TIME SERIES MODELLING PROJECT

BUSINESS REPORT

Report by:

Tushar Babbar
PGPDSBA Online
DSBA Dec20 Group 5

Table of Contents

Table of Figures

Problem 1: Linear Regression

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

Data Dictionary:

Variable Name	*Description*
Sparkling	-- Sales dataset for Sparkling Dataset
Rose	-- Sales dataset for Rose Dataset
YearMonth	-- YearMonth indexes for sales .

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

HEAD OF THE DATA

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471

TAIL OF THE DATA

	YearMonth	Sparkling
182	1995-03	1897
183	1995-04	1862
184	1995-05	1670
185	1995-06	1688
186	1995-07	2031

1.1 Creating the Time Stamps and adding to the data frame to make it a Time Series Data

```
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')
```

1.1.1 Adding the time stamp to the data frame and setting it as index.

	Sparkling	Time_Stamp
0	1686	1980-01-31
1	1591	1980-02-29
2	2304	1980-03-31
3	1712	1980-04-30
4	1471	1980-05-31

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

1.1.2 Plot the Time Series to understand the behaviour of the data.

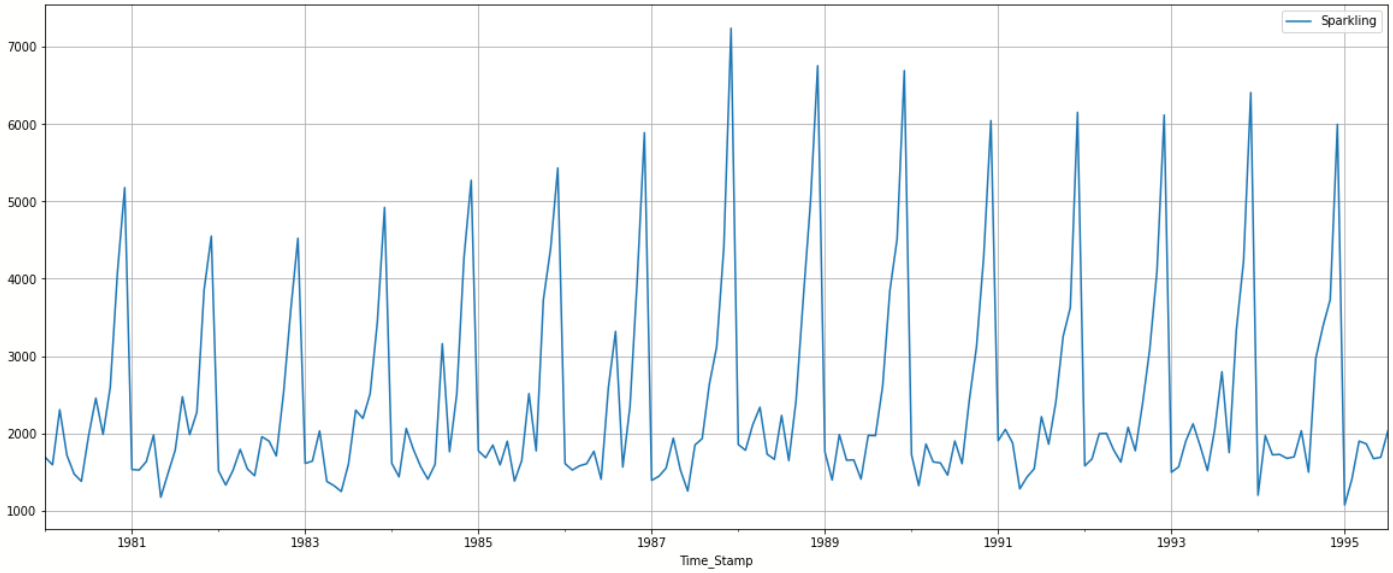


Figure 1(a)

- *As observed there is a slight trend in the data as well as seasonality.*

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

HEAD OF THE DATA

	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

TAIL OF THE DATA

	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

1A.1 Creating the Time Stamps and adding to the data frame to make it a Time Series Data

```
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')
```

1A.2 Adding the time stamp to the data frame and setting it as index.

	Rose	Time_Stamp
0	112.0	1980-01-31
1	118.0	1980-02-29
2	129.0	1980-03-31
3	99.0	1980-04-30
4	116.0	1980-05-31

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

1A.3 Plot the Time Series to understand the behaviour of the data.

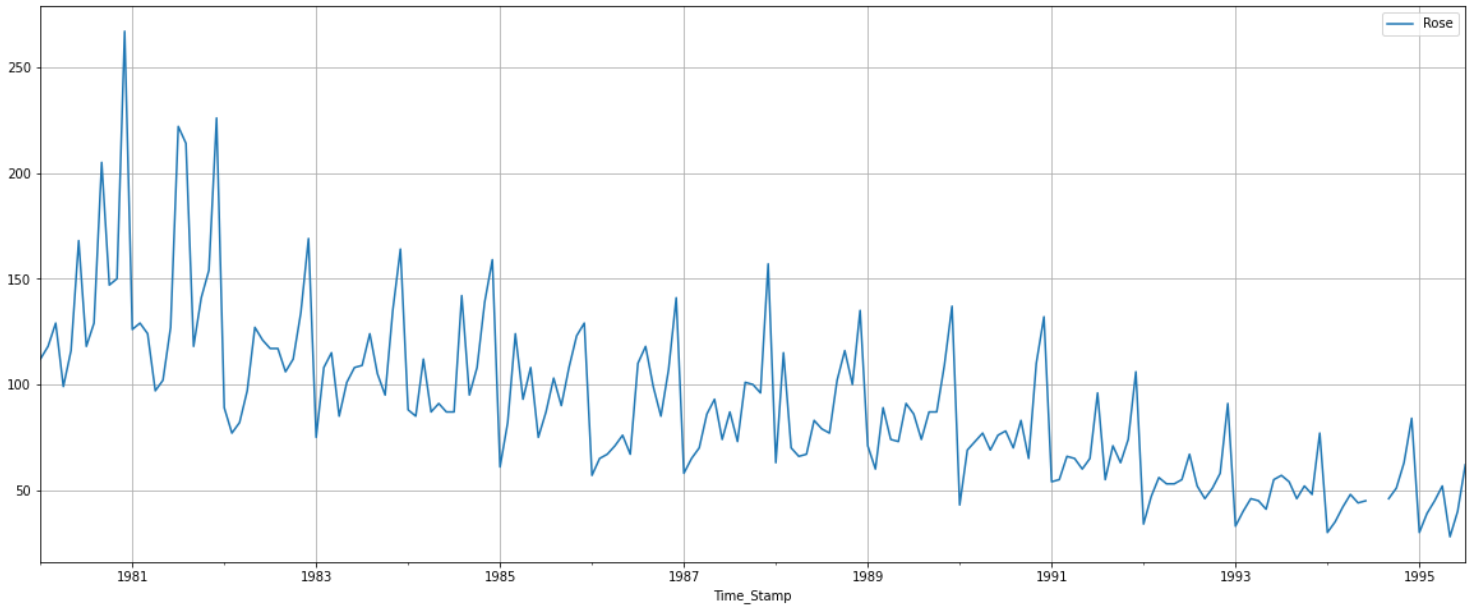


Figure 1(b)

- *As observed there is a slight trend in the data as well as seasonality.*

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

Check the basic measures of descriptive statistics

	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000
50%	1874.000
75%	2549.000
max	7242.000

- **Sparkling has no null Values.**

	Rose
count	185.000
mean	90.395
std	39.175
min	28.000
25%	63.000
50%	86.000
75%	112.000
max	267.000

- **Rose has 2 null values, which will affect our model.**
- **To remove the null values, we have interpolated our dataset with method as 'linear'.**

2.1 PLOT A BOXPLOT TO UNDERSTAND THE SPREAD ACROSS DIFFERENT YEARS AND WITHIN DIFFERENT MONTHS ACROSS YEARS.

SPARKLING DATASET

YEARLY BOXPLOT

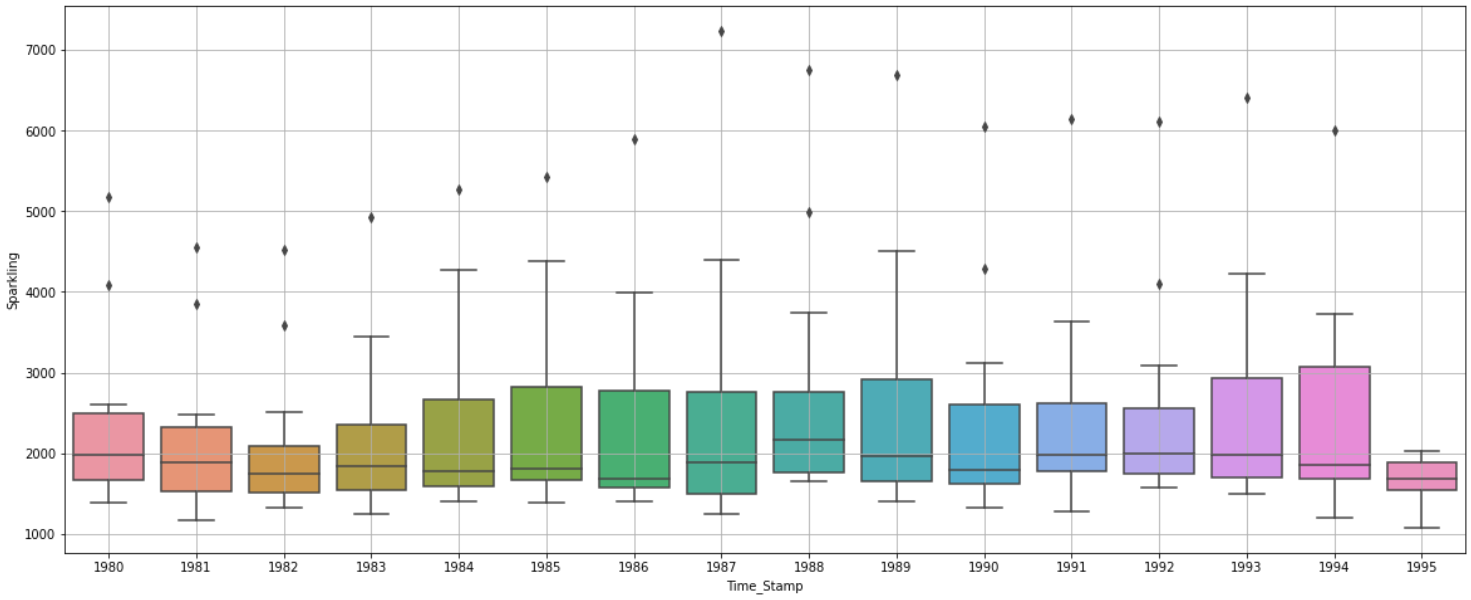


Figure 2(a)

- It can be observed that initially the yearly data has a upward trend, with the highest median value being in the year 1993.

MONTHLY BOXPLOT

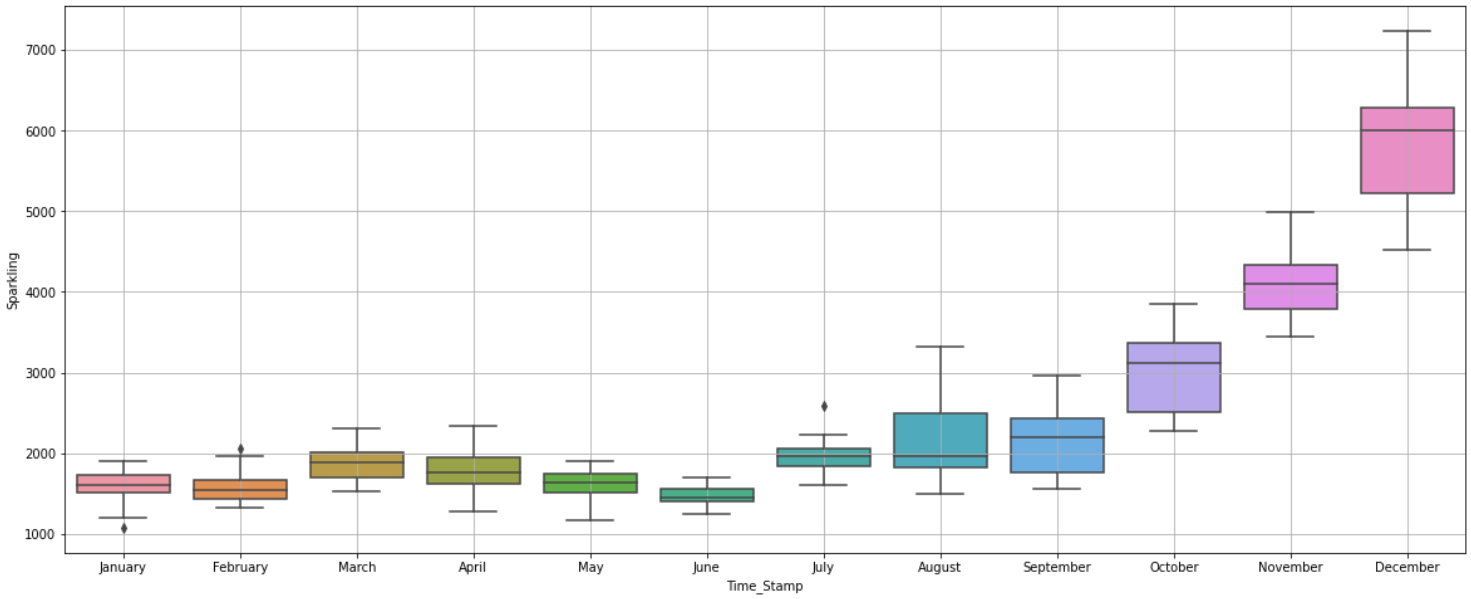


Figure 2(b)

- It can be observed that monthly data has an exponential increasing trend, with the highest median value being in the month of December.

ROSE DATASET

YEARLY BOXPLOT

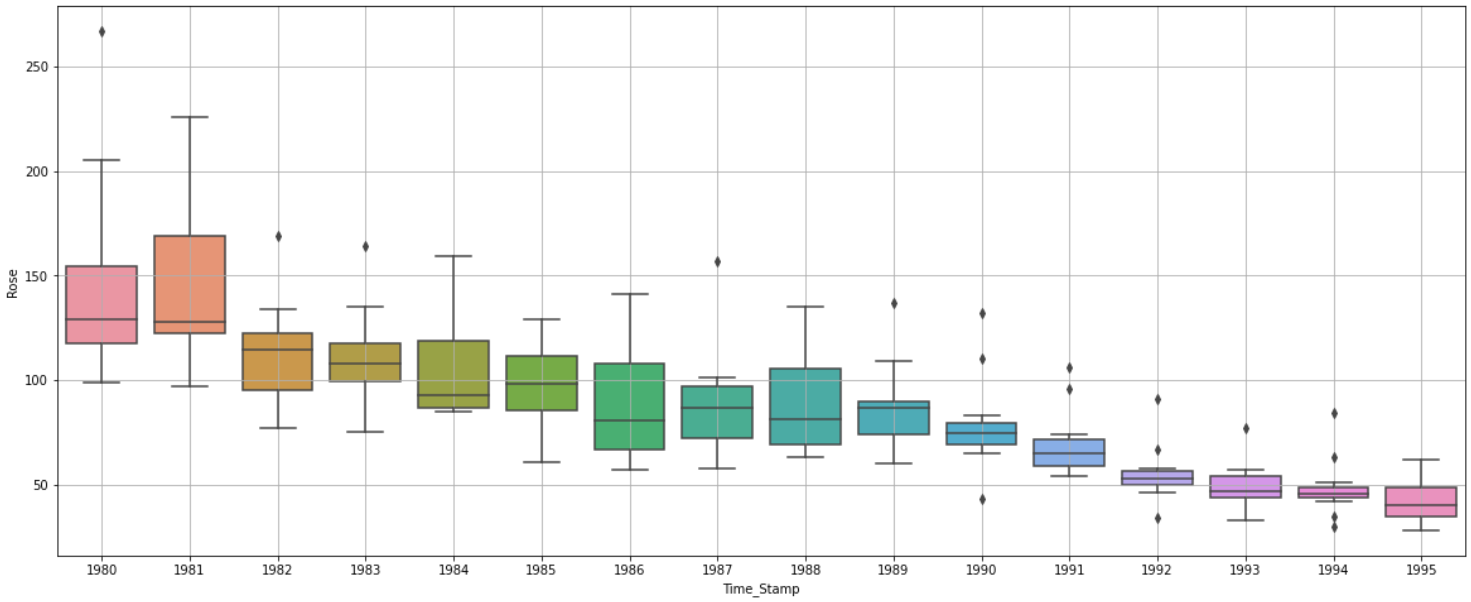


Figure 2(c)

- It can be observed that the yearly data has a exponentially decreasing trend, with the lowest median value being in the year 1995.

MONTHLY BOXPLOT

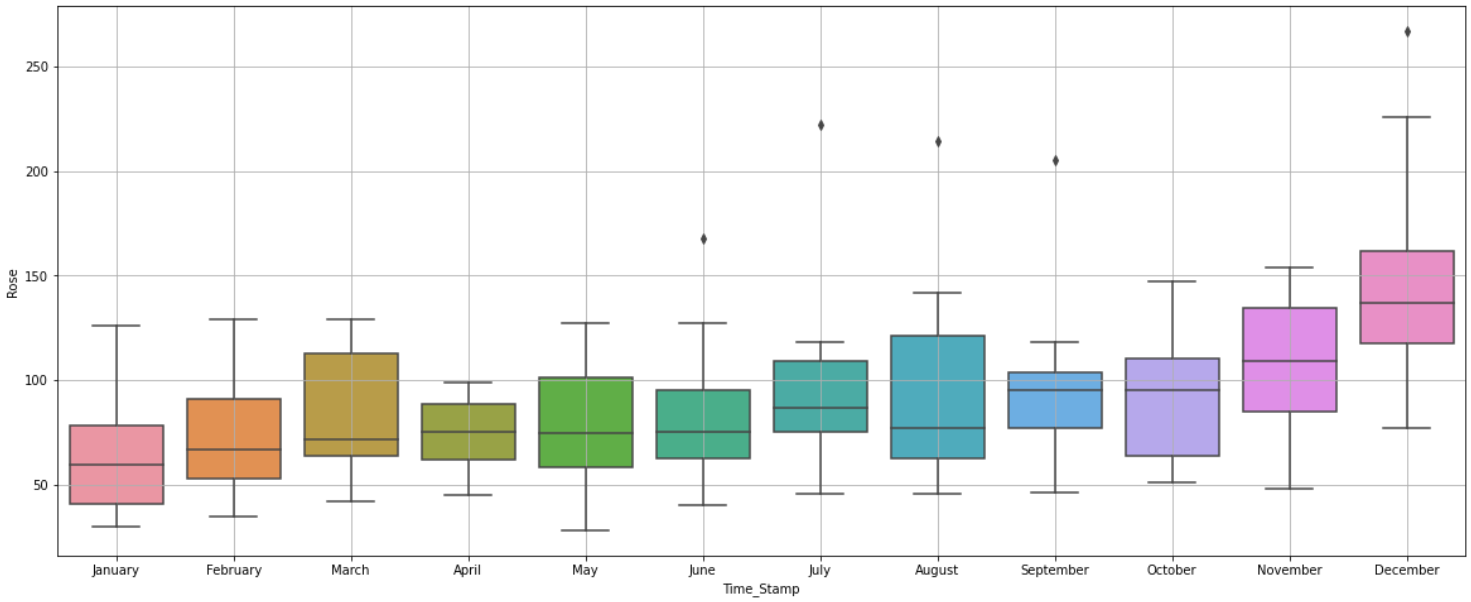


Figure 2(d)

- It can be observed that monthly data has an increasing trend, with the highest median value being in the month of December.

2.2 PLOT A TIME SERIES MONTHPLOT TO UNDERSTAND THE SPREAD ACROSS DIFFERENT YEARS AND WITHIN DIFFERENT MONTHS ACROSS YEARS

SPARKLING

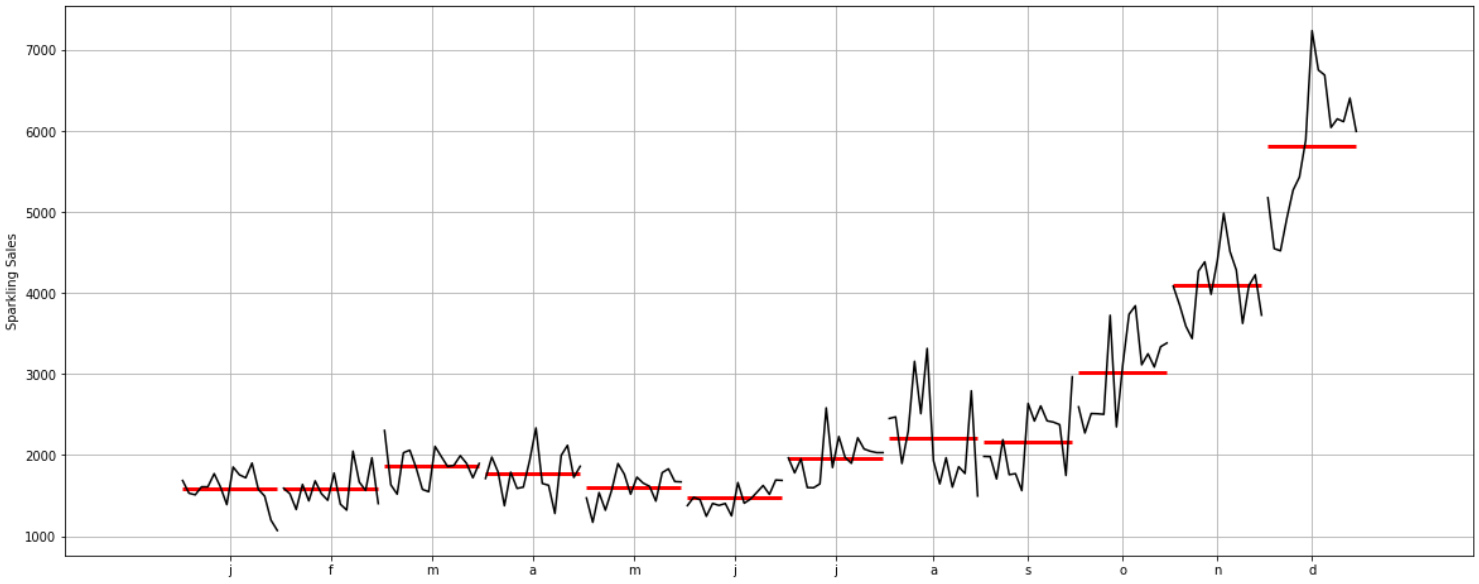


Figure 2.2(a)

This plot shows us the behaviour of the Time Series ('Sparkling' in this case) across various months. The red line is the median value, which is the maximum for December.

ROSE

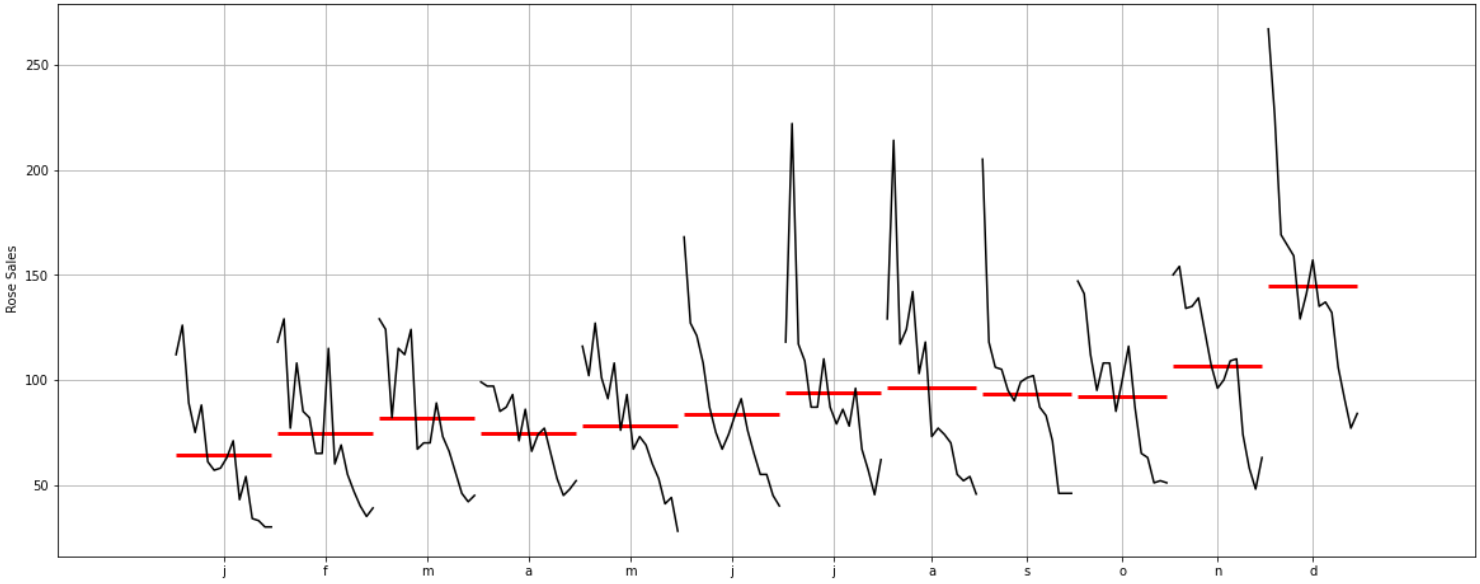


Figure 2.2(b)

This plot shows us the behaviour of the Time Series ('Rose' in this case) across various months. The red line is the median value, which is the maximum for December.

PLOT A GRAPH OF MONTHLY SPARKLING SALES ACROSS YEARS.

Time_Stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_Stamp												
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

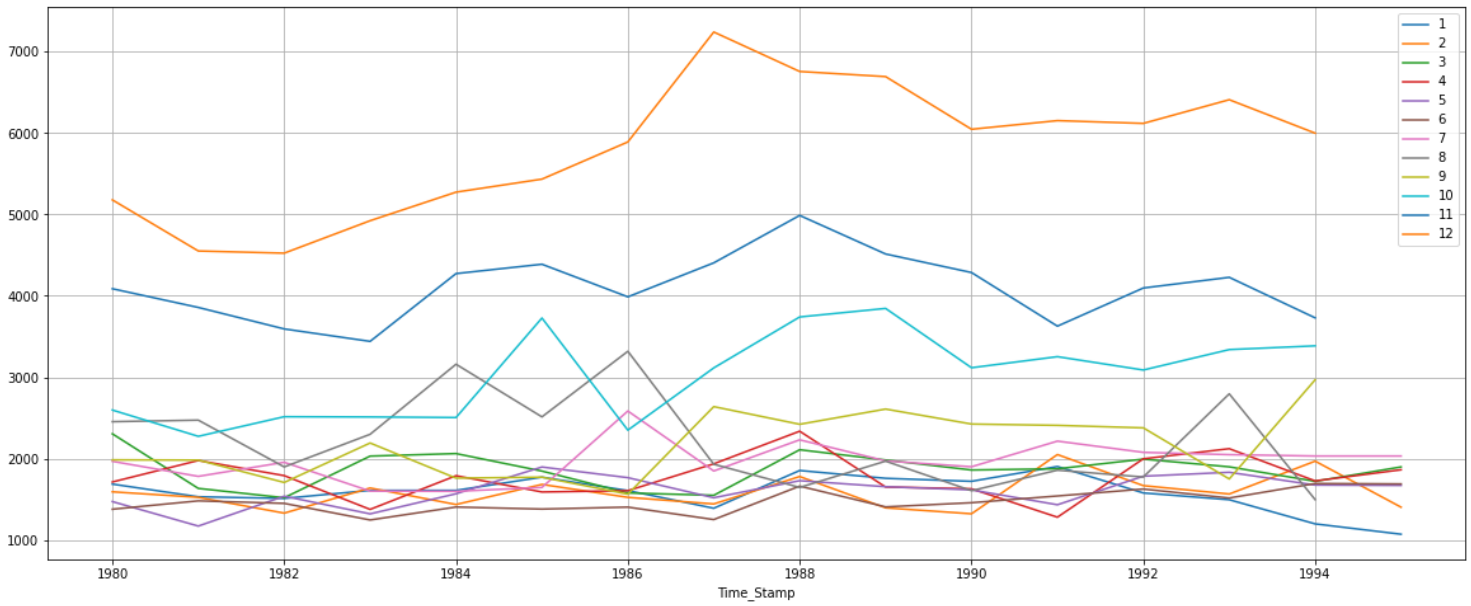


Figure 2.2(c)

- Similar observation, the orange line i.e. December holds the maximum sales.

PLOT A GRAPH OF MONTHLY ROSE SALES ACROSS YEARS.

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

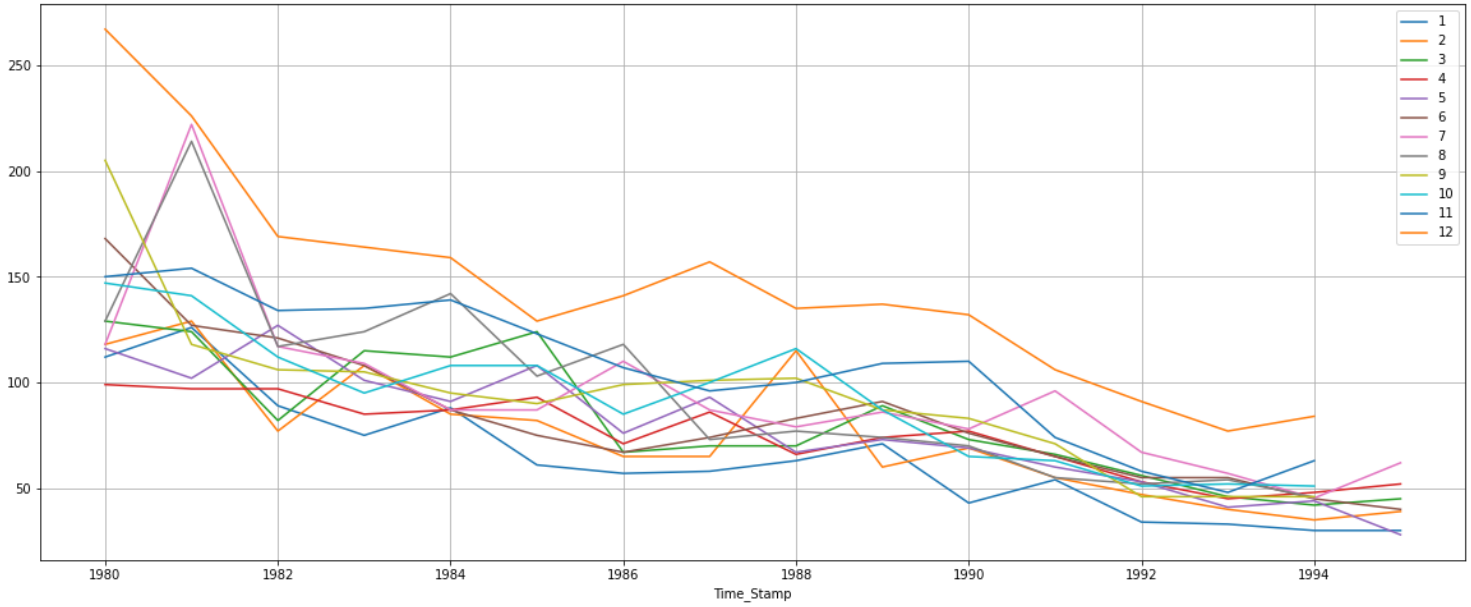


Figure 2.2(d)

- Similar observation, the orange line i.e. December holds the maximum sales.

PLOT THE EMPIRICAL CUMULATIVE DISTRIBUTION.

SPARKLING

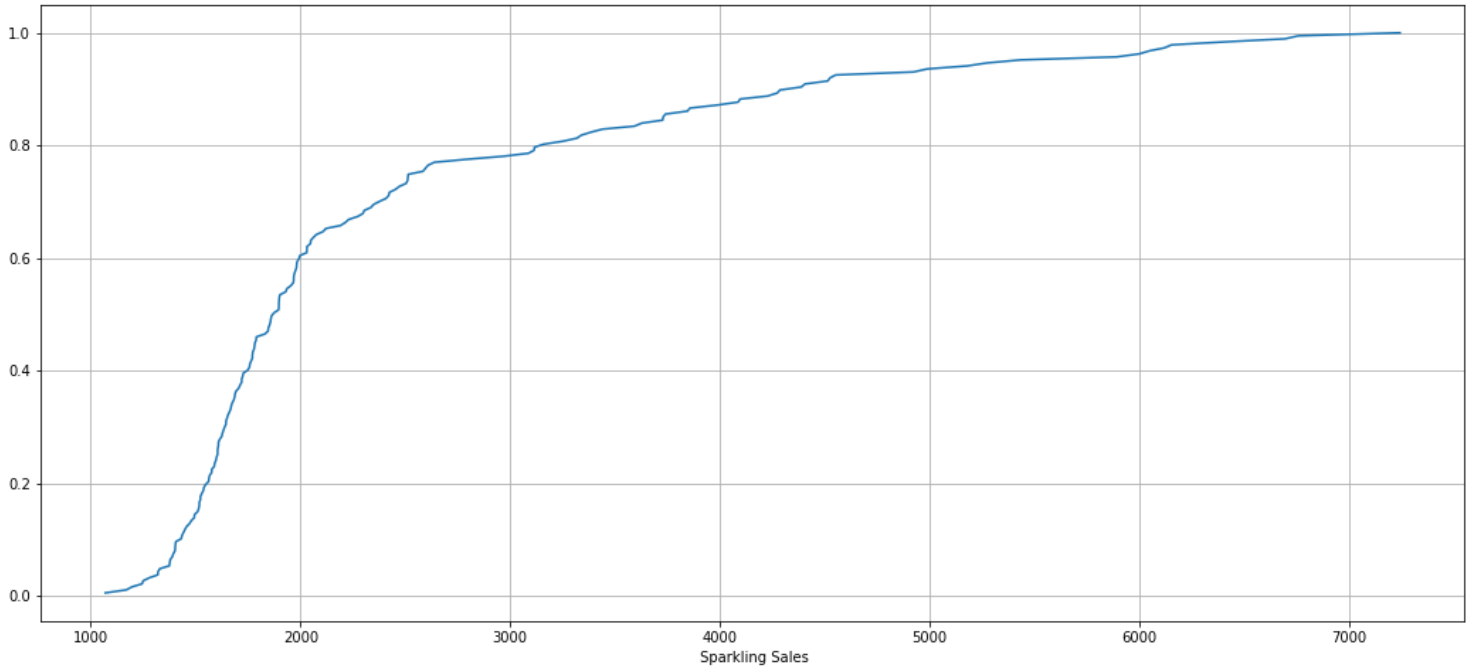


Figure 2.3(a)

ROSE

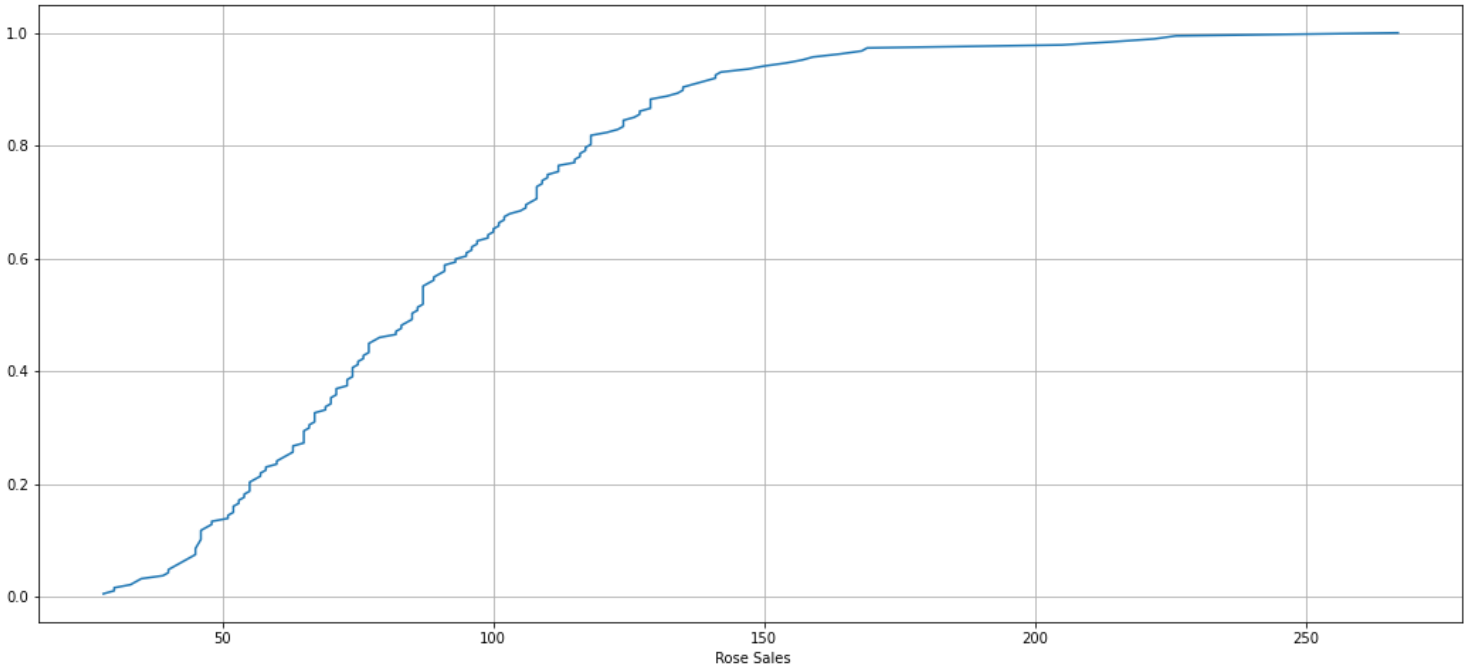


Figure 2.3(b)

PLOT THE AVERAGE SPARKLING SALES PER MONTH AND THE MONTH ON MONTH PERCENTAGE CHANGE OF SPARKLING SALES.

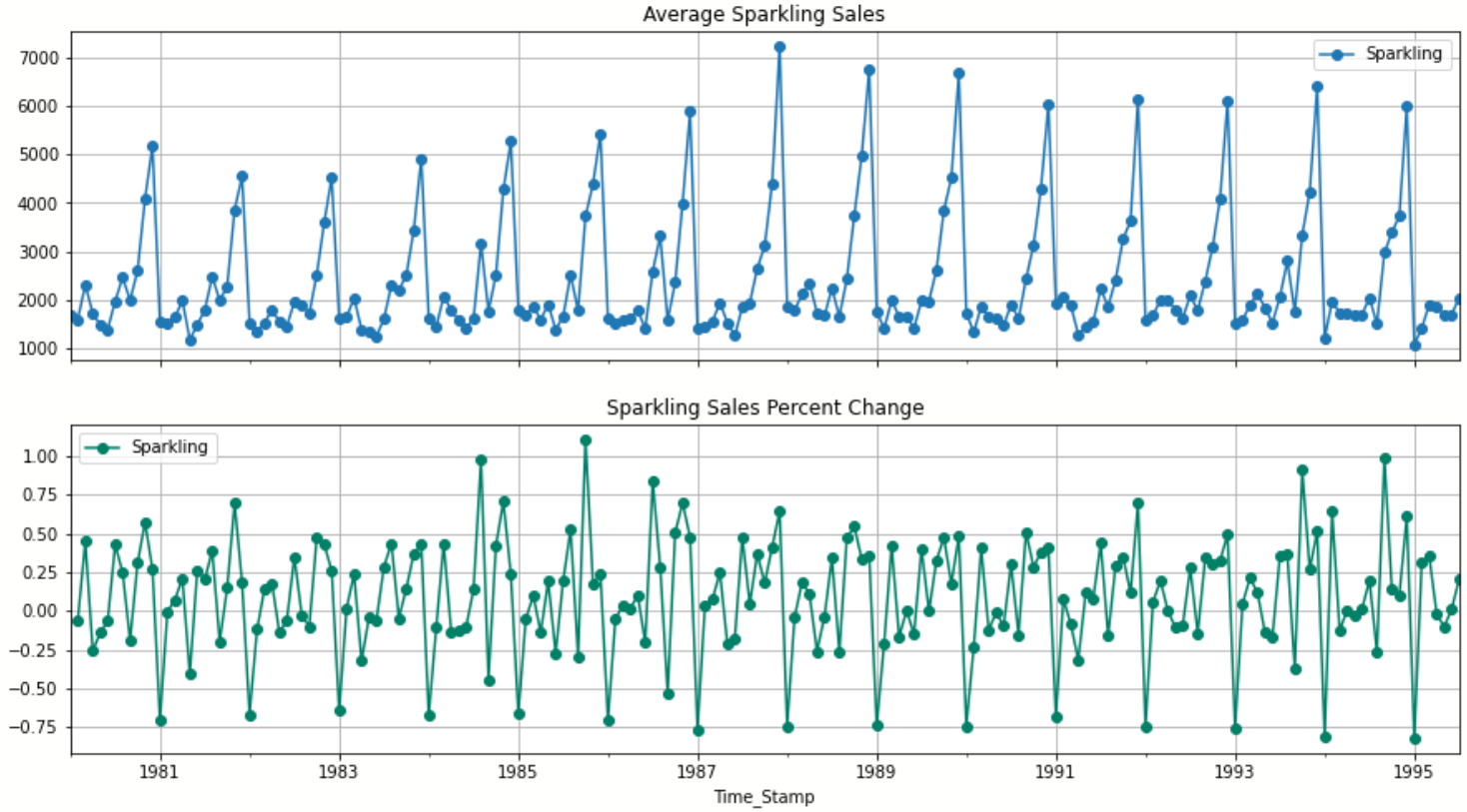


Figure 2.4(a)

- The maximum average sales recorded is 7000.
- The Percentage Change varies between the range of -1 to 1 which shows less fluctuation and is a good indication.

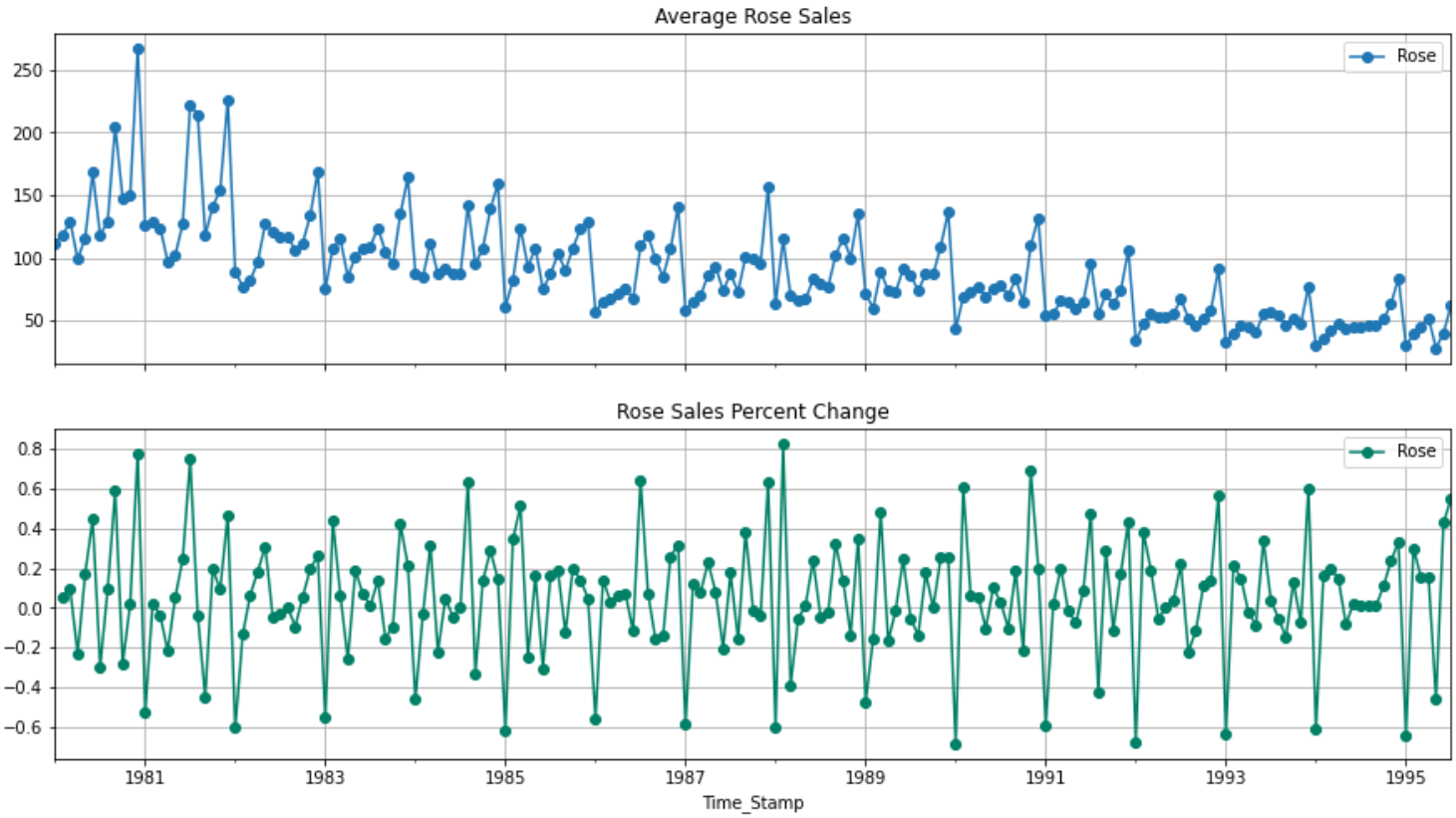


Figure 2.4(b)

- The maximum average sales recorded is around 275.
- The Percentage Change varies between the range of -0.7 to 0.8 which shows less fluctuation and is a good indication.

DECOMPOSE THE TIME SERIES AND PLOT THE DIFFERENT COMPONENTS.

ADDITIVE DECOMPOSITION

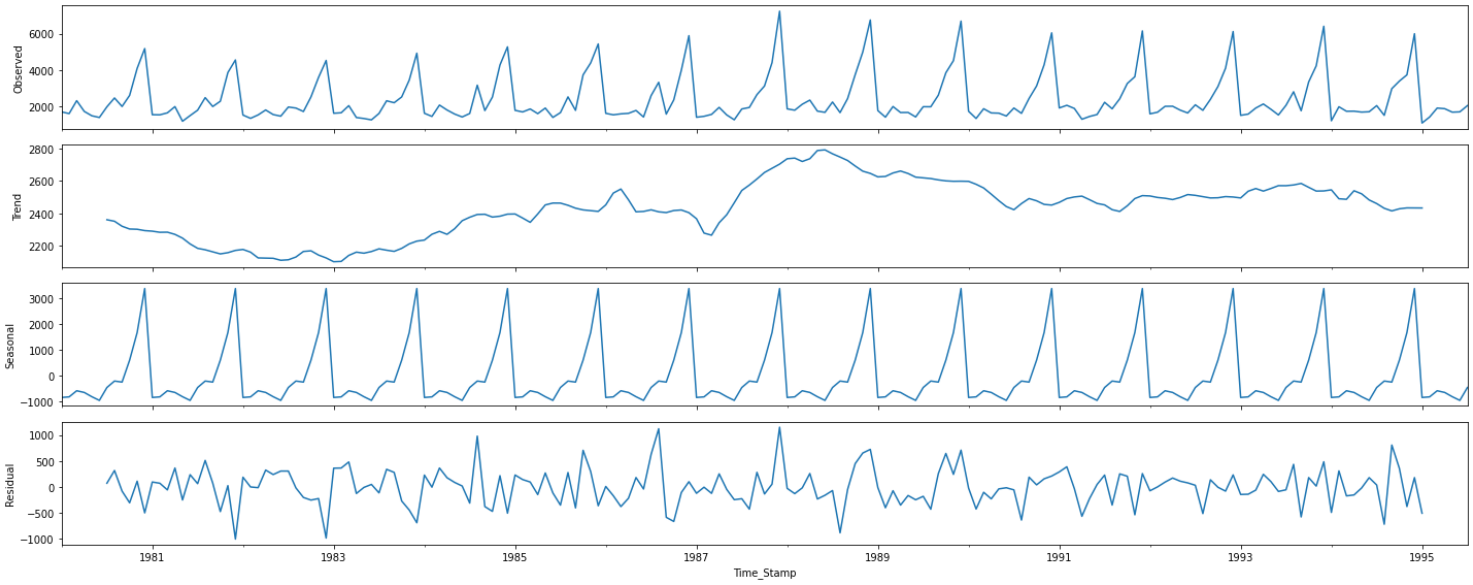


Figure 2.5(a)

- We can observe a slight upward trend present in our data.
- There's definitely seasonality present in our data which will be calculated later.
- The Residuals lie in the range of -1000 to 1000 therefore its unlikely we will choose this decomposition method.

```
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    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
Name: Sparkling, dtype: float64

Seasonality
 Time_Stamp
1980-01-31    -854.260599
1980-02-29    -830.350678
1980-03-31    -592.356630
1980-04-30    -658.490559
1980-05-31    -824.416154
1980-06-30    -967.434011
1980-07-31    -465.502265
1980-08-31    -214.332821
1980-09-30    -254.677265
1980-10-31     599.769957
1980-11-30    1675.067179
1980-12-31    3386.983846
Name: Sparkling, 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     70.835599
1980-08-31    315.999487
1980-09-30    -81.864401
1980-10-31   -307.353290
1980-11-30    109.891154
1980-12-31   -501.775513
Name: Sparkling, dtype: float64
```

MULTIPLICATIVE DECOMPOSITION

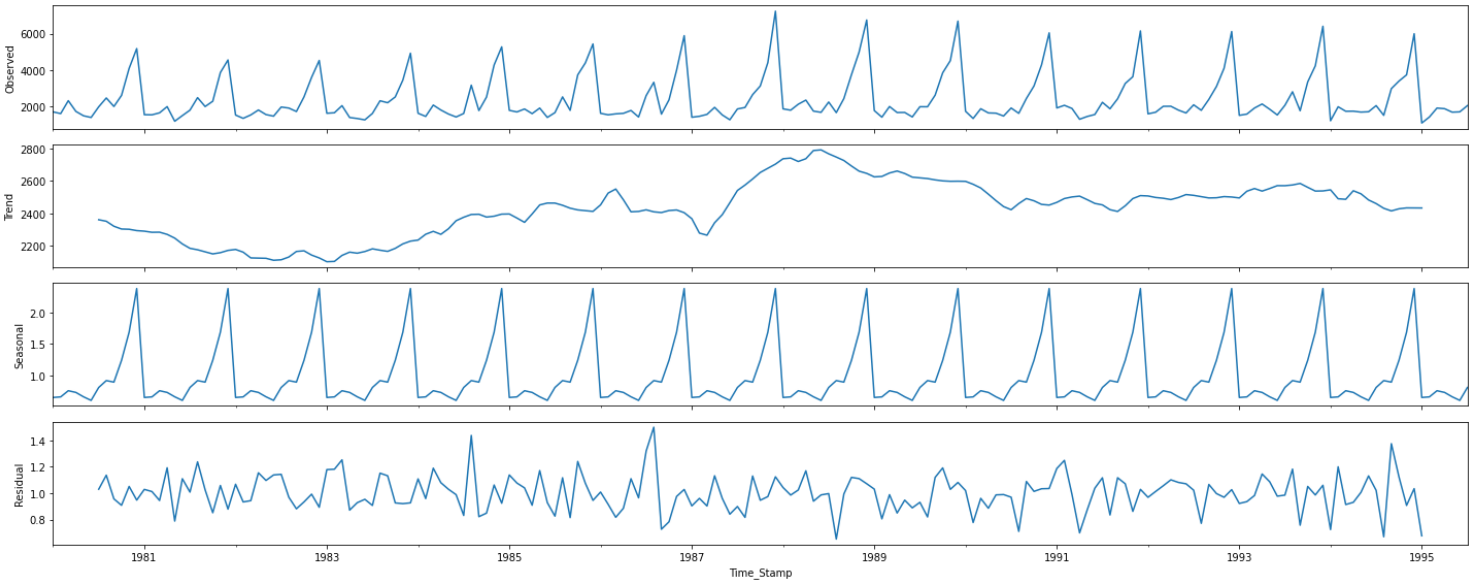


Figure 2.5(b)

- We can observe a slight upward trend present in our data.
- There's definitely seasonality present in our data which will be calculated later.
- The Residuals lie in the range of 0.6 to 1.6 therefore we will choose this decomposition method.

```
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    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
Name: Sparkling, dtype: float64

Seasonality
 Time_Stamp
1980-01-31    0.649843
1980-02-29    0.659214
1980-03-31    0.757440
1980-04-30    0.730351
1980-05-31    0.660609
1980-06-30    0.603468
1980-07-31    0.809164
1980-08-31    0.918822
1980-09-30    0.894367
1980-10-31    1.241789
1980-11-30    1.690158
1980-12-31    2.384776
Name: Sparkling, 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     1.029230
1980-08-31     1.135407
1980-09-30     0.955954
1980-10-31     0.907513
1980-11-30     1.050423
1980-12-31     0.946770
Name: Sparkling, dtype: float64
```

CHECK FOR STATIONARITY OF THE WHOLE TIME SERIES SPARKLING DATA.

To Check Stationarity we'll define our Null (Ho) and Alternate (Ha) hypothesis

Ho = The Time Series is Not Stationary

Ha = The Time Series is Stationary

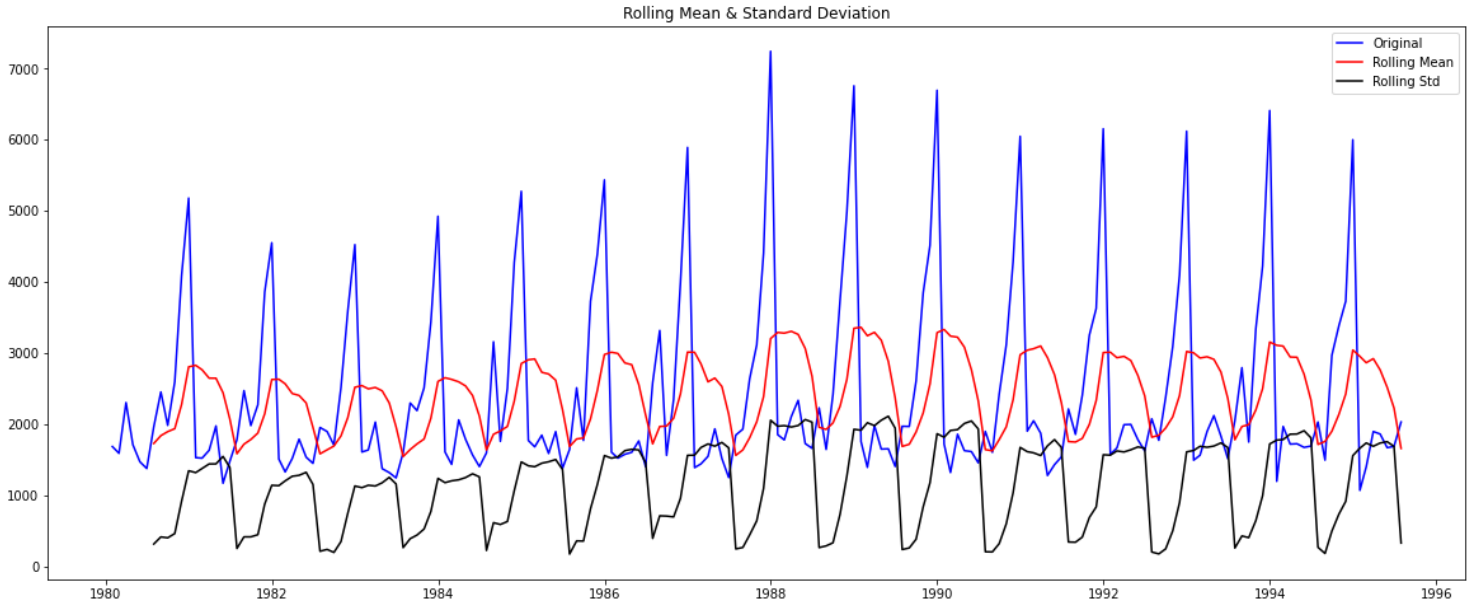


Figure 2.5(c)

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

- Since p-value > alpha (0.05). Ho holds true, i.e Time Series is not Stationary

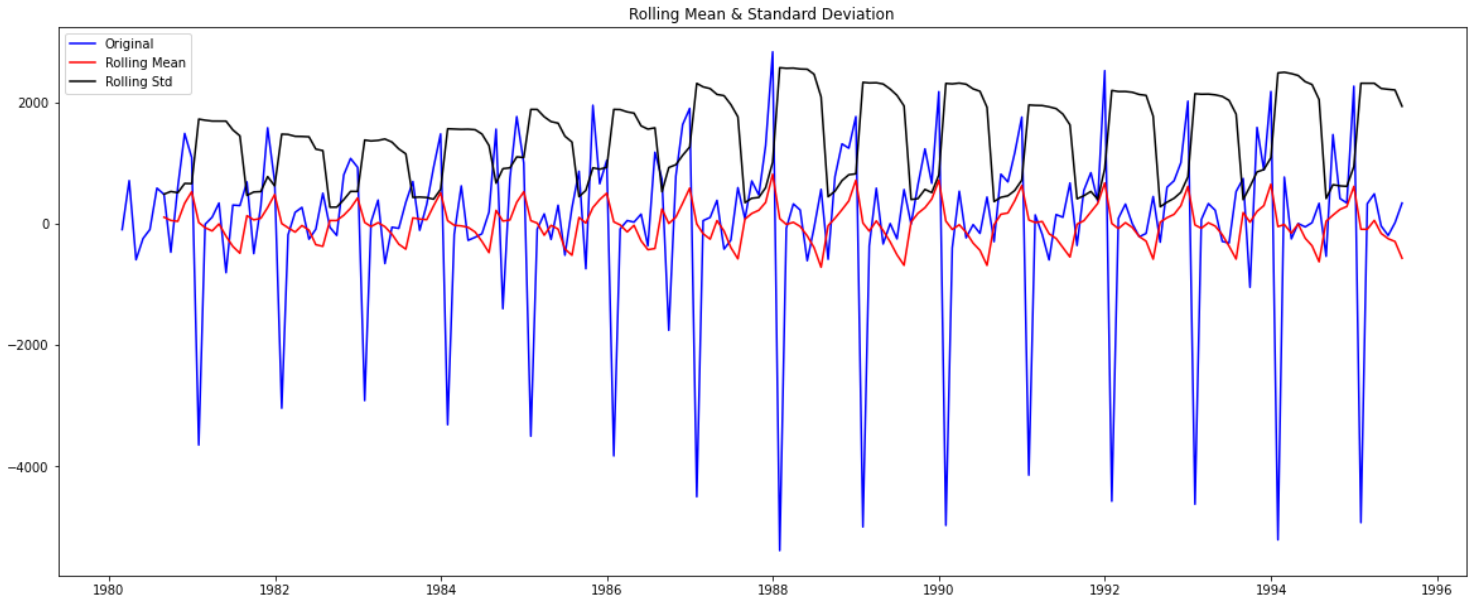


Figure 2.5(d)

```
Results of Dickey-Fuller Test:
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical Value (10%)             -2.575653
dtype: float64
```

- Since p-value < alpha (0.05). Ho is rejected, i.e Time Series is Stationary

ROSE DATASET

ADDITIONAL DECOPOSITION

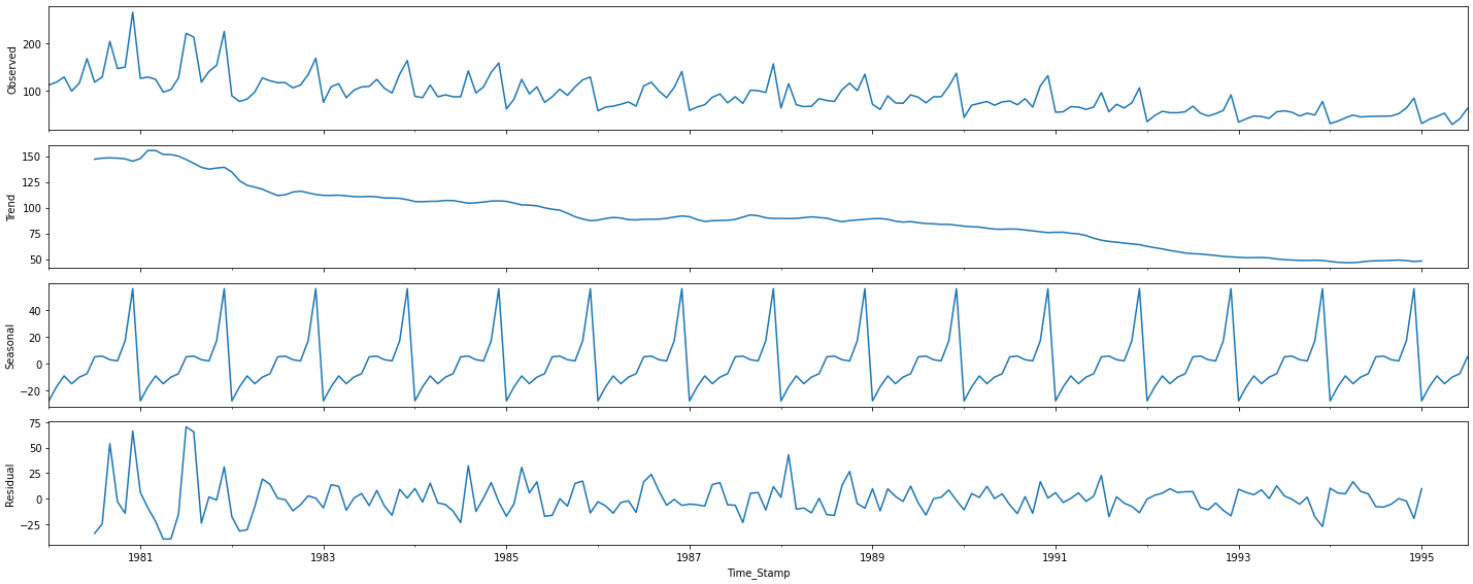


Figure 2.6(a)

- We can observe a downward trend present in our data.
- There's definitely seasonality present in our data which will be calculated later.
- The Residuals lie in the range of -25 to 75 therefore its unlikely we will choose this decomposition method.

```
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: Rose, 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: Rose, 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: Rose, dtype: float64
```

MULTIPLICATIVE DECOMPOSITION

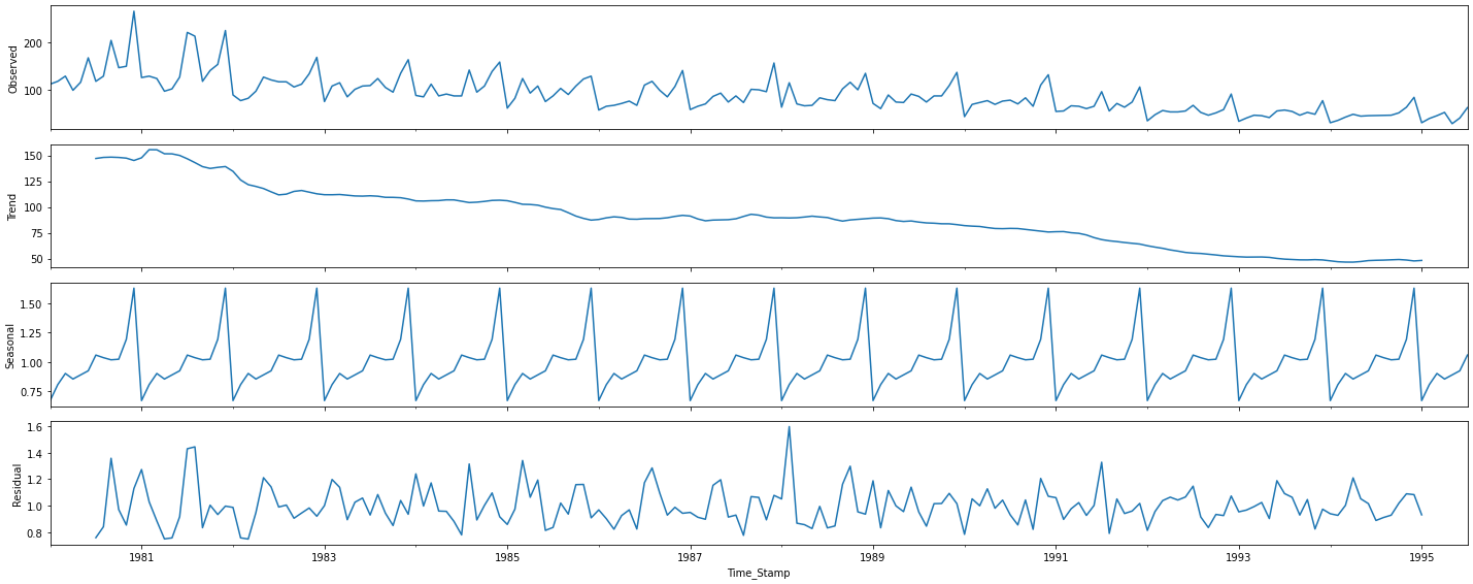


Figure 2.6(b)

- We can observe a downward trend present in our data.
- There's definitely seasonality present in our data which will be calculated later.
- The Residuals lie in the range of 0.6 to 1.6 therefore we will choose this decomposition method.

```
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: Rose, dtype: float64

Seasonality
 Time_Stamp
1980-01-31    0.670111
1980-02-29    0.806163
1980-03-31    0.901164
1980-04-30    0.854024
1980-05-31    0.889415
1980-06-30    0.923985
1980-07-31    1.058038
1980-08-31    1.035881
1980-09-30    1.017648
1980-10-31    1.022573
1980-11-30    1.192349
1980-12-31    1.628646
Name: Rose, 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     0.758258
1980-08-31     0.840720
1980-09-30     1.357674
1980-10-31     0.970771
1980-11-30     0.853378
1980-12-31     1.129646
Name: Rose, dtype: float64
```

CHECK FOR STATIONARITY OF THE WHOLE TIME SERIES ROSE DATA.

To Check Stationarity we'll define our Null (Ho) and Alternate (Ha) hypothesis

Ho = The Time Series is Not Stationary

Ha = The Time Series is Stationary

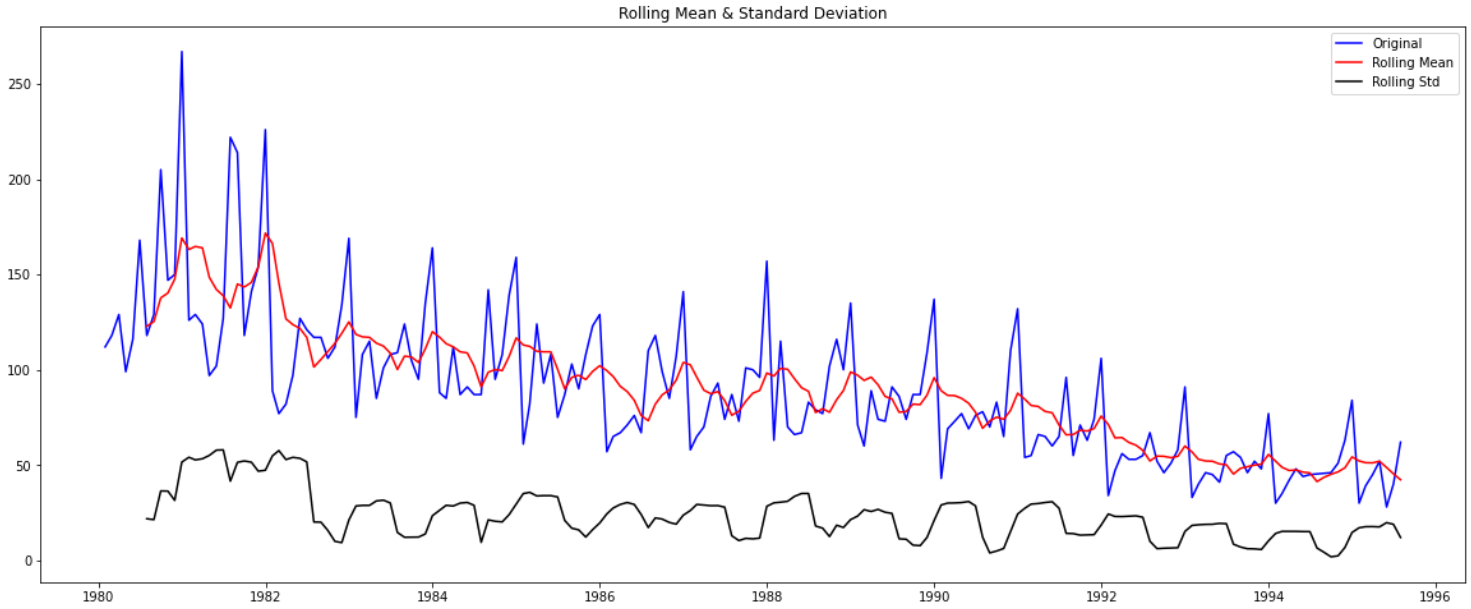


Figure 2.6(c)

```
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
```

- Since p-value > alpha (0.05). Ho holds true, i.e Time Series is not Stationary

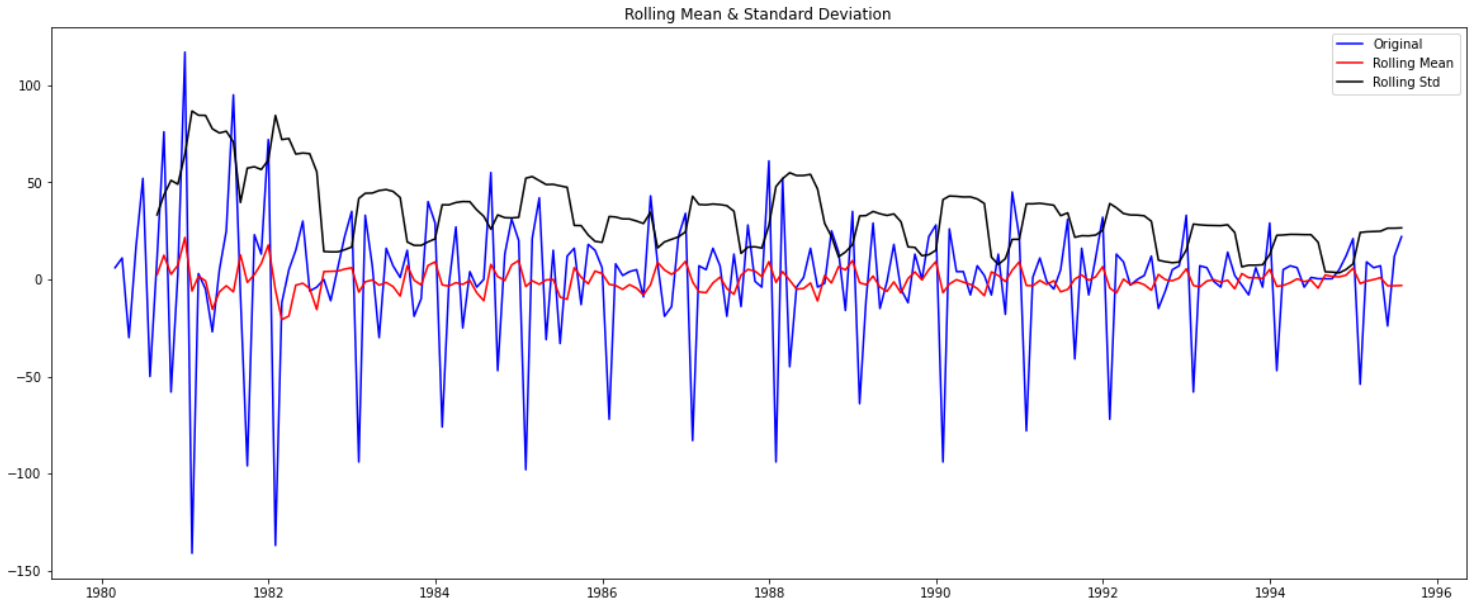


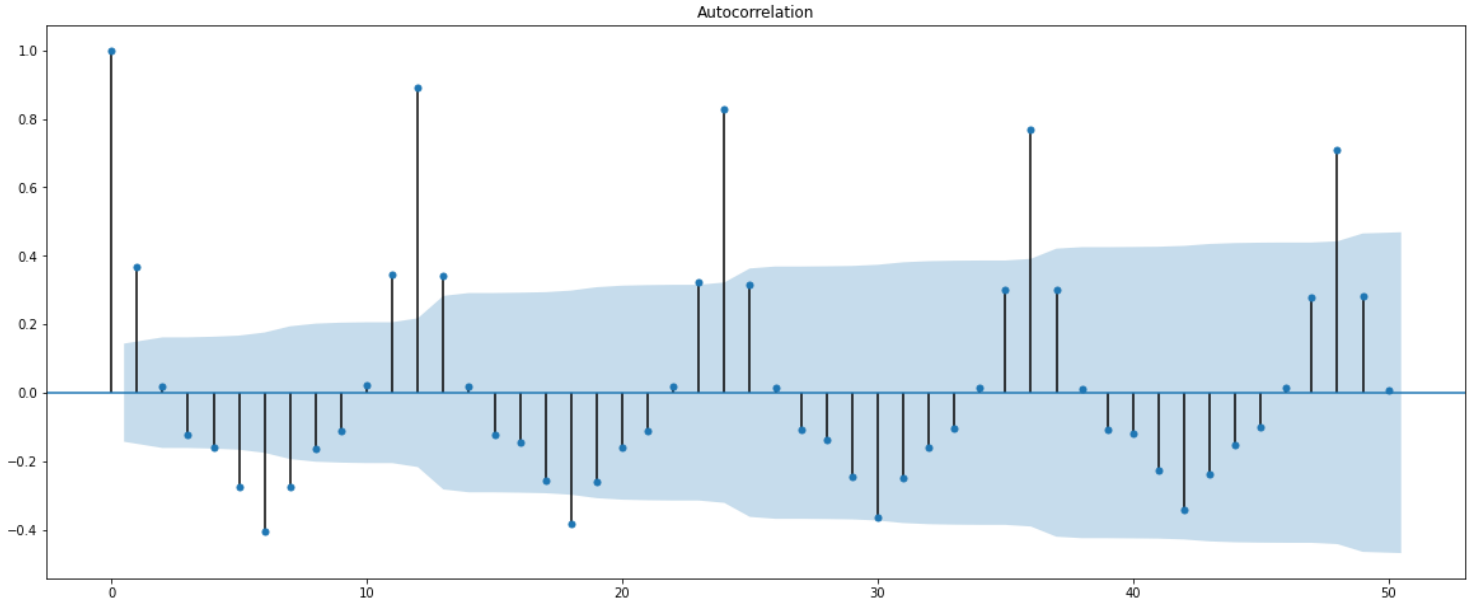
Figure 2.6(d)

```
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
```

- Since p-value < alpha (0.05). Ho is rejected, i.e Time Series is Stationary

PLOT THE AUTOCORRELATION AND THE PARTIAL AUTOCORRELATION FUNCTION PLOTS ON THE WHOLE DATA

Sparkling - ACF



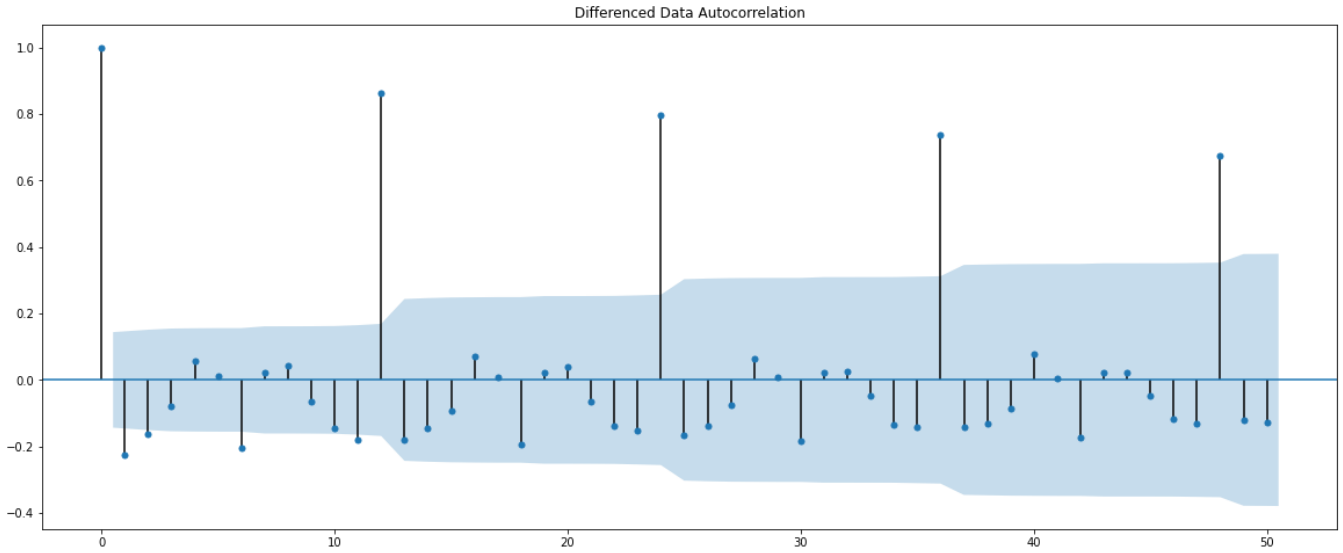
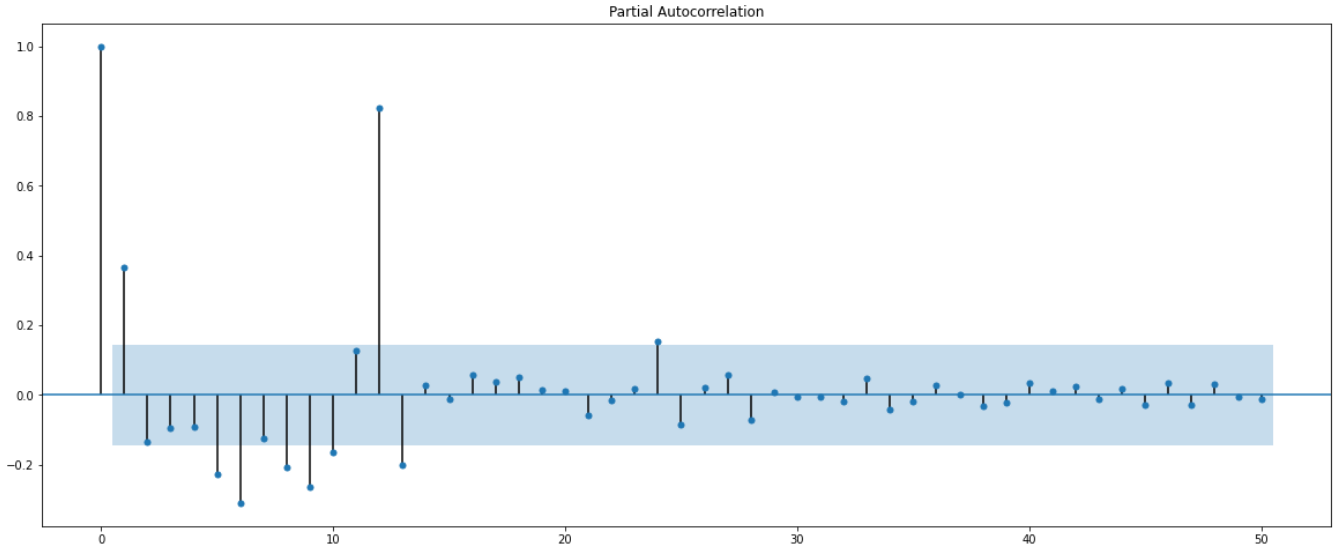


Figure 2.7(a)

PACF



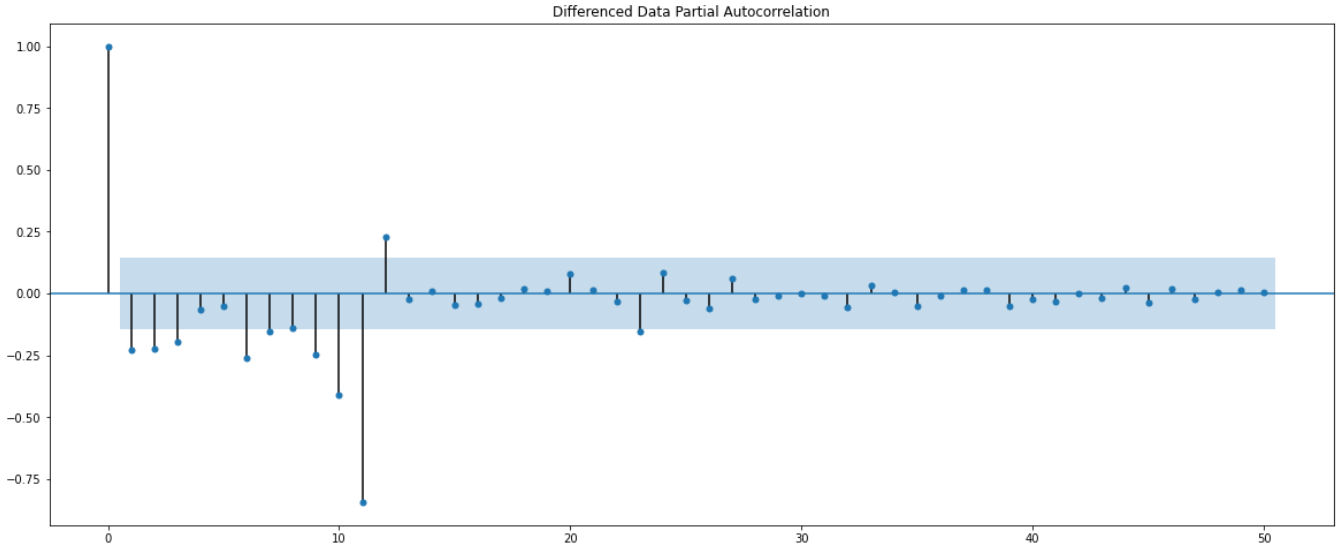


Figure 2.7(b)

The ACF cuts off after lag 2 and PACF after lag 3, therefore our p=3 and q=2 which will be used to build ARIMA/SARIMA models.

Rose - ACF

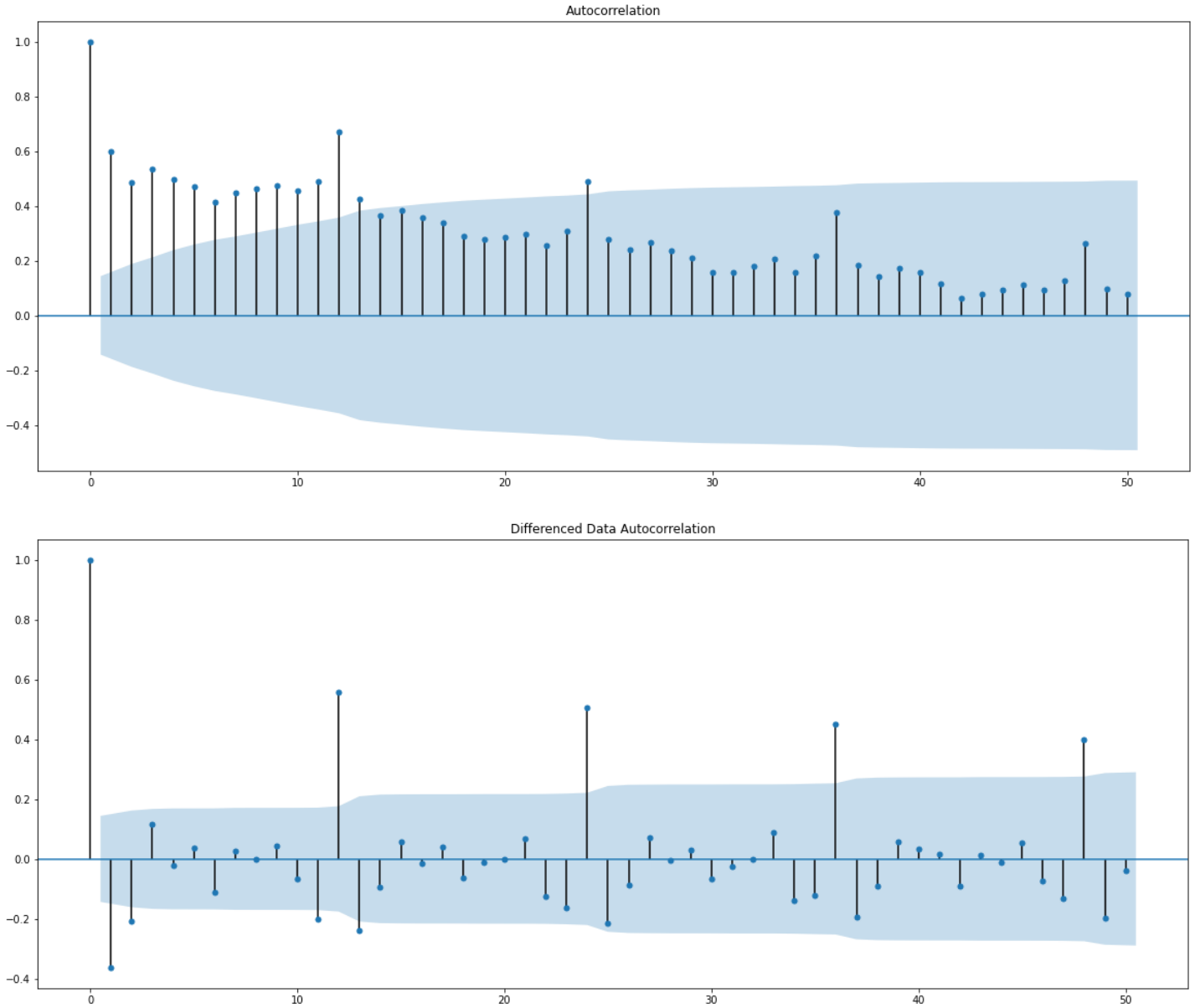
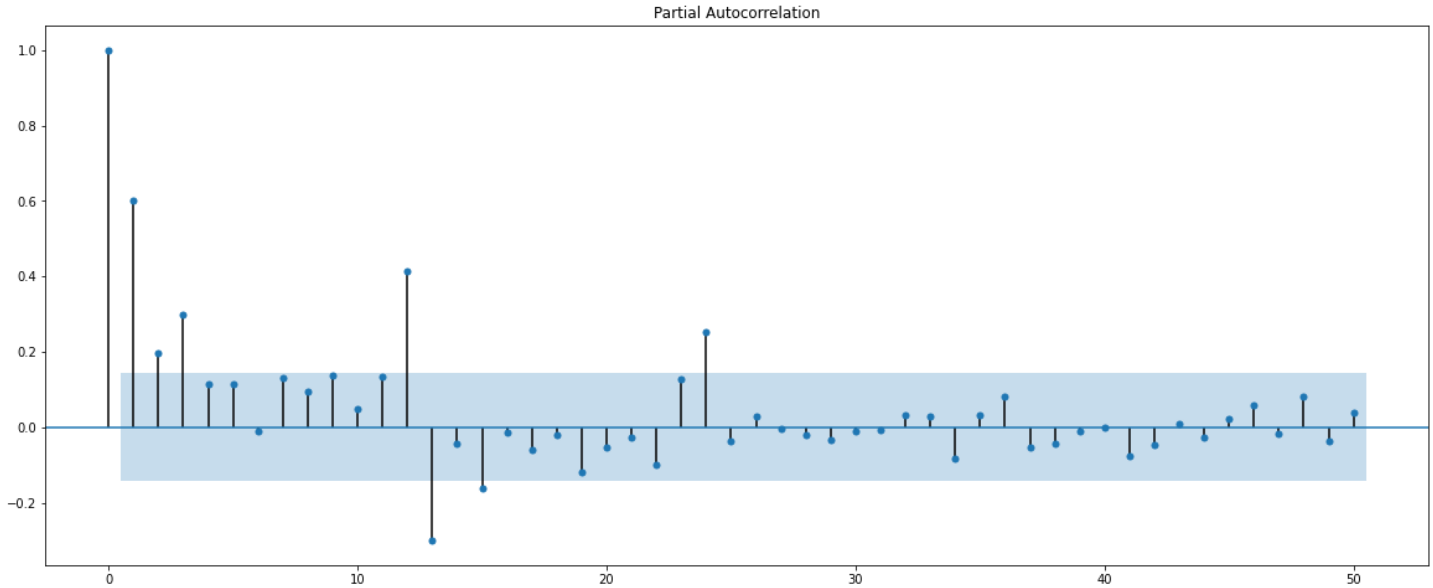


Figure 2.7(c)

PACF



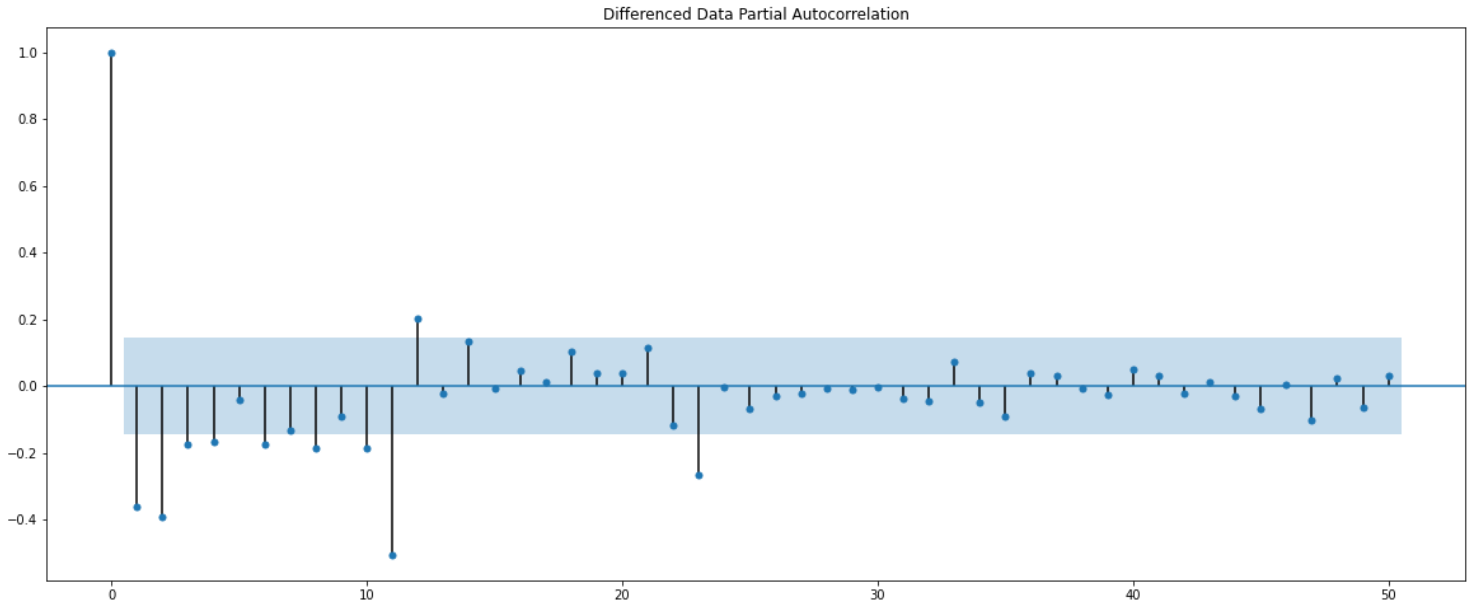


Figure 2.7(d)

The ACF cuts off after lag 2 and PACF after lag 4, therefore our p=3 and q=2 which will be used to build ARIMA/SARIMA models.

We have taken p as 3 and not 4 is to reduce model complexity.

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

First few rows of Training Data

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Last few rows of Training Data

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

First few rows of Test Data

Time_Stamp	Sparkling
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432

Last few rows of Test Data

Time_Stamp	Sparkling
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

*The total number of rows present in the **'training'** dataset above is : 132*
*The total number of rows present in the **'testing'** dataset above is : 55*

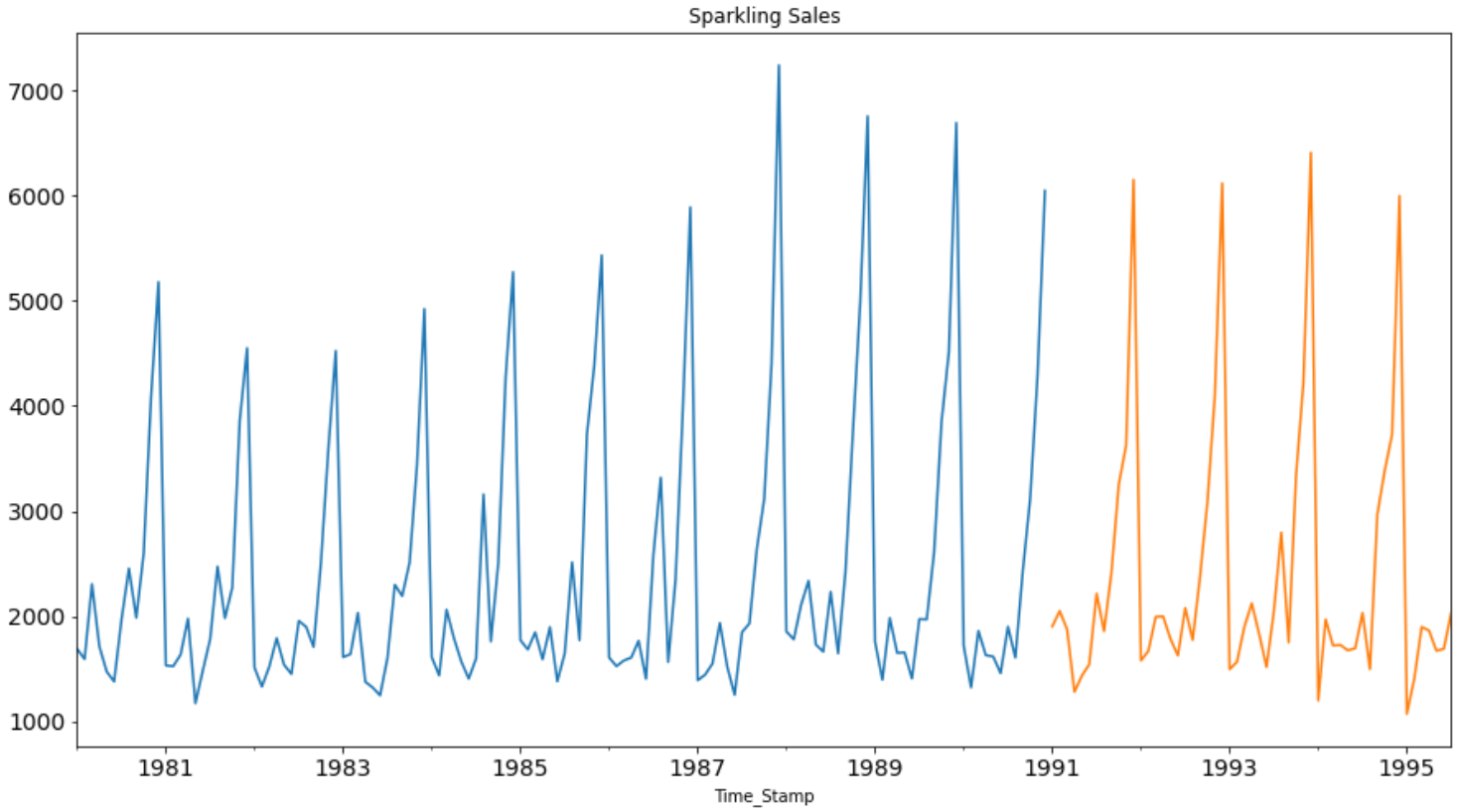


Figure 3(a)

First few rows of Training Data

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

Last few rows of Training Data

Time_Stamp	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

First few rows of Test Data

Time_Stamp	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Last few rows of Test Data

Time_Stamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

*The total number of rows present in the **'training'** dataset above is : 132*
*The total number of rows present in the **'testing'** dataset above is : 55*

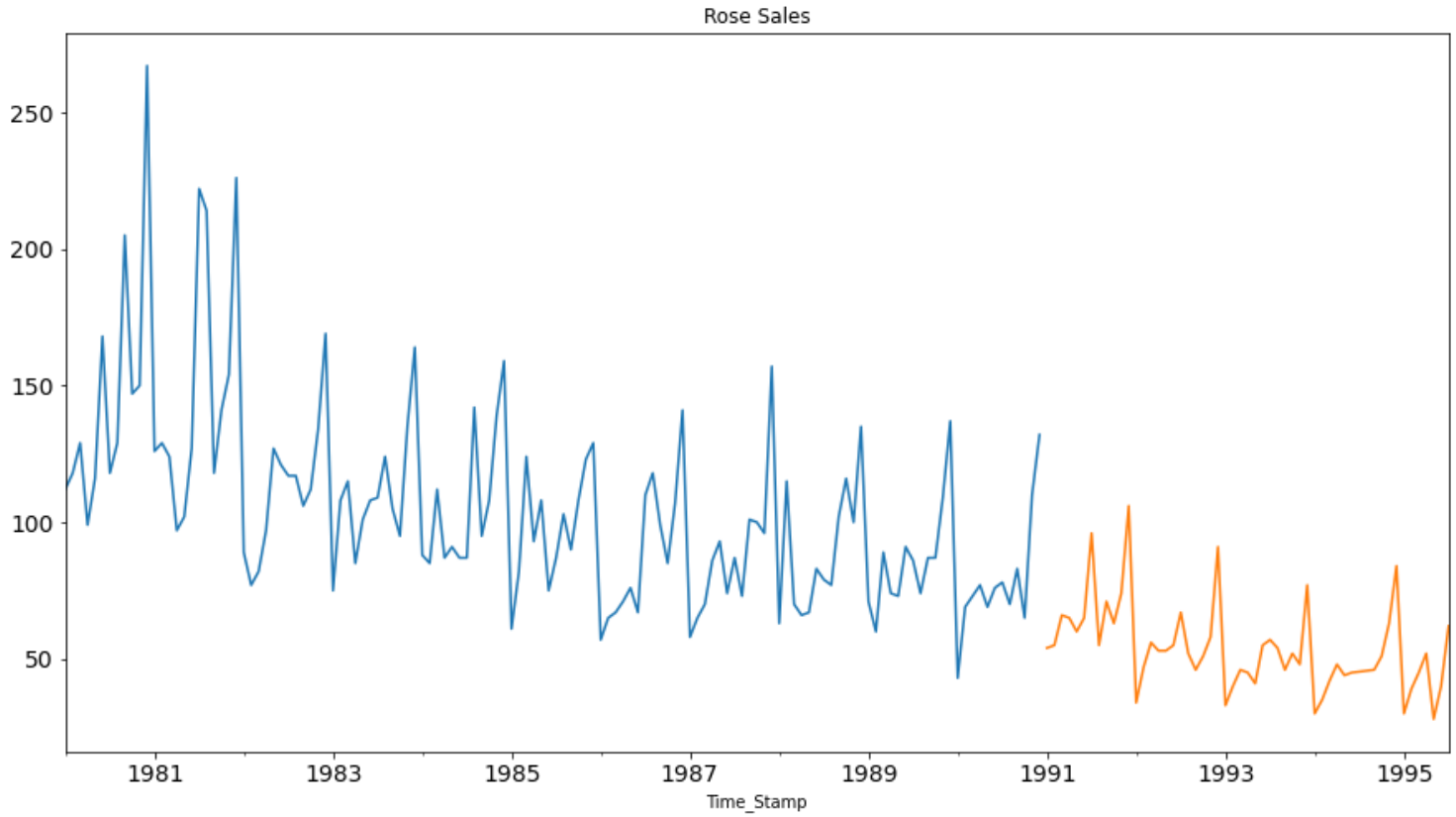


Figure 3(b)

4. BUILD VARIOUS EXPONENTIAL SMOOTHING MODELS ON THE TRAINING DATA AND EVALUATE THE MODEL USING RMSE ON THE TEST DATA.

BUILDING DIFFERENT MODELS AND COMPARING THE ACCURACY METRICS.

- MODEL 1: LINEAR REGRESSION

SPARKLING

```
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, 34, 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, 183,
184, 185, 186, 187]
```

We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

First few rows of Training Data

Time_Stamp	Sparkling	time
1980-01-31	1686	1
1980-02-29	1591	2
1980-03-31	2304	3
1980-04-30	1712	4
1980-05-31	1471	5

Last few rows of Training Data

Time_Stamp	Sparkling	time
1990-08-31	1605	128
1990-09-30	2424	129
1990-10-31	3116	130
1990-11-30	4286	131
1990-12-31	6047	132

First few rows of Test Data

Time_Stamp	Sparkling	time
1991-01-31	1902	133
1991-02-28	2049	134
1991-03-31	1874	135
1991-04-30	1279	136
1991-05-31	1432	137

Last few rows of Test Data

Time_Stamp	Sparkling	time
1995-03-31	1897	183
1995-04-30	1862	184
1995-05-31	1670	185
1995-06-30	1688	186
1995-07-31	2031	187

Now that our training and test data has been modified, let us go ahead use $LinearRegression$ to build the model on the training data and test the model on the test data.

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

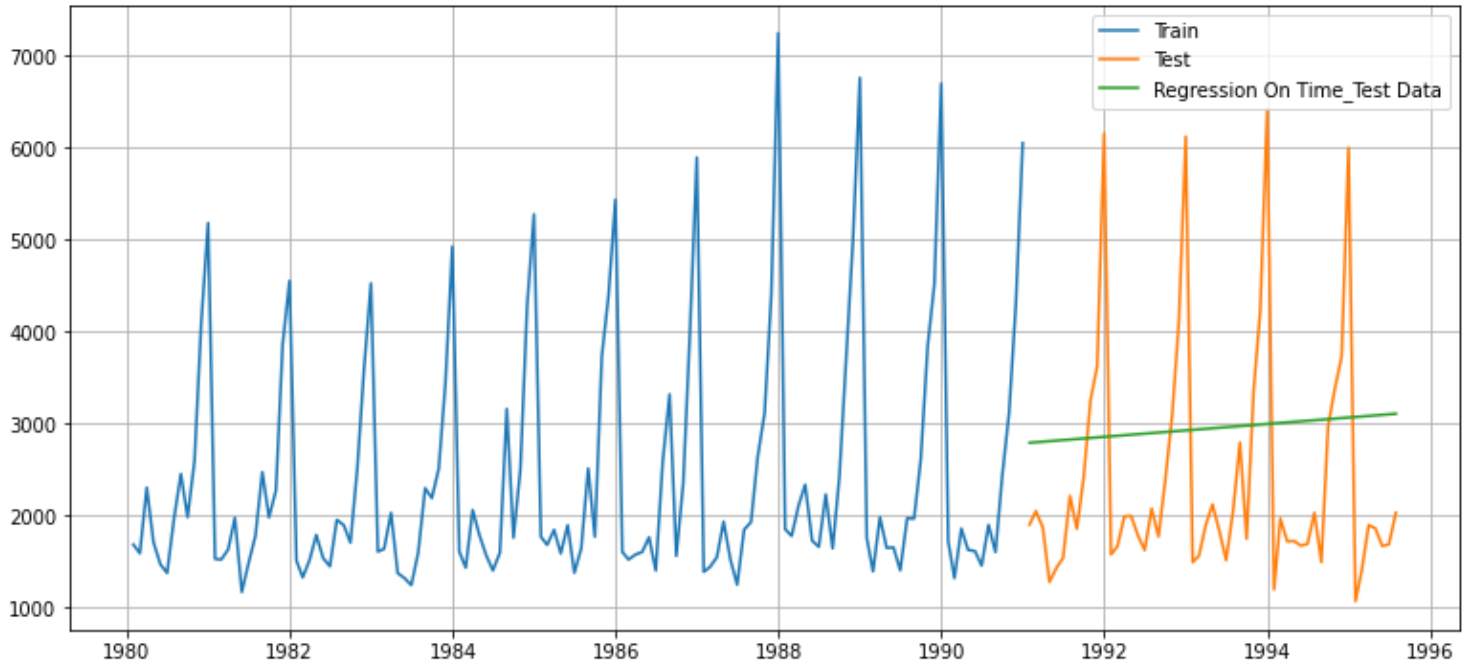


Figure 4.1(a)

DEFINING THE ACCURACY METRICS.

MODEL EVALUATION

- ***For RegressionOnTime forecast on the Test Data, RMSE is 1389.135***

ROSE

```
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, 34, 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, 183, 184, 185, 186, 187]
```

We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

First few rows of Training Data

Time_Stamp	Rose	time
1980-01-31	112.0	1
1980-02-29	118.0	2
1980-03-31	129.0	3
1980-04-30	99.0	4
1980-05-31	116.0	5

Last few rows of Training Data

Time_Stamp	Rose	time
1990-08-31	70.0	128
1990-09-30	83.0	129
1990-10-31	65.0	130
1990-11-30	110.0	131
1990-12-31	132.0	132

First few rows of Test Data

Time_Stamp	Rose	time
1991-01-31	54.0	133
1991-02-28	55.0	134
1991-03-31	66.0	135
1991-04-30	65.0	136
1991-05-31	60.0	137

Last few rows of Test Data

Time_Stamp	Rose	time
1995-03-31	45.0	183
1995-04-30	52.0	184
1995-05-31	28.0	185
1995-06-30	40.0	186
1995-07-31	62.0	187

Now that our training and test data has been modified, let us go ahead use $LinearRegression$ to build the model on the training data and test the model on the test data.

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

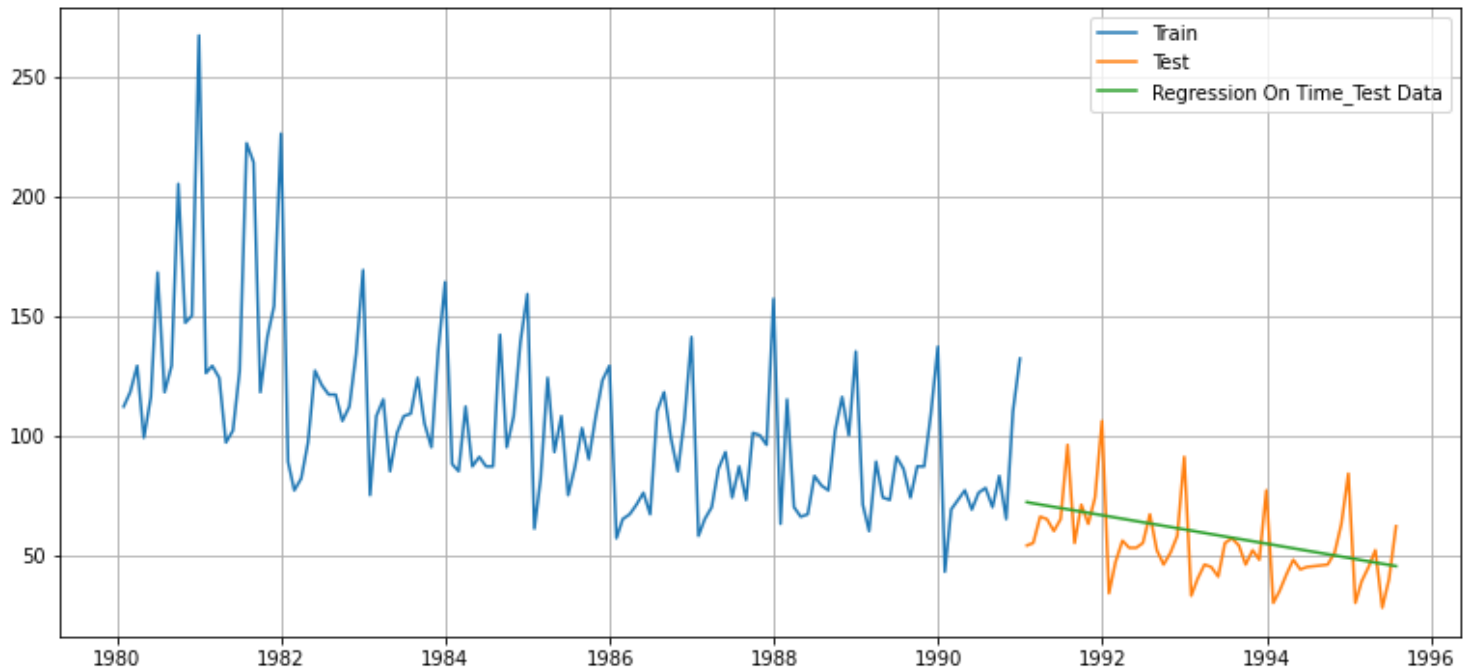


Figure 4.1(b)

DEFINING THE ACCURACY METRICS.

MODEL EVALUATION

- ***For RegressionOnTime forecast on the Test Data, RMSE is <u>15.269</u>***

- ## MODEL 2: NAIVE APPROACH: $\hat{y}\ t{+}1{=}yt$

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

SPARKLING

Tail of the Training set.

	Sparkling
Time_Stamp	
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

Head of the Naïve Model.

```
Time_Stamp
1991-01-31    6047
1991-02-28    6047
1991-03-31    6047
1991-04-30    6047
1991-05-31    6047
Name: naive, dtype: int64
```

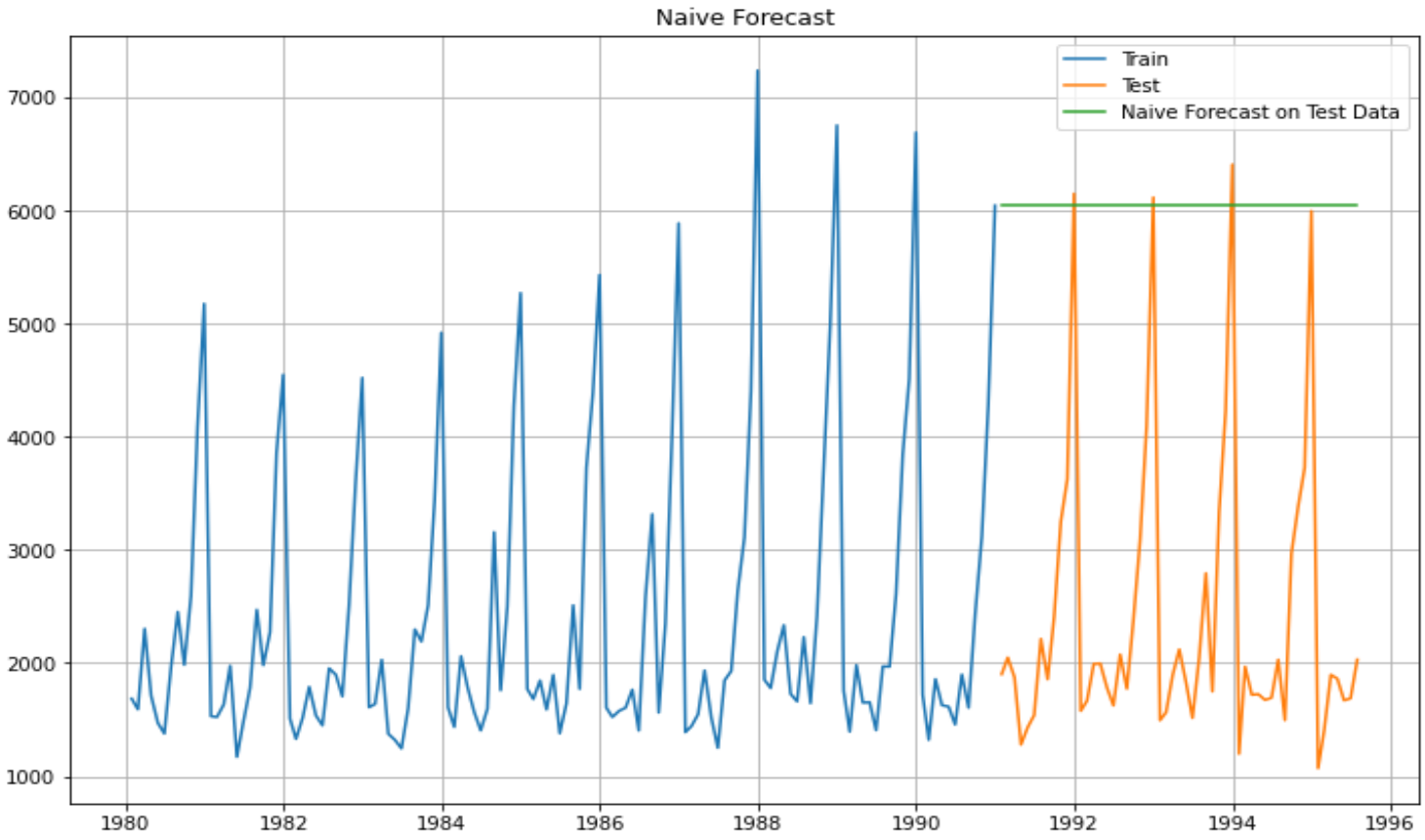


Figure 4.2(a)

MODEL EVALUATION

- **For NaïveModel forecast on the Test Data, RMSE is 3864.279**

ROSE

Tail of the Training set.

	Rose
Time_Stamp	
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

Head of the Naïve Model.

```
Time_Stamp
1991-01-31    132.0
1991-02-28    132.0
1991-03-31    132.0
1991-04-30    132.0
1991-05-31    132.0
```

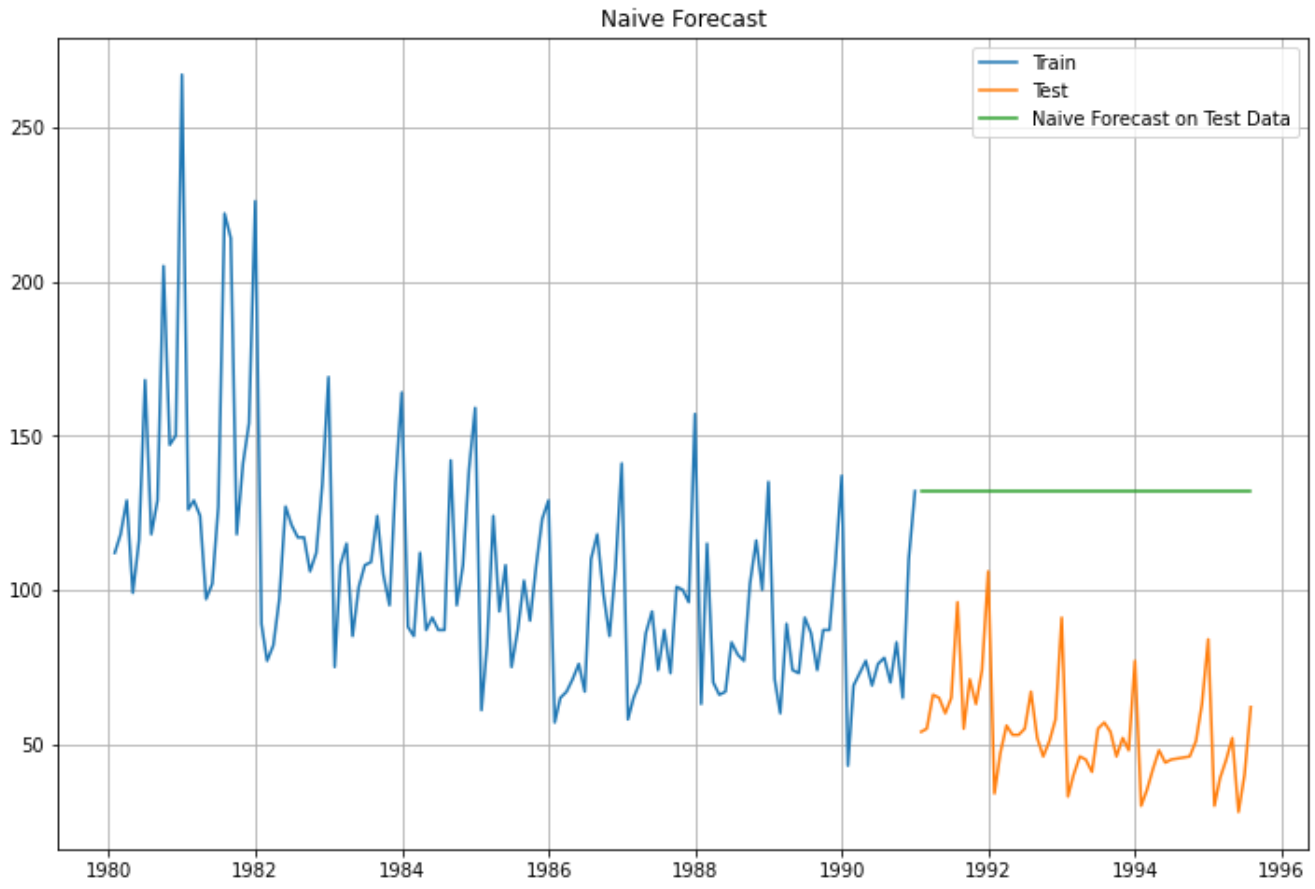


Figure 4.2(b)

MODEL EVALUATION

- **For NaiveModel forecast on the Test Data, RMSE is 79.719**

• METHOD 3: SIMPLE AVERAGE

For this particular simple average method, we will forecast by using the average of the training values

SPARKLING

Time_Stamp	Sparkling	mean_forecast
1991-01-31	1902	2403.780303
1991-02-28	2049	2403.780303
1991-03-31	1874	2403.780303
1991-04-30	1279	2403.780303
1991-05-31	1432	2403.780303

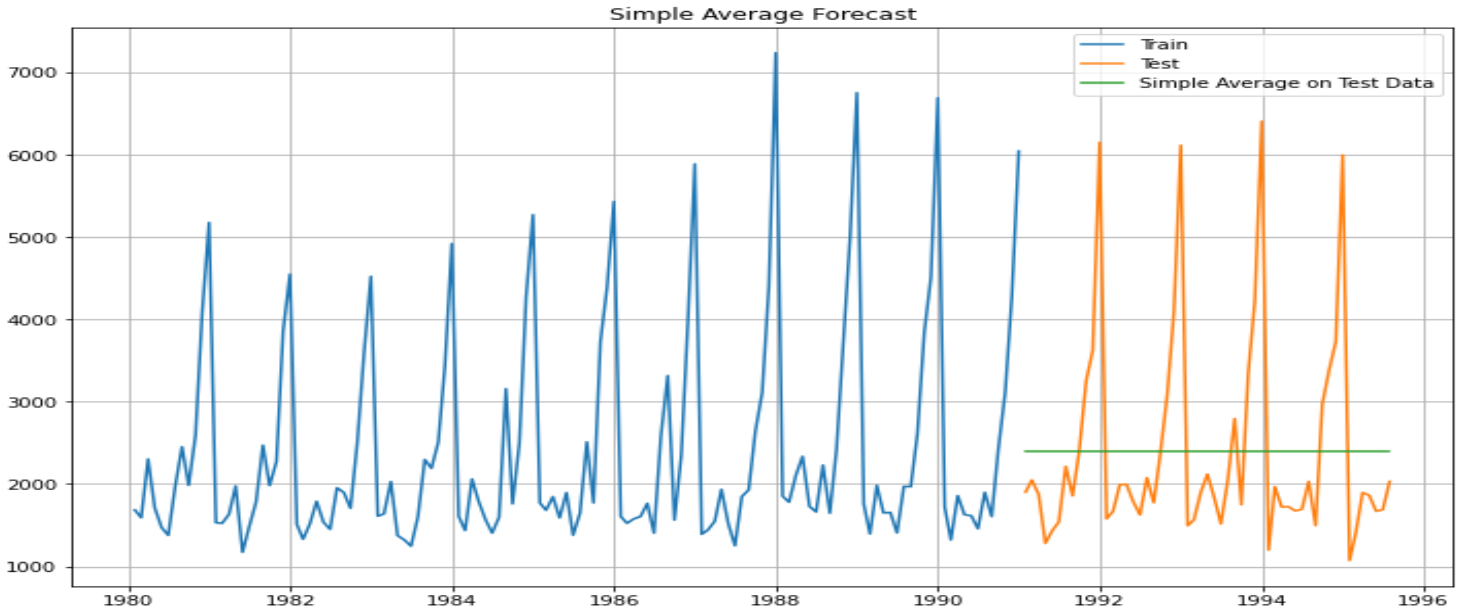


Figure 4.3(a)

MODEL EVALUATION

- ***For Simple Average forecast on the Test Data, RMSE is 1275.082***

ROSE

Time_Stamp	Rose	mean_forecast
1991-01-31	54.0	104.939394
1991-02-28	55.0	104.939394
1991-03-31	66.0	104.939394
1991-04-30	65.0	104.939394
1991-05-31	60.0	104.939394

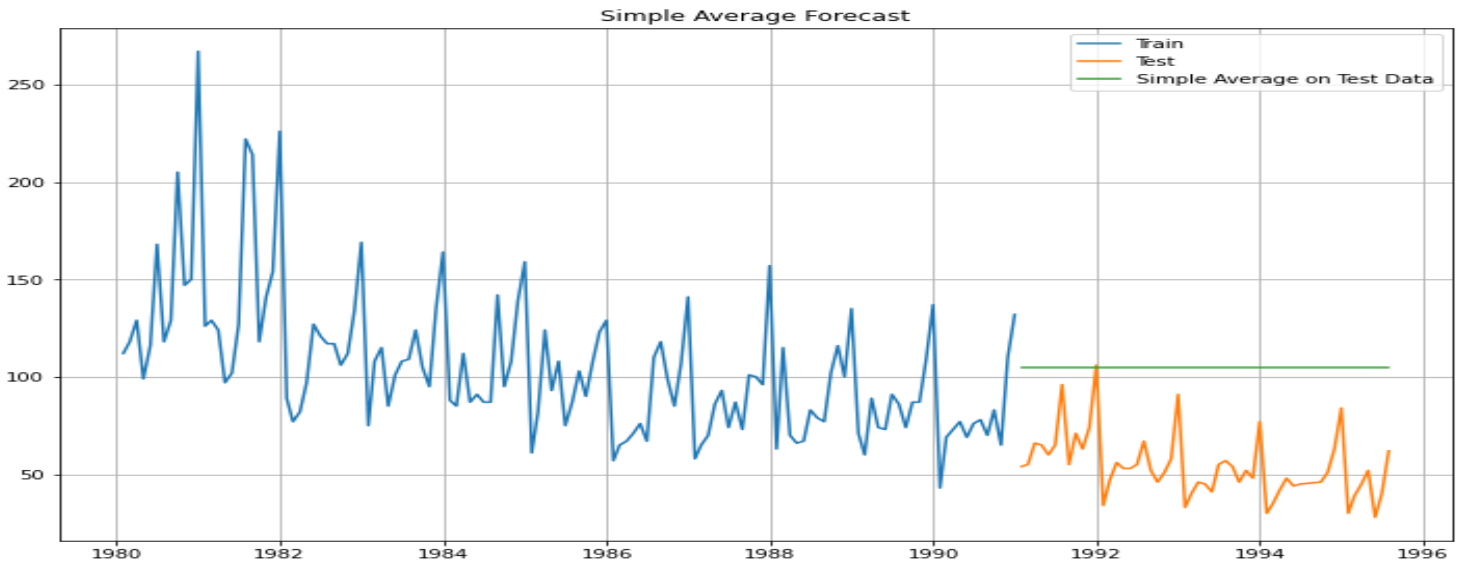


Figure 4.3(b)

Model Evaluation

- **For Simple Average forecast on the Test Data, RMSE is 53.461**

• Method 4: Moving Average(MA)

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

For Moving Average, we are going to average over the entire data.

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Trailing moving averages

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	1686	NaN	NaN	NaN	NaN
1980-02-29	1591	1638.5	NaN	NaN	NaN
1980-03-31	2304	1947.5	NaN	NaN	NaN
1980-04-30	1712	2008.0	1823.25	NaN	NaN
1980-05-31	1471	1591.5	1769.50	NaN	NaN
1980-06-30	1377	1424.0	1716.00	1690.166667	NaN
1980-07-31	1966	1671.5	1631.50	1736.833333	NaN
1980-08-31	2453	2209.5	1816.75	1880.500000	NaN
1980-09-30	1984	2218.5	1945.00	1827.166667	1838.222222
1980-10-31	2596	2290.0	2249.75	1974.500000	1939.333333

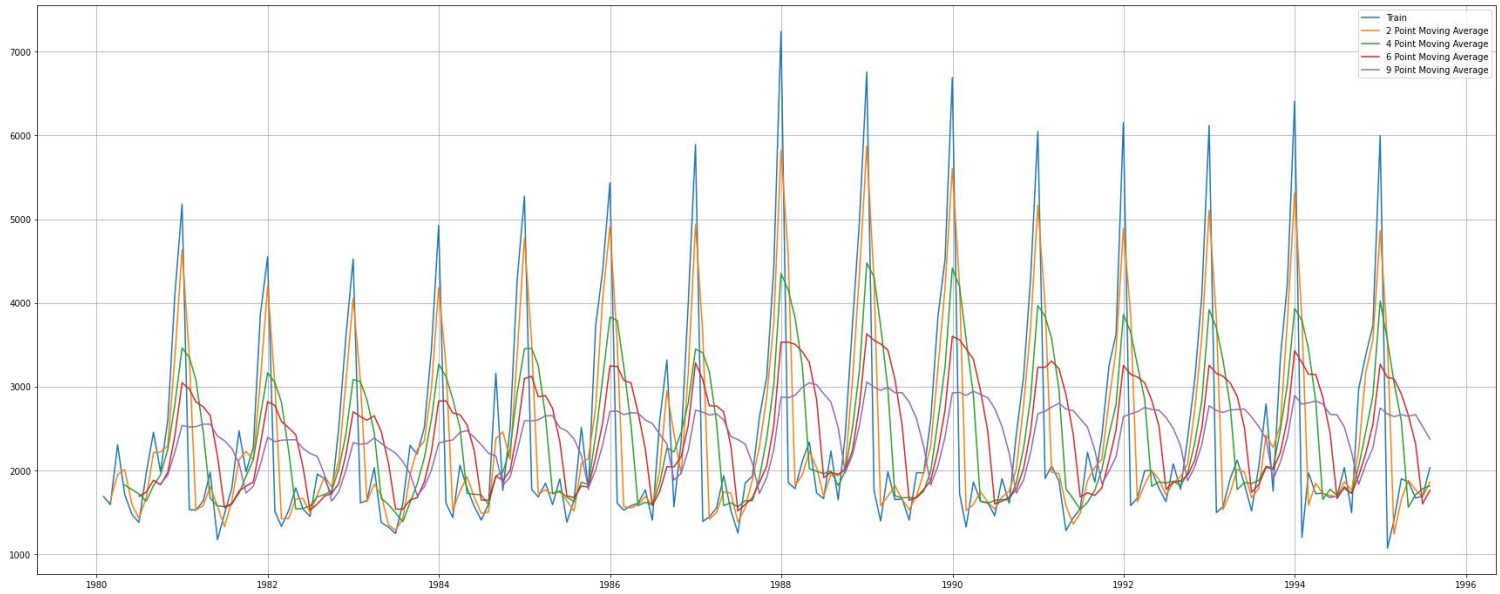


Figure 4.4(a)

Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.

- TRAINING SET

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	1686	NaN	NaN	NaN	NaN
1980-02-29	1591	1638.5	NaN	NaN	NaN
1980-03-31	2304	1947.5	NaN	NaN	NaN
1980-04-30	1712	2008.0	1823.25	NaN	NaN
1980-05-31	1471	1591.5	1769.50	NaN	NaN

- TESTING SET

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1991-01-31	1902	3974.5	3837.75	3230.000000	2705.666667
1991-02-28	2049	1975.5	3571.00	3304.000000	2753.888889
1991-03-31	1874	1961.5	2968.00	3212.333333	2800.222222
1991-04-30	1279	1576.5	1776.00	2906.166667	2731.333333
1991-05-31	1432	1355.5	1658.50	2430.500000	2712.111111

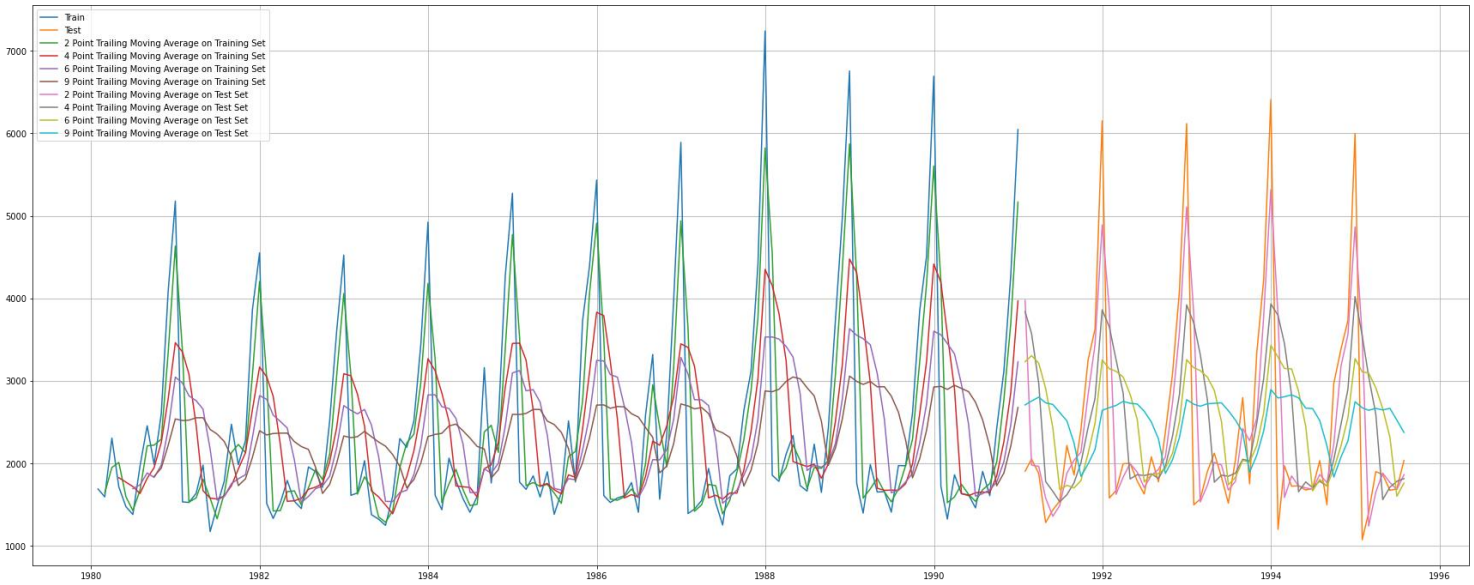


Figure 4.4(b)

MODEL EVALUATION

Done only on the test data

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

ROSE

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

TRAILING MOVING AVERAGES

Time_Stamp	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.50	NaN	NaN
1980-05-31	116.0	107.5	115.50	NaN	NaN
1980-06-30	168.0	142.0	128.00	123.666667	NaN
1980-07-31	118.0	143.0	125.25	124.666667	NaN
1980-08-31	129.0	123.5	132.75	126.500000	NaN
1980-09-30	205.0	167.0	155.00	139.166667	132.666667
1980-10-31	147.0	176.0	149.75	147.166667	136.555556

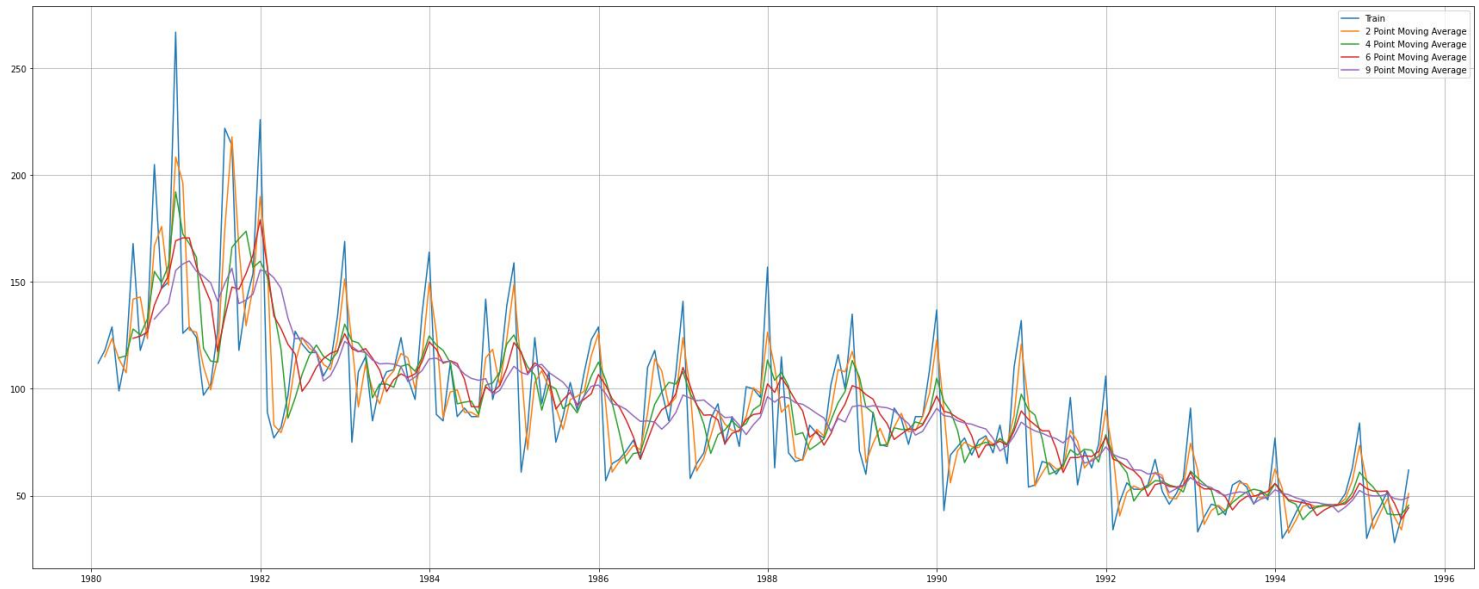


Figure 4.4(c)

Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.

- TRAINING SET

Time_Stamp	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.5	NaN	NaN
1980-05-31	116.0	107.5	115.5	NaN	NaN

- TESTING SET

Time_Stamp	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1991-01-31	54.0	93.0	90.25	85.666667	81.888889
1991-02-28	55.0	54.5	87.75	83.166667	80.333333
1991-03-31	66.0	60.5	76.75	80.333333	79.222222
1991-04-30	65.0	65.5	60.00	80.333333	77.777778
1991-05-31	60.0	62.5	61.50	72.000000	76.666667

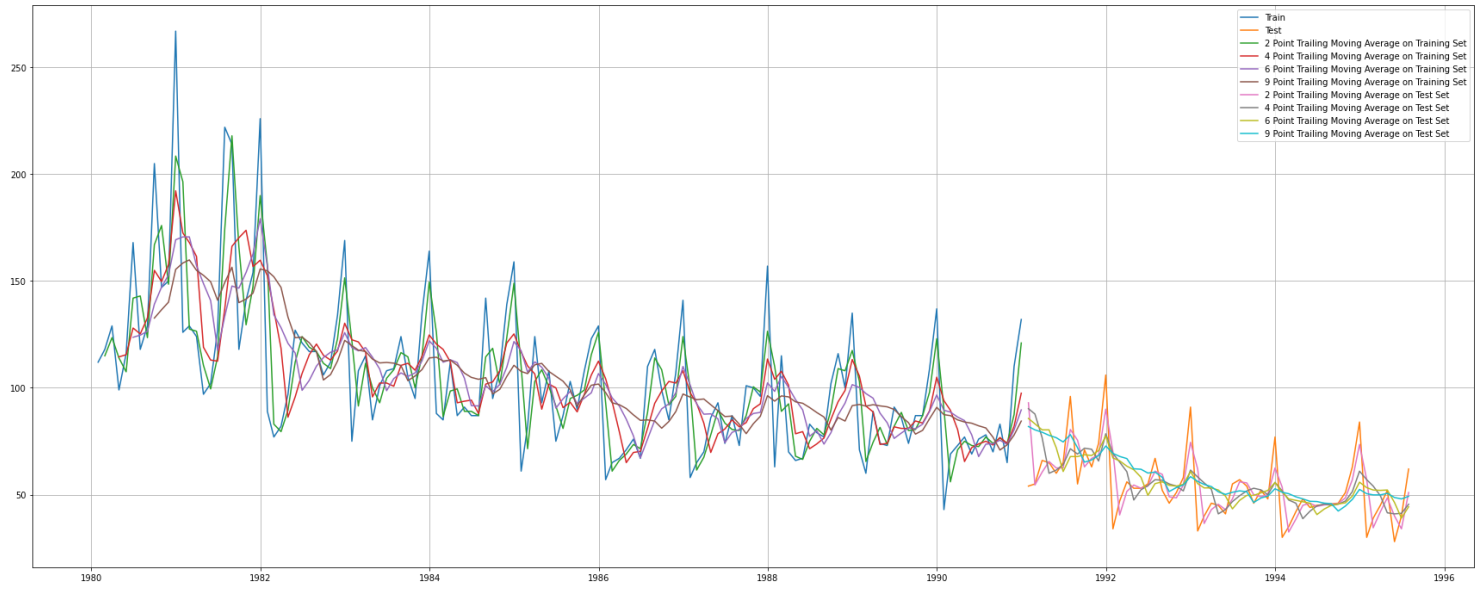


Figure 4.4(d)

Model Evaluation

Done only on the test data

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

BEFORE WE GO ON TO BUILD THE VARIOUS EXPONENTIAL SMOOTHING MODELS, LET US PLOT ALL THE MODELS AND COMPARE THE TIME SERIES PLOTS.

SPARKLING

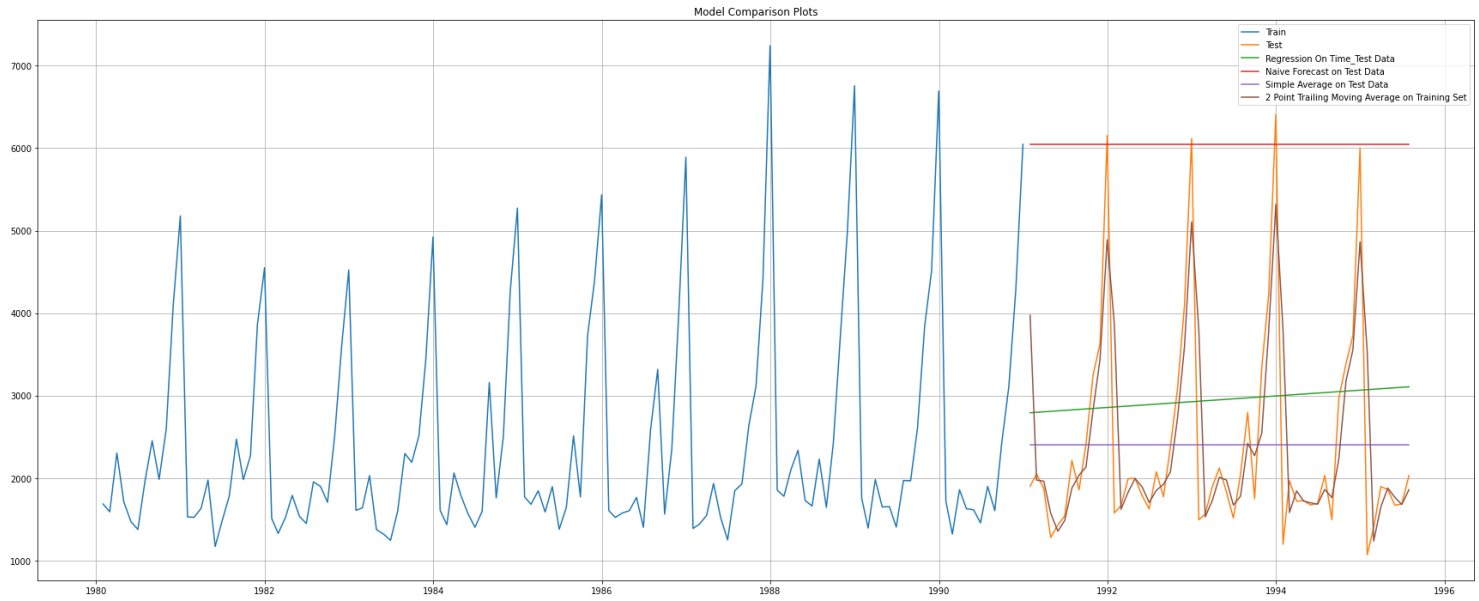


Figure 4.4(e)

ROSE

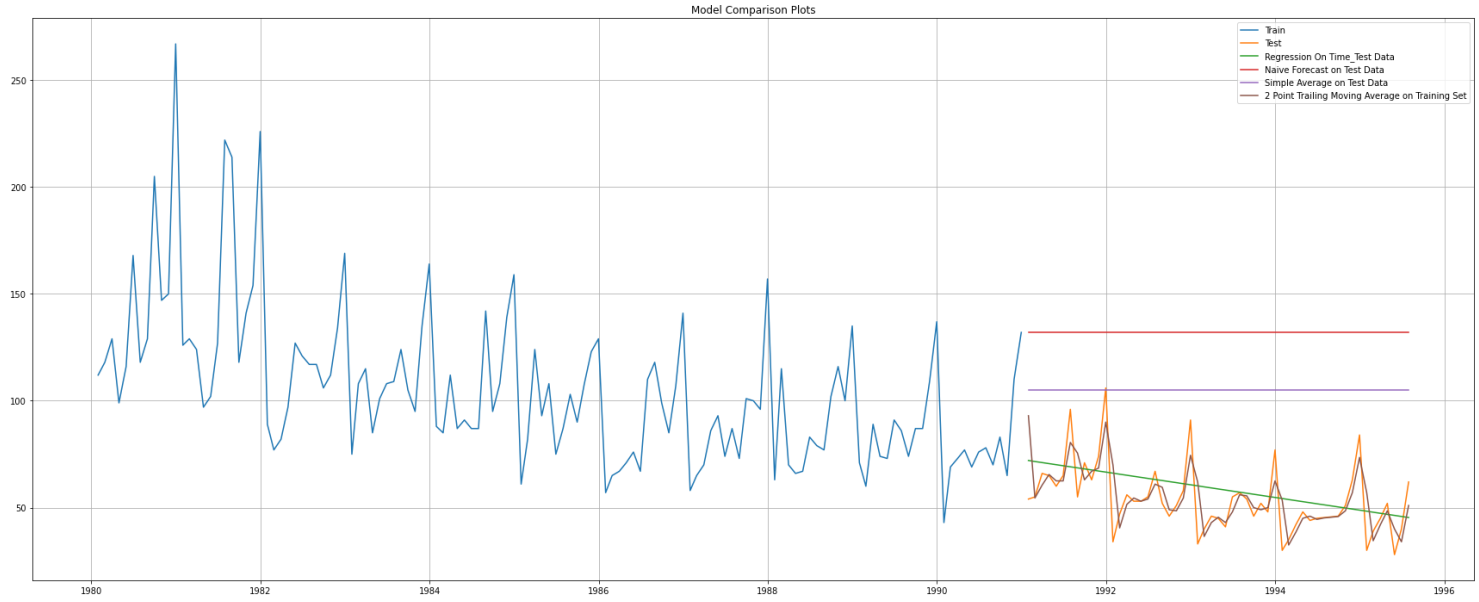


Figure 4.4(f)

- METHOD 5: SIMPLE EXPONENTIAL SMOOTHING

Automatically choosen parameters.

```
{'damping_slope': nan,
 'initial_level': 2403.7828696439005,
 'initial_seasons': array([], dtype=float64),
 'initial_slope': nan,
 'lamda': None,
 'remove_bias': False,
 'smoothing_level': 0.0,
 'smoothing_seasonal': nan,
 'smoothing_slope': nan,
 'use_boxcox': False}
```

Time_Stamp	Sparkling	predict
1991-01-31	1902	2403.78287
1991-02-28	2049	2403.78287
1991-03-31	1874	2403.78287
1991-04-30	1279	2403.78287
1991-05-31	1432	2403.78287

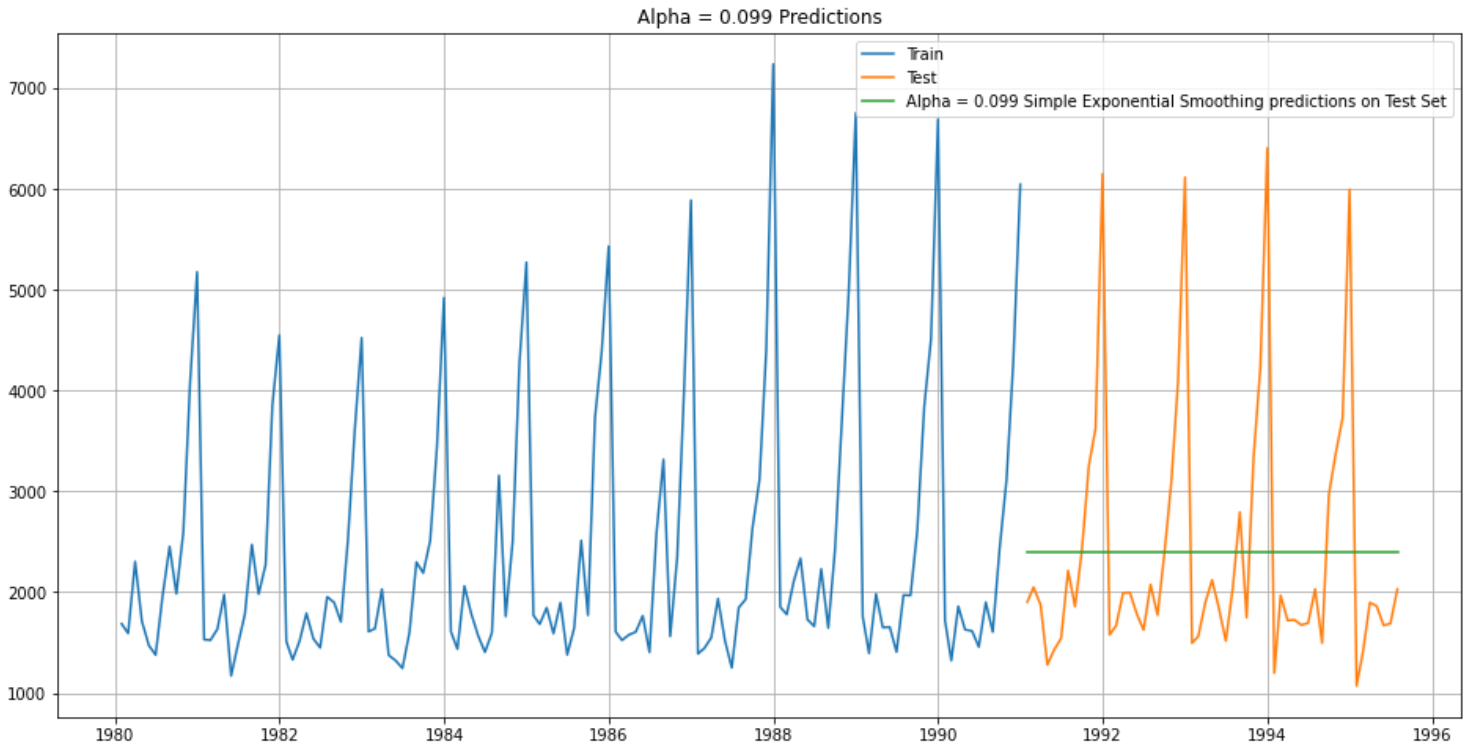


Figure 4.5(a)

MODEL EVALUATION FOR α = 0.099 : SIMPLE EXPONENTIAL SMOOTHING

- For Alpha =0.099 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1275.082

ROSE

```
{'damping_slope': nan,
 'initial_level': 134.38712015111975,
 'initial_seasons': array([], dtype=float64),
 'initial_slope': nan,
 'lamda': None,
 'remove_bias': False,
 'smoothing_level': 0.09875003987520162,
 'smoothing_seasonal': nan,
 'smoothing_slope': nan,
 'use_boxcox': False}
```

Time_Stamp	Rose	predict
1991-01-31	54.0	87.105003
1991-02-28	55.0	87.105003
1991-03-31	66.0	87.105003
1991-04-30	65.0	87.105003
1991-05-31	60.0	87.105003

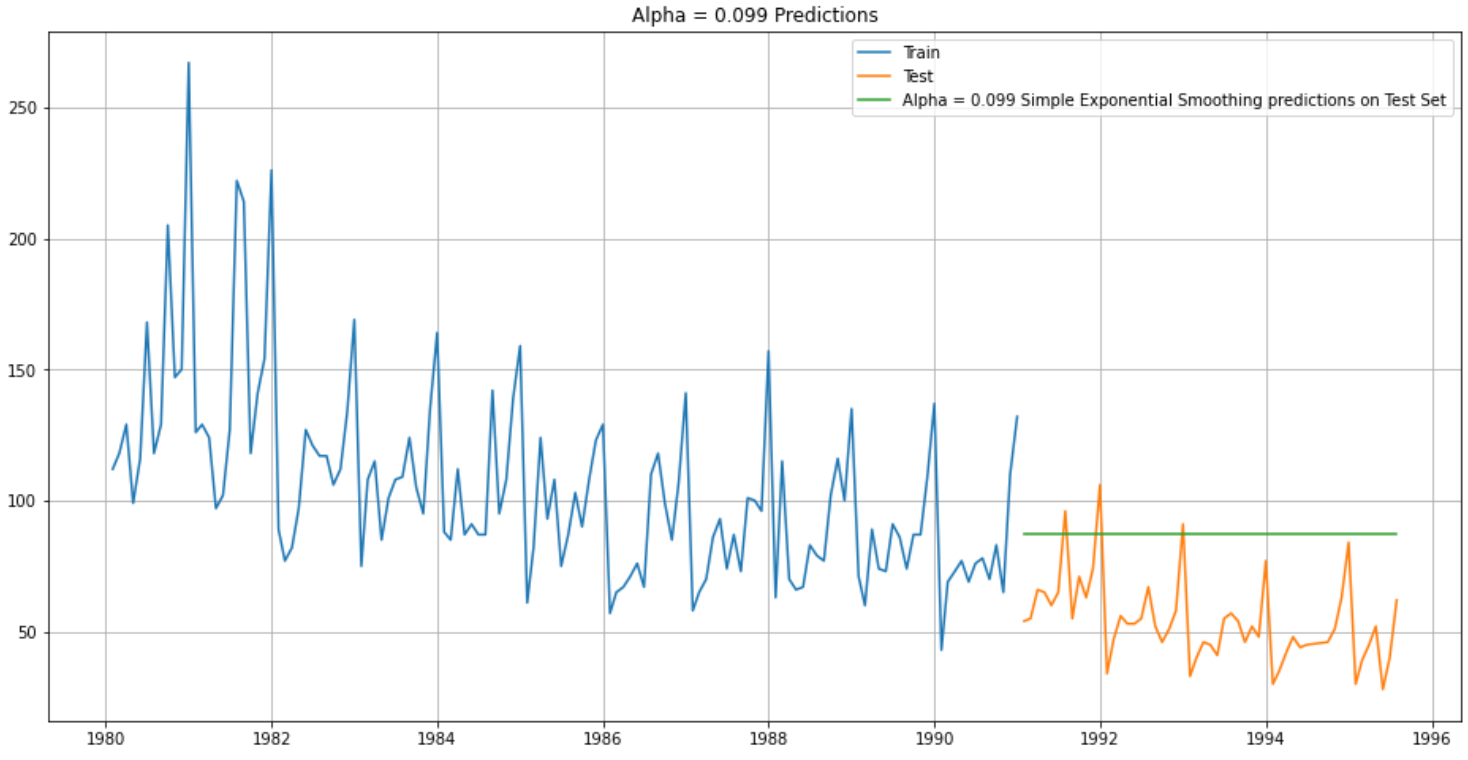


Figure 4.5(b)

MODEL EVALUATION FOR α = 0.099 : SIMPLE EXPONENTIAL SMOOTHING

- For Alpha = 0.099 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.796

Setting different alpha values.

Remember, the higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again.

We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

MODEL EVALUATION

SPARKLING

	Alpha Values	Train RMSE Sparkling	Test RMSE Sparkling
0	0.3	1359.511747	1935.507132
14	0.3	1359.511747	1935.507132
1	0.4	1352.588879	2311.919615
15	0.4	1352.588879	2311.919615
13	0.9	2638.989894	2603.403036
12	0.8	2639.223515	2606.403808
11	0.7	2639.458642	2610.048513

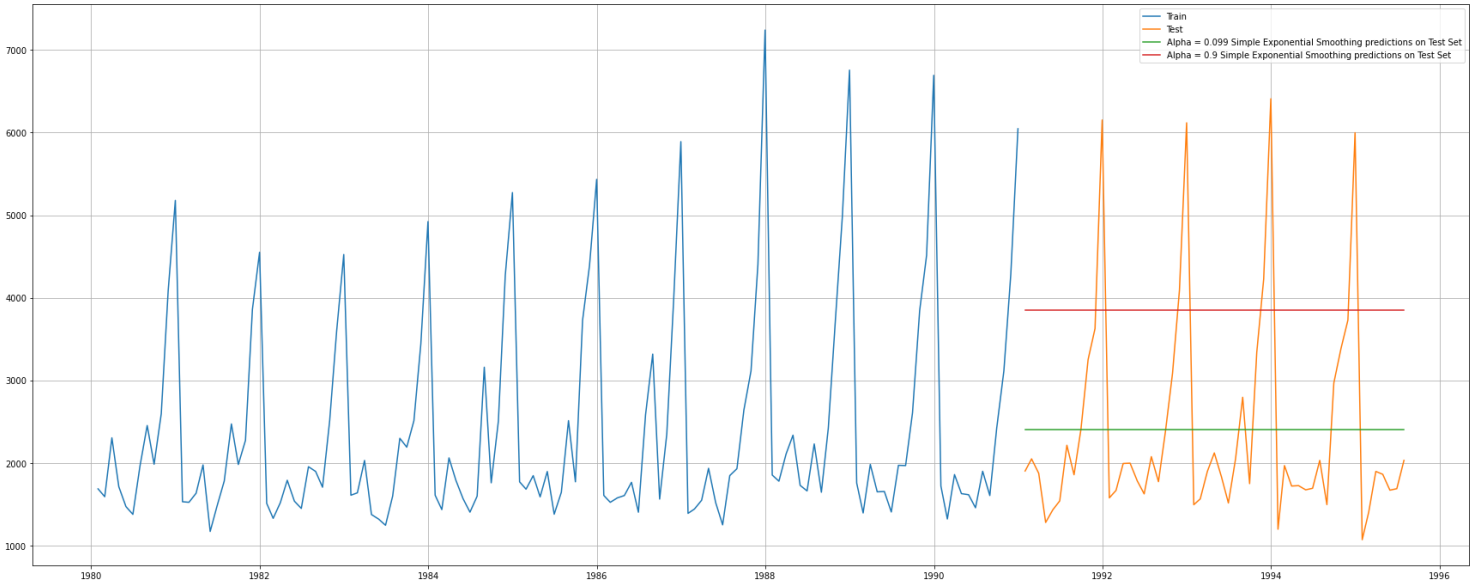


Figure 4.5(c)

- *AS OBSERVED, ALPHA = 0.3 IS THE ONE GIVING US THE LOWEST RMSE VALUE.*

ROSE

	Alpha Values	Train RMSE Rose	Test RMSE Rose
0	0.3	32.470164	47.504821
1	0.4	33.035130	53.767406
2	0.5	33.682839	59.641786
3	0.6	34.441171	64.971288
4	0.7	35.323261	69.698162
5	0.8	36.334596	73.773992
6	0.9	37.482782	77.139276

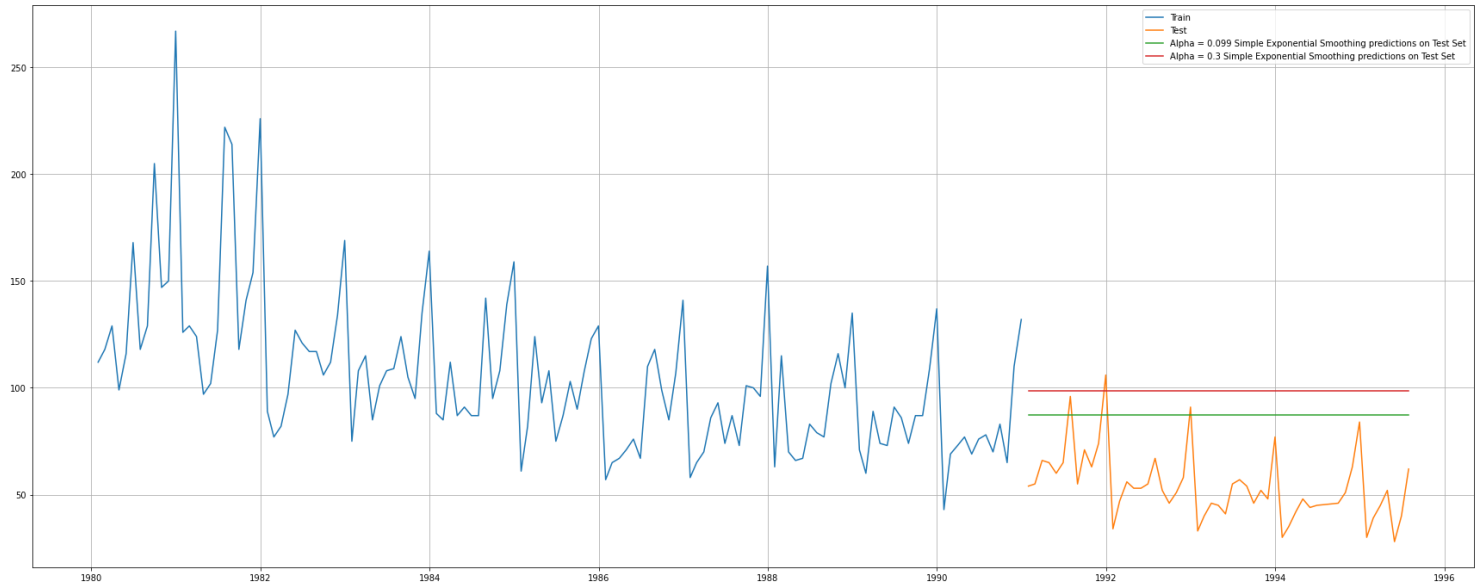


Figure 4.5(d)

- *AS OBSERVED, ALPHA = 0.3 IS THE ONE GIVING US THE LOWEST RMSE VALUE.*

- METHOD 6: DOUBLE EXPONENTIAL SMOOTHING (HOLT'S MODEL)

SPARKLING

	Alpha Values	Beta Values	Train RMSE Sparkling	Test RMSE Sparkling
0	0.3	0.3	1592.292788	18259.110704
8	0.4	0.3	1569.338606	23878.496940
1	0.3	0.4	1682.573828	26069.841401
16	0.5	0.3	1530.575845	27095.532414
24	0.6	0.3	1506.449870	29070.722592

- Here we are getting the lowest RMSE value for alpha=0.3 and beta=0.3.

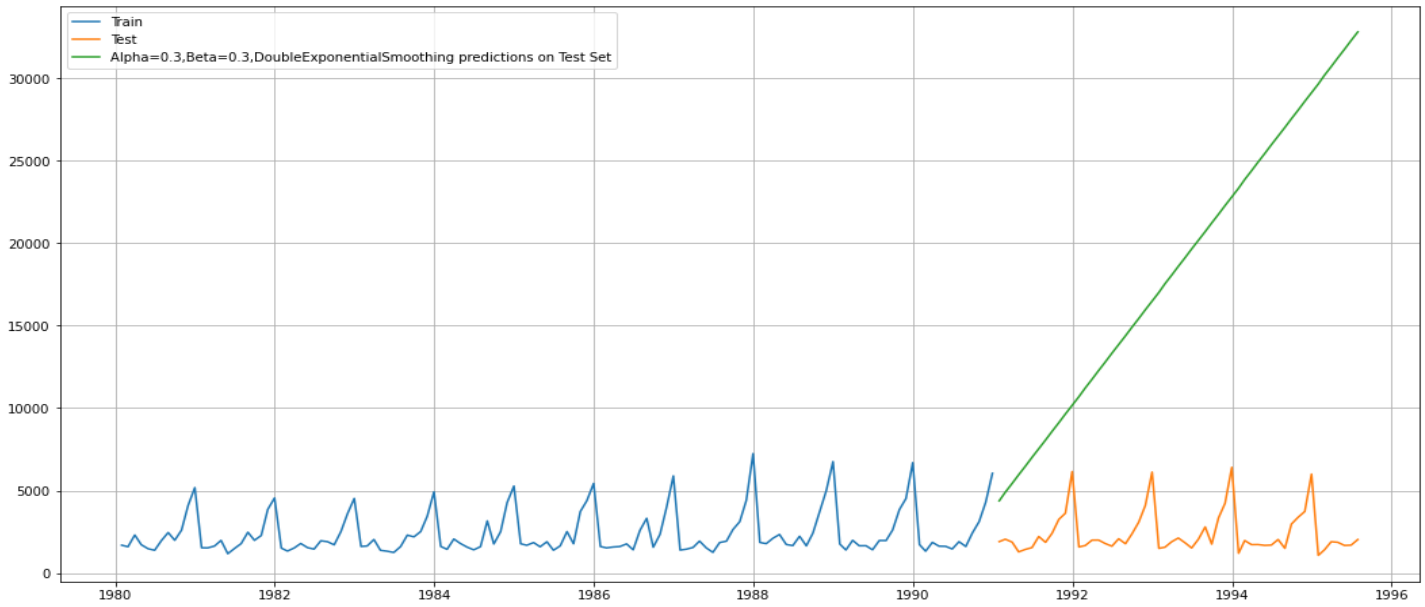


Figure 4.6(a)

ROSE

	Alpha Values	Beta Values	Train RMSE Rose	Test RMSE Rose
0	0.3	0.3	35.944983	265.567594
8	0.4	0.3	36.749123	339.306534
1	0.3	0.4	37.393239	358.750942
16	0.5	0.3	37.433314	394.272629
24	0.6	0.3	38.348984	439.296033

- Here we are getting the lowest RMSE value for alpha=0.3 and beta=0.3.

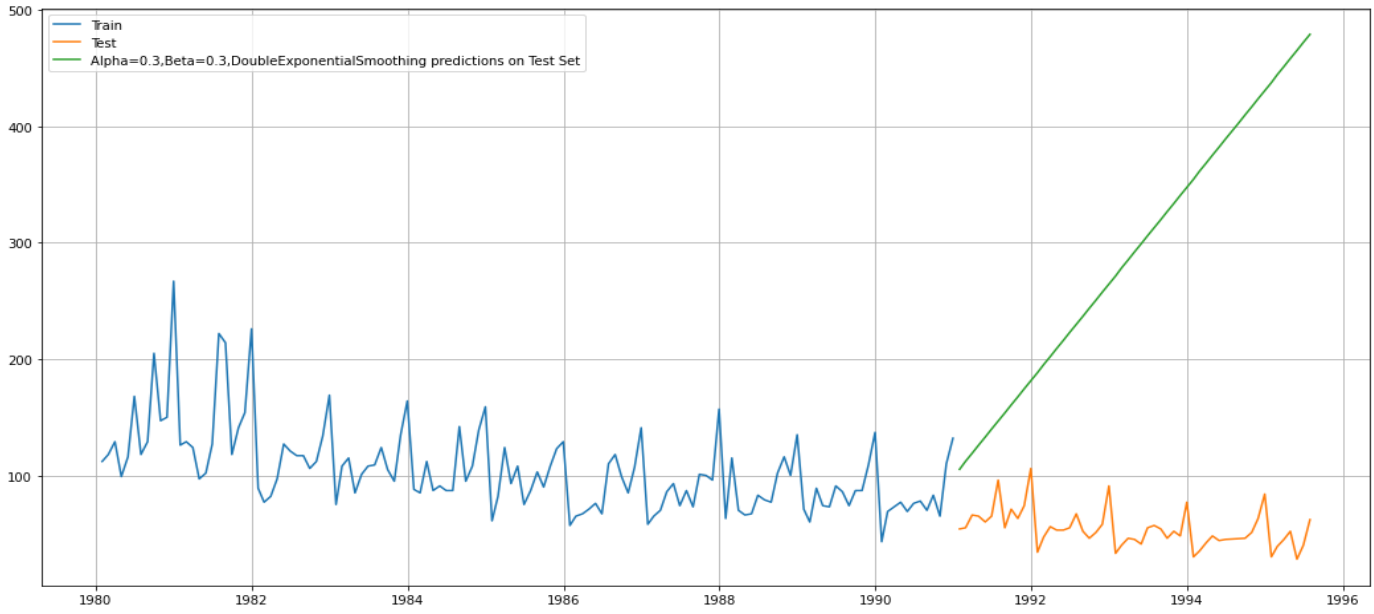


Figure 4.6(b)

- METHOD 7: TRIPLE EXPONENTIAL SMOOTHING (HOLT - WINTER'S MODEL)

Three parameters α , β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model

SPARKLING

```
{'damping_slope': nan,
 'initial_level': 1639.9993399027126,
 'initial_seasons': array([1.00841739, 0.96899632, 1.24171643, 1.13206069,
0.93984119,
        0.93813873, 1.22454059, 1.54419134, 1.27332629, 1.63190096,
        2.4826116 , 3.11820572]),
 'initial_slope': 4.884660841988308,
 'lamda': None,
 'remove_bias': False,
 'smoothing_level': 0.15443784802011637,
 'smoothing_seasonal': 0.37116865308973673,
 'smoothing_slope': 7.4133171248060435e-28,
 'use_boxcox': False}
```

Time_Stamp	Sparkling	auto_predict
1991-01-31	1902	1602.230902
1991-02-28	2049	1374.018912
1991-03-31	1874	1807.607439
1991-04-30	1279	1704.811690
1991-05-31	1432	1602.617906

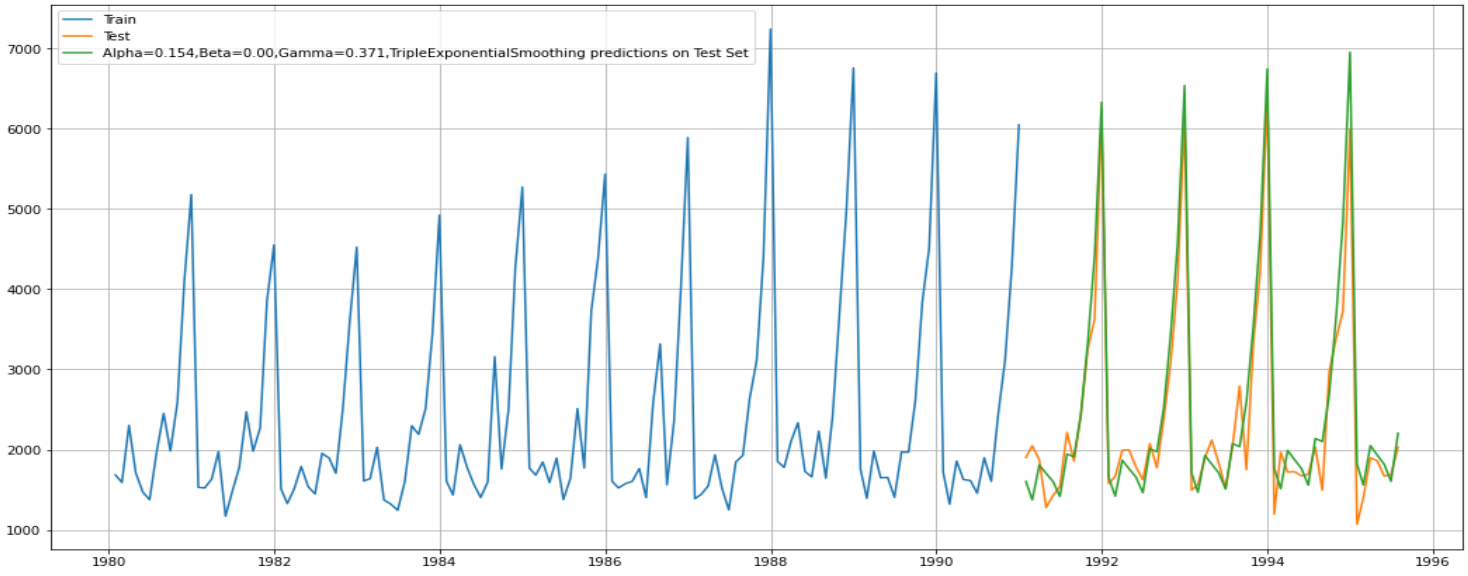


Figure 4.7(a)

MODEL EVALUATION

- For Alpha=0.154,Beta=0.00,Gamma=0.371, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 384.198

ROSE

```
{'damping_slope': nan,
 'initial_level': 76.65565186546128,
 'initial_seasons': array([1.47550231, 1.65927093, 1.80572588, 1.58888782,
1.77822665,
        1.92604314, 2.11649409, 2.25135146, 2.11690519, 2.08112772,
        2.40927212, 3.30448044]),
 'initial_slope': 0.0,
 'lamda': None,
 'remove_bias': False,
 'smoothing_level': 0.10609635974778751,
 'smoothing_seasonal': 0.0,
 'smoothing_slope': 0.048438458440198374,
 'use_boxcox': False}
```

Time_Stamp	Rose	auto_predict
1991-01-31	54.0	56.674338
1991-02-28	55.0	63.471271
1991-03-31	66.0	68.788789
1991-04-30	65.0	60.277826
1991-05-31	60.0	67.180381

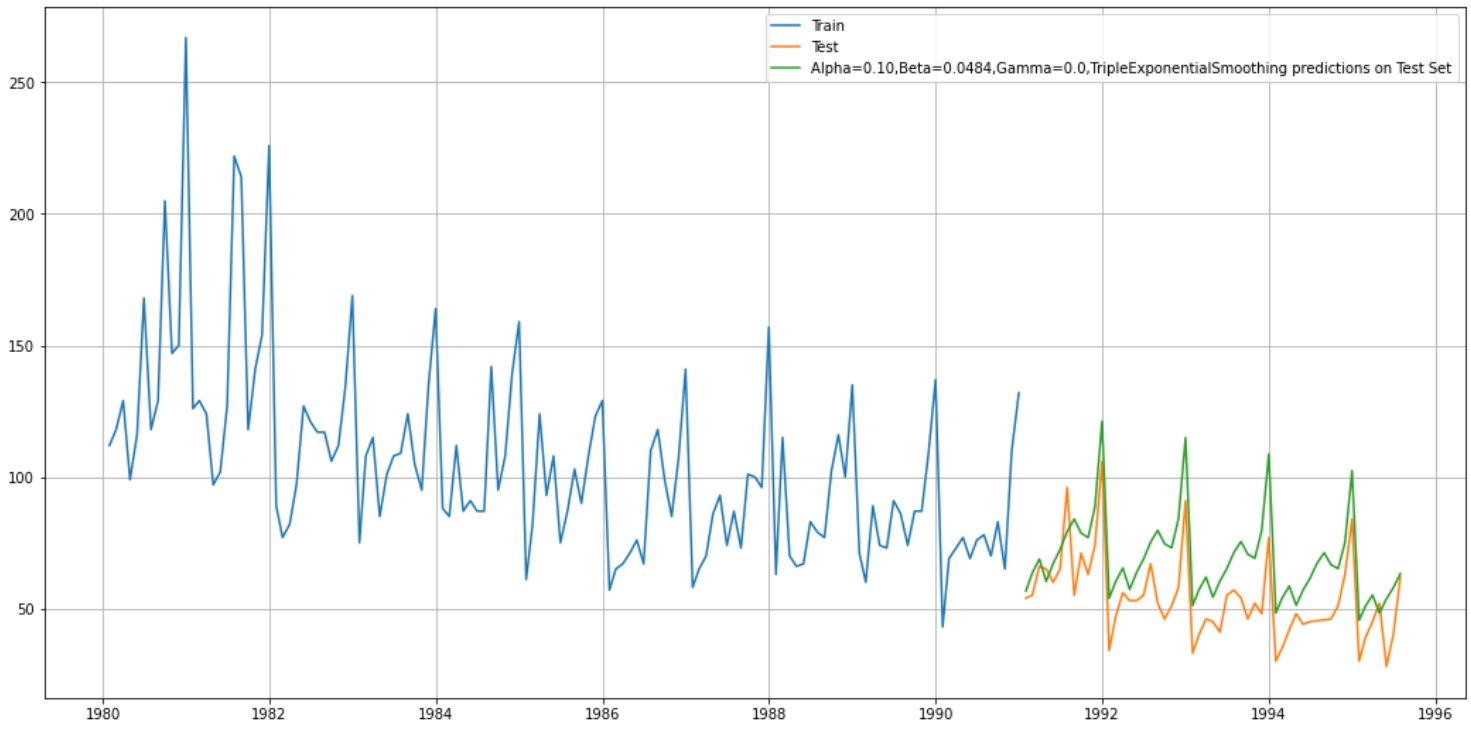


Figure 4.7(b)

MODEL EVALUATION

- For Alpha=0.10,Beta=0.0484,Gamma=0.0, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 17.369

Setting different alpha, beta and gamma values.

Remember, the higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again.

We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

SPARKLING

	Alpha Values	Beta Values	Gamma Values	Train RMSE Sparkling	Test RMSE Sparkling
0	0.3	0.3	0.3	404.513320	392.786198
8	0.3	0.4	0.3	424.828055	410.854547
65	0.4	0.3	0.4	435.553595	421.409170
296	0.7	0.8	0.3	700.317756	518.188752
130	0.5	0.3	0.5	498.239915	542.175497

- Here we are getting the lowest RMSE value for alpha=0.3, beta=0.3 and gamma=0.3.

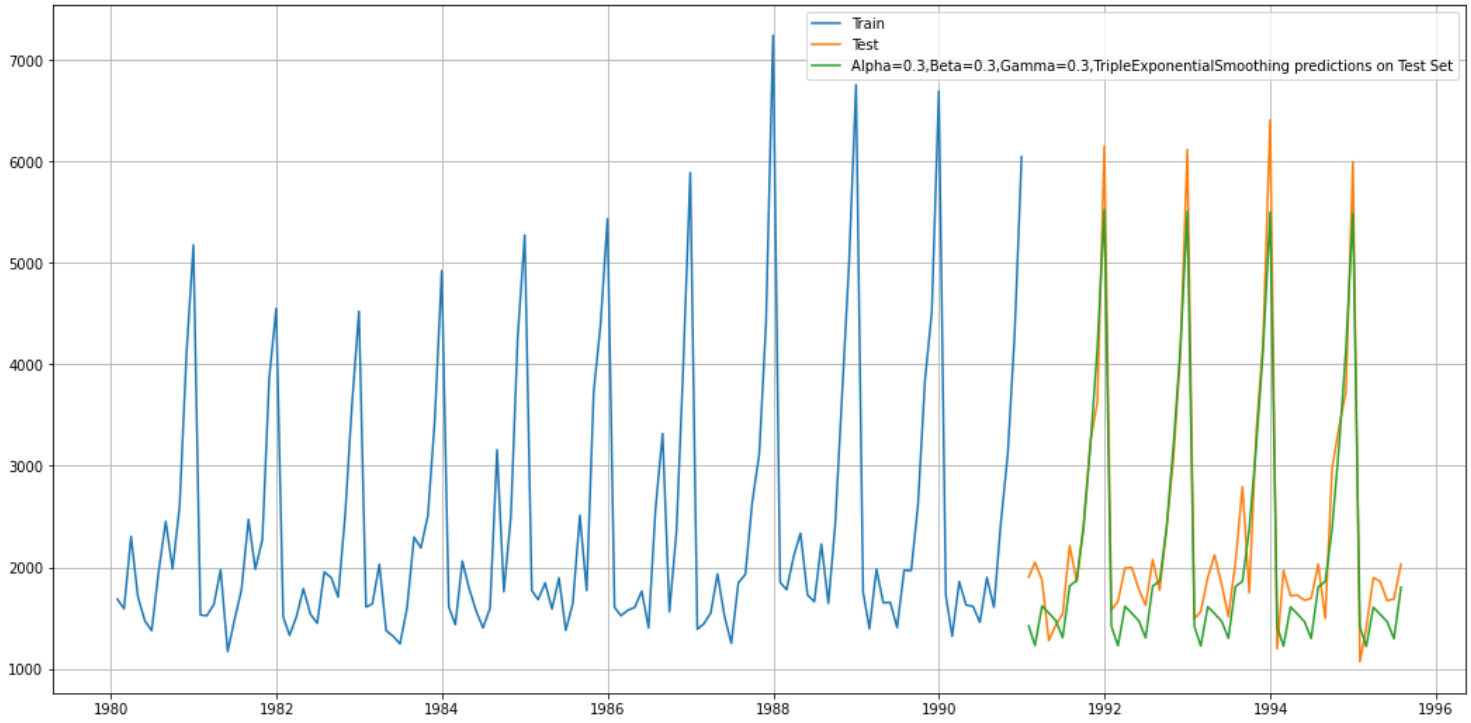


Figure 4.7(c)

ROSE

	Alpha Values	Beta Values	Gamma Values	Train RMSE Rose	Test RMSE Rose
8	0.3	0.4	0.3	28.111886	10.945435
1	0.3	0.3	0.4	27.399095	11.201633
69	0.4	0.3	0.8	32.601491	12.615607
16	0.3	0.5	0.3	29.087520	14.414604
131	0.5	0.3	0.6	32.144773	16.720720

- Here we are getting the lowest RMSE value for alpha=0.3, beta=0.4 and gamma=0.3.

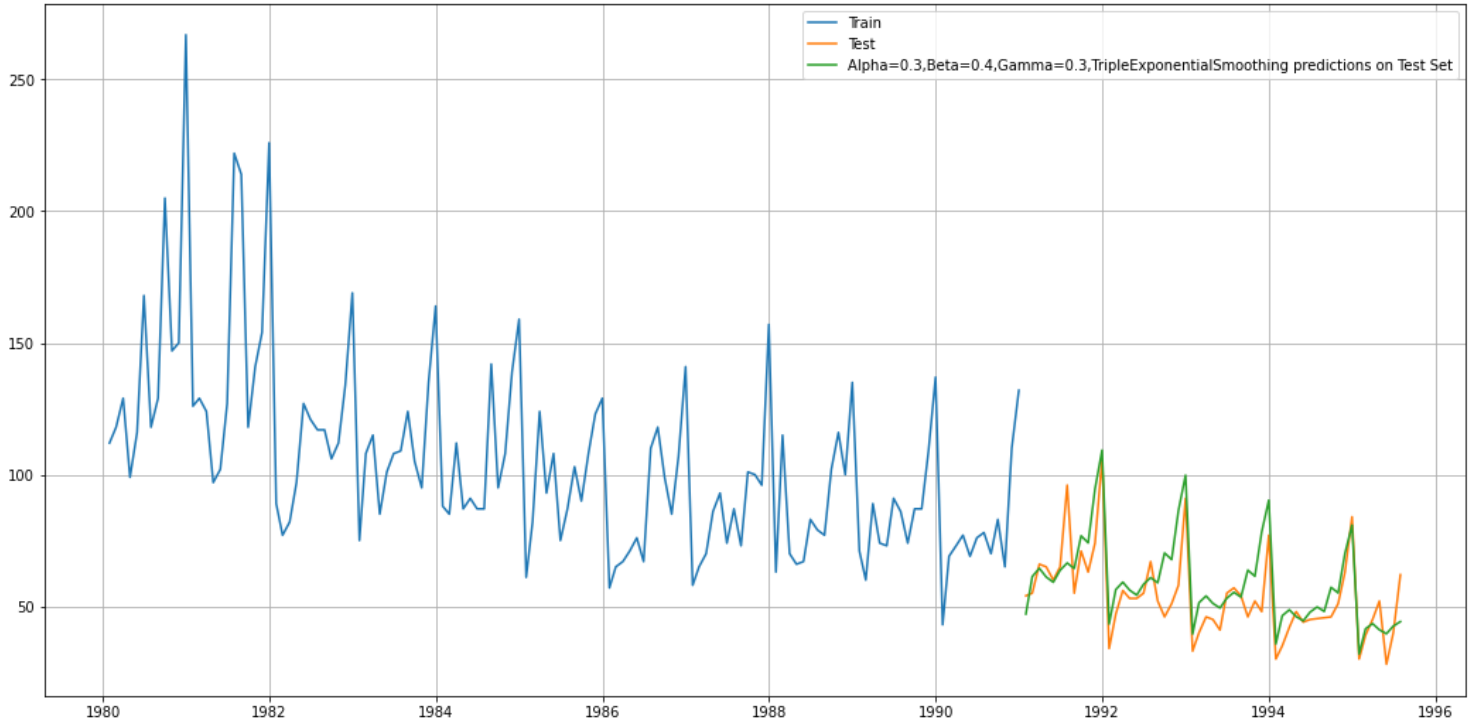


Figure 4.7(d)

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

SPARKLING

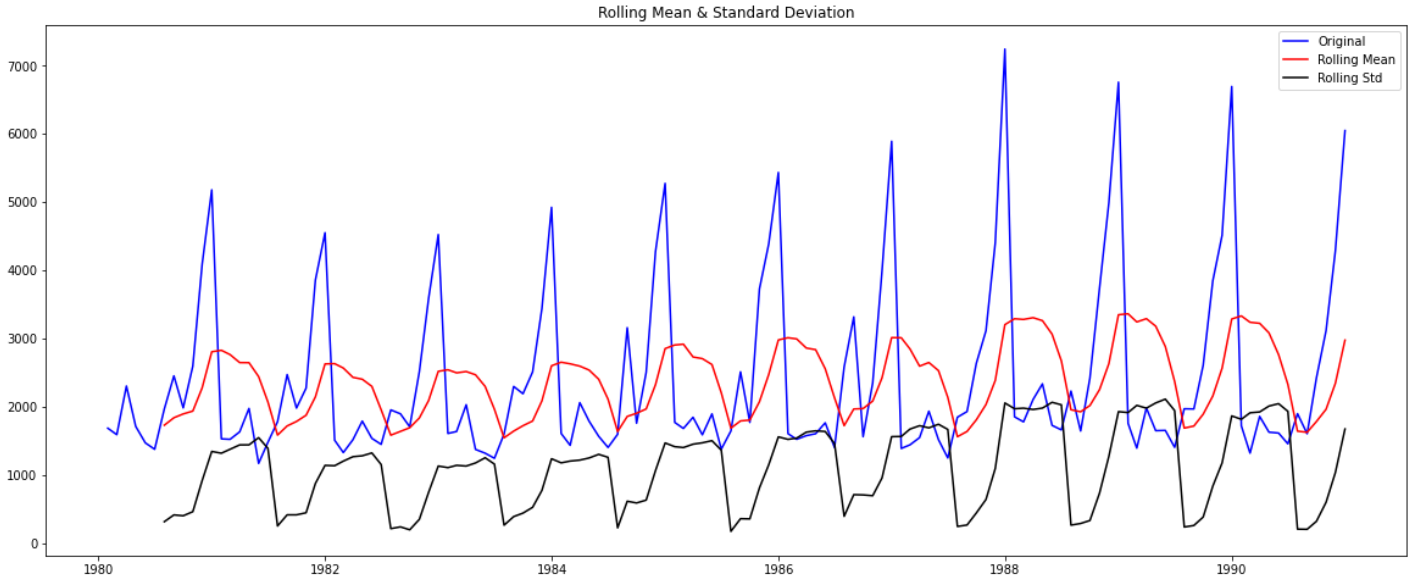


Figure 5.1(a)

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.208926
p-value                          0.669744
#Lags Used                      12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
```

- **We see that the series is not stationary at α = 0.05. Since p-value > 0.05**

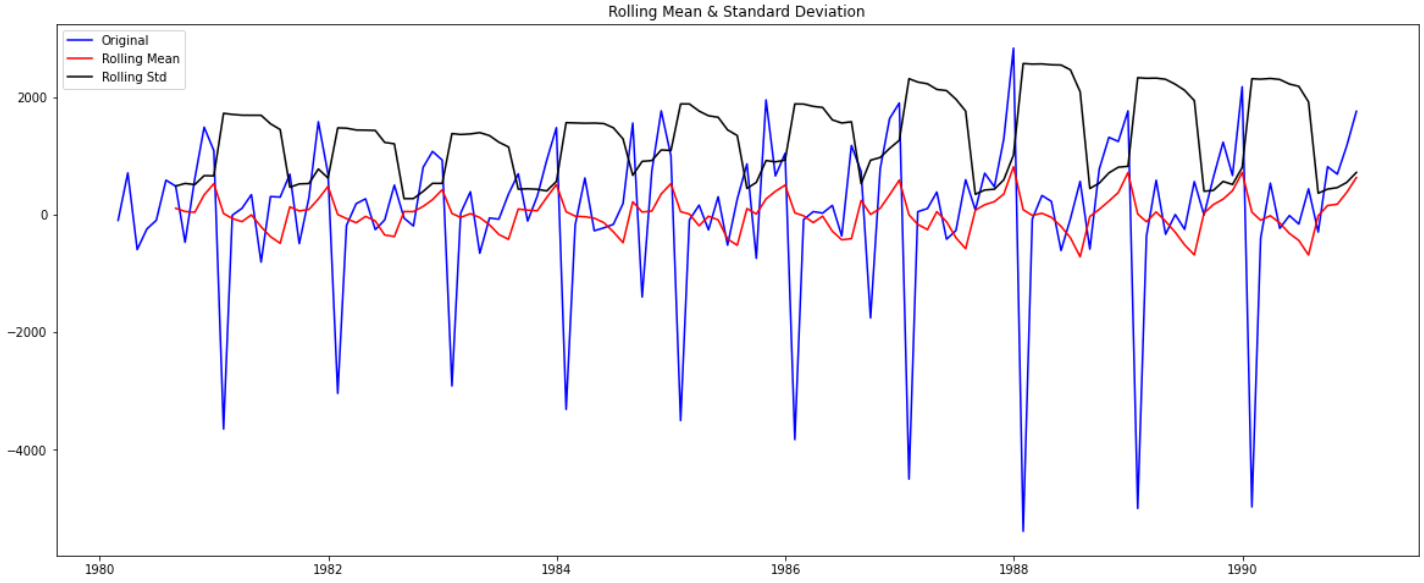


Figure 5.1(b)

```
Results of Dickey-Fuller Test:
Test Statistic                -8.005007e+00
p-value                        2.280104e-12
#Lags Used                     1.100000e+01
Number of Observations Used    1.190000e+02
Critical Value (1%)           -3.486535e+00
Critical Value (5%)           -2.886151e+00
Critical Value (10%)          -2.579896e+00
```

- **We see that after taking a difference of order 1 the series have become stationary at α = 0.05.**
- **Since p-value < 0.05**

Note: If the series is non-stationary, stationarize the Time Series by taking a difference of the Time Series. Then we can use this particular differenced series to train the ARIMA models. We do not need to worry about stationarity for the Test Data because we are not building any models on the Test Data, we are evaluating our models over there. You can look at other kinds of transformations as part of making the time series stationary like taking logarithms.

ROSE

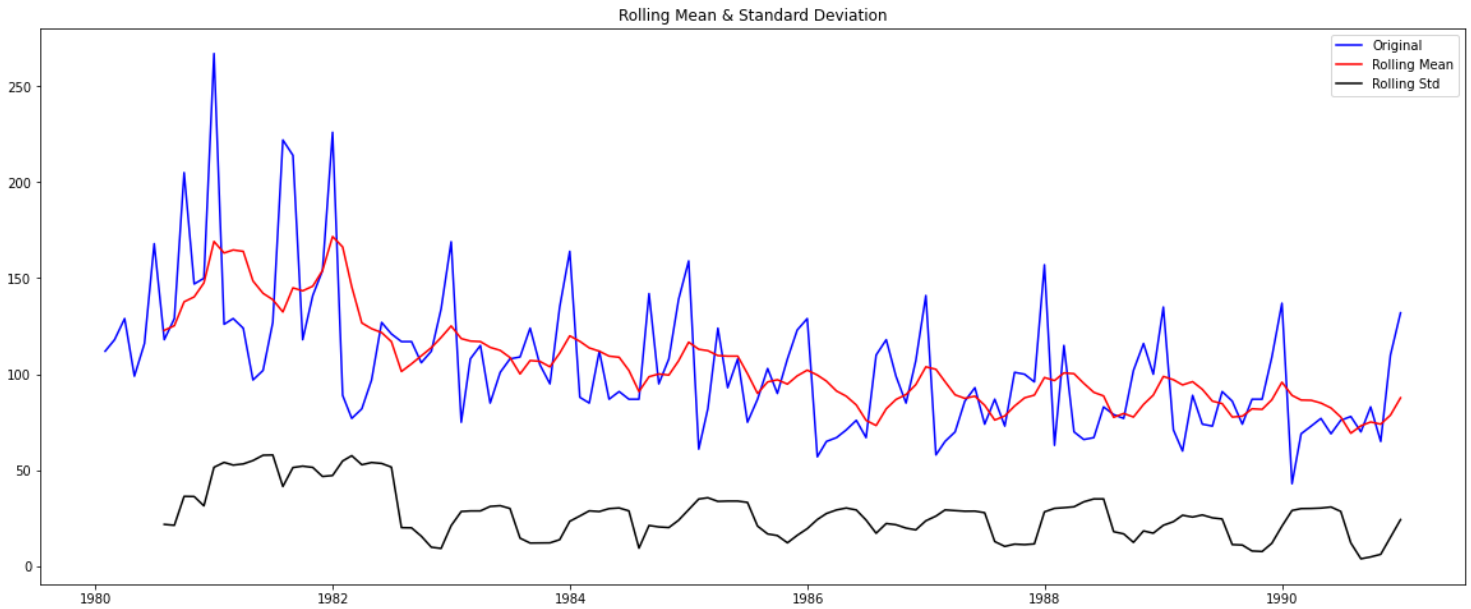


Figure 5.2(a)

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
```

- **We see that the series is not stationary at α = 0.05. Since p-value > 0.05**

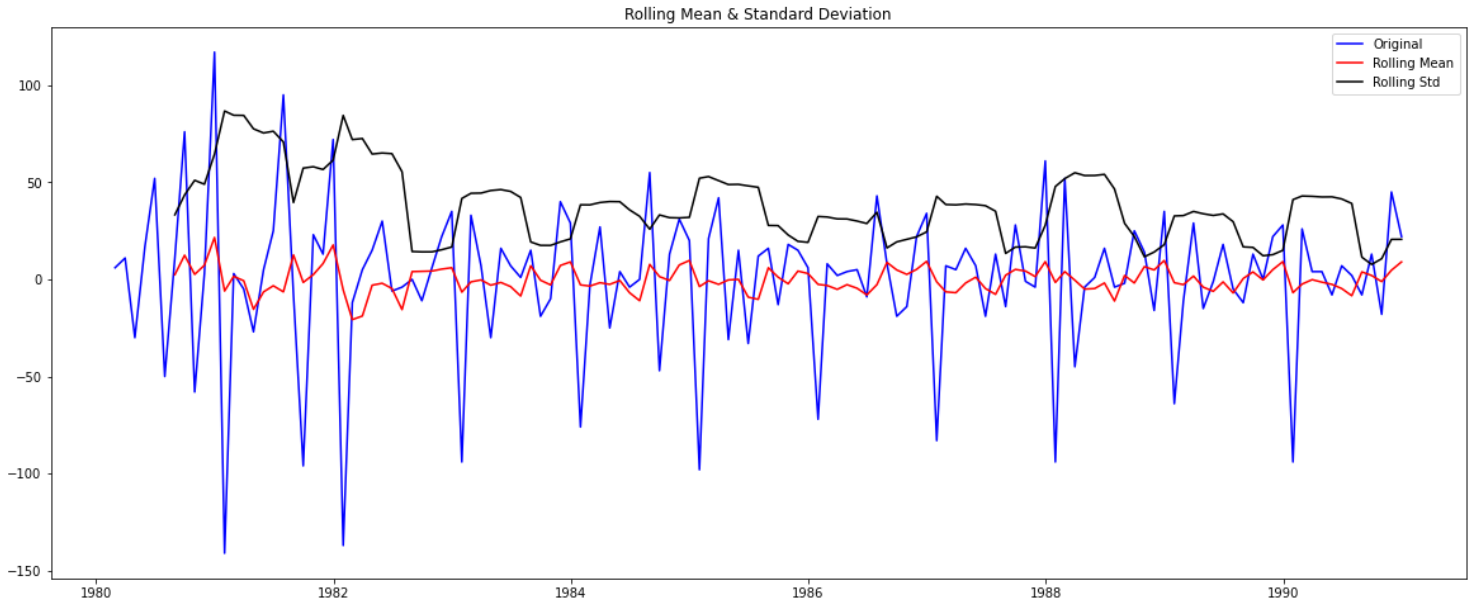


Figure 5.2(b)

```
Results of Dickey-Fuller Test:
Test Statistic                -6.592372e+00
p-value                        7.061944e-09
#Lags Used                     1.200000e+01
Number of Observations Used    1.180000e+02
Critical Value (1%)           -3.487022e+00
Critical Value (5%)           -2.886363e+00
Critical Value (10%)          -2.580009e+00
```

- **We see that after taking a difference of order 1 the series have become stationary at α = 0.05.**
- **Since p-value < 0.05**

6. BUILD AN AUTOMATED VERSION OF THE ARIMA/SARIMA MODEL IN WHICH THE PARAMETERS ARE SELECTED USING THE LOWEST AKAIKE INFORMATION CRITERIA (AIC) ON THE TRAINING DATA AND EVALUATE THIS MODEL ON THE TEST DATA USING RMSE.

- ARIMA

SPARKLING

```
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)
```

	param	AIC
8	(2, 1, 2)	2210.616145
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035601
1	(0, 1, 1)	2264.906437
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

WE will choose the lowest AIC value for best results, here (2,1,2)

