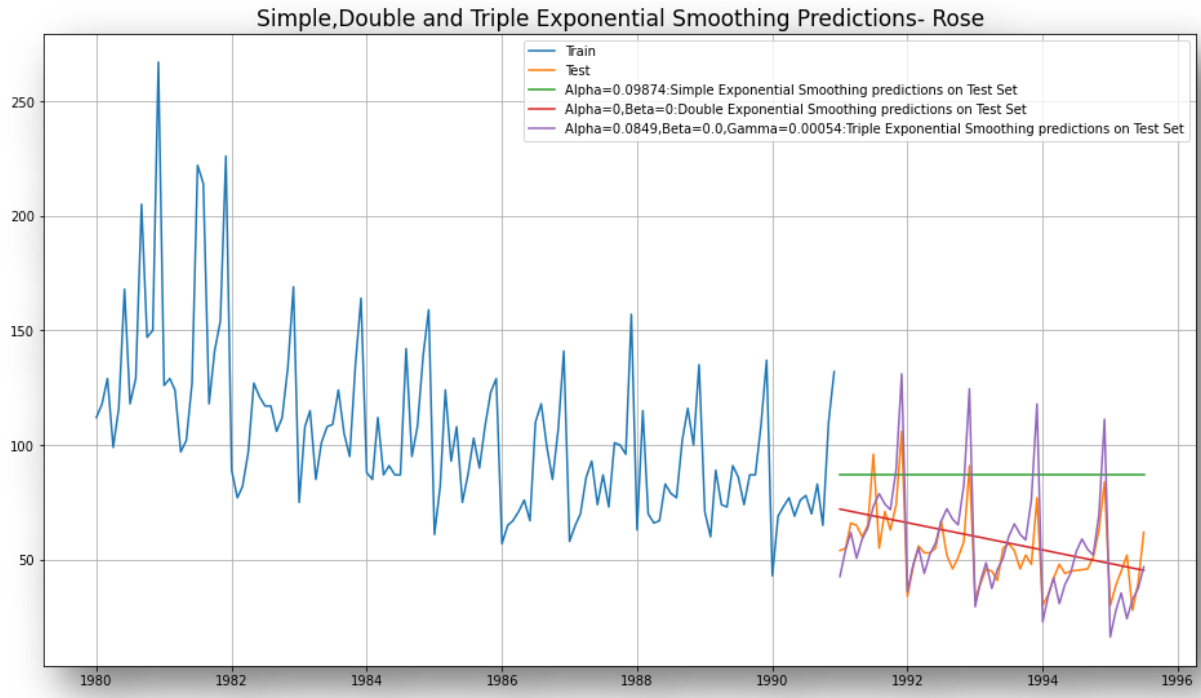
- **The Best Parameters for Dataset Rose:**

```
Holt model Exponential Smoothing Estimated Parameters :

{'smoothing_level': 1.9086427682180844e-08, 'smoothing_trend': 7.302464353829351e-09, 'smoothing_seasonal': nan, 'damping_tren
d': nan, 'initial_level': 137.81629861505857, 'initial_trend': -0.4943753249082896, 'initial_seasons': array([], dtype=float6
4), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

- **The Best Parameters for Dataset Sparkling:**

```
Holt model Exponential Smoothing Estimated Parameters :

{'smoothing_level': 0.6638769092832238, 'smoothing_trend': 9.966251357628782e-05, 'smoothing_seasonal': nan, 'damping_trend': n
an, 'initial_level': 1502.5681711003654, 'initial_trend': 29.020225552837097, 'initial_seasons': array([], dtype=float64), 'use
_boxcox': False, 'lamda': None, 'remove_bias': False}
```

FIGURE 20: TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS, ADDITIVE SEASONALITY - ETS (A, A, A)



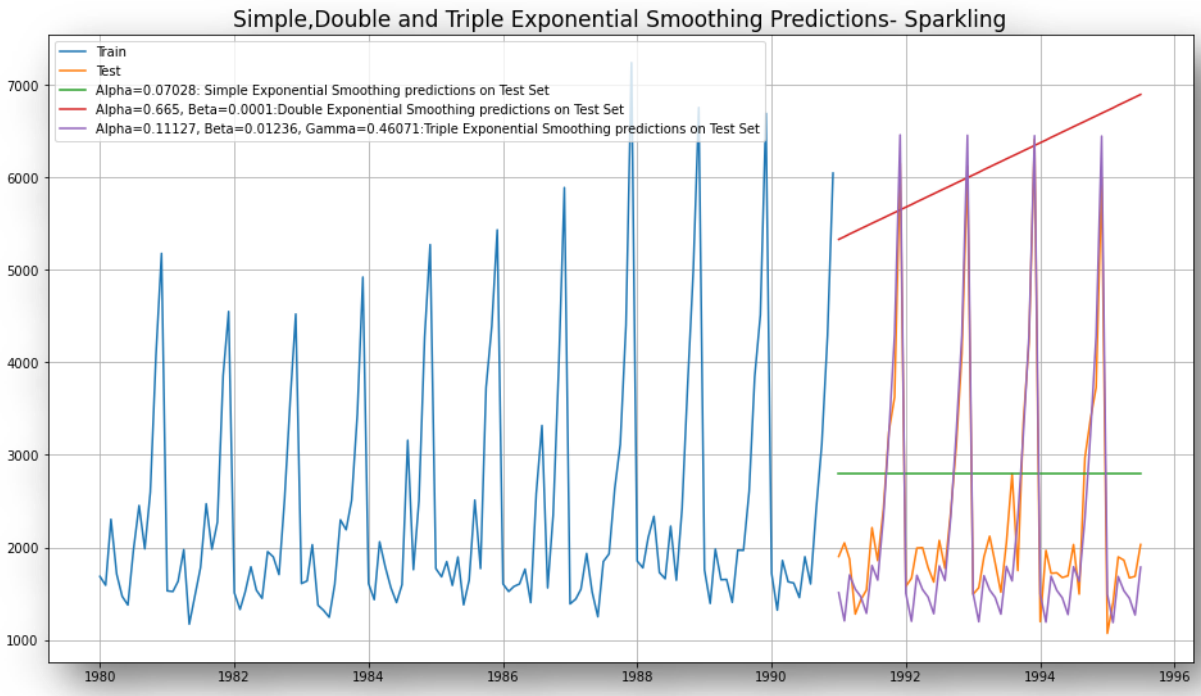


TABLE 20: TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS, ADDITIVE SEASONALITY - ETS (A, A, A) RMSE

	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	36.796228	1338.012144
Double Exponential Smoothing	15.269328	3949.993290
Triple Exponential Smoothing (Additive Season)	14.265713	379.695686

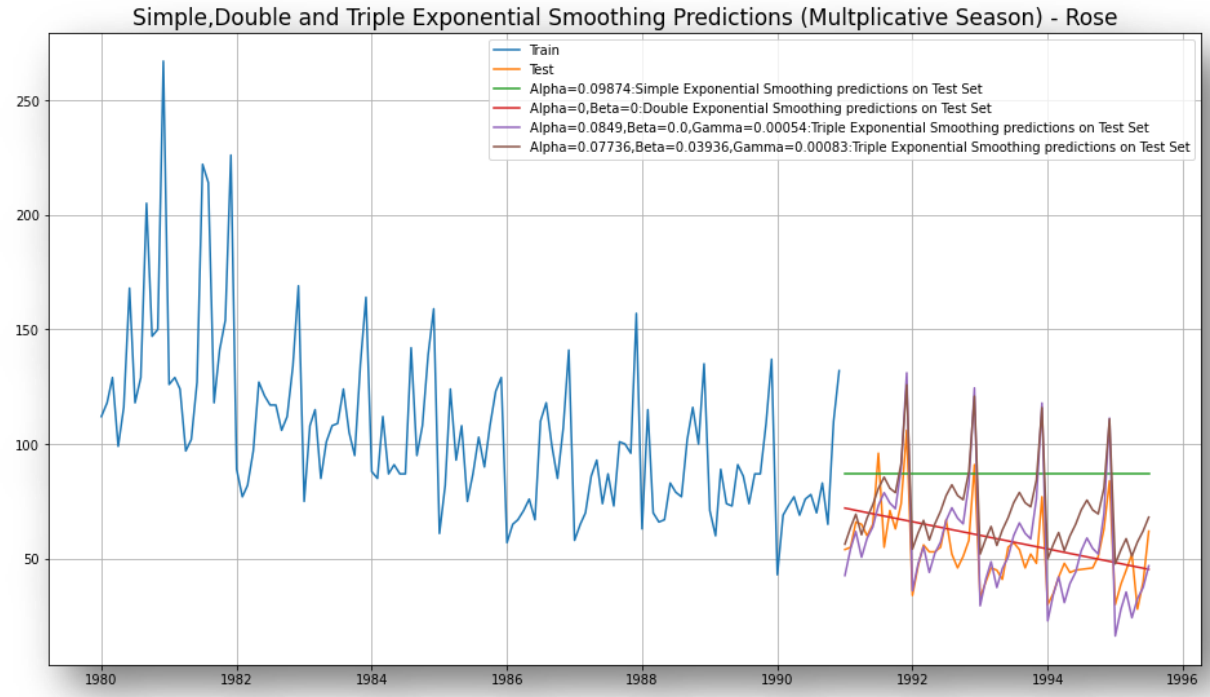
- In this model the trend and seasonality has picked up very well, but still the 2-point Trailing M.A still performs better than the others.
- **The Best Parameters for Dataset Rose:**

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.08830330642635406, 'smoothing_trend': 6.730635331927582e-05, 'smoothing_seasonal': 0.004455138229351625,
'damping_trend': nan, 'initial_level': 146.88752868155674, 'initial_trend': -0.5492163940406024, 'initial_seasons': array([-31.
12207537, -18.81171138, -10.86052241, -21.52235816,
       -12.68359535,  -7.17529564,   2.7456236 ,   8.84900094,
         4.85724354,   2.9520333 ,  21.05004912,  63.29916317]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

- **The Best Parameters for Dataset Sparkling:**

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.10005373820823961, 'smoothing_trend': 0.010034490652580457, 'smoothing_seasonal': 0.5095957543425532, 'da
mping_trend': nan, 'initial_level': 2364.584774604334, 'initial_trend': -0.016752880078245408, 'initial_seasons': array([-653.8
2559323, -736.67734144, -368.25456128, -483.63906084,
       -826.15467946, -832.96819741, -386.3751117 ,   91.82676187,
       -261.32455153,  265.38968222, 1580.26233564, 2619.56221896]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

FIGURE 21: TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS, MULTIPLICATIVE SEASONALITY - ETS (A, A, M)



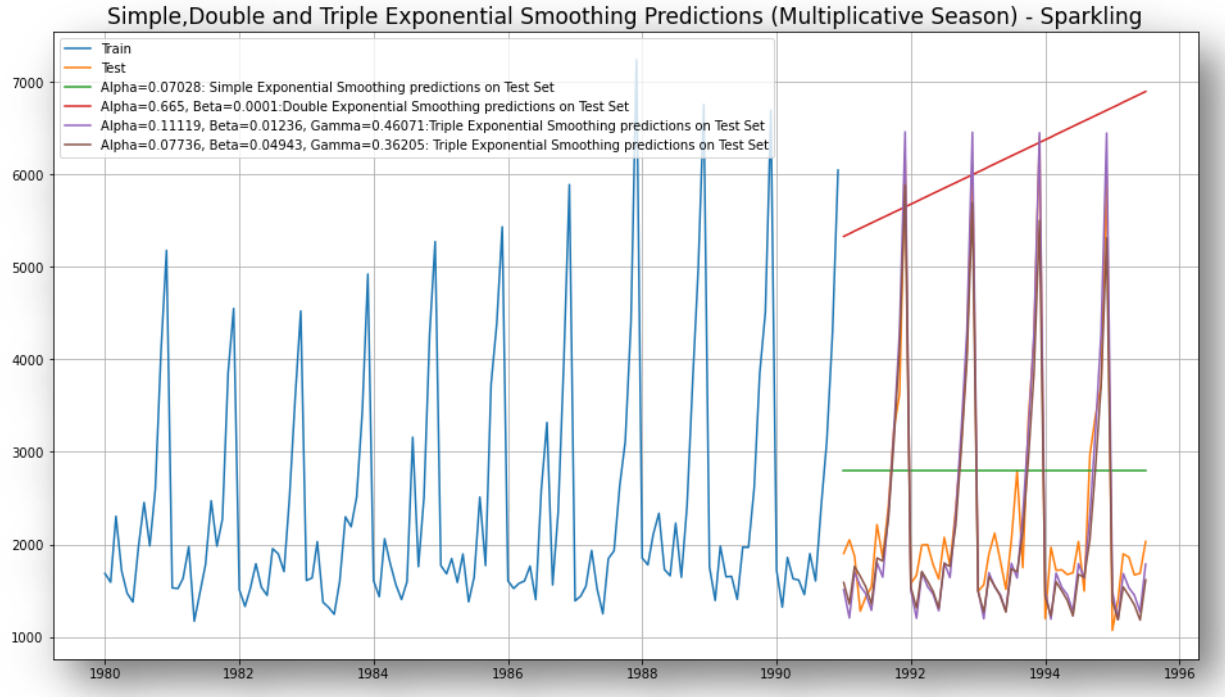


TABLE 21: TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE TRENDS, MULTIPLICATIVE SEASONALITY - ETS (A, A, M) RMSE

	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	36.796228	1338.012144
Double Exponential Smoothing	15.269328	3949.993290
Triple Exponential Smoothing (Additive Season)	14.265713	379.695686
Triple Exponential Smoothing (Multiplicative Season)	20.190998	406.510170

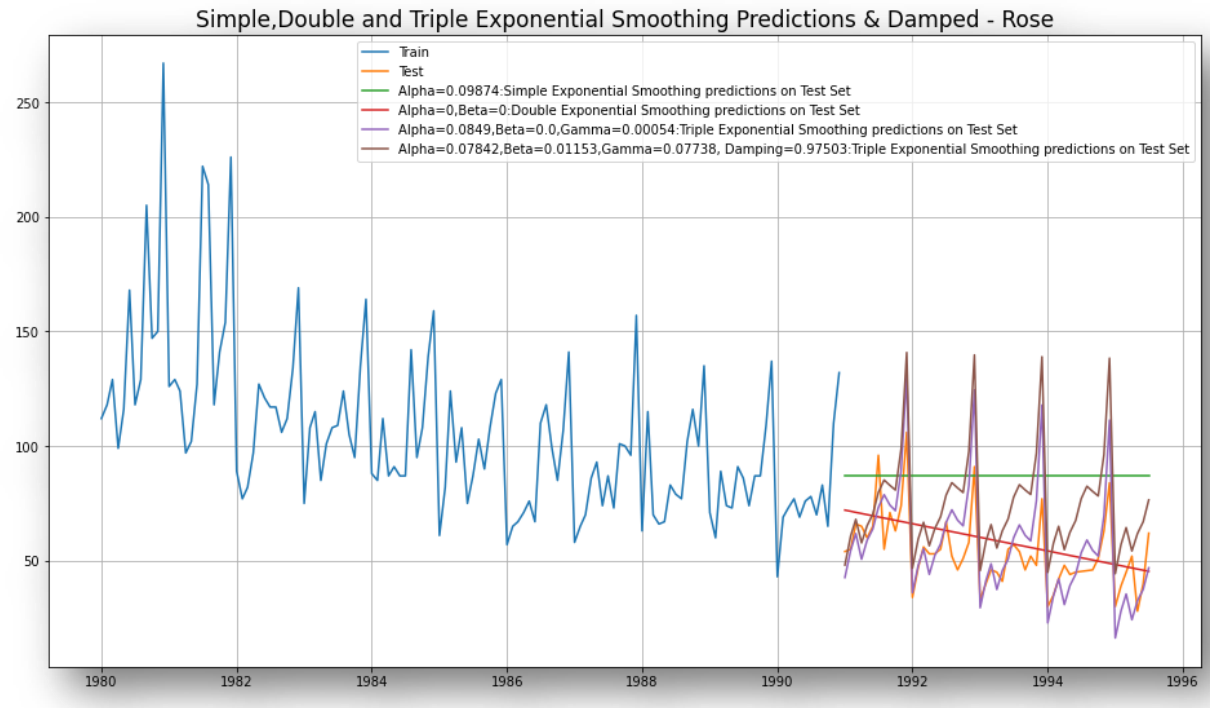
- **The Best Parameters for Dataset Rose:**

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.07132109562890512, 'smoothing_trend': 0.04553831096563722, 'smoothing_seasonal': 8.356711212063695e-07,
'damping_trend': nan, 'initial_level': 134.25655591779326, 'initial_trend': -0.8038265942903572, 'initial_seasons': array([0.83
746068, 0.94985307, 1.03812083, 0.90732186, 1.02043162,
       1.11131741, 1.22228039, 1.30104211, 1.23132915, 1.20610008,
       1.40577823, 1.93832412]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

- **The Best Parameters for Dataset Sparkling:**

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.11194572287706502, 'smoothing_trend': 0.04979454913988668, 'smoothing_seasonal': 0.3616765678435302, 'dam
ping_trend': nan, 'initial_level': 2356.340229937152, 'initial_trend': -10.519480221963526, 'initial_seasons': array([0.7146511
8, 0.68302129, 0.90263858, 0.80589958, 0.65660325,
       0.65654363, 0.88525948, 1.132562  , 0.92225104, 1.21110112,
