Time Series Forecasting Project

Name: Swetha Kunapuli

Batch & Course: PGP-DSBA

Online June Batch

Date: 11/2/2022

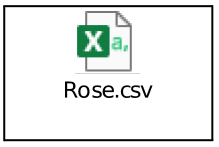
Table of Contents

PROBLEM

Executive Summary:

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

Data set for the Problem:



Data Introduction:

The purpose of this whole exercise is to explore the dataset and is recommended for learning and practicing our skills using Time series forecasting analysis.

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

Sample of the Sparkling dataset:

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

Table: 1

Sample of the Rose dataset:

	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

Table:2

Creating the Time Stamps with monthly frequency and adding to the data frame to make it a Time series data.

Sparkling

```
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')
```

Fig:1

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

Table:3

Rose

```
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')
```

Fig:2

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

Table:4

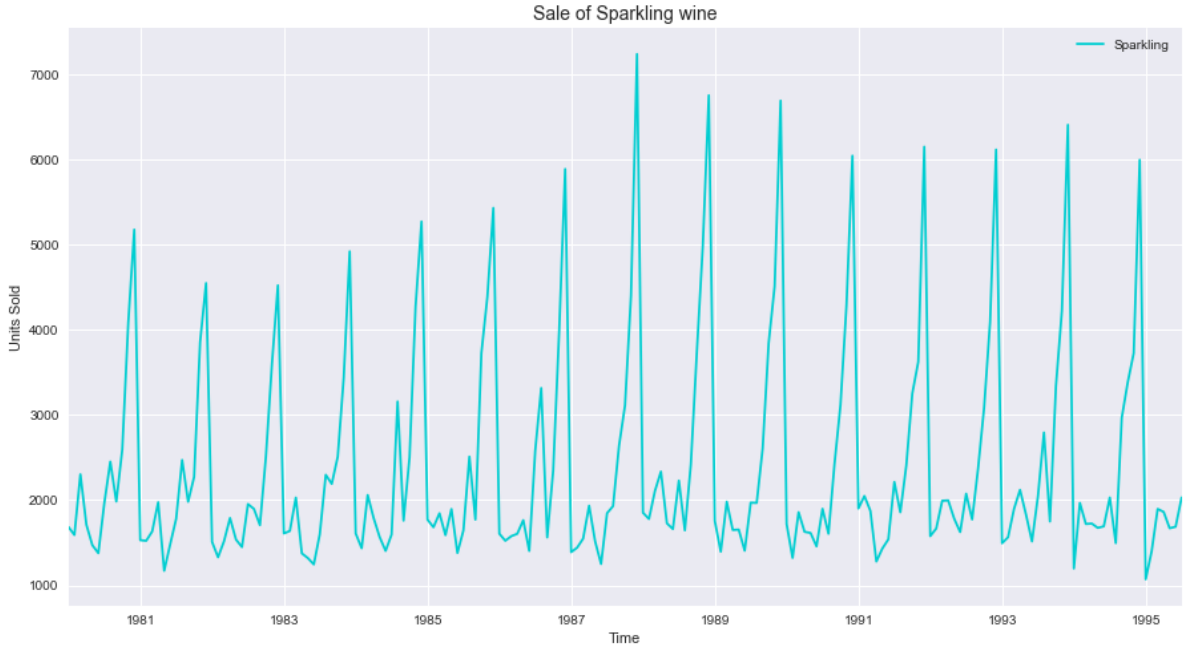


Fig:3

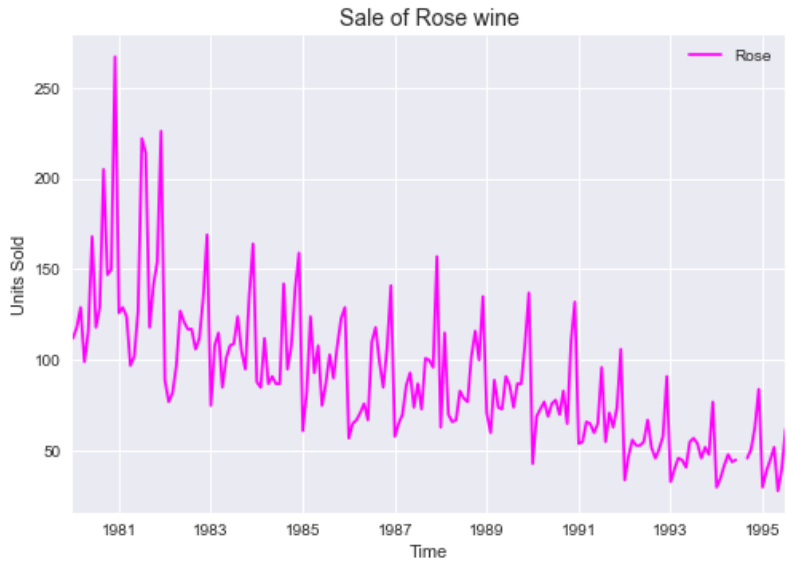


Fig:4

Observations:

- Monthly sales of two types of wines, such as Sparkling and Rose are given, for a period from January 1980 to July 1995.

- The given data files are read as-is and a date-range has been applied to the data as an index.

- Both the datasets show significant seasonality. While the sale of Rose shows an evident downward trend, Sparkling doesn't show any consistent trend but has upward and downward slopes during the time period.

- While Sparkling wine has been consistently favoured over the years by customers, the demand for Rose has fallen out-of-favour over the years.

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

Sparkling

Check for types of variables in the data frame:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Fig:5

Check for statistical description:

	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

Table:5

Check for missing values in the dataset

```
Sparkling    0
dtype: int64
```

Fig:6

ECDF Plot

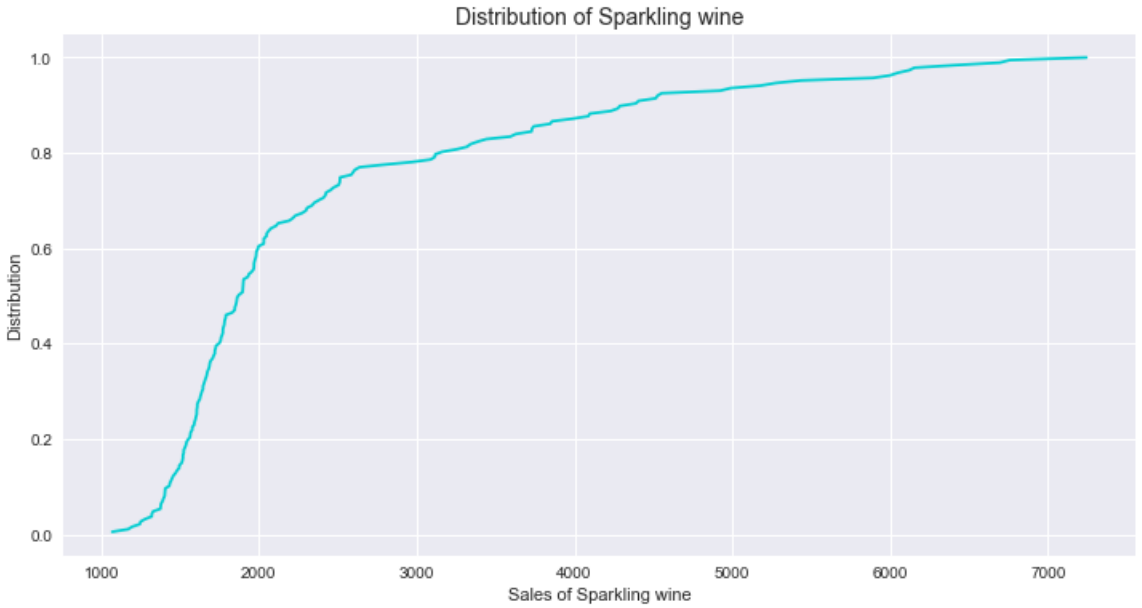


Fig:7

Yearly Box Plot

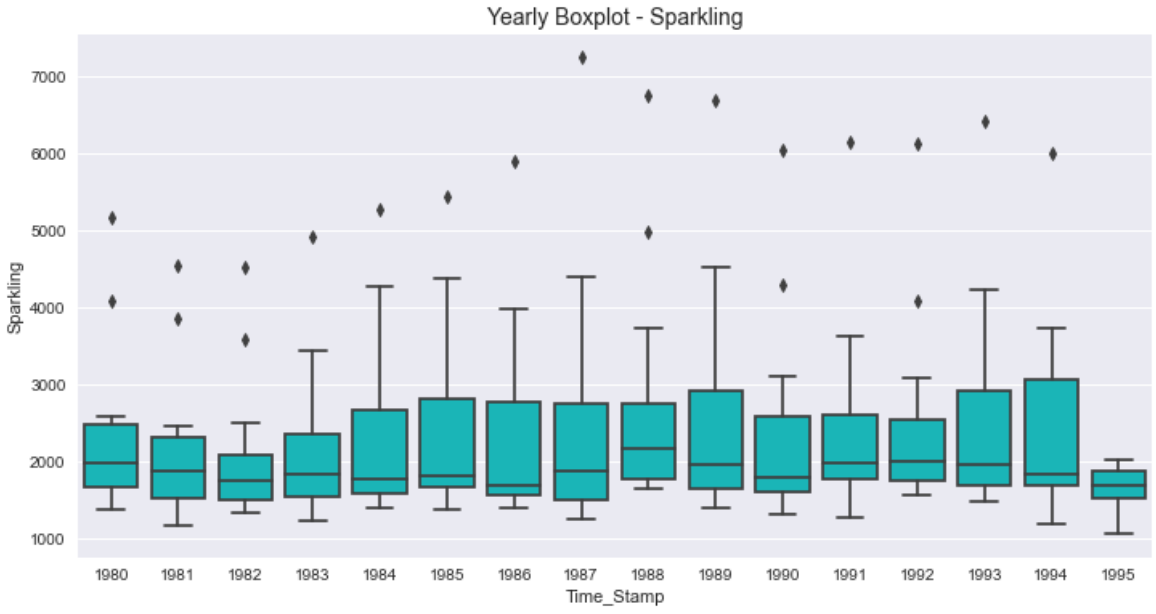


Fig:8

Monthly Box Plot

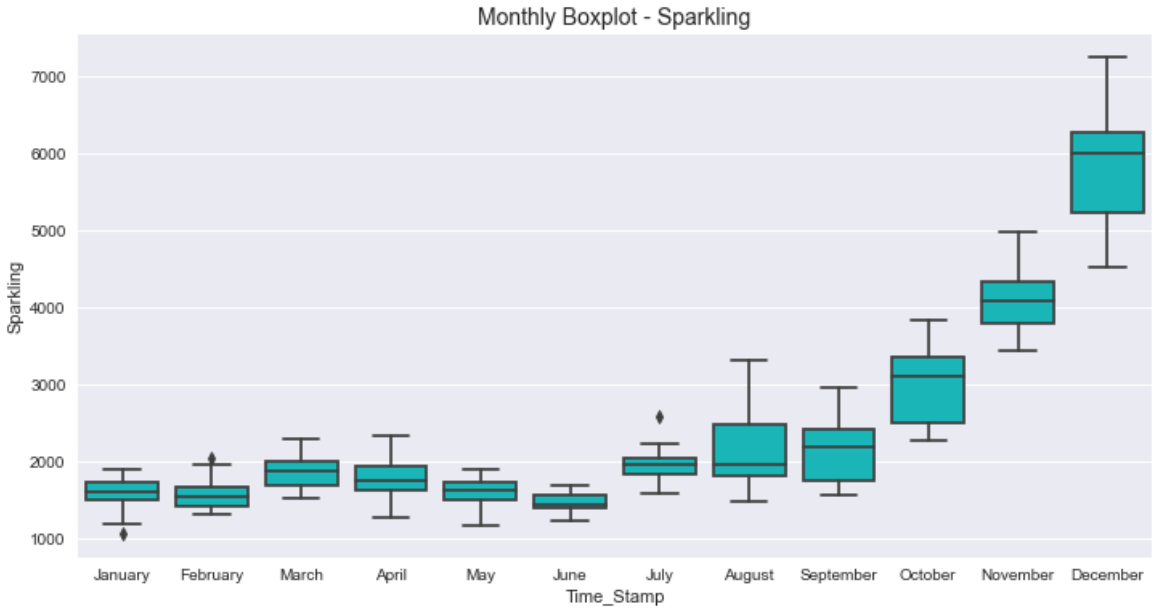


Fig:9

Observations:

- The descriptive summary of the data shows that on an average 2402 units of Sparkling wines were sold each month in the given period of time. 50% of monthly sales varied from 1605 units to 2549 units. Maximum sale reported in a month is 7242 units.

- The Empirical CDF plot shows that, in 80% of months, at least 3000 units of Sparkling wine were sold

- There are no missing values in the sparkling dataset.

- The yearly box plot shows that the average sale of Sparkling has been more or less consistent across the period, at or a little below 2000 units.

- The outliers in the yearly boxplot most probably represent the seasonal sale during the seasonal months

- The monthly box plot shows a clear seasonality during the festive seasonal months of October, November, and December, which peaks in December. The sale tanks in the month of June.

Plot a time series month plot to understand the spread of sales across different years and within different months across years.

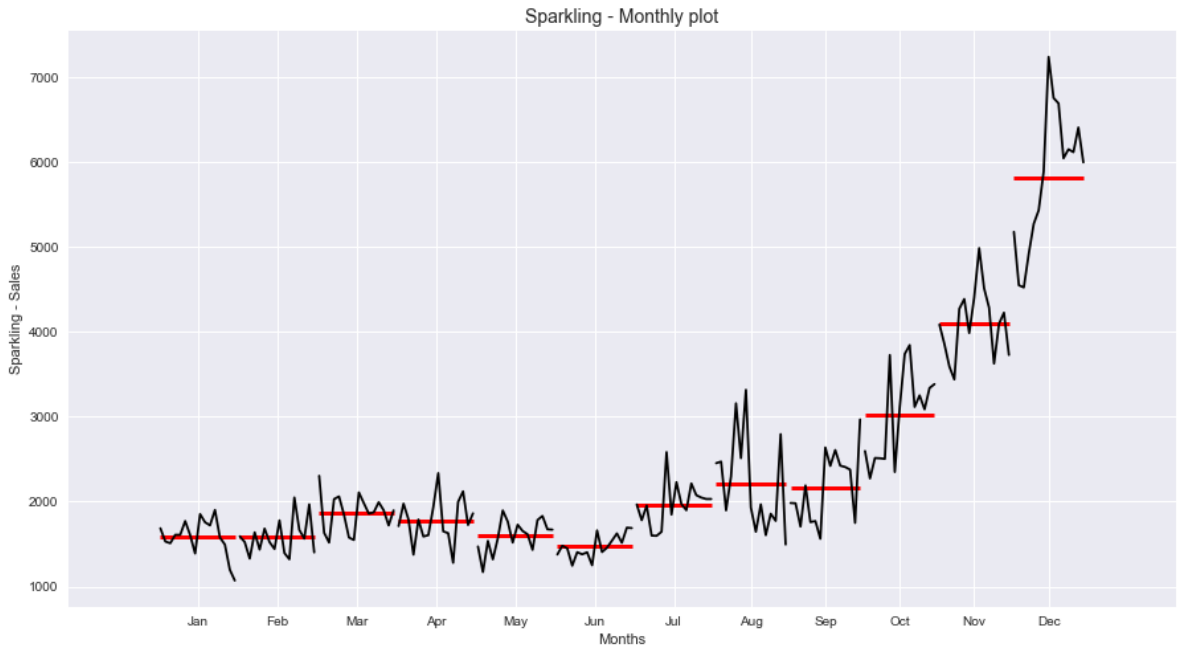


Fig:10

Observations:

This plot shows us the behaviour of the Time Series across various months. The red line is the median value.

- The monthly plot for Sparkling shows the mean and variation of units sold each month over the years. Sales in seasonal months show a higher variation than in the lean months.

- Sales in December with a mean few points below 6000, varied from 7400 to 4500 units over the years. Whereas sales in November vary from 3500 units to 5000 units and sales in October vary from 2500 to 4000 units.

- The lean months from January till September shows more or less a consistent sale around 2000 units

Monthly sales across years

Time_Stamp	1	2	3	4	5	6	7	8	9	10	11	1
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.
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.

Table:6

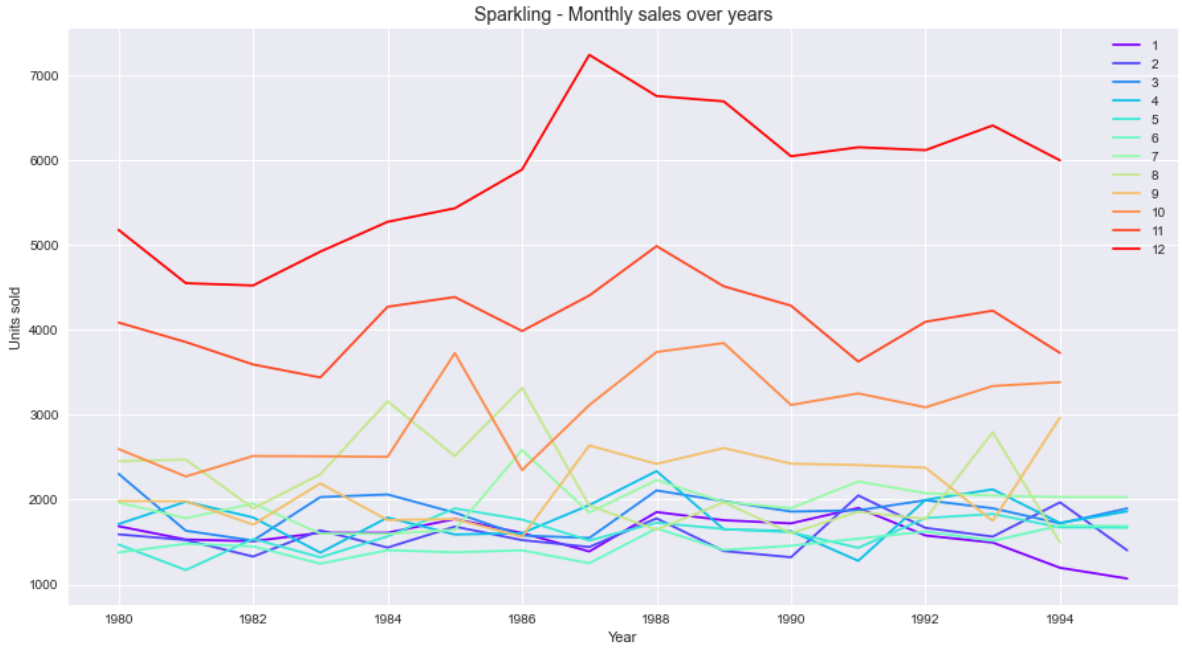


Fig:11

Observations:

- The plot of monthly sales over the years also shows the seasonality component of the time-series, with October, November and December selling exponentially higher volumes.

- The highest volume of Sparkling wines was sold in December,1987 and the least of December sales was in 1981. Post 1987 December sales are around an average 6500 units, which was around 5000 in the early 80's.

- The seasonal sale since 1990 has been more or less consistent around 6000 units in December, 4000 units in November and 3000 units in October.

- Sales for the months from January to July are seen to be consistent across the years, compared to the rest of the months.

Rose

Check for types of variables in the data frame

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Fig:12

Check for statistical description:

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

Table:7

Check for missing values in the dataset

```
Rose    2
dtype: int64
```

Fig:13

Observations:

- The Rose time-series had values missing for two months in 1994, which were imputed using interpolation (linear method).

- Rose data after interpolation for the year 1994 is given below as well as the plot.

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

Table:8

```
Time_Stamp
1994-01-31    30.000000
1994-02-28    35.000000
1994-03-31    42.000000
1994-04-30    48.000000
1994-05-31    44.000000
1994-06-30    45.000000
1994-07-31    45.336957
1994-08-31    45.673913
1994-09-30    46.000000
1994-10-31    51.000000
1994-11-30    63.000000
1994-12-31    84.000000
Name: Rose, dtype: float64
```

Table:9

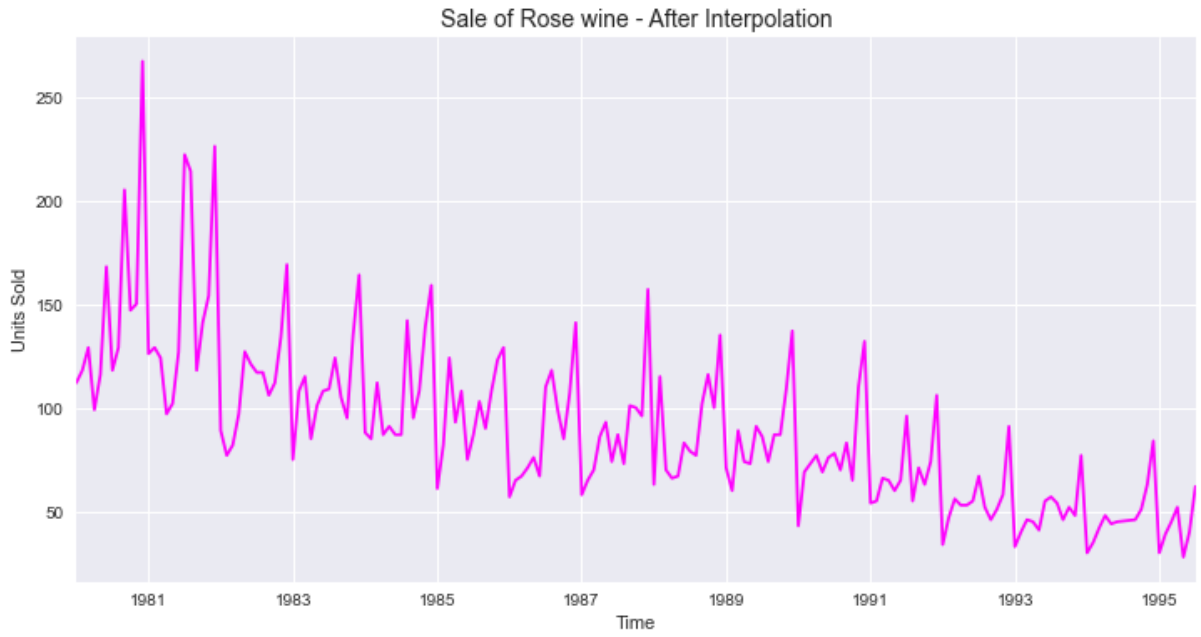


Fig:14

ECDF Plot

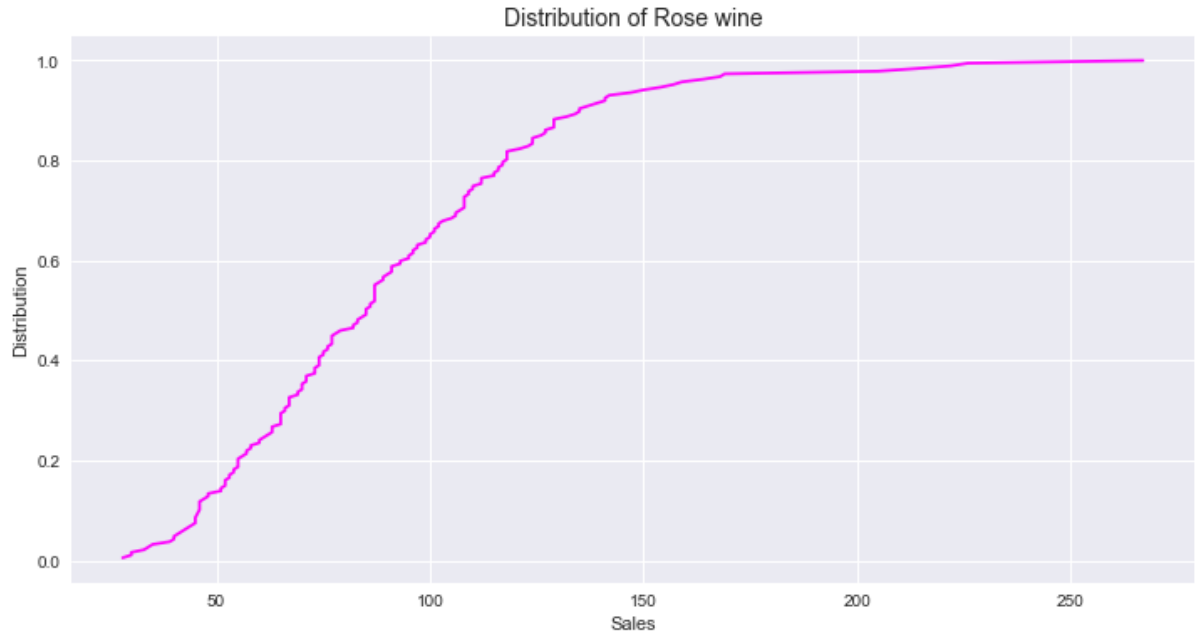


Fig:15

Yearly Box Plot

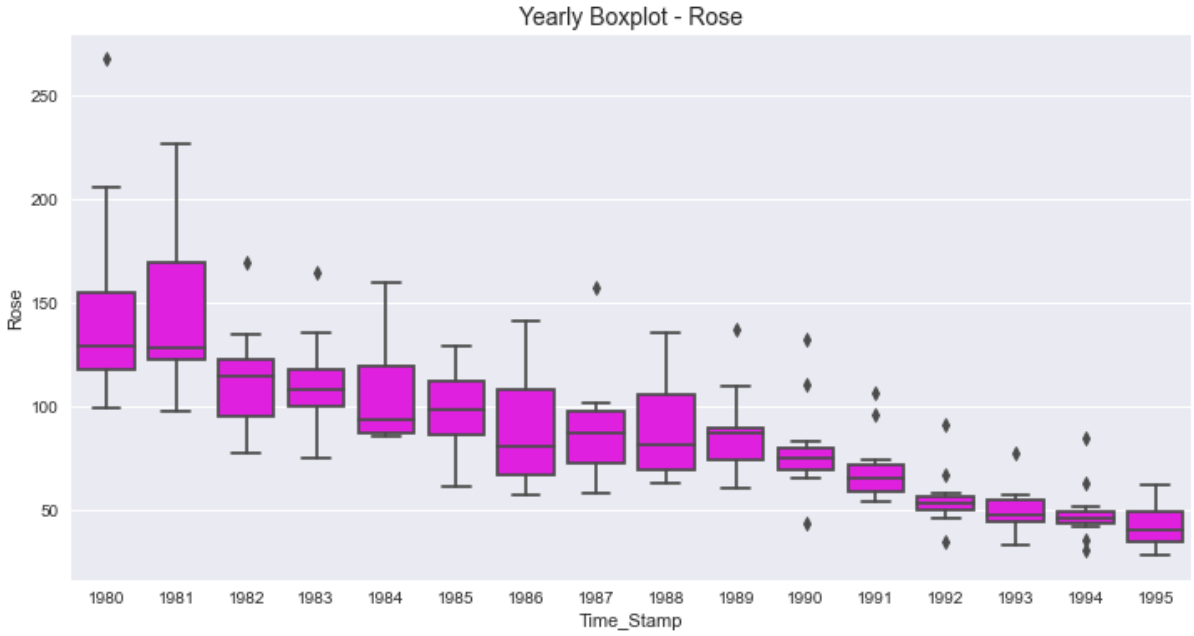


Fig:16

Monthly Box Plot

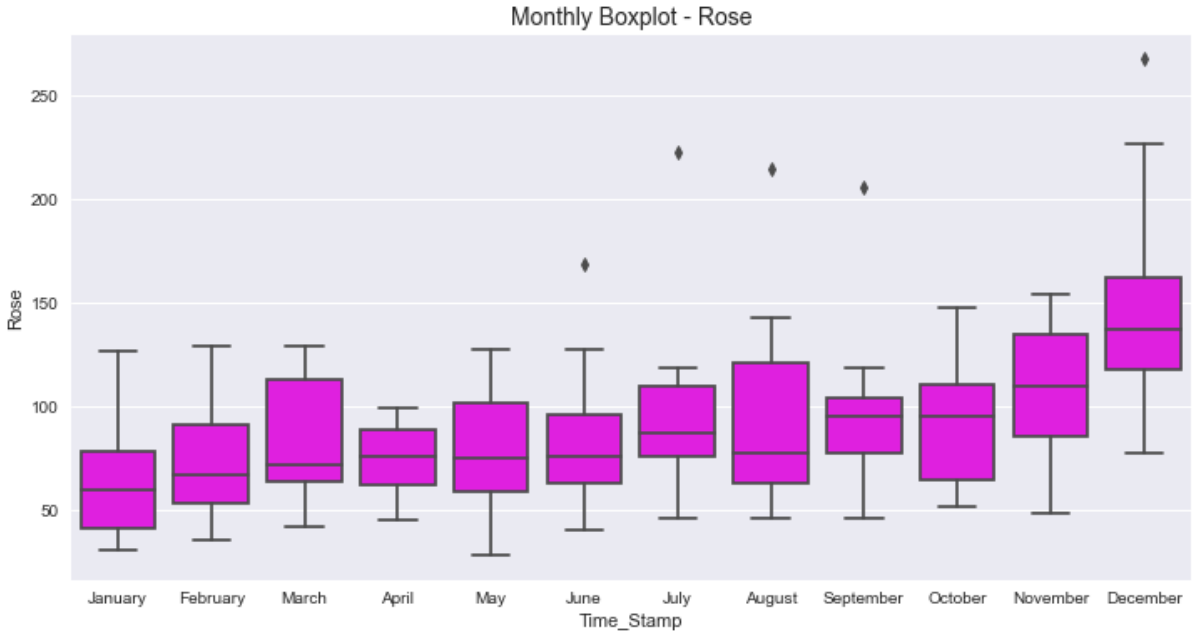


Fig:17

Observations:

- The descriptive summary of the data shows that on an average 90 units of Rose wines were sold each month in the given period of time. 50% of monthly sales varied from 63 units to 112 units. Maximum sales reported in a month is 267 units and a minimum of 28 units.

- The Empirical CDF plot shows that, in 80% of months, at least 120 units of Rose wine were sold.

- The yearly box plot shows the average sale of Rose wine moving according to the downward trend in sales over the years. The outliers over the upper bound in the yearly boxplot most probably represent the seasonal sale during the seasonal months.

- The monthly box plot shows a clear seasonality during the seasonal months of November and December. Though the sale tanks in the month of January, it picks up in the due course of the year.

- Average sale in December is around 140 units, November is around 110 units and October is around 90 units.

Plot a time series month plot to understand the spread of sales across different years and within different months across years.

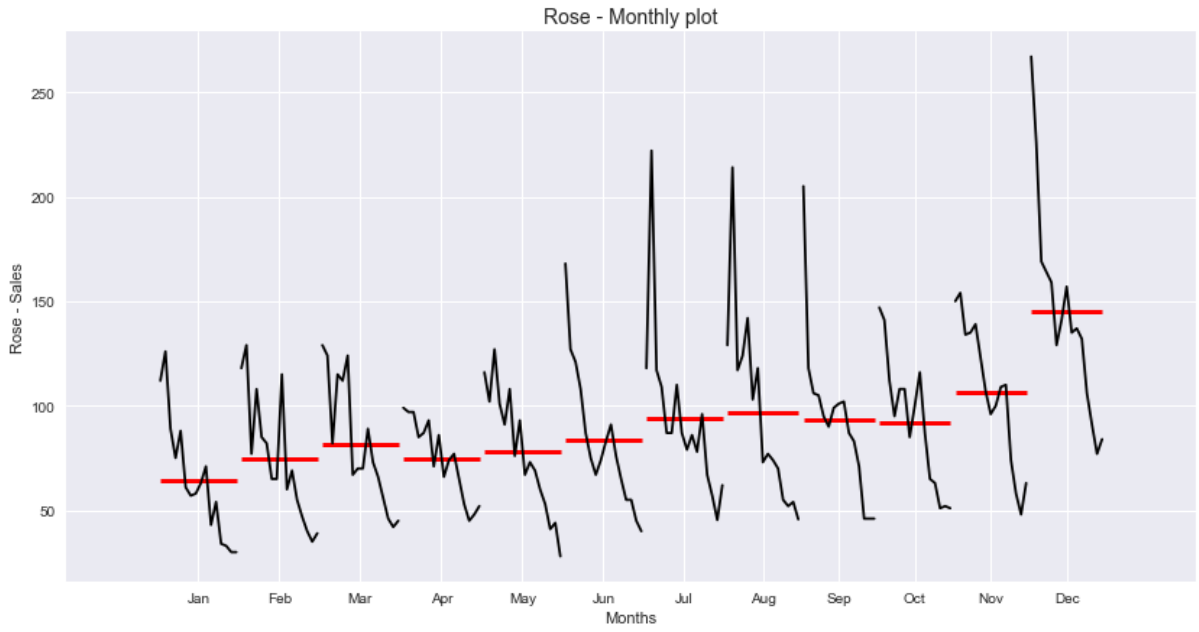


Fig:18

Observations:

This plot shows us the behaviour of the Time Series across various months. The red line is the median value.

- The monthly plot for Rose shows the mean and variation of units sold each month over the years. Sales in months such as July, August, September and December show a higher variation than the rest.
- Sales in December with a mean few points below 100, varied from 75 to 270 units over the years. Whereas the average sale is less than or closer to 100 units (above 50) for the rest of the year.

Monthly 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

Table:10

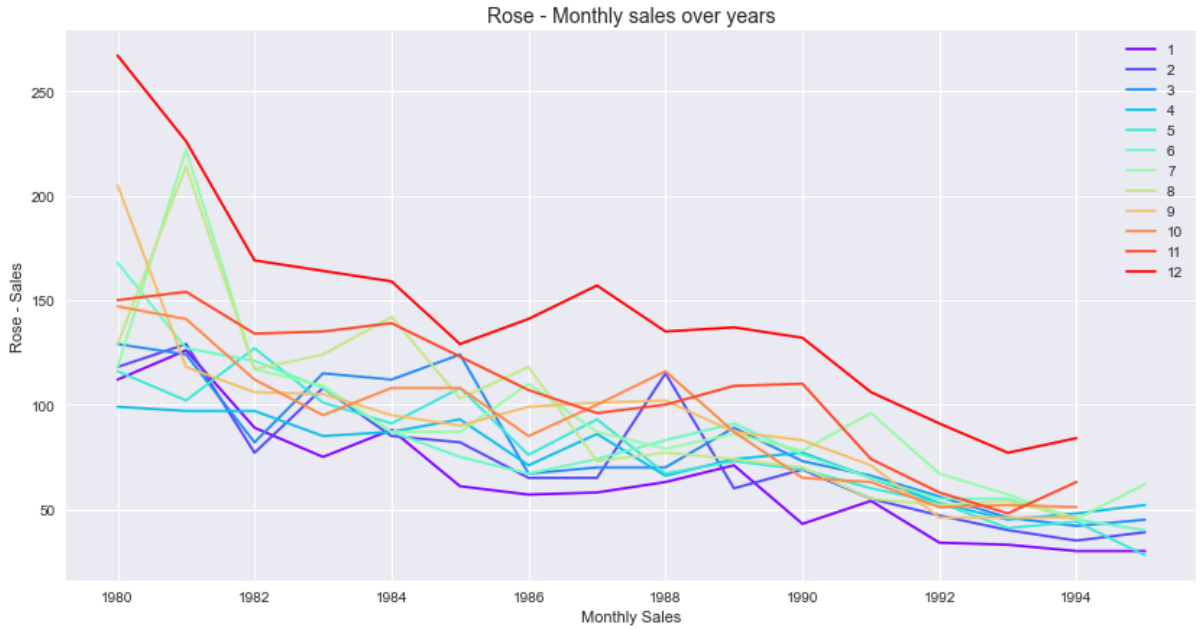


Fig:19

Observations:

- The plot of monthly sales over the years also shows the seasonality component of the time-series, with November and December selling exponentially higher volumes than other months.

- The highest volume of Rose wines was sold in December,1980 and the lowest December sale was in 1993. Though December sales picked after 1983, it consistently dipped after 1987.

TIME SERIES DECOMPOSITION

If the seasonality and residual components are independent of the trend, then you have an additive series. If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then you have a multiplicative series.

Sparkling

Additive Decomposition

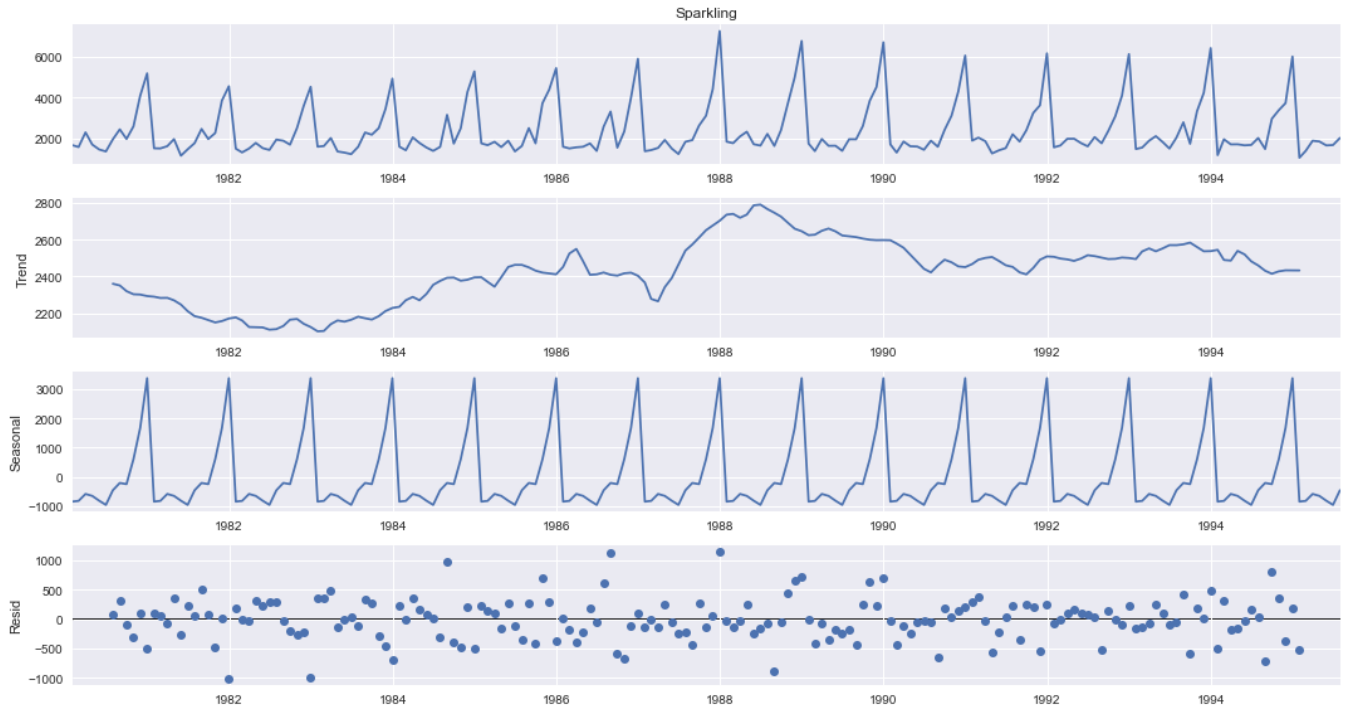


Fig:20

Additive Decomposition tables of Trend, Seasonality, Residual.

```
Trend
 Time_Stamp
1980-01-31            NaN
1980-02-29            NaN
1980-03-31            NaN
1980-04-30            NaN
1980-05-31            NaN
1980-06-30            NaN
1980-07-31    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: trend, dtype: float64
```

Table:11

```
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: seasonal, dtype: float64
```

Table:12

```
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: resid, dtype: float64
```

Table:13

Deseasonalized Time Stamp for Additive Decomposition

```
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    2431.502265
1980-08-31    2667.332821
1980-09-30    2238.677265
1980-10-31    1996.230043
1980-11-30    2411.932821
1980-12-31    1792.016154
dtype: float64
```

Table:14

Deseasonalized Time Series Plot of Additive Decomposition

Fig:21

Multiplicative Decomposition

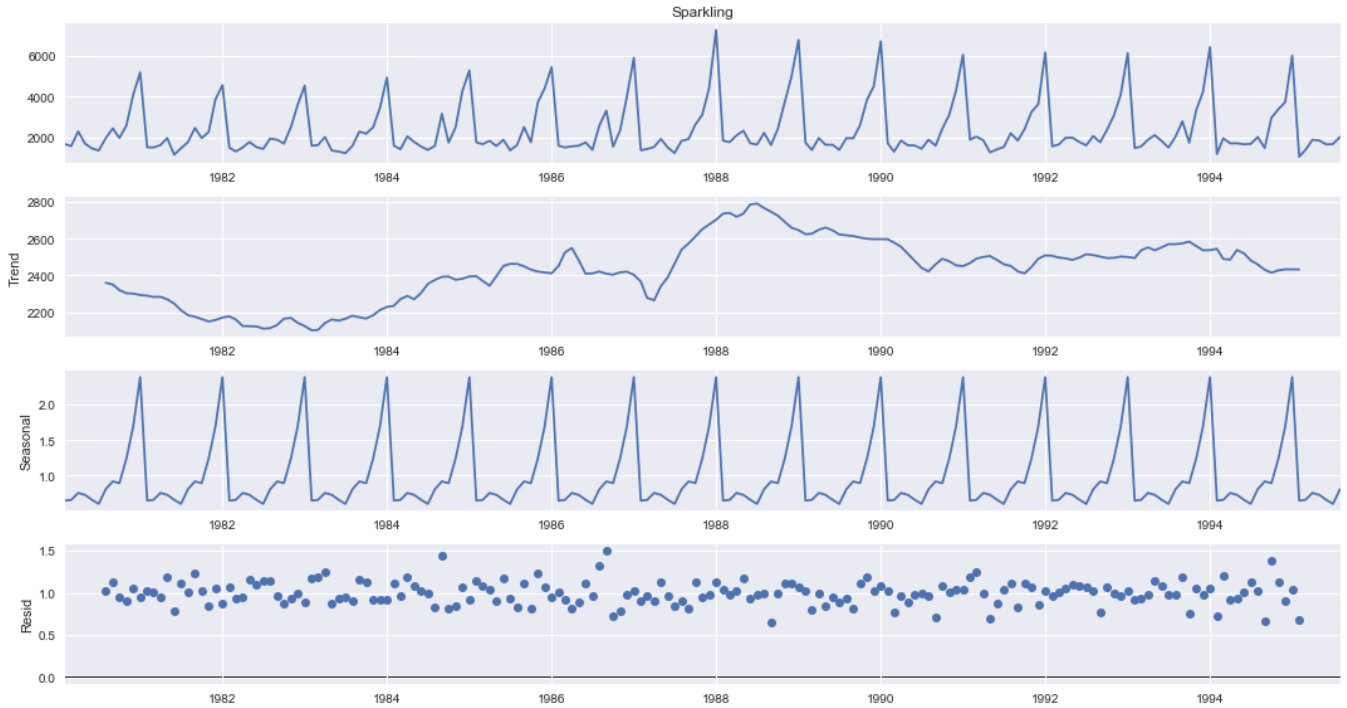


Fig:22

Multiplicative Decomposition tables of Trend, Seasonality, Residual

```
Trend
 Time_Stamp
1980-01-31            NaN
1980-02-29            NaN
1980-03-31            NaN
1980-04-30            NaN
1980-05-31            NaN
1980-06-30            NaN
1980-07-31    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: trend, dtype: float64
```

Table:15

```
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: seasonal, dtype: float64
```

Table:16

```
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: resid, dtype: float64
```

Table:17

Deseasonalized Time Stamp for Multiplicative Decomposition

```
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    2361.695896
1980-08-31    2352.468741
1980-09-30    2321.497620
1980-10-31    2304.490847
1980-11-30    2303.092089
1980-12-31    2294.738436
dtype: float64
```

Table:18

Deseasonalized Time Series Plot for Multiplicative Decomposition

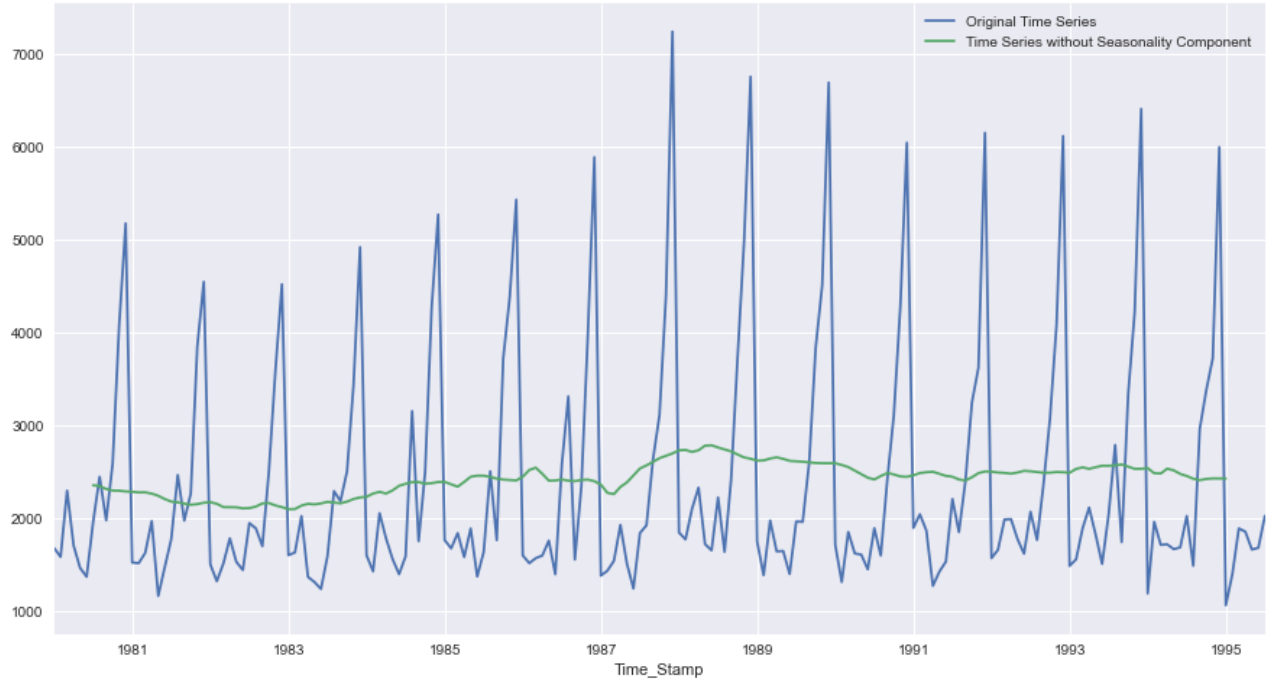


Fig:23

Observations:

• The decomposition plots of Sparkling wine sales are given above.

• As the altitude of the seasonal peaks in the observed plot is changing according to the change in trend, the time-series is assumed to be 'multiplicative'.

• The plot of the trend component does not show a consistent trend, but an intermediate period shows an upward slope which becomes consistent in the late half of the time-series.

• The additive model shows the seasonality with a variance of 3000 units and the multiplicative model shows a variance of 30%.

• The residual shows a pattern of high variability across the period of time-series, which is more or less consistent in both additive and multiplicative decompositions. For the multiplicative series, we see that a lot of residuals are located around 1.

• The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 10%.

• If the seasonality and residual components are independent of the trend, then you have an additive series. If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then you have a multiplicative series.

Rose

Additive Decomposition

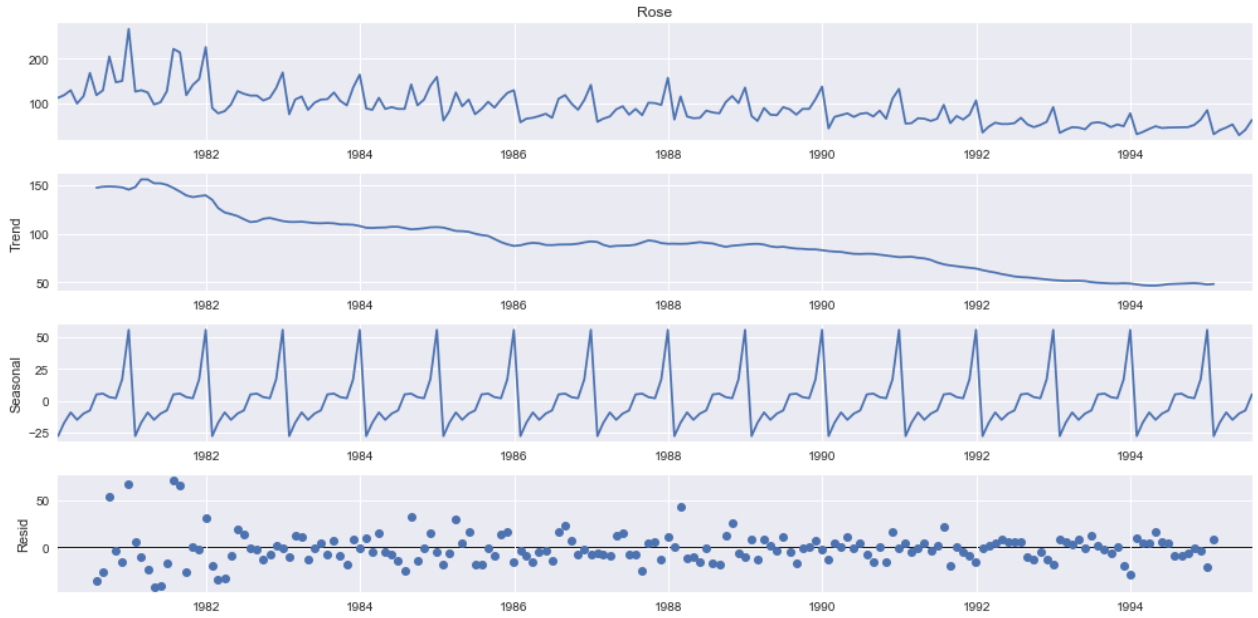


Fig:24

Additive Decomposition tables of Trend, Seasonality, Residual

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

Table:19

```
Seasonality
 Time_Stamp
1980-01-31   -27.908708
1980-02-29   -17.435675
1980-03-31    -9.285895
1980-04-30   -15.098395
1980-05-31   -10.196609
1980-06-30    -7.678752
1980-07-31     4.897089
1980-08-31     5.500109
1980-09-30     2.774625
1980-10-31     1.871848
1980-11-30    16.846848
1980-12-31    55.713514
Name: seasonal, dtype: float64
```

Table:20

```
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   -33.980423
1980-08-31   -24.625109
1980-09-30    53.850375
1980-10-31    -2.955181
1980-11-30   -14.263514
1980-12-31    66.161486
Name: resid, dtype: float64
```

Table:21

Deseasonalized Time Stamp for Additive Decomposition

```
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    113.102911
1980-08-31    123.499891
1980-09-30    202.225375
1980-10-31    145.128152
1980-11-30    133.153152
1980-12-31    211.286486
dtype: float64
```

Table:22

Deseasonalized Time Series Plot for Additive Decomposition

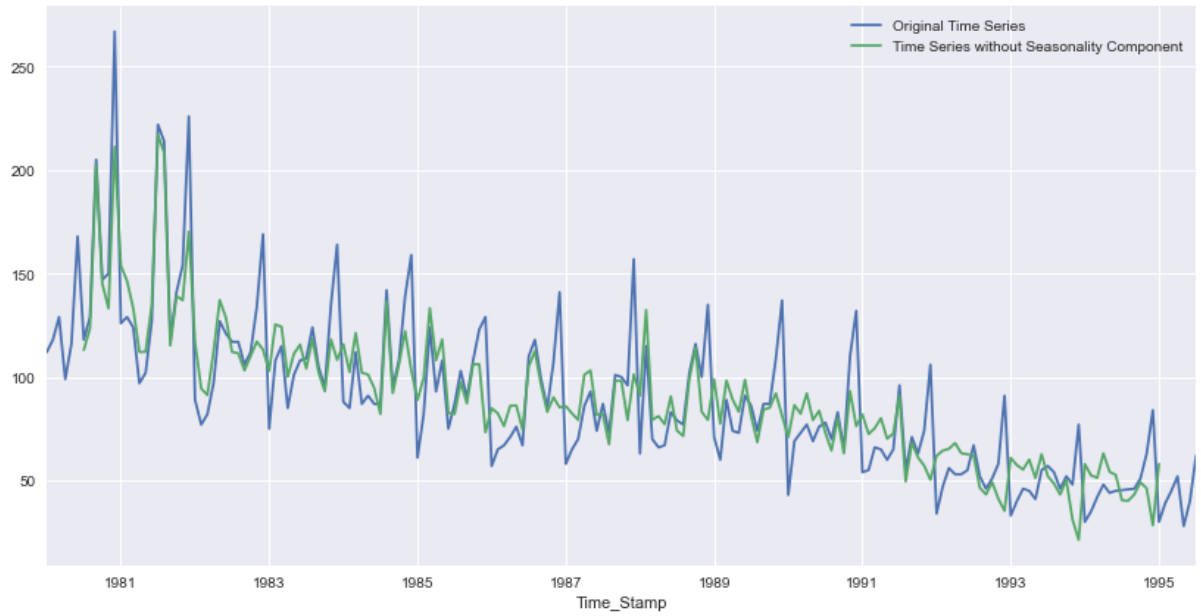


Fig:25

Multiplicative Decomposition

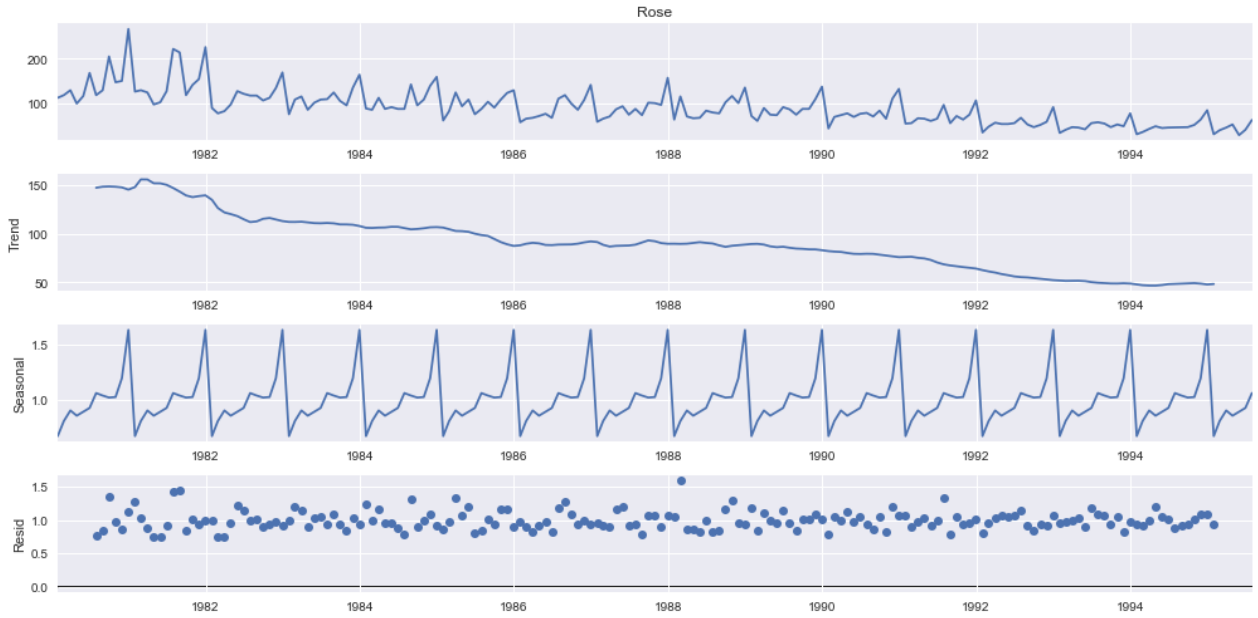


Fig:26

Multiplicative Decomposition tables of Trend, Seasonality, Residual

Time_Stamp	Trend
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	147.083333
1980-08-31	148.125000
1980-09-30	148.375000
1980-10-31	148.083333
1980-11-30	147.416667
1980-12-31	145.125000

Name: trend, dtype: float64

Table:23

```
Seasonality
 Time_Stamp
1980-01-31    0.670111
1980-02-29    0.806163
1980-03-31    0.901163
1980-04-30    0.854023
1980-05-31    0.889414
1980-06-30    0.923984
1980-07-31    1.058042
1980-08-31    1.035890
1980-09-30    1.017647
1980-10-31    1.022572
1980-11-30    1.192347
1980-12-31    1.628644
Name: seasonal, dtype: float64
```

Table:24

```
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    0.758256
1980-08-31    0.840713
1980-09-30    1.357675
1980-10-31    0.970772
1980-11-30    0.853379
1980-12-31    1.129647
Name: resid, dtype: float64
```

Table:25

Deseasonalized Time Stamp for Multiplicative Decomposition

```
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    147.841589
1980-08-31    148.965713
1980-09-30    149.732675
1980-10-31    149.054105
1980-11-30    148.270046
1980-12-31    146.254647
dtype: float64
```

Table:26

Deseasonalized Time Series Plot for Multiplicative Decomposition

Fig:27

Observations:

• The observed plot of the decomposition diagrams shows visible annual seasonality and a downward trend. The early period of the plot shows higher variation than in the later periods.

• The trend diagram shows a downward trend overall. Exponential dips can be seen between 1981 and 1983 and later from 1991 to 1993.

• Seasonal components are quite visible and consistent in both the observed and seasonal charts of the diagrams. The additive chart shows variance in seasonality from -20 to 50 units and the multiplicative model shows variance of 16%.

• The residuals show a pattern of high variability across the period of time-series, which is more or less consistent in both additive and multiplicative decompositions.

• The variance in residuals shows higher variance in the early period of the series, which explains the higher variance in the observed plot at the same time period.

• The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 15%.

• As the seasonality peaks are consistently reducing its altitude in line with trend, the series can be treated as multiplicative in model building.

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

- The train and test datasets are created with the year 1991 as the starting year for test data, using the index.year property of the time series index.

- The plots of the Sparkling and Rose time-series as train and test are given below.

Sparkling

```
train=df_spark[df_spark.index.year < 1991]
test=df_spark[df_spark.index.year >= 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

Table:27

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

Table:28

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

Table:29

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

Table:30

```
print(train.shape)
print(test.shape)
```

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

Fig:28

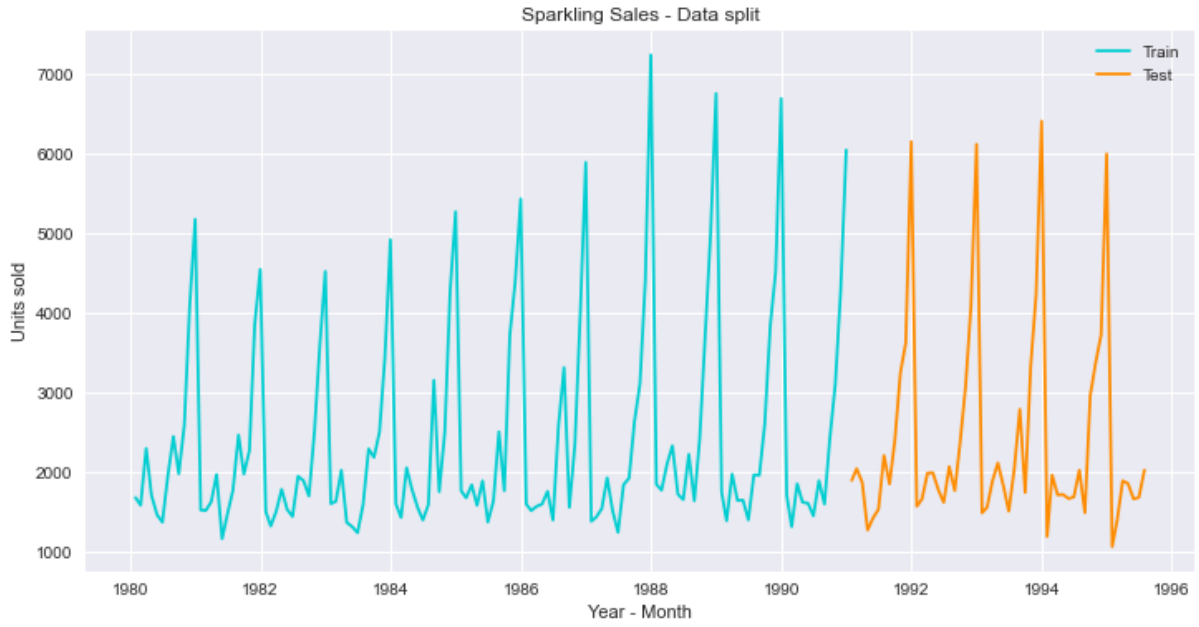


Fig:29

Rose

```
train=df_rose[df_rose.index.year < 1991]
test=df_rose[df_rose.index.year >= 1991]
```

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

	Rose
Time_Stamp	
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

Table:31

Last few rows of Training Dat

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

Table:32

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

Table:33

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

Table:34

```
print(train.shape)
print(test.shape)
```

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

Fig:30

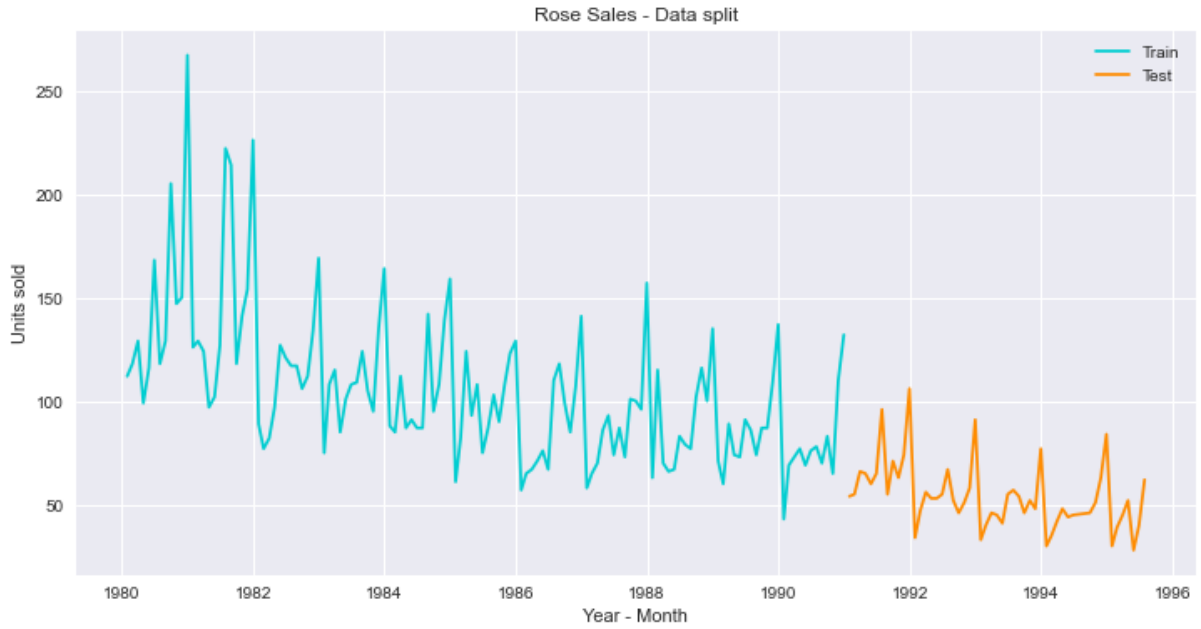


Fig:31

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

LINEAR REGRESSION

- To regress the sale of Sparkling and Rose wines, numerical time instance orders for both training and test were generated, and the values added to the respective dataset.

Sparkling

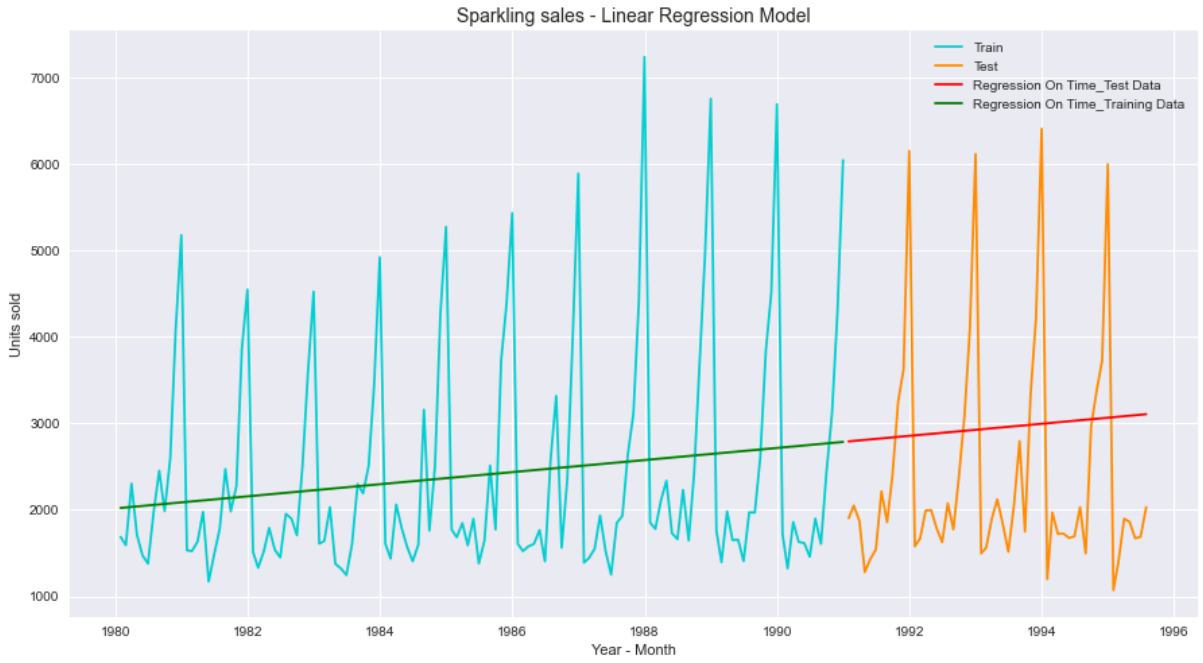


Fig:32

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15

Table:35

Observations:

- The linear regression plots show a gradual upward trend in the forecast of Sparkling wine, consistent with the observed trend which was not visually apparent.

- The RMSE and MAPE values for the Test data set above 50% of the forecast are erroneous.

Rose

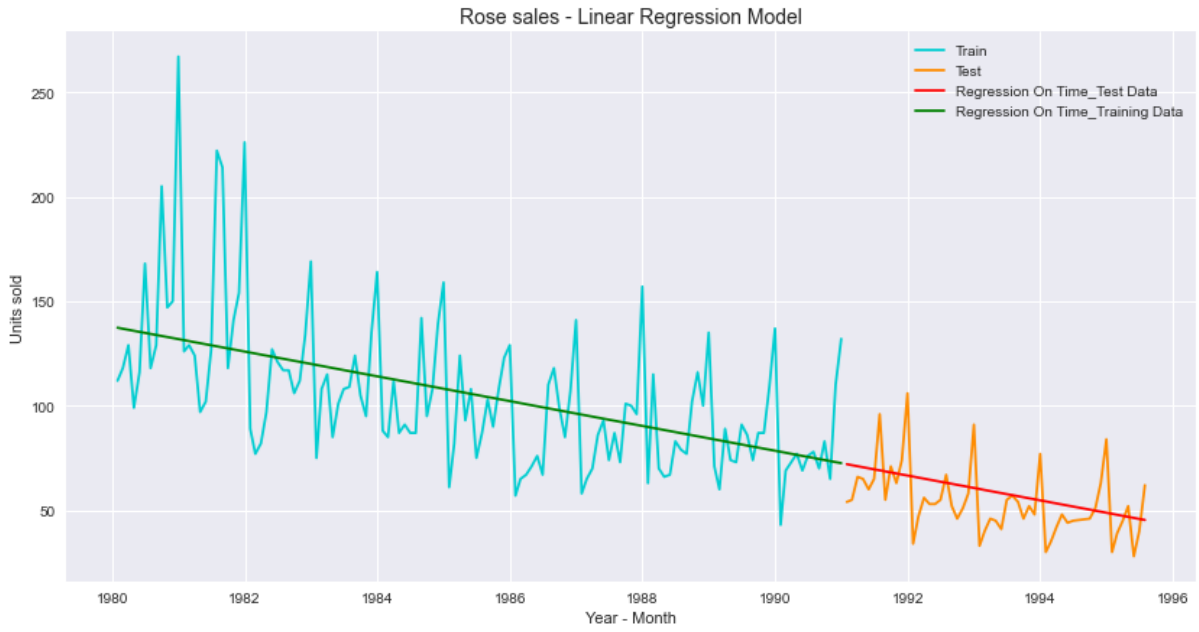


Fig:33

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82

Table:36

Observations:

- