TIME SERIES

FORECASTING

[JYOTI RANJAN PADHIARY]

Catalog

TIME SERIES FORECASTING.. 1
[JYOTI RANJAN PADHIARY]... 1
WINE SALES DATA... 7
ROSE WINE... 7
1. Read the data as an appropriate Time Series data and plot the data..................7
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition...8
3. Split the data into training and test. The test data should start in 1991............12
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,naive forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.¶......................... 14
Model-1 Linear Regression..14
Model-2 Naive Approach.. 15
Model-3 Simple Average...15
Model-4 Moving Average.. 16
Model -5 Simple Exponential smoothing...18
Model 6: Holt - ETS(A, A, N) - Holt's linear method with additive errors - Rose Wine Sales.. 18
Double Exponential Smoothing - Rose... 18
Model 7: Triple Exponential Smoothing (Holt - Winter's Model).............. 19
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.¶.............20
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............. 22
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......... 28
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.¶... 33
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.. 34
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales............................. 36
SPARKLING WINE SALES DATA...37
1. Read the data as an appropriate Time Series data and plot the data.................37
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition...38
Decomposition...40
3. Split the data into training and test. The test data should start in 1991............41
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,naive forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE........................... 42
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.................47
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............. 49
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......... 59
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..62
10.Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales............................. 63

Table 1 -First 10 observations of Rose Data.. 7
Table 2 -Last 10 observations of Rose Data...7
Table 3 -Plot of Rose time Series data.. 8
Table 4 -Description of Rose Data.. 8
Table 5 -Yearly Box Plot of Rose Data... 8
Table 6 -Monthly Box Plot of Rose Data..9
Table 7 -Month wise Rose wine sales data of every year..................................... 9
Table 8 -Monthly Sales across years for Rose data...10
Table 9 -Residual plot of additive decomposition...11
Table 10 - Multiplicative decomposition of rose data...12
Table 11 -Residual plot of multiplicative decomposition....................................12
Table 12 -First and last rows of train and test rose data...................................... 13
Table 13 -Train and test data plot.. 13
Table 14 -Linear Regression model on Rose Data.. 14
Table 15 -Naive Approach on Rose Data.. 15
Table 16 -Simple average forecast on Rose data...15
Table 17 -Moving average forecast on Rose data... 16
Table 18 -Moving average forecast on Rose data train and test..........................17
Table 19 -RMSE Values of trailing moving average 17
Table 20 -All model graph...17
Table 21 -Forecasting using SES model on rose data... 18
Table 22 -Simple and Double exponential smoothing predictions......................18
Table 23 -Tripple Exponential smoothing on Rose data (Additive model)..........19
Table 24 -Tripple Exponential smoothing (Multiplicative model).......................19
Table 25 -Best model 2-pt moving average 20
Table 26 -Stationary data plot... 21
Table 27 - RoseTrain data stationary check 22
Table 28 -Stationary rose train data...22
Table 29 -AIC Value for different values.. 23
Table 30 -AIC Values in ascending order..24
Table 31 -SARIMAX result (2,1,3)...24
Table 32 -Rose data diagnostic plot.. 25
Table 33 -ARIMA Plot on train and test rose data.. 25
Table 34 -ACF Plot of rose train data..26
Table 35 -AIC Values for different values and seasonality................................. 26
Table 36 -SARIMAX model on Rose data..27
Table 37 -SRIMA Model(2,1,3) (2,0,3,6)...28
Table 38 -Rose train ACF graph..28
Table 39 -ARIMA MODEL mannual..29
Table 40 -Mannual ARIMA diagnostic plot..29
Table 41 -ARIMA (2,1,2)..29
Table 42 -Mannual SARIMA result summary.. 31
Table 43 -Mannual SARIMA diagnostic plot..31
Table 44 -Mannual SARIMA (2,1,2)(3,1,2,12) result summary......................... 32
Table 45 -Mannual SARIMA (2,1,2) (3,1,2,12) diagnostic plot......................... 32
Table 46 -All model RMSE & MAPE score... 33
Table 47 -Best model graph for Rose data.. 34
Table 48 -Forecast on best model..34
Table 49 -second best model forecast on Rose data..35

Table 50 -head of sparkling data.. 37
Table 51 -tail of sparkling data...38
Table 52 -Time series plot of sparkling data.. 38
Table 53 -Sparkling data description...38
Table 54 -Year on year box plot of sparkling data... 39
Table 55 -Monthly box plot of sparkling data..39
Table 56 -Yearly sales across the month of sparkling data.................................40
Table 57 -Sparkling data multiplicative decomposition.....................................41
Table 58 -Sparkling train and test data plot...41
Table 59 -Linear regression on sparkling data.. 42
Table 60 -Naive approach forecast on sparkling data... 43
Table 61 -Simple average forecast on sparkling data..43
Table 62 -Moving average forecast on whole sparkling data.............................44
Table 63 -Moving average forecast on train and test data..................................44
Table 64 -RMSE Value of trailing moving average 45
Table 65 -All model graph...45
Table 66 -SES model forecast on sparkling data 45
Table 67 -Double Exponential smoothing on sparkling data.............................. 46
Table 68 -TES additive model on sparkling data.. 46
Table 69 -TES multiplicative forecast on sparkling data.................................... 47
Table 70 -Stationary sparkling data..48
Table 71 -Stationary train data of sparkling.. 49
Table 72 -AIC score table for auto arima... 51
Table 73 -Auto ARIMA result summary.. 51
Table 74 -Auto ARIMA diagnostic plot... 52
Table 75 -RMSE value of auto ARIMA...52
Table 76 -Auto ARIMA forecast on sparkling data..52
Table 77 -ACF plot sparkling train data..53
Table 78 -PACF plot sparkling train data.. 53
Table 79 -Sorted table of AIC scores auto SARIMA.. 54
Table 80 -Auto SARIMA result summary..54
Table 81 -Auto SARIMA diagonostic plot..55
Table 82 -RMSE and MAPE value for auto SARIMA on sparkling data...........55
Table 83 -Auto SARIMA(2,1,3)(2,0,3,6)...55
Table 84 -Sparkling train data ACF plot.. 56
Table 85 -Sparkling train data PACF plot... 56
Table 86 -Mannual SARIMA result summary.. 57
Table 87 -mannual SARIMA diagnostic plot..57
Table 88 -Mannual SARIMA 1st model forecast on sparkling data................... 58
Table 89 -Mannual SARIMA 2nd model forecast on sparkling data..................58
Table 90 -RMSE score of mannual SARIMA models on sparkling data............58
Table 91 -Sparkling train data ACF plot.. 59
Table 92 -Sparkling train data PACF plot... 59
Table 93 -Mannual ARIMA result summary sparkling data............................... 60
Table 94 -Mannual ARIMA sparkling data diagnostic plot................................ 61
Table 95 -mannual ARIMA forecast sparkling data 61
Table 96 -Sparkling data All model RMSE score... 62
Table 97 -Sparkling data forecast on best model.. 63

WINE SALES DATA

ROSE WINE

Problem statement : For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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.

Rose wine sales data

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

There are two columns in the data , one is “Rose” another is “Year Month”. We have 187 observations starting from 1980-01-01 to 1995-07-01, of total 15years.

There are two missing values in the data-set which we have imputed using the interpolate function.

Below are head and tail of the data-set

Table 1-First 10 observations of Rose Data

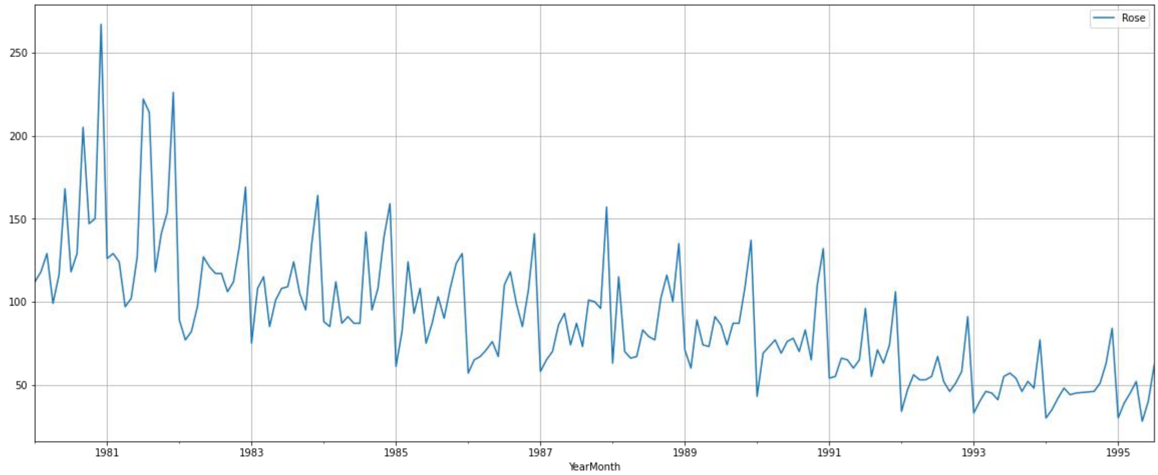
	Rose
YearMonth	
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Table 2-Last 10 observations of Rose Data

	Rose
YearMonth	
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

Below we have plot the entire time series data

Table 3-Plot of Rose time Series data

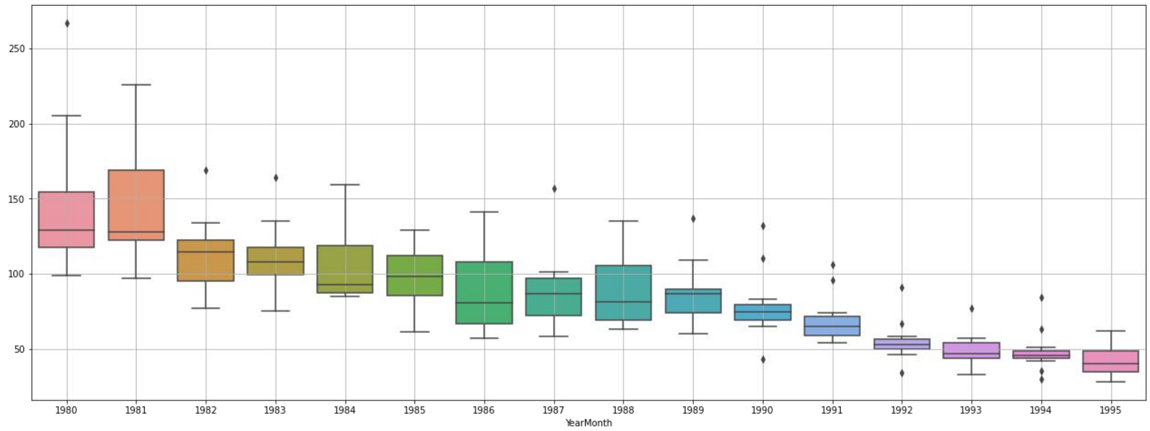


2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Table 4-Description of Rose Data

	Rose
count	187.000000
mean	89.914439
std	39.238325
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Table 5-Yearly Box Plot of Rose Data



we have plot the yearly box plot above and we can see there are some Outliers present.

Trend is also there, it is decreasing.

Table 6-Monthly Box Plot of Rose Data

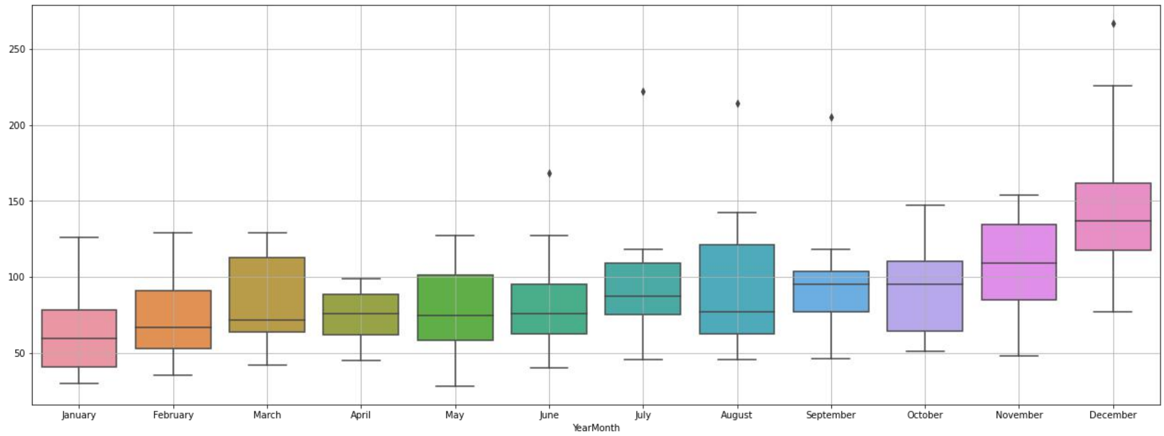
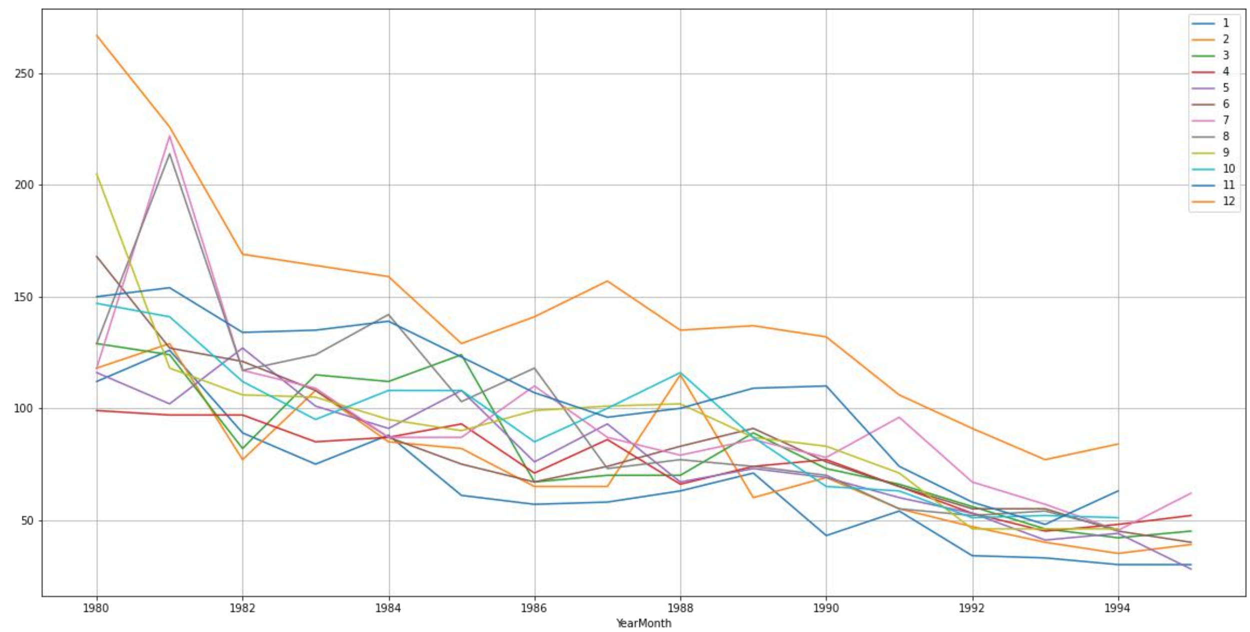


Table 7-Month wise Rose wine sales data of every year

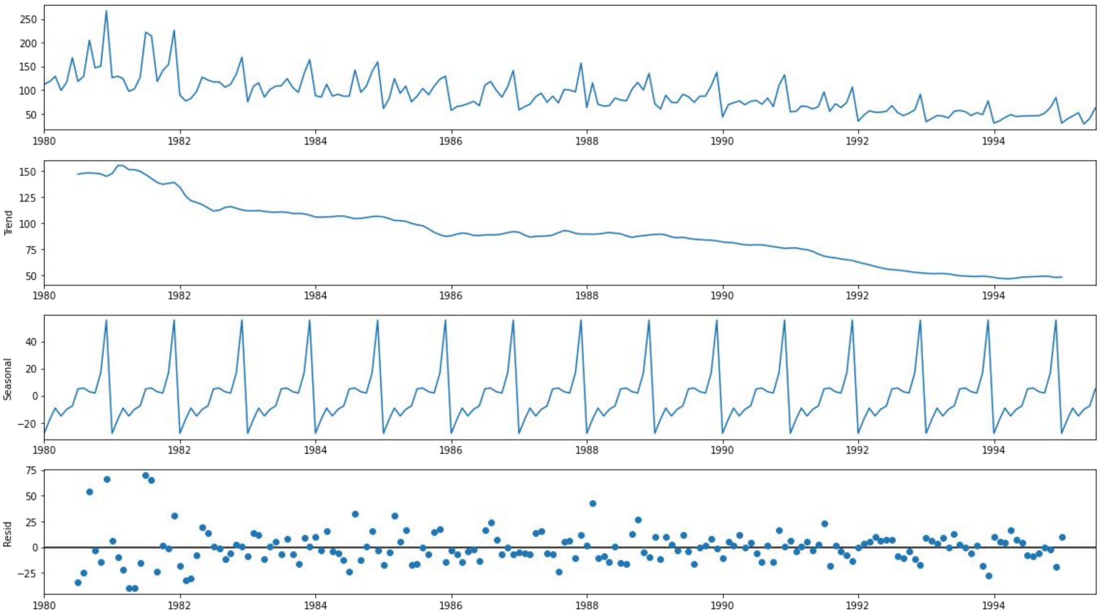
YearMonth	April	August	December	February	January	July	June	March	May	November	October	September
YearMonth												
1980	99.0	129.000000	267.0	118.0	112.0	118.000000	168.0	129.0	116.0	150.0	147.0	205.0
1981	97.0	214.000000	226.0	129.0	126.0	222.000000	127.0	124.0	102.0	154.0	141.0	118.0
1982	97.0	117.000000	169.0	77.0	89.0	117.000000	121.0	82.0	127.0	134.0	112.0	106.0
1983	85.0	124.000000	164.0	108.0	75.0	109.000000	108.0	115.0	101.0	135.0	95.0	105.0
1984	87.0	142.000000	159.0	85.0	88.0	87.000000	87.0	112.0	91.0	139.0	108.0	95.0
1985	93.0	103.000000	129.0	82.0	61.0	87.000000	75.0	124.0	108.0	123.0	108.0	90.0
1986	71.0	118.000000	141.0	65.0	57.0	110.000000	67.0	67.0	76.0	107.0	85.0	99.0
1987	86.0	73.000000	157.0	65.0	58.0	87.000000	74.0	70.0	93.0	96.0	100.0	101.0
1988	66.0	77.000000	135.0	115.0	63.0	79.000000	83.0	70.0	67.0	100.0	116.0	102.0
1989	74.0	74.000000	137.0	60.0	71.0	86.000000	91.0	89.0	73.0	109.0	87.0	87.0
1990	77.0	70.000000	132.0	69.0	43.0	78.000000	76.0	73.0	69.0	110.0	65.0	83.0
1991	65.0	55.000000	106.0	55.0	54.0	96.000000	65.0	66.0	60.0	74.0	63.0	71.0
1992	53.0	52.000000	91.0	47.0	34.0	67.000000	55.0	56.0	53.0	58.0	51.0	46.0
1993	45.0	54.000000	77.0	40.0	33.0	57.000000	55.0	46.0	41.0	48.0	52.0	46.0
1994	48.0	45.666667	84.0	35.0	30.0	45.333333	45.0	42.0	44.0	63.0	51.0	46.0
1995	52.0	NaN	NaN	39.0	30.0	62.000000	40.0	45.0	28.0	NaN	NaN	NaN

Table 8-Monthly Sales across years for Rose data



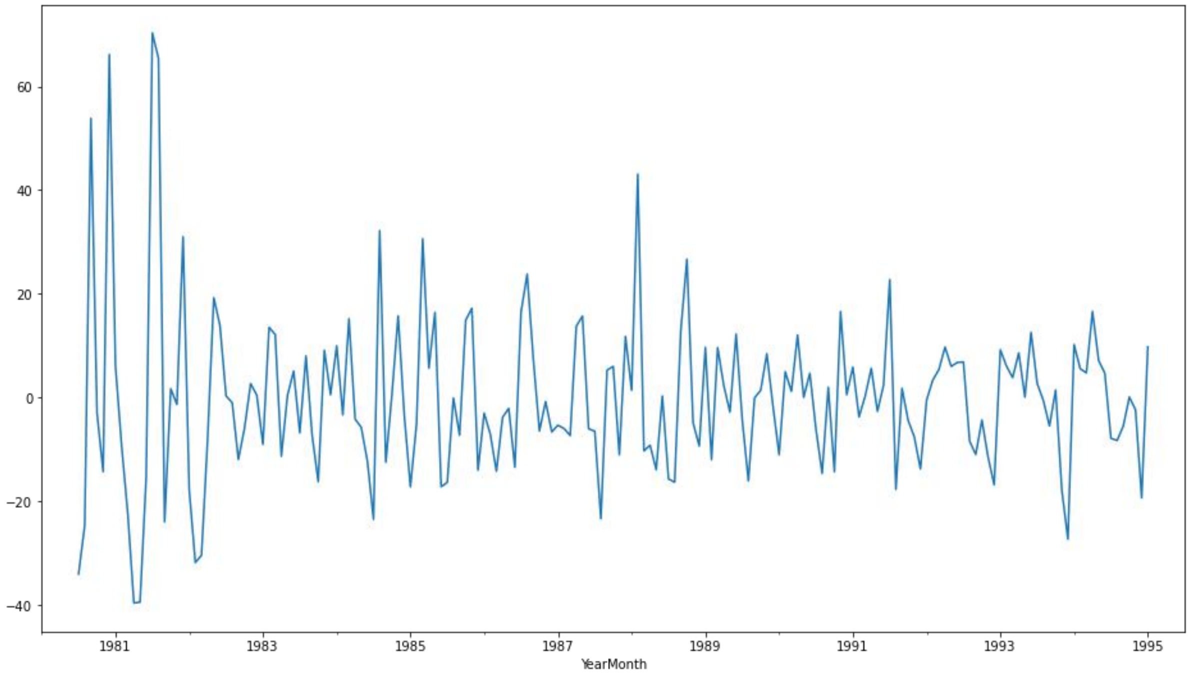
Decomposition

Additive



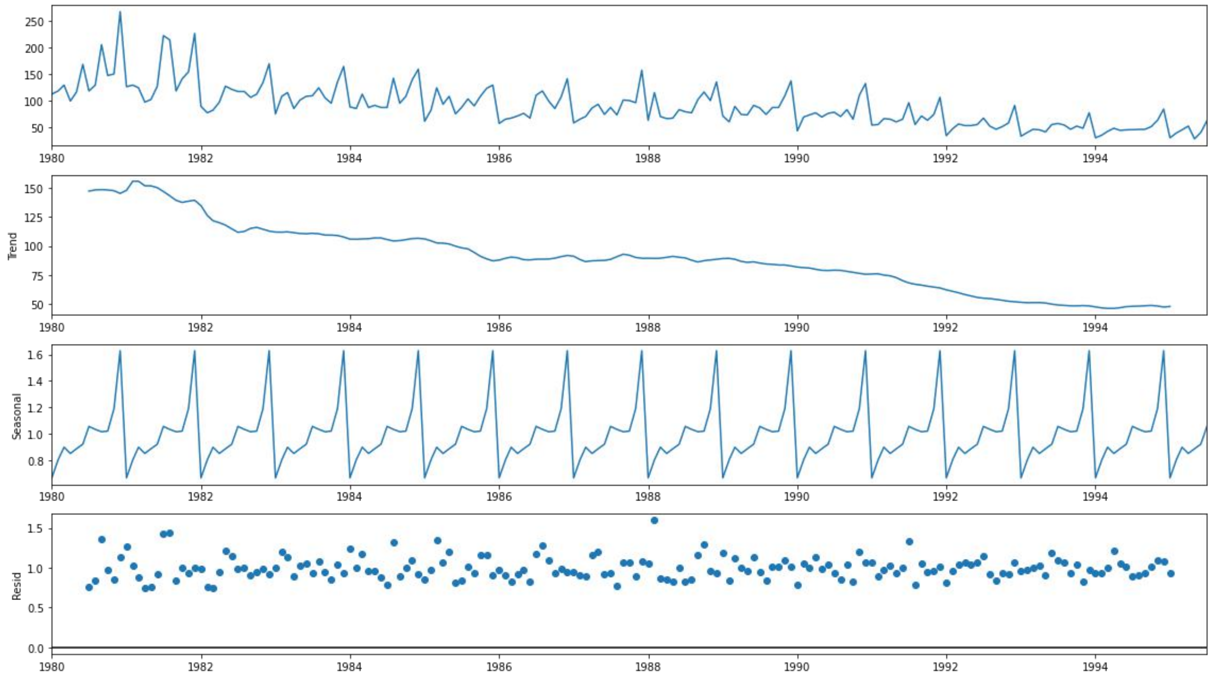
Above additive decomposition tells us that there is a trend and seasonality present in the data.

Table 9-Residual plot of additive decomposition



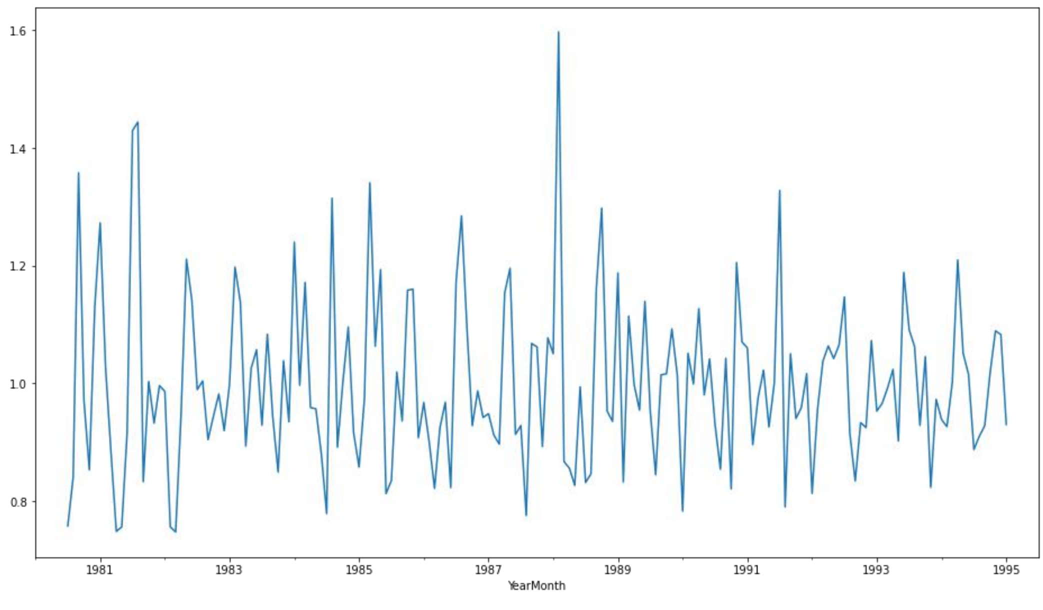
Multiplicative decomposition

Table 10- Multiplicative decomposition of rose data



Above multiplicative decomposition tells us that there is a trend and seasonality in the data.

Table 11-Residual plot of multiplicative decomposition



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

We have split the data set into train and test , the train data is having 132 observations and test data having 55 observations.

Table 12-First and last rows of train and test rose data

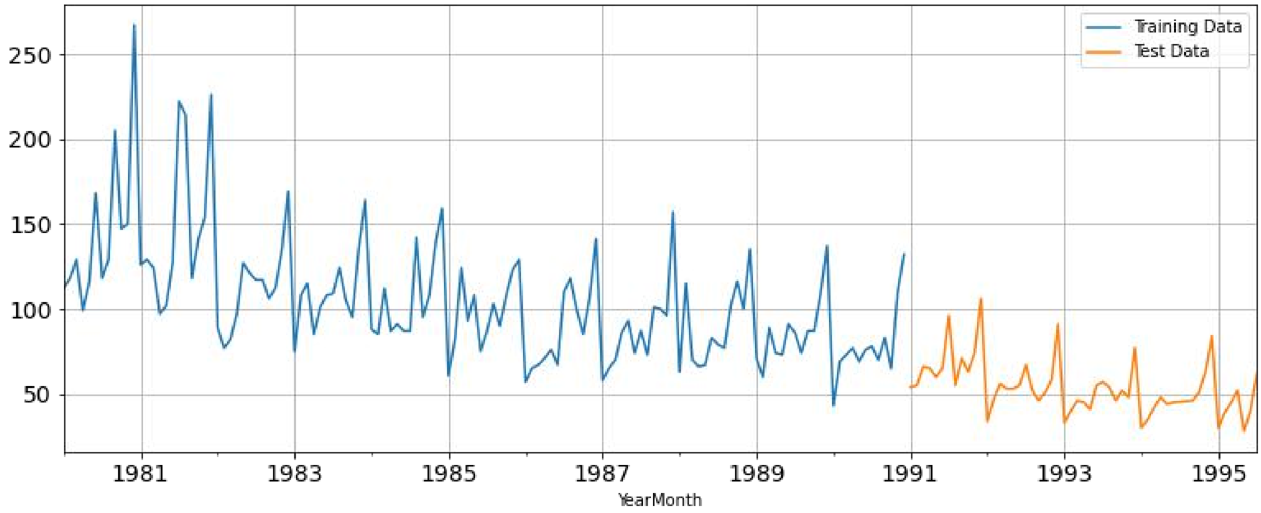
```
First few rows of Training Data
              Rose
YearMonth
1980-01-01   112.0
1980-02-01   118.0
1980-03-01   129.0
1980-04-01    99.0
1980-05-01   116.0

Last few rows of Training Data
              Rose
YearMonth
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

First few rows of Test Data
             Rose
YearMonth
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 Test Data
             Rose
YearMonth
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
```

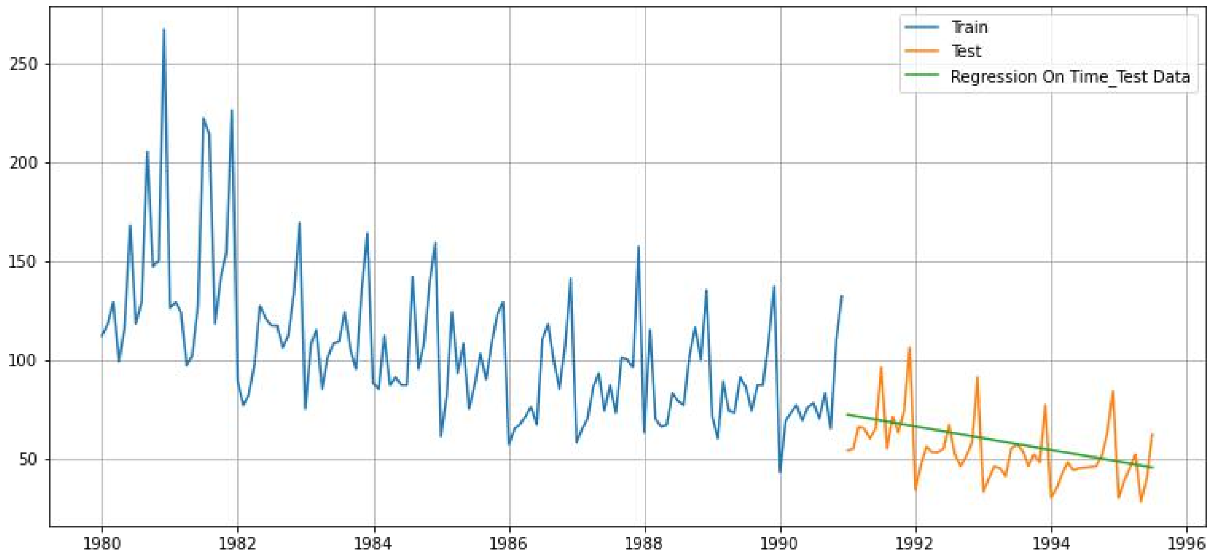
Table 13-Train and test data plot



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, naive forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.¶

Model-1 Linear Regression

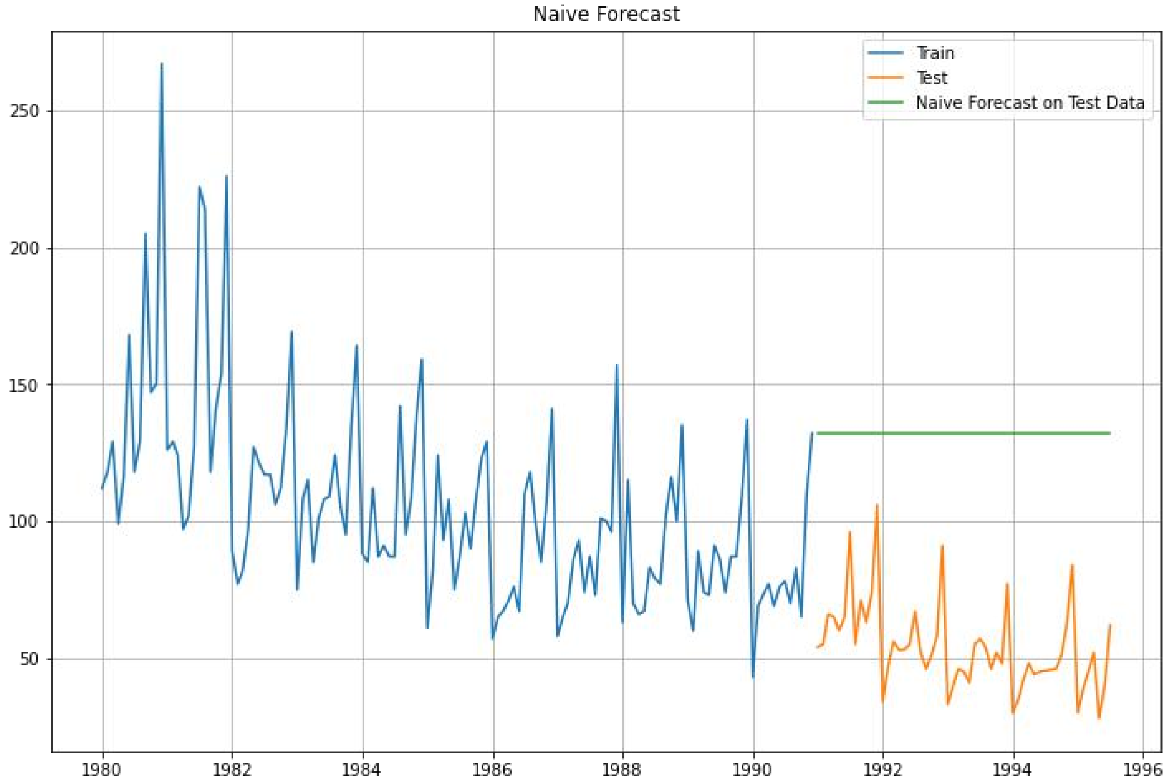
Table 14-Linear Regression model on Rose Data



The RMSE value here is 15.26

Model-2 Naive Approach

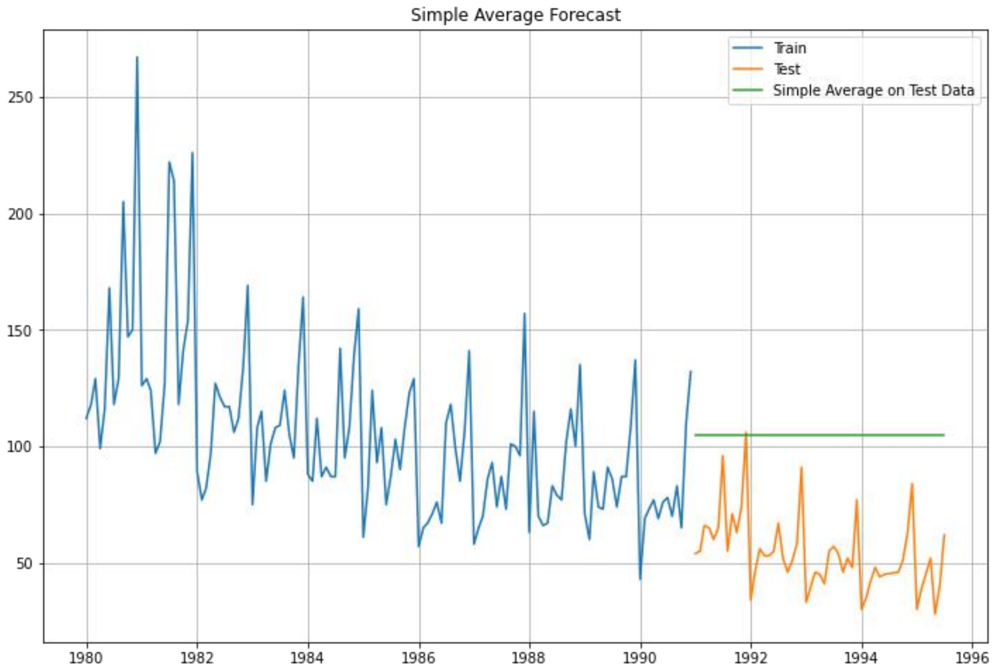
Table 15-Naive Approach on Rose Data



The RMSE Value for Naive approach is 79.71

Model-3 Simple Average

Table 16-Simple average forecast on Rose data



The RMSE value here is 53.46

Model-4 Moving Average

Table 17-Moving average forecast on Rose data

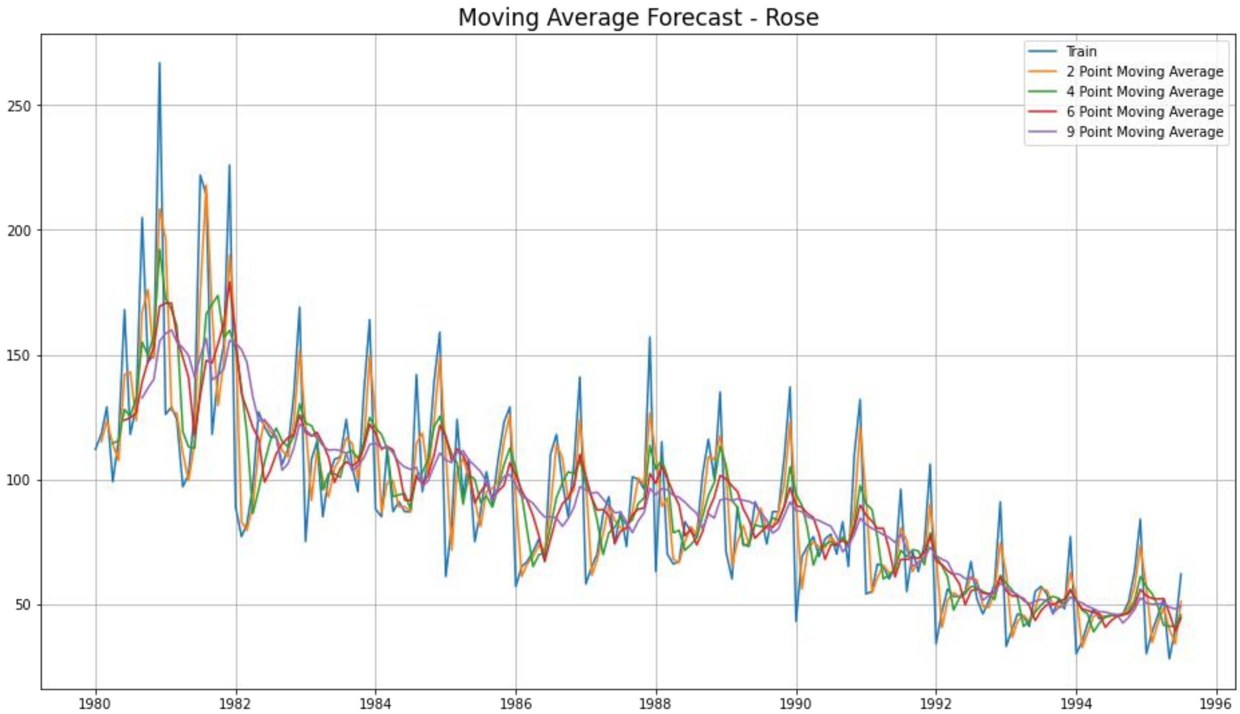
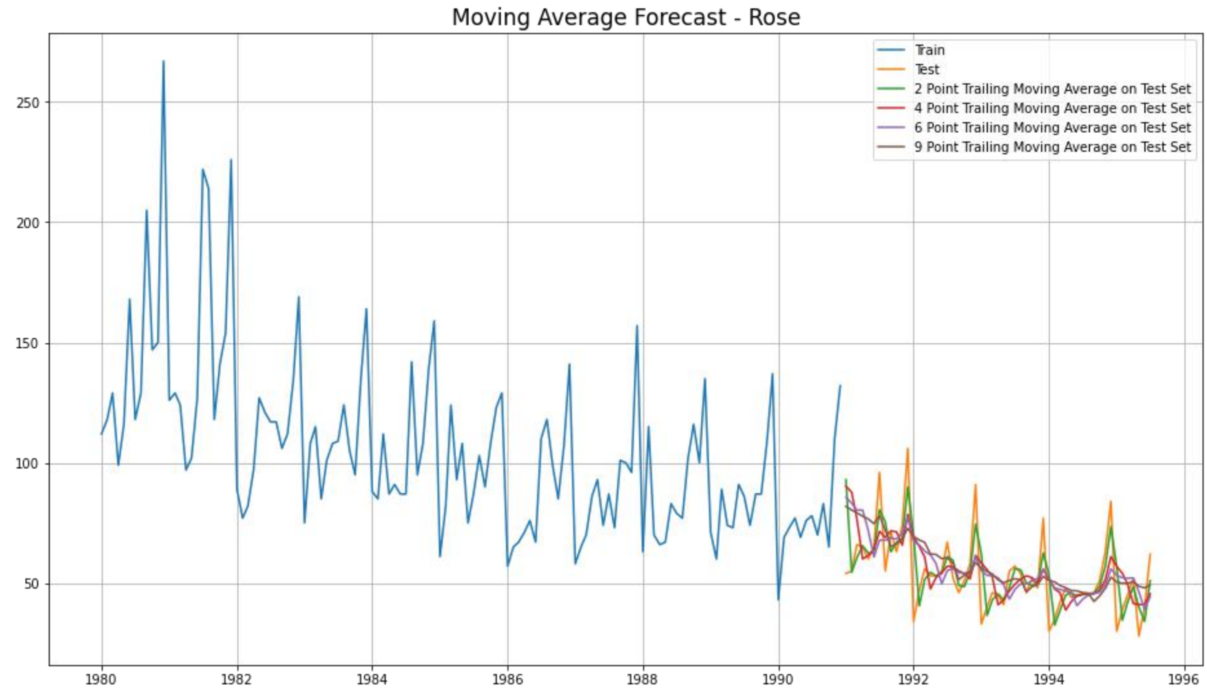


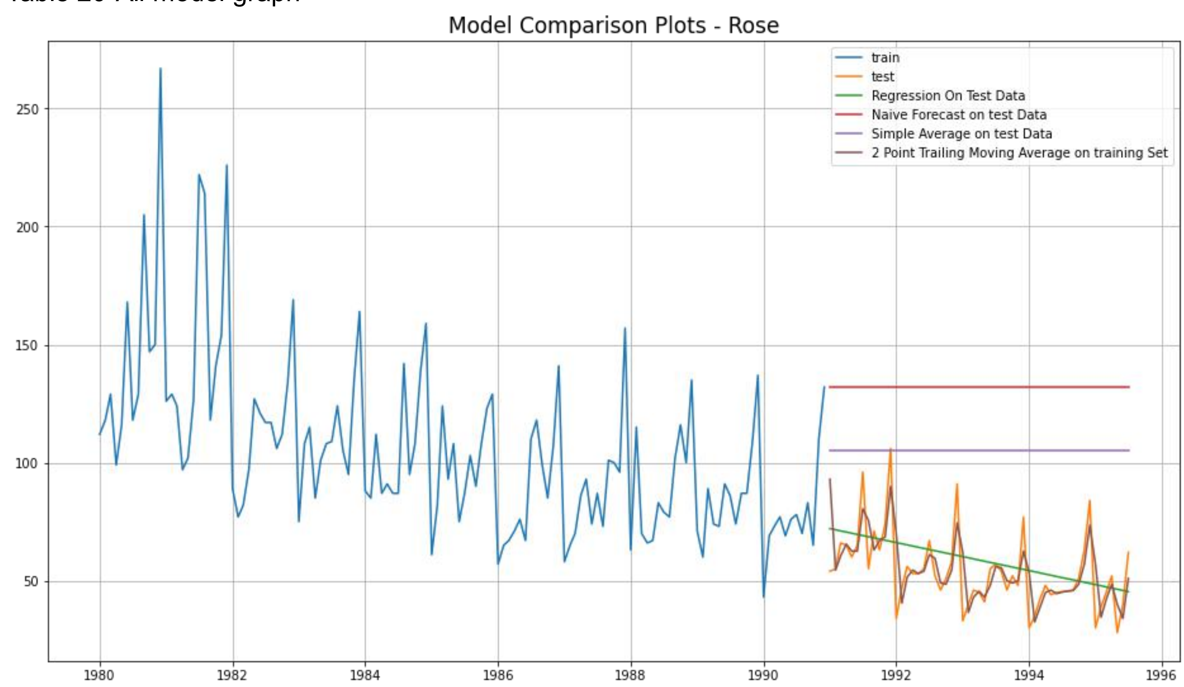
Table 18-Moving average forecast on Rose data train and test




```
For 2 point Moving Average Model forecast on Rose Training Data,  RMSE is 11.529
For 4 point Moving Average Model forecast on Rose Training Data,  RMSE is 14.451
For 6 point Moving Average Model forecast on Rose Training Data,  RMSE is 14.566
For 9 point Moving Average Model forecast on Rose Training Data,  RMSE is 14.728
```

Table 19-RMSE Values of trailing moving average

Table 20-All model graph



Model -5 Simple Exponential smoothing

Table 21-Forecasting using SES model on rose data

```
1991-01-01    87.104999
1991-02-01    87.104999
1991-03-01    87.104999
1991-04-01    87.104999
1991-05-01    87.104999
Freq: MS, dtype: float64
```

```
For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is
 36.796
```

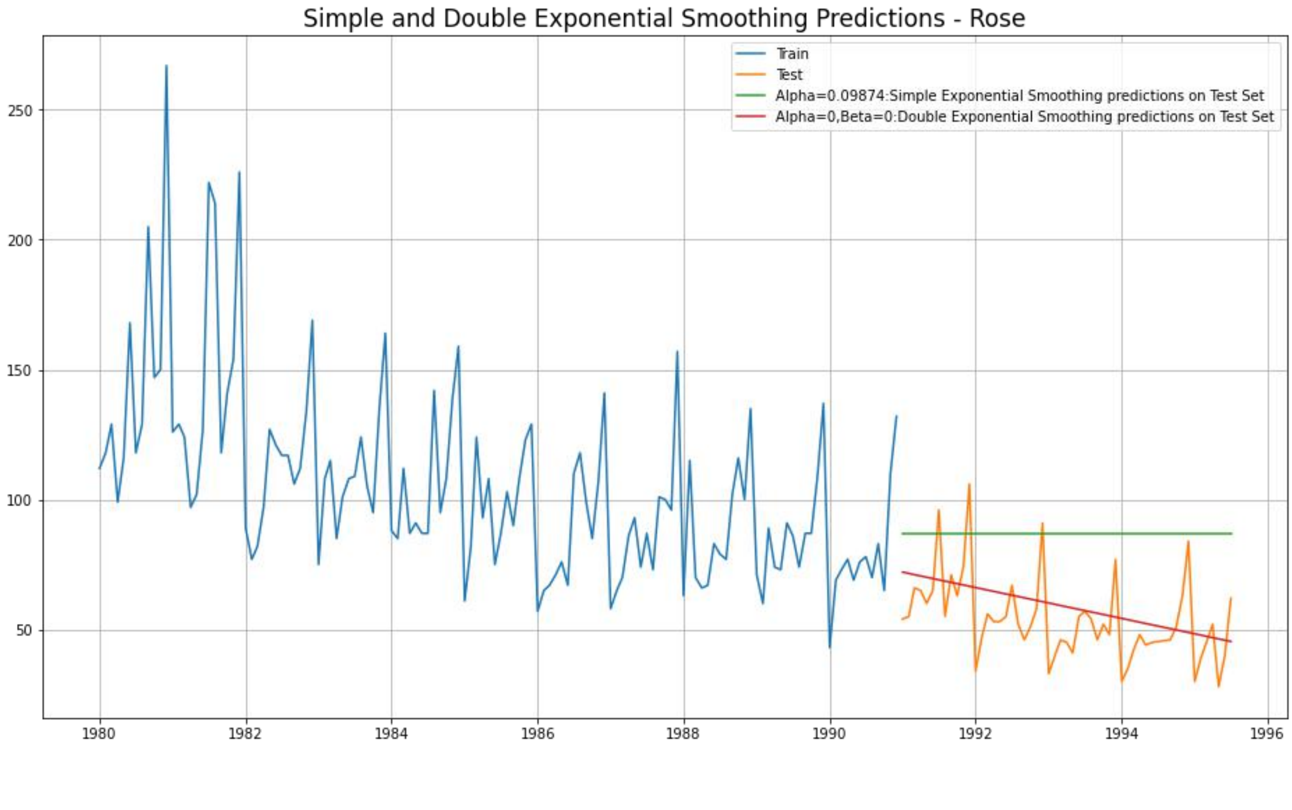
Model 6: Holt - ETS(A, A, N) - Holt's linear method with additive errors - Rose Wine Sales.

Double Exponential Smoothing - Rose

```
Holt model Exponential Smoothing Estimated Parameters :

{'smoothing_level': 1.4901247095597348e-08, 'smoothing_trend': 7.3896641488640725e-09,
 'smoothing_seasonal': 3e21
-+,    'damping_trend': nan, 'initial_level': 137.81551313502814, 'initial_trend': -0.
4943777717865305, 'initial_seasons': array([], dtype=float64), 'use_boxcox': False, 'l
amda': None, 'remove_bias': False}
```

Table 22-Simple and Double exponential smoothing predictions



We see that the double exponential smoothing is picking up the trend component along with the level component as well.

The RMSE Value is 15.26

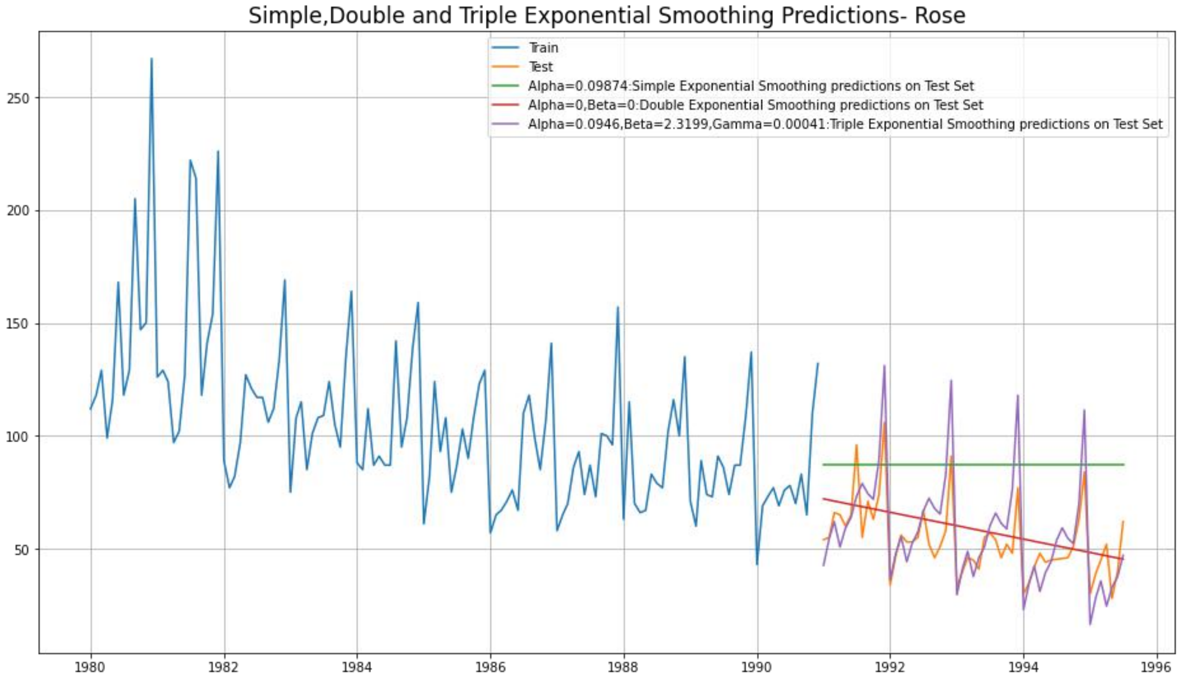
Model 7: Triple Exponential Smoothing (Holt - Winter's Model).

Additive model

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

{'smoothing_level': 0.09467987567540882, 'smoothing_trend': 2.31999683285252e-05, 'smo
othing_seasonal': 0.0004175285691922314, 'damping_trend': nan, 'initial_level': 146.40
142527639352, 'initial_trend': -0.5464913833622084, 'initial_seasons': array([-31.1926
8548, -18.83344765, -10.84745053, -21.48718886,
       -12.67654312,  -7.19154248,   2.65454402,   8.80233514,
         4.79913097,   2.91389547,  21.00157004,  63.18716583]), 'use_boxcox': False,
'lamda': None, 'remove_bias': False}
```

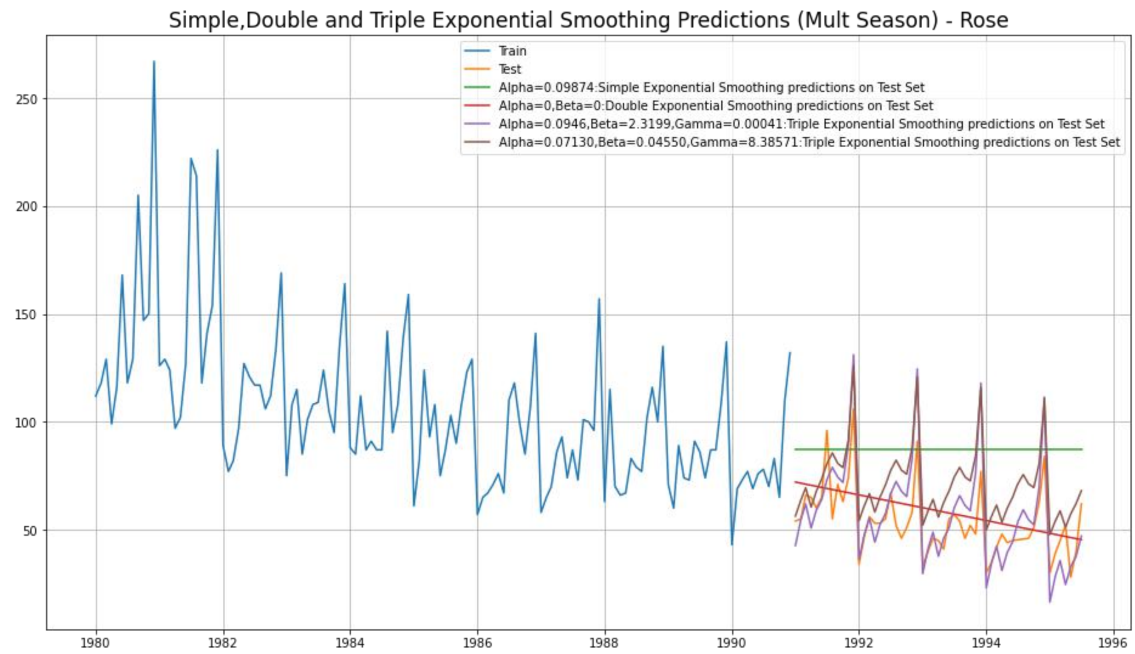
Table 23-Tripple Exponential smoothing on Rose data (Additive model)



Here the RMSE Value is found to be 14.27

Multiplictive model

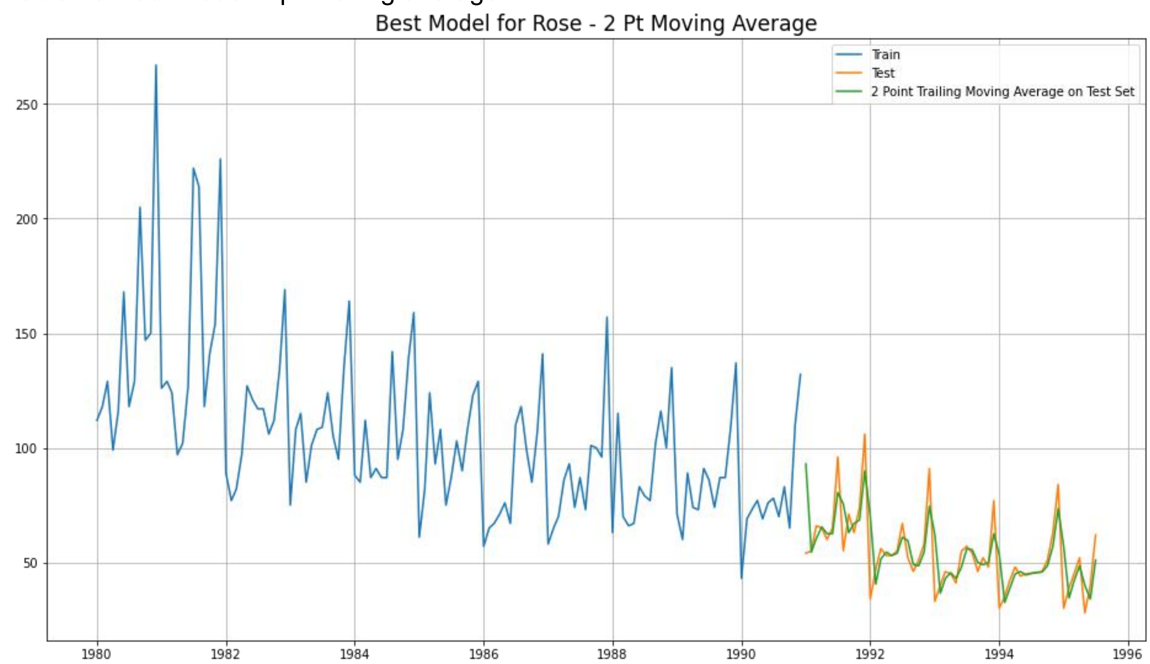
Table 24-Tripple Exponential smoothing (Multiplicative model)



Here the RMSE Value is 20.18

As of now we can see that 2 point trailing moving average model performs the best.

Table 25-Best model 2-pt moving average



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.¶

Checking Stationarity of ROSE Wine sales Data. The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary. The hypothesis in a simple form for the ADF test is:

- $H0$: The Time Series has a unit root and is thus non-stationary.
- $H1$: The Time Series does not have a unit root and is thus stationary. We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value. Hypothesis for Statistical Test for Rose wine sales.
- H0: The Time series of Rose Wine has a unit root and is thus a Non- stationary.
- H1: The Time series of Rose Wine does not have a unit root and is thus Stationary.

We have performed ADF test to check stationarity and below is the result

```
DF test statistic is -2.240
DF test p-value is 0.467137162779315
Number of lags used 13
```

So the data-set is non stationary as p>0.05

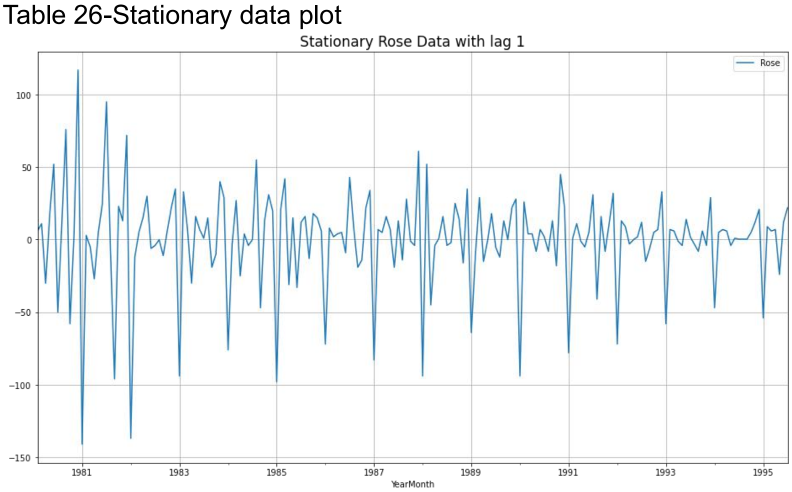
We have to differentiate the data to make it stationary

After differentiating the p<0.05

```
DF test statistic is -8.162
DF test p-value is 3.01597611582779e-11
Number of lags used 12
```

So now it has become stationary , lets see this by plotting it

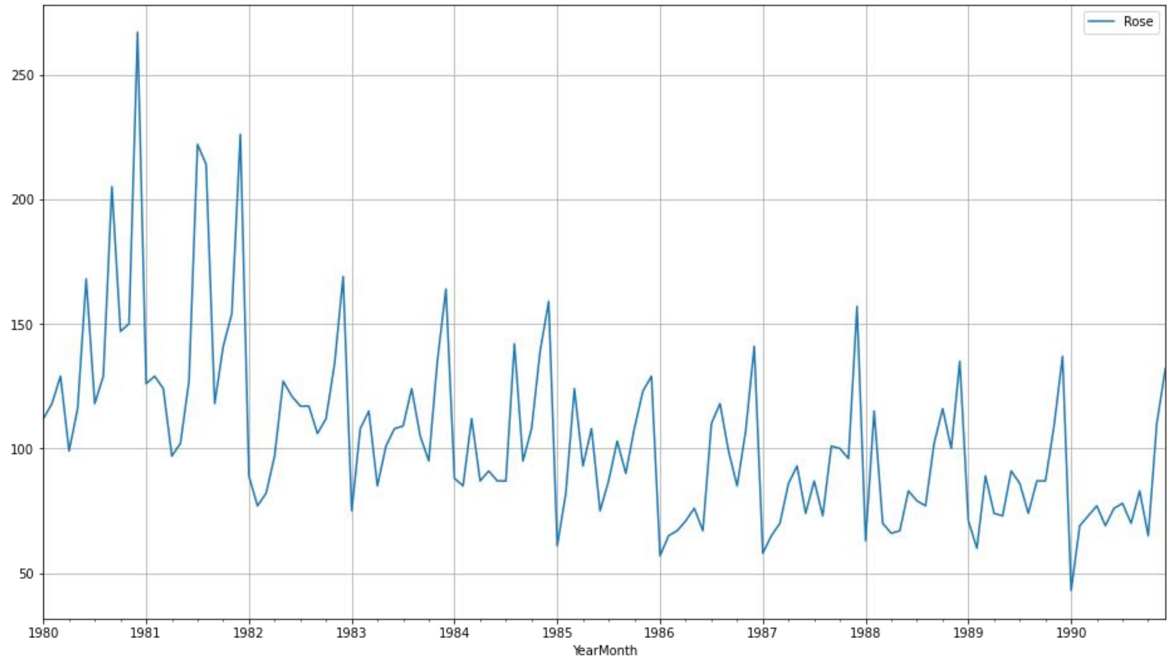
Table 26-Stationary data plot



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.

Checking stationarity condition for ARIMA/SARIMA model in train data.

Table 27- RoseTrain data stationary check



Looking at the above plot it looks like the train data is non stationary

We have to differentiate it to make it stationary as from ADF test the p>0.05

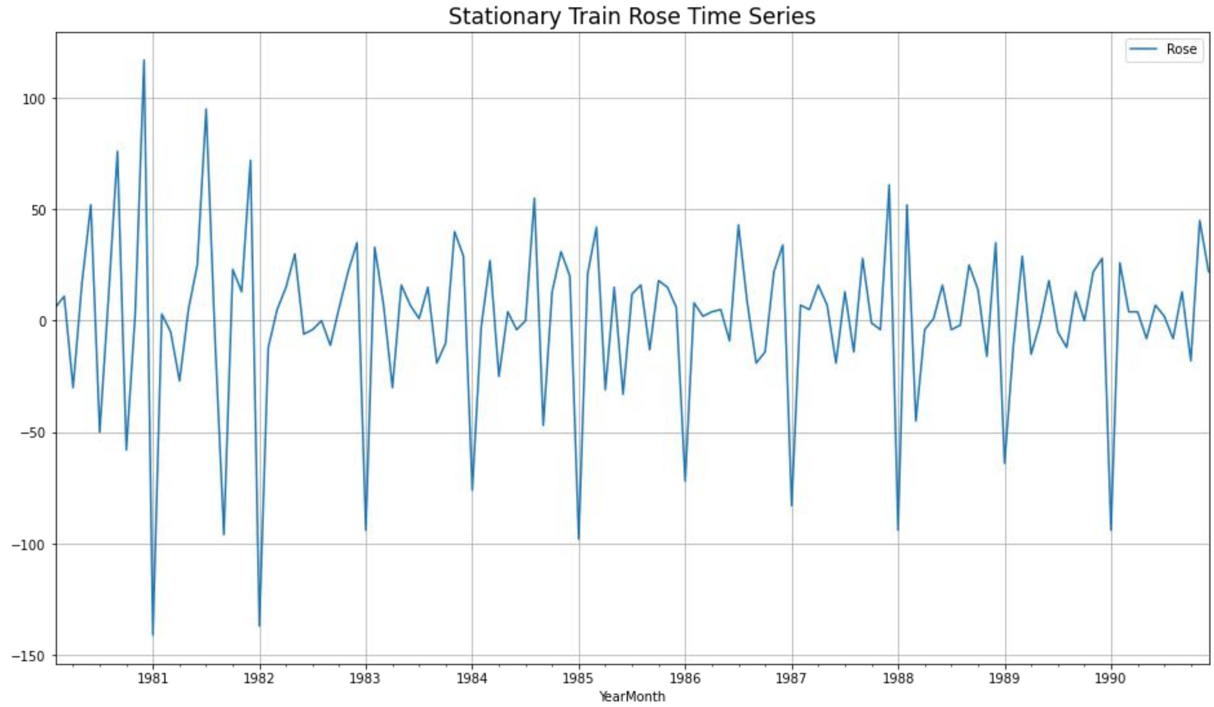
```
DF test statistic is -1.686
DF test p-value is 0.7569093051047072
Number of lags used 13
```

After differentiating we get the below value of p<0.05

```
DF test statistic is -6.804
DF test p-value is 3.894831356782219e-08
Number of lags used 12
```

Now the train data is stationary.

Table 28-Stationary rose train data



Building model

ARIMA

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
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)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

Table 29-AIC Value for different values

```
ARIMA(0, 1, 0) - AIC:1333.1546729124348
ARIMA(0, 1, 1) - AIC:1282.3098319748299
ARIMA(0, 1, 2) - AIC:1279.6715288535793
ARIMA(0, 1, 3) - AIC:1280.5453761734657
ARIMA(1, 1, 0) - AIC:1317.3503105381546
ARIMA(1, 1, 1) - AIC:1280.5742295380073
ARIMA(1, 1, 2) - AIC:1279.8707234231906
ARIMA(1, 1, 3) - AIC:1281.8707223309966
ARIMA(2, 1, 0) - AIC:1298.6110341604983
ARIMA(2, 1, 1) - AIC:1281.507862186851
ARIMA(2, 1, 2) - AIC:1281.8707222264593
ARIMA(2, 1, 3) - AIC:1274.6948122121792
ARIMA(3, 1, 0) - AIC:1297.4810917271661
ARIMA(3, 1, 1) - AIC:1282.4192776271907
ARIMA(3, 1, 2) - AIC:1283.720740597711
ARIMA(3, 1, 3) - AIC:1278.6579946366096
```

Table 30-AIC Values in ascending order

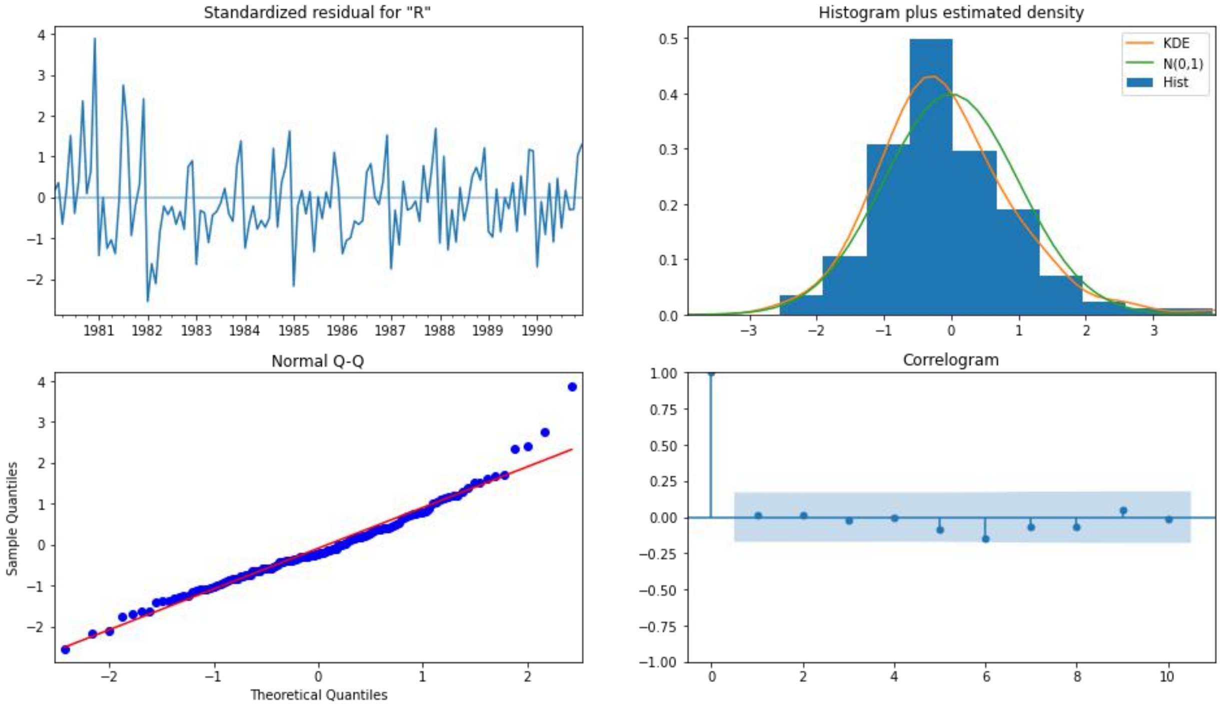
	param	AIC
11	(2, 1, 3)	1274.694812
15	(3, 1, 3)	1278.657995
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

So we will build the model with (2,1,3) which is giving the lowest AIC value

Table 31-SARIMAX result (2,1,3)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.347
Date:                Sun, 03 Jul 2022   AIC                           1274.695
Time:                        15:17:16   BIC                           1291.946
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6780      0.084    -20.035      0.000      -1.842      -1.514
ar.L2         -0.7288      0.084     -8.702      0.000      -0.893      -0.565
ma.L1          1.0447      0.670      1.559      0.119      -0.269       2.358
ma.L2         -0.7721      0.136     -5.696      0.000      -1.038      -0.506
ma.L3         -0.9048      0.608     -1.487      0.137      -2.097       0.288
sigma2       859.3109    564.947      1.521      0.128    -247.966    1966.587
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.48
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```

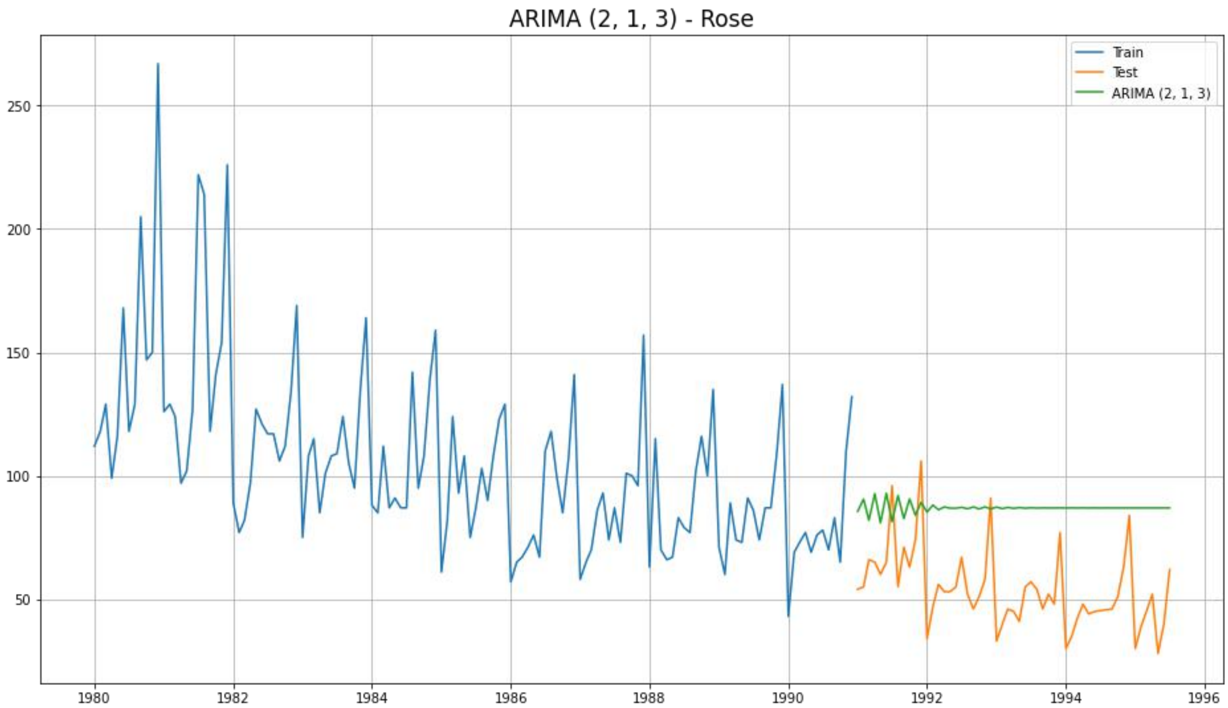
Table 32-Rose data diagnostic plot



The RMSE and MAPE values are as per below

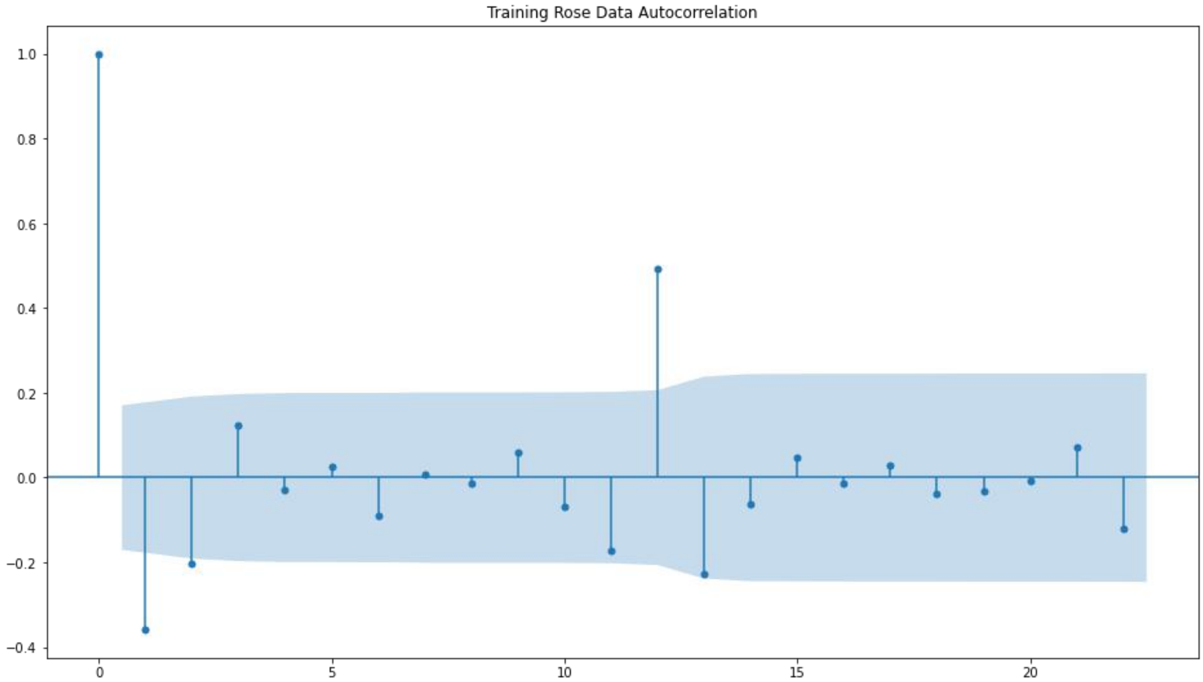
	Test RMSE	Test MAPE
ARIMA(2,1,3)	36.811514	75.835805

Table 33-ARIMA Plot on train and test rose data



Checking the ACF PACF plots

Table 34-ACF Plot of rose train data



SARIMA

Here we will take seasonality into consideration and take it as 6 months seasonality

```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (0, 1, 3)(0, 0, 3, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (1, 1, 3)(1, 0, 3, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
Model: (2, 1, 3)(2, 0, 3, 6)
Model: (3, 1, 0)(3, 0, 0, 6)
Model: (3, 1, 1)(3, 0, 1, 6)
Model: (3, 1, 2)(3, 0, 2, 6)
Model: (3, 1, 3)(3, 0, 3, 6)
```

Table 35-AIC Values for different values and seasonality

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	951.744302
59	(0, 1, 3)	(2, 0, 3, 6)	952.073632
251	(3, 1, 3)	(2, 0, 3, 6)	952.582104
191	(2, 1, 3)	(3, 0, 3, 6)	953.205617
123	(1, 1, 3)	(2, 0, 3, 6)	953.684951

We will now build our model using (2,1,3) and (2,0.3,6) which gives the lowest AIC value.

And test the result.

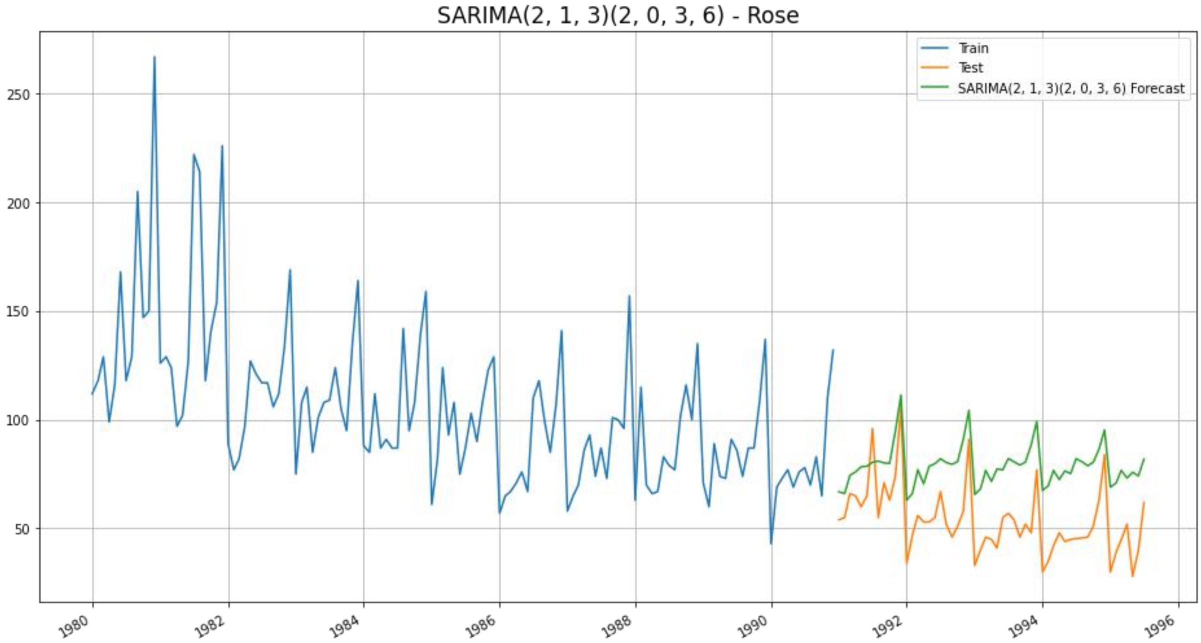
Table 36-SARIMAX model on Rose data

```
                                    SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -464.872
Date:                           Sun, 03 Jul 2022   AIC                            951.744
Time:                                   15:21:06   BIC                            981.349
Sample:                               01-01-1980   HQIC                           963.750
                                    - 12-01-1990
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5027      0.083     -6.081      0.000      -0.665      -0.341
ar.L2         -0.6627      0.084     -7.919      0.000      -0.827      -0.499
ma.L1         -0.3713    186.636     -0.002      0.998    -366.171     365.428
ma.L2          0.2032    117.314      0.002      0.999    -229.729     230.135
ma.L3         -0.8319    155.219     -0.005      0.996    -305.055     303.392
ar.S.L6       -0.0837      0.049     -1.719      0.086      -0.179       0.012
ar.S.L12       0.8099      0.052     15.464      0.000       0.707       0.913
ma.S.L6        0.1699      0.247      0.689      0.491      -0.314       0.653
ma.S.L12      -0.5643      0.198     -2.847      0.004      -0.953      -0.176
ma.S.L18       0.1709      0.142      1.199      0.230      -0.108       0.450
sigma2       260.8272   4.87e+04      0.005      0.996   -9.52e+04    9.57e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.72   Jarque-Bera (JB):                 4.77
Prob(Q):                              0.40   Prob(JB):                         0.09
Heteroskedasticity (H):               0.54   Skew:                            -0.36
Prob(H) (two-sided):                  0.06   Kurtosis:                         3.73
===================================================================================
```

The RMSE and MAPE values of SARIMA model is plotted below.

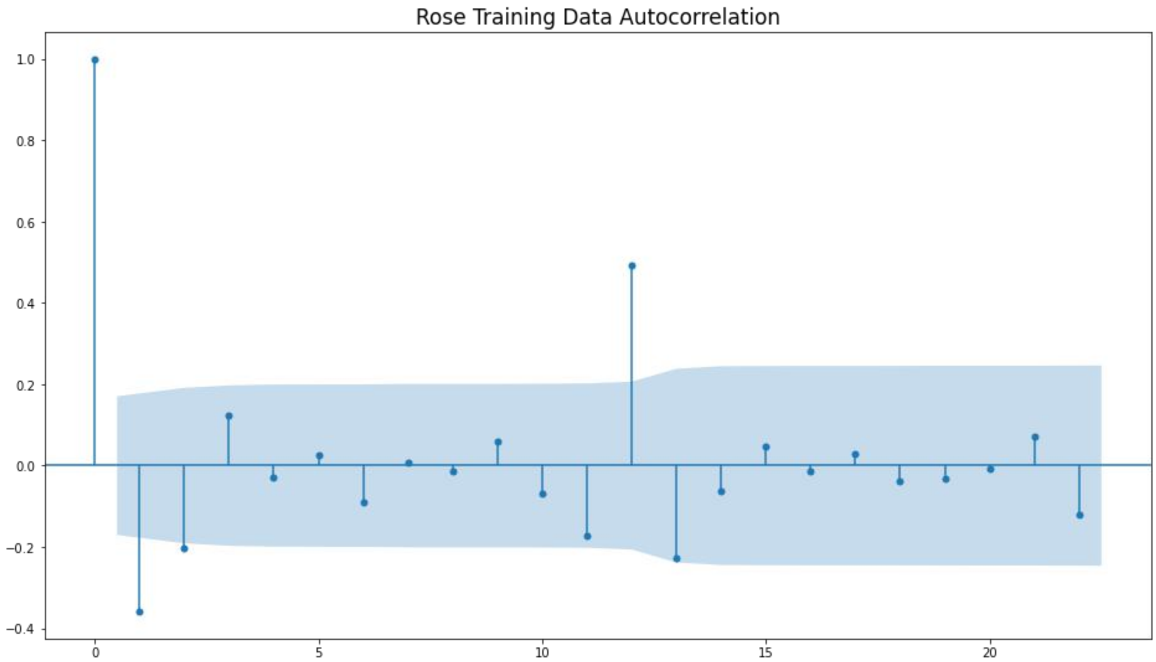
	Test RMSE	Test MAPE
ARIMA(2,1,3)	36.811514	75.835805
SARIMA(2, 1, 3)(2, 0, 3, 6)	27.124116	55.239893

Table 37-SRIMA Model(2,1,3) (2,0,3,6)



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.

Table 38-Rose train ACF graph



Here, we have taken alpha=0.05.

The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off

The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off

By looking at the above plots, we will take the value of p and q to be 2 and 2 respectively.¶

Table 39-ARIMA MODEL mannual

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -635.935
Date:                Sun, 03 Jul 2022   AIC                           1281.871
Time:                        15:21:41   BIC                           1296.247
Sample:                    01-01-1980   HQIC                          1287.712
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4540      0.469     -0.969      0.333      -1.372       0.464
ar.L2          0.0001      0.170      0.001      0.999      -0.334       0.334
ma.L1         -0.2541      0.459     -0.554      0.580      -1.154       0.646
ma.L2         -0.5984      0.430     -1.390      0.164      -1.442       0.245
sigma2       952.1601     91.424     10.415      0.000     772.973    1131.347
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                34.16
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.37   Skew:                             0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.94
===================================================================================
```

Table 40-Mannual ARIMA diagnostic plot

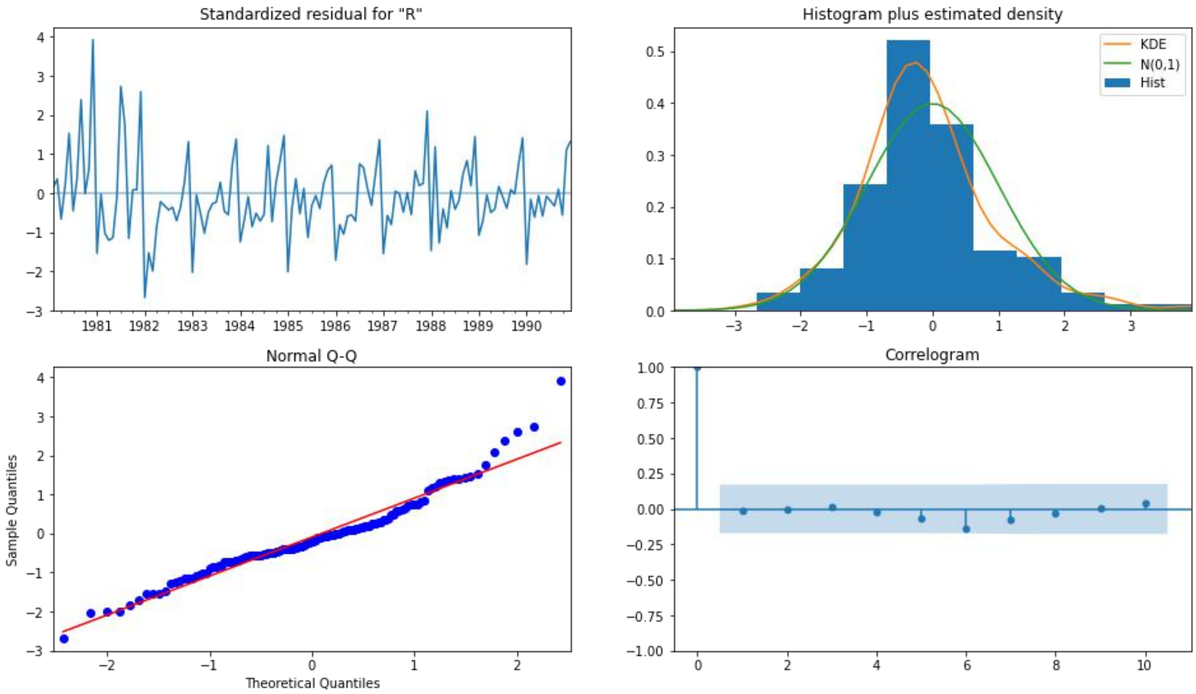
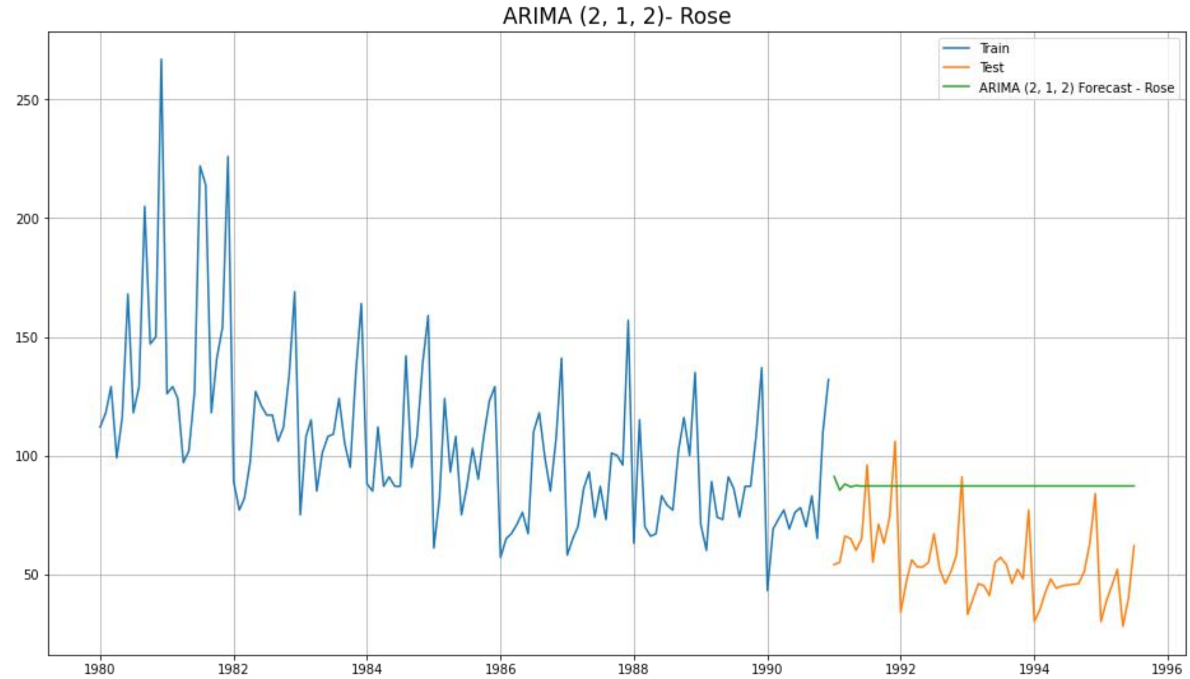
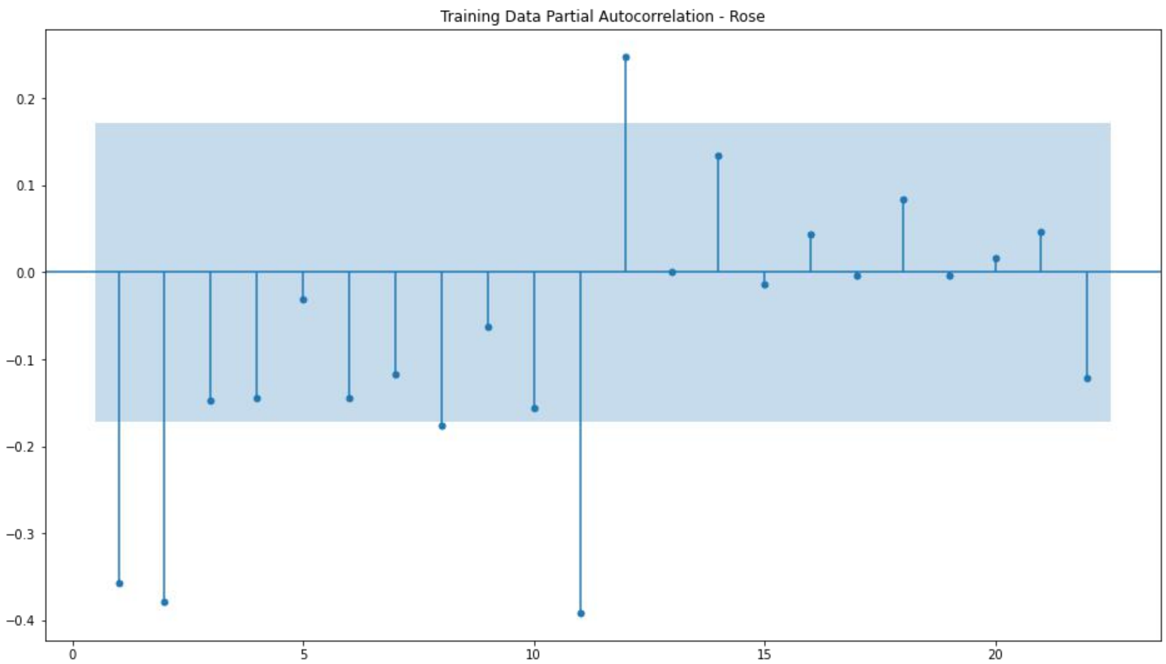


Table 41-ARIMA (2,1,2)



Checking ARIMA Model based on ACF PACF for mannual SARIMA



Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12

We are taking the p value to be 2 and the q value also to be 2 as the parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 3

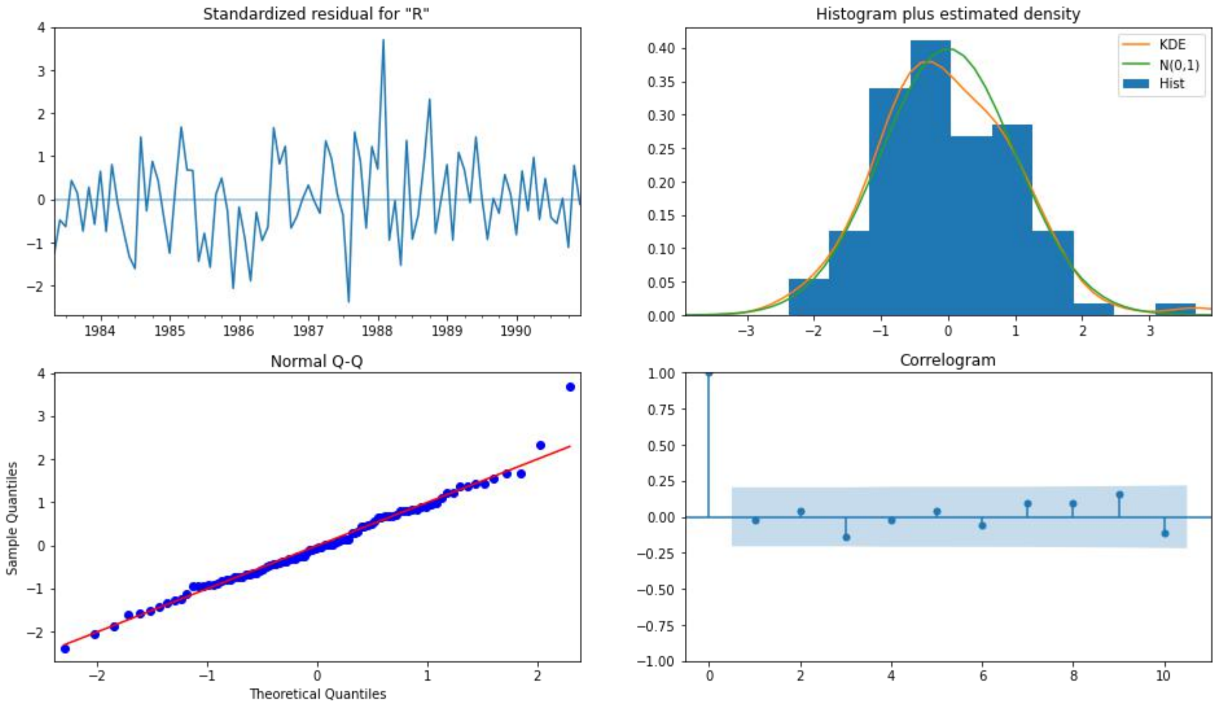
We have built the model using the order as p , q , d as 2 , 1 , 2 and seasonal parameter as (2, 1 ,2 , 12)

Below is the result summary

Table 42-Mannual SARIMA result summary

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -379.498
Date:                            Sun, 03 Jul 2022   AIC                            776.996
Time:                                    15:21:46   BIC                            799.692
Sample:                                01-01-1980   HQIC                           786.156
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.8551      0.146     -5.838      0.000      -1.142      -0.568
ar.L2         -0.0022      0.125     -0.017      0.986      -0.247       0.242
ma.L1          0.0120      0.184      0.065      0.948      -0.348       0.372
ma.L2         -0.9435      0.150     -6.292      0.000      -1.237      -0.650
ar.S.L12       0.0347      0.185      0.188      0.851      -0.328       0.397
ar.S.L24      -0.0459      0.029     -1.599      0.110      -0.102       0.010
ma.S.L12      -0.7223      0.333     -2.172      0.030      -1.374      -0.071
ma.S.L24      -0.0772      0.212     -0.364      0.716      -0.493       0.339
sigma2       192.1802     39.479      4.868      0.000     114.802     269.558
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                 7.06
Prob(Q):                              0.86   Prob(JB):                         0.03
Heteroskedasticity (H):               0.87   Skew:                             0.45
Prob(H) (two-sided):                  0.71   Kurtosis:                         4.01
===================================================================================
```

Table 43-Mannual SARIMA diagnostic plot



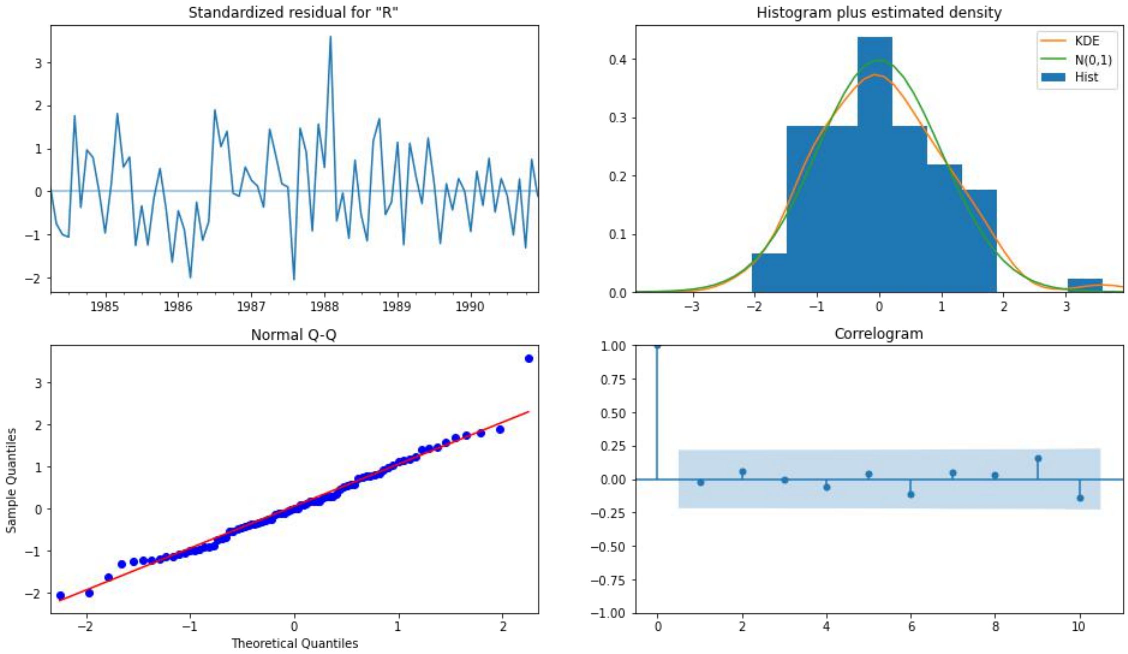
We have built another model taking the seasonal parameter as (3, 1 , 2 , 12)

Below is the result summary

Table 44-Mannual SARIMA (2,1,2)(3,1,2,12) result summary

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                132
Model:             SARIMAX(2, 1, 2)x(3, 1, 2, 12)   Log Likelihood                -334.893
Date:                              Sun, 03 Jul 2022   AIC                            689.786
Time:                                      15:22:00   BIC                            713.730
Sample:                                  01-01-1980   HQIC                           699.392
                                       - 12-01-1990
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7091      0.403      1.758      0.079      -0.081       1.500
ar.L2         -0.1501      0.176     -0.854      0.393      -0.494       0.194
ma.L1         -1.6100      0.422     -3.820      0.000      -2.436      -0.784
ma.L2          0.6498      0.396      1.639      0.101      -0.127       1.427
ar.S.L12      -0.0423      0.234     -0.181      0.856      -0.501       0.416
ar.S.L24      -0.0169      0.158     -0.107      0.915      -0.327       0.293
ar.S.L36   -4.883e-06      0.067  -7.34e-05      1.000      -0.130       0.130
ma.S.L12      -0.8406     49.632     -0.017      0.986     -98.118      96.437
ma.S.L24      -0.1606      7.738     -0.021      0.983     -15.326      15.005
sigma2       185.5463   9199.789      0.020      0.984   -1.78e+04    1.82e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                 4.60
Prob(Q):                              0.82   Prob(JB):                         0.10
Heteroskedasticity (H):               0.63   Skew:                             0.48
Prob(H) (two-sided):                  0.24   Kurtosis:                         3.67
===================================================================================
```

Table 45-Mannual SARIMA (2,1,2) (3,1,2,12) diagnostic plot



The RMSE and MAPE Value of the models are plotted below

	Test RMSE	Test MAPE
ARIMA(2,1,2)	36.871197	76.056213
SARIMA(2,1,2)(3,1,2,12)	15.357192	22.955567

Here there is no change in mannual ARIMA vs Auto ARIMA but in mannual SARIMA there is a difference.

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.¶

Table 46-All model RMSE & MAPE score

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	NaN
NaiveModel	79.718773	NaN
SimpleAverageModel	53.460570	NaN
2pointTrailingMovingAverage	11.529278	NaN
4pointTrailingMovingAverage	14.451403	NaN
6pointTrailingMovingAverage	14.566327	NaN
9pointTrailingMovingAverage	14.727630	NaN
SimpleExponentialSmoothing	36.796242	NaN
Double Exponential Smoothing	15.268957	NaN
Triple Exponential Smoothing (Additive Season)	14.278440	NaN
Triple Exponential Smoothing (Multiplicative Season)	20.189764	NaN
ARIMA(2,1,3)	36.811514	75.835805
SARIMA(2, 1, 3)(2, 0, 3, 6)	27.124116	55.239893
SARIMA(2,1,2)(3,1,2,12)	15.357192	22.955567
ARIMA(2,1,2)	15.357192	22.955567

The 2-point trailing moving average model performs the best.

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.

Table 47-Best model graph for Rose data

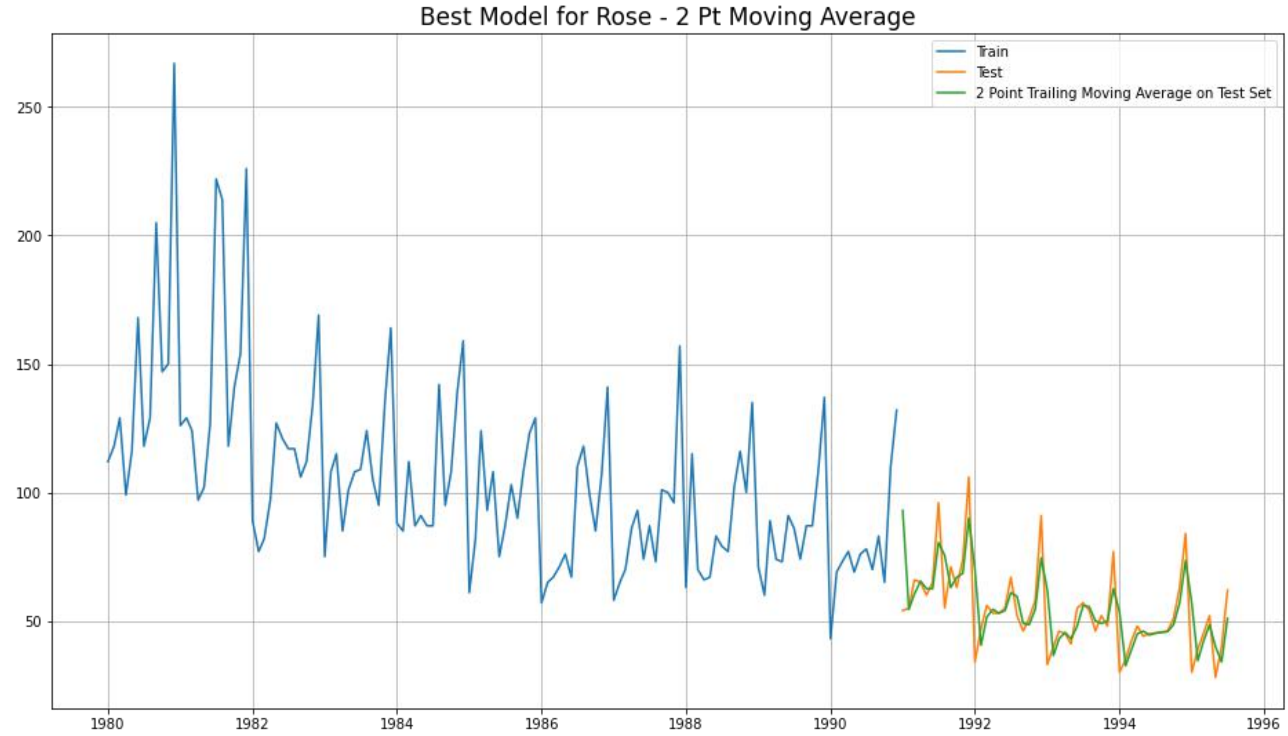
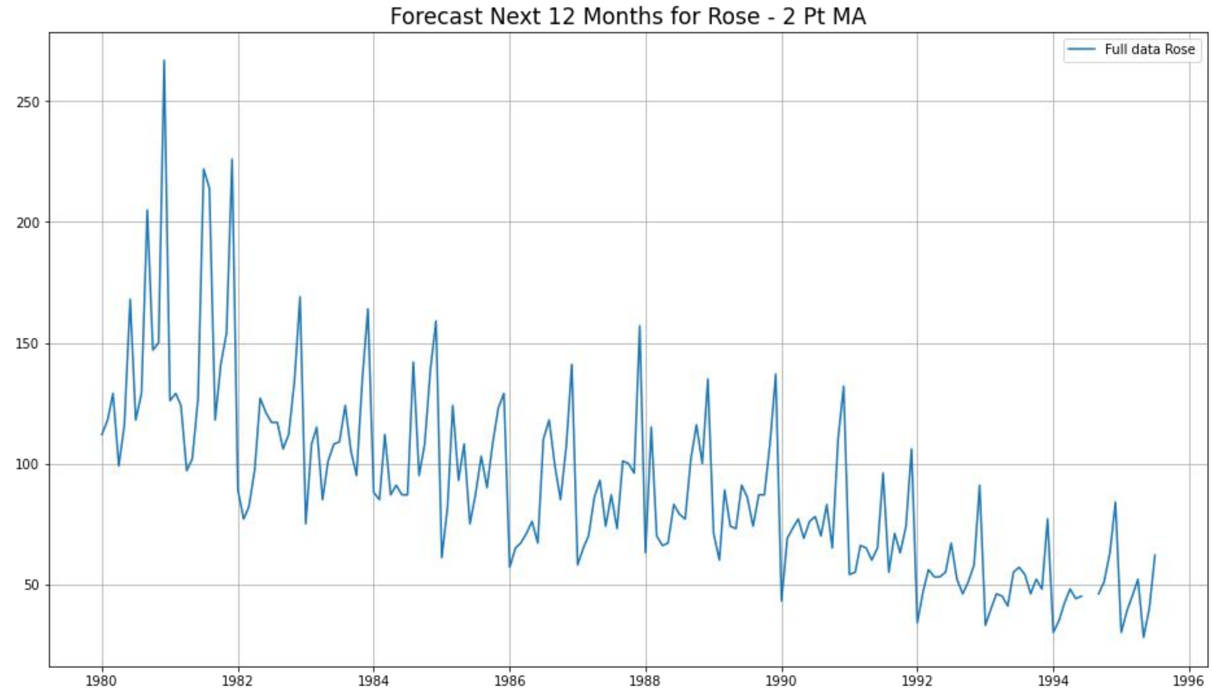


Table 48-Forecast on best model

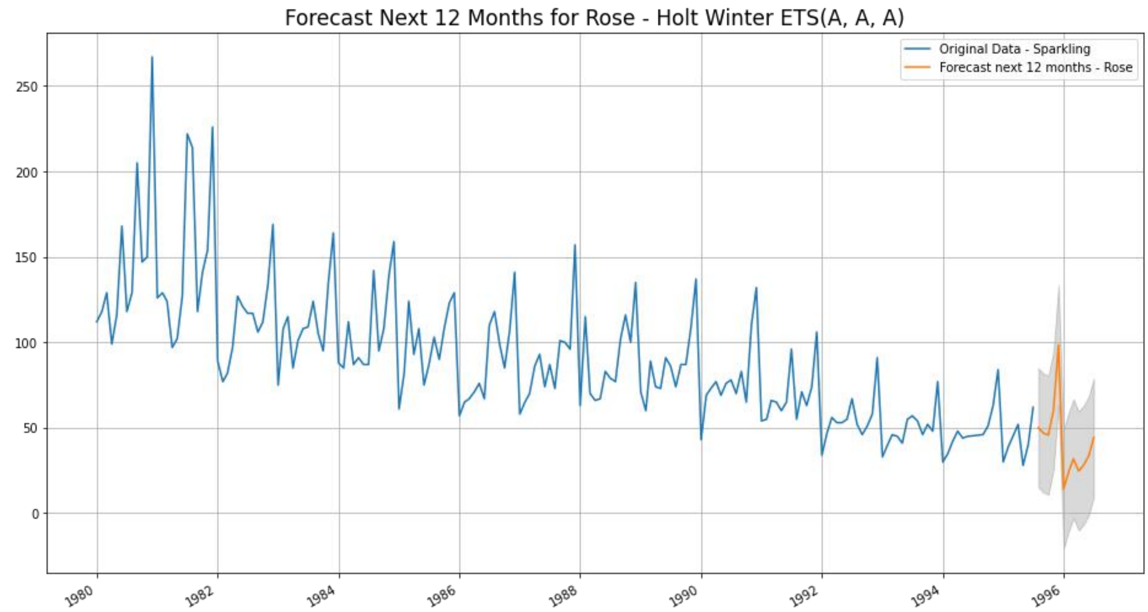


But we are not very satisfied with the forecast plot as the graph is not looking continuous

So we would like to plot the second best model i.e TES additive

Below is the forecast using TES additive model

Table 49-second best model forecast on Rose data



So looking at the above graph we can say that TES additive model gives us the most accurate forecast for ROSE wine sales

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

The best model here is found to be TES additive , and the forecast has been done for next 1 yr.

As we can see that there is going to be an upward trend in demand and hence the production team has to be prepared with proper plan to meet the demand with raw material procurement and logistics arrangement.

Company can decrease the price to generate more demand if it is a cash cow product.

Over the year the trend it decreasing so company can decide to close this product also if it is not a core product.

SPARKLING WINE SALES DATA

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

There are 187 observations from 1980-01-01 to 1995-07-01 and we don’t have any missing data here.

Below are head and tail of the data

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

Table 50-head of sparkling data

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

Table 51-tail of sparkling data

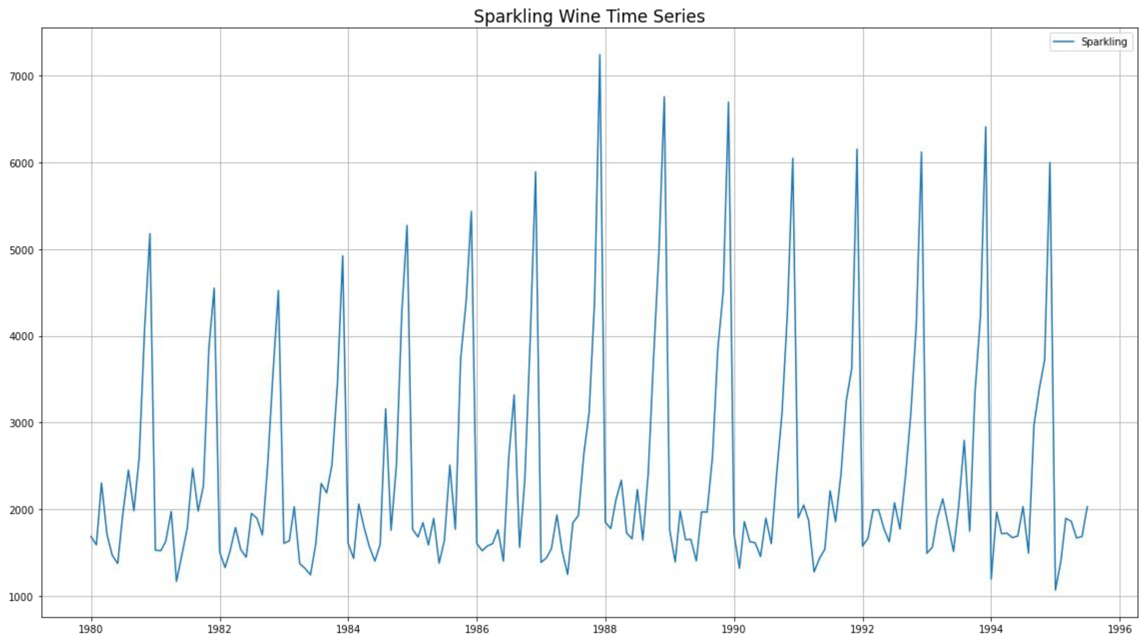


Table 52-Time series plot of sparkling data

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

	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

Table 53-Sparkling data description

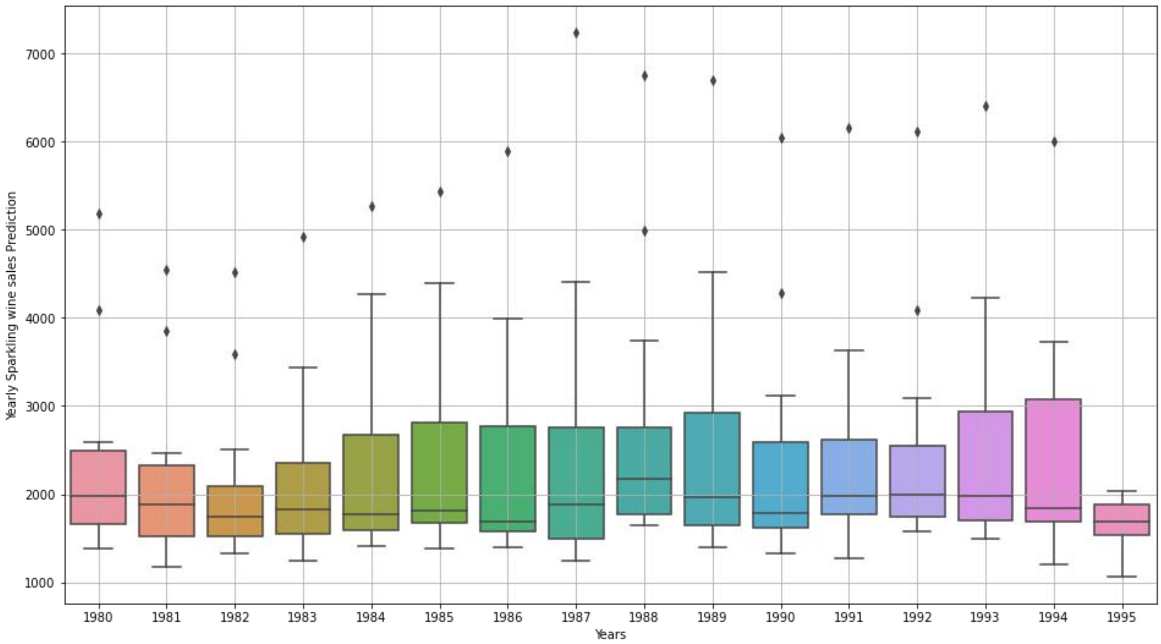


Table 54-Year on year box plot of sparkling data

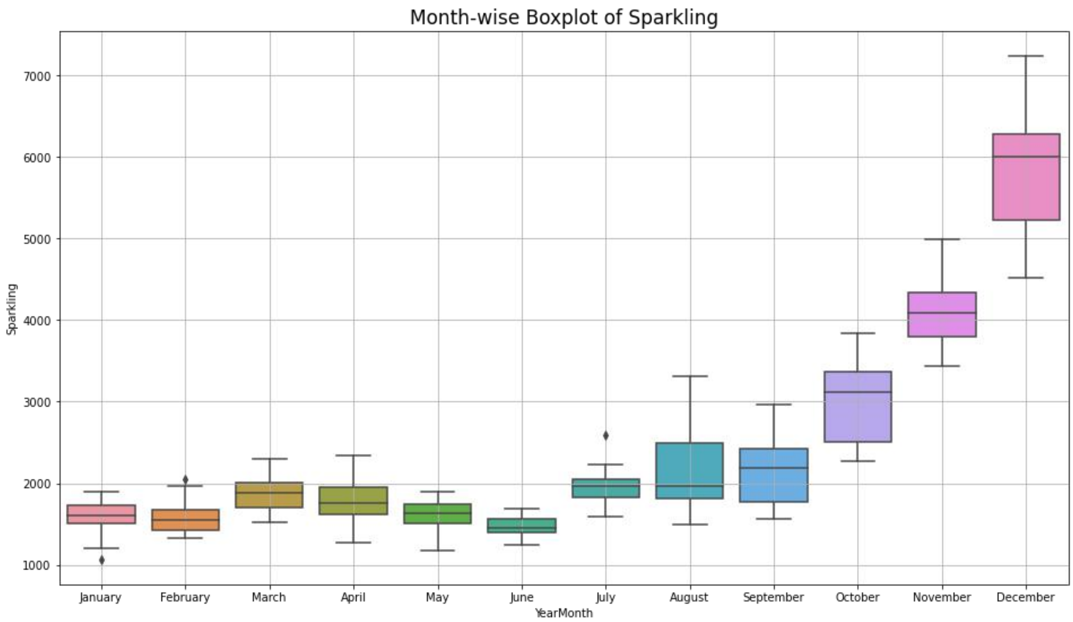


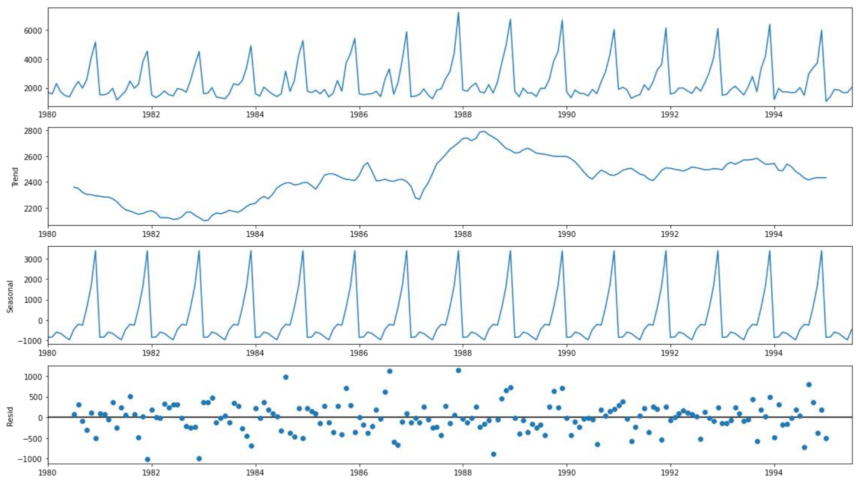
Table 55-Monthly box plot of sparkling data

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

Table 56-Yearly sales across the month of sparkling data

Decomposition

Additive model



There is seasonality but no trend

Multiplicative model

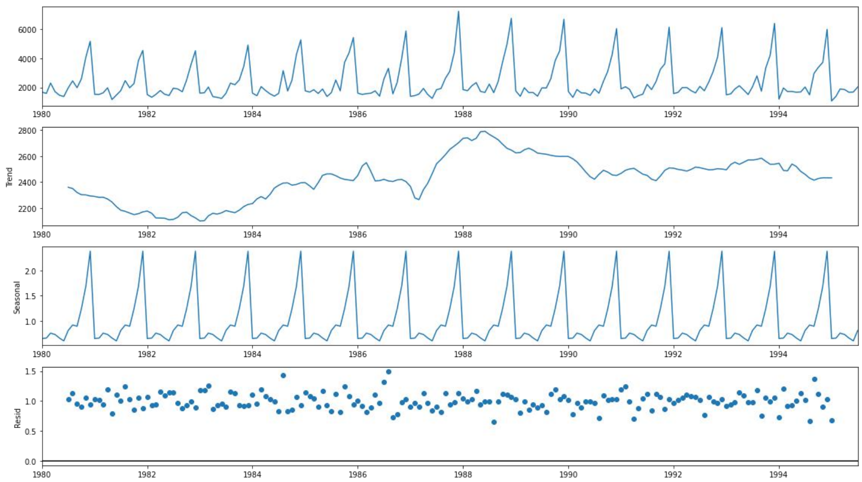


Table 57-Sparkling data multiplicative decomposition

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

Data is splitted from 1991 in train and test.

Before 1991 train and after that it is test data.

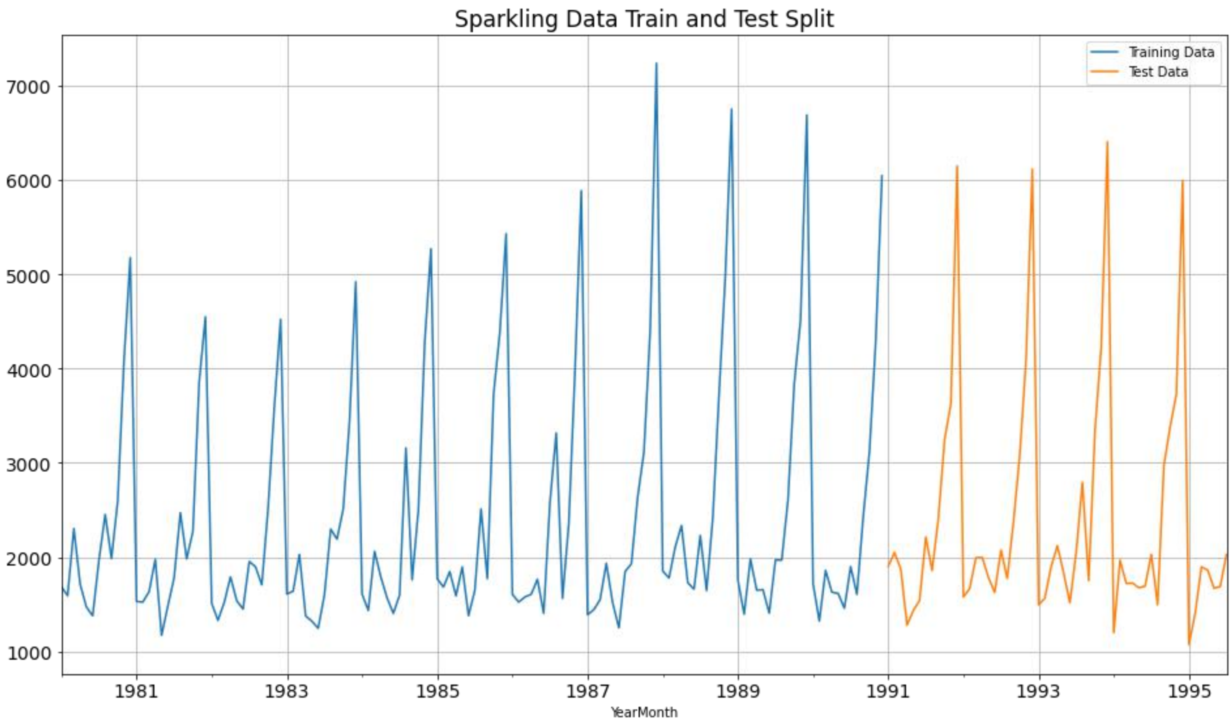


Table 58-Sparkling train and test data plot

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, naive forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Model-1 (Linear regression)

We divided the data into train and test and then implemented the LR model, below is the plot of forecast

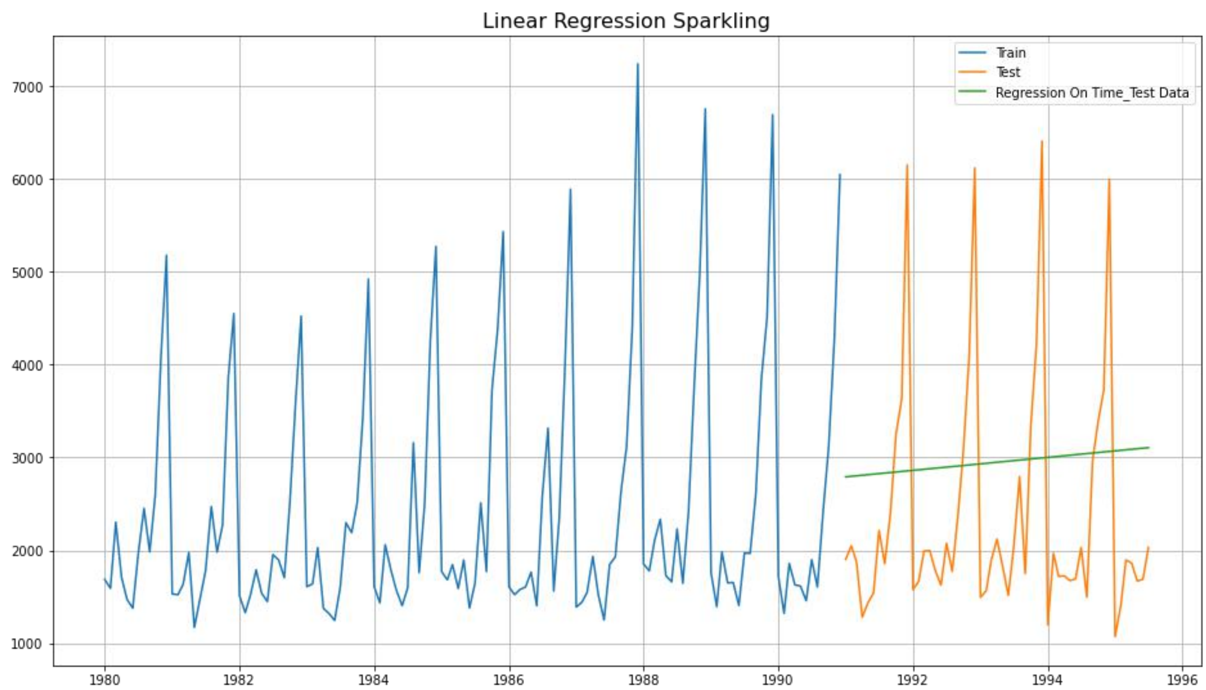


Table 59-Linear regression on sparkling data

The RMSE value is 1389

Model-2 (Naive Approach)

We have built the model and forecast the future values , below is the plot

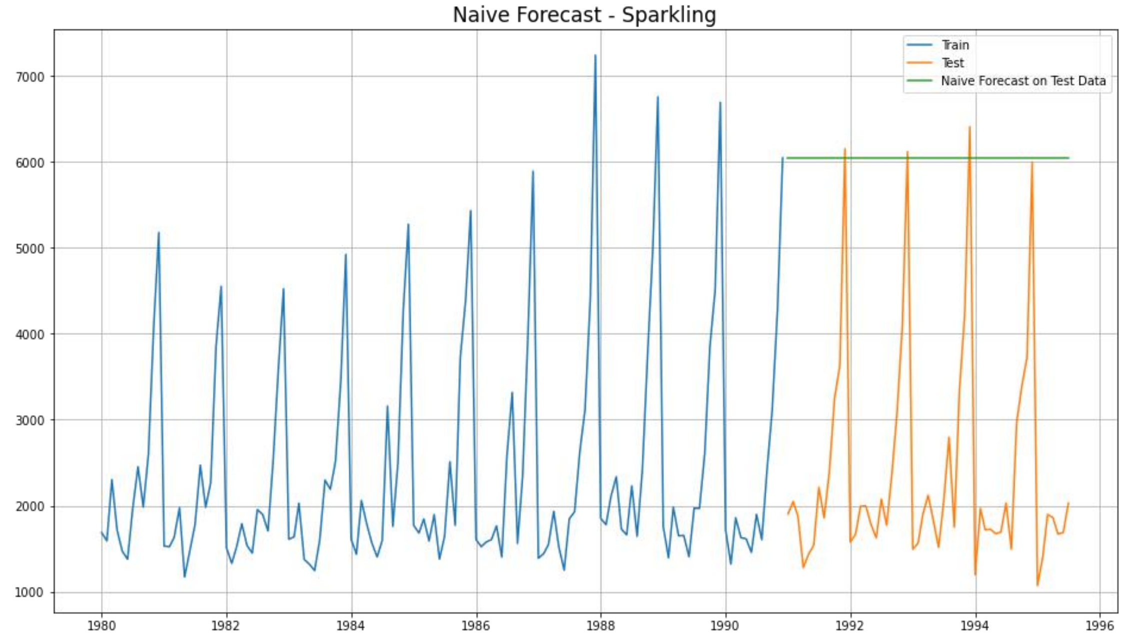


Table 60-Naive approach forecast on sparkling data

The RMSE value found to be 3864

Model-3 (Simple average)

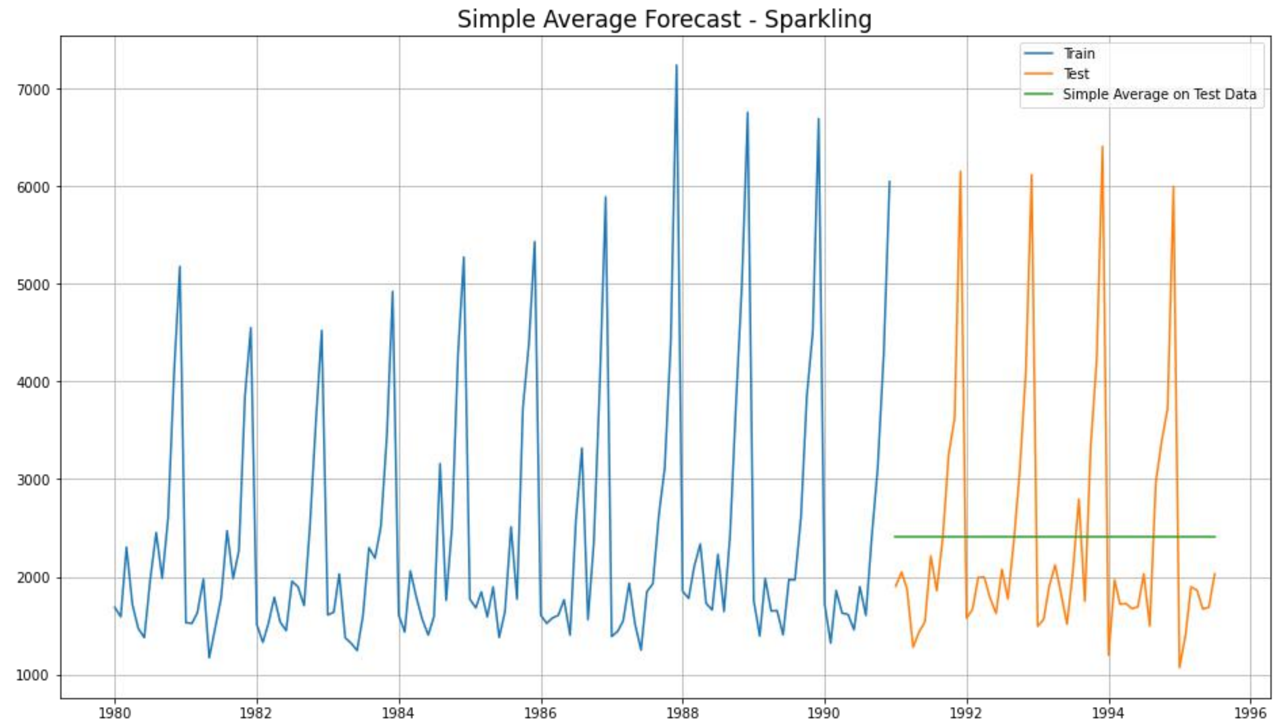


Table 61-Simple average forecast on sparkling data

The RMSE value found to be 1275

Model-4 (Moving average)

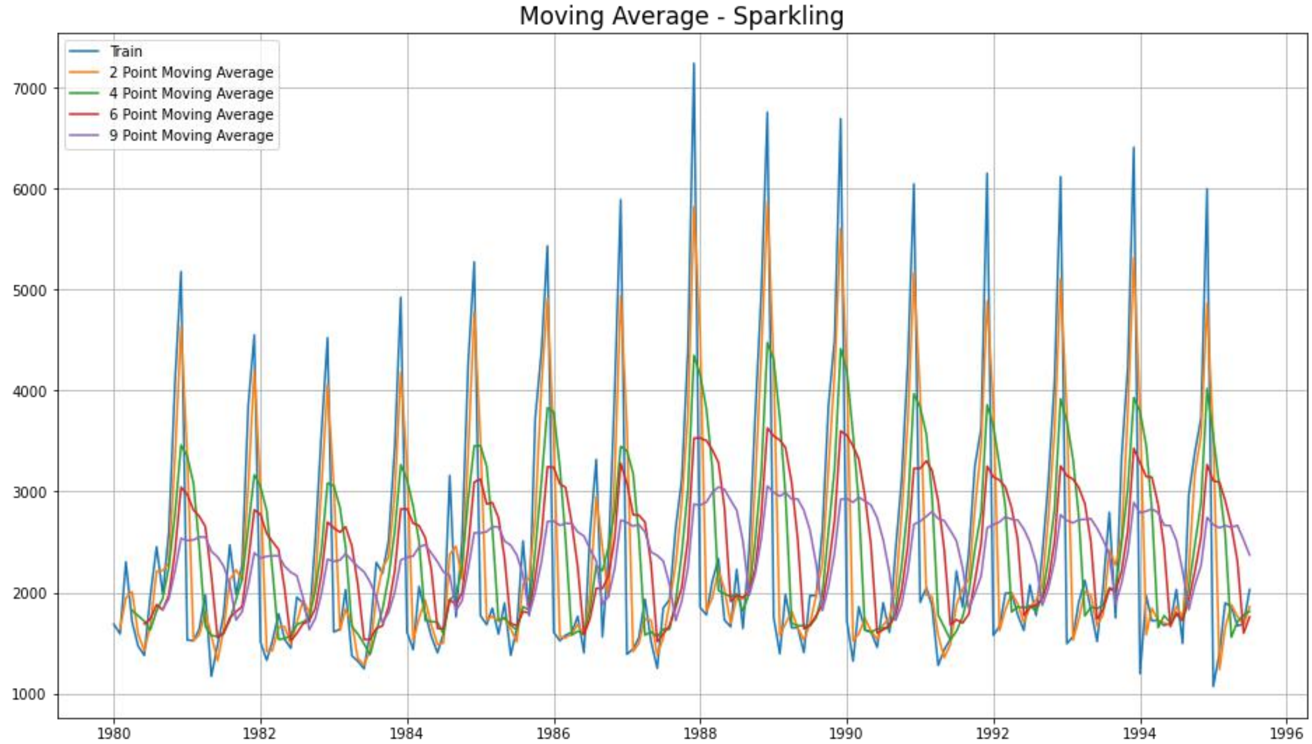


Table 62-Moving average forecast on whole sparkling data

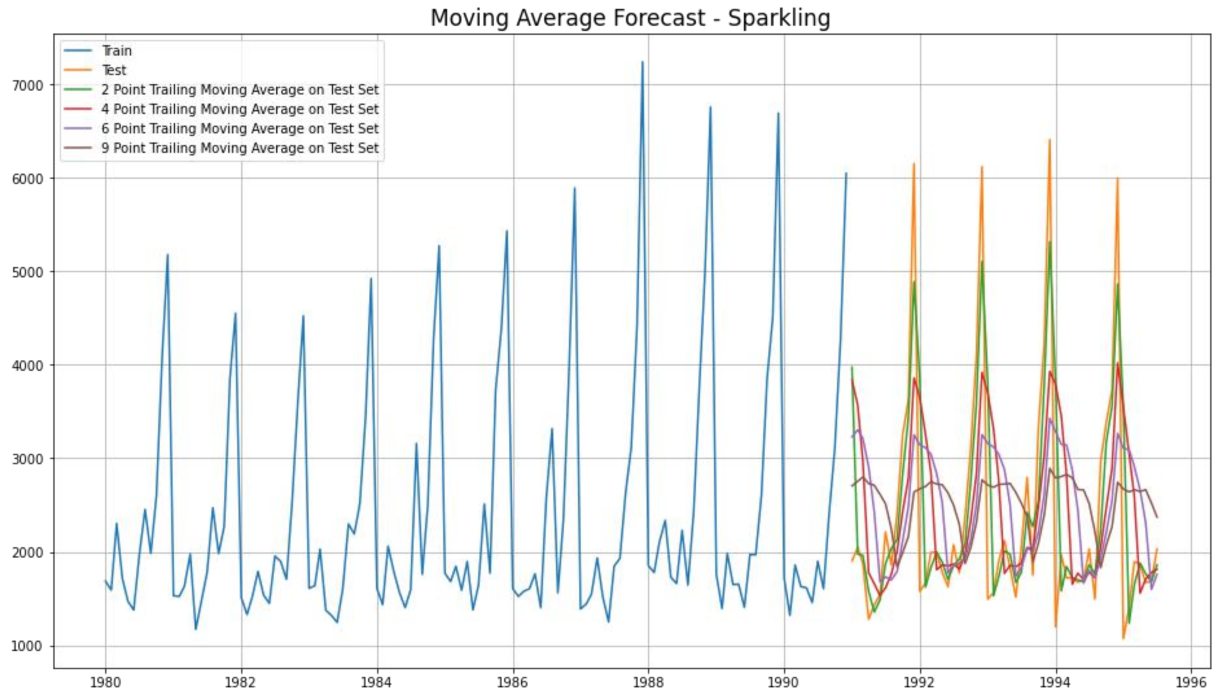


Table 63-Moving average forecast on train and test data

Below is the RMSE value of each trailing moving average

```
For 2 point Moving Average Model forecast on Sparkling Training Data,  RMSE is 813.401
For 4 point Moving Average Model forecast on Sparkling Training Data,  RMSE is 1156.590
For 6 point Moving Average Model forecast on Sparkling Training Data,  RMSE is 1283.927
For 9 point Moving Average Model forecast on Sparkling Training Data,  RMSE is 1346.278
```

Table 64-RMSE Value of trailing moving average

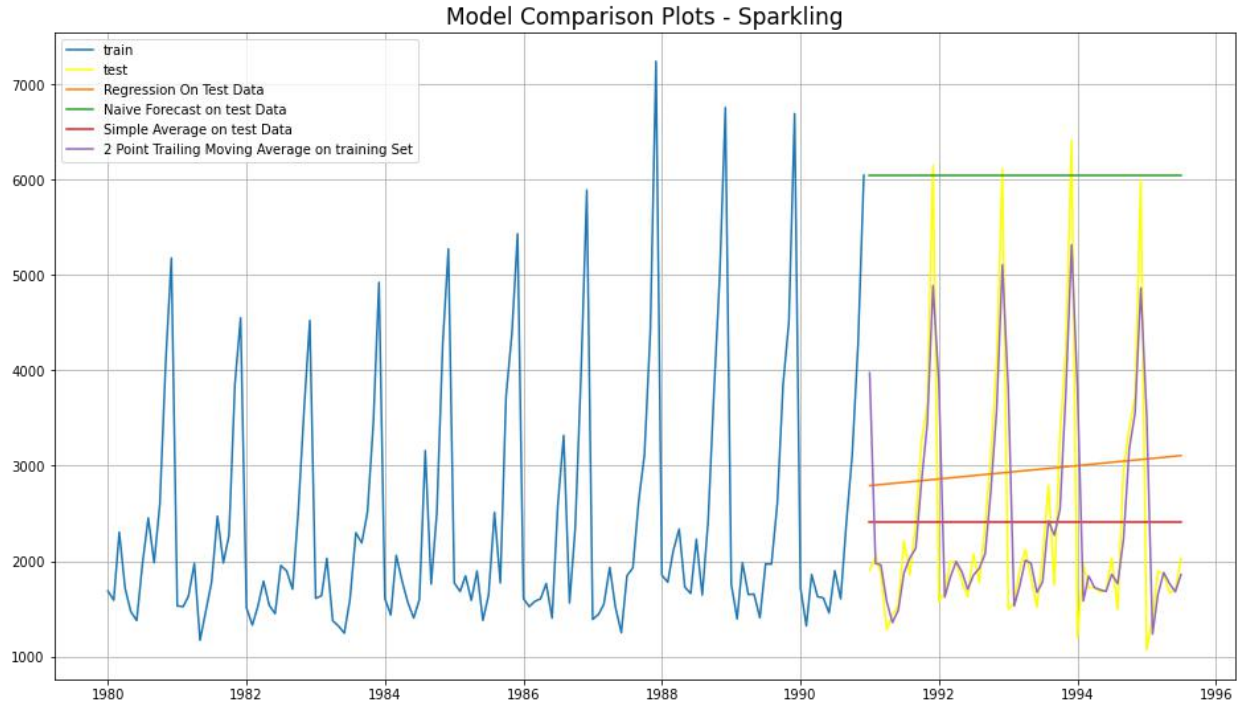


Table 65-All model graph

2-pt trailing moving average is doing best till now.

Simple exponential smoothing model

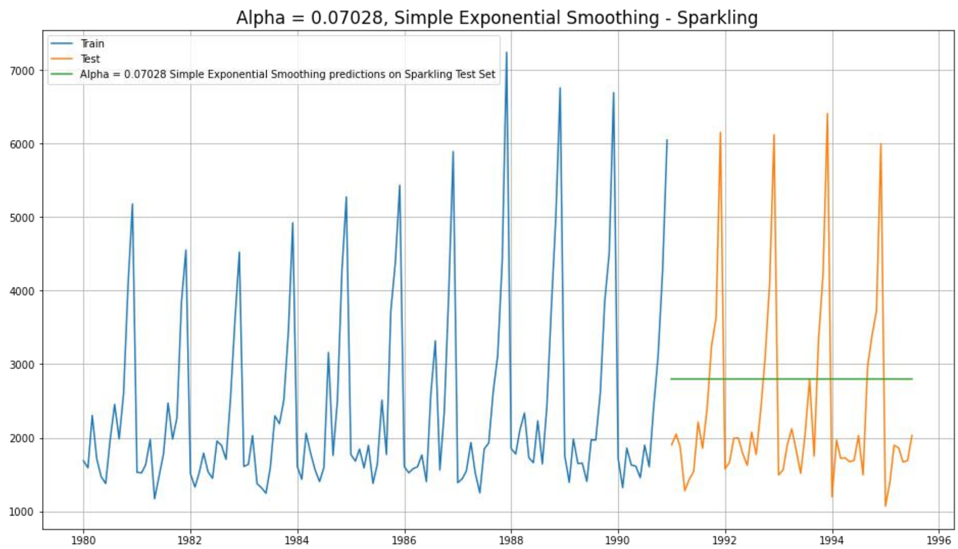


Table 66-SES model forecast on sparkling data

Model-6 Holt ETS (A , A, N)- Holt's linear method with additive errors

Double exponential smoothing

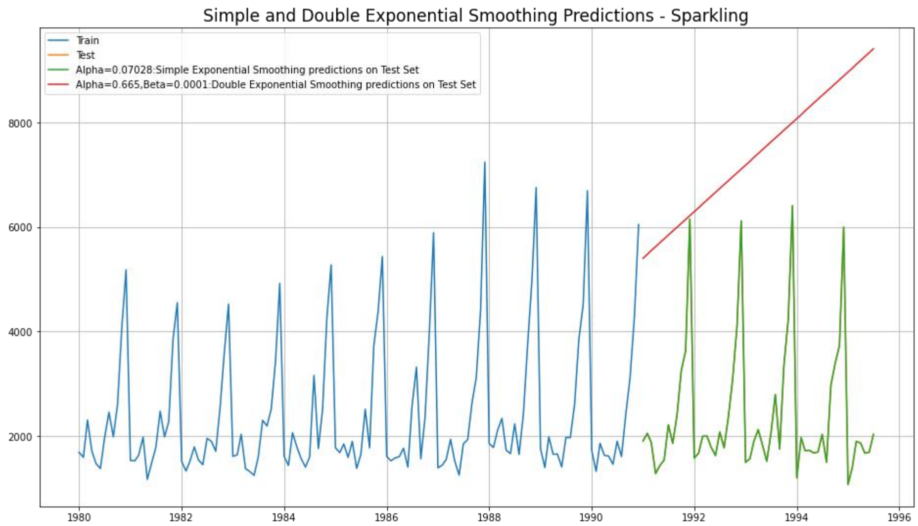


Table 67-Double Exponential smoothing on sparkling data

The RMSE score is 5291

Tripple exponential smoothing

Additive model

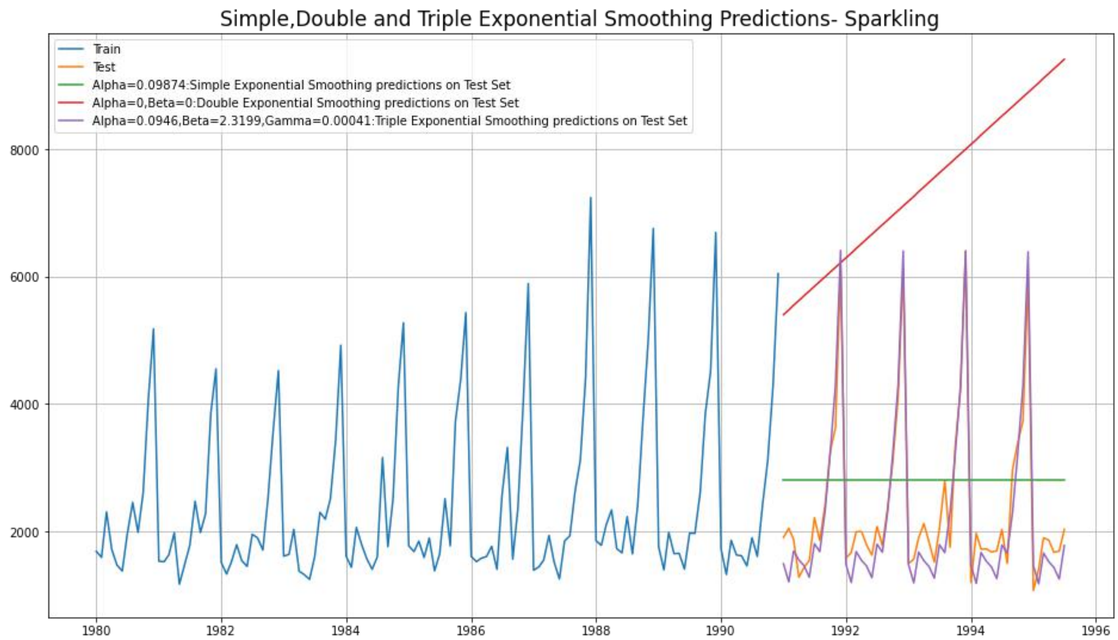


Table 68-TES additive model on sparkling data

The RMSE value here is 378.625

Tripple exponential smoothing

Multiplicative model

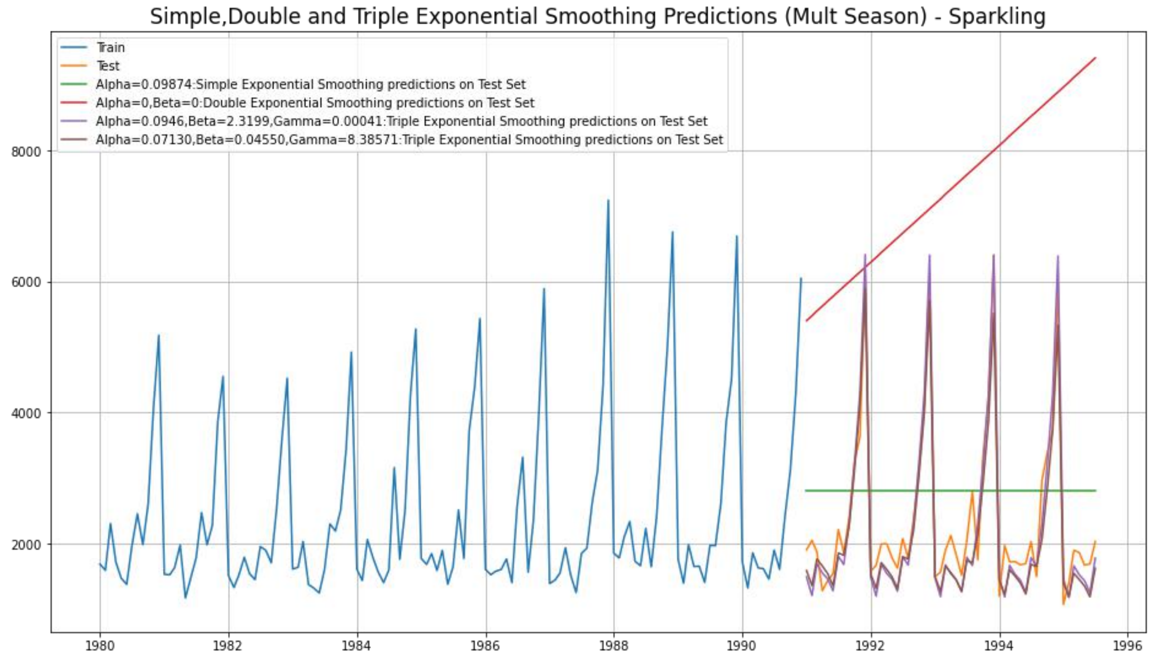


Table 69-TES multiplicative forecast on sparkling data

The RMSE score is 402 for this model

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.

Checking Stationarity of sparkling Wine sales Data. The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary. The hypothesis in a simple form for the ADF test is:

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the

Hypothesis for Statistical Test for sparkling wine sales.

H0: The Time series of sparkling Wine has a unit root and is thus a Non- stationary.

H1: The Time series of sparkling Wine does not have a unit root and is thus Stationary.

We have done ADF test to check the stationarity , below is the test result

```
DF test statistic is -1.798
DF test p-value +is 0.705595845993243
Number of lags used 12
```

Here p>0.05 hence data is non stationary

We have to do differentiation to make it stationary

Below is the result after differentiation

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

Now p<0.05 hence the data has become stationary now

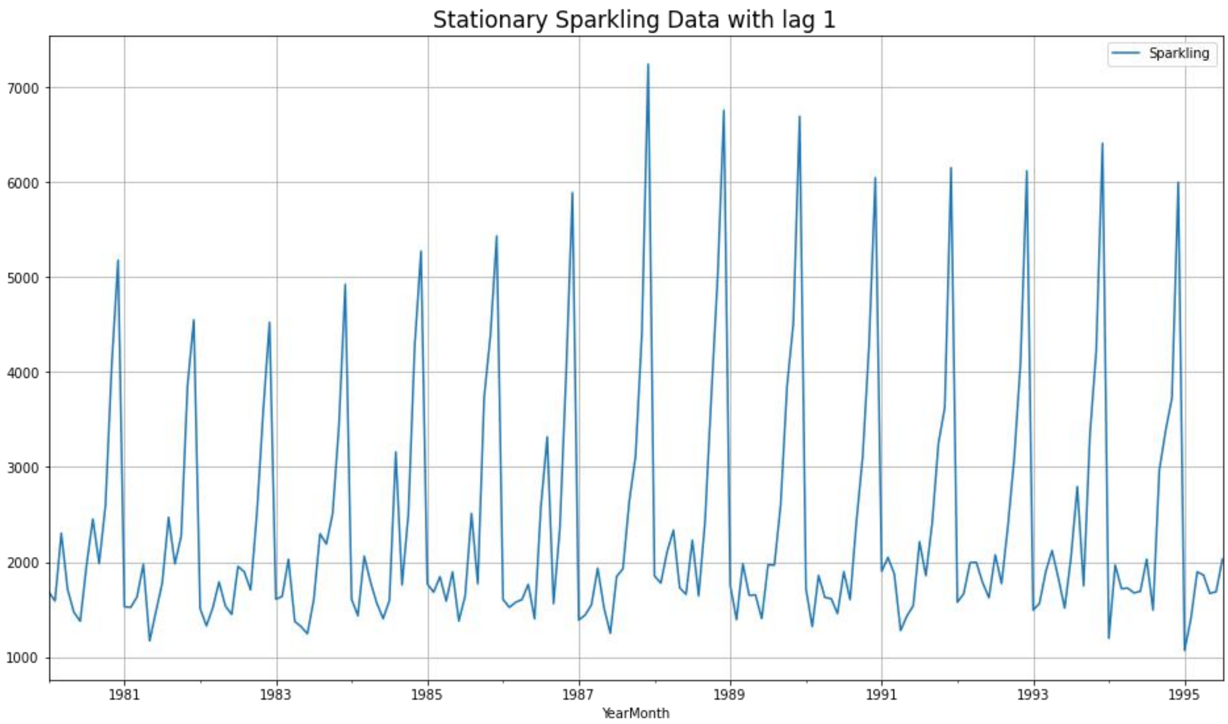


Table 70-Stationary sparkling data

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.

Before building ARIMA/SARIMA we have to check the stationarity of the train data

```
DF test statistic is -2.062
DF test p-value is 0.5674110388593686
Number of lags used 12
```

After the ADF test p>0.05 hence train data is non stationary

Now we have to differentiate the train data to make it stationary

```
DF test statistic is -7.968
DF test p-value is 8.4792106555143e-11
Number of lags used 11
```

P<0.05 now so the train data has now become stationary now.

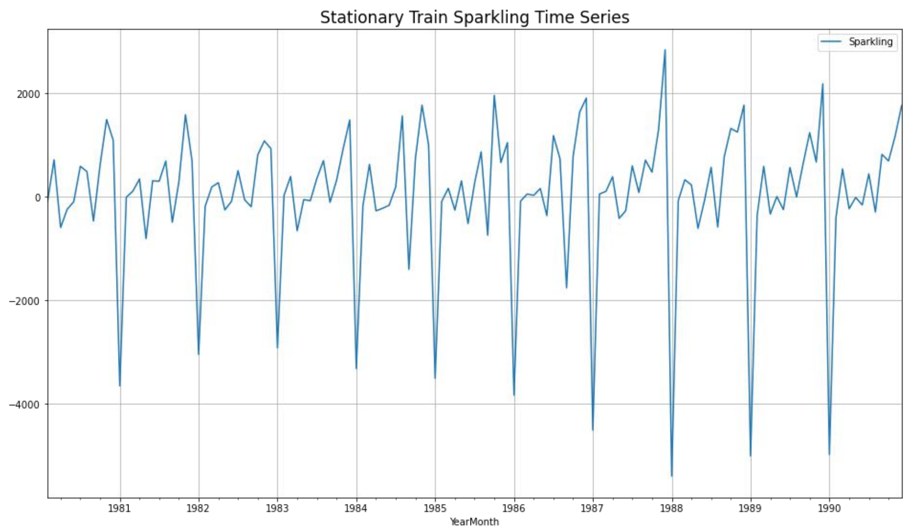


Table 71-Stationary train data of sparkling

Auto ARIMA model

Combinations of p , q , d

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
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)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

AIC values for different p , q , d

```
ARIMA(0, 1, 0) - AIC:2267.6630357855465
ARIMA(0, 1, 1) - AIC:2263.060015592133
ARIMA(0, 1, 2) - AIC:2234.4083231291497
ARIMA(0, 1, 3) - AIC:2233.9948577420064
ARIMA(1, 1, 0) - AIC:2266.6085393190097
ARIMA(1, 1, 1) - AIC:2235.7550946704996
ARIMA(1, 1, 2) - AIC:2234.5272004521435
ARIMA(1, 1, 3) - AIC:2235.6078163476477
ARIMA(2, 1, 0) - AIC:2260.36574396809
ARIMA(2, 1, 1) - AIC:2233.777626308103
ARIMA(2, 1, 2) - AIC:2213.5092123973454
ARIMA(2, 1, 3) - AIC:2232.9305188226363
ARIMA(3, 1, 0) - AIC:2257.72337899794
ARIMA(3, 1, 1) - AIC:2235.4988428729475
ARIMA(3, 1, 2) - AIC:2230.7666102201465
ARIMA(3, 1, 3) - AIC:2221.461700336058
```

The lowest AIC score parameter

	param	AIC
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.461700
14	(3, 1, 2)	2230.766610
11	(2, 1, 3)	2232.930519
9	(2, 1, 1)	2233.777626

Table 72-AIC score table for auto arima

Result summary

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 03 Jul 2022   AIC                           2213.509
Time:                        15:22:30   BIC                           2227.885
Sample:                    01-01-1980   HQIC                          2219.351
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.782      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.741      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.217      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                14.46
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================
```

Table 73-Auto ARIMA result summary

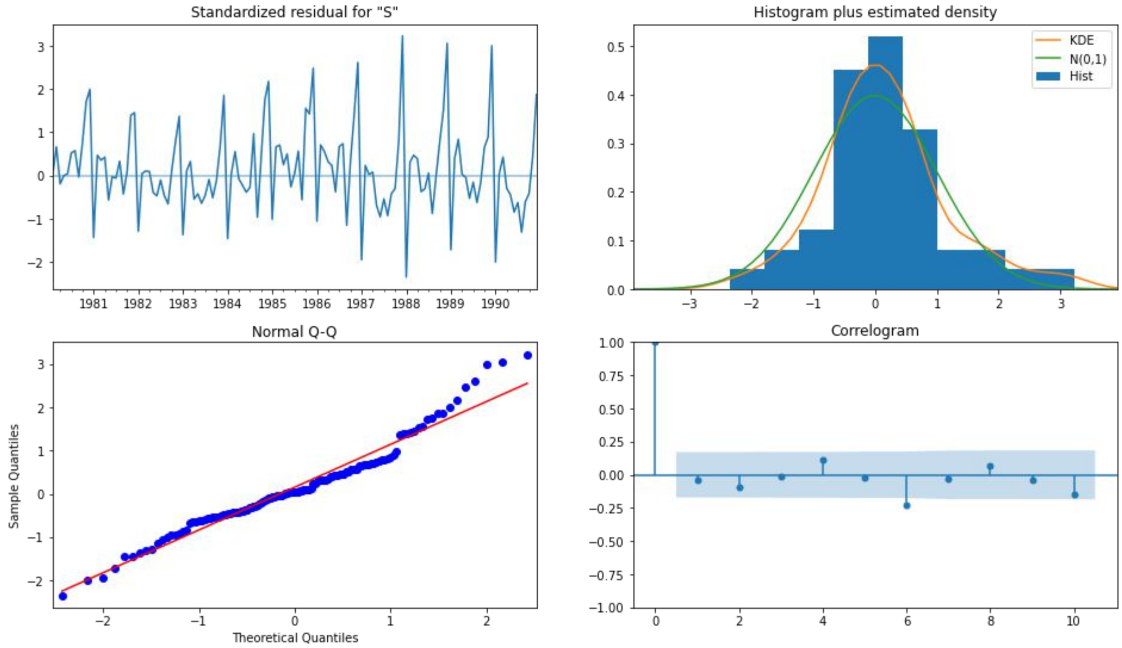


Table 74-Auto ARIMA diagnostic plot

The model evaluation

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(2,1,2)	1299.980353	47.100122

Table 75-RMSE value of auto ARIMA

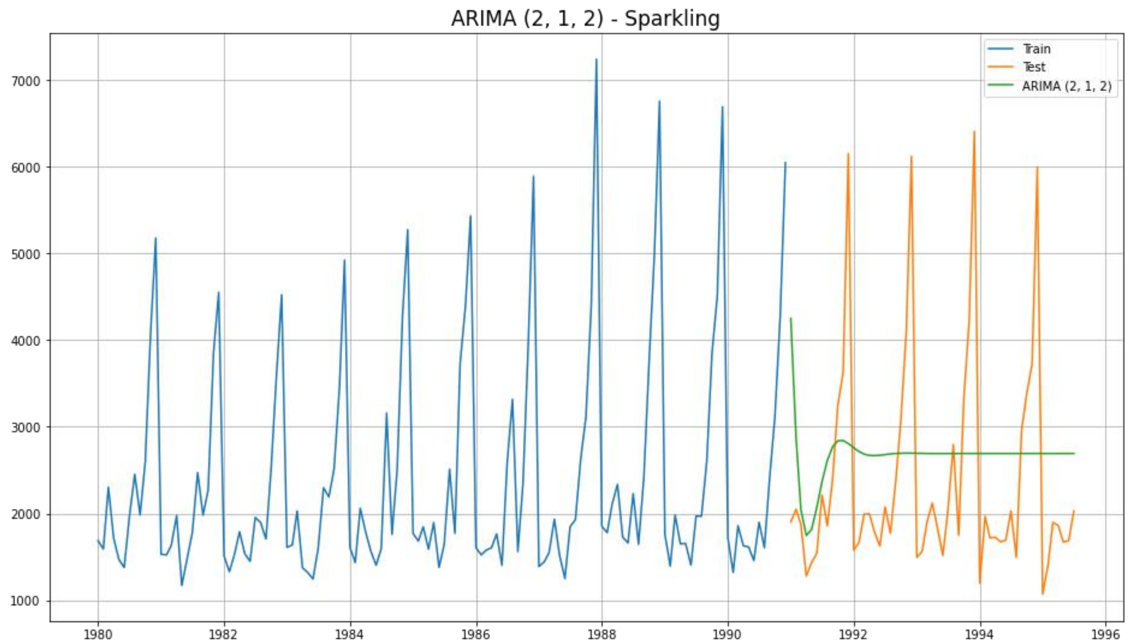


Table 76-Auto ARIMA forecast on sparkling data

Let us look at ACF and PACF plot

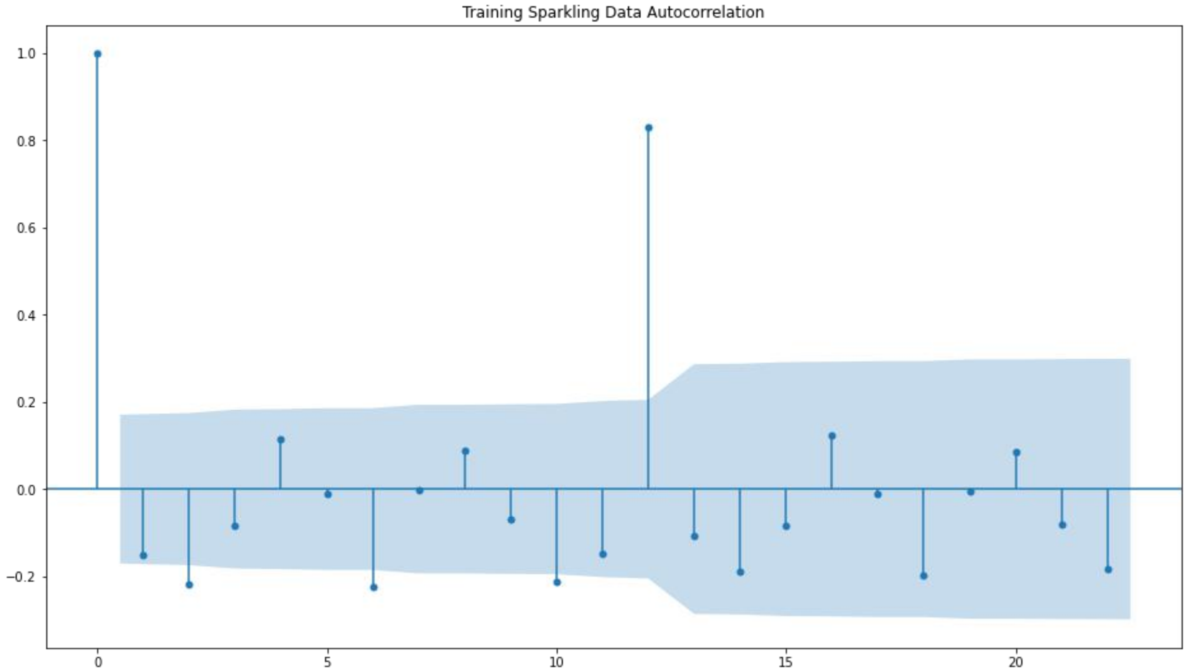


Table 77-ACF plot sparkling train data

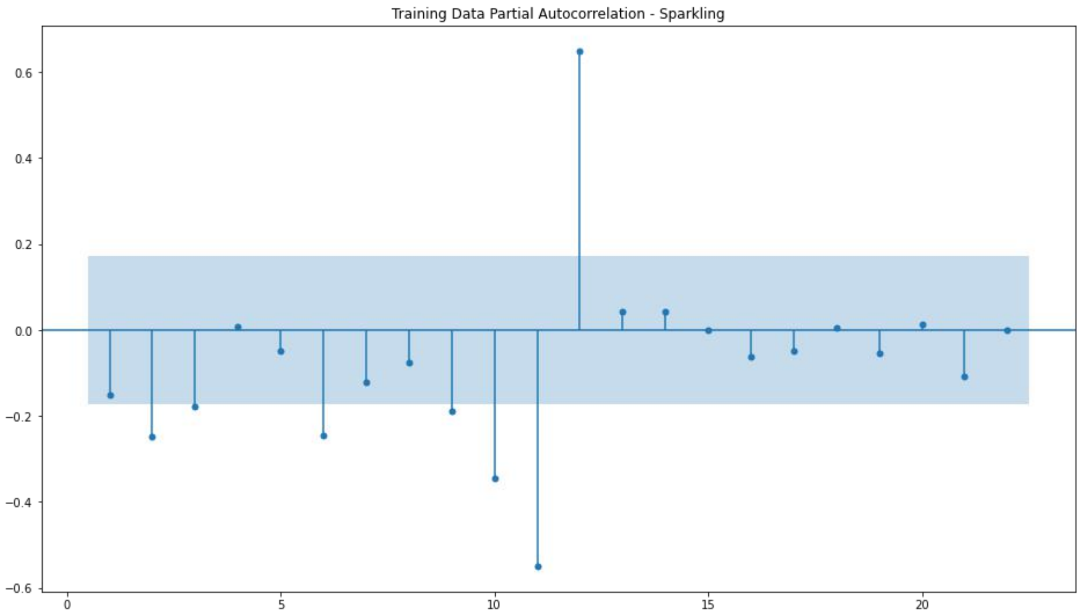


Table 78-PACF plot sparkling train data

Auto SARIMA

```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (0, 1, 3)(0, 0, 3, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (1, 1, 3)(1, 0, 3, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
Model: (2, 1, 3)(2, 0, 3, 6)
Model: (3, 1, 0)(3, 0, 0, 6)
Model: (3, 1, 1)(3, 0, 1, 6)
Model: (3, 1, 2)(3, 0, 2, 6)
Model: (3, 1, 3)(3, 0, 3, 6)
```

We have taken seasonality as 6

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	1629.605578
59	(0, 1, 3)	(2, 0, 3, 6)	1633.327862
123	(1, 1, 3)	(2, 0, 3, 6)	1633.964363
251	(3, 1, 3)	(2, 0, 3, 6)	1634.617343
63	(0, 1, 3)	(3, 0, 3, 6)	1635.054422

Table 79-Sorted table of AIC scores auto SARIMA

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -803.803
Date:                            Sun, 03 Jul 2022   AIC                           1629.606
Time:                                    15:30:38   BIC                           1659.210
Sample:                                01-01-1980   HQIC                          1641.611
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.7138      0.074    -23.298      0.000      -1.858      -1.570
ar.L2         -0.7799      0.081     -9.574      0.000      -0.939      -0.620
ma.L1          1.0702      0.547      1.957      0.050      -0.002       2.142
ma.L2         -0.7236      0.153     -4.734      0.000      -1.023      -0.424
ma.L3         -0.8829      0.463     -1.906      0.057      -1.791       0.025
ar.S.L6       -0.0063      0.027     -0.232      0.817      -0.060       0.047
ar.S.L12       1.0419      0.020     51.758      0.000       1.002       1.081
ma.S.L6       -0.6954      0.527     -1.319      0.187      -1.728       0.338
ma.S.L12      -1.2092      0.330     -3.668      0.000      -1.855      -0.563
ma.S.L18       0.1445      0.362      0.399      0.690      -0.566       0.855
sigma2      6.252e+04    1.5e-05   4.16e+09      0.000    6.25e+04    6.25e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 7.89
Prob(Q):                              0.75   Prob(JB):                         0.02
Heteroskedasticity (H):               1.39   Skew:                             0.25
Prob(H) (two-sided):                  0.32   Kurtosis:                         4.22
===================================================================================
```

Table 80-Auto SARIMA result summary

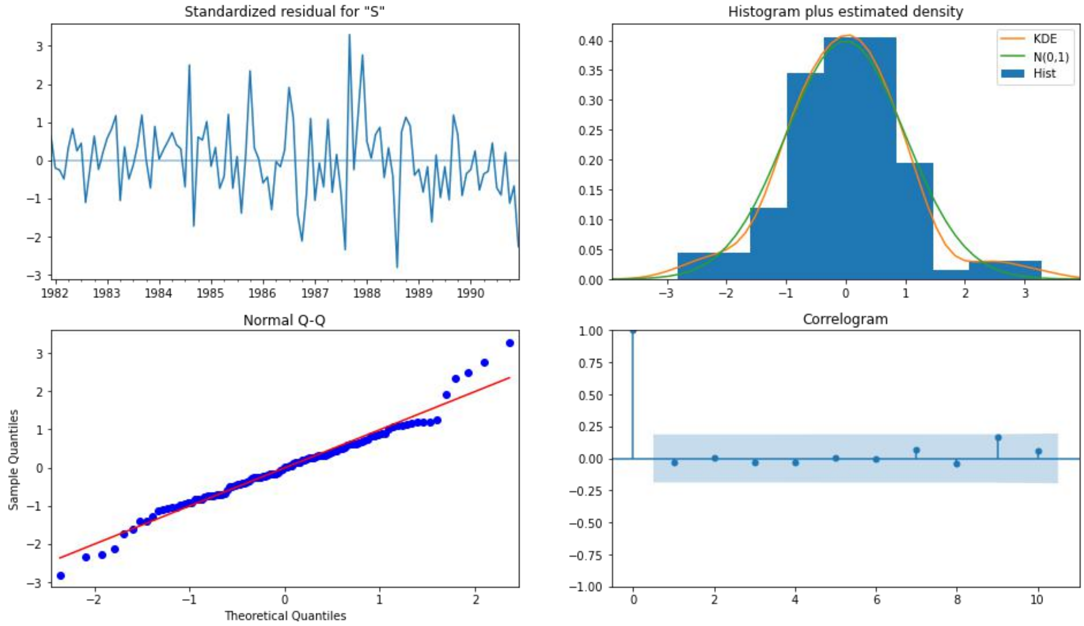


Table 81-Auto SARIMA diagonostic plot

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(2,1,2)	1299.980353	47.100122
SARIMA(2, 1, 3)(2, 0, 3, 6)	772.439807	34.531975

Table 82-RMSE and MAPE value for auto SARIMA on sparkling data

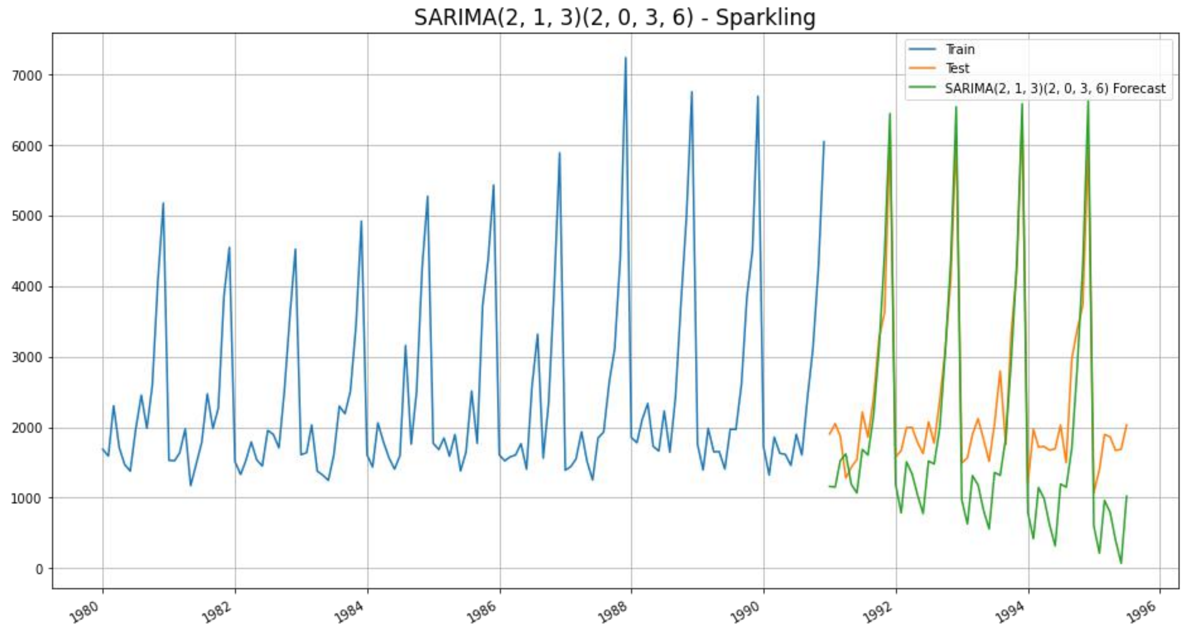


Table 83-Auto SARIMA(2,1,3)(2,0,3,6)

Let us look at ACF and PACF plot again

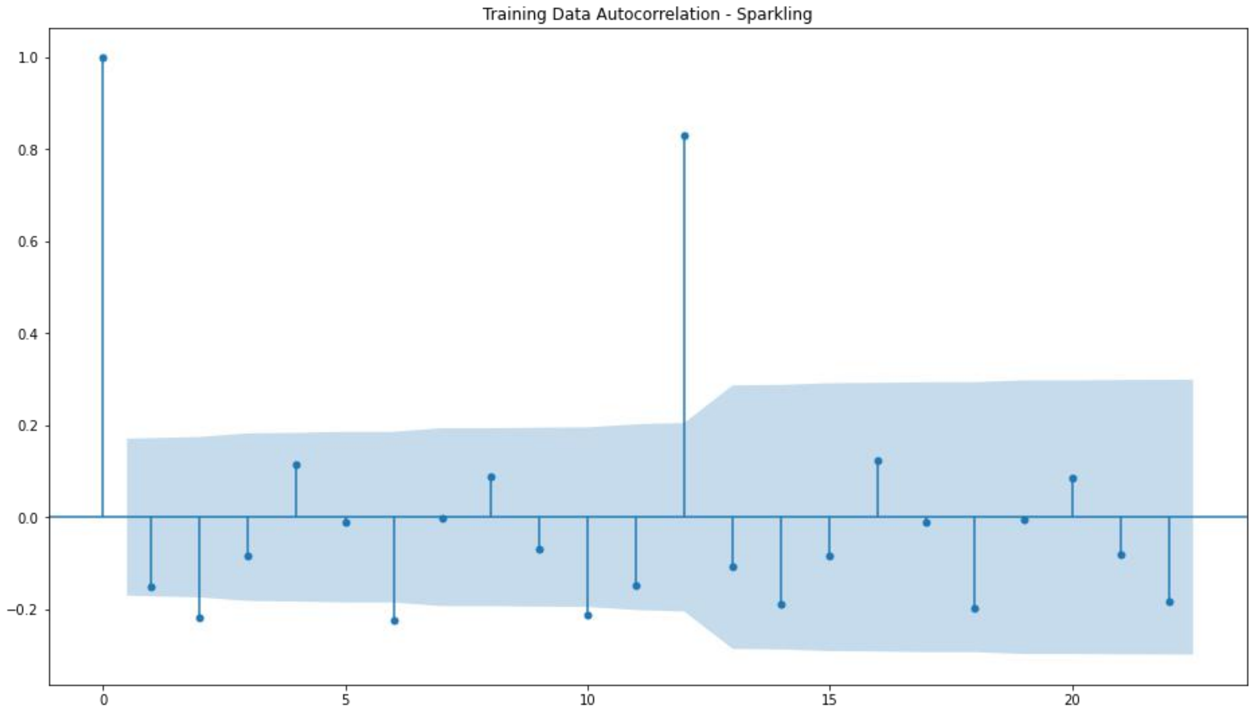


Table 84-Sparkling train data ACF plot

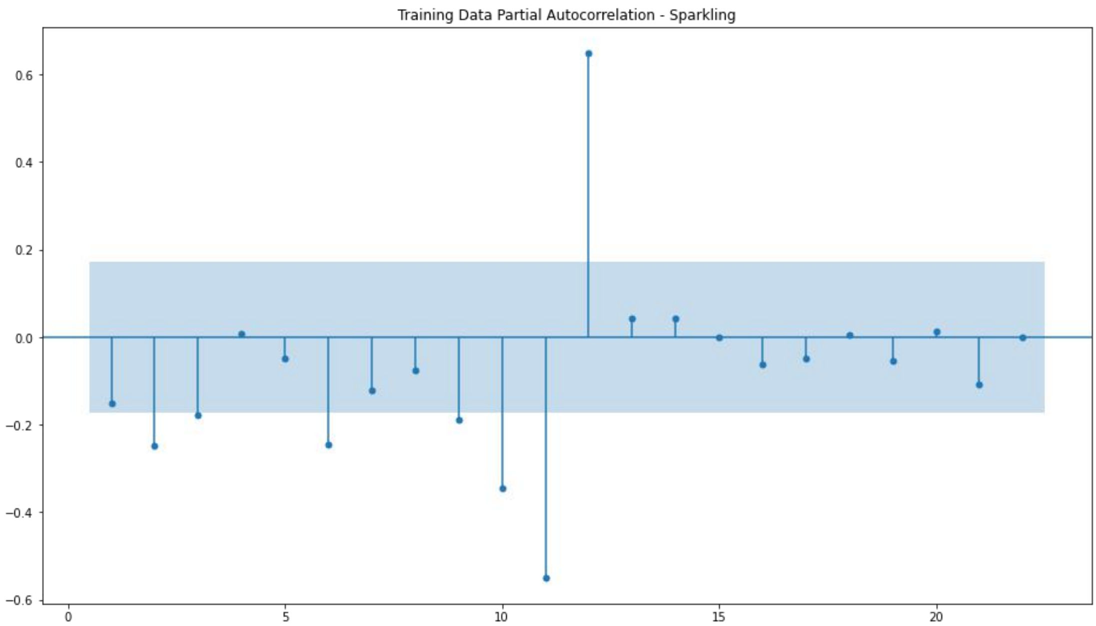


Table 85-Sparkling train data PACF plot

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12

We are taking the p value to be 0 and the q value also to be 0 as the parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                             Sparkling   No. Observations:              132
Model:             SARIMAX(0, 1, 0)x(1, 1, [1], 12)   Log Likelihood             -811.162
Date:                               Sun, 03 Jul 2022   AIC                        1628.324
Time:                                       15:30:40   BIC                        1636.315
Sample:                                   01-01-1980   HQIC                       1631.563
                                        - 12-01-1990
Covariance Type:                                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12       0.1482      0.223      0.664      0.507      -0.289       0.586
ma.S.L12      -0.5732      0.217     -2.640      0.008      -0.999      -0.148
sigma2      2.577e+05   2.63e+04      9.806      0.000    2.06e+05    3.09e+05
===================================================================================
Ljung-Box (L1) (Q):                  13.54   Jarque-Bera (JB):                27.17
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.73   Skew:                             0.59
Prob(H) (two-sided):                  0.36   Kurtosis:                         5.19
===================================================================================
```

Table 86-Mannual SARIMA result summary

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 3.

We have built another mannual SARIMA model with parameters (0,1,0)(2,1,2,12)

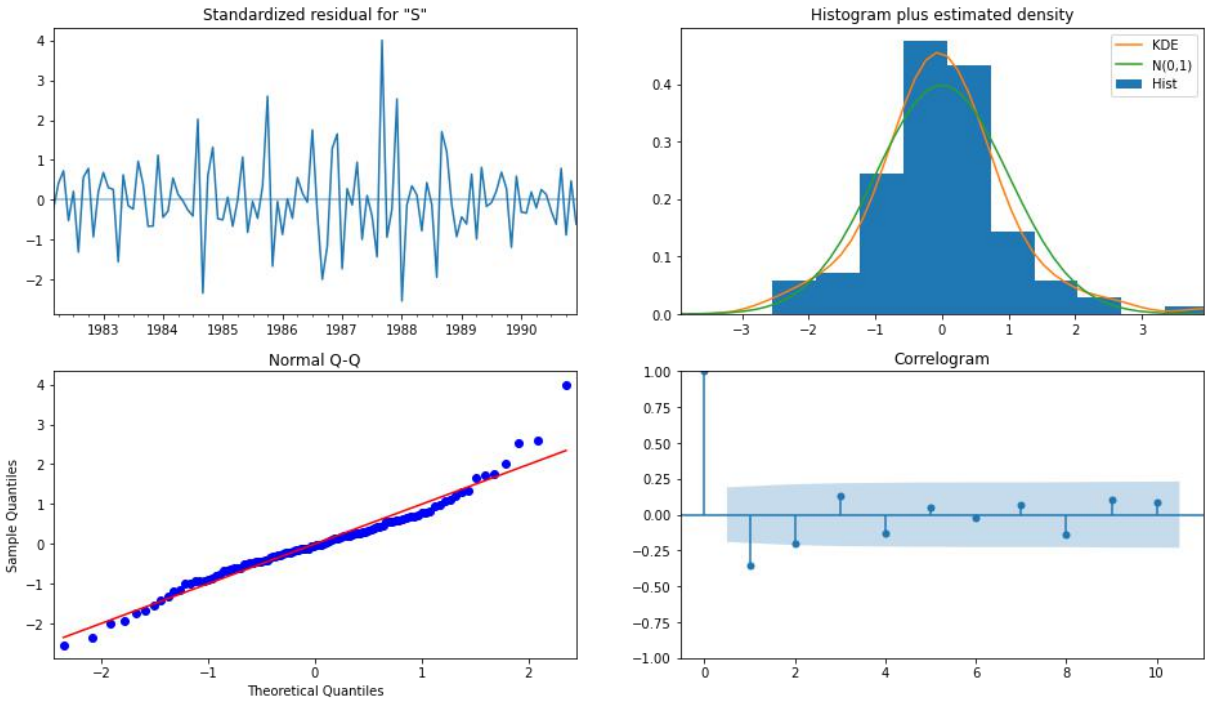


Table 87-mannual SARIMA diagnostic plot

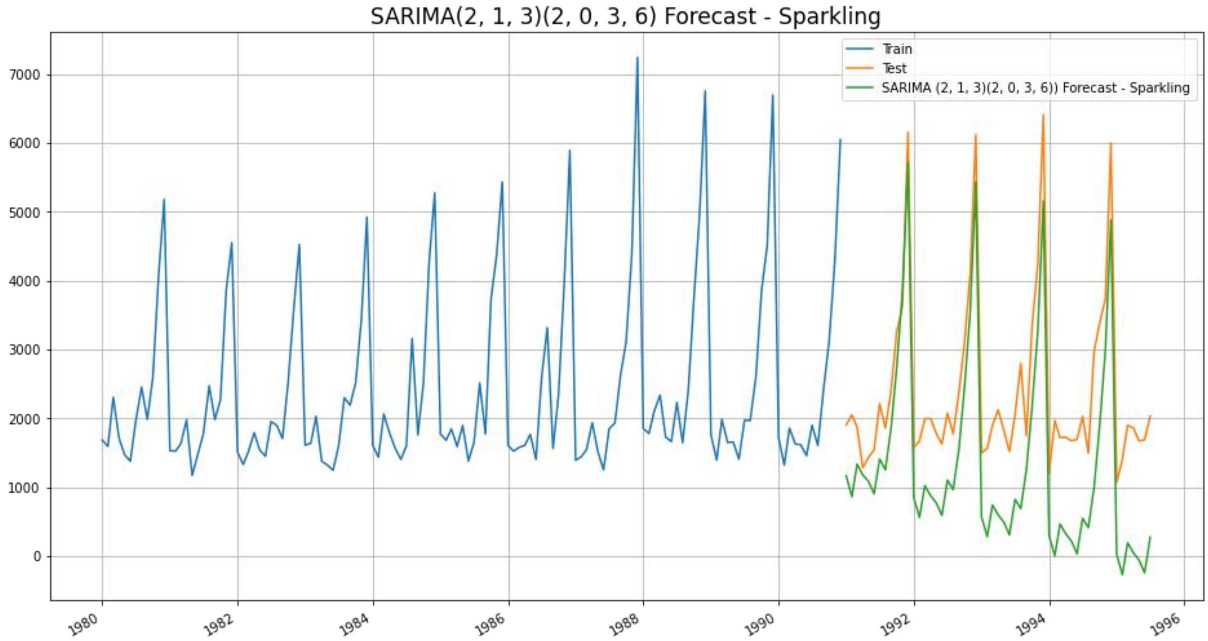


Table 88-Mannual SARIMA 1st model forecast on sparkling data

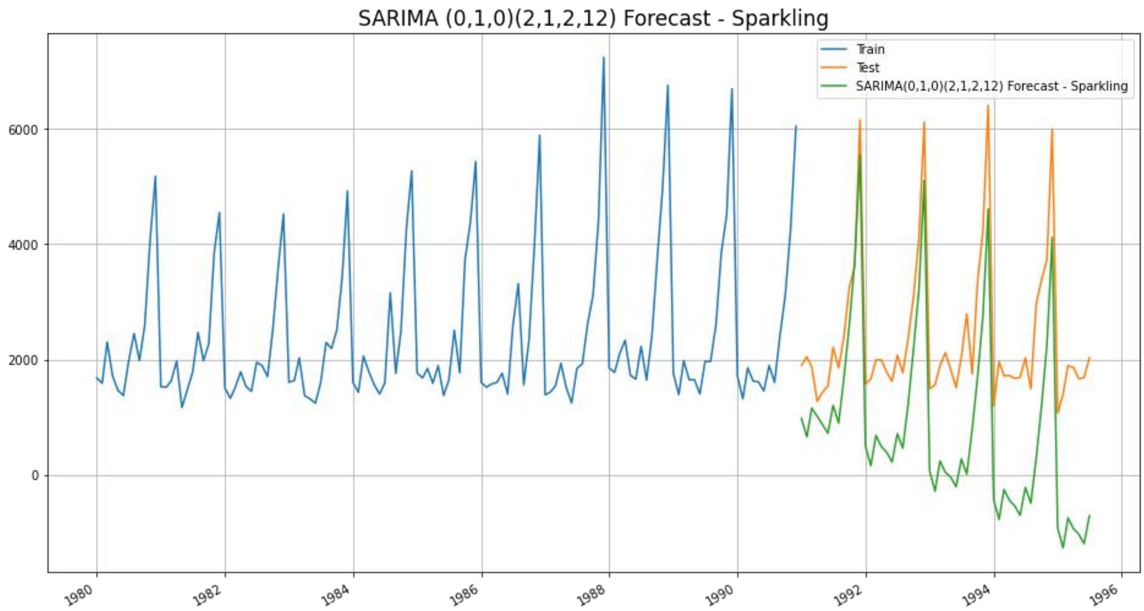


Table 89-Mannual SARIMA 2nd model forecast on sparkling data

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(2,1,2)	1299.980353	47.100122
SARIMA(2, 1, 3)(2, 0, 3, 6)	772.439807	34.531975
SARIMA(0,1,0)(1,1,1,12)	1757.726403	81.785213

Table 90-RMSE score of mannual SARIMA models on sparkling data

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.

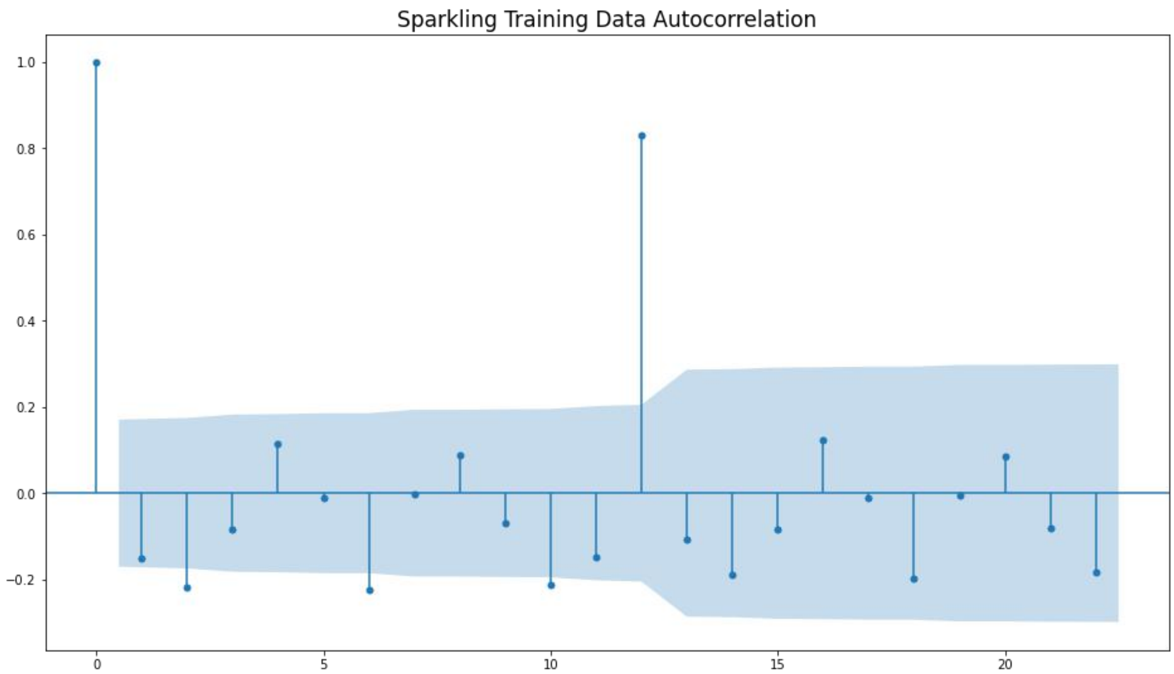


Table 91-Sparkling train data ACF plot

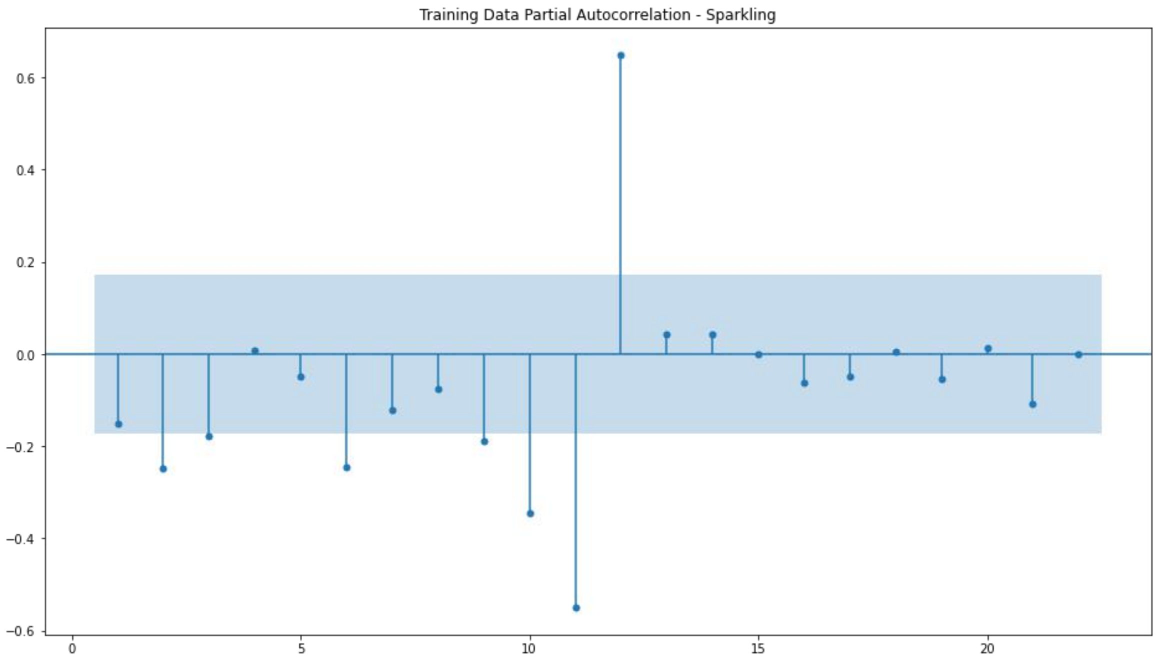


Table 92-Sparkling train data PACF plot

Here, we have taken alpha=0.05.

The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off

The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off

By looking at the above plots, we will take the value of p and q to be 0 and 0 respectively.¶

Mannual ARIMA result summary

```
                               SARIMAX Results                                
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.832
Date:                Sun, 03 Jul 2022   AIC                           2267.663
Time:                        15:30:44   BIC                           2270.538
Sample:                    01-01-1980   HQIC                          2268.831
                         - 12-01-1990                                         
Covariance Type:                  opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      1.885e+06   1.29e+05     14.658      0.000    1.63e+06    2.14e+06
===================================================================================
Ljung-Box (L1) (Q):                   3.07   Jarque-Bera (JB):               198.83
Prob(Q):                              0.08   Prob(JB):                         0.00
Heteroskedasticity (H):               2.46   Skew:                            -1.92
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.65
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Table 93-Mannual ARIMA result summary sparkling data

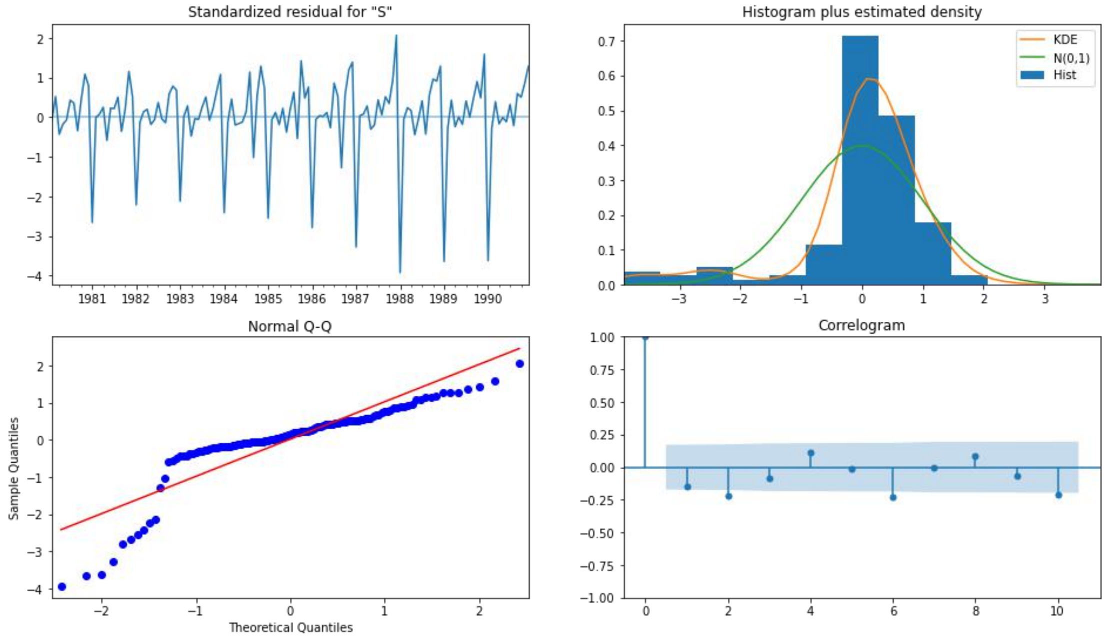


Table 94-Mannual ARIMA sparkling data diagnostic plot

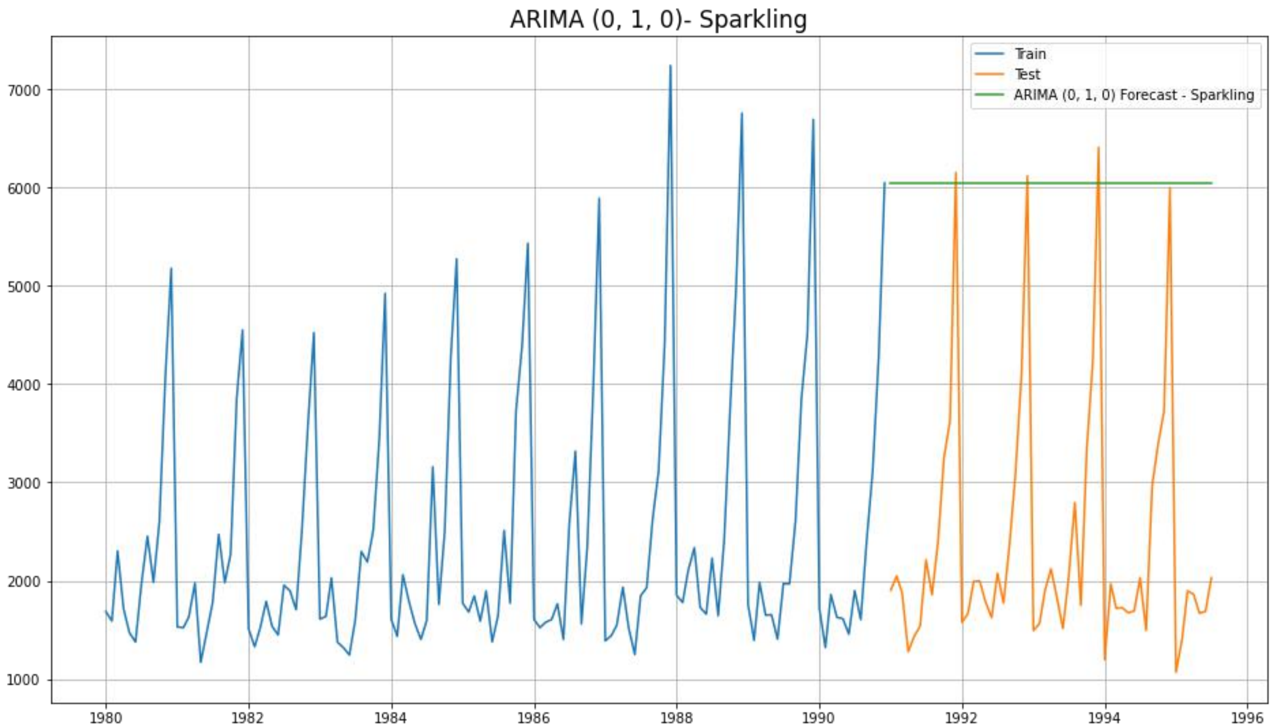


Table 95-mannual ARIMA forecast sparkling data

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.

	Test RMSE Sparkling	Test MAPE Sparkling
RegressionOnTime	1389.135175	NaN
NaiveModel	3864.279352	NaN
SimpleAverageModel	1275.081804	NaN
2pointTrailingMovingAverage	813.400684	NaN
4pointTrailingMovingAverage	1156.589694	NaN
6pointTrailingMovingAverage	1283.927428	NaN
9pointTrailingMovingAverage	1346.278315	NaN
Simple Exponential Smoothing	1338.000861	NaN
Double Exponential Smoothing	5291.879833	NaN
Triple Exponential Smoothing (Additive Season)	378.625883	NaN
Triple Exponential Smoothing (Multiplicative Season)	2647.181960	NaN
ARIMA(2,1,2)	1299.980353	47.100122
SARIMA(2, 1, 3)(2, 0, 3, 6)	772.439807	34.531975
SARIMA(0,1,0)(1,1,1,12)	1757.726403	81.785213
ARIMA(0,1,0)	3864.279352	201.327650

Table 96-Sparkling data All model RMSE score

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.

So we have selected TES additive model as per RMSE values to forecast, below is the plot.

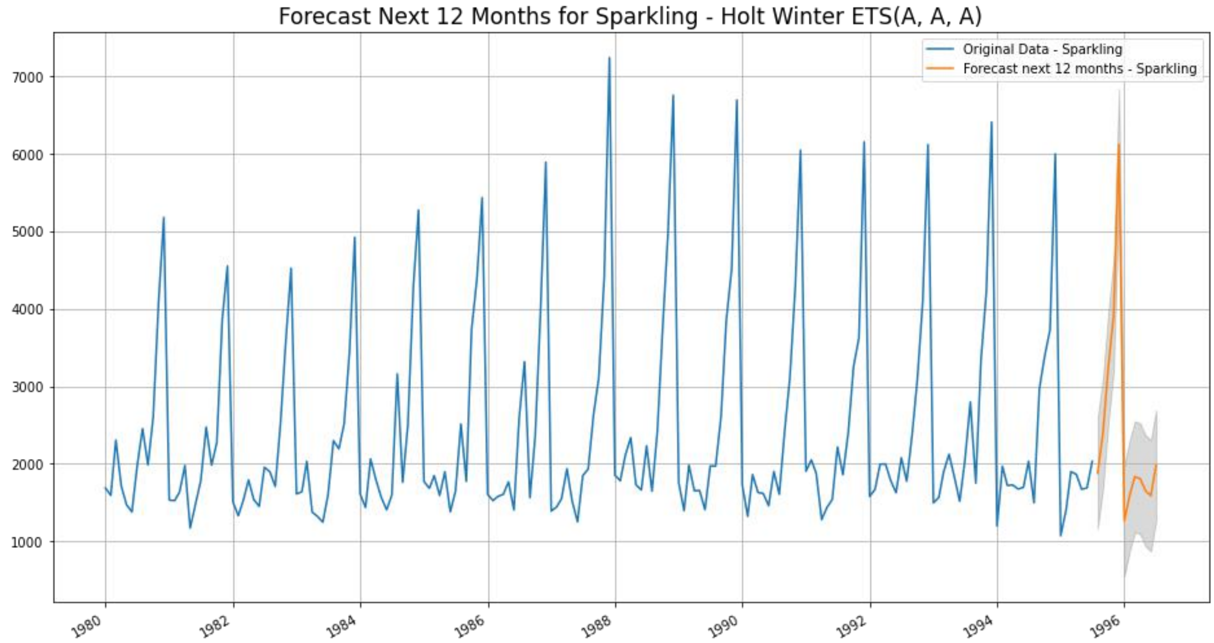


Table 97-Sparkling data forecast on best model

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

The best model is Tripple exponential smoothing with additive model.

We have plot the forecast for next one year and it looks like that there will be a repeat of the seasonality in future and current trend will continue.

But to forecast more accurately we need more data.

The trend is more or less stable so demand is consistent for this product, company can increase price to generate more revenue.

The production team has to sit and devise a plan to meet the requirement by timely procurement of raw material and logistics planning.