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.308
Method:                       css-mle   S.D. of innovations            1012.081
Date:                Tue, 10 Aug 2021   AIC                            2210.616
Time:                        11:28:38   BIC                            2227.867
Sample:                    02-29-1980   HQIC                           2217.626
                         - 12-31-1990
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     5.5859      0.516     10.826      0.000       4.575       6.597
ar.L1.D.Sparkling         1.2699      0.074     17.047      0.000       1.124       1.416
ar.L2.D.Sparkling        -0.5601      0.074     -7.617      0.000      -0.704      -0.416
ma.L1.D.Sparkling        -1.9991      0.042    -47.175      0.000      -2.082      -1.916
ma.L2.D.Sparkling         0.9991      0.042     23.591      0.000       0.916       1.082
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1336           -0.7073j            1.3361           -0.0888
AR.2            1.1336           +0.7073j            1.3361            0.0888
MA.1            1.0005           -0.0009j            1.0005           -0.0001
MA.2            1.0005           +0.0009j            1.0005            0.0001
-----------------------------------------------------------------------------
/usr/local/lib/python3.7/dist-packages/statsmodels/base/model.py:492:
HessianInversionWarning: Inverting hessian failed, no bse or cov_params available
  'available', HessianInversionWarning)
```

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

forecast Returns

forecast : array Array of out of sample forecasts

stderr : array Array of the standard error of the forecasts.

conf_int : array 2d array of the confidence interval for the forecast

- The RMSE value for ARIMA(2,1,2) is 1374.976339330213

ROSE

- 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)

	param	AIC
2	(0, 1, 2)	1276.835376
5	(1, 1, 2)	1277.359224
4	(1, 1, 1)	1277.775759
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

WE WILL CHOOSE THE LOWEST AIC VALUE FOR BEST RESULTS, HERE (0,1,2)

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(0, 1, 2)	Log Likelihood	-634.418
Method:	css-mle	S.D. of innovations	30.167
Date:	Tue, 10 Aug 2021	AIC	1276.835
Time:	11:28:39	BIC	1288.336
Sample:	02-29-1980 - 12-31-1990	HQIC	1281.509

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4885	0.085	-5.742	0.000	-0.655	-0.322
ma.L1.D.Rose	-0.7601	0.101	-7.499	0.000	-0.959	-0.561
ma.L2.D.Rose	-0.2398	0.095	-2.518	0.013	-0.427	-0.053

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.0001	+0.0000j	1.0001	0.0000
MA.2	-4.1696	+0.0000j	4.1696	0.5000

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

- The RMSE value for ARIMA(0,1,2) is 15.618703067489148

- SARIMA

SPARKLING

Let us look at the ACF plot once more to understand the seasonal parameter for the SARIMA model.

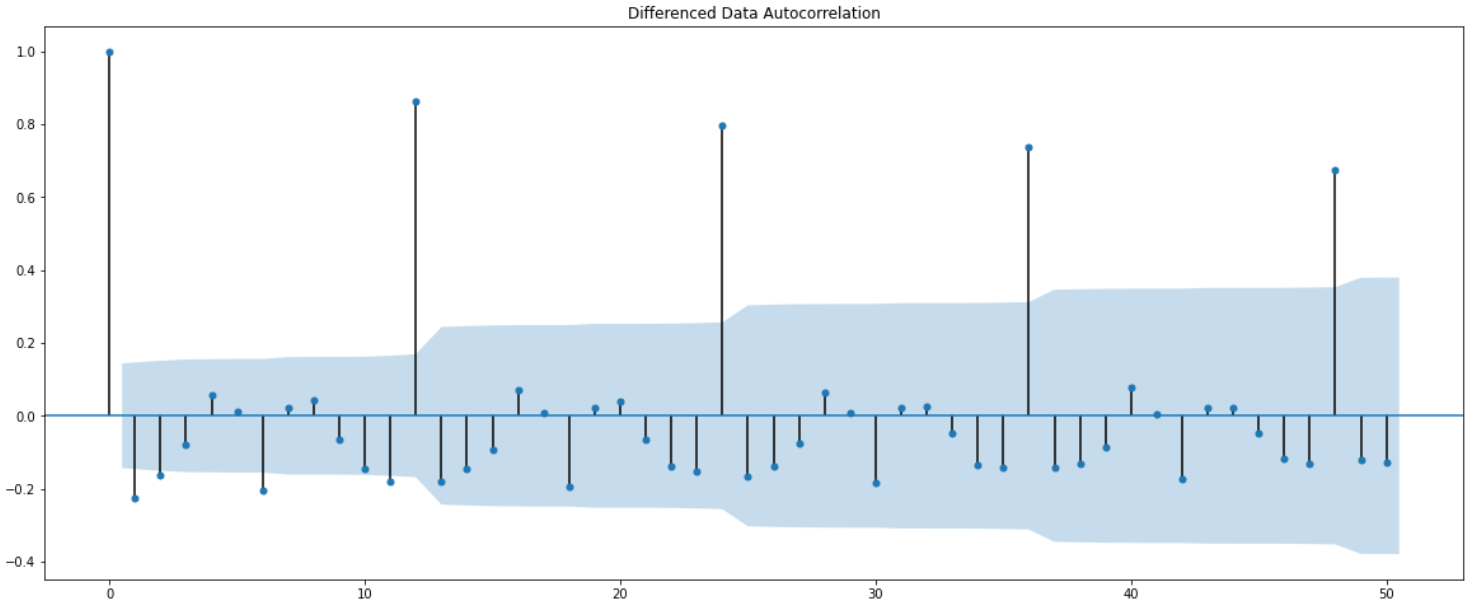


Figure 6.1(a)

We see that there can be a seasonality of 6 as well as 12. We will run our auto SARIMA models by setting seasonality both as 6 and 12.

- Setting the seasonality as 6 for the first iteration of the auto SARIMA model.

Top 5 Results sorted by AIC values.

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.678706
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888804
80	(2, 1, 2)	(2, 0, 2, 6)	1729.192578
17	(0, 1, 1)	(2, 0, 2, 6)	1741.696451
44	(1, 1, 1)	(2, 0, 2, 6)	1743.379779

Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	855.839
Date:	Tue, 10 Aug 2021	AIC	1727.679
Time:	11:29:13	BIC	1749.707
Sample:	0 - 132	HQIC	1736.621
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6452	0.286	-2.258	0.024	-1.205	-0.085
ma.L1	-0.1065	0.250	-0.427	0.670	-0.596	0.383
ma.L2	-0.7008	0.202	-3.474	0.001	-1.096	-0.305
ar.S.L6	-0.0045	0.027	-0.165	0.869	-0.057	0.049
ar.S.L12	1.0361	0.018	56.097	0.000	1.000	1.072
ma.S.L6	0.0675	0.152	0.444	0.657	-0.231	0.366
ma.S.L12	-0.6125	0.093	-6.591	0.000	-0.795	-0.430
sigma2	1.448e+05	1.71e+04	8.466	0.000	1.11e+05	1.78e+05

Ljung-Box (Q):	28.93	Jarque-Bera (JB):	25.23
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	2.63	Skew:	0.47
Prob(H) (two-sided):	0.00	Kurtosis:	5.09

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

DIAGNOSTIC PLOT

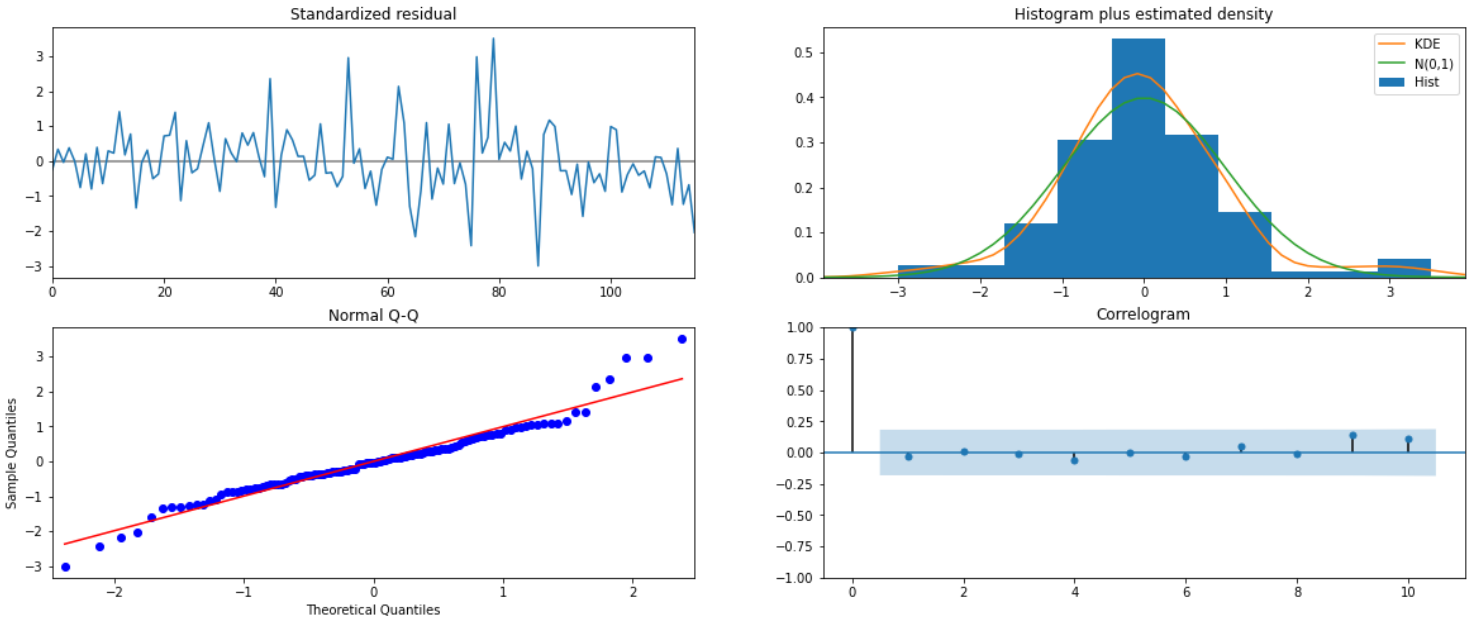


Figure 6.1(b)

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

Predict on the Test Set using this model and evaluate the model.

Summary Frame of the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1330.291793	380.590238	584.348633	2076.234953
1	1177.259923	392.141600	408.676509	1945.843336
2	1625.818625	392.336089	856.854022	2394.783229
3	1546.373979	397.740941	766.816060	2325.931898
4	1308.662018	398.960289	526.714220	2090.609816

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

- The RMSE value for SARIMA(1,1,2)(2,0,2,6) is 626.933219231813

ROSE

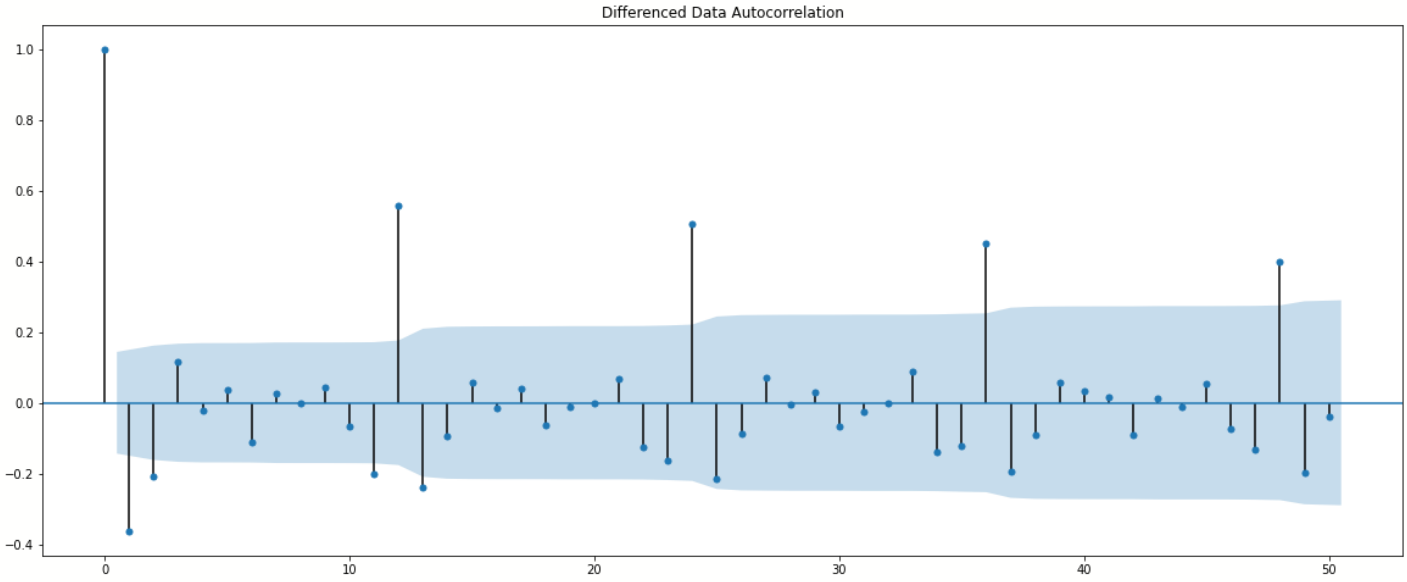


Figure 6.2(a)

- Setting the seasonality as 6 for the first iteration of the auto SARIMA model.

Top 5 Results sorted by AIC values.

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655819
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.228247
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	512.828
Date:	Wed, 11 Aug 2021	AIC	1041.656
Time:	14:28:36	BIC	1063.685
Sample:	0 - 132	HQIC	1050.598
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5938	0.152	-3.896	0.000	-0.892	-0.295
ma.L1	-0.1955	964.803	-0.000	1.000	-1891.175	1890.784
ma.L2	-0.8045	776.176	-0.001	0.999	-1522.082	1520.473
ar.S.L6	-0.0626	0.035	-1.765	0.078	-0.132	0.007
ar.S.L12	0.8451	0.039	21.886	0.000	0.769	0.921
ma.S.L6	0.2226	311.195	0.001	0.999	-609.708	610.153
ma.S.L12	-0.7775	241.893	-0.003	0.997	-474.879	473.324
sigma2	335.2008	3.49e+05	0.001	0.999	-6.83e+05	6.84e+05

Ljung-Box (Q):	15.89	Jarque-Bera (JB):	56.68
Prob(Q):	1.00	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).

DIAGNOSTIC PLOT

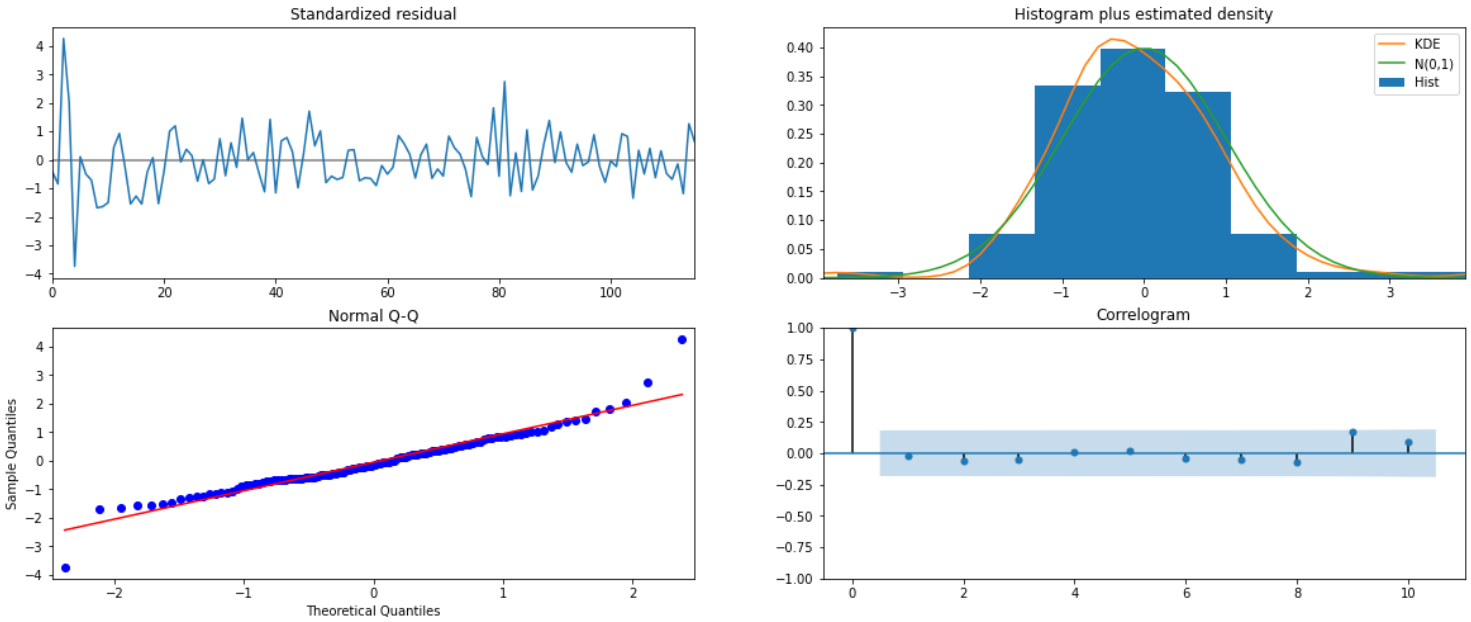


Figure 6.2(b)

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

	y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.843009	18.848543	25.900545	99.785474
1	67.632101	19.300677	29.803470	105.460732
2	74.747794	19.413252	36.698519	112.797068
3	71.326411	19.476177	33.153806	109.499017
4	76.017208	19.484445	37.828398	114.206018

- The RMSE value for SARIMA(1,1,2)(2,0,2,6) is 26.136428629473784

- SETTING THE SEASONALITY AS 12 FOR THE SECOND ITERATION OF THE AUTO SARIMA MODEL.

Top 5 Results sorted by AIC values for SPARKLING.

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.934564
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121563
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340409

Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	770.792
Date:	Wed, 11 Aug 2021	AIC	1555.584
Time:	14:29:59	BIC	1574.095
Sample:	0	HQIC	1563.083
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6282	0.255	-2.464	0.014	-1.128	-0.128
ma.L1	-0.1040	0.225	-0.463	0.643	-0.545	0.336
ma.L2	-0.7277	0.154	-4.736	0.000	-1.029	-0.427
ar.S.L12	1.0439	0.014	72.837	0.000	1.016	1.072
ma.S.L12	-0.5550	0.098	-5.663	0.000	-0.747	-0.363
ma.S.L24	-0.1354	0.120	-1.133	0.257	-0.370	0.099
sigma2	1.506e+05	2.03e+04	7.401	0.000	1.11e+05	1.9e+05

Ljung-Box (Q):	23.02	Jarque-Bera (JB):	11.72
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	1.47	Skew:	0.36
Prob(H) (two-sided):	0.26	Kurtosis:	4.48

Diagnostic Plot

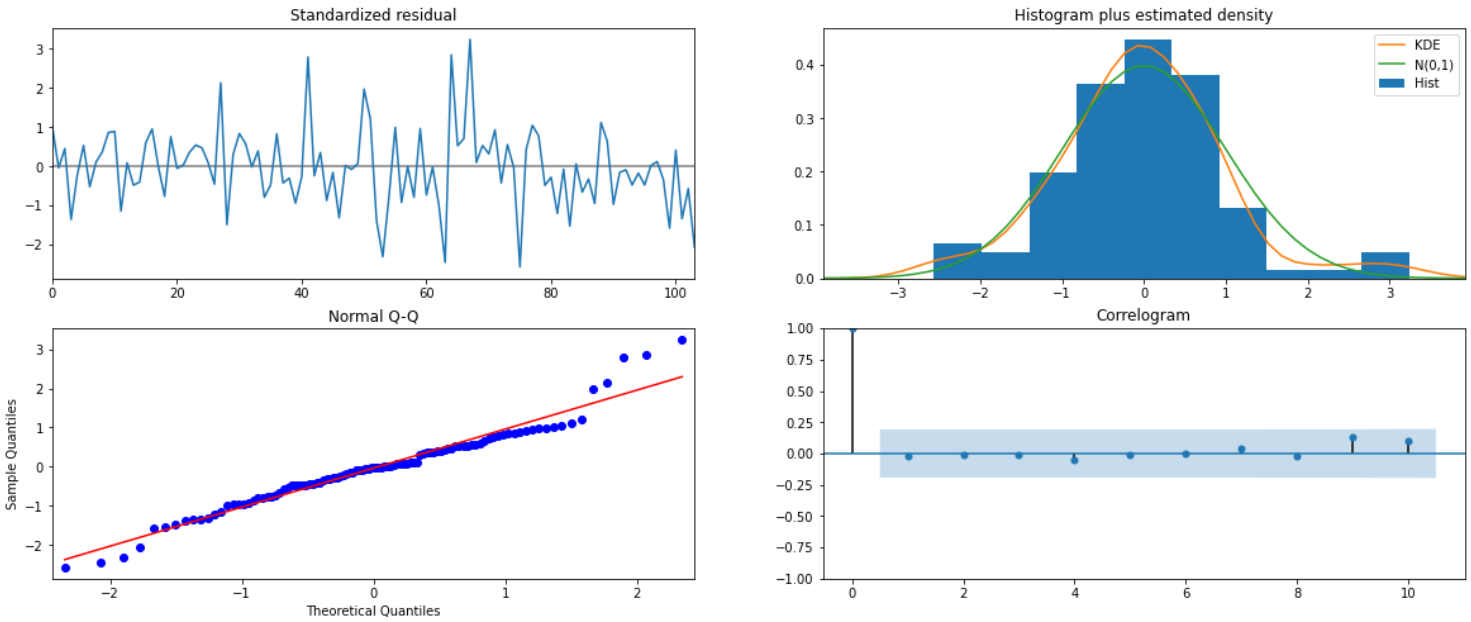


Figure 6.3(a)

Similar to the last iteration of the model where the seasonality parameter was taken as 6, here also we see that the model diagnostics plot does not indicate any remaining information that we can get.

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1327.401257	388.341132	566.266625	2088.535890
1	1315.149557	402.001939	527.240235	2103.058879
2	1621.594351	401.995550	833.697550	2409.491151
3	1598.902813	407.233028	800.740745	2397.064882
4	1392.703609	407.962500	593.111803	2192.295415

- The RMSE value for SARIMA(1,1,2)(1,0,2,12) is 528.6025130673597

Top 5 Results sorted by AIC values for ROSE.

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.900383
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	436.969
Date:	Wed, 11 Aug 2021	AIC	887.938
Time:	14:30:57	BIC	906.448
Sample:	0 - 132	HQIC	895.437
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8427	189.890	-0.004	0.996	-373.020	371.334
ma.L2	-0.1573	29.833	-0.005	0.996	-58.628	58.314
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	251.3137	4.77e+04	0.005	0.996	-9.33e+04	9.38e+04

Ljung-Box (Q):	24.56	Jarque-Bera (JB):	2.33
Prob(Q):	0.97	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).

Diagnostic Plot

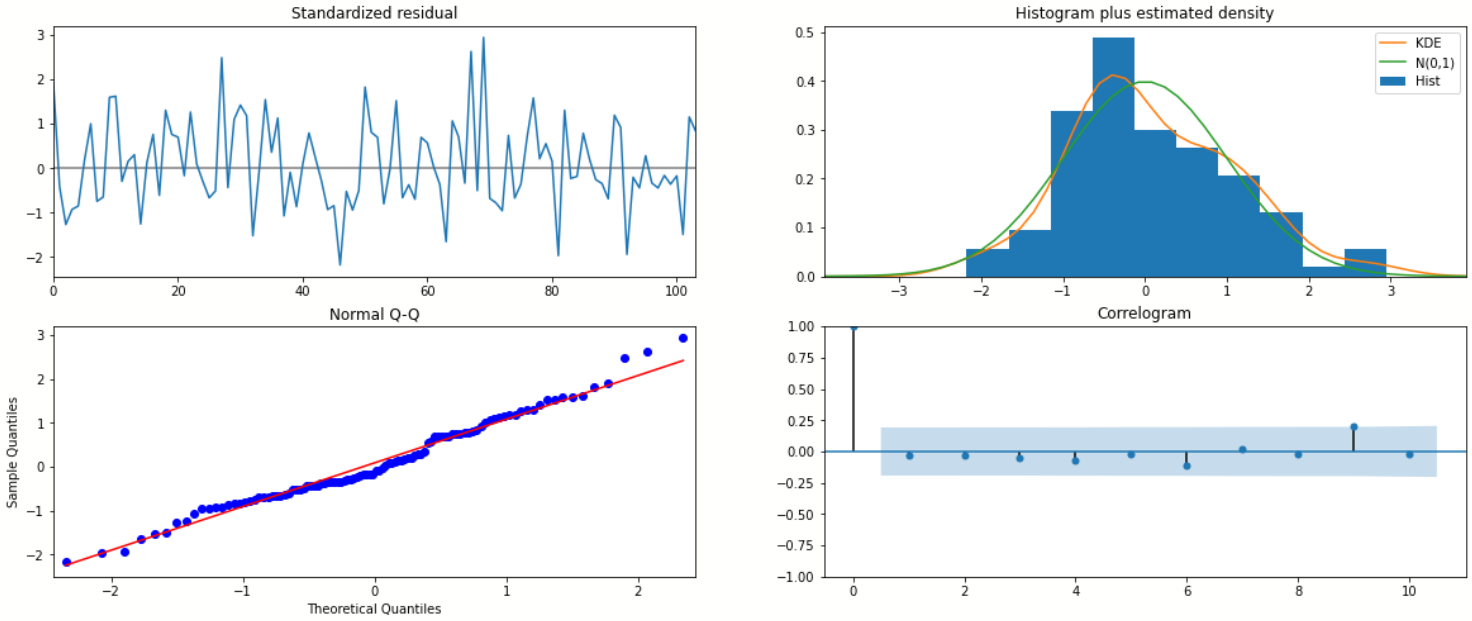


Figure 6.3(b)

Similar to the last iteration of the model where the seasonality parameter was taken as 6, here also we see that the model diagnostics plot does not indicate any remaining information that we can get.

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

	y	mean	mean_se	mean_ci_lower	mean_ci_upper
0		62.867264	15.928501	31.647976	94.086552
1		70.541190	16.147659	38.892360	102.190020
2		77.356411	16.147656	45.707586	109.005236
3		76.208814	16.147656	44.559989	107.857639
4		72.747398	16.147656	41.098573	104.396223

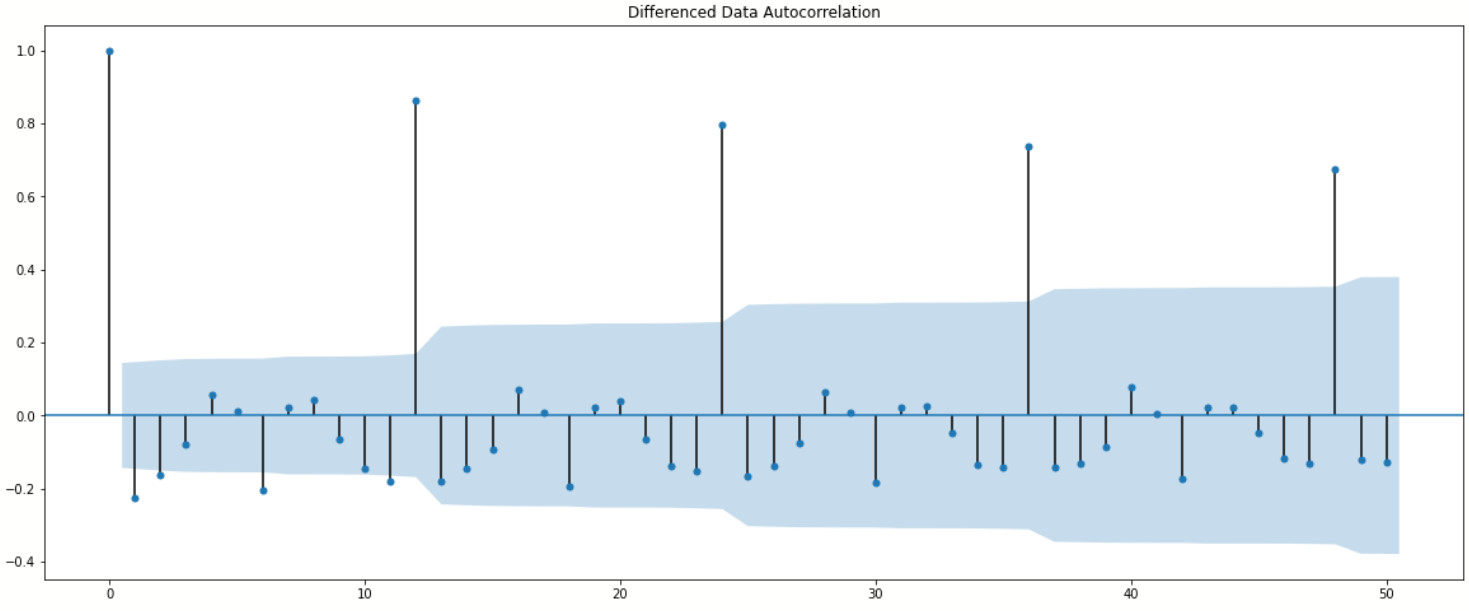
- The RMSE value for SARIMA(0,1,2)(2,0,2,12) is 26.928361908025416

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.

SPARKLING

- ARIMA

Here we will be using the differenced time series dataset .



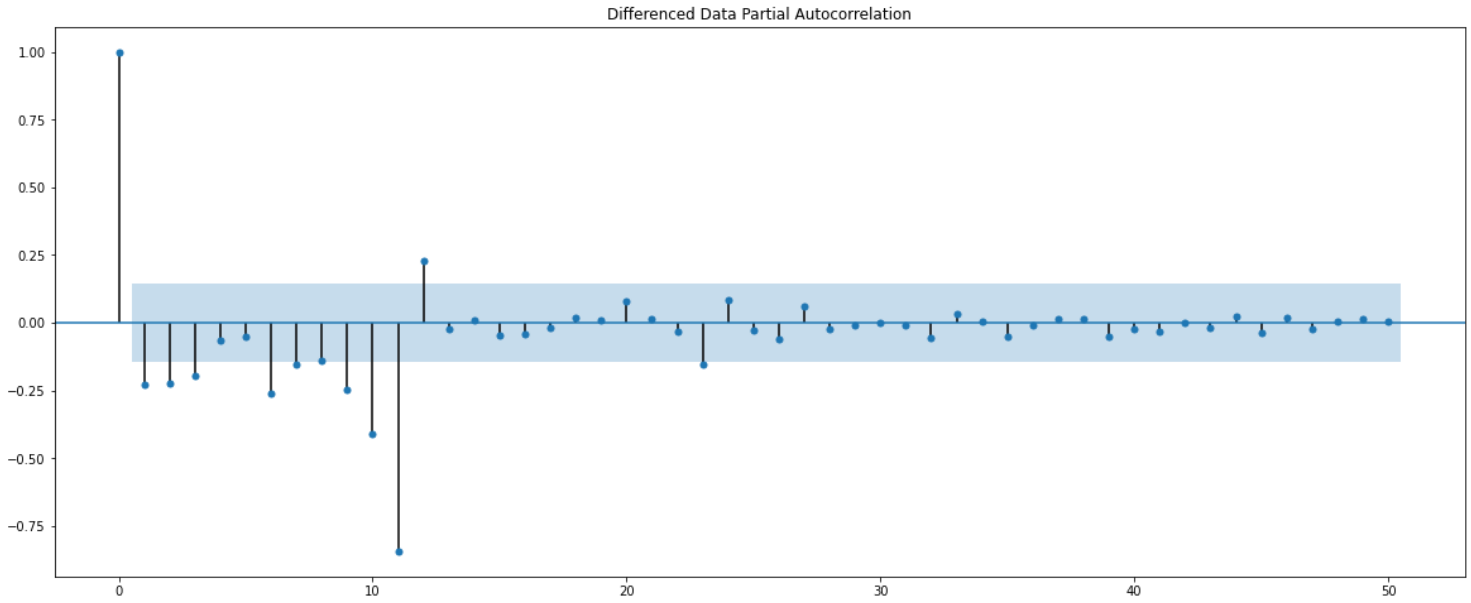


Figure 7.1(a)

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 3.

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

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(3, 1, 2)   Log Likelihood               -1107.464
Method:                       css-mle   S.D. of innovations           1106.180
Date:                Wed, 11 Aug 2021   AIC                           2228.928
Time:                        14:30:59   BIC                           2249.054
Sample:                    02-29-1980   HQIC                          2237.106
                         - 12-31-1990
==============================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                 5.9886      3.644      1.643      0.103      -1.154      13.131
ar.L1.D.Sparkling    -0.4420   8.18e-06  -5.41e+04      0.000      -0.442      -0.442
ar.L2.D.Sparkling     0.3081   2.58e-05   1.19e+04      0.000       0.308       0.308
ar.L3.D.Sparkling    -0.2499    2.2e-05  -1.13e+04      0.000      -0.250      -0.250
ma.L1.D.Sparkling    -0.0007      0.020     -0.034      0.973      -0.040       0.038
ma.L2.D.Sparkling    -0.9993      0.020    -50.450      0.000      -1.038      -0.961
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -1.0000           -0.0000j            1.0000           -0.5000
AR.2            1.1165           -1.6600j            2.0005           -0.1558
AR.3            1.1165           +1.6600j            2.0005            0.1558
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.0007           +0.0000j            1.0007            0.5000
-----------------------------------------------------------------------------
```

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

- The RMSE value for ARIMA (3,1,2) is 1379.1826761518028

ROSE

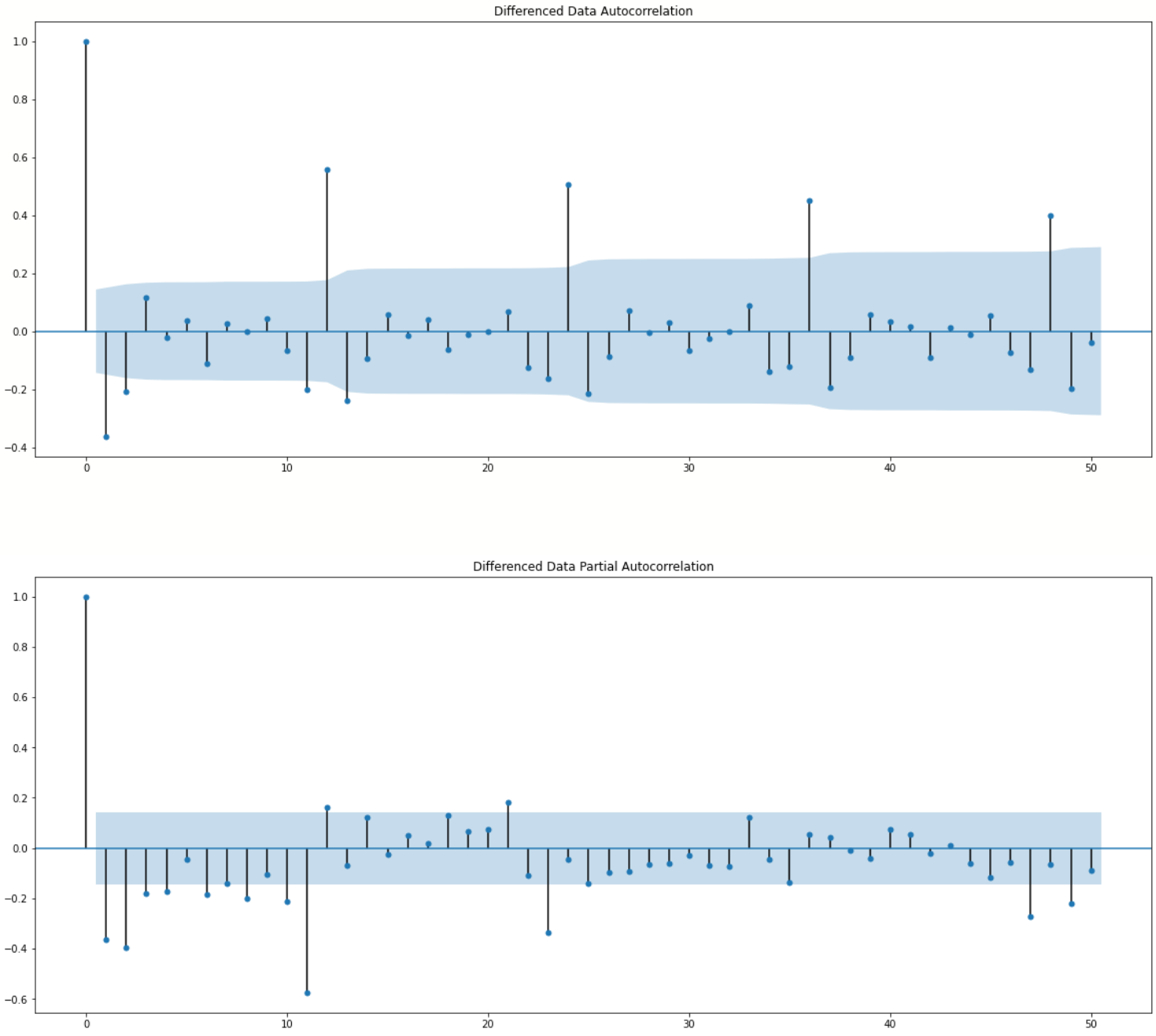


Figure 7.1(b)

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. But we will take its value as 3 to prevent a complex model.

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

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(3, 1, 2)	Log Likelihood	-633.485
Method:	css-mle	S.D. of innovations	29.950
Date:	Thu, 12 Aug 2021	AIC	1280.969
Time:	11:02:14	BIC	1301.096
Sample:	02-29-1980 - 12-31-1990	HQIC	1289.148

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4883	0.085	-5.723	0.000	-0.656	-0.321
ar.L1.D.Rose	-0.3558	0.332	-1.071	0.286	-1.007	0.296
ar.L2.D.Rose	0.0279	0.120	0.232	0.817	-0.208	0.264
ar.L3.D.Rose	0.0598	0.104	0.577	0.565	-0.143	0.263
ma.L1.D.Rose	-0.4141	0.325	-1.274	0.205	-1.051	0.223
ma.L2.D.Rose	-0.5857	0.324	-1.810	0.073	-1.220	0.049

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.8009	-1.4473j	2.3104	-0.3923
AR.2	-1.8009	+1.4473j	2.3104	0.3923
AR.3	3.1348	-0.0000j	3.1348	-0.0000
MA.1	1.0001	+0.0000j	1.0001	0.0000
MA.2	-1.7072	+0.0000j	1.7072	0.5000

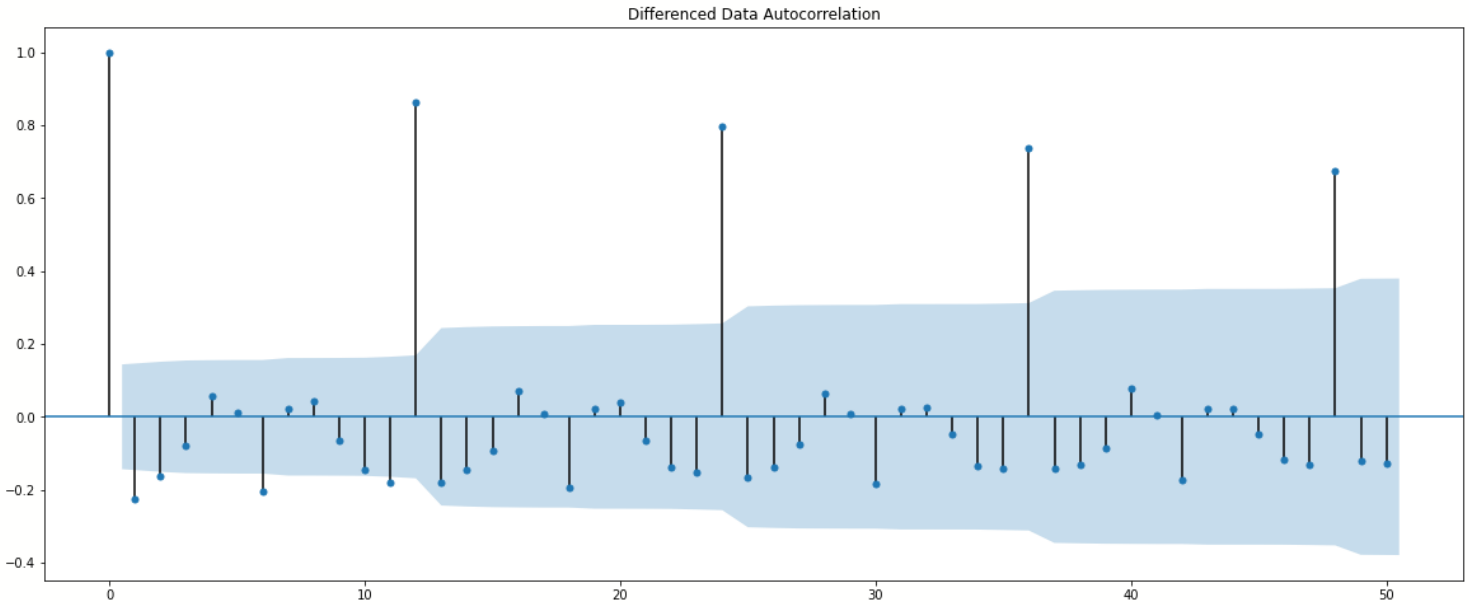
PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

- The RMSE value for ARIMA (3,1,2) is 15.52288708797049

- SARIMA

SPARKLING

Let us look at the ACF and the PACF plots once more.



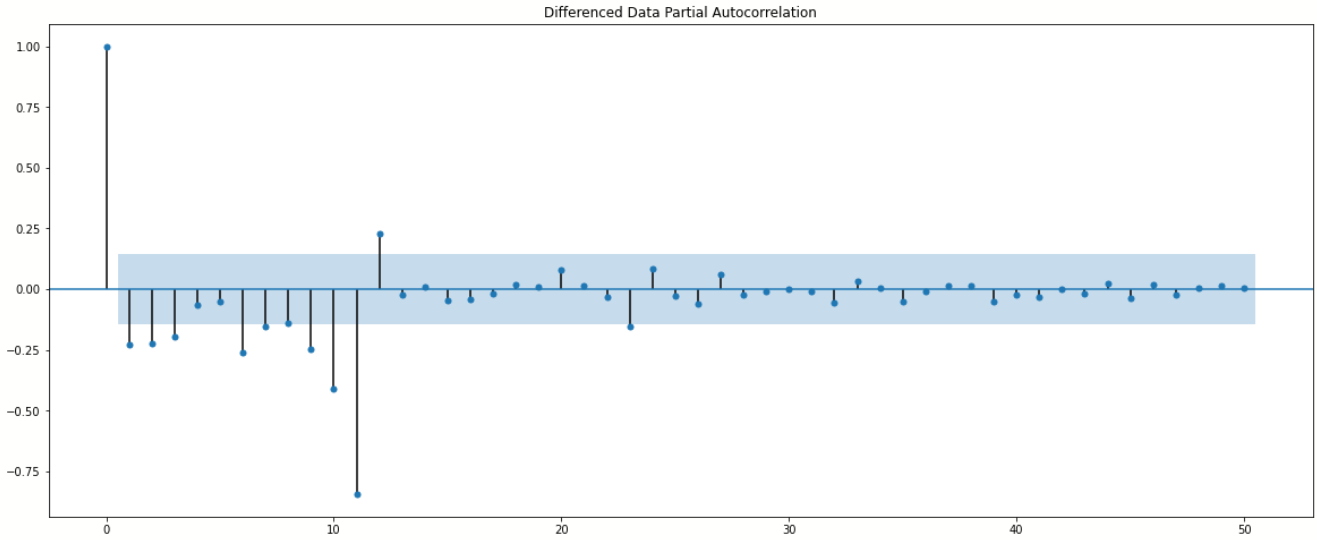


Figure 7.2(a)

We see that our ACF plot at the seasonal interval (6) does begin to taper off.

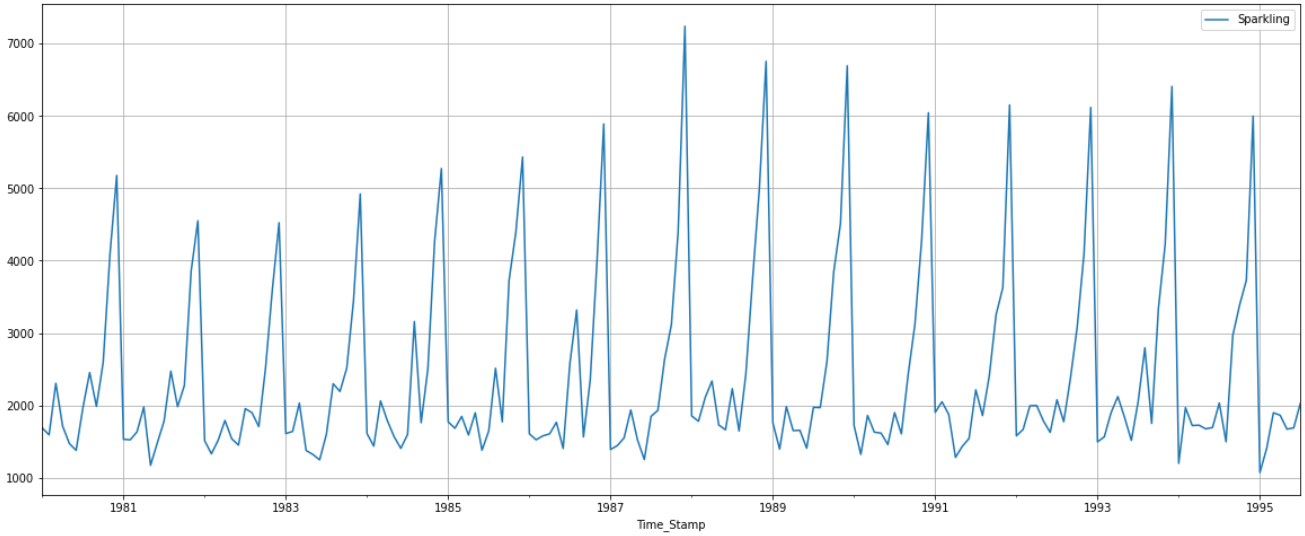


Figure 7.2(b)

- Seasonality = 6

We see that there is a little trend and seasonality. So, now we take a seasonal differencing and check the series.

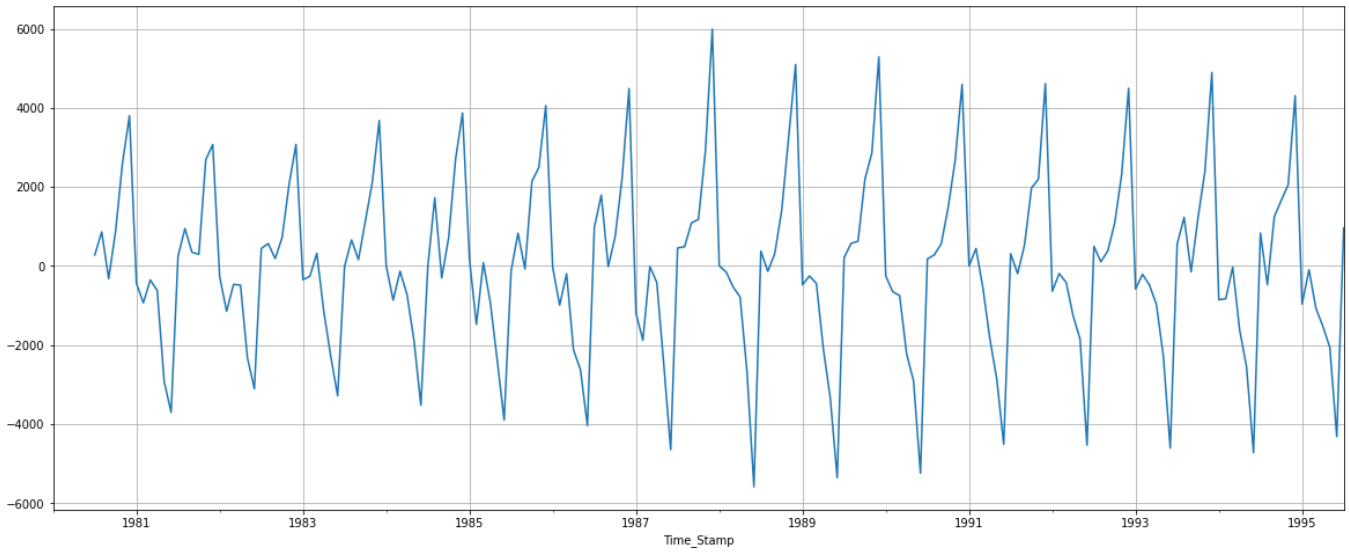


Figure 7.2(c)

Now we see that there is almost no trend present in the data. Seasonality is only present in the data. Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.

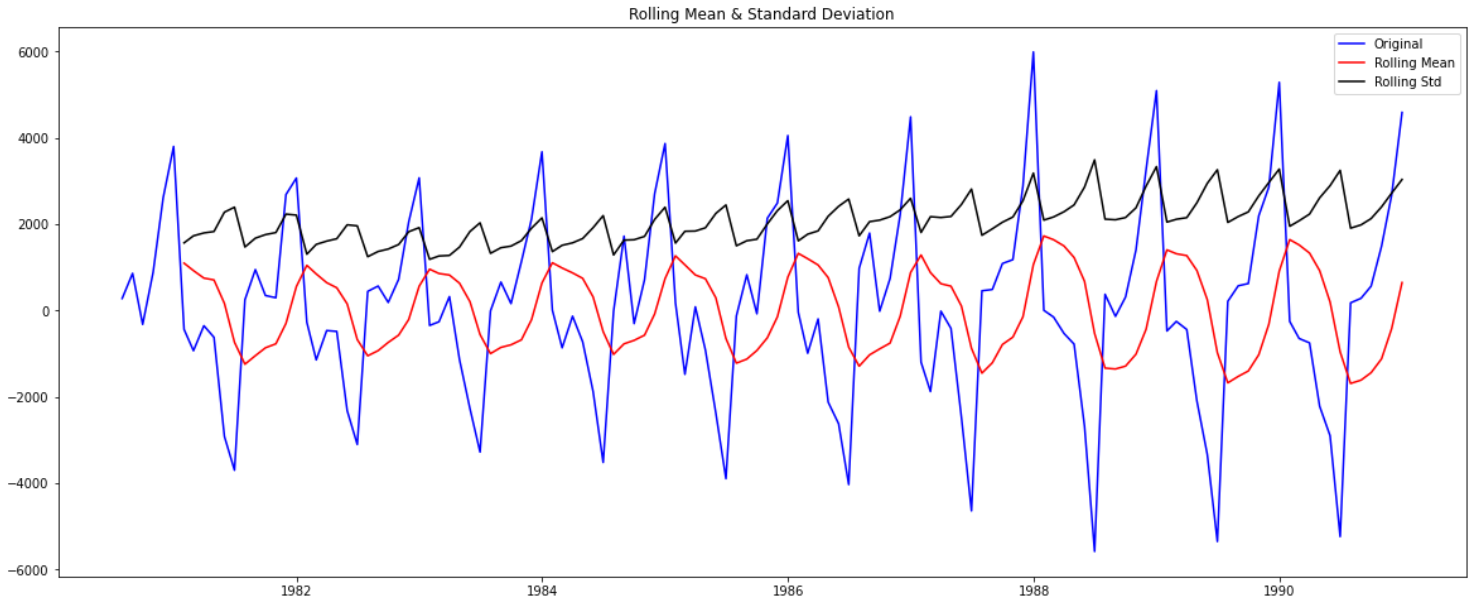
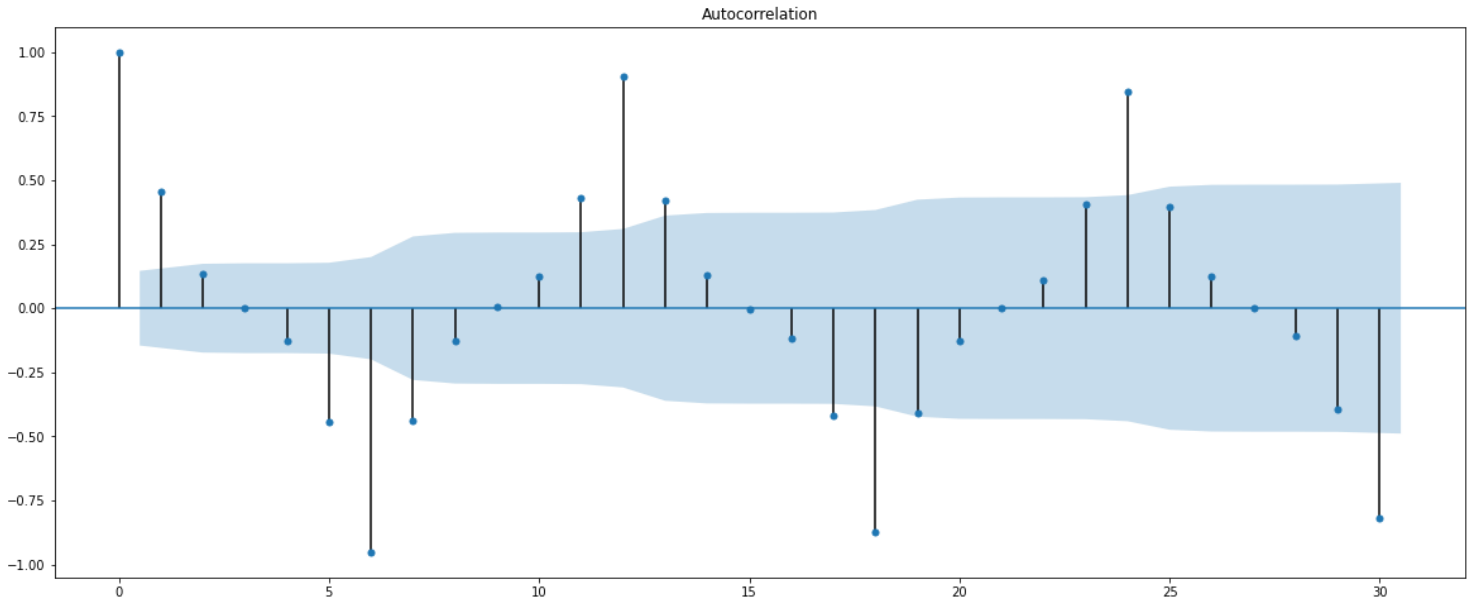


Figure 7.2(d)

```
Results of Dickey-Fuller Test:
Test Statistic                  -8.181919e+00
p-value                          8.088278e-13
#Lags Used                       6.000000e+00
Number of Observations Used      1.190000e+02
Critical Value (1%)             -3.486535e+00
Critical Value (5%)             -2.886151e+00
Critical Value (10%)            -2.579896e+00
```

As observed we have a p-value lesser than our alpha 0.05. This explains that our time series is stationary therefore we won't difference it further.

Checking the ACF and the PACF plots for the new modified Time Series



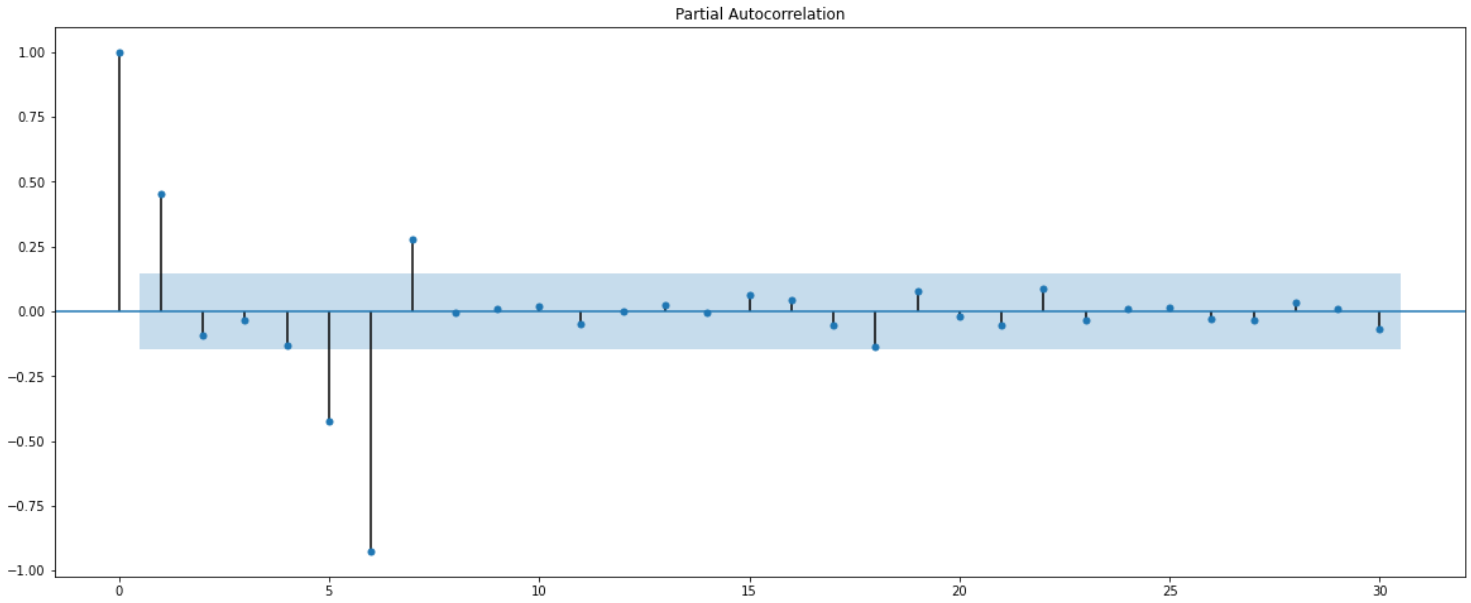


Figure 7.2(e)

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 6.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 1.

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 to prevent model complexity.

Remember to check the ACF and the PACF plots only at multiples of 6 (since 6 is the seasonal period).

```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(1, 0, 3, 6)   Log Likelihood                 845.298
Date:                           Thu, 12 Aug 2021   AIC                           1710.596
Time:                                   11:48:59   BIC                           1737.601
Sample:                                        0   HQIC                          1721.549
                                             - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6986      0.105     -6.664      0.000      -0.904      -0.493
ar.L2         -0.0459      0.129     -0.357      0.721      -0.298       0.206
ar.L3         -0.3418      0.091     -3.737      0.000      -0.521      -0.163
ma.L1          0.0326      1.722      0.019      0.985      -3.342       3.407
ma.L2         -0.9675      1.655     -0.584      0.559      -4.212       2.277
ar.S.L6       -1.0288      0.010   -104.687      0.000      -1.048      -1.010
ma.S.L6        1.3230      1.691      0.783      0.434      -1.991       4.637
ma.S.L12       0.8141      0.587      1.386      0.166      -0.337       1.965
ma.S.L18       0.4979      0.809      0.616      0.538      -1.087       2.083
sigma2      2.404e+05   1.36e-05   1.77e+10      0.000     2.4e+05     2.4e+05
==========================================================================================
Ljung-Box (Q):                      110.87   Jarque-Bera (JB):                 4.27
Prob(Q):                              0.00   Prob(JB):                         0.12
Heteroskedasticity (H):               2.21   Skew:                             0.38
Prob(H) (two-sided):                  0.02   Kurtosis:                         3.58
==========================================================================================
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Diagnostic Plot

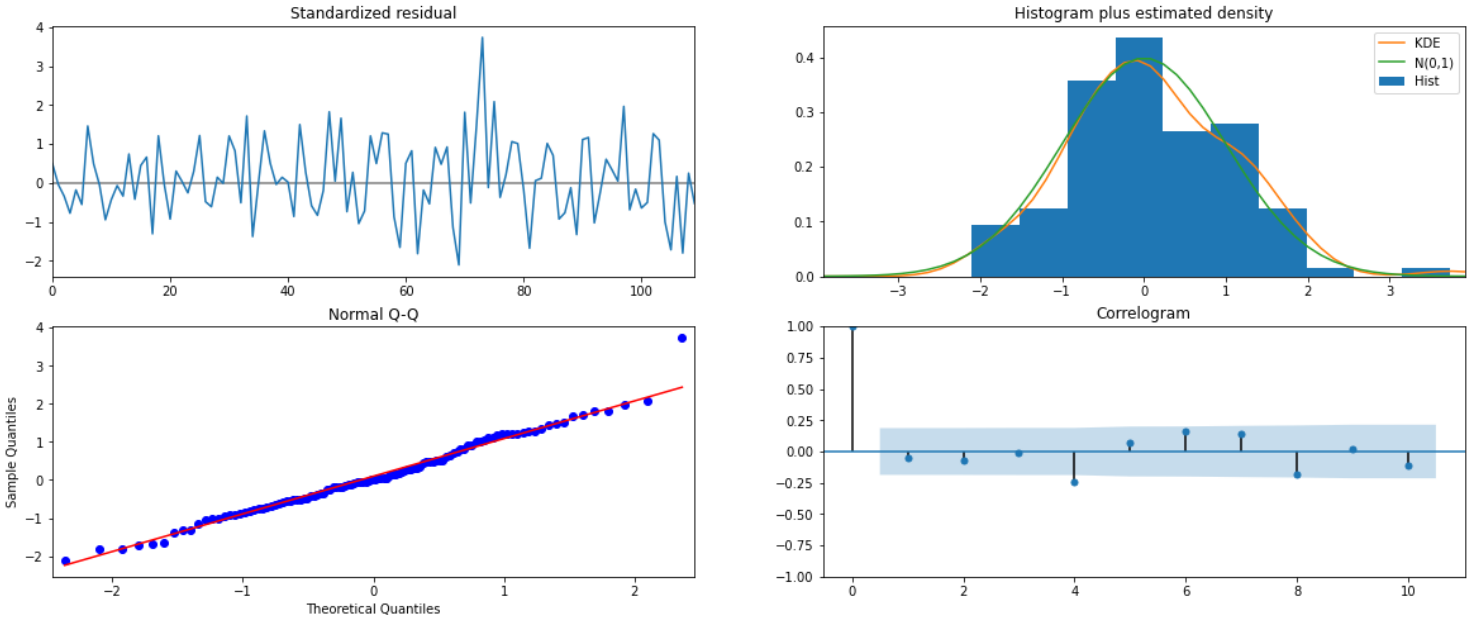


Figure 7.2(f)

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

DATA SUMMARY

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1963.295813	505.709120	972.124150	2954.467475
1	1630.593769	531.259466	589.344350	2671.843189
2	1934.592530	541.662668	872.953208	2996.231852
3	1586.197130	546.614345	514.852701	2657.541558
4	950.971966	547.537833	-122.182468	2024.126400

- The RMSE value for ARIMA (3,1,2) is 902.7907259447834

- Seasonality = 12

We see that there is a little trend and seasonality. So, now we take a seasonal differencing and check the series.

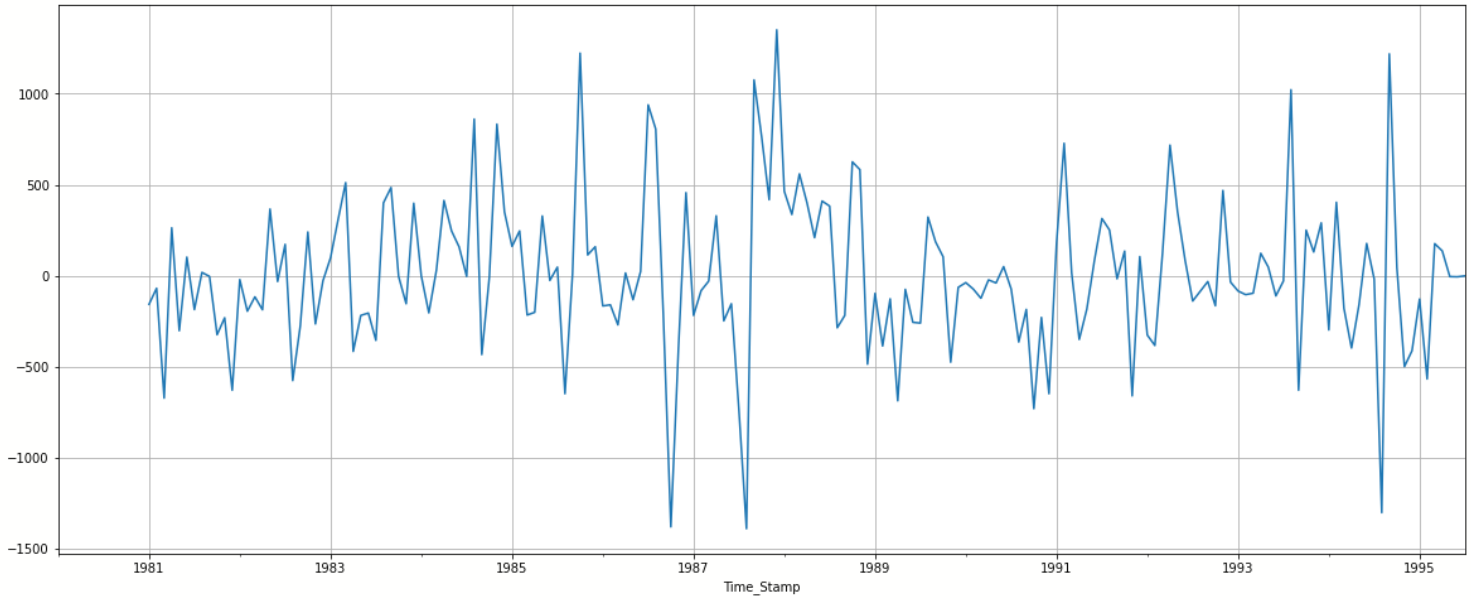


Figure 7.3(a)

Now we see that there is almost no trend present in the data. Seasonality is only present in the data.

Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.

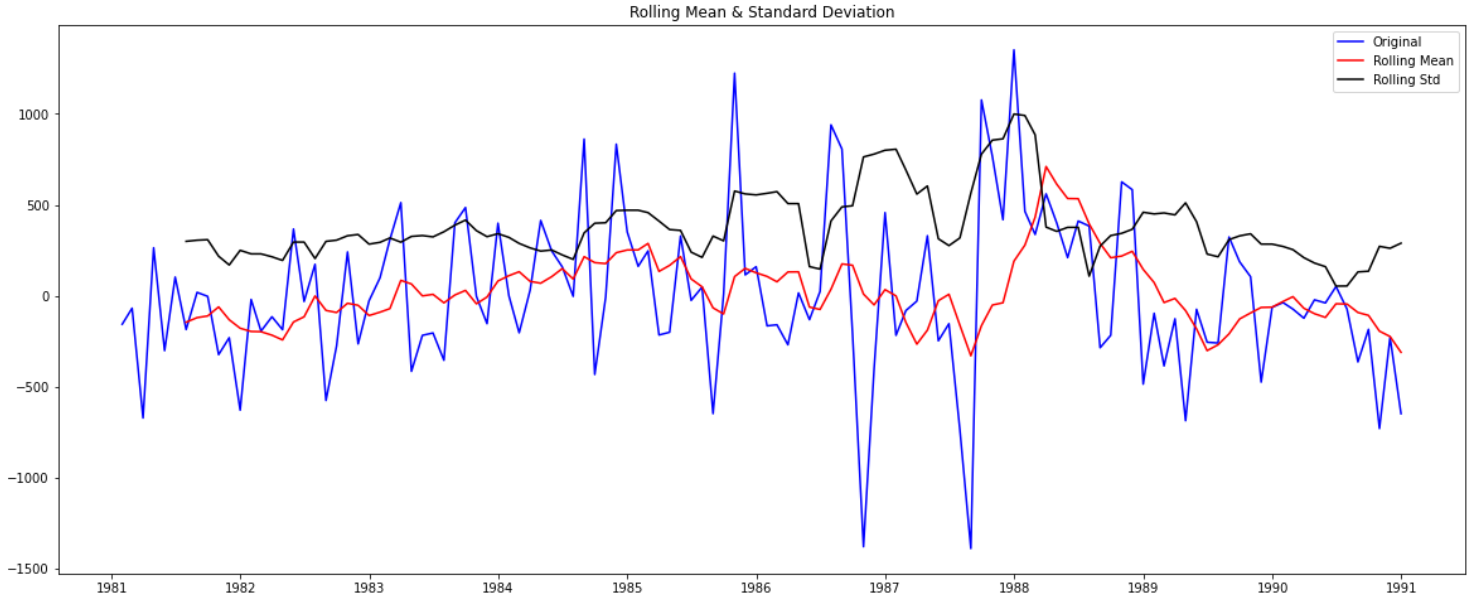


Figure 7.3(b)

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.136812
p-value                          0.023946
#Lags Used                      11.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
dtype: float64
```

As observed we have a p-value lesser than our alpha 0.05. This explains that our time series is stationary therefore we won't difference it further.

Checking the ACF and the PACF plots for the new modified Time Series

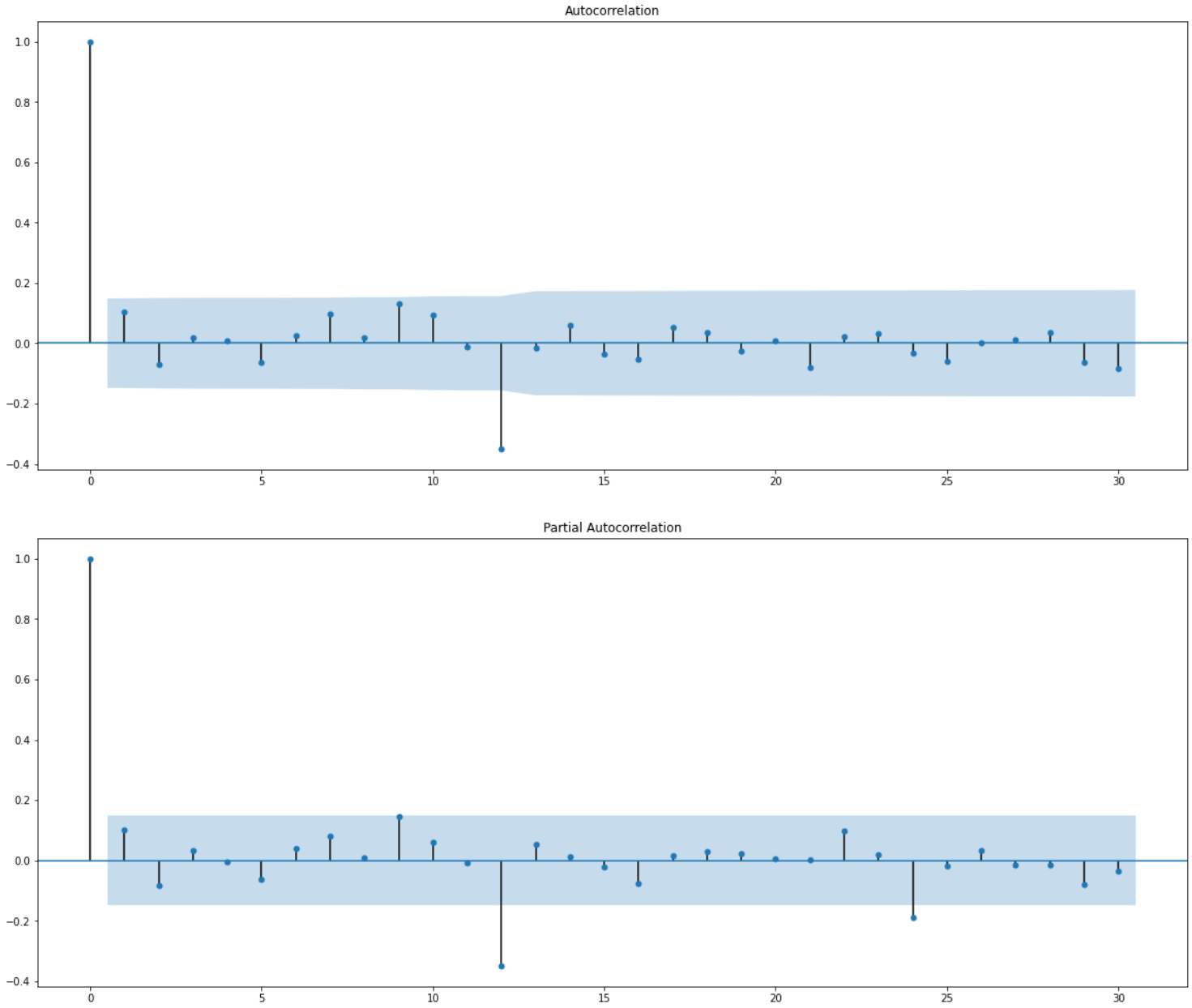


Figure 7.3(c)

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 2.

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 1.

Remember to check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period).

Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 2)x(2, 0, 1, 12)	Log Likelihood	771.395
Date:	Thu, 12 Aug 2021	AIC	1560.790
Time:	12:32:45	BIC	1584.589
Sample:	0 - 132	HQIC	1570.432
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6518	0.313	-2.081	0.037	-1.266	-0.038
ar.L2	-0.0463	0.178	-0.260	0.795	-0.396	0.303
ar.L3	0.0172	0.134	0.129	0.898	-0.245	0.279
ma.L1	-0.3499	0.244	-1.432	0.152	-0.829	0.129
ma.L2	-0.8706	0.331	-2.633	0.008	-1.519	-0.223
ar.S.L12	1.0340	0.226	4.578	0.000	0.591	1.477
ar.S.L24	0.0067	0.226	0.030	0.976	-0.436	0.450
ma.S.L12	-0.5825	0.229	-2.546	0.011	-1.031	-0.134
sigma2	1.228e+05	2.58e+04	4.761	0.000	7.22e+04	1.73e+05

Ljung-Box (Q):	24.26	Jarque-Bera (JB):	11.01
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	1.53	Skew:	0.33
Prob(H) (two-sided):	0.22	Kurtosis:	4.45

Diagnostic Plot

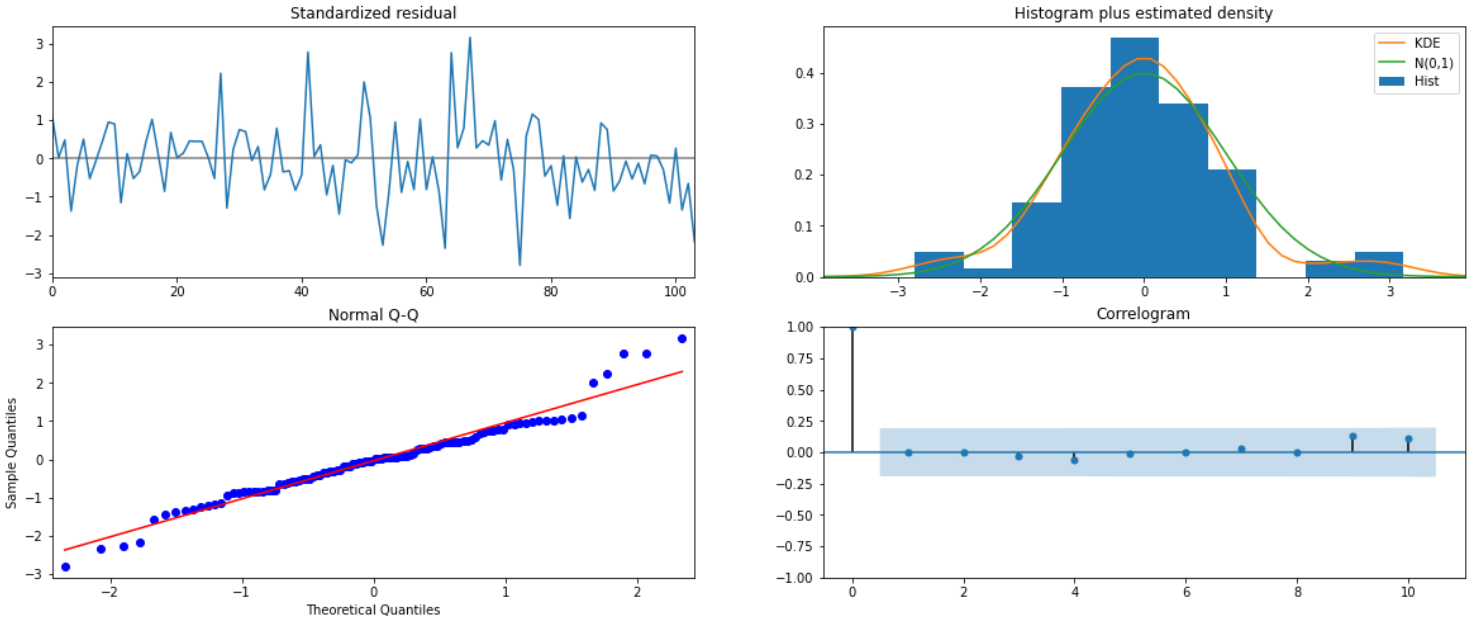


Figure 7.3(d)

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

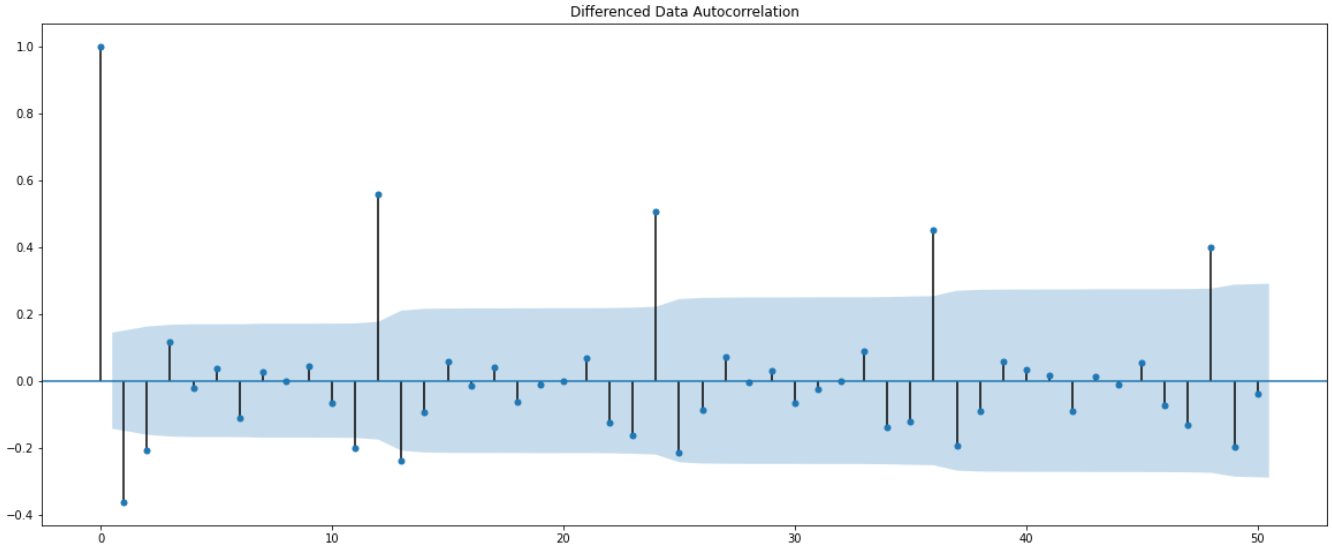
PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

DATA SUMMARY

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1309.966523	393.927789	537.882245	2082.050801
1	1257.747627	404.488599	464.964542	2050.530713
2	1555.138035	404.489335	762.353506	2347.922565
3	1540.082836	412.389501	731.814268	2348.351405
4	1341.694655	413.241561	531.756078	2151.633232

- The RMSE value for ARIMA (3,1,2) is 611.0380372403605

ROSE



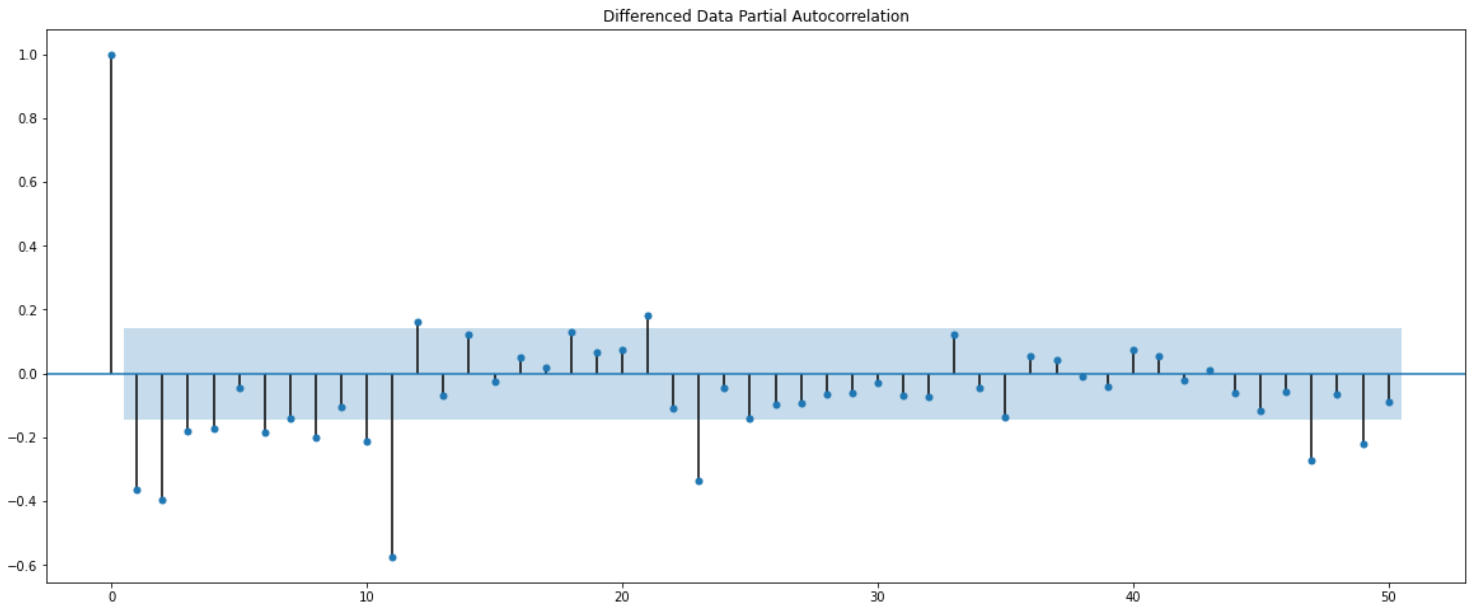


Figure 7.4(a)

We see that our ACF plot at the seasonal interval (6) does begin to taper off.

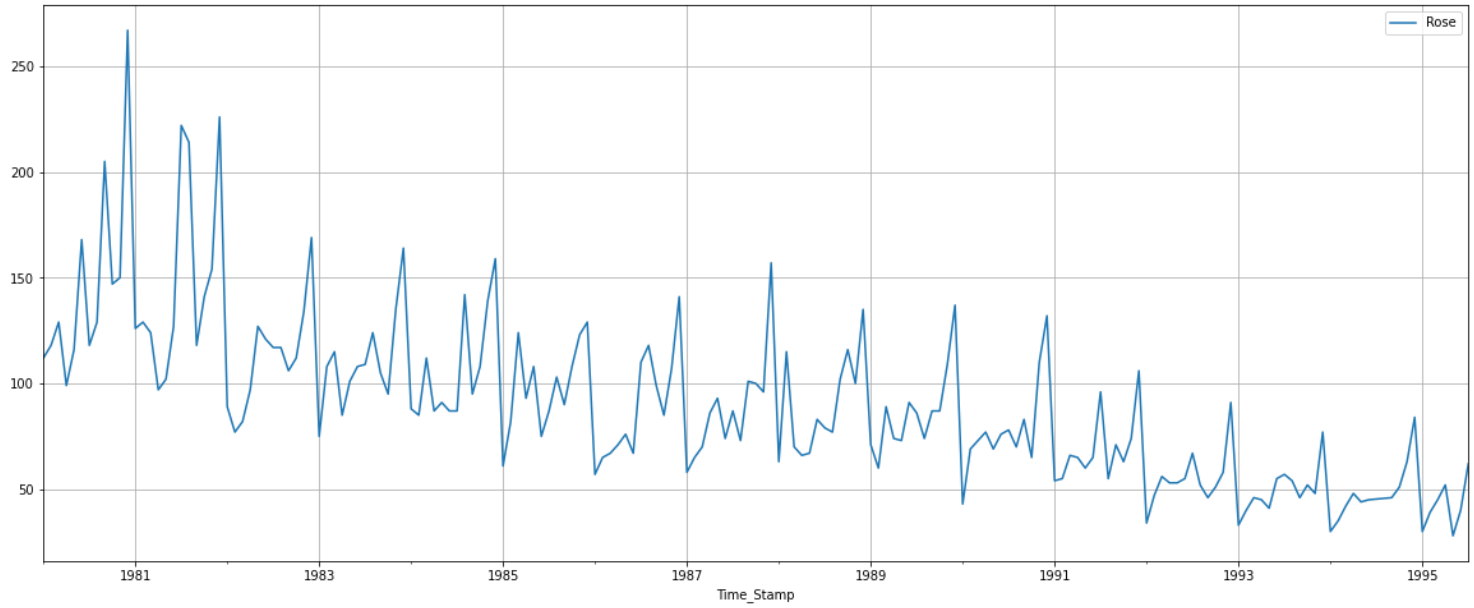


Figure 7.4(b)

- Seasonality = 6

We see that there is a downward trend and seasonality. So, now we take a seasonal differencing and check the series.

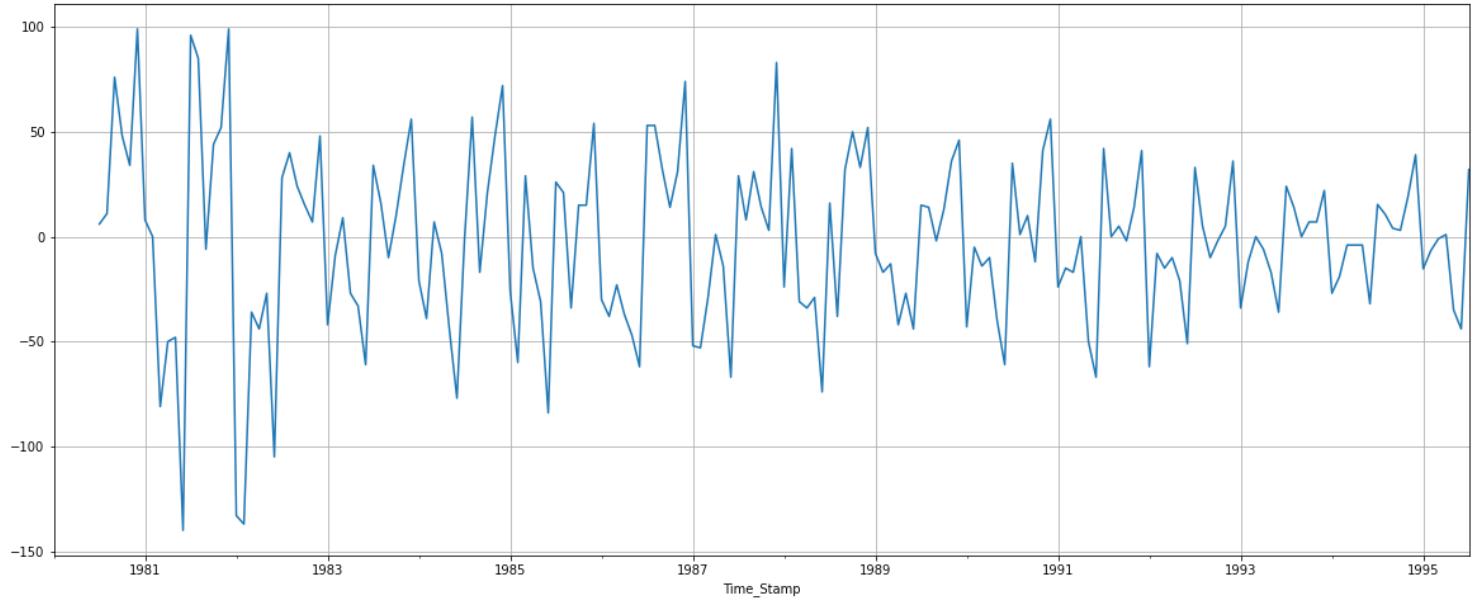


Figure 7.5(a)

Now we see that there is almost no trend present in the data. Seasonality is only present in the data. Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.

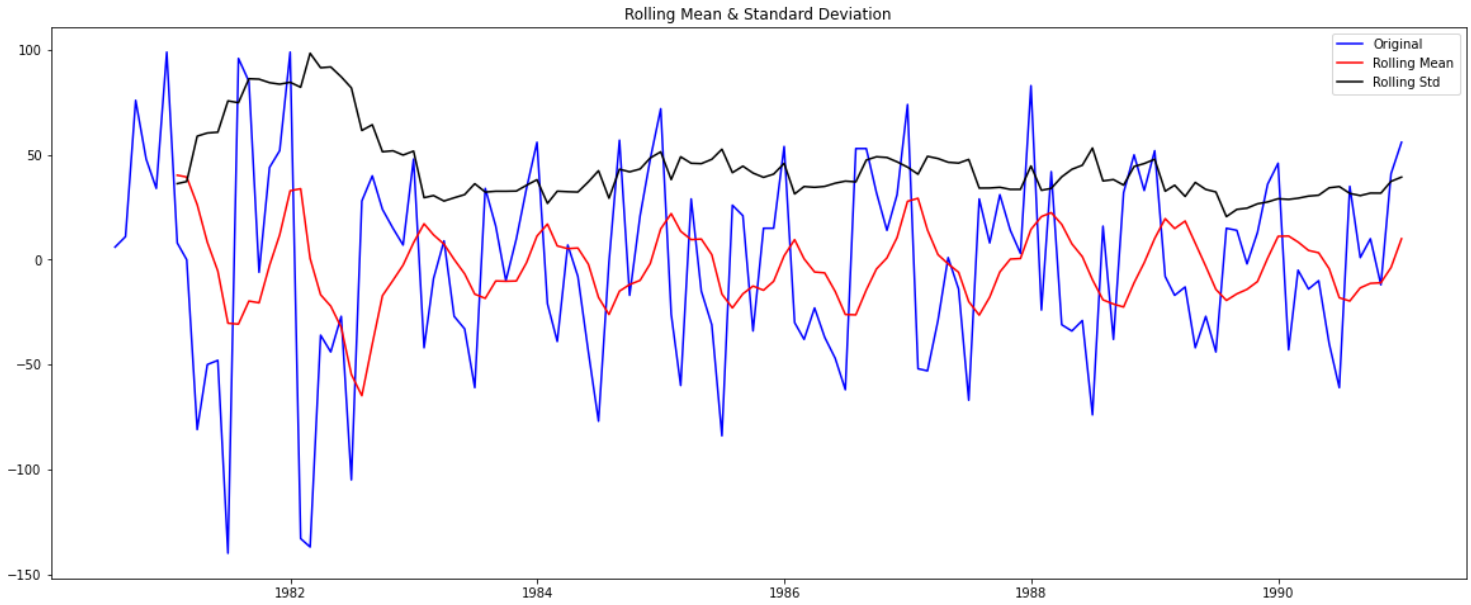
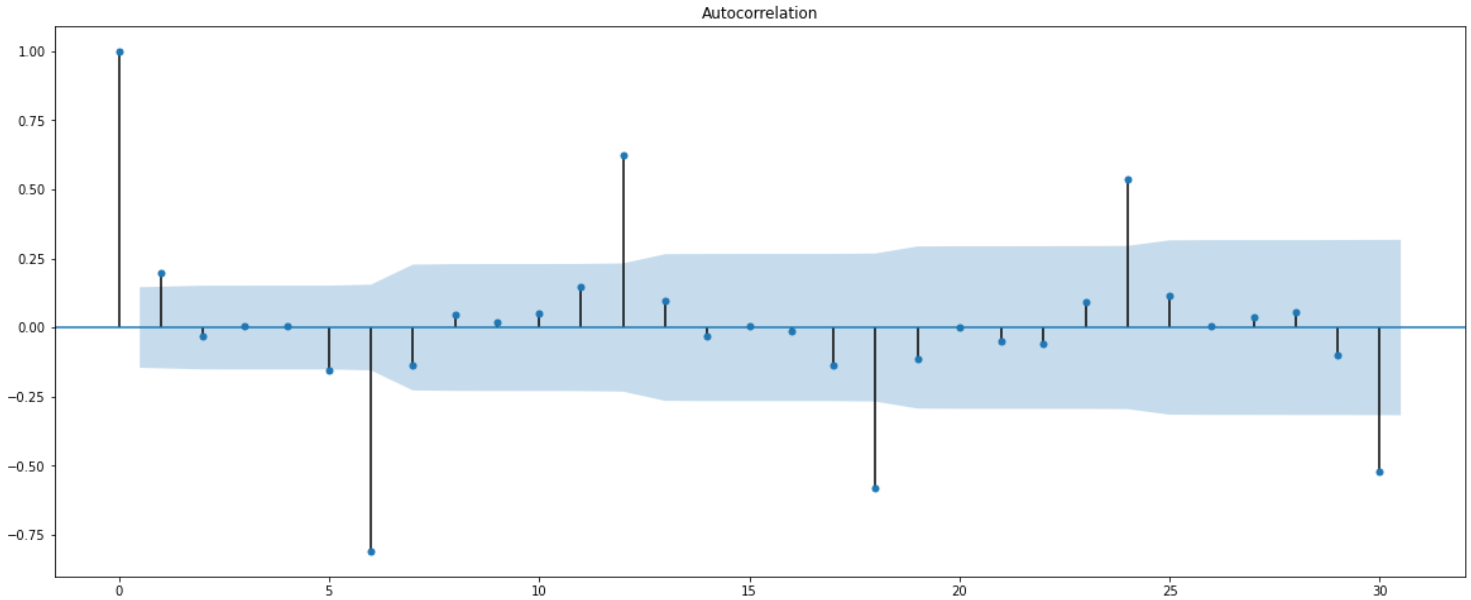


Figure 7.5(b)

```
Results of Dickey-Fuller Test:
Test Statistic                  -7.442449e+00
p-value                          5.956534e-11
#Lags Used                       7.000000e+00
Number of Observations Used      1.180000e+02
Critical Value (1%)             -3.487022e+00
Critical Value (5%)             -2.886363e+00
Critical Value (10%)            -2.580009e+00
```

As observed we have a p-value lesser than our alpha 0.05. This explains that our time series is stationary therefore we won't difference it further.