       1.8820382 , 2.38194187]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

FIGURE 22: TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE DAMPED TRENDS, ADDITIVE SEASONALITY - ETS (A, AD, A)



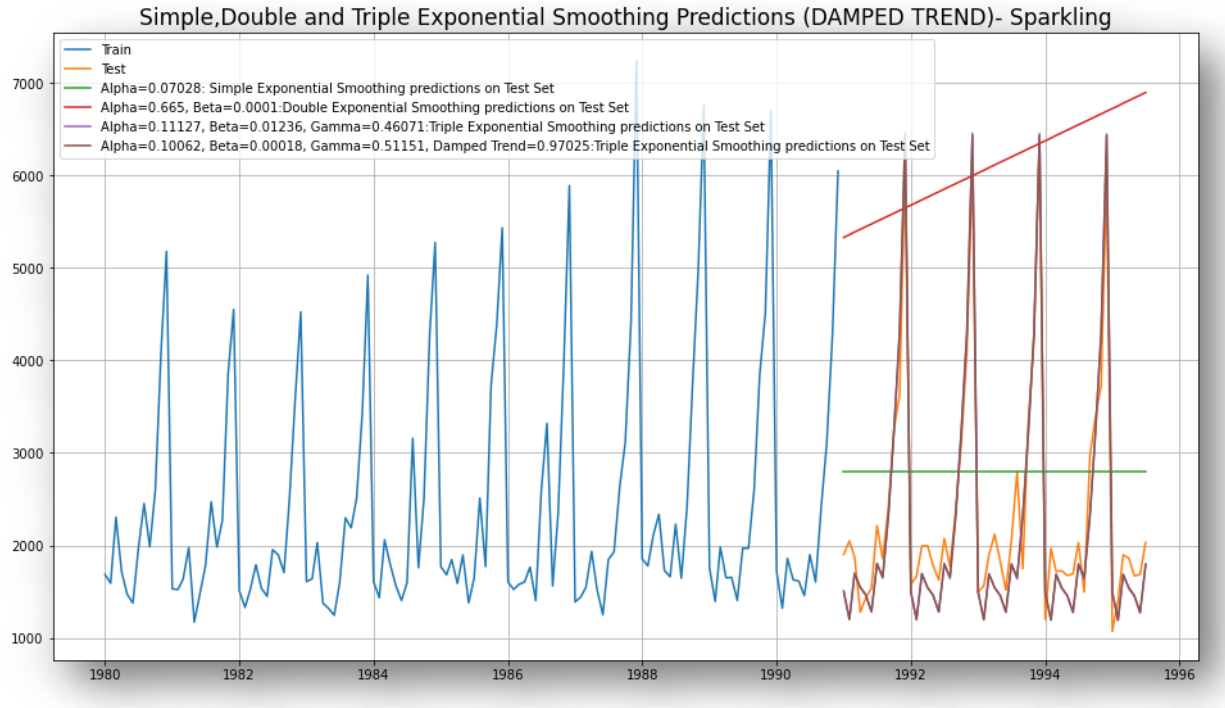


TABLE 22: TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE DAMPED TRENDS, ADDITIVE SEASONALITY - ETS (A, AD, A) RMSE

	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	36.796228	1338.012144
Double Exponential Smoothing	15.269328	3949.993290
Triple Exponential Smoothing (Additive Season)	14.265713	379.695686
Triple Exponential Smoothing (Multiplicative Season)	20.190998	406.510170
Triple Exponential Smoothing (Additive Season, Damped Trend)	25.660960	379.695686

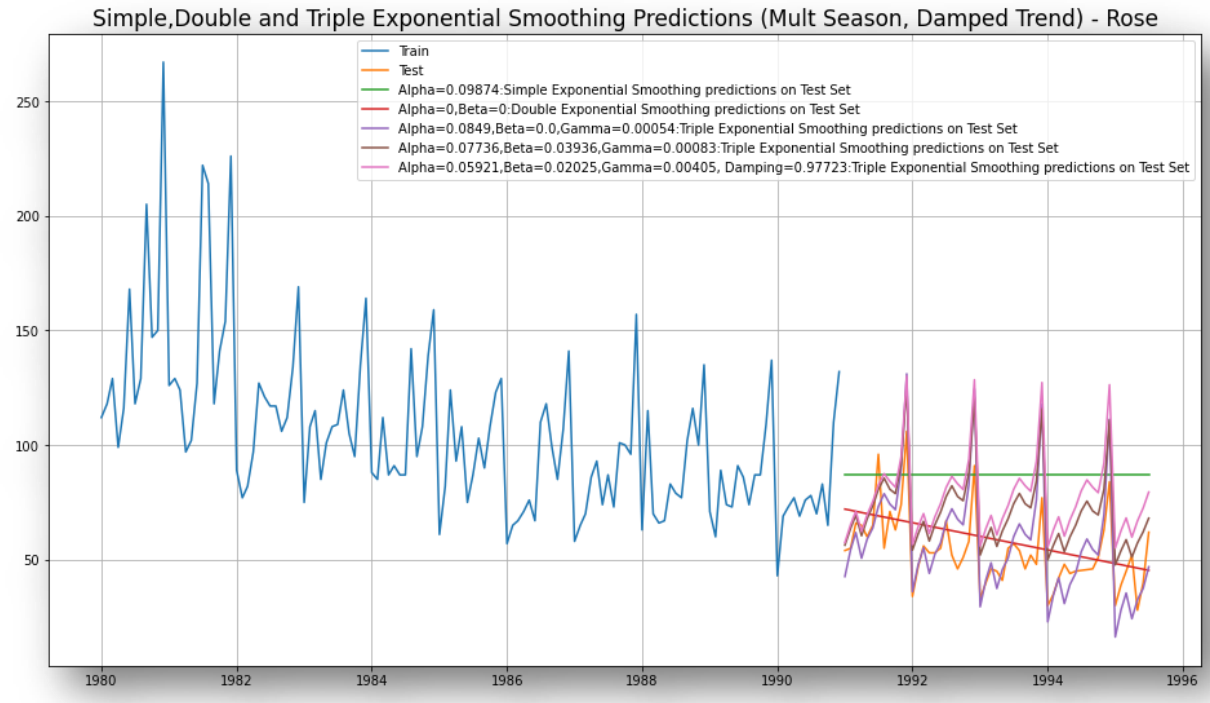
- **The Best Parameters for Dataset Rose:**

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.07848124317186382, 'smoothing_trend': 0.011340806743727252, 'smoothing_seasonal': 0.07684518378585875, 'd
amping_trend': 0.9774271738153909, 'initial_level': 153.42198975882621, 'initial_trend': -1.4893377479142316, 'initial_season
s': array([-30.39209525, -18.87017253, -10.88179619, -22.89643544,
       -13.58410591,  -6.64926869,   3.49196919,  10.51827686,
         6.1270493 ,   3.40480373,  21.30972085,  66.57108112]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

- **The Best Parameters for Dataset Sparkling:**

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.1057383191297317, 'smoothing_trend': 0.00014115807384965632, 'smoothing_seasonal': 0.48697109949814266,
'damping_trend': 0.9796782568408767, 'initial_level': 2361.578047768269, 'initial_trend': -1.9467315454588507, 'initial_season
s': array([-645.76716436, -730.40122176, -382.32172765, -478.92581014,
       -817.89058936, -824.7179298 , -385.27429991,   83.30682471,
       -250.13814372,  268.73180207, 1562.43394424, 2606.0934519 ]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

FIGURE 23: TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE DAMPED TRENDS, MULTIPLICATIVE SEASONALITY - ETS (A, AD, M).



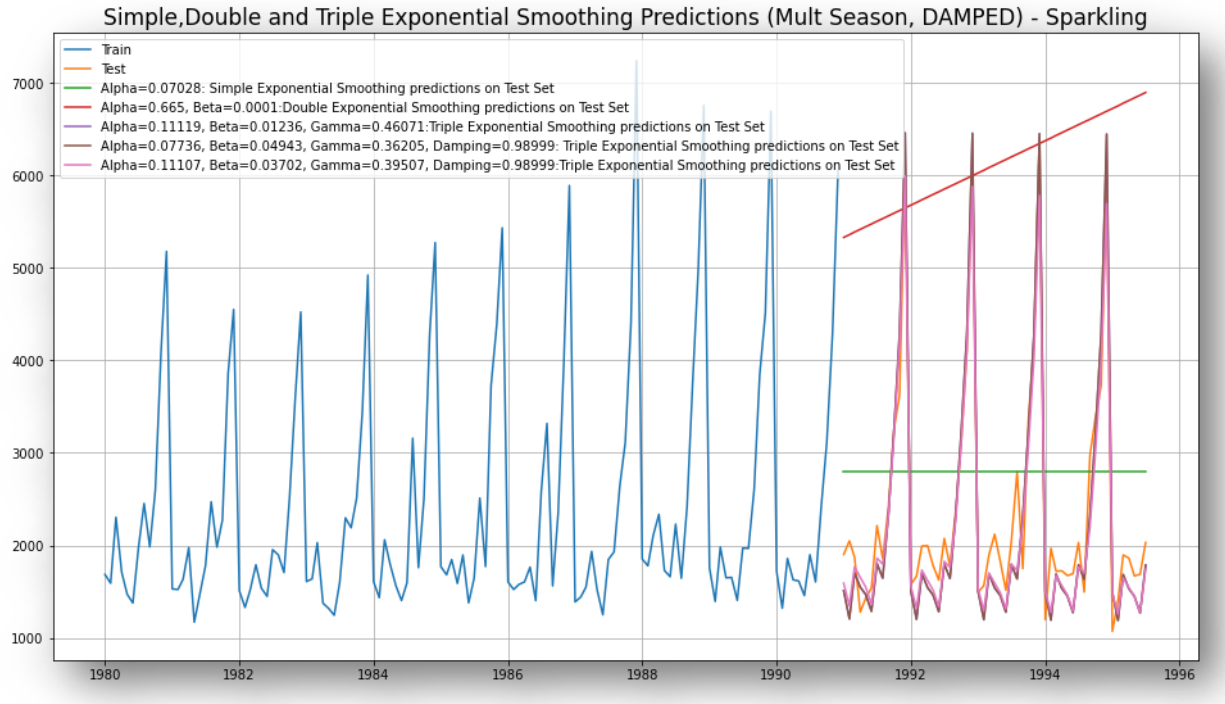


TABLE 23: TRIPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS, ADDITIVE DAMPED TRENDS, MULTIPLICATIVE SEASONALITY - ETS (A, AD, M) RMSE

	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	36.796228	1338.012144
Double Exponential Smoothing	15.269328	3949.993290
Triple Exponential Smoothing (Additive Season)	14.265713	379.695686
Triple Exponential Smoothing (Multiplicative Season)	20.190998	406.510170
Triple Exponential Smoothing (Additive Season, Damped Trend)	25.660960	379.695686
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)	26.295981	352.443335

- **The Best Parameters for Dataset Rose:**

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.0002560473477695915, 'smoothing_trend': 3.3814318768645266e-07, 'smoothing_seasonal': 0.00016960815725286
792, 'damping_trend': 0.9790757197728251, 'initial_level': 165.96547270159138, 'initial_trend': -1.9265981623841522, 'initial_s
easons': array([0.70787477, 0.80549593, 0.87787457, 0.7704003 , 0.86151917,
       0.93514919, 1.0250353 , 1.08732193, 1.04699153, 1.01656203,
       1.18353599, 1.62323663]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

- **The Best Parameters for Dataset Sparkling:**

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.1110714622899222, 'smoothing_trend': 0.03702380844127912, 'smoothing_seasonal': 0.39507957727536136, 'dam
ping_trend': 0.9899999825826437, 'initial_level': 2356.5418308172734, 'initial_trend': -9.179892630347588, 'initial_seasons': a
rray([0.713876  , 0.68479146, 0.89985055, 0.80522628, 0.65413878,
       0.65498002, 0.88128754, 1.12310179, 0.91373324, 1.1919948 ,
       1.848147  , 2.33628145]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

- **We conclude that models with least RMSE:**
    1. **Best Model for Rose: 2 Pt Moving Average**
    2. **Best Model for Sparkling: Holt-Winter Damped Trend ETS (A, Ad, M)**

FIGURE 24: BEST MODEL FOR ROSE

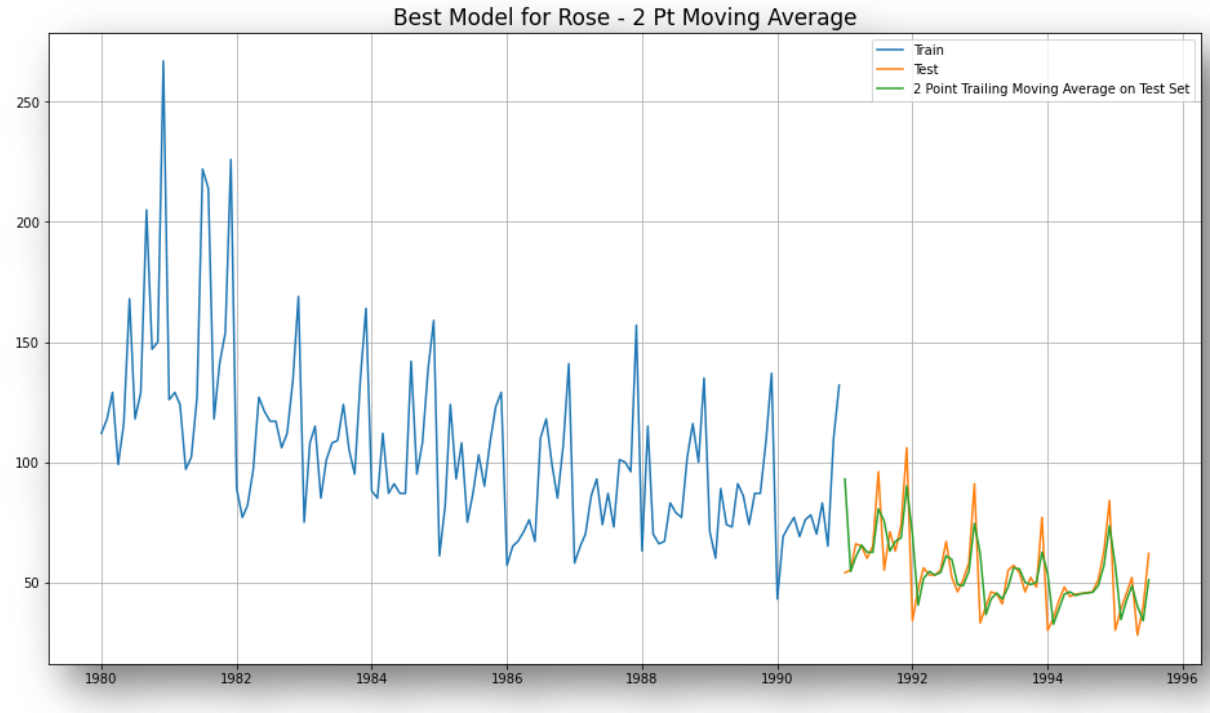
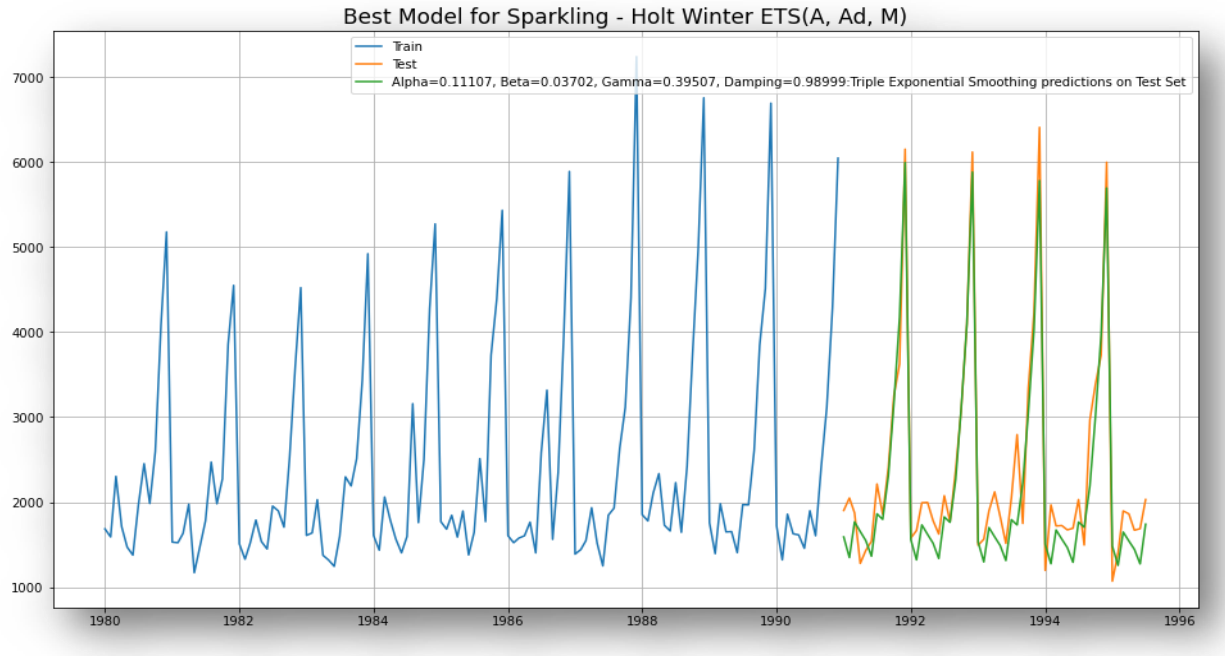


FIGURE 25: BEST MODEL FOR SPARKLING



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.

- In ADF Test, if p-value < alpha then we reject the Null Hypothesis and hence conclude that given Time Series is Stationary
- If p-value > alpha then we fail to reject the Null Hypothesis and hence conclude that given Time Series is Not Stationary
- If the T.S is not stationary then we apply one level of differencing and check for stationarity and if further there is no stationarity then we shall again apply one more level of differencing.
- Usually with two level of differencing the model becomes Stationary.
- Once it is stationary we apply ARIMA/SARIMA to the model.
- **Applying one level we get the following output for Rose Dataset:**

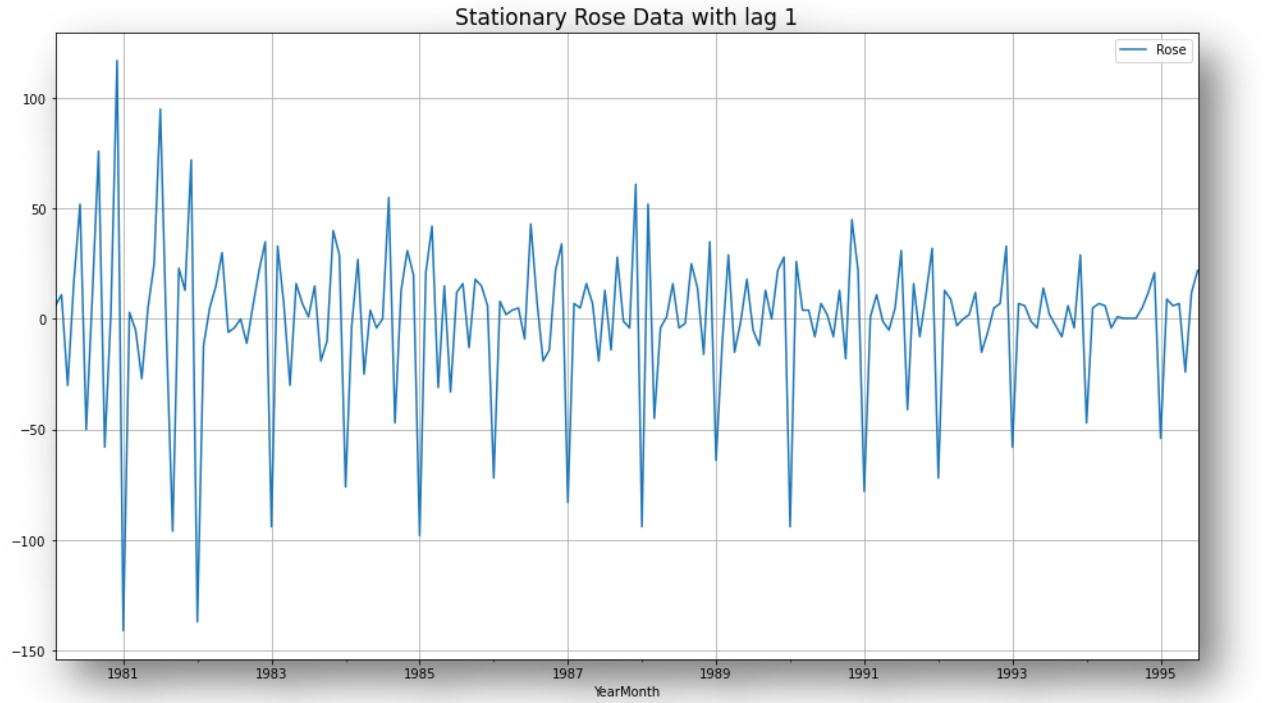
```
DF test statistic is -2.240
DF test p-value is 0.4671371627793208
Number of lags used 13
```

- **Applying two level (due non stationarity in first attempt) we get the following output for Rose Dataset:**

```
DF test statistic is -8.162
DF test p-value is 3.015976115826596e-11
Number of lags used 12
```

- **The Rose Dataset is stationary and ready for further forecasting.**

FIGURE 26: STATIONARITY OF ROSE WITH LAG 1



- **Applying one level we get the following output for Sparkling Dataset:**

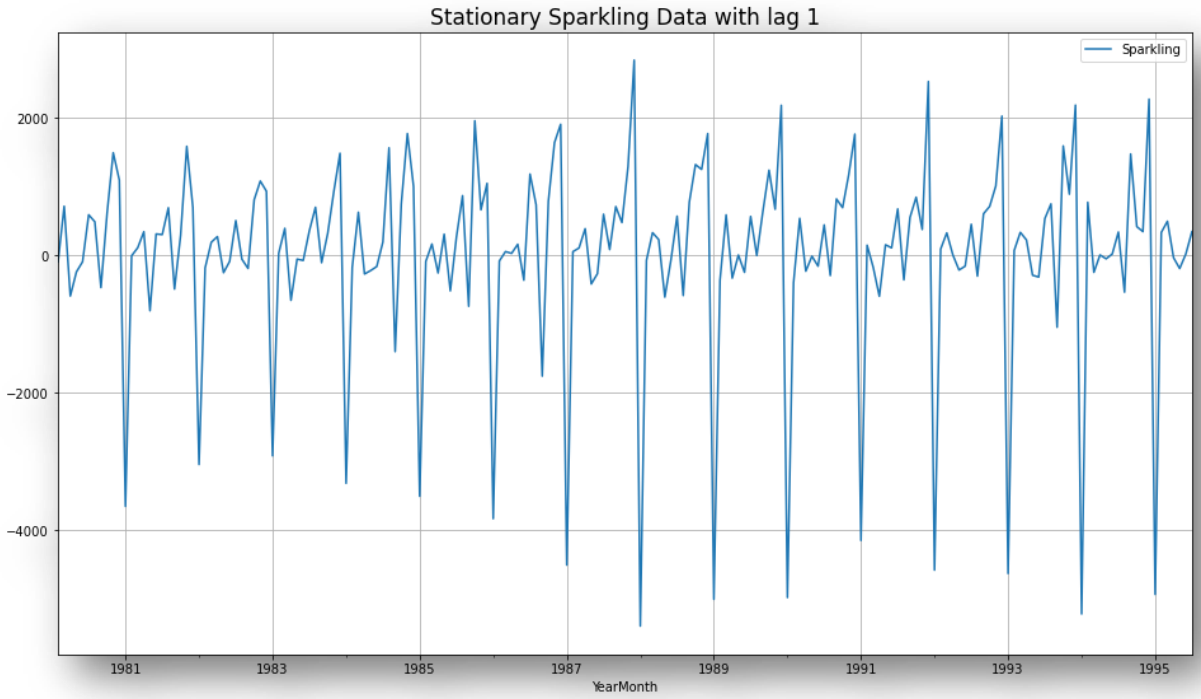
```
DF test statistic is -1.798
DF test p-value is 0.7055958459932035
Number of lags used 12
```

- **Applying two level (due non stationarity in first attempt) we get the following output for Sparkling Dataset:**

```
DF test statistic is -44.912
DF test p-value is 0.0
Number of lags used 10
```

- **The Sparkling Dataset is stationary and ready for further forecasting.**

FIGURE 27: STATIONARITY OF SPARKLING WITH LAG 1



Q.6. Build an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

Note: The data has some seasonality so ideally we should build a SARIMA model. But let's test on the lowest Akaike Information Criteria, which one suits best - ARIMA or SARIMA.

- After the stationarity check, we apply ARIMA automated / SARIMA automated.
- ARIMA automated: Here a grid is made including all possible combinations of p, d, q, where p =Range of q = 0 to 3 and d (constant) =1
- From that grid we choose that combination with the least Akaike Information Criteria (AIC).
- Then ARIMA is fitted into the train set and forecasted on the test set.
- Lastly, the RMSE is checked. **ARIMA AUTOMATED**

TABLE 24: AKAIKE INFORMATION CRITERIA FOR ROSE

```
ARIMA(0, 1, 0) - AIC:1333.1546729124348
ARIMA(0, 1, 1) - AIC:1282.3098319748315
ARIMA(0, 1, 2) - AIC:1279.6715288535752
ARIMA(0, 1, 3) - AIC:1280.5453761734652
ARIMA(1, 1, 0) - AIC:1317.3503105381526
ARIMA(1, 1, 1) - AIC:1280.5742295380076
ARIMA(1, 1, 2) - AIC:1279.8707234231915
ARIMA(1, 1, 3) - AIC:1281.870722330997
ARIMA(2, 1, 0) - AIC:1298.6110341604908
ARIMA(2, 1, 1) - AIC:1281.507862186858
ARIMA(2, 1, 2) - AIC:1281.8707222264168
```

```
ARIMA(2, 1, 3) - AIC:1274.6953190416875
ARIMA(3, 1, 0) - AIC:1297.4810917271702
ARIMA(3, 1, 1) - AIC:1282.419277627203
ARIMA(3, 1, 2) - AIC:1283.720740597714
ARIMA(3, 1, 3) - AIC:1278.6543993387522
```

TABLE 25: EXTRACTING THE LEAST AIC FOR ROSE

	param	AIC
11	(2, 1, 3)	1274.695319
15	(3, 1, 3)	1278.654399
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

- From the above it is seen that, the best parameter is (2,1,3) with an AIC of 1274.69 which is least amongst all.

FIGURE 28: DIAGNOSTIC PLOT FOR ROSE

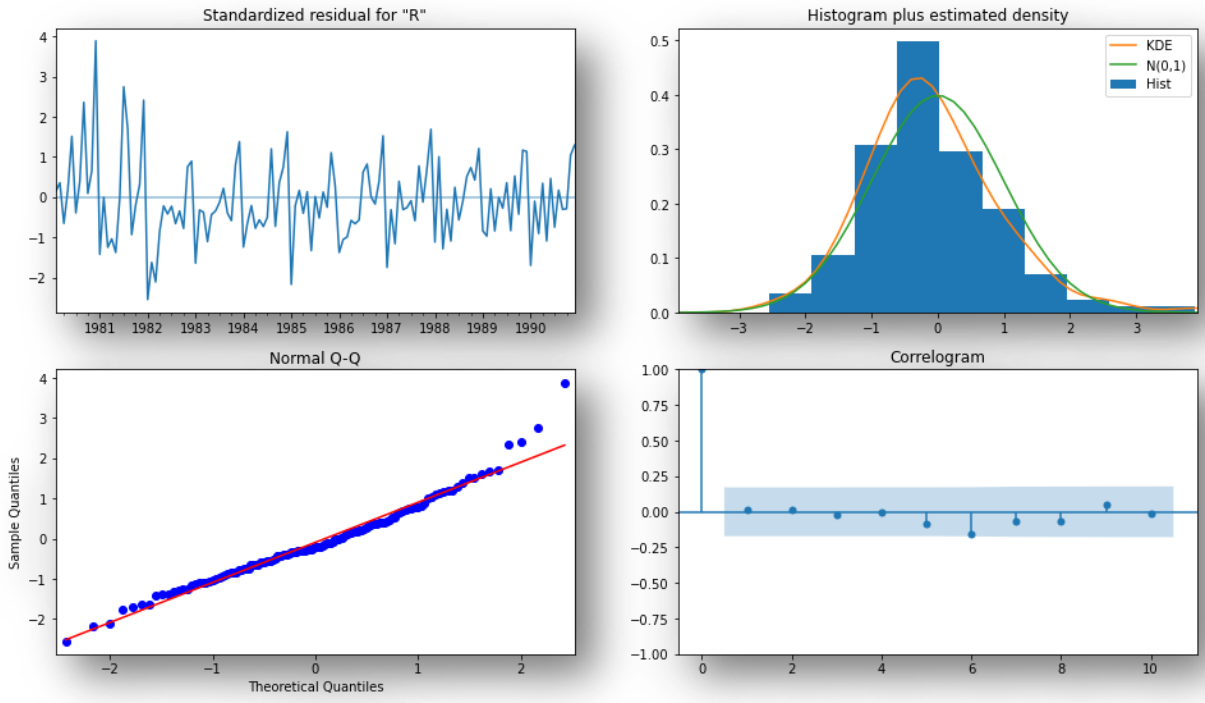


FIGURE 29: FORECASTED PLOT FOR ROSE (ARIMA)

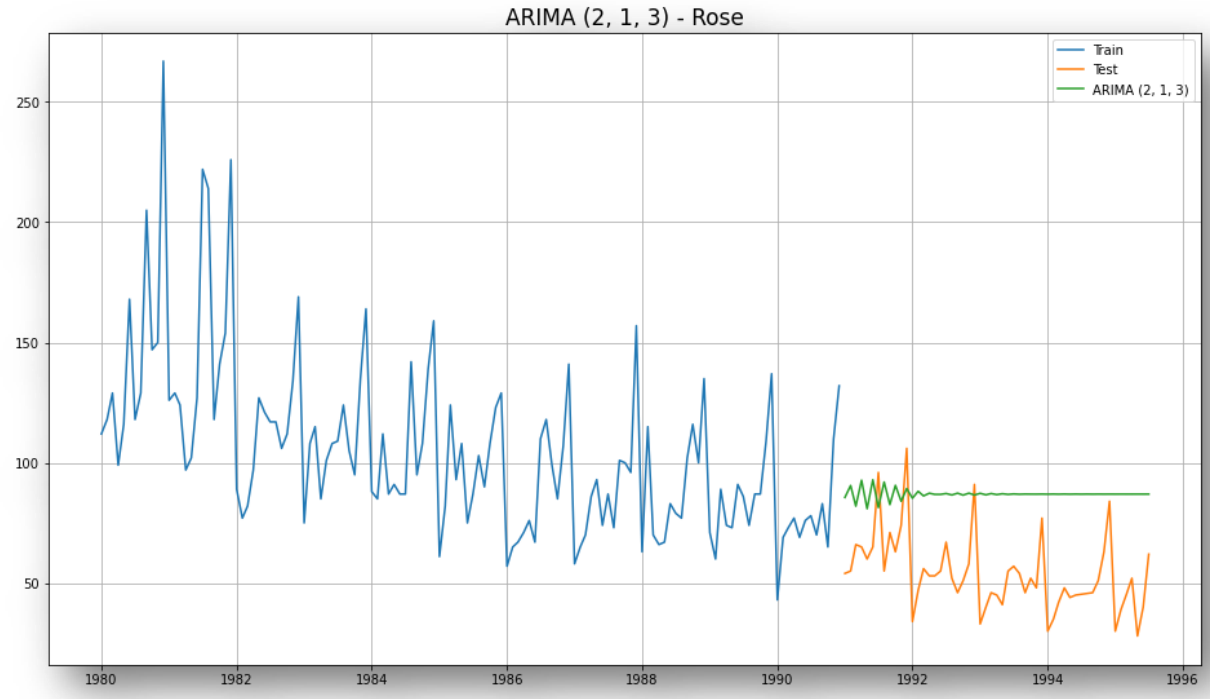


TABLE 26: RMSE AND MAPE SCORE

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,1,3)	36.816347	75.845836

TABLE 27: AKAIKE INFORMATION CRITERIA -SPARKLING

```
ARIMA(0, 1, 0) - AIC:2267.6630357855465
ARIMA(0, 1, 1) - AIC:2263.060015591336
ARIMA(0, 1, 2) - AIC:2234.408323130674
ARIMA(0, 1, 3) - AIC:2233.9948577476116
ARIMA(1, 1, 0) - AIC:2266.6085393190087
ARIMA(1, 1, 1) - AIC:2235.7550946742404
ARIMA(1, 1, 2) - AIC:2234.5272004519366
```

```
ARIMA(1, 1, 3) - AIC:2235.6078101124103
ARIMA(2, 1, 0) - AIC:2260.365743968097
ARIMA(2, 1, 1) - AIC:2233.7776262581274
ARIMA(2, 1, 2) - AIC:2213.50921703971
```

```
ARIMA(2, 1, 3) - AIC:2232.983057575394
ARIMA(3, 1, 0) - AIC:2257.72337899794
ARIMA(3, 1, 1) - AIC:2235.4989865071907
```

```
ARIMA(3, 1, 2) - AIC:2230.7572943437854
ARIMA(3, 1, 3) - AIC:2221.4519770502657
```

EXTRACTING THE LEAST AIC - SPARKLING

- From table 26 we can see that, parameter (2,1,2) with the least AIC of 2213.51 is the best parameter for the dataset.

FIGURE 31: DIAGNOSTIC PLOT FOR SPARKLING

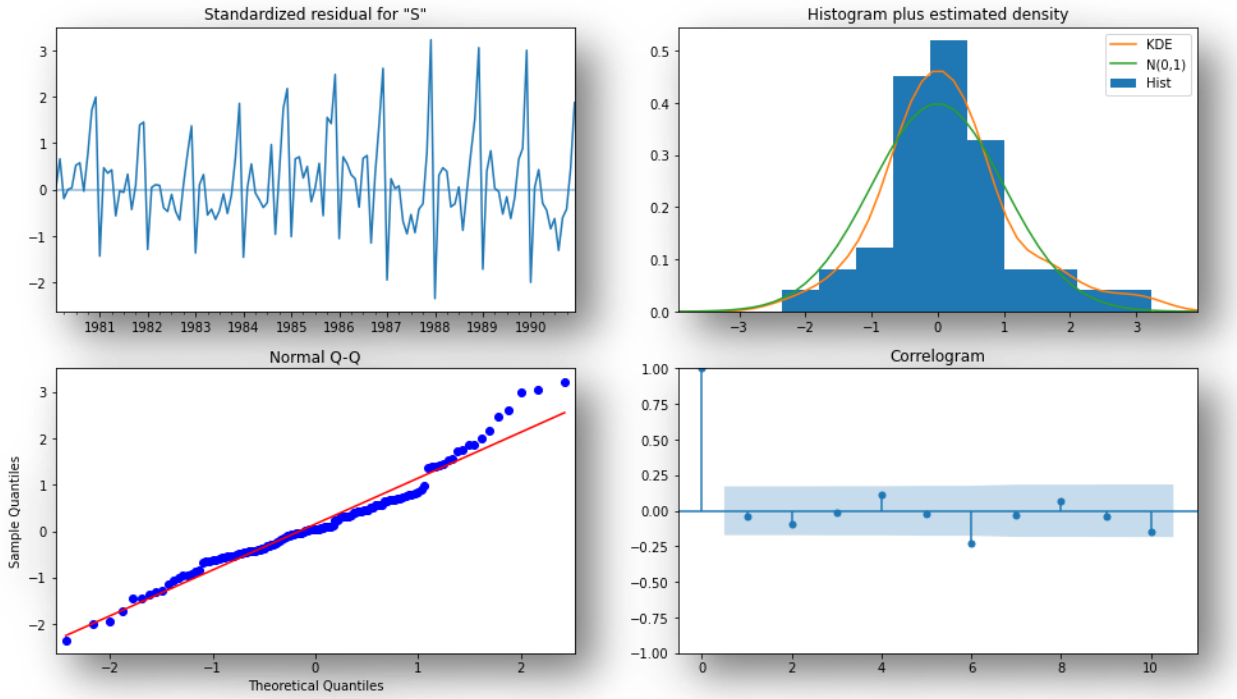


FIGURE 31: FORECASTED PLOT FOR SPARKLING

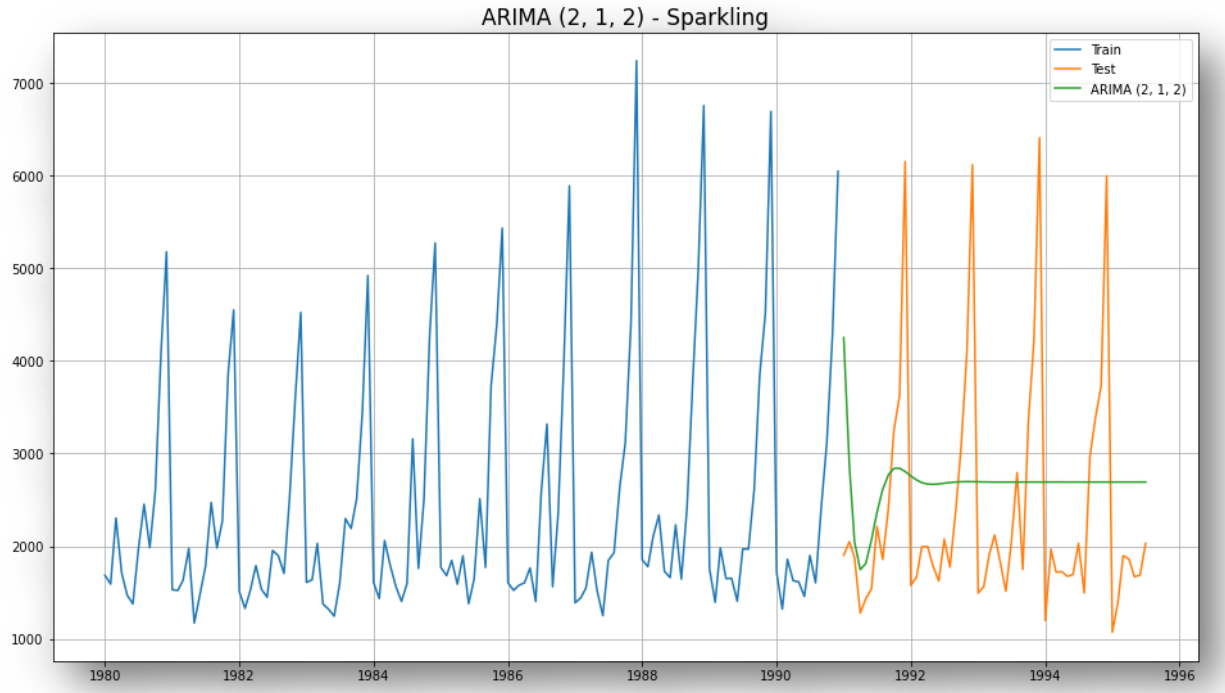


TABLE 28: RMSE AND MAPE SCORE

	RMSE	MAPE
ARIMA(2,1,2)	1299.980373	47.100017

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.
- 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.

TABLE 29: EXTRACTING THE LEAST AIC (SARIMA) - ROSE

	param	seasonal	AIC
222	(3, 1, 1)	(3, 0, 2, 12)	774.400287
238	(3, 1, 2)	(3, 0, 2, 12)	774.880935
220	(3, 1, 1)	(3, 0, 0, 12)	775.426699
221	(3, 1, 1)	(3, 0, 1, 12)	775.495330
252	(3, 1, 3)	(3, 0, 0, 12)	775.561019

FIGURE 32: DIAGNOSTIC PLOT (SARIMA) - ROSE

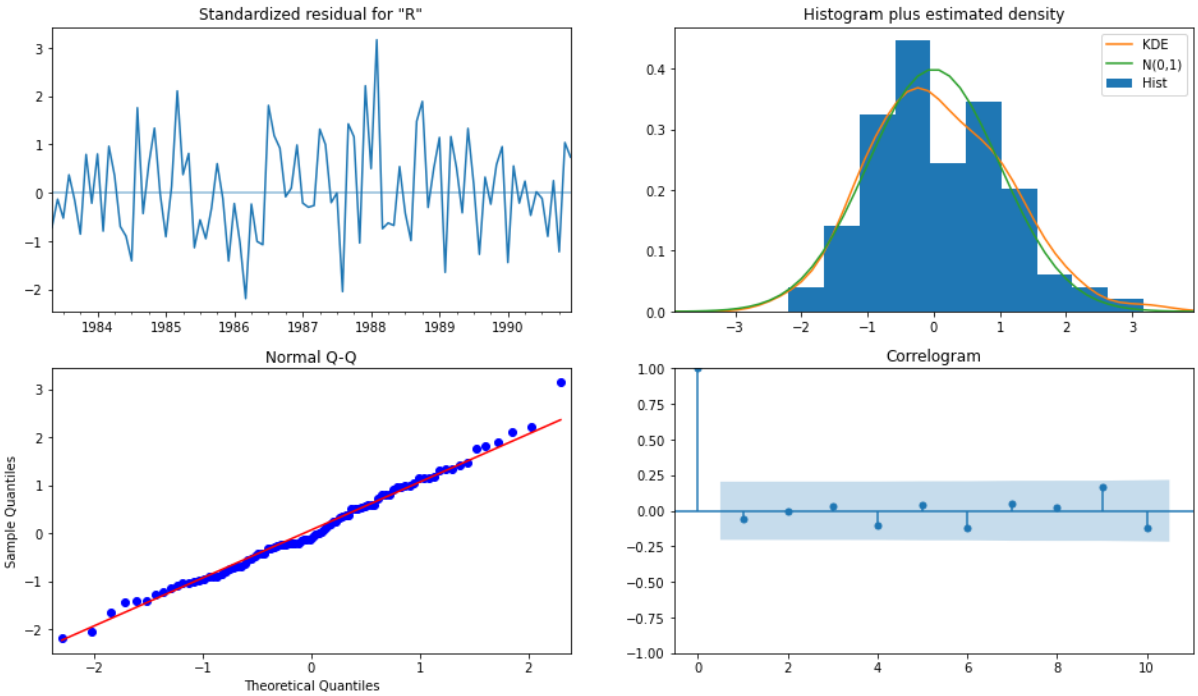


TABLE 30: RMSE AND MAPE SCORE (SARIMA) - ROSE

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,1,3)	36.816347	75.845836
ARIMA(2,1,2)	36.871197	76.056213
SARIMA(3, 1, 1)(3, 0, 2, 12)	18.882257	36.376727

FIGURE 33: FORECASTED PLOT (SARIMA) - ROSE

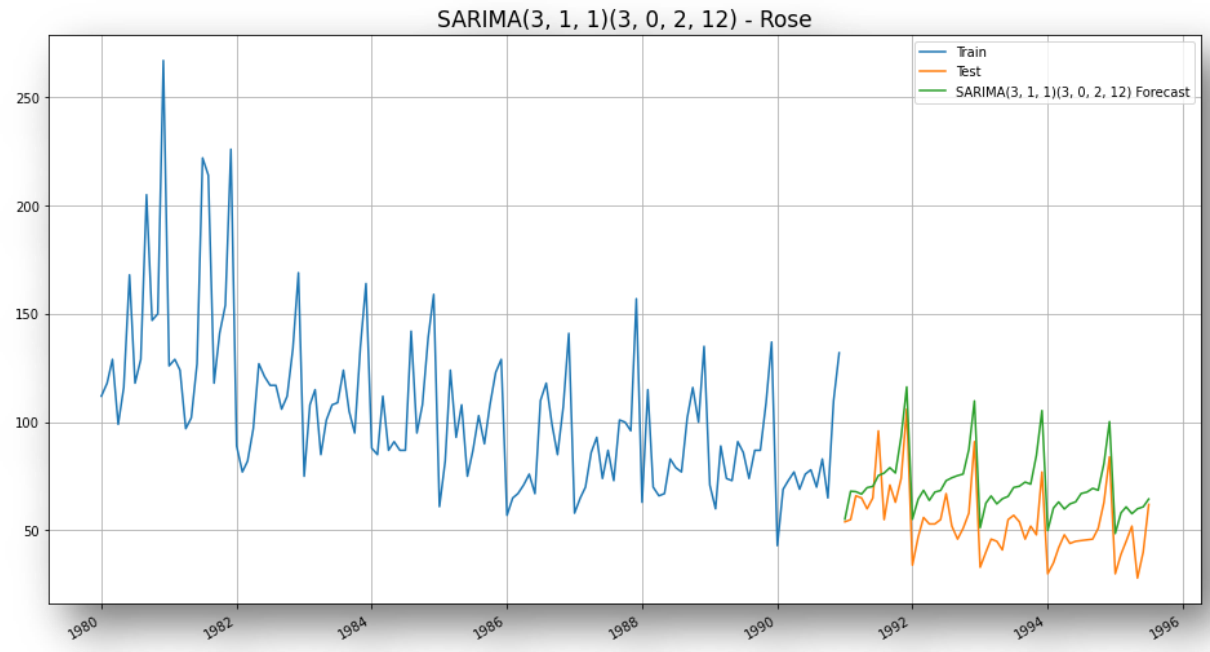


TABLE 31: EXTRACTING THE LEAST AIC (SARIMA) - SPARKLING

	param	seasonal	AIC
99	(1, 1, 2)	(0, 0, 3, 12)	14.000000
147	(2, 1, 1)	(0, 0, 3, 12)	14.000000
215	(3, 1, 1)	(1, 0, 3, 12)	208.047774
151	(2, 1, 1)	(1, 0, 3, 12)	347.672942
251	(3, 1, 3)	(2, 0, 3, 12)	349.867995
87	(1, 1, 1)	(1, 0, 3, 12)	718.878545
51	(0, 1, 3)	(0, 0, 3, 12)	1018.308738
220	(3, 1, 1)	(3, 0, 0, 12)	1387.788331

- From the above table we can observe that the best parameter (3,1,1) (3,0,0,12) with the least AIC of 1387.78.

FIGURE 34: DIAGNOSTIC PLOT (SARIMA) - SPARKLING

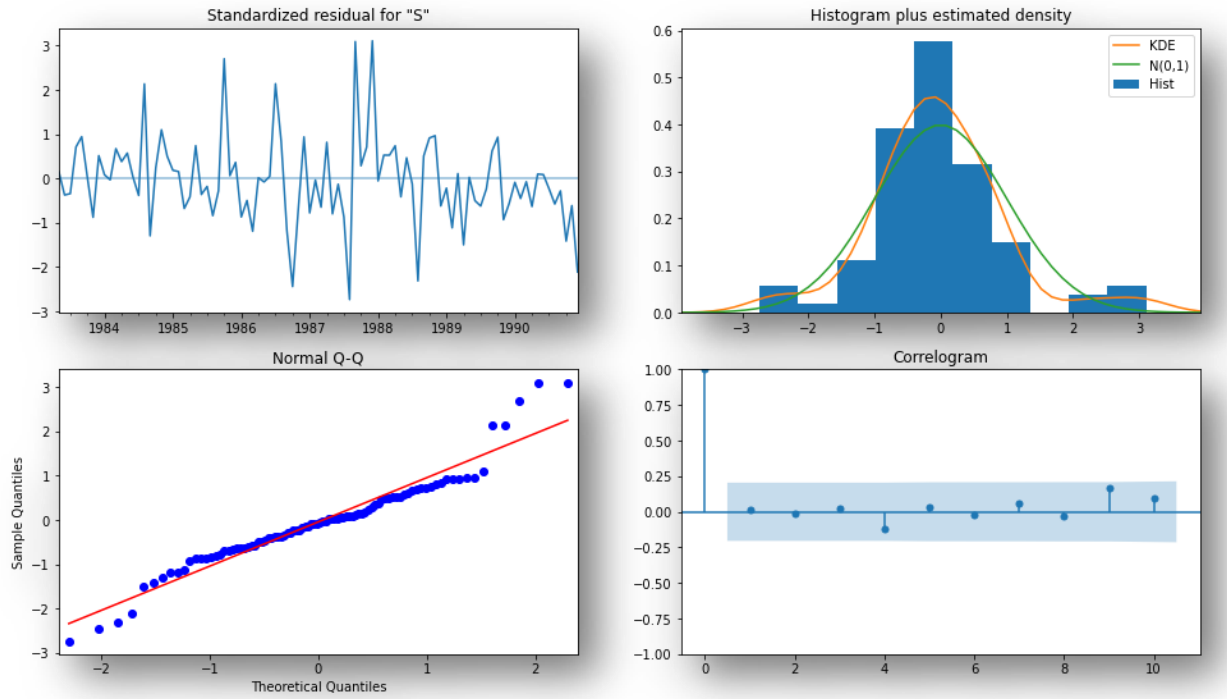


FIGURE 35: FORECASTED PLOT (SARIMA) - SPARKLING

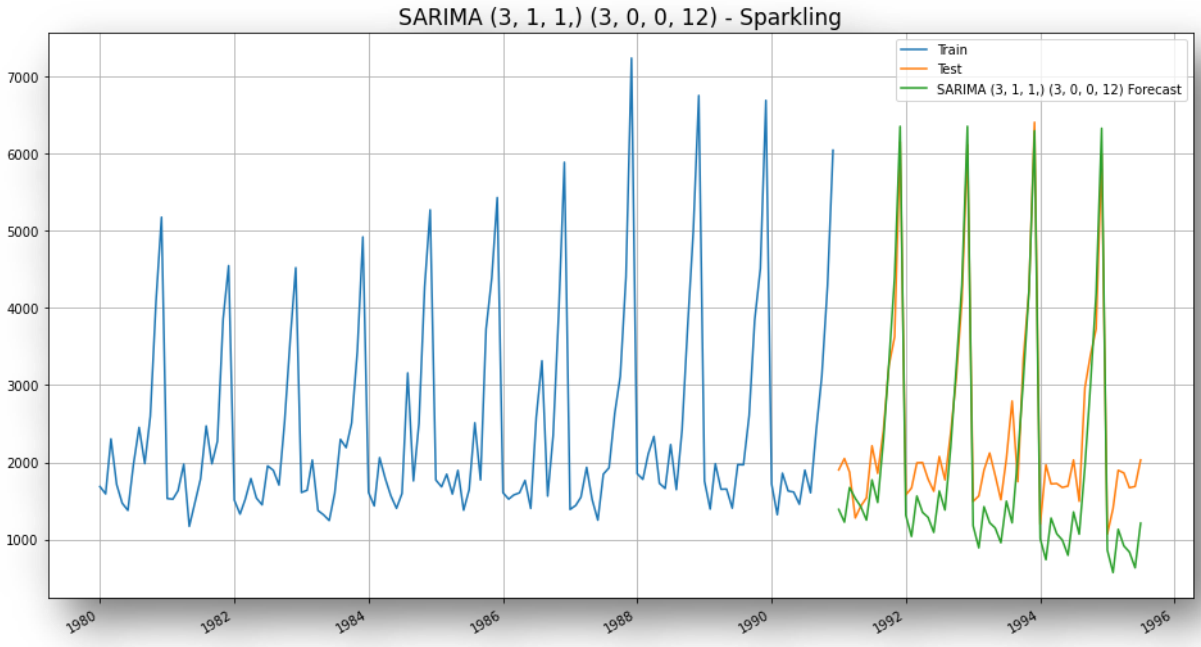


TABLE 32: RMSE AND MAPE SCORE (SARIMA) - SPARKLING

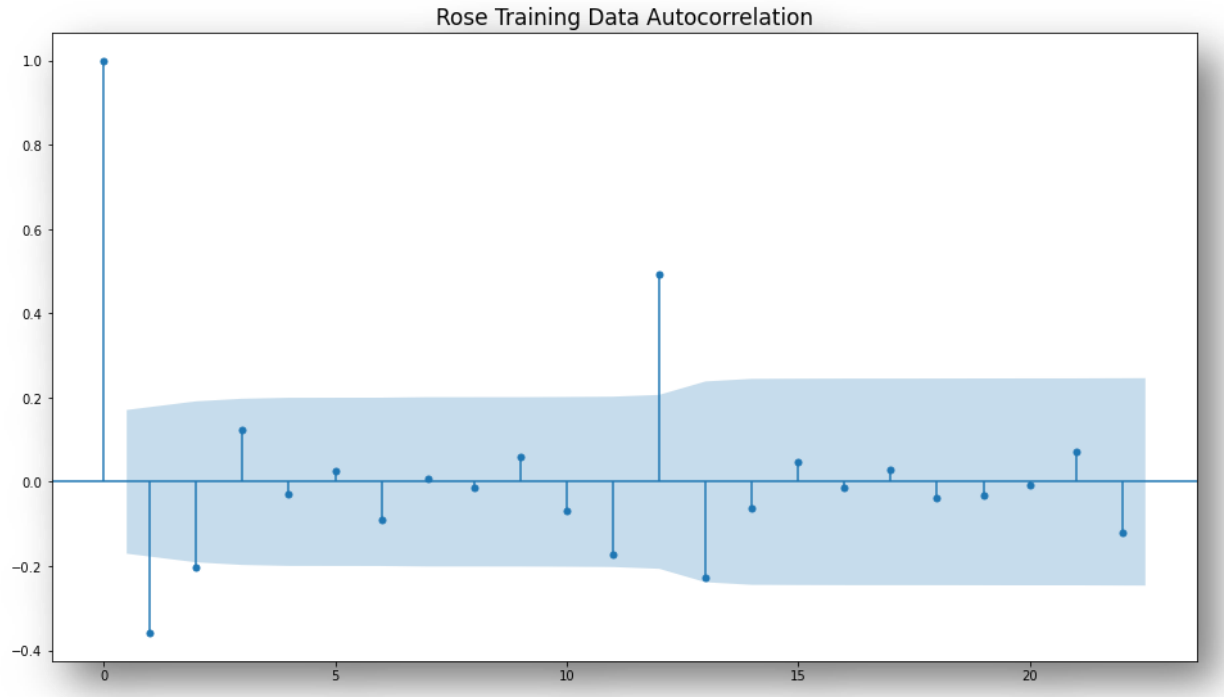
	RMSE	MAPE
ARIMA(2,1,2)	1299.980373	47.100017
ARIMA(0,1,0)	3864.279352	201.327650
SARIMA(3,1,1)(3,0,2,12)	601.244396	25.870721
SARIMA(3,1,1)(3,0,0,12)	35.911261	54.872536

- 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.

- Autocorrelation refers to how correlated a time series is with its past values. e.g., with also with 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.
- 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.

FIGURE 35: AUTO CORRELATION



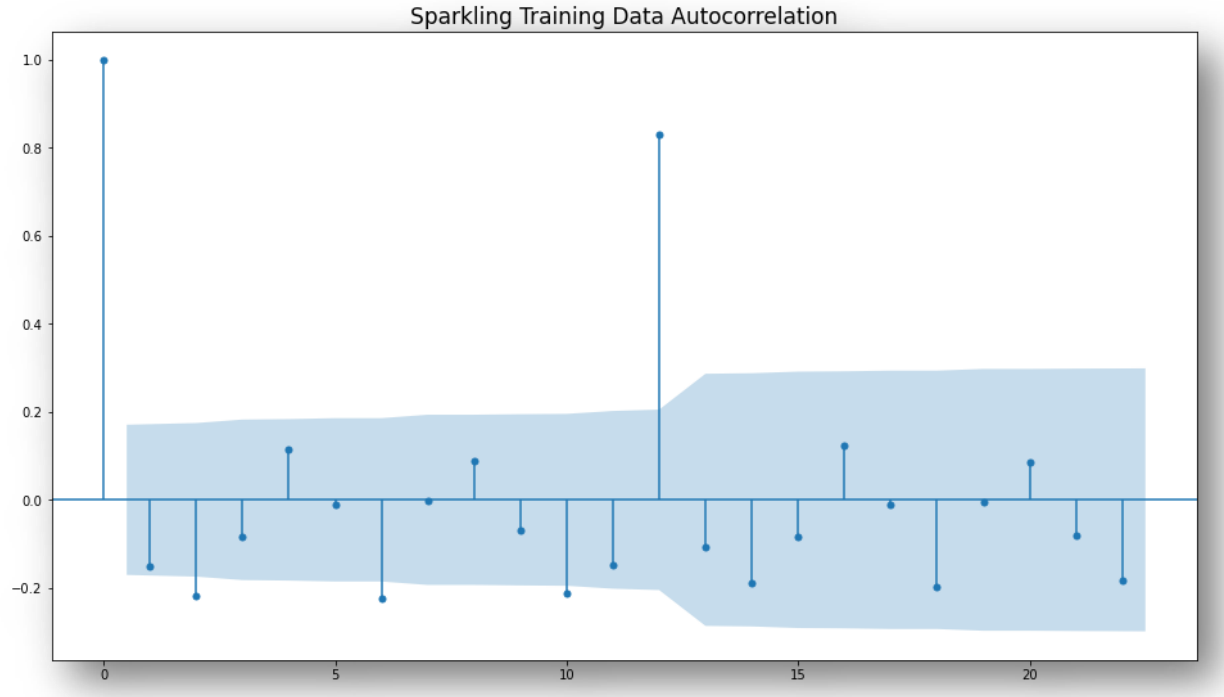
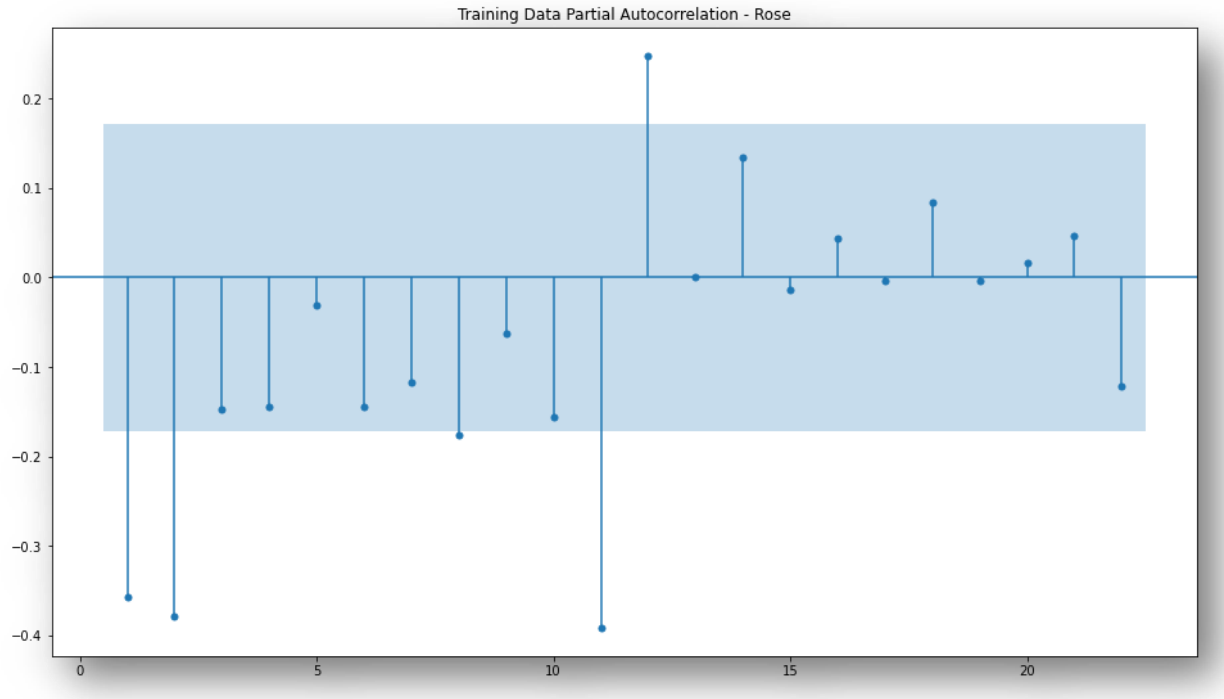


FIGURE 36: PARTIAL AUTO CORRELATION



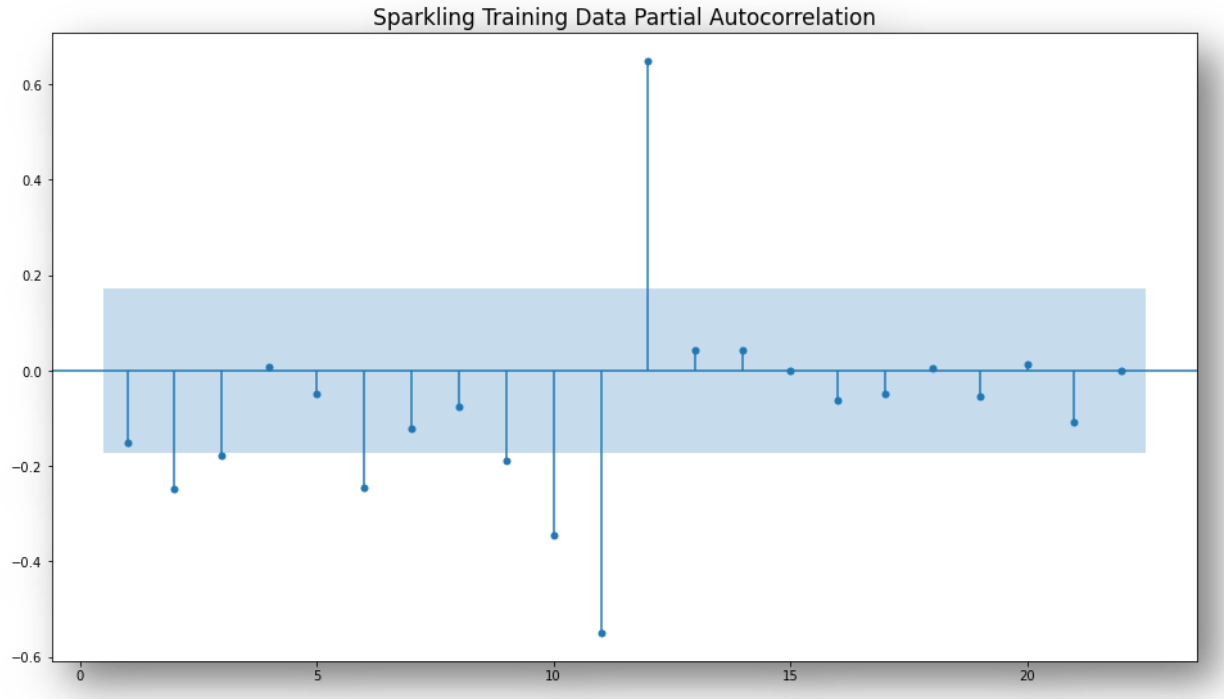


FIGURE 38: MANUAL ARIMA ROSE-DIAGNOSTIC PLOT

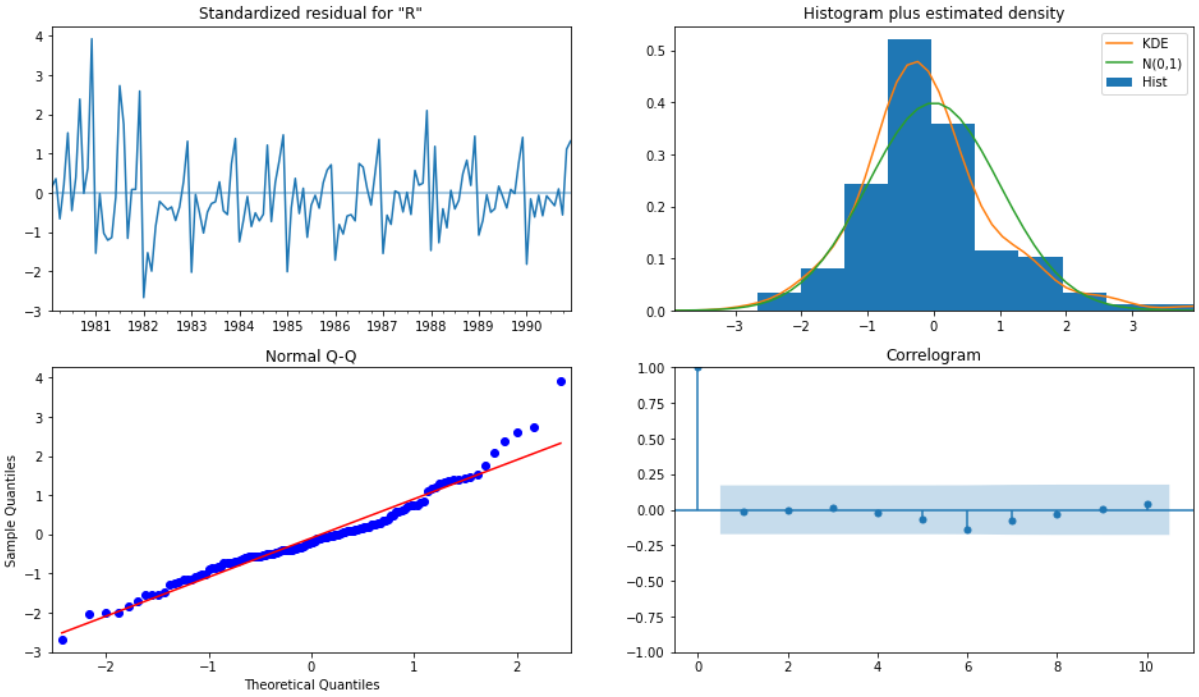


FIGURE 39: FORECASTED PLOT MANUAL ARIMA- ROSE

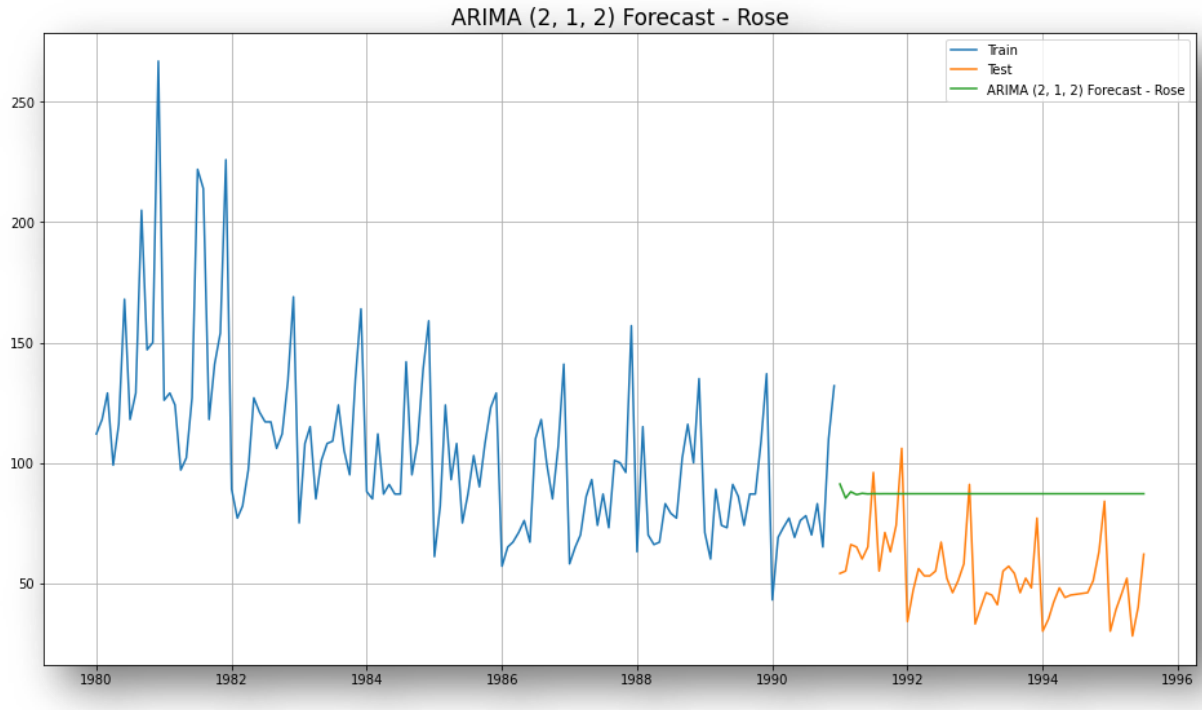


TABLE 33: RMSE AND MAPE SCORE (MANUAL ARIMA) - ROSE

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,1,3)	36.816347	75.845836
ARIMA(2,1,2)	36.871197	76.056213

FIGURE 40: MANUAL SARIMA ROSE-DIAGNOSTIC PLOT

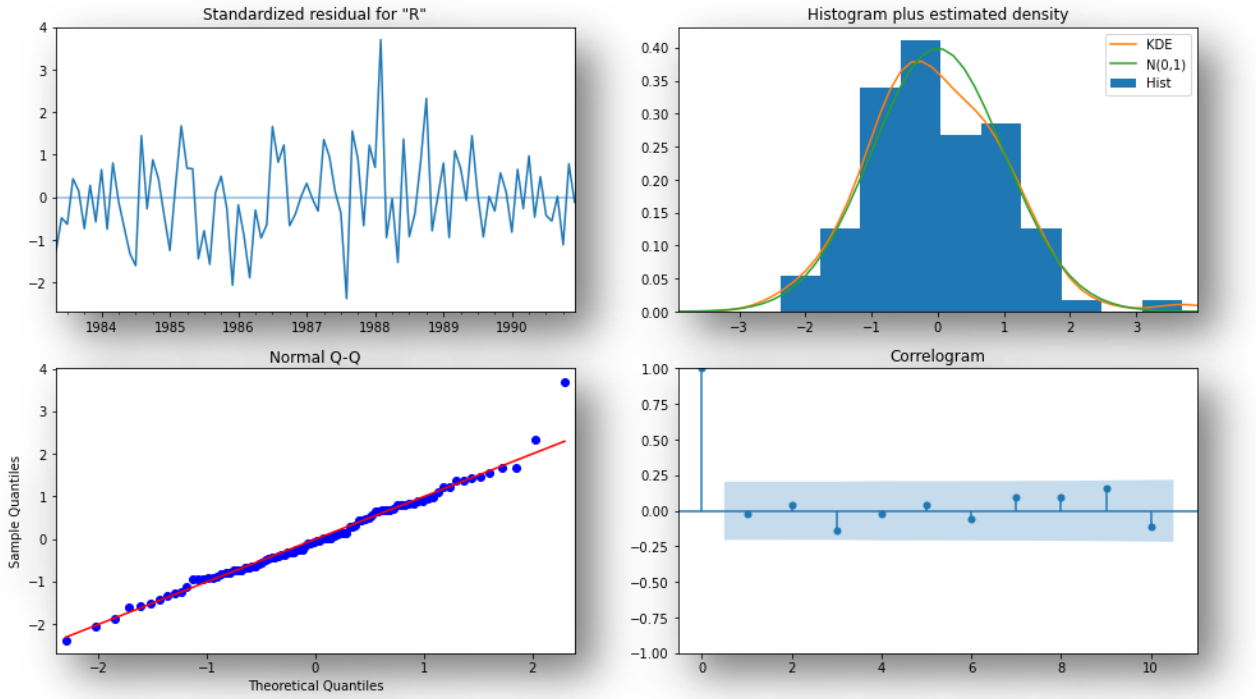


FIGURE 41: MANUAL SARIMA ROSE-FORECAST PLOT

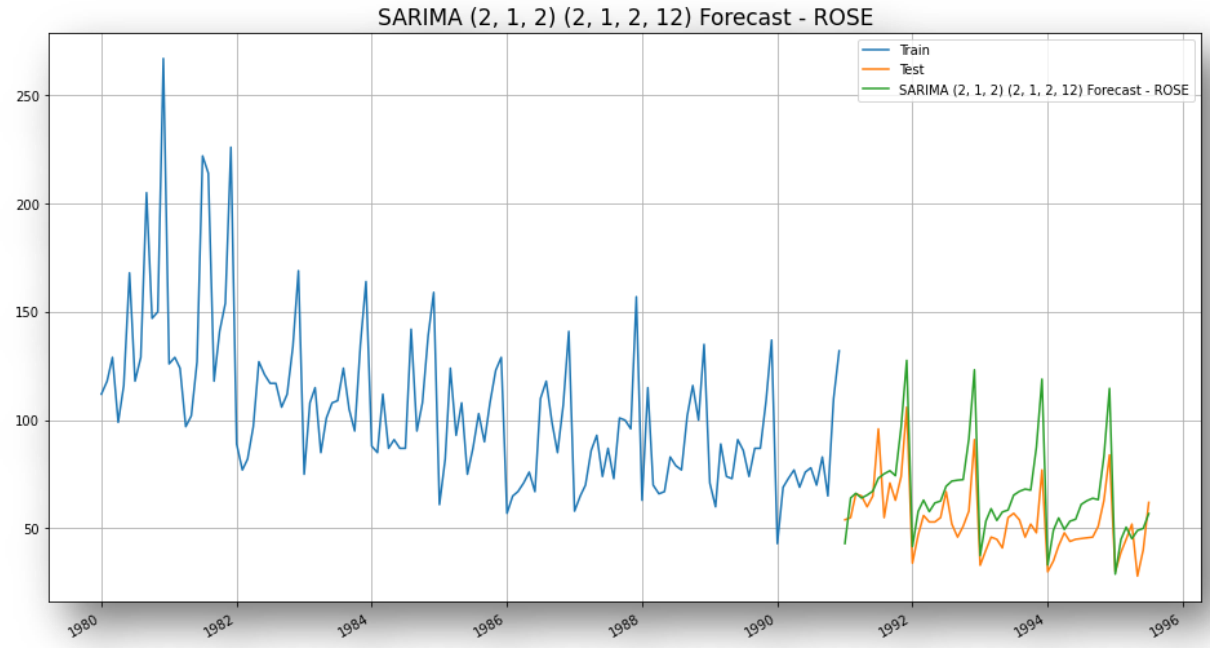


TABLE 34: RMSE AND MAPE SCORE (MANUAL SARIMA) - ROSE

RMSE:	16.551081104195777
MAPE:	25.47804525274619

FIGURE 42: MANUAL SARIMA ROSE-DIAGNOSTIC PLOT

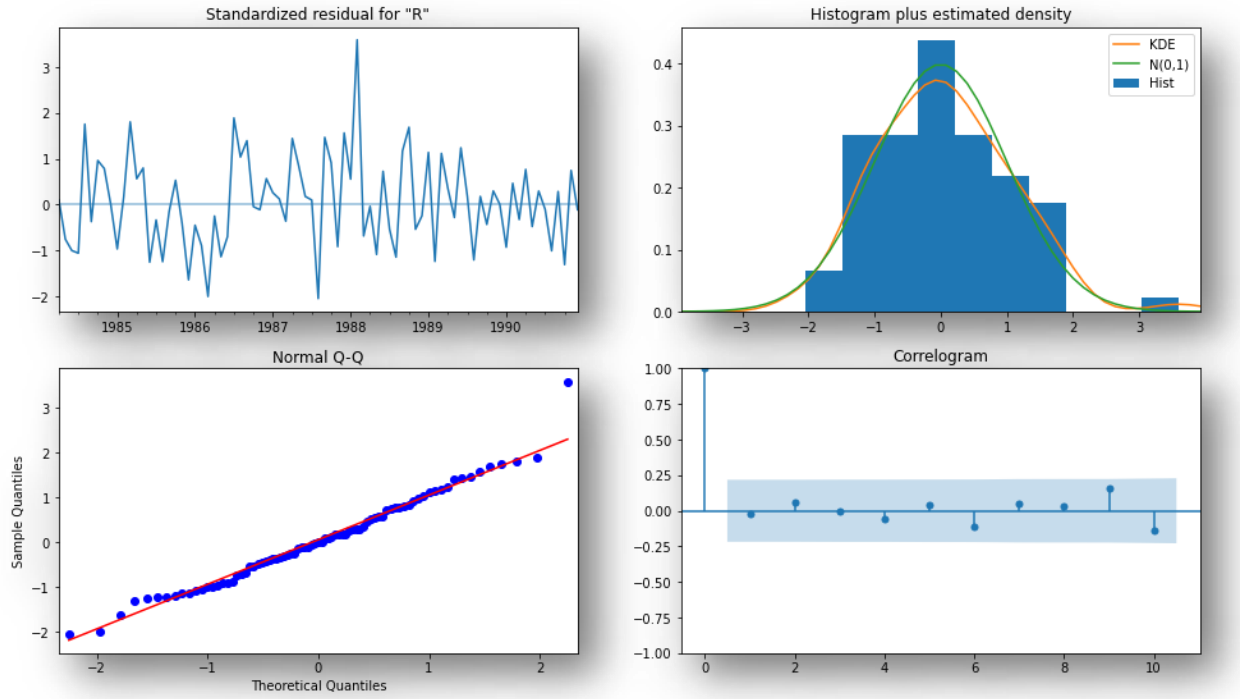


FIGURE 43: MANUAL SARIMA ROSE-FORECASTED PLOT

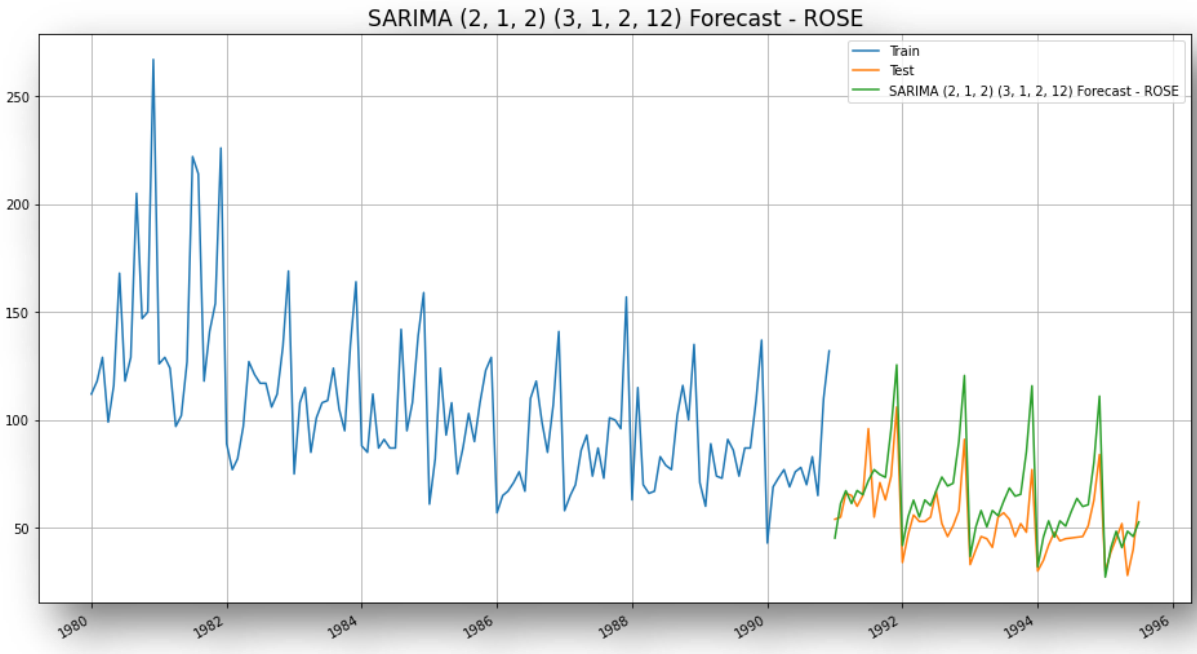


TABLE 35: RMSE AND MAPE SCORE (MANUAL SARIMA) - ROSE

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,1,3)	36.816347	75.845836
ARIMA(2,1,2)	36.871197	76.056213
SARIMA(3, 1, 1)(3, 0, 2, 12)	18.882257	36.376727
SARIMA(2,1,2)(3,1,2,12)	15.359185	22.960942

FIGURE 44: MANUAL ARIMA SPARKLING-DIAGNOSTIC PLOT

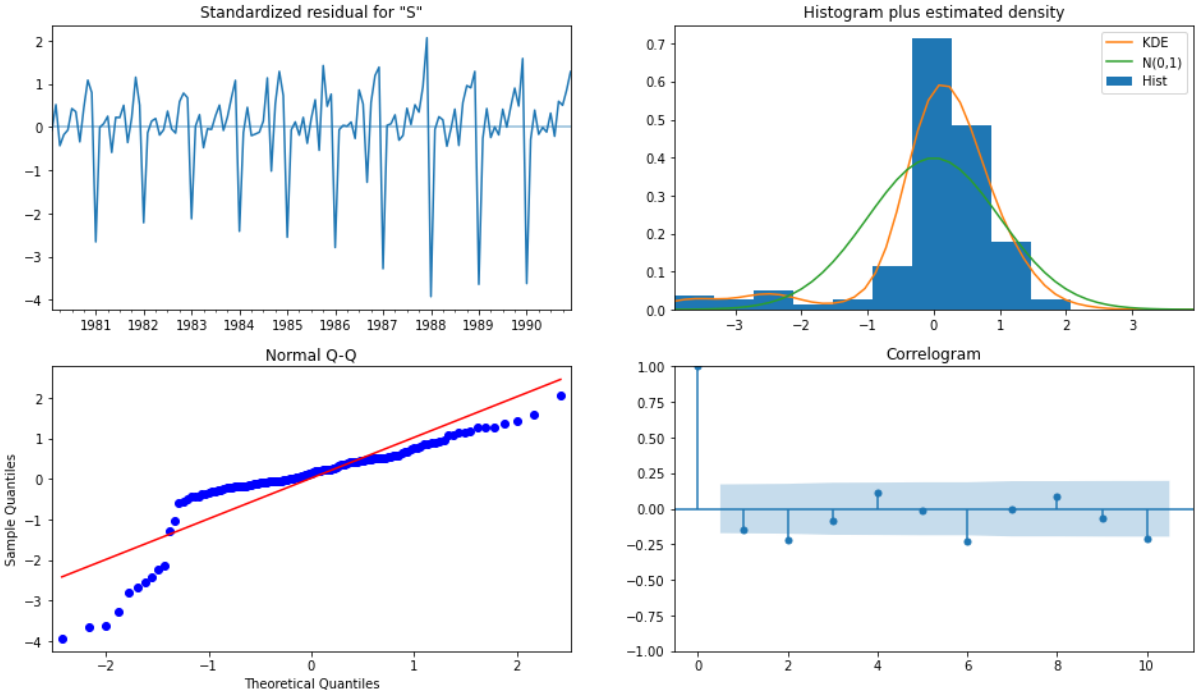


FIGURE 45: MANUAL ARIMA SPARKLING-FORECASTED PLOT

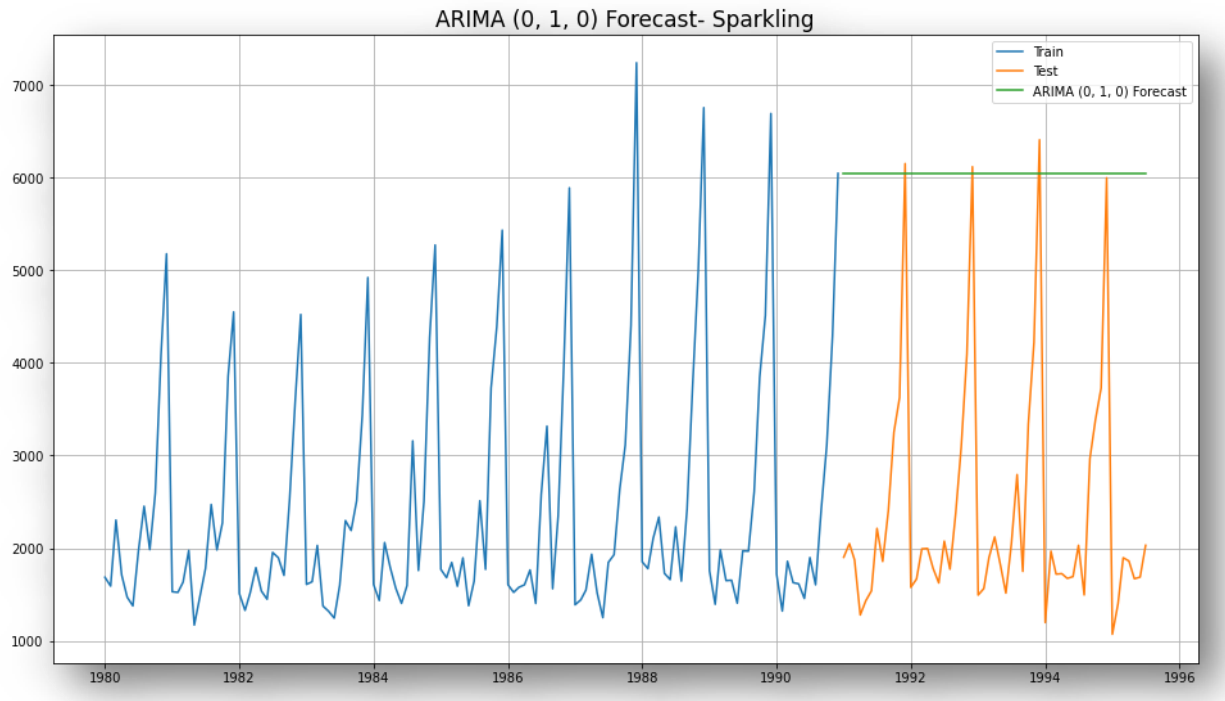


TABLE 36: RMSE AND MAPE SCORE (MANUAL ARIMA) – SPARKLING

	RMSE	MAPE
ARIMA(2,1,2)	1299.980373	47.100017
ARIMA(0,1,0)	3864.279352	201.327650

FIGURE 46: MANUAL SARIMA SPARKLING-DIAGNOSTIC PLOT

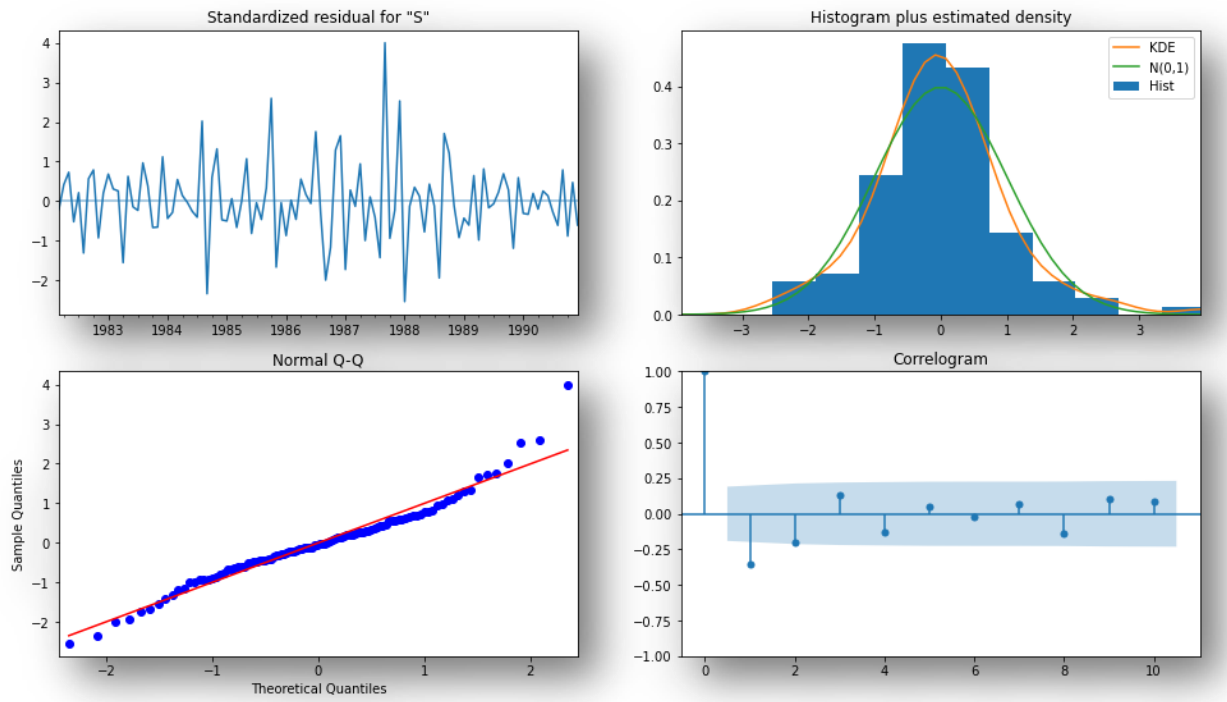


FIGURE 47: MANUAL SARIMA SPARKLING-FORECAST PLOT

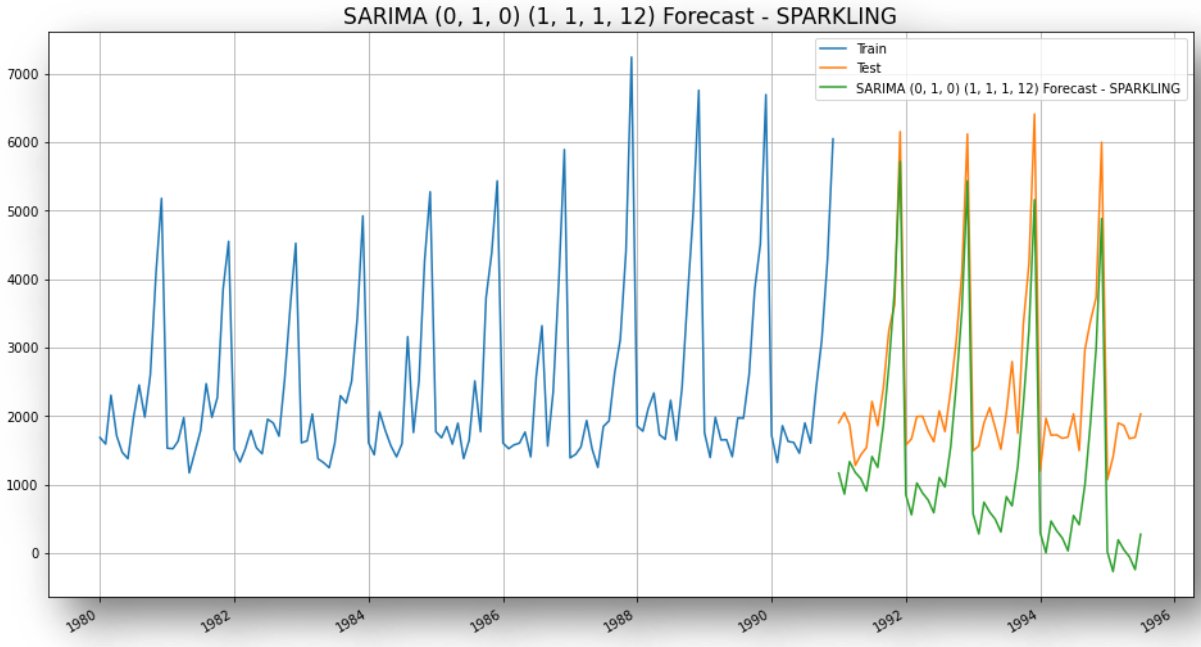


FIGURE 48: MANUAL SARIMA SPARKLING-DIAGNOSTIC PLOT

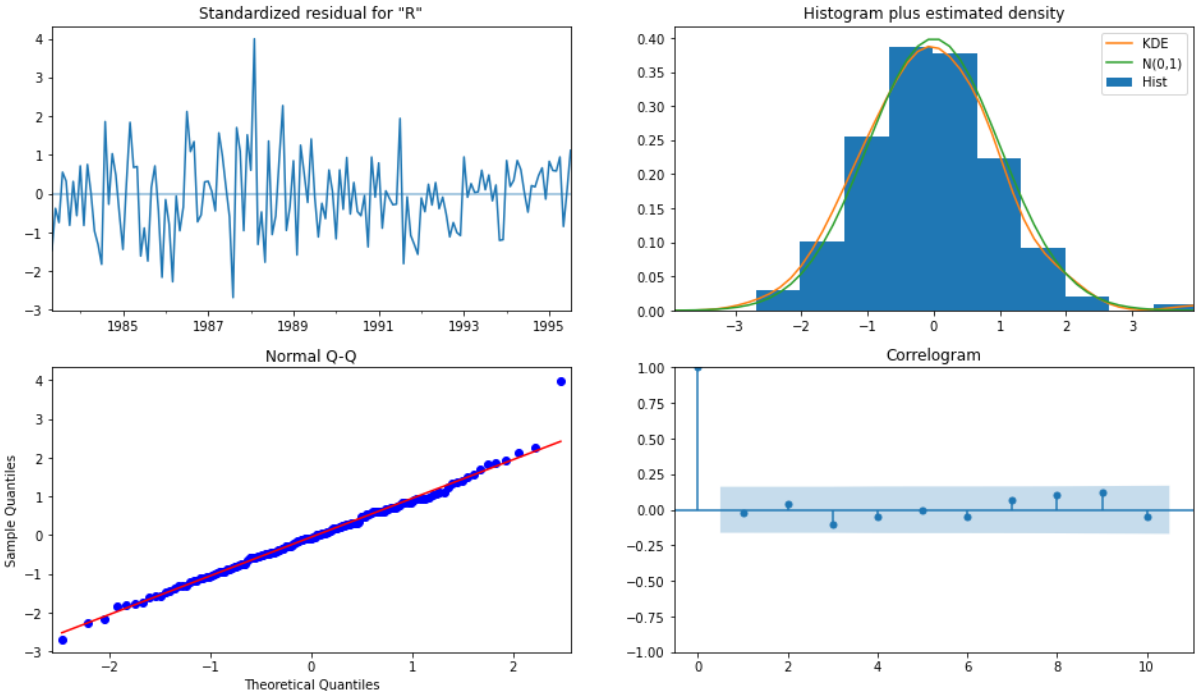


FIGURE 49: MANUAL SARIMA SPARKLING-FORECAST PLOT

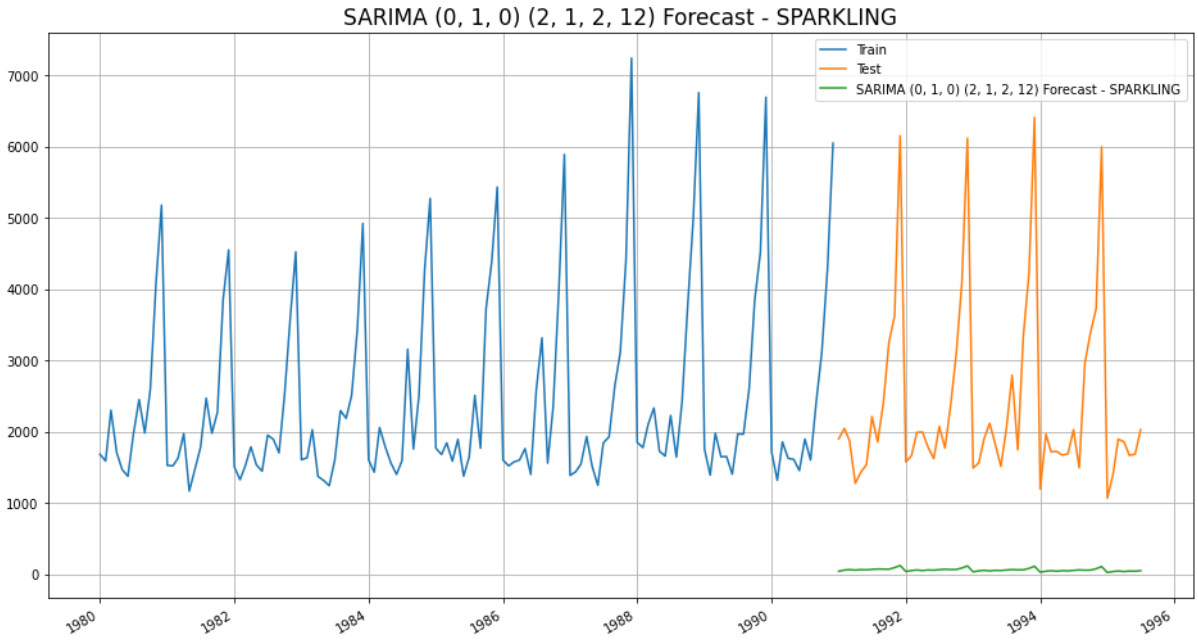


TABLE 37: RMSE AND MAPE SCORE (MANUAL SARIMA) – SPARKLING

	RMSE	MAPE
ARIMA(2,1,2)	1299.980373	47.100017
ARIMA(0,1,0)	3864.279352	201.327650
SARIMA(3,1,1)(3,0,2,12)	601.244396	25.870721
SARIMA(0,1,0)(3,1,2,12)	1189.835783	54.872536
SARIMA(0,1,0)(2,1,2,12)	2652.266482	97.103927

FIGURE 50: MANUAL SARIMA SPARKLING-DIAGNOSTIC PLOT

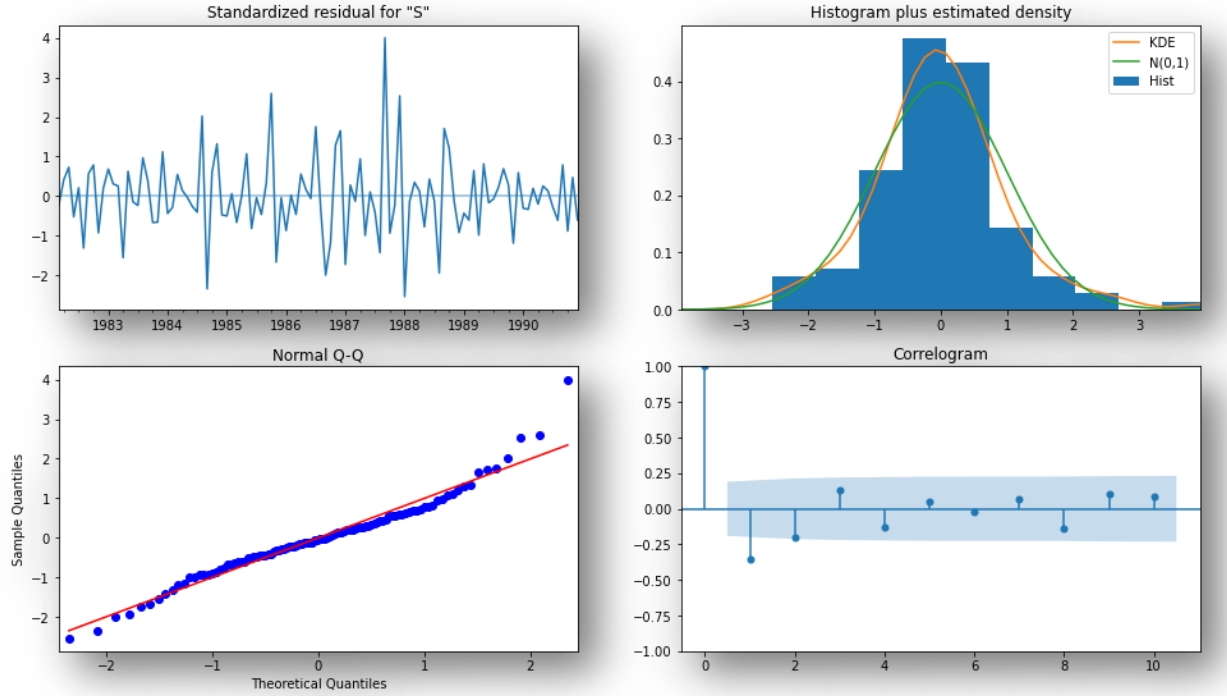


FIGURE 51: MANUAL SARIMA SPARKLING-FORECAST PLOT

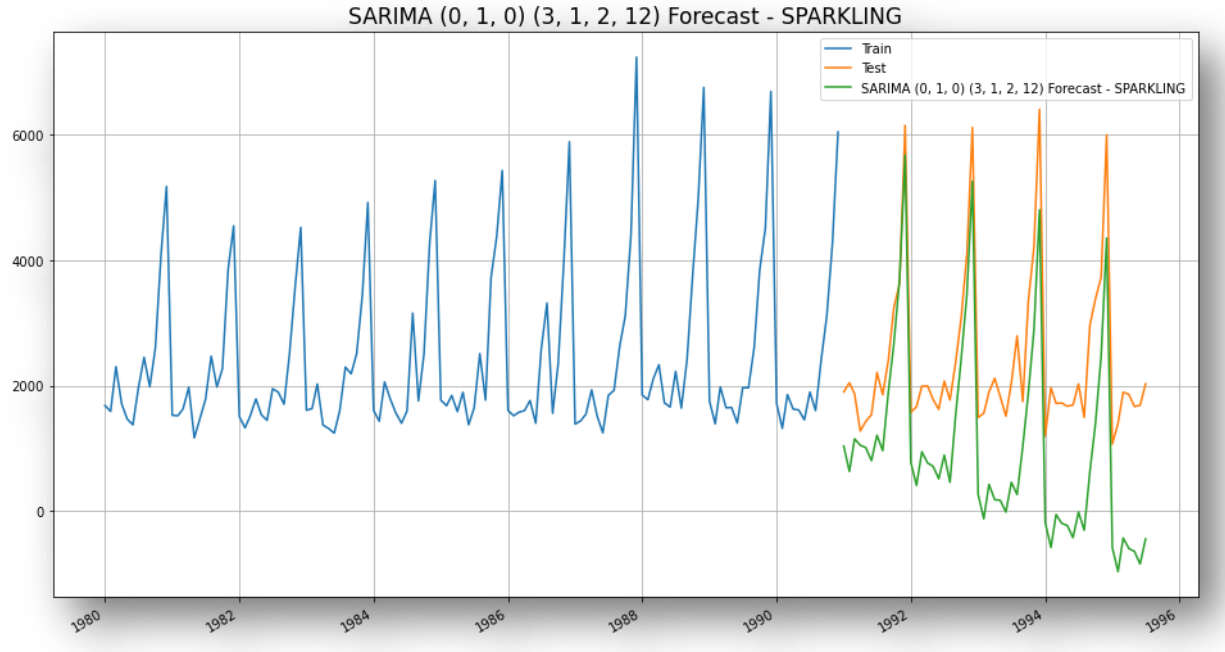


TABLE 38: RMSE AND MAPE SCORE (MANUAL SARIMA) – SPARKLING

RMSE:	1189.8357829699708
MAPE:	54.87253569306072

- 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)**

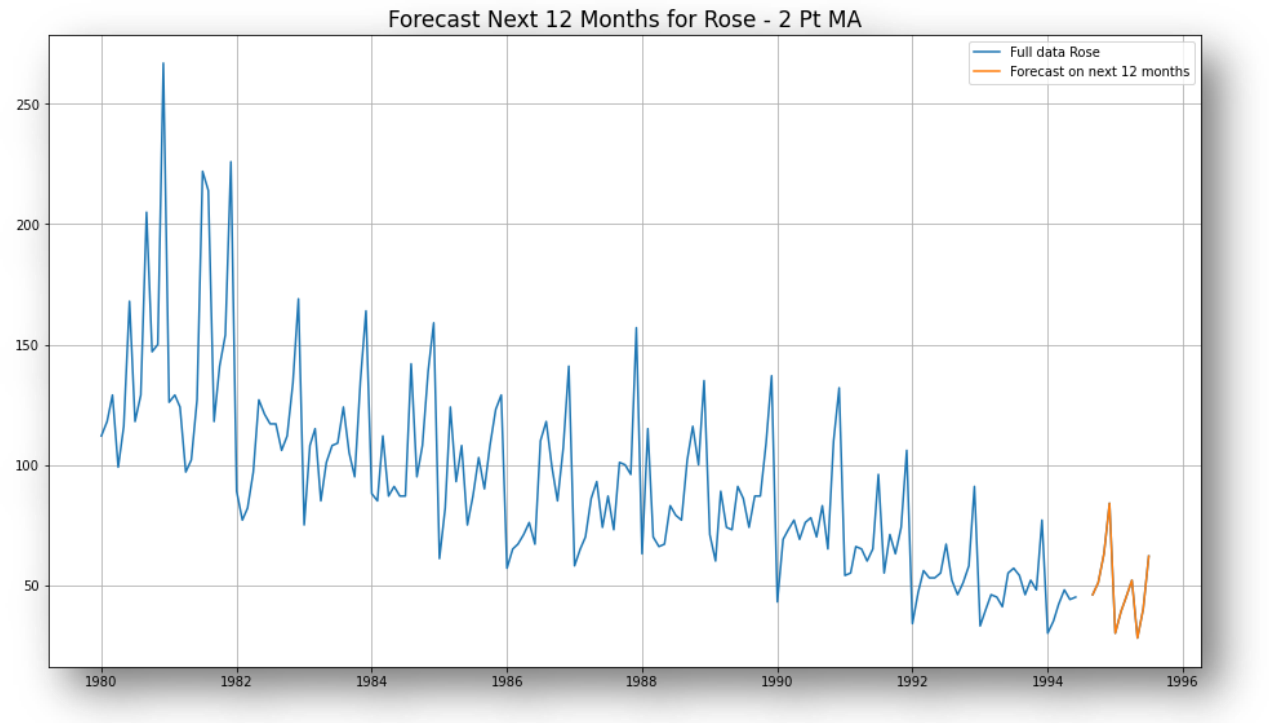
Q.8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

TABLE 39: ROSE DATASET		
MODEL	**PARAMETERS**	**RMSE**
2pointTrailingMovingAverage		11.53
Triple Exponential Smoothing (Additive Season)	α = 0.0849 β = 5.52e−6 ≈ 0.00 γ = 0.00054	14.24
4pointTrailingMovingAverage		14.25
6pointTrailingMovingAverage		14.57
9pointTrailingMovingAverage		14.73
RegressionOnTime		15.27
Double Exponential Smoothing	α = 1.49e−8 ≈ 0.00 β = 5.488e−9 ≈ 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)	α = 0.07736 β = 0.03936 γ = 0.00083	19.11
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)	α = 0.05921 β = 0.0205 γ = 0.00405	25.99
Triple Exponential Smoothing (Additive Season, Damped Trend)	α = 0.07842 β = 0.01153 γ = 0.07738	26.04
Simple Exponential Smoothing	α = 0.09874	36.8
ARIMA (2,1,3)		36.81
ARIMA (2,1,2)		36.87
Simple Average Model		53.46
Naïve Model		79.72

TABLE 40: SPARKLING DATASET		
MODEL	**PARAMETERS**	**RMSE**
Triple Exponential Smoothing (Multiplicative Season, Damped Trend)	α = 0.11107 β = 0.03702 γ = 0.39507	352.45
Triple Exponential Smoothing (Additive Season)	α = 0.1112 β = 0.01236 γ = 0.46071	378.63
Triple Exponential Smoothing (Additive Season, Damped Trend)	α = 0.10062 β = 0.00018 γ = 0.51151	378.63
Triple Exponential Smoothing (Multiplicative Season)	α = 0.11119 β = 0.04943 γ = 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
Simple Average Model		1275.08
6pointTrailingMovingAverage		1283.93
ARIMA (2,1,2)		1299.98
Simple Exponential Smoothing	α = 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
Naïve Model		3864.28
ARIMA (0,1,0)		3864.28
Double Exponential Smoothing	α = 0.665 β = 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.

FIGURE 52: ROSE FORECAST NEXT 12 MONTHS - 2 PT MOVING AVERAGE



- The prediction in the above plot doesn't seems to be predictable, hence opting for the second-best model.

FIGURE 53: ROSE FORECAST NEXT 12 MONTHS - TRIPLE EXPONENTIAL SMOOTHING ETS (A, A, A)

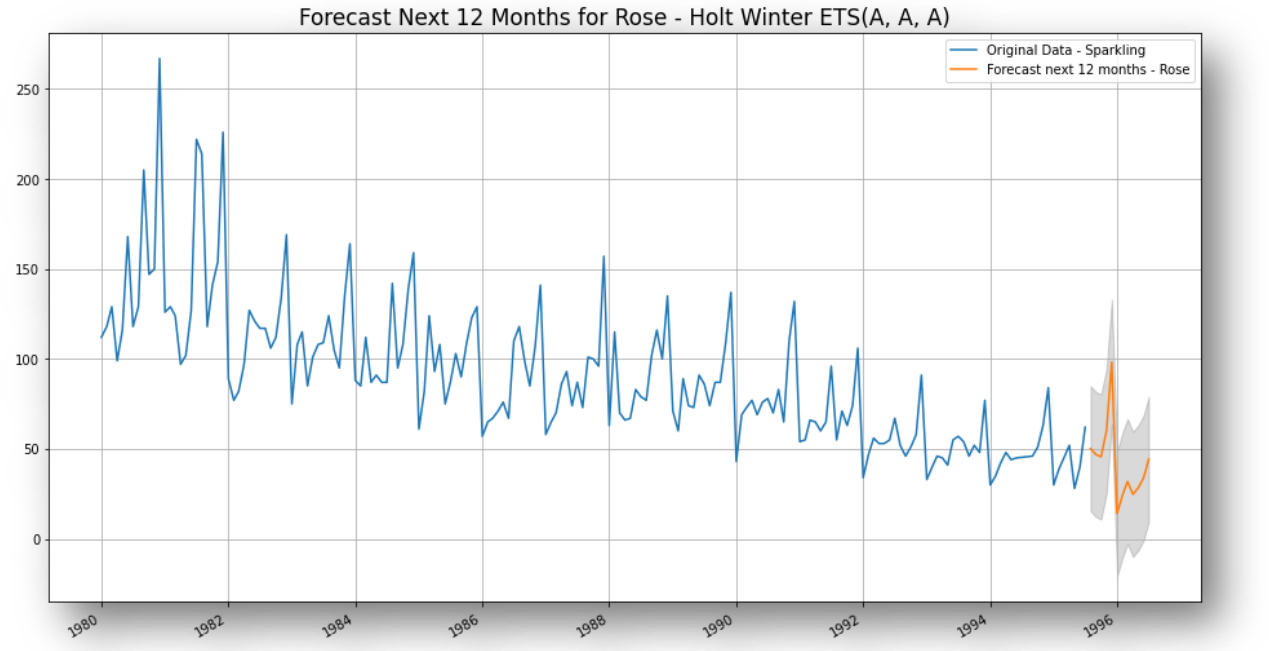
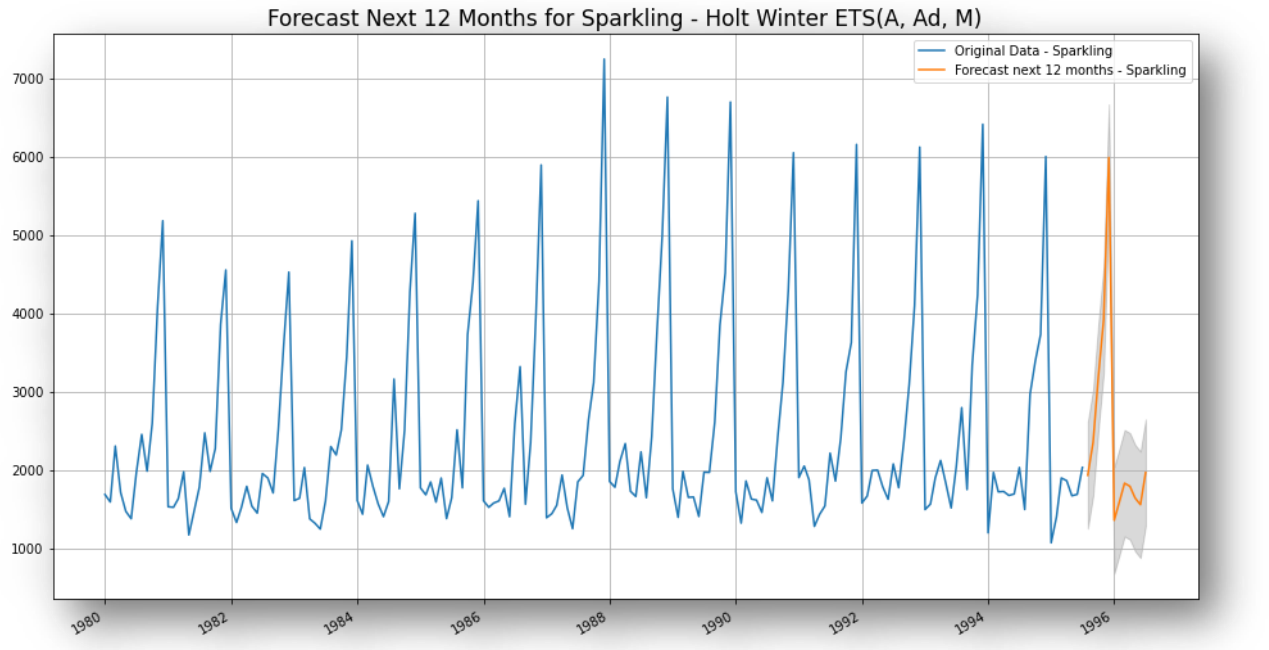


FIGURE 54: SPARKLING FORECAST NEXT 12 MONTHS - TRIPLE EXPONENTIAL SMOOTHING ETS (A, AD, M) - DAMPED TREND, MULTIPLICATIVE SEASONALITY



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 - Forecast Models:**

1. Top 2 best models as per lowest Test RMSE were found to be - 2 Pt Moving Average and Holt-Winters - Additive Seasonality & Trend
2. Holt-Winters seems to give a consistent forecast with respect to the data
3. 2 Pt Moving Average model, when used for forecasting do not seem to give good predictions. Forecast values level out after a few iterations
4. Hence, for final forecast of Rose Wine Sales - we choose Holt-Winters.

- **Rose Wine Sales -Observations:**

1. Rose wine shows a clear trend of declining sales since 1980. This shows decline in popularity of this variant of wine.
2. There is also an instant crashing slump in sales in the first quarter of every year from Jan.
3. 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.

- **Rose Wine Sales - Suggestions:**

1. Company should bring in a new brand along with the existing Rose wine or a different version of it.
2. During the non- season times or when the demand is going increase the company should start marketing by doing ad campaigns etc.
3. The ad campaign can also attract new or first-time drinkers, who will only be attracted by the marketing and no other reason such as taste.
4. During the season time the company should keep there stocks loaded up.
5. In case there is still downward slope in the sales of Rose wines then the company should invest more on the product development.
6. Since there is a decline in the sale of the product over a period of time, more data will be required to get more insights to the issue.

- **Sparkling Wine Sales - Forecast Models:**

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

- **Sparkling Wine Sales – Observation:**

1. Sparkling wine sales don't show any upward or downward trend, it shows flat sales over long term range.
2. There is downward fall from January similar to Rose wines, may be due to the after effects of the holidays.
3. There is increase in the sales during the mid-year July-August.
4. There is very high spike in sales seen in the last quarter of every year from October to December - This might be due to the Holiday season in this period. Highest peak in sales is seen in December every year. December sales are almost 3 times of Sep sales.

- **Sparkling Wine Sales - Suggestions:**

1. Sparkling wine has a great popularity when compared to Rose wines, it can be witnessed in the end of the year sales.
2. Sparkling will also need marketing but not heavy as what Rose wines is required.
3. Since the product is already popular, the best the company can do introduce new bottle sizes, new shaped bottles which can attract the customer with the packaging.
4. The company should load up the stock during the end of the year as that is when there sales are high.
5. Though the monthly sales are good but the yearly sales are not satisfying as they are stagnant, hence more data will be required to analyse .