TIME SERIES FORECASTING BUSINESS REPORT – ROSE WINE SALES

TEJAS PADEKAR
PGP-DSBA Online
FEB' 22
Date: 20/02/2022

CONTENTS:

List of Figures

List of Tables

PROBLEM 2

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines (Sparkling & Rose). As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Rose Wine Sales in the 20th century.

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

Dataset Head:

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

Dataset Tail:

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

The dataset contents 187 observations across 02 columns in total.

Date Time Index:

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

Time Stamp:

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

We do not require the column of "YearMonth" as we have created a Time Stamp for the same and made it as our index column as well. Hence, we have dropped "YearMonth" from our dataset.

Missing Values:

```
YearMonth    0
Rose         2
dtype: int64
```

	YearMonth	Rose
174	1994-07	NaN
175	1994-08	NaN

We observe that there a 2 months of July and August 1994 where the Rose Wine sales figures are missing.

Dataset Description:

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

We observe from a historical record of 187 months since Jan 1980 until July 1995 that average sales over the period of Rose wine was 90 bottles. The least being 28 bottles and highest being 267 bottles. The data however is missing the sales for the 2 months of July and August 1994 which we will fill accurately through interpolate function later. Now, we have our data ready for the Time Series Analysis.

Time Series Plot

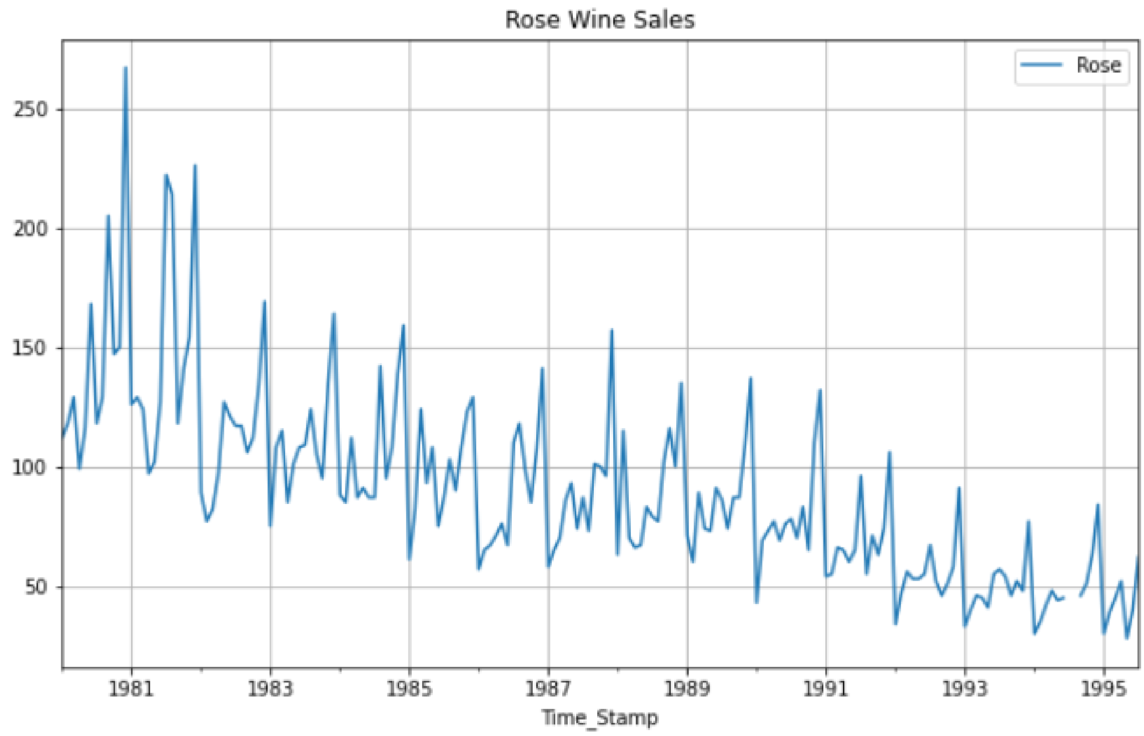


Figure 1: Time Series Plot

For above figure, we observe presence of trend and seasonality throughout the time series. The trend is downward indicating that the sales figures of the Rose Wine have been dropping over the years. We can also observe a slight cut between the plot line between the years 1994 and 1995 due to the missing values.

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

Yearly Boxplot:

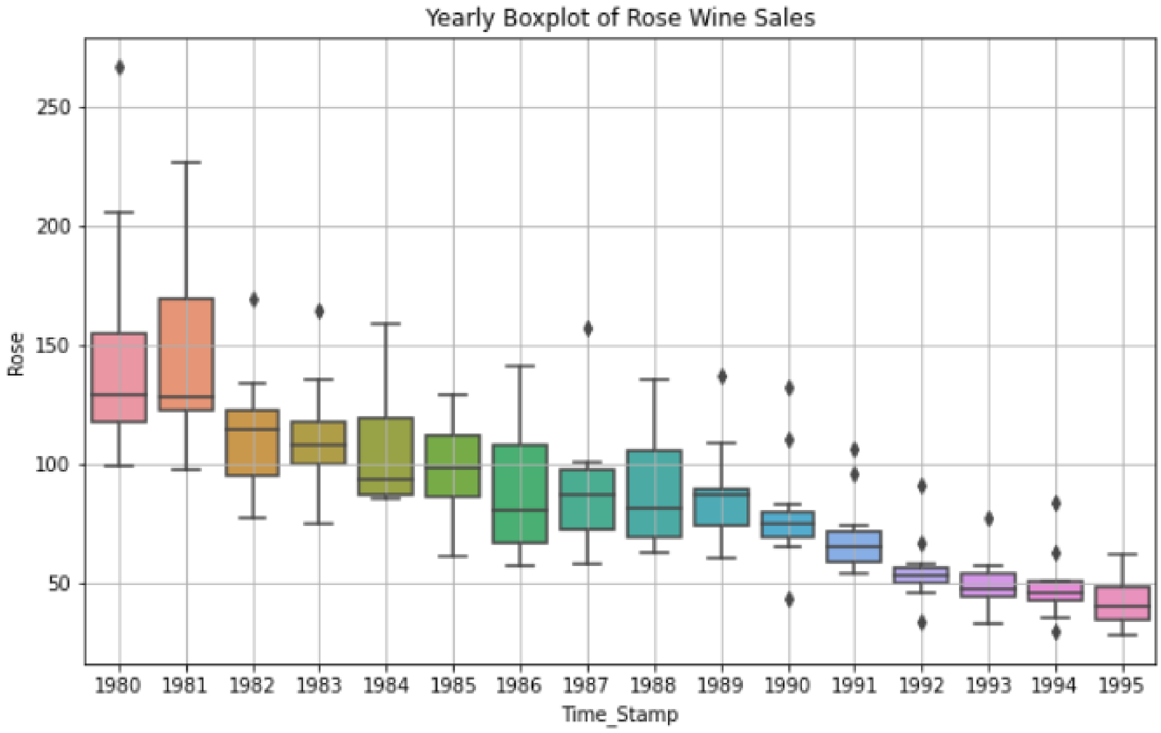


Figure 2 – Yearly Boxplot of Rose Wine Sales

The sales figures can be seen to pretty much ranges between 25 bottles to 175 bottles from 1982 till 1995 with outlier present in some years. The sales figures were higher in 1980 and 1981 after which there has been a significant drop in the sales figures. The sales have been on a continuous drop in most years dropping to the lowest in 1995 which can be clearly seen from the downward trend.

Monthly Boxplot:

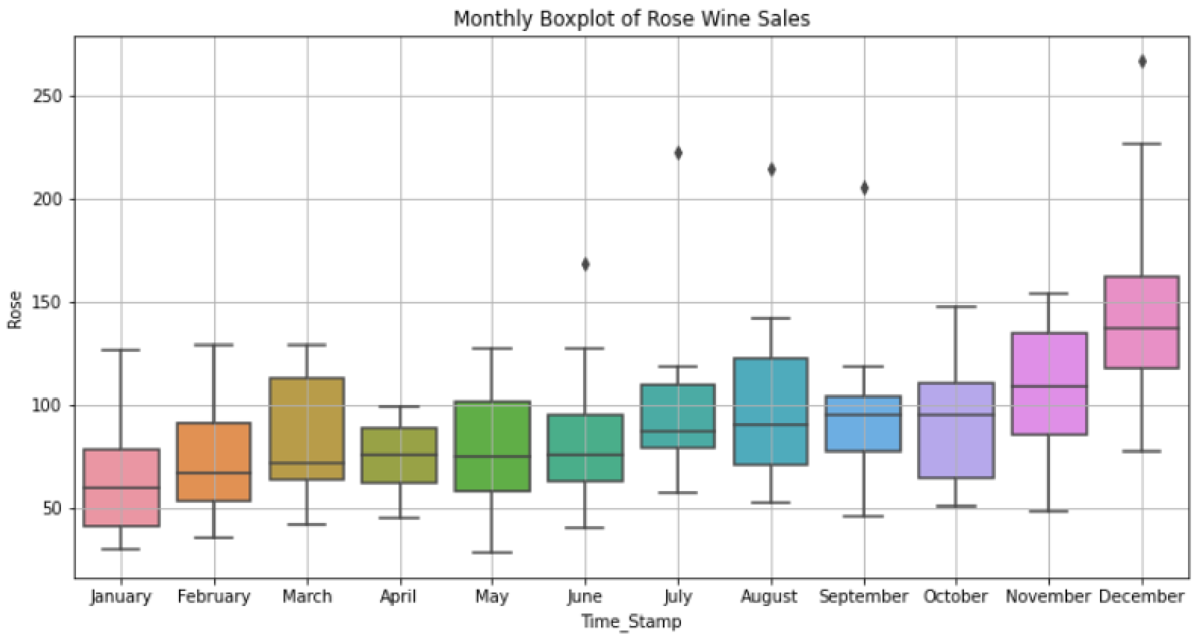


Figure 3 – Monthly Boxplot of Rose Wine Sales

The sales have been better in the 2nd half of the year throughout the time series especially from July onwards. The least sales have come in the month of April and highest from December throughout various years!

Outliers are present for June, July, August, September and December months.

Monthly Boxplot with Median Values as Red Line:

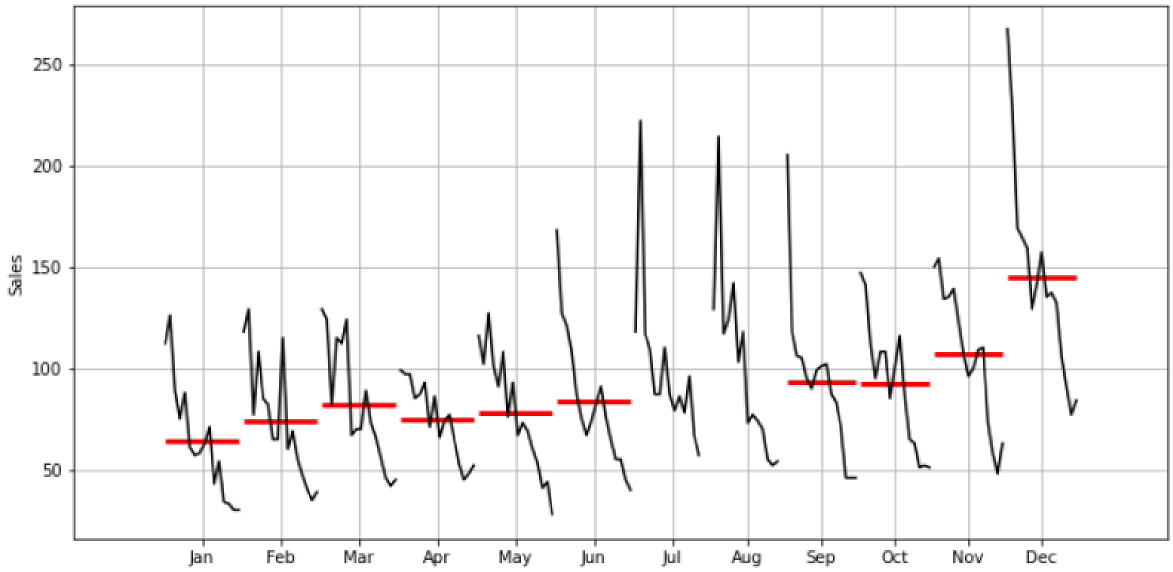


Figure 4 – Monthly Boxplot with Median Values as Red Line

The median values are within 50 and 100 bottles until October and 100 to 150 bottles for November and December. There is actually a huge spike in the median value for December compared to November . Also, we can see that the median values for July and August are not seen as there are missing values for these months in the time series dataset.

Pivot Table Monthly Sales:

Time_Stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_Stamp												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	NaN	NaN	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

Figure 5 – Pivot Table of Rose Wine Sales

Monthly Sales across Years:

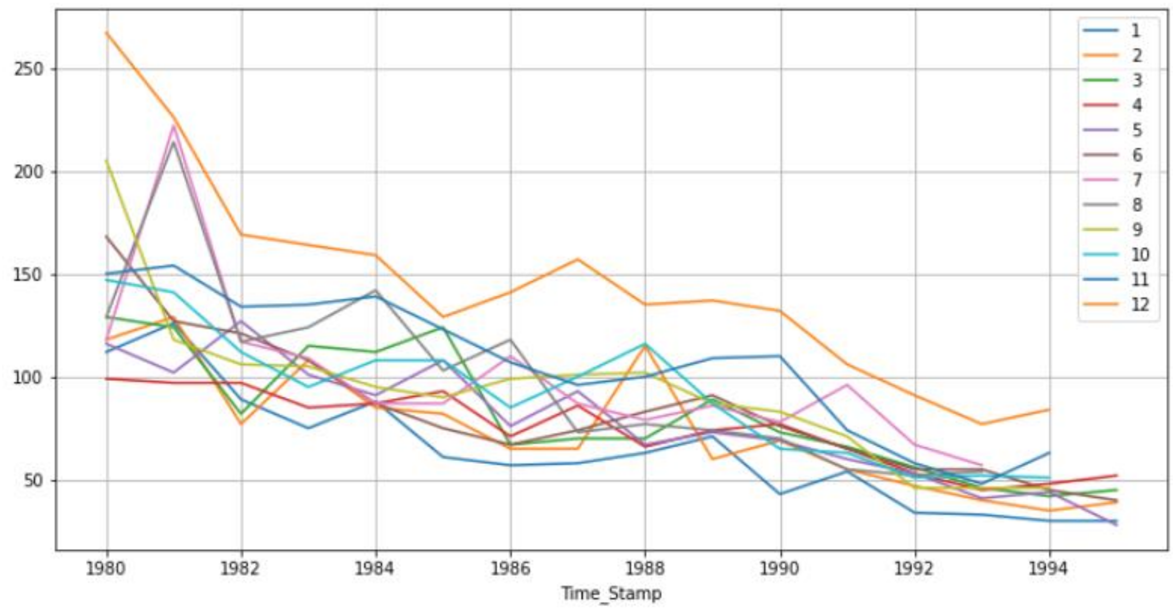


Figure 6 –Rose Wine Monthly Sales across Years

December has the highest sales across all years

Empirical Cumulative Distribution:

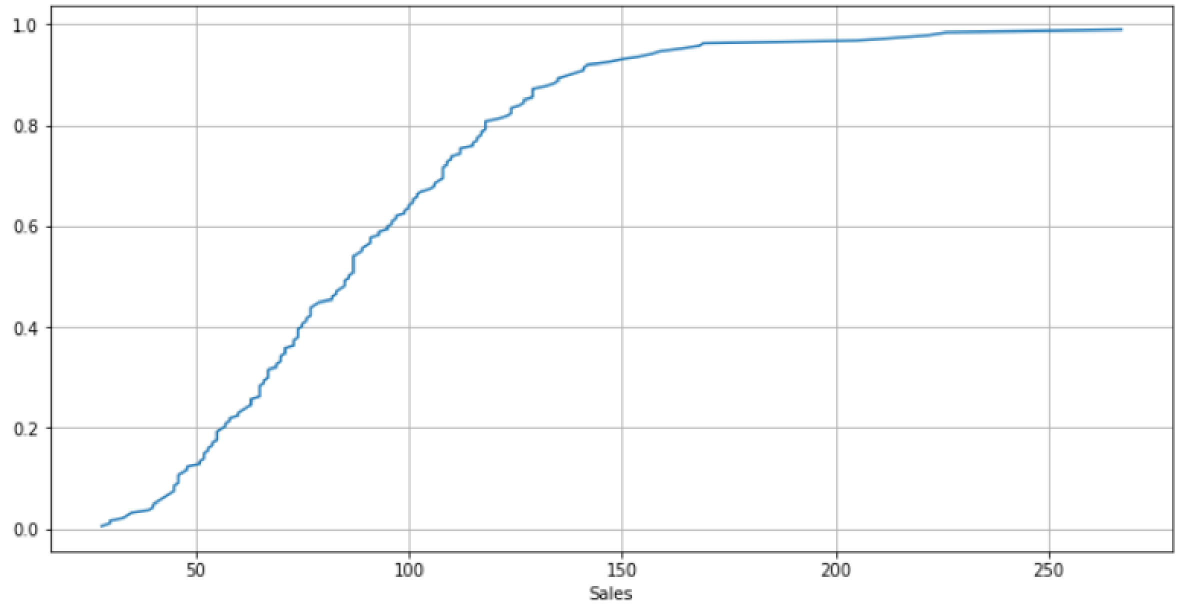


Figure 7 –Rose Wine Sales Empirical Cumulative Distribution Plot

Average Sales and Percentage Change of Sales:

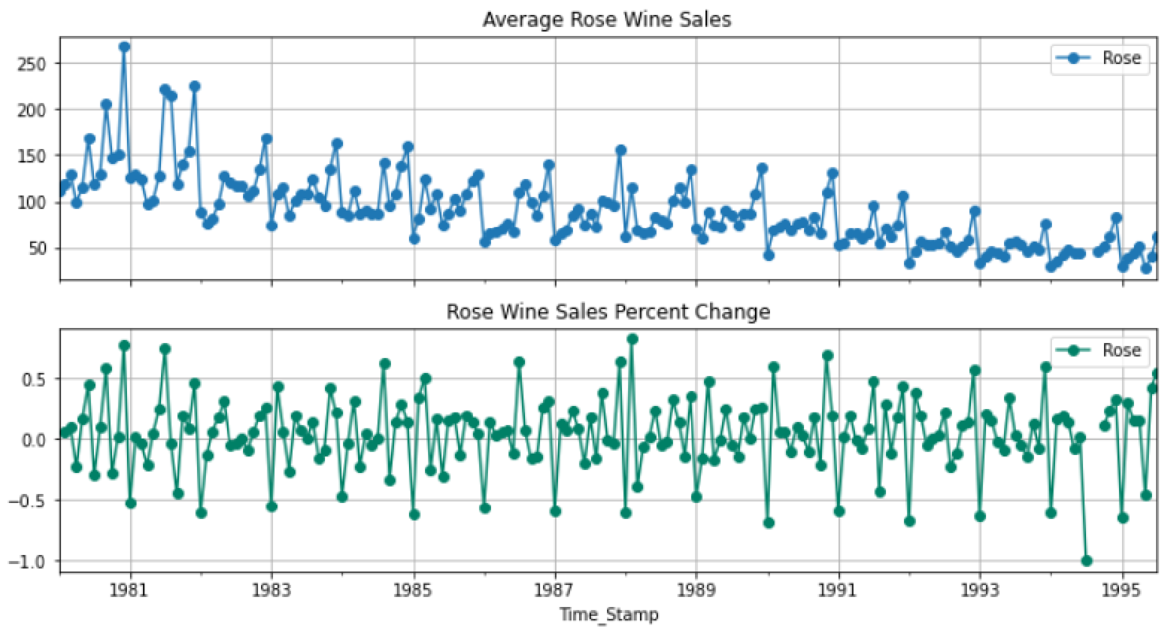


Figure 8 –Average Sales and Percentage Change of Sales

Sum of Sales of each year in Plot:

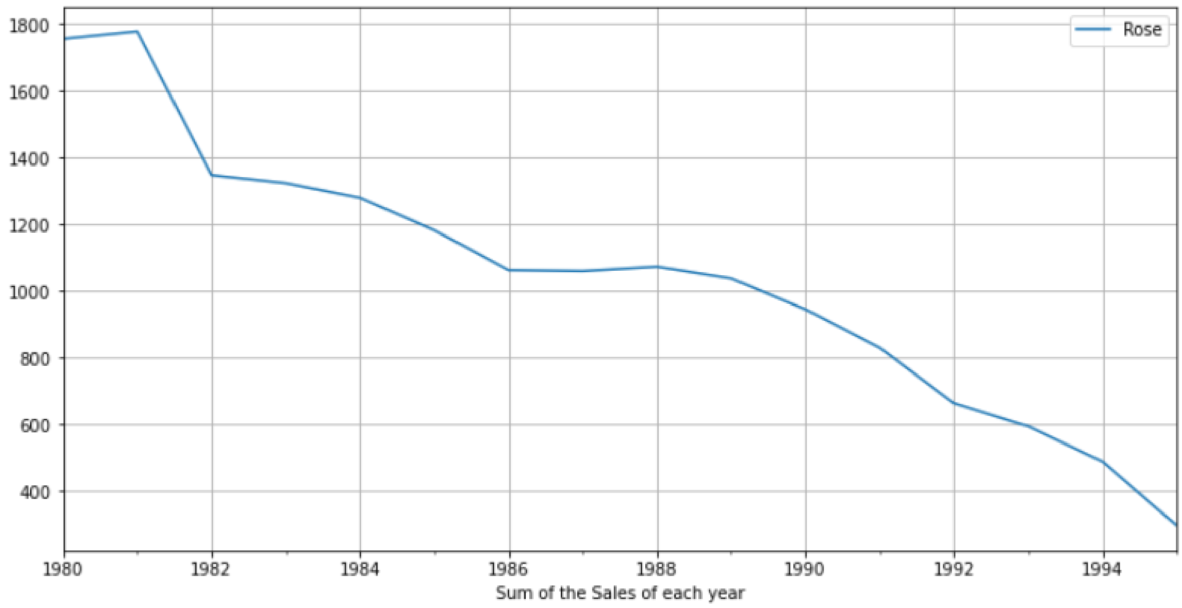


Figure 9 –Sum of Rose Wine Sales for each Year

The sales have dropped over the years, only exception being 1981 where they have slightly increased.

Sum of Sales of each year in Figures:

	Rose
Time_Stamp	
1980-12-31	1758.0
1981-12-31	1780.0
1982-12-31	1348.0
1983-12-31	1324.0
1984-12-31	1280.0
1985-12-31	1183.0
1986-12-31	1063.0
1987-12-31	1060.0
1988-12-31	1073.0
1989-12-31	1038.0
1990-12-31	945.0
1991-12-31	830.0
1992-12-31	663.0
1993-12-31	594.0
1994-12-31	488.0
1995-12-31	296.0

The highest sales were recorded in the year 1981 with total 1780 bottles sold and the least sales were for 1995 with total 296 bottles sold within the first seven months. On a 12 monthly basis, the least sales can be seen for the year 1994 with 488 bottles sold in total.

Mean of Sales of each year in Plot:

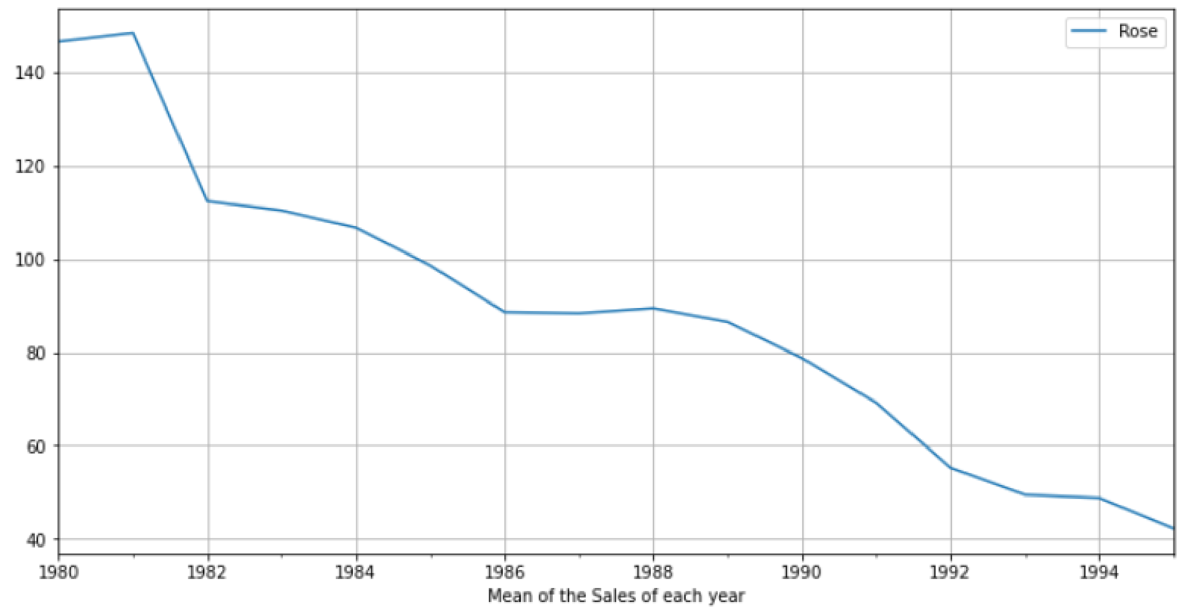


Figure 10 –Mean of Sparkling Wine Sales for each Year

Mean of Sales of each year in Figures:

	Rose
Time_Stamp	
1980-12-31	146.500000
1981-12-31	148.333333
1982-12-31	112.333333
1983-12-31	110.333333
1984-12-31	106.666667
1985-12-31	98.583333
1986-12-31	88.583333
1987-12-31	88.333333
1988-12-31	89.416667
1989-12-31	86.500000
1990-12-31	78.750000
1991-12-31	69.166667
1992-12-31	55.250000
1993-12-31	49.500000
1994-12-31	48.800000
1995-12-31	42.285714

The mean sales ranged between 42 and 146.5 from 1980 till 1995. With lowest in 1995 and highest in 1981 similar to earlier observation for sum of monthly sales.

Interpolate Missing Values:

Time_Stamp	Rose
1994-01-31	30.0
1994-02-28	35.0
1994-03-31	42.0
1994-04-30	48.0
1994-05-31	44.0
1994-06-30	45.0
1994-07-31	45.0
1994-08-31	45.0
1994-09-30	46.0
1994-10-31	51.0
1994-11-30	63.0
1994-12-31	84.0

We have interpolated the missing values for July and August 1994 as the same sales figures for June of 45 bottles. Now, that we have interpolated the missing values, we can now decompose the time series.

Decomposing the Time Series: Additive Method

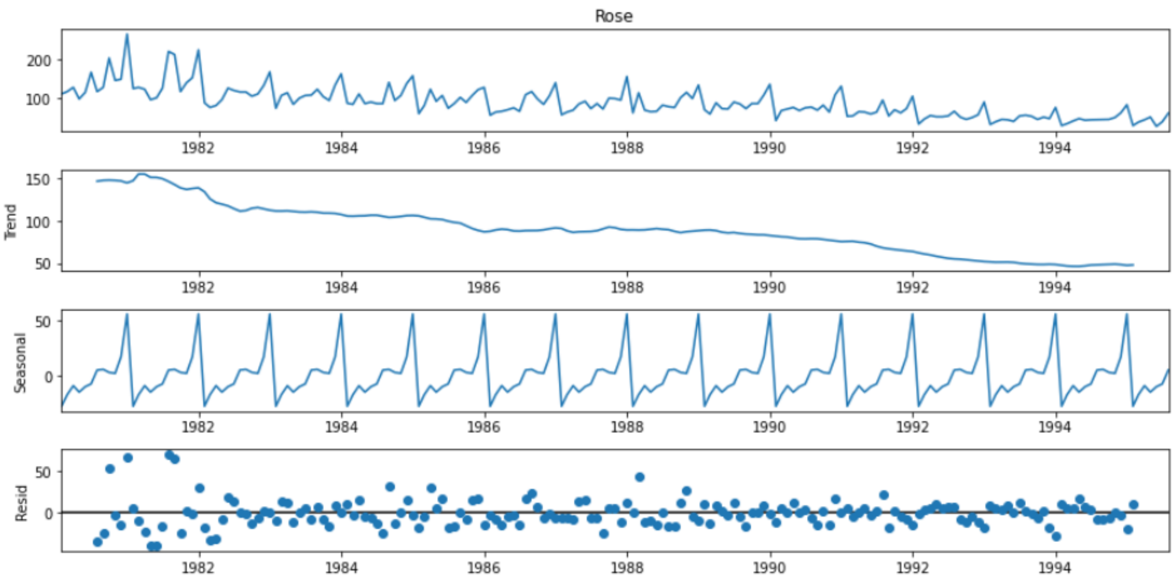


Figure 11 – Additive Decomposition

As per the 'additive' decomposition, we see that there is a pronounced trend until 1991. There is a seasonality as well. A lot of residuals are located around 0 from the plot of the residuals in the decomposition.

Trend, Seasonality and Residual:

```
Trend
 Time_Stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
1980-11-30    147.416667
1980-12-31    145.125000
Name: trend, dtype: float64
```

```
Seasonality
 Time_Stamp
1980-01-31   -27.908647
1980-02-29   -17.435632
1980-03-31    -9.285830
1980-04-30   -15.098330
1980-05-31   -10.196544
1980-06-30    -7.678687
1980-07-31     4.896908
1980-08-31     5.499686
1980-09-30     2.774686
1980-10-31     1.871908
1980-11-30    16.846908
1980-12-31    55.713575
Name: seasonal, dtype: float64
```

```
Residual
 Time_Stamp
1980-01-31          NaN
1980-02-29          NaN
1980-03-31          NaN
1980-04-30          NaN
1980-05-31          NaN
1980-06-30          NaN
1980-07-31   -33.980241
1980-08-31   -24.624686
1980-09-30    53.850314
1980-10-31    -2.955241
1980-11-30   -14.263575
1980-12-31    66.161425
Name: resid, dtype: float64
```

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

```
Training Data Shape  (132, 1)
Testing Data Shape  (55, 1)
```

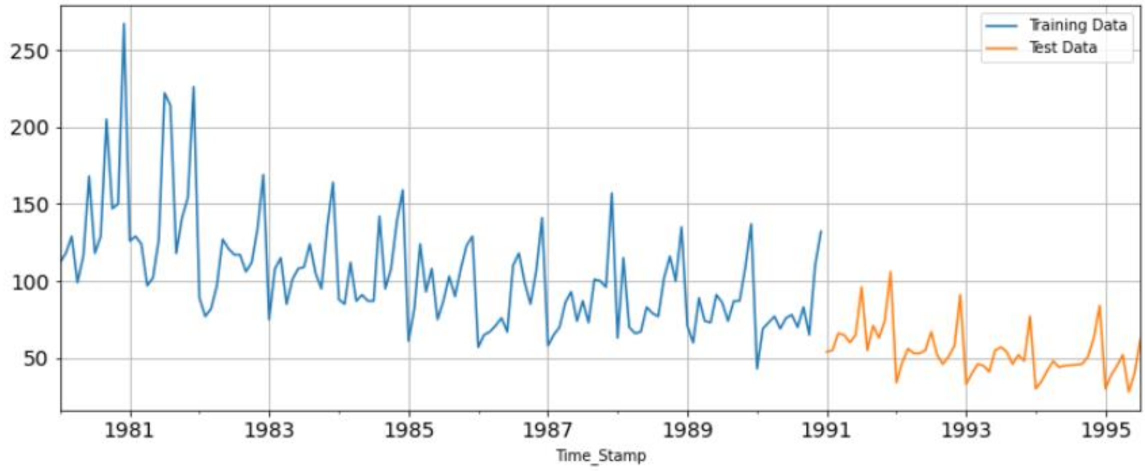


Figure 12 – Train Test Dataset

Training and Test Time Instances:

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

We see that we have successfully the generated the numerical time instance order for both the training and test set.

2.3 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.

MODEL 1 – Linear Regression Model:

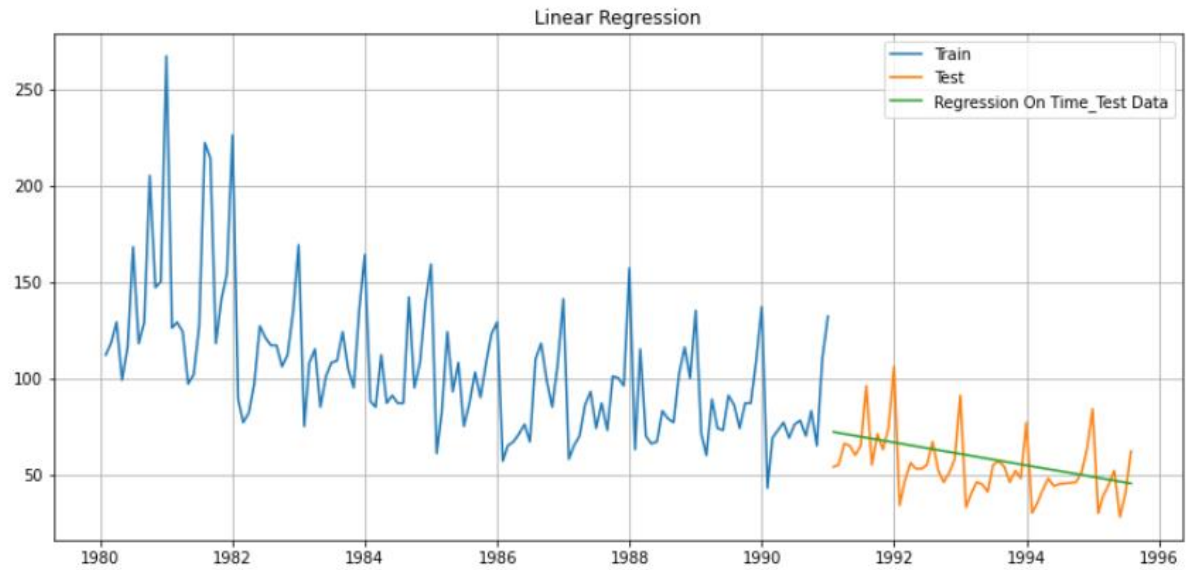


Figure 13 – Linear Regression Plot

	Test RMSE
RegressionOnTime	15.268955

MODEL 2 – Naïve Model:

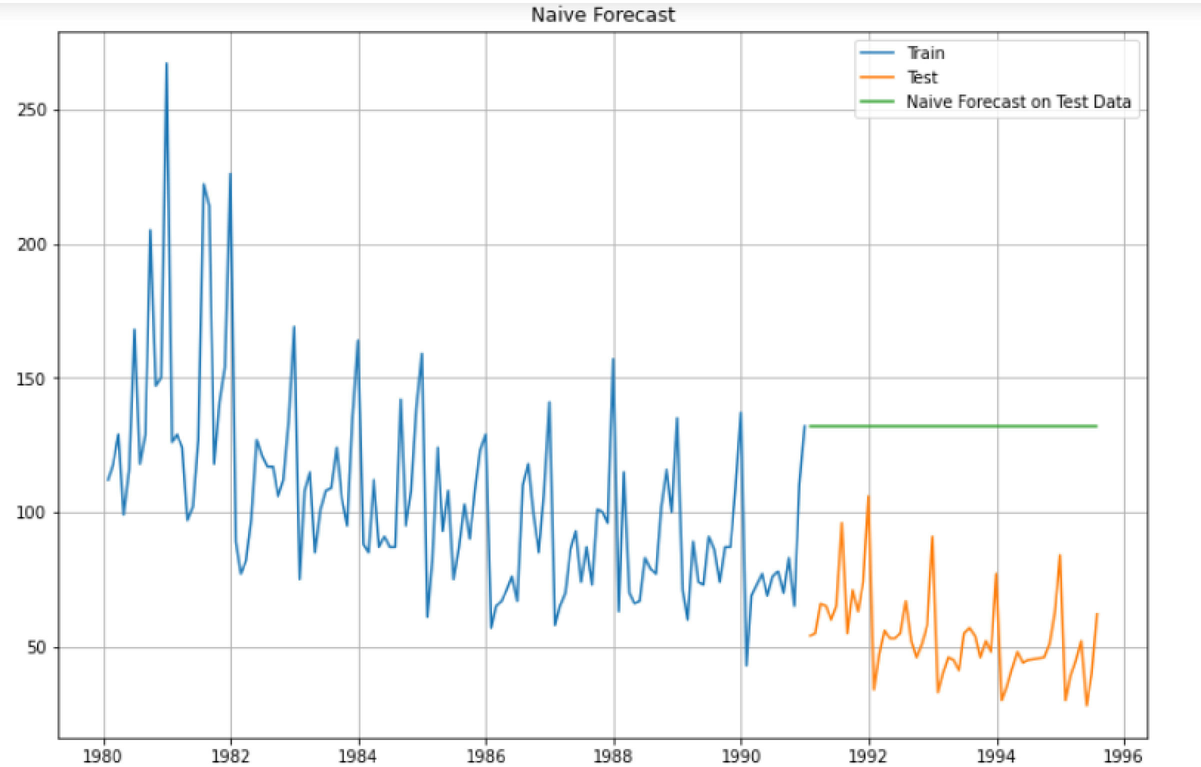


Figure 14 – Naïve Forecast Plot

	Test RMSE
NaiveModel	79.718773

MODEL 3 – Simple Average Model:

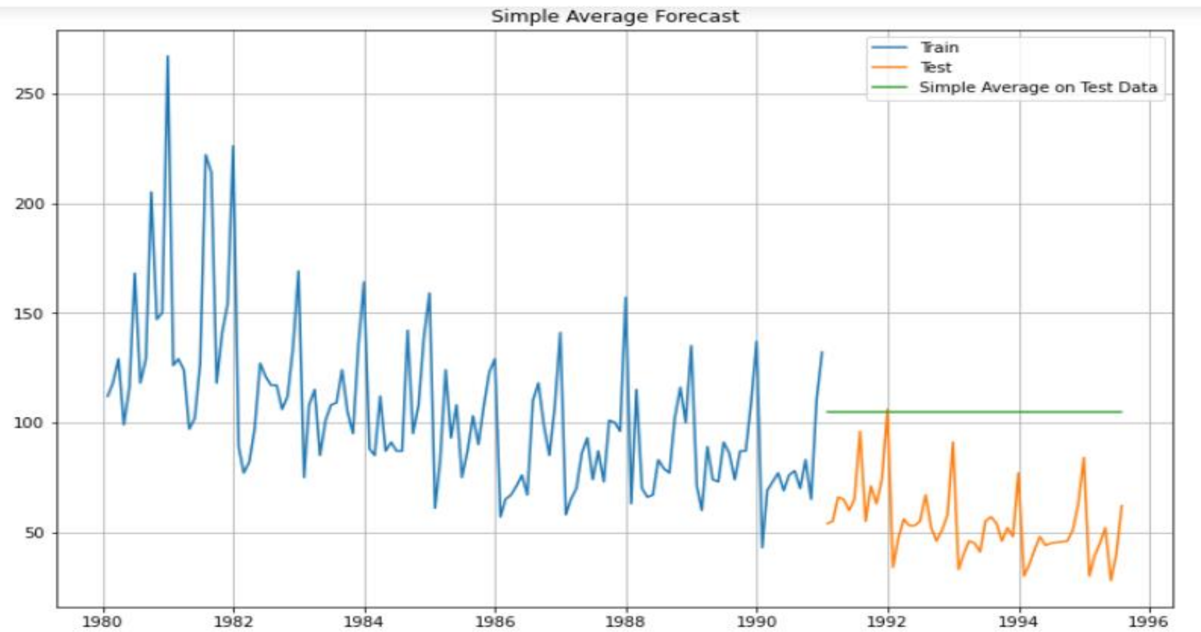


Figure 15 – Simple Average Forecast

	Test RMSE
SimpleAverageModel	53.46057

MODEL 4 – Moving Average Model:

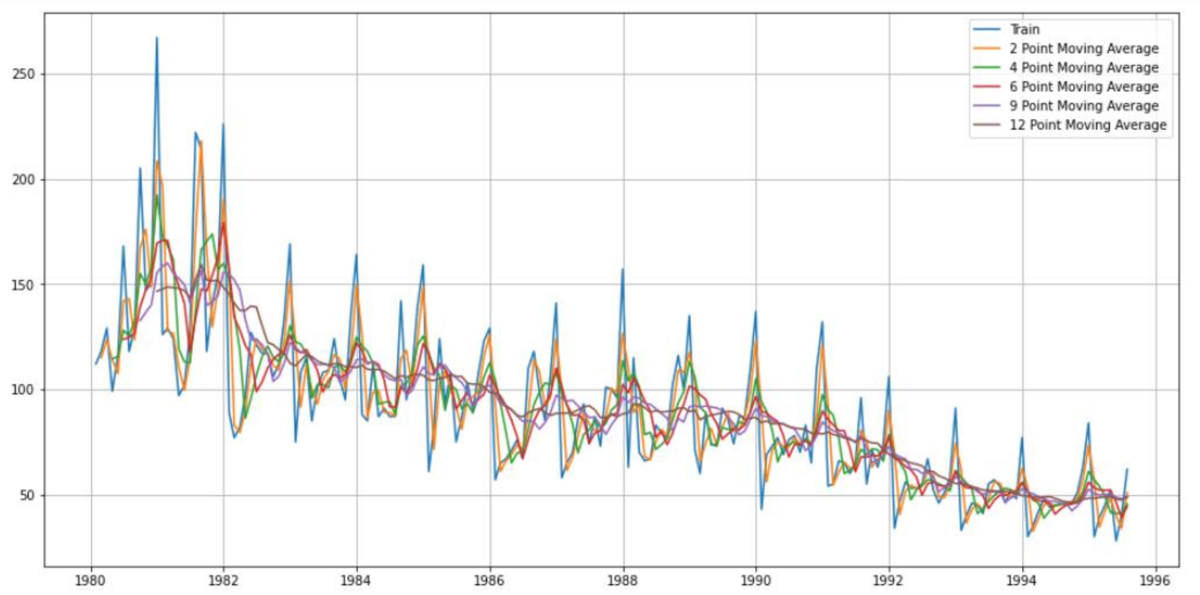


Figure 16 – Point-Wise Moving Average Forecast

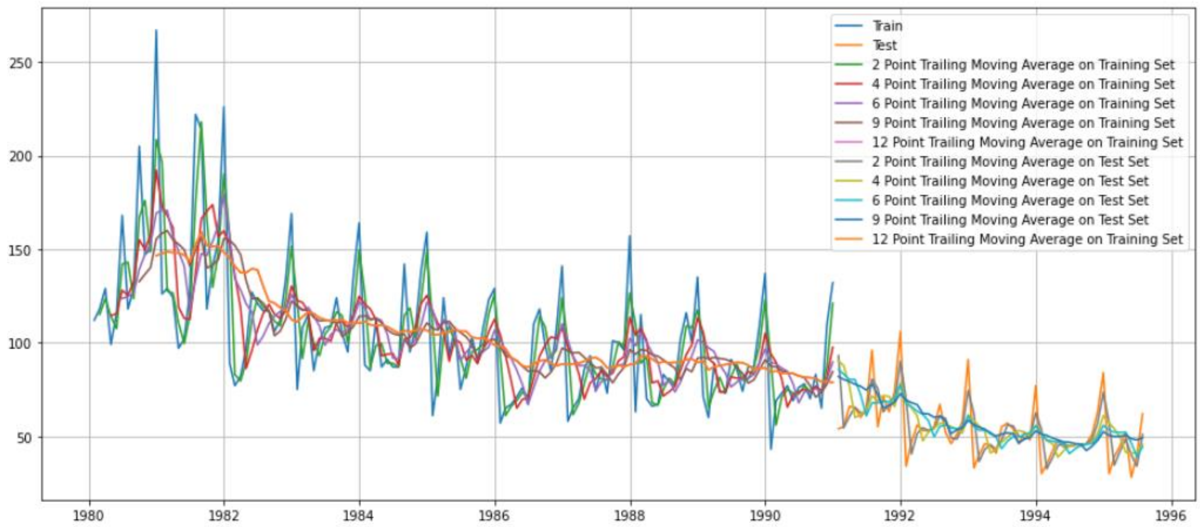


Figure 17 – Point-Wise Trailing Moving Average Forecast

	Test RMSE
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
12pointTrailingMovingAverage	15.236052

Plotting of all the models:

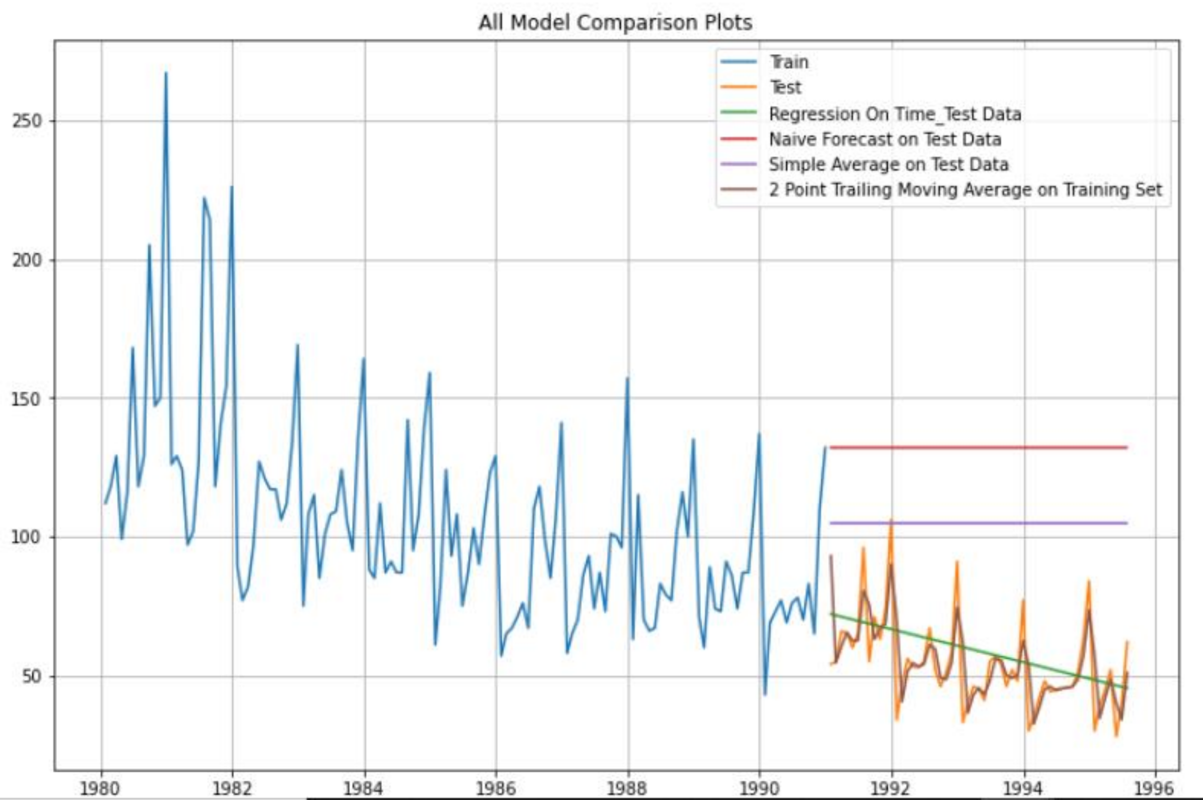


Figure 18 – All Model Comparison Plots

MODEL 5 – Simple Exponential Smoothing Model:

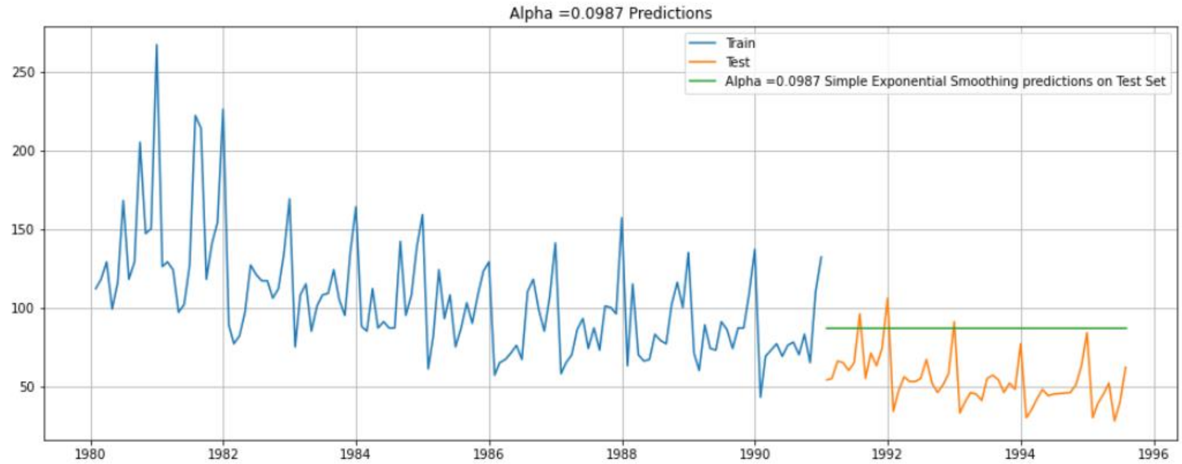


Figure 19 – Simple Exponential Smoothing Model

	Test RMSE
Alpha=0.0987,SimpleExponentialSmoothing	36.796243

MODEL 6 – Double Exponential Smoothing:

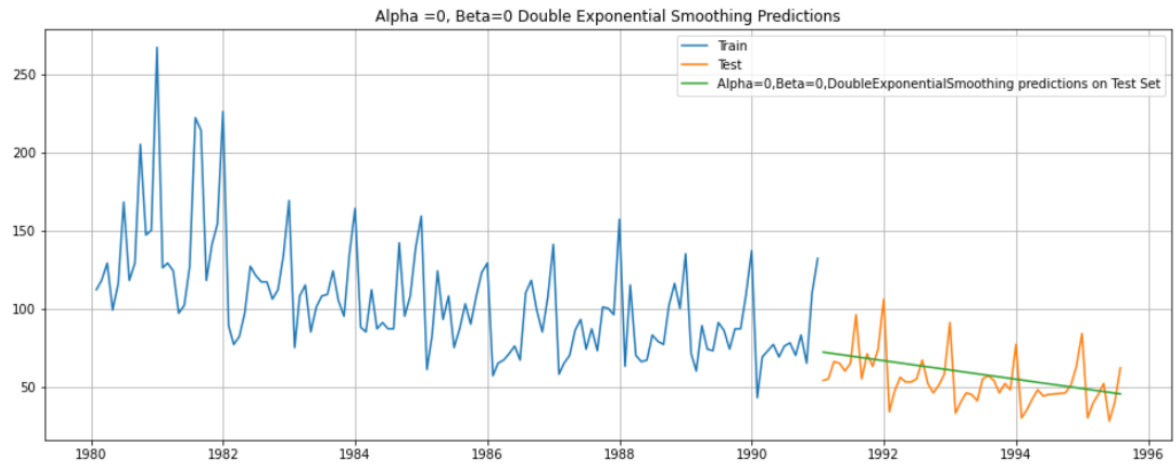


Figure 20 – Double Exponential Smoothing Model

	Test RMSE
For Alpha =0, Beta = 0 DoubleExponentialSmoothing	15.268961

MODEL 7 – Triple Exponential Smoothing Additive:

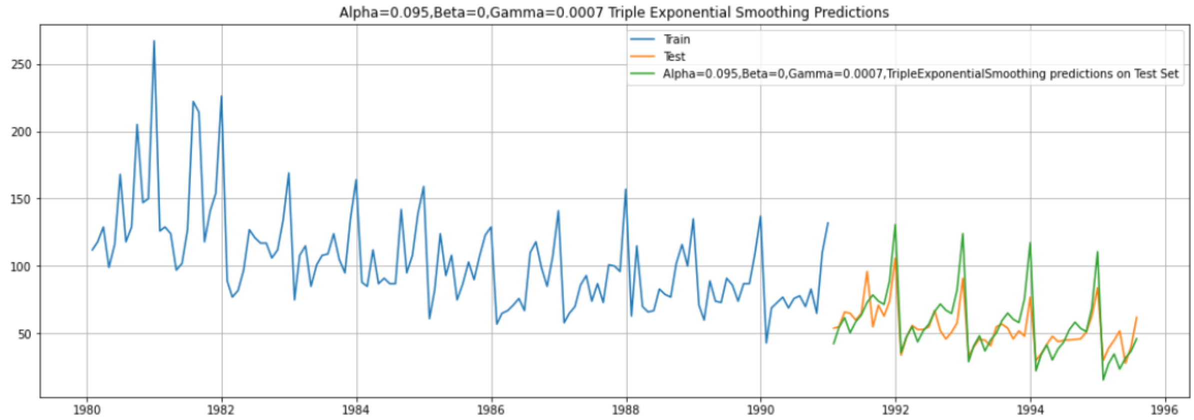


Figure 21 – Triple Exponential Smoothing Additive Model

	Test RMSE
Alpha=0.095,Beta=0,Gamma=0.0007,TripleExponentialSmoothing	14.176738

MODEL 8 – Triple Exponential Smoothing Multiplicative:

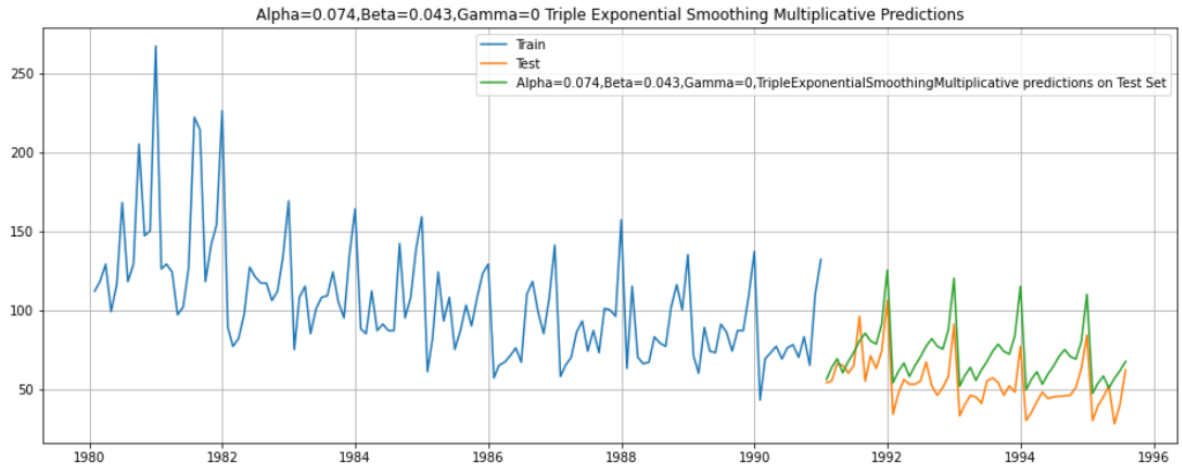


Figure 22 – Triple Exponential Smoothing Multiplicative Model

	Test RMSE
Alpha=0.074,Beta=0.043,Gamma=0 TripleExponentialSmoothingMultiplicative	19.741738

Performance using RSME on Test Data:

	Test RMSE
RegressionOnTime	15.268955
SimpleAverageModel	53.460570
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
12pointTrailingMovingAverage	15.236052
Alpha=0.0987,SimpleExponentialSmoothing	36.796243
For Alpha =0.1, Beta = 0 DoubleExponentialSmoothing	15.268961
Alpha=0.095,Beta=0,Gamma=0.0007,TripleExponentialSmoothing	14.176738
NaiveModel	79.718773
Alpha=0.074,Beta=0.043,Gamma=0 TripleExponentialSmoothingMultiplicative	19.741738

From the observations so far, we can clearly see that Trailing Moving Average models have performed better than the other models as they have a lower RMSE on the test data. The 2 point Trailing Moving Average Model has performed the best due to its lower RMSE of 11.52.

2.4 Build all the 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.

Hypothesis for Statistical Test:

Null Hypothesis – H0 = Time Series is not Stationary
Alternative Hypothesis – HA = Time Series is Stationary

Stationarity Check Using Dickey-Fuller Test:

```
Results of Dickey-Fuller Test:
Test Statistic                 -1.876699
p-value                         0.343101
#Lags Used                     13.000000
Number of Observations Used   173.000000
Critical Value (1%)            -3.468726
Critical Value (5%)            -2.878396
Critical Value (10%)           -2.575756
dtype: float64
```

We observe the Time Series is non-stationary for alpha = 0.05 as the p-value is > alpha at 0.34 .
Hence, we fail to reject the null hypothesis.

Let us take a difference of order 1 and check whether the Time Series is stationary or not.

Stationarity Check Using Dickey-Fuller Test by taking difference of Order 1:

```
Results of Dickey-Fuller Test:
Test Statistic                -8.044392e+00
p-value                        1.810895e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

We observe the Time Series is now stationary for alpha = 0.05 as the p-value at less than alpha.

2.5 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.

MODEL 9 – ARIMA Model:

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

Above is some combination of different parameters of p and q in the range of 0 and 2

ARIMA AIC SCORES for Parameters in range of 0 & 2:

	param	AIC
2	(0, 1, 2)	1279.671529
5	(1, 1, 2)	1279.870723
4	(1, 1, 1)	1280.574230
7	(2, 1, 1)	1281.507862
8	(2, 1, 2)	1281.870722
1	(0, 1, 1)	1282.309832
6	(2, 1, 0)	1298.611034
3	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

We can see the best AIC is for ARIMA (0,1,2) of 1279.67. Below, we will built our ARIMA Model for this parameter and check the performance.

ARIMA (0,1,2) Model Results:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -636.836
Date:                Wed, 02 Mar 2022   AIC                           1279.672
Time:                        00:08:56   BIC                           1288.297
Sample:                    01-31-1980   HQIC                          1283.176
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.6970      0.072     -9.689      0.000      -0.838      -0.556
ma.L2         -0.2042      0.073     -2.794      0.005      -0.347      -0.061
sigma2       965.8407     88.305     10.938      0.000     792.766    1138.915
===================================================================================
Ljung-Box (L1) (Q):                   0.14   Jarque-Bera (JB):                39.24
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               0.36   Skew:                             0.82
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.13
===================================================================================

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

Figure 23 – Arima Model

	RMSE
ARIMA(0,1,2)	37.30648

MODEL 10 – SARIMA Model with seasonality 6:

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 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: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

Above is some combination of different parameters of p and q in the range of 0 and 2

SARIMA AIC SCORES for Parameters in range of 0 & 2:

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655818
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.220389
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

We can see the best AIC is for SARIMA (1,1,2) (2,0,2,6) of 1041.65. Below we will built our SARIMA Model for this parameter and check the performance.

SARIMA (1,1,2) (2,0,2,6) Model Results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -512.828
Date:                           Wed, 02 Mar 2022   AIC                           1041.656
Time:                                   00:15:34   BIC                           1063.685
Sample:                                        0   HQIC                          1050.598
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5939      0.152     -3.914      0.000      -0.891      -0.296
ma.L1         -0.1954    188.566     -0.001      0.999    -369.777     369.387
ma.L2         -0.8046    151.765     -0.005      0.996    -298.258     296.649
ar.S.L6       -0.0625      0.035     -1.794      0.073      -0.131       0.006
ar.S.L12       0.8451      0.039     21.889      0.000       0.769       0.921
ma.S.L6        0.2226    188.635      0.001      0.999    -369.495     369.940
ma.S.L12      -0.7774    146.598     -0.005      0.996    -288.104     286.549
```

```
sigma2        335.1965       0.906     369.902       0.000      333.420      336.973
===================================================================================
Ljung-Box (L1) (Q):                   0.07   Jarque-Bera (JB):                56.68
Prob(Q):                              0.78   Prob(JB):                         0.00
Heteroskedasticity (H):               0.47   Skew:                             0.52
Prob(H) (two-sided):                  0.02   Kurtosis:                         6.26
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.37e+21. Standard errors may be unstable.
```

Figure 24 – Sarima Model with seasonality 6

Summary Frame for Alpha = 0.05:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.839941	18.848279	25.897993	99.781888
1	67.629885	19.300121	29.802343	105.457427
2	74.746081	19.412682	36.697924	112.794238
3	71.324859	19.475628	33.153329	109.496389
4	76.016791	19.483907	37.829034	114.204548

	RMSE
SARIMA(1, 1, 2)(2,0,2,6)	26.134254

MODEL 11 – SARIMA Model with seasonality 12:

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

Above is some combination of different parameters of p and q in the range of 0 and 2

SARIMA AIC SCORES for Parameters in range of 0 & 2:

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
53	(1, 1, 2)	(2, 0, 2, 12)	896.686897
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

We can see the best AIC is for SARIMA (0,1,2) (2,0,2,12) of 887.94. Below we will built our SARIMA Model for this parameter and check the performance.

SARIMA (0,1,2) (2,0,2,12) Model Results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Wed, 02 Mar 2022   AIC                            887.938
Time:                                    00:24:11   BIC                            906.448
Sample:                                         0   HQIC                           895.437
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.512     -0.004      0.996    -372.279     370.593
ma.L2         -0.1573     29.773     -0.005      0.996     -58.512      58.197
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3136   4.76e+04      0.005      0.996   -9.31e+04    9.36e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.75   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================

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

Figure 25 – Sarima Model with Seasonality 12

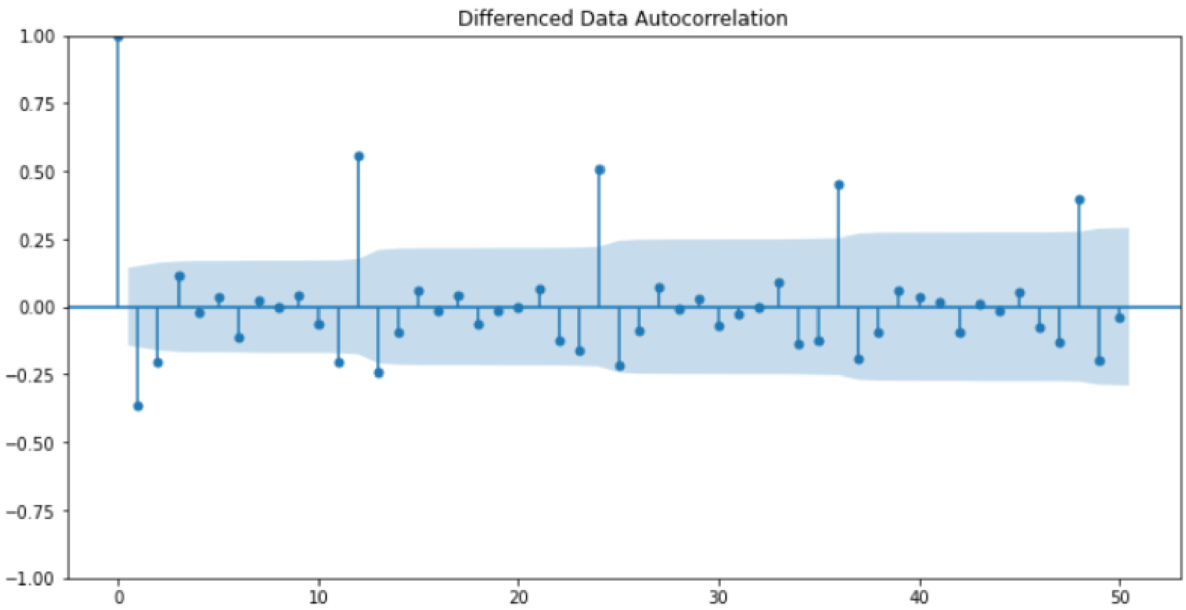
Summary Frame for Alpha = 0.05:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.867261	15.928500	31.647975	94.086547
1	70.541189	16.147658	38.892362	102.190017
2	77.356410	16.147655	45.707587	109.005233
3	76.208813	16.147655	44.559990	107.857637
4	72.747397	16.147655	41.098574	104.396220

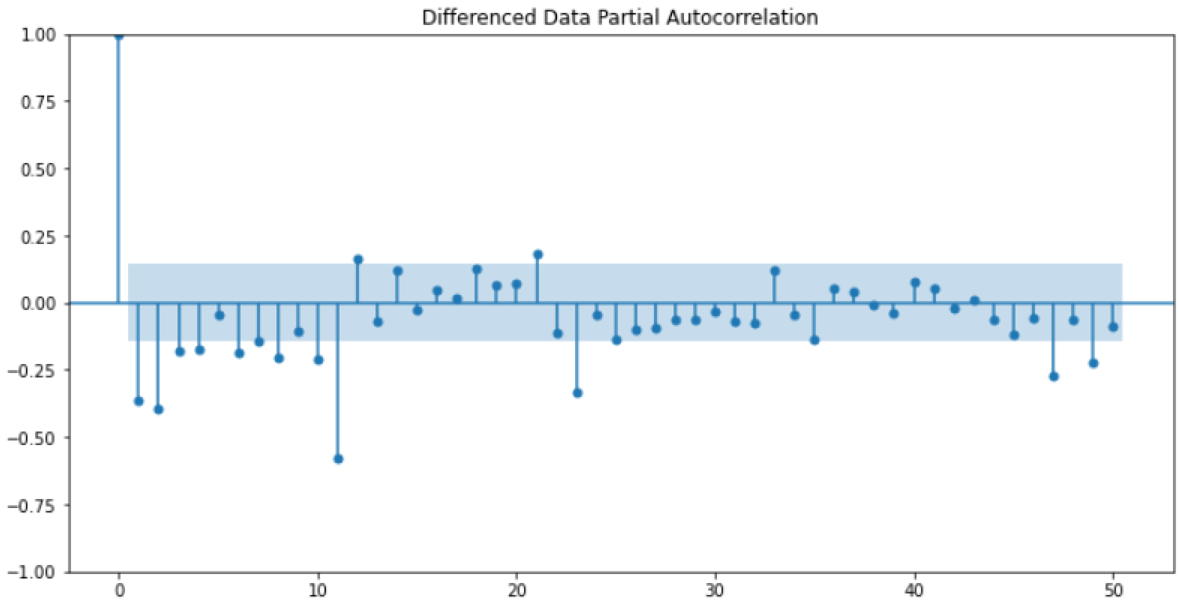
	RMSE
SARIMA(0, 1, 2)(2, 0, 2, 12)	26.928361

2.6 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.

ACF Plot:



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 to 4.
- The Moving-Average parameter in an ARIMA model is ‘q’ which comes from the significant lag before the ACF plot cuts-off to 2.

MODEL 12 – Manual ARIMA Model based on cut-0ff point (4,1,2):

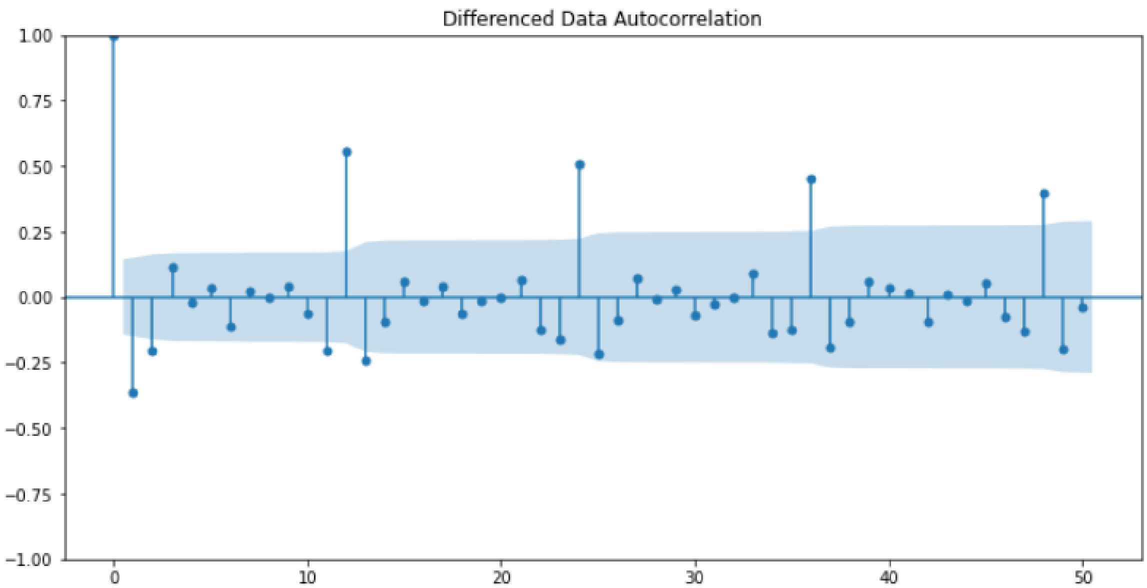
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(4, 1, 2)   Log Likelihood                -635.859
Date:                Sun, 06 Mar 2022   AIC                           1285.718
Time:                        12:51:01   BIC                           1305.845
Sample:                    01-31-1980   HQIC                          1293.896
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3838      0.923     -0.416      0.677      -2.192       1.425
ar.L2          0.0046      0.258      0.018      0.986      -0.502       0.511
ar.L3          0.0414      0.113      0.366      0.714      -0.180       0.263
ar.L4         -0.0054      0.177     -0.031      0.976      -0.353       0.342
ma.L1         -0.3239      0.933     -0.347      0.729      -2.153       1.505
ma.L2         -0.5407      0.874     -0.619      0.536      -2.254       1.172
sigma2       951.1524     93.870     10.133      0.000     767.170    1135.135
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                32.85
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.37   Skew:                             0.77
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.91
===================================================================================

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

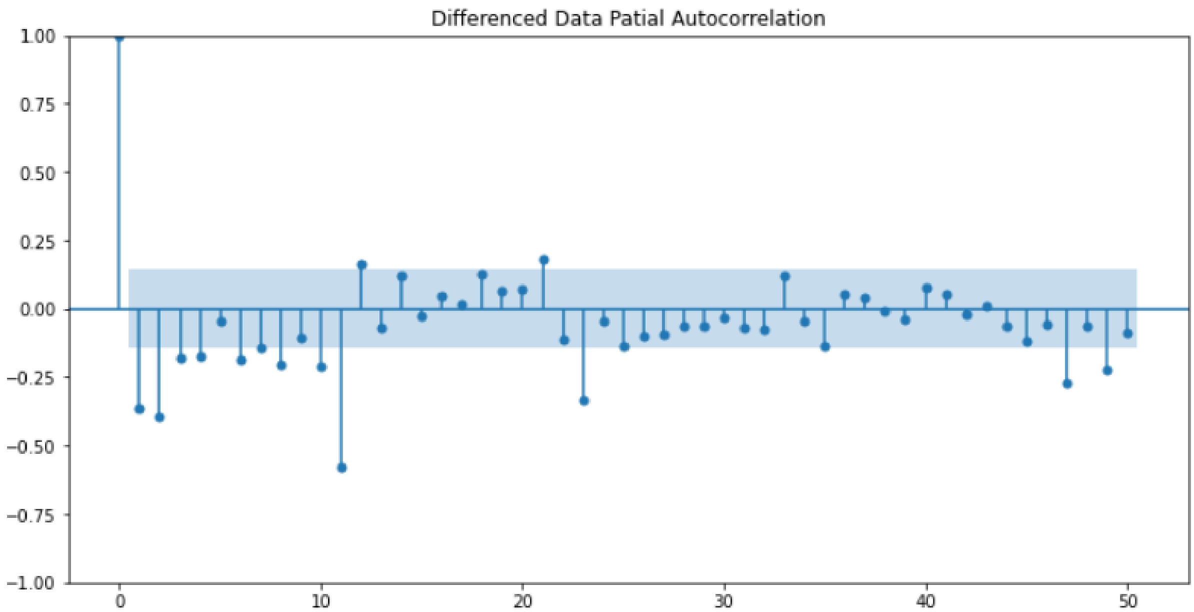
Figure 26 – Arima Model based on ACF and PACF cut-off points

	RMSE
ARIMA(4,1,2)	37.037639

MODEL – 13 Manual SARIMA Model based on cut-0ff point (4,1,2):

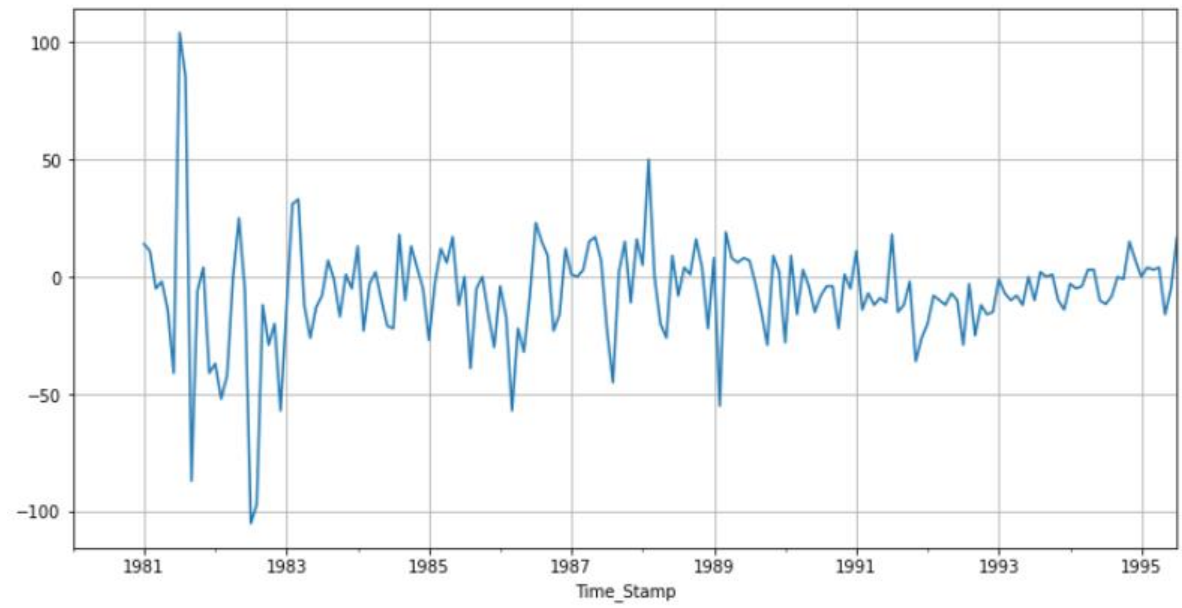
ACF Plot:



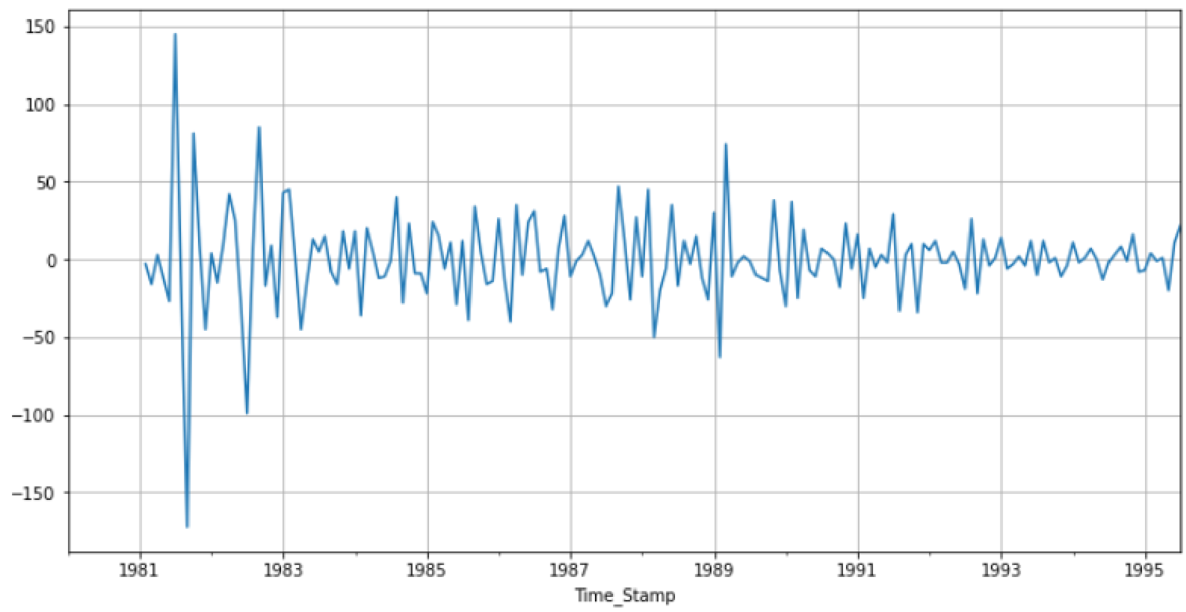


We see that there can be a seasonality of 12. We will run our auto SARIMA models by setting seasonality for 12.

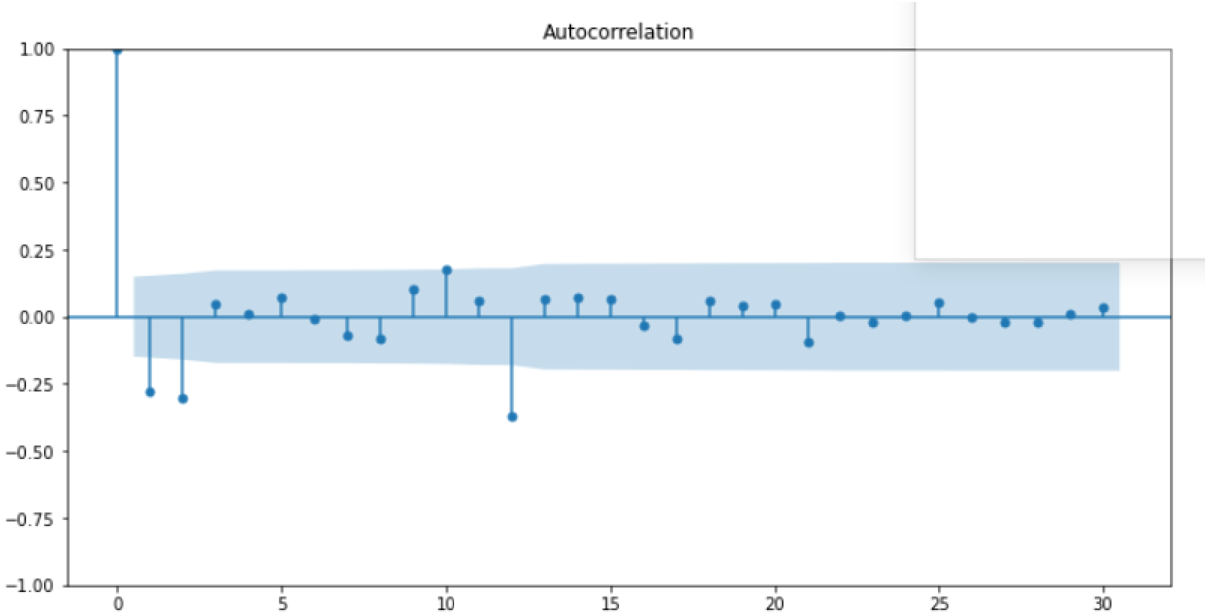
Plot with Seasonal Difference of 12:



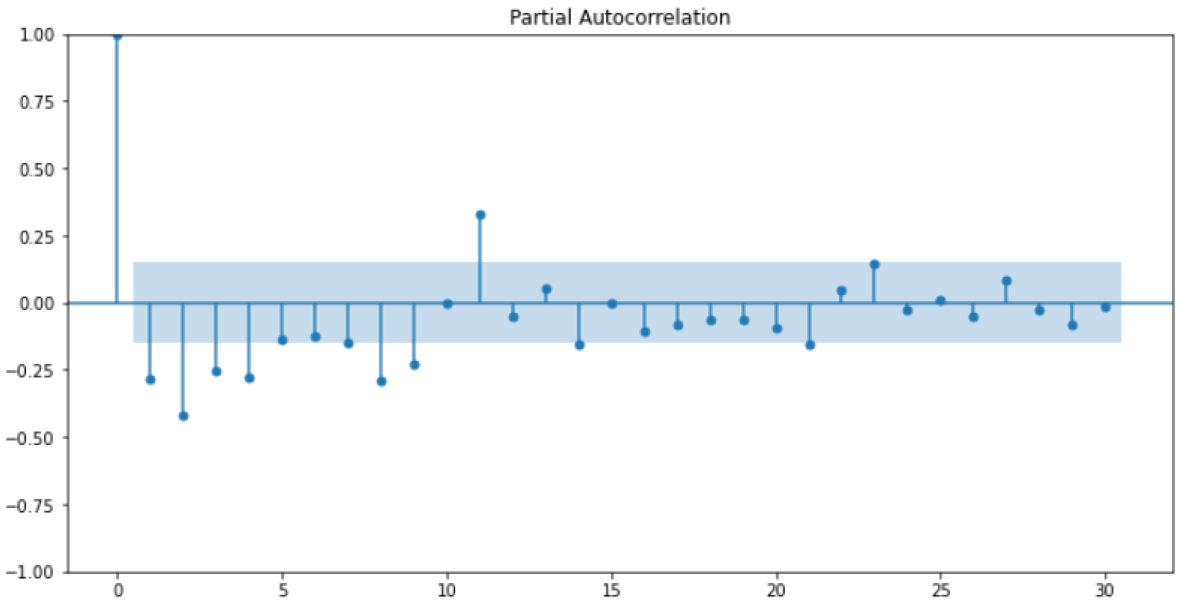
Plot without Trend and Only Seasonality:



ACF plot for the new modified Time Series with Stationarity:



PACF plot for the new modified Time Series with Stationarity:



Manual SARIMA (4,1,2) (4,1,2,12) Model Results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(4, 1, 2)x(4, 1, 2, 12)   Log Likelihood                -277.661
Date:                            Sun, 06 Mar 2022   AIC                            581.322
Time:                                    13:16:32   BIC                            609.983
Sample:                                         0   HQIC                           592.663
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9742      0.199     -4.900      0.000      -1.364      -0.585
ar.L2         -0.1123      0.285     -0.394      0.693      -0.670       0.446
ar.L3         -0.1044      0.277     -0.377      0.706      -0.647       0.438
ar.L4         -0.1285      0.162     -0.794      0.427      -0.446       0.189
ma.L1          0.1605    299.757      0.001      1.000    -587.353     587.674
ma.L2         -0.8395    251.668     -0.003      0.997    -494.099     492.420
ar.S.L12      -0.1441      0.364     -0.396      0.692      -0.858       0.569
ar.S.L24      -0.3596      0.227     -1.587      0.113      -0.804       0.085
ar.S.L36      -0.2153      0.106     -2.039      0.041      -0.422      -0.008
ar.S.L48      -0.1195      0.093     -1.281      0.200      -0.302       0.063
ma.S.L12      -0.5158      0.343     -1.502      0.133      -1.189       0.157
ma.S.L24       0.2085      0.373      0.559      0.576      -0.523       0.940
sigma2       215.3526   6.46e+04      0.003      0.997   -1.26e+05    1.27e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                 2.41
Prob(Q):                              0.86   Prob(JB):                         0.30
Heteroskedasticity (H):               0.49   Skew:                             0.32
Prob(H) (two-sided):                  0.10   Kurtosis:                         3.68
===================================================================================

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

Figure 27 – Sarima Model with Seasonality 12 based on ACF and PACF cut-off points

Summary Frame for Alpha = 0.05:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	46.385325	14.770634	17.435414	75.335237
1	62.932846	14.989764	33.553449	92.312244
2	63.527810	14.999436	34.129455	92.926165
3	66.473786	15.179628	36.722261	96.225311
4	63.540811	15.180505	33.787567	93.294055

	RMSE
SARIMA(4,1,2)(4,1,2,12)	17.529039

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

MODELS	TEST RMSE
2pointTrailingMovingAverage	11.529278
Alpha=0.095,Beta=0,Gamma=0.0007,TripleExponentialSmoothing	14.176738
4pointTrailingMovingAverage	14.451403

6pointTrailingMovingAverage	**14.566327**
9pointTrailingMovingAverage	**14.72763**
12pointTrailingMovingAverage	**15.236052**
RegressionOnTime	**15.268955**
For Alpha =0, Beta = 0 DoubleExponentialSmoothing	**15.268961**
SARIMA(4,1,2)(4,1,2,12)	**17.529039**
SARIMA(1, 1, 2)(2,0,2,6)	**26.134254**
Alpha=0.074,Beta=0.043,Gamma=0 TripleExponentialSmoothingMultiplicative	**19.741738**
SARIMA(0, 1, 2)(2, 0, 2, 12)	**26.928361**
Alpha=0.0987,SimpleExponentialSmoothing	**36.796243**
ARIMA(4,1,2)	**37.037639**
ARIMA(0,1,2)	**37.30648**
Simple Average	**53.46057**
Naive Model	**79.718773**

Table 1 – All Models with RMSE

2.8 Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

The best model built is **2pointTrailingMovingAverage** with Test RMSE of **11.529278.** However, moving average models are somewhat outdated and there are new models such as Arima/Sarima which are more robust and more advanced. Also, it is somewhat incorrect to use a 2point trailing moving average to future forecast 12 months. Hence, we will be building our model on the 2nd best optimum model, "**Alpha=0.095,Beta=0,Gamma=0.0007,TripleExponentialSmoothing"** which has the Test RMSE as **14.176** .Now we will built the best optimum full model on the same parameters

Future 12 Months Sales Forecast :

```
1995-08-31    49.878500
1995-09-30    46.705164
1995-10-31    45.439534
1995-11-30    60.040383
1995-12-31    98.313260
1996-01-31    13.835254
1996-02-29    24.144658
1996-03-31    31.704142
1996-04-30    24.511248
1996-05-31    27.880899
1996-06-30    33.379915
1996-07-31    44.016469
Freq: M, dtype: float64
```

Future 12 Months Forecast Plot:

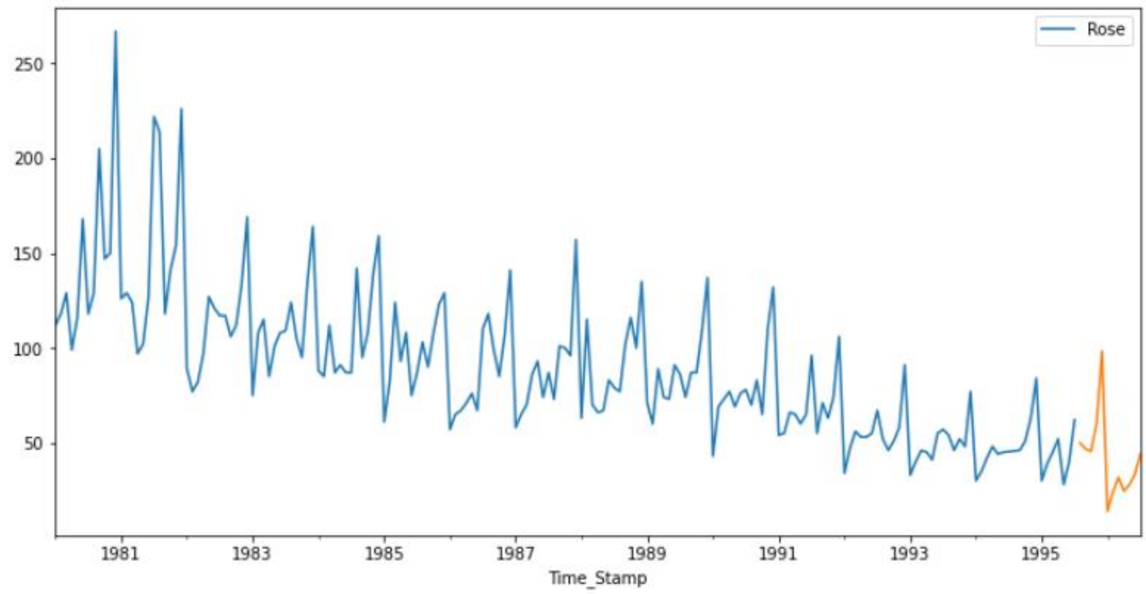


Figure 28 – Future 12 months Forecast Plot

Summary Frame of next 12 months for Alpha = 0.05:

	lower_CI	prediction	upper_ci
1995-08-31	15.161545	49.878500	84.595455
1995-09-30	11.988209	46.705164	81.422119
1995-10-31	10.722579	45.439534	80.156489
1995-11-30	25.323428	60.040383	94.757338
1995-12-31	63.596305	98.313260	133.030215
1996-01-31	-20.881701	13.835254	48.552209
1996-02-29	-10.572297	24.144658	58.861613
1996-03-31	-3.012813	31.704142	66.421097
1996-04-30	-10.205707	24.511248	59.228203
1996-05-31	-6.836056	27.880899	62.597854
1996-06-30	-1.337040	33.379915	68.096870
1996-07-31	9.299514	44.016469	78.733424

Forecast for next 12 months along with confidence band:

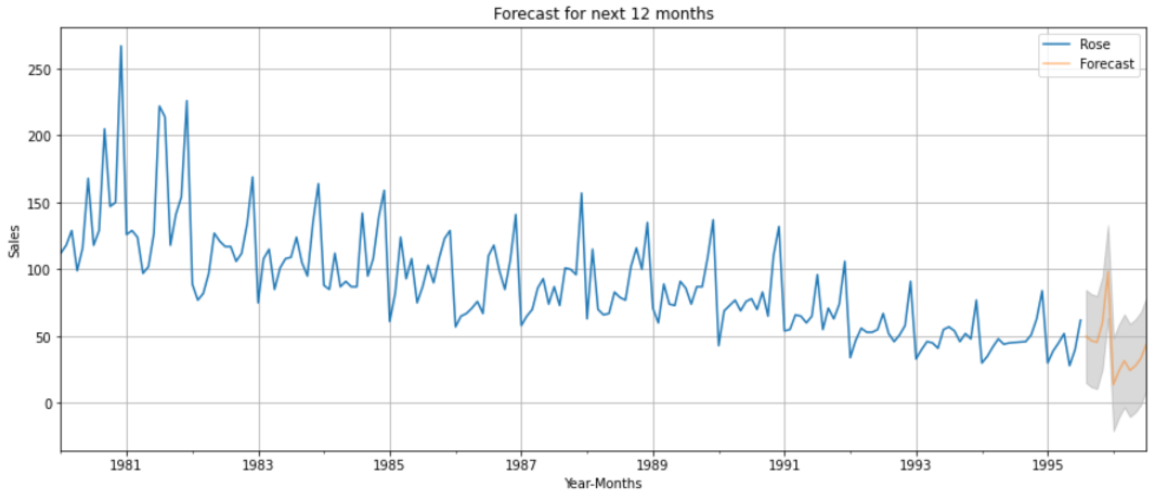


Figure 29 – Future 12 months Forecast with Confidence Bands

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

- The sales for Rose Wine has been dropping over the years. Hence, there is a need to reintroduce this variant and maybe with a little bit of refreshing touch, perhaps a different breed of grapes could be tried or cultivation at more cooler places, indoor cultivation at specific temperatures could work.

- It also could be evaluated if transportation of the grapes/bottles is having any effect on the quality of the wine or if there could be modifications made to the barrels that are used.

- The future forecast shows a slight uptick which is a good sign but it still remains a question if the popularity of Rose wine is sliding down or there is something which can turn around and increase the popularity.

- The company could also look at marketing strategies during peak season to boost sales even further which can maybe cover up for the lower sales during off-season or/and they can come up with impressive offers during off-season to increase the sales