Checking the ACF and the PACF plots for the new modified Time Series



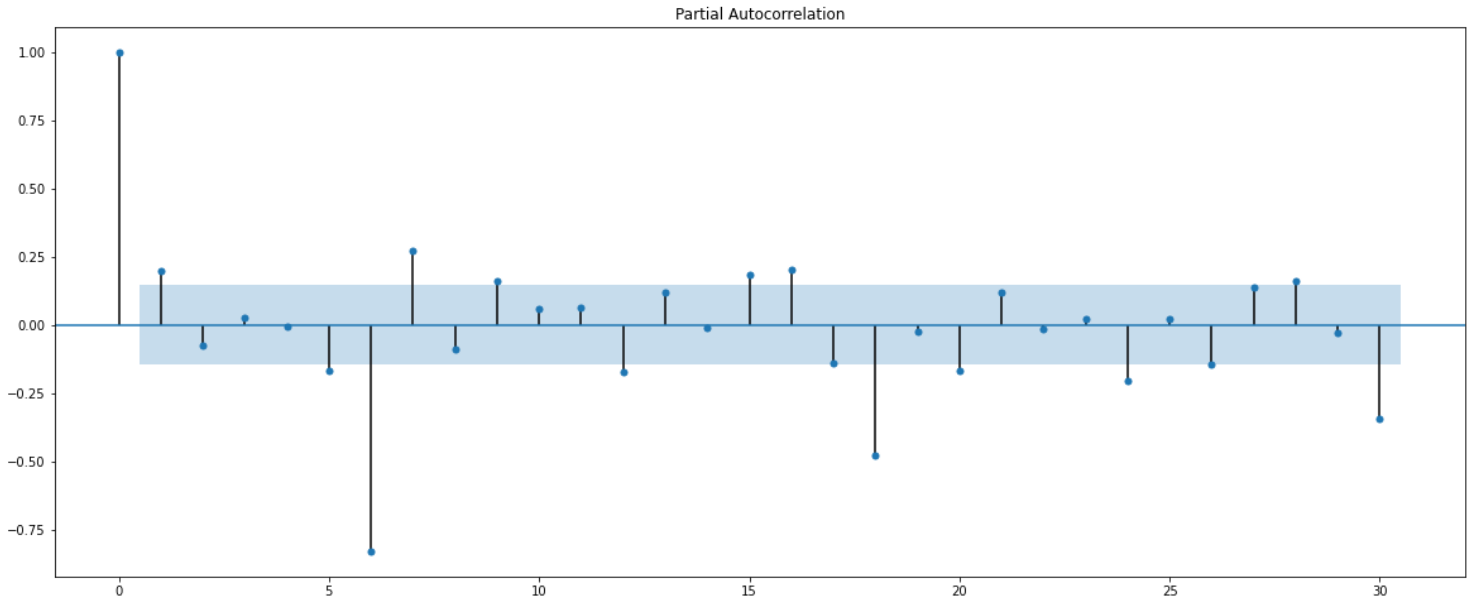


Figure 7.5(c)

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 6.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 3 to reduce complexity.

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 to reduce complexity.

Remember to check the ACF and the PACF plots only at multiples of 6 (since 6 is the seasonal period).

Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 2)x(3, 0, 3, 6)	Log Likelihood	470.135
Date:	Thu, 12 Aug 2021	AIC	964.271
Time:	13:07:01	BIC	996.676
Sample:	0	HQIC	977.415
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1695	0.660	0.257	0.797	-1.124	1.463
ar.L2	-0.1613	0.214	-0.755	0.450	-0.580	0.257
ar.L3	-0.0990	0.194	-0.510	0.610	-0.479	0.281
ma.L1	-0.9294	543.081	-0.002	0.999	-1065.348	1063.489
ma.L2	-0.0706	38.142	-0.002	0.999	-74.827	74.686
ar.S.L6	-0.1251	0.128	-0.975	0.329	-0.376	0.126
ar.S.L12	0.7216	0.058	12.347	0.000	0.607	0.836
ar.S.L18	0.0361	0.120	0.301	0.764	-0.199	0.271
ma.S.L6	0.1218	0.207	0.587	0.557	-0.285	0.528
ma.S.L12	-0.2746	0.127	-2.170	0.030	-0.523	-0.027
ma.S.L18	0.0773	0.159	0.487	0.626	-0.234	0.388
sigma2	287.4563	1.56e+05	0.002	0.999	-3.06e+05	3.06e+05

Ljung-Box (Q):	27.85	Jarque-Bera (JB):	4.79
Prob(Q):	0.93	Prob(JB):	0.09
Heteroskedasticity (H):	0.82	Skew:	-0.06
Prob(H) (two-sided):	0.54	Kurtosis:	4.02

Diagnostic Plot

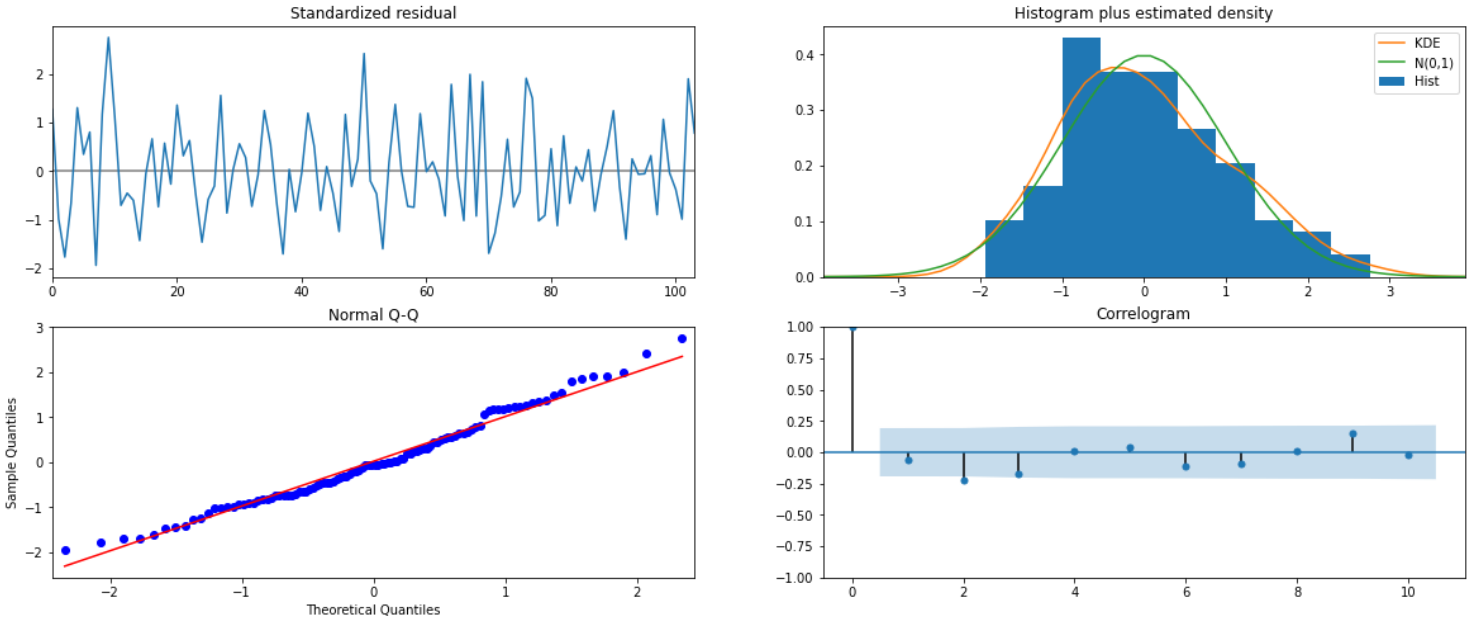


Figure 7.5(d)

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

DATA SUMMARY

	y	mean	mean_se	mean_ci_lower	mean_ci_upper
0		65.318738	17.031776	31.937070	98.700406
1		70.705074	17.551763	36.304250	105.105898
2		74.968885	17.649271	40.376949	109.560821
3		76.843507	17.827611	41.902032	111.784981
4		75.606380	17.831093	40.658079	110.554680

- The RMSE value for ARIMA (3,1,2) is 29.93440417797009

- Seasonality = 12

We see that there is a little trend and seasonality. So, now we take a seasonal differencing and check the series.

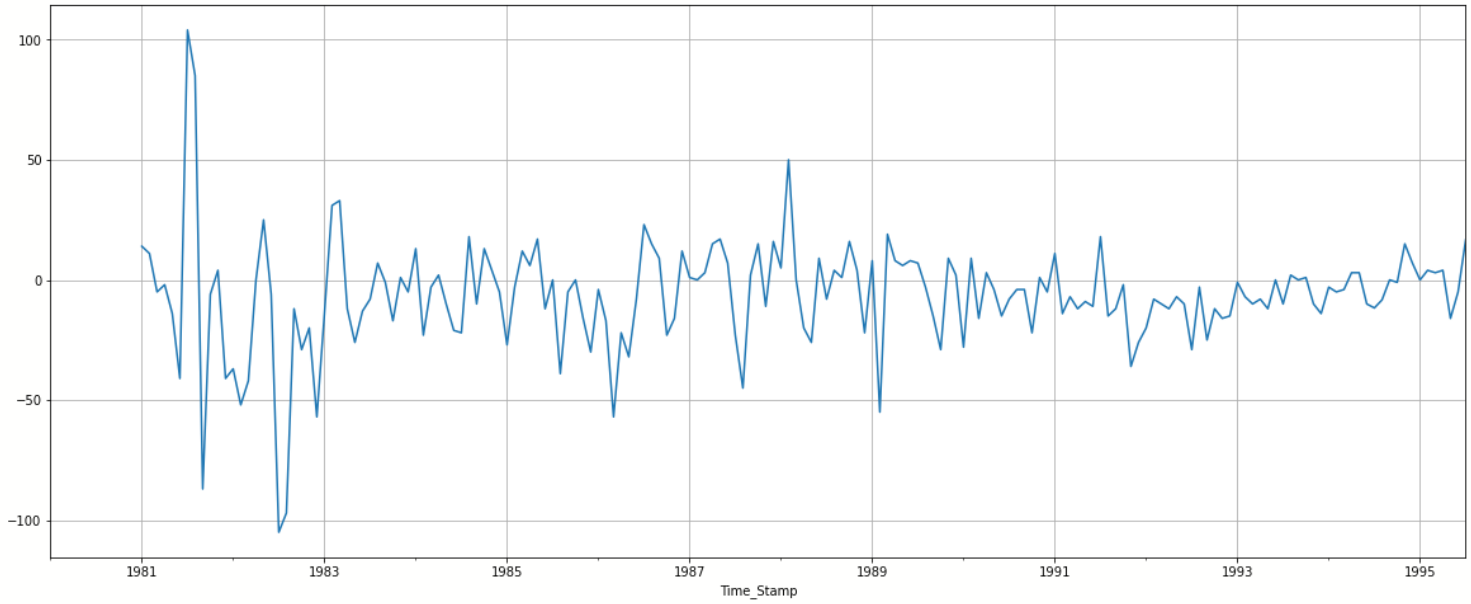


Figure 7.6(a)

Now we see that there is almost no trend present in the data. Seasonality is only present in the data. Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.

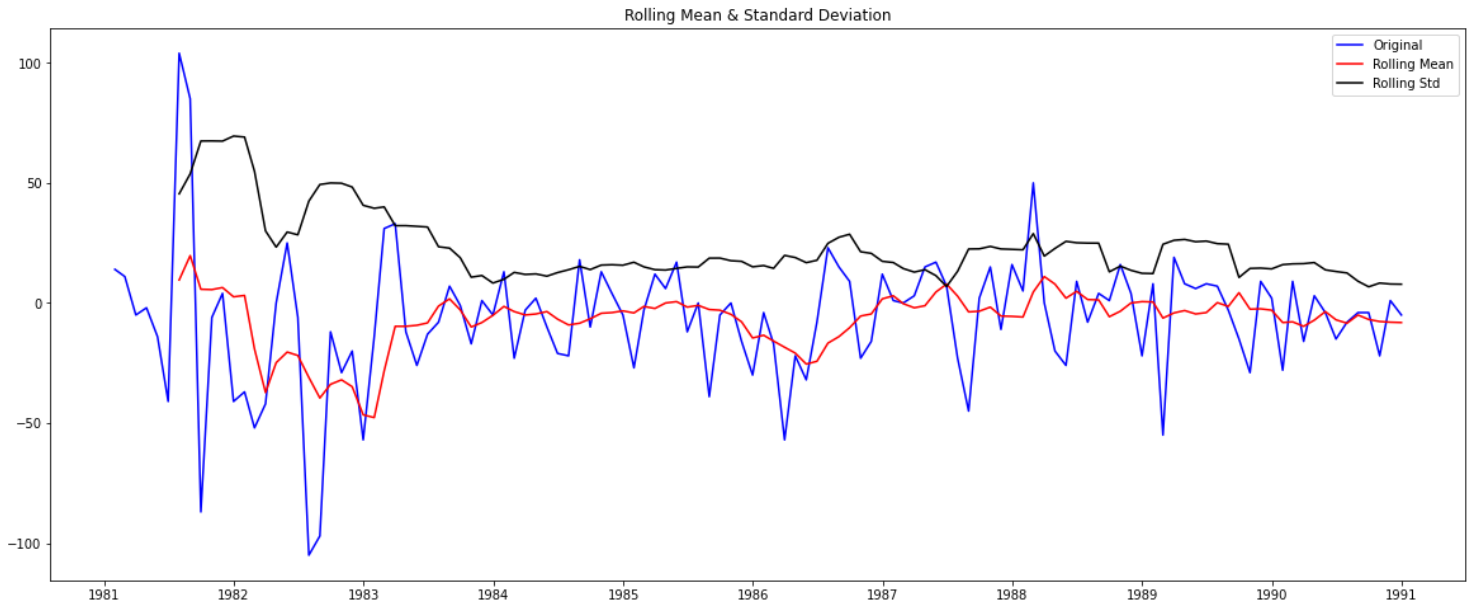


Figure 7.6(b)

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.619482
p-value                          0.005399
#Lags Used                      11.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
```

As observed we have a p-value lesser than our alpha 0.05. This explains that our time series is stationary therefore we won't difference it further.

Checking the ACF and the PACF plots for the new modified Time Series

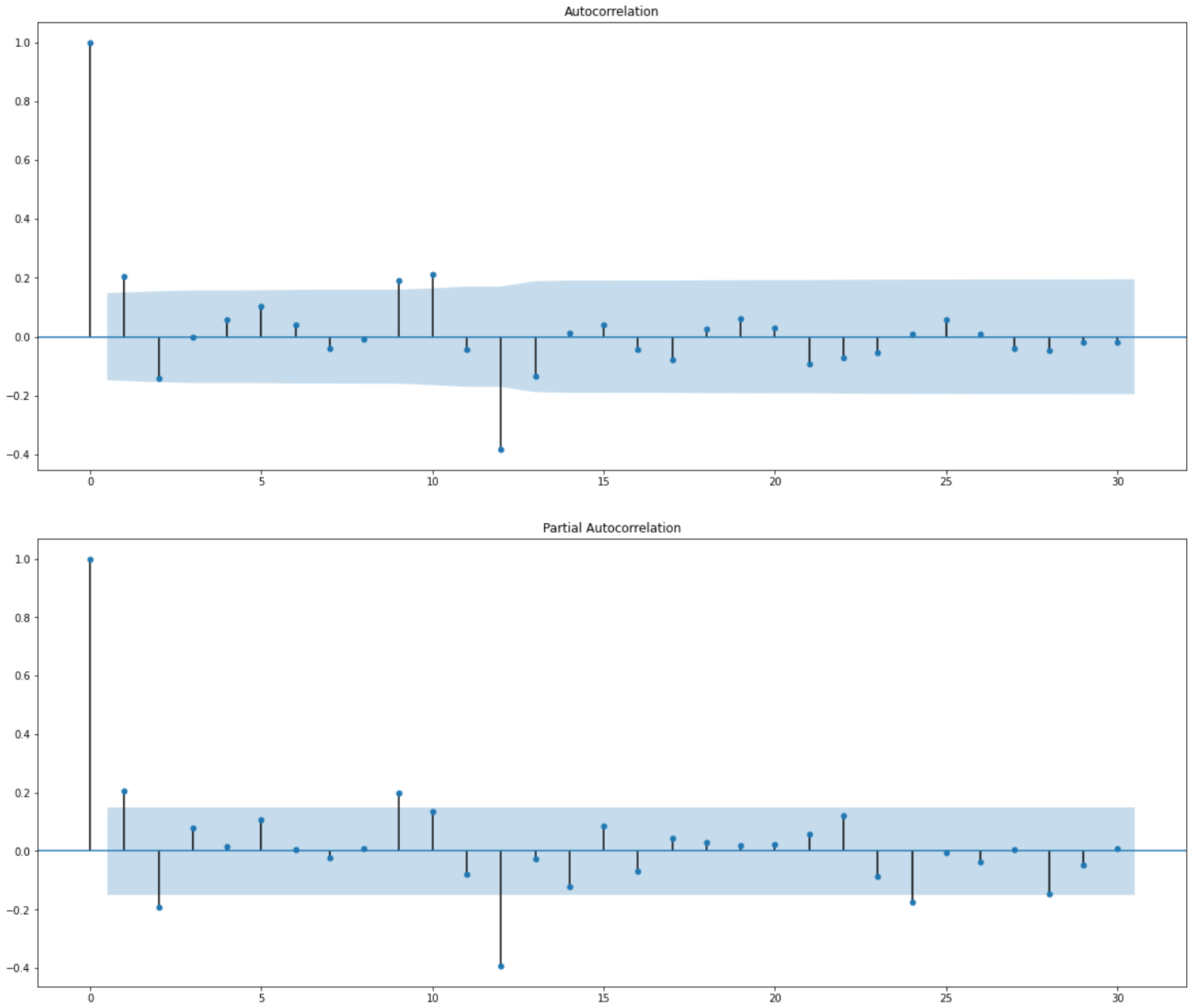


Figure 7.6(c)

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 2.

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 1.

Remember to check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period).

Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 2)x(2, 0, 1, 12)	Log Likelihood	436.164
Date:	Thu, 12 Aug 2021	AIC	890.328
Time:	13:33:52	BIC	914.127
Sample:	0 - 132	HQIC	899.970
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1437	0.780	0.184	0.854	-1.384	1.672
ar.L2	-0.0757	0.159	-0.477	0.634	-0.387	0.235
ar.L3	-0.0817	0.152	-0.537	0.591	-0.380	0.217
ma.L1	-1.0058	550.867	-0.002	0.999	-1080.684	1078.673
ma.L2	0.0058	3.425	0.002	0.999	-6.708	6.719
ar.S.L12	0.3383	0.080	4.214	0.000	0.181	0.496
ar.S.L24	0.2808	0.069	4.090	0.000	0.146	0.415
ma.S.L12	0.1322	0.131	1.010	0.312	-0.124	0.389
sigma2	247.0643	1.36e+05	0.002	0.999	-2.66e+05	2.67e+05

Ljung-Box (Q):	25.01	Jarque-Bera (JB):	2.71
Prob(Q):	0.97	Prob(JB):	0.26
Heteroskedasticity (H):	0.96	Skew:	0.36
Prob(H) (two-sided):	0.91	Kurtosis:	3.31

Diagnostic Plot

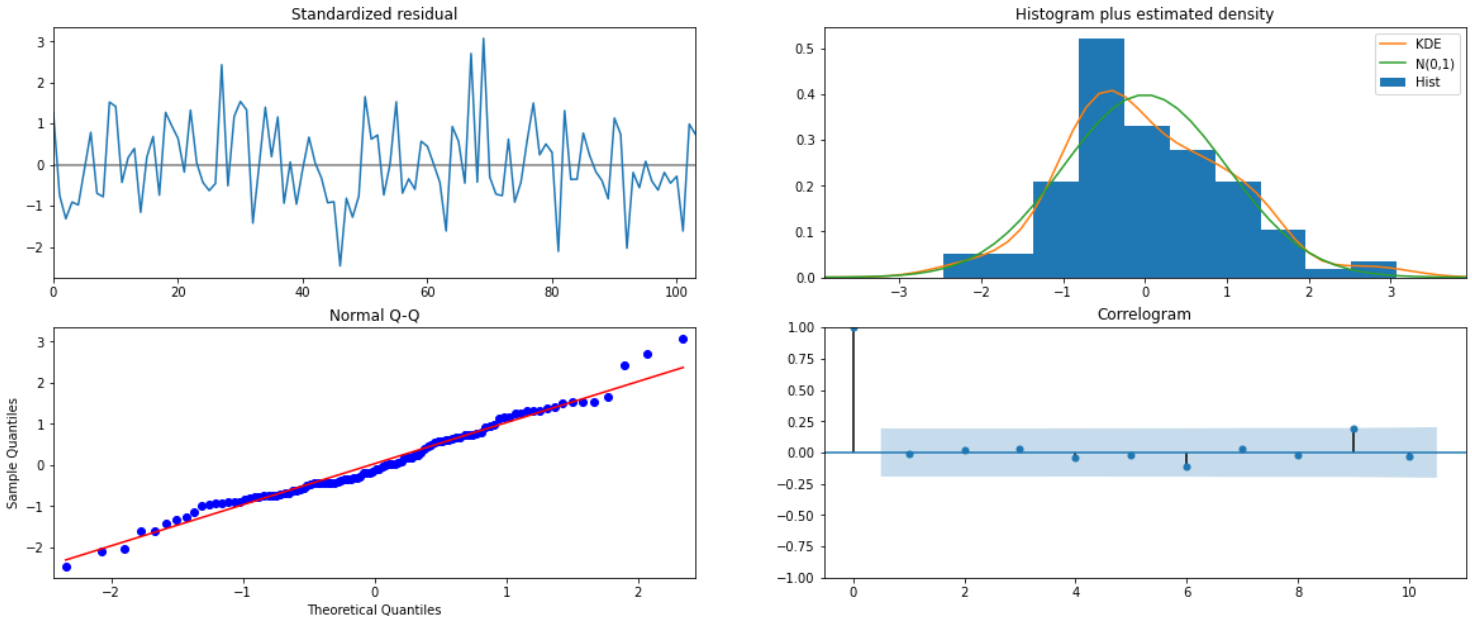


Figure 7.6(d)

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

PREDICT ON THE TEST SET USING THIS MODEL AND EVALUATE THE MODEL.

DATA SUMMARY

	y	mean	mean_se	mean_ci_lower	mean_ci_upper
0		63.980778	15.790699	33.031576	94.929980
1		67.323908	15.960043	36.042798	98.605018
2		76.841578	15.975306	45.530553	108.152604
3		76.993560	16.038050	45.559559	108.427561
4		73.616041	16.038632	42.180900	105.051182

- The RMSE value for ARIMA (3,1,2) is 27.992337075309443

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.

- SPARKLING

	Test RMSE Sparkling
Alpha=0.10,Beta=0.00,Gamma=0.371,TripleExponentialSmoothing	384.197750
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.786198
SARIMA(1,1,2)(1,0,2,12)	528.602513
SARIMA(3,1,2)(2,0,1,12)	611.038037
SARIMA(1,1,2)(2,0,2,6)	626.892244
2pointTrailingMovingAverage	813.400684
SARIMA(3,1,2)(1,0,3,6)	902.790726
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
Alpha=0.099,SimpleExponentialSmoothing	1275.081813
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
ARIMA(2,1,2)	1374.696495
ARIMA(3,1,2)	1379.182676
RegressionOnTime	1389.135175
Alpha=0.3,SimpleExponentialSmoothing	2603.403036
NaiveModel	3864.279352
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704

- ROSE

	Test RMSE Rose
Alpha=0.3,Beta=0.4,Gamma=0.3,TripleExponentialSmoothing	10.945435
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
RegressionOnTime	15.268955
ARIMA(3,1,2)	15.522887
ARIMA(0,1,2)	15.618093
Alpha=0.10,Beta=0.048,Gamma=0.0,TripleExponentialSmoothing	17.369488
SARIMA(1,1,2)(2,0,2,6)	26.136429
SARIMA(0,1,2)(2,0,2,12)	26.928362
SARIMA(3,1,2)(2,0,1,12)	27.992337
SARIMA(3,1,2)(3,0,3,6)	29.934404
Alpha=0.099,SimpleExponentialSmoothing	36.796246
Alpha=0.3,SimpleExponentialSmoothing	47.504821
SimpleAverageModel	53.460570
NaiveModel	79.718773
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.567594

OBSERVATION

- For Sparkling Dataset, Triple Exponential Smoothing (Holt Winter's) Autofit Model gives us the lowest RMSE value, hence it will be used for building our Optimum Model.

- For Rose Dataset, Triple Exponential Smoothing (Holt Winter's) Bestfit Model gives us the lowest RMSE value where alpha=0.3,beta=0.4 and gamma=0.3 hence it will be used for building our Optimum Model.

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.

- SPARKLING

Full Data model Summary

```
                       ExponentialSmoothing Model Results
==============================================================================
Dep. Variable:                      endog   No. Observations:                  187
Model:             ExponentialSmoothing   SSE                       22523264.581
Optimized:                         True   AIC                           2219.704
Trend:                         Additive   BIC                           2271.402
Seasonal:                Multiplicative   AICC                          2223.775
Seasonal Periods:                    12   Date:                 Thu, 12 Aug 2021
Box-Cox:                          False   Time:                         14:12:53
Box-Cox Coeff.:                    None
==============================================================================
                            coeff                 code              optimized
------------------------------------------------------------------------------
smoothing_level            0.0681476                alpha                 True
smoothing_slope            0.0681476                 beta                 True
smoothing_seasonal         0.2568278                gamma                 True
initial_level              1580.0005                  l.0                 True
initial_slope               0.000000                  b.0                 True
initial_seasons.0          1.0476792                  s.0                 True
initial_seasons.1          1.0068122                  s.1                 True
initial_seasons.2          1.2620726                  s.2                 True
initial_seasons.3          1.1703427                  s.3                 True
initial_seasons.4          0.9727799                  s.4                 True
initial_seasons.5          0.9526434                  s.5                 True
initial_seasons.6          1.2620336                  s.6                 True
initial_seasons.7          1.5960288                  s.7                 True
initial_seasons.8          1.3169898                  s.8                 True
initial_seasons.9          1.7611764                  s.9                 True
initial_seasons.10         2.6440116                 s.10                 True
initial_seasons.11         3.4137902                 s.11                 True
------------------------------------------------------------------------------
/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/holtwinters.py:712:
ConvergenceWarning: Optimization failed to converge. Check mle_retvals.
  ConvergenceWarning)
```

FORECASTING AND PREDICTING 12 Months into the future.

Top 5 rows of Predicted values

```
1995-08-31    1949.155797
1995-09-30    2344.802057
1995-10-31    3180.835470
1995-11-30    3937.976866
1995-12-31    5991.077374
Freq: M, dtype: float64
```

- RMSE of the Full Model is 347.0522513114097

Calculating the 95% confidence bands for better plotting.

	lower_CI	prediction	upper_ci
1995-08-31	1267.109074	1949.155797	2631.202521
1995-09-30	1662.755334	2344.802057	3026.848781
1995-10-31	2498.788747	3180.835470	3862.882193
1995-11-30	3255.930142	3937.976866	4620.023589
1995-12-31	5309.030651	5991.077374	6673.124097

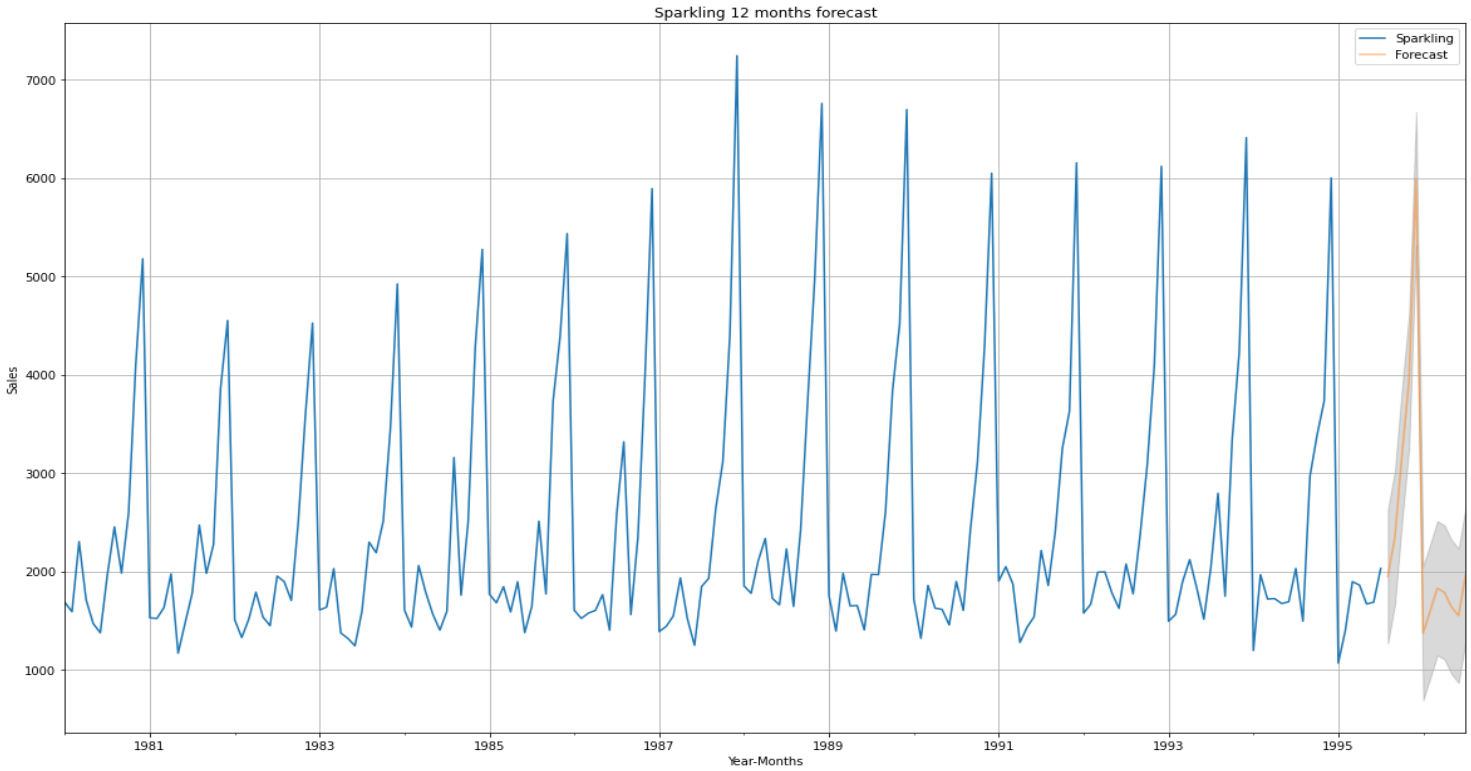


Figure 9(a)

- In the plot, the Orange region is our 12 months forecasted value and the shaded region surrounding are the confidence intervals which explains how much variation can be accepted for 95% accuracy of our forecasted values.

- ROSE

Full Data model Summary

```
                       ExponentialSmoothing Model Results
==============================================================================
Dep. Variable:                      endog   No. Observations:                  187
Model:             ExponentialSmoothing   SSE                         110117.712
Optimized:                         True   AIC                          1224.723
Trend:                         Additive   BIC                          1276.420
Seasonal:                Multiplicative   AICC                         1228.794
Seasonal Periods:                    12   Date:                 Thu, 12 Aug 2021
Box-Cox:                          False   Time:                         15:23:46
Box-Cox Coeff.:                    None
==============================================================================
                            coeff                 code              optimized
------------------------------------------------------------------------------
smoothing_level            0.3000000                alpha                 False
smoothing_slope            0.4000000                 beta                 False
smoothing_seasonal         0.3000000                gamma                 False
initial_level             64.000000                  l.0                  True
initial_slope              0.1527778                  b.0                  True
initial_seasons.0          1.7500000                  s.0                  True
initial_seasons.1          1.8437500                  s.1                  True
initial_seasons.2          2.0156250                  s.2                  True
initial_seasons.3          1.5468750                  s.3                  True
initial_seasons.4          1.8125000                  s.4                  True
initial_seasons.5          2.6250000                  s.5                  True
initial_seasons.6          1.8437500                  s.6                  True
initial_seasons.7          2.0156250                  s.7                  True
initial_seasons.8          3.2031250                  s.8                  True
initial_seasons.9          2.2968750                  s.9                  True
initial_seasons.10         2.3437500                 s.10                  True
initial_seasons.11         4.1718750                 s.11                  True
```

FORECASTING AND PREDICTING 12 Months into the future.

Top 5 rows of Predicted values

```
1995-08-31    40.074648
1995-09-30    39.245730
1995-10-31    41.312369
1995-11-30    48.067469
1995-12-31    68.735454
Freq: M, dtype: float64
```

- RMSE of the Full Model is 24.266535961104537

Calculating the 95% confidence bands for better plotting.

	lower_CI	prediction	upper_ci
1995-08-31	-7.559277	40.074648	87.708574
1995-09-30	-8.388196	39.245730	86.879656
1995-10-31	-6.321556	41.312369	88.946295
1995-11-30	0.433543	48.067469	95.701394
1995-12-31	21.101528	68.735454	116.369379

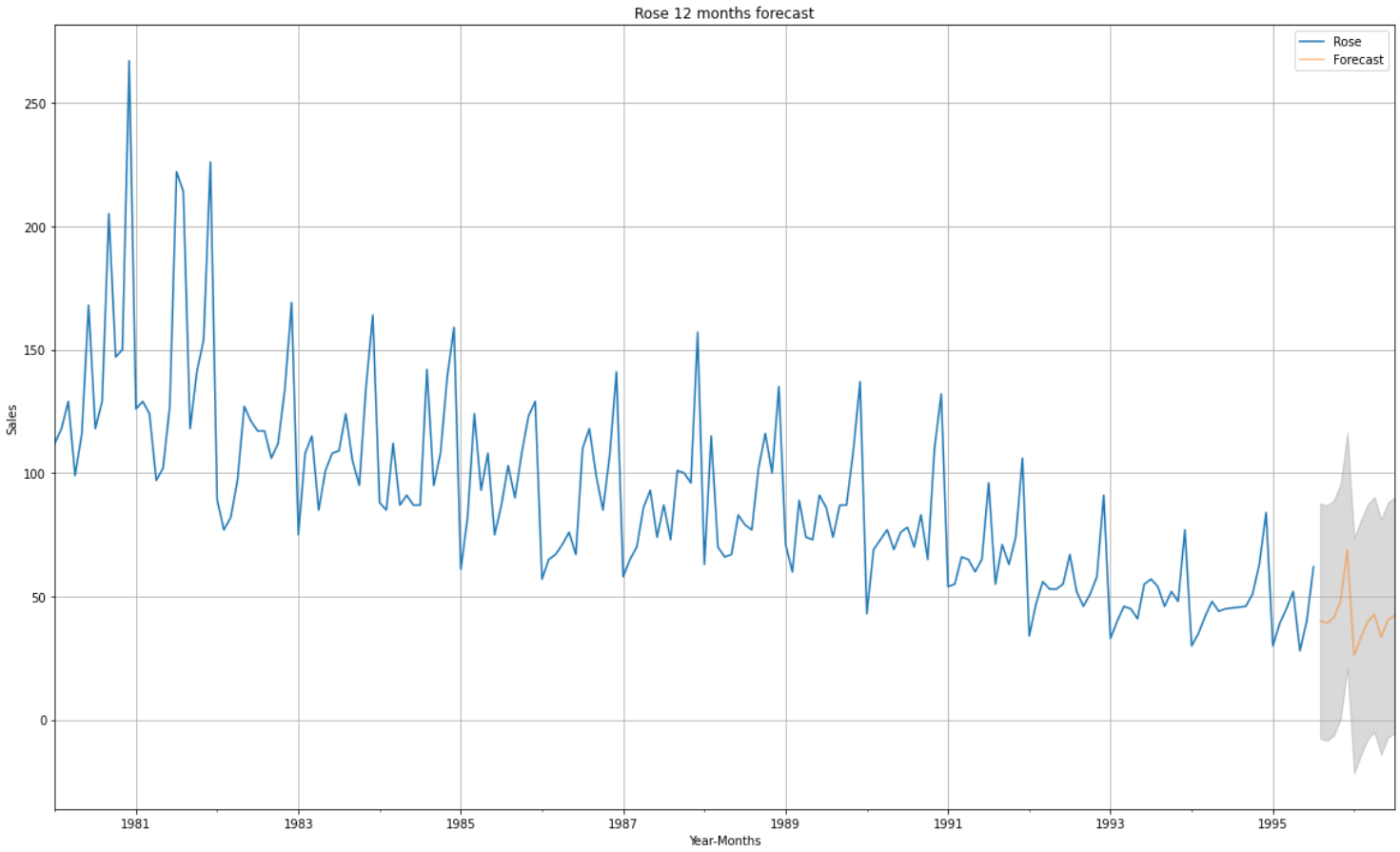


Figure 9(b)

- In the plot, the Orange region is our 12 months forecasted value and the shaded region surrounding are the confidence intervals which explains how much variation can be accepted for 95% accuracy of our forecasted values.

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

- As an analyst in the ABC Estate Wines, I was tasked to analyse and forecast Wine Sales in the 20th century of the individual datasets provided for Sparkling and Rose wine sales.
- For Sparkling when we analyze the time series, there was an exponential monthly increase in the wine sales however there was little to none yearly trend present in our data which means the year round sales haven't improved as compared to previous years.
- For Rose when we analyze the time series, there was an exponential monthly increase in the wine sales however there was a downward yearly trend present in our data which means the year round sales have dropped with each year.

- Both the datasets have trend and seasonality present in them explaining that **our data is not stationary** which is the basis of our **Null hypothesis**.
- This trend and seasonality is fixed by differencing our time series once. This makes it safe to conclude that our **time series is now stationary** which is the basis of our **Alternate Hypothesis**.
- This can also be observed from observing the p-value achieved with the value of alpha=0.05. If our test **p-value < alpha(0.05)** our **Null Hypothesis is rejected** stating that time series is **stationary**.
- If our test **p-value > alpha(0.05)** our **Null Hypothesis is accepted** stating that time series is **not stationary**.
- Also, observing the residual pattern from the decomposition it can be observed that multiplicative decomposition is the best way to go.
- The ACF and PACF gives us the order for Moving Average and Auto Regressive model values which explains the significance of variables for our model.
- RMSE value is the value which explains the error present it the predictions of our models and obviously the best model would be one with the lowest RMSE value.
- After building numerous models for both datasets it can be observed that Triple Exponential Smoothing (Holt Winter's) Model is the best/optimum performing model for both datasets giving us the best predictions.
- The only Difference being the Autofit TES model gives us the lowest RMSE for Sparkling dataset.
- However, for Rose Dataset we will manually fit the alpha, beta and gamma values for the TES model.

Measures that the company should be taking for future sales are:

- More data should be captured on the wine such as consumer age, to better understand what kind of audience to target.
- Brand recognition is important more emphasis for Rose as they have a downward trend, this can be done by smart marketing of their product by increasing wine tasting initiatives as it is the easiest and most efficient way to achieve recognition.
- Decreasing the alcohol content and increasing the organic products used in wine as the society has become a little sceptical and moving towards a healthy lifestyle.
- Fully Utilizing the modern techniques is something that should be used to its full potential for better and fast production.
- Once the wine has gained recognition it'll be a huge advantage as import/export will bring in more sales for the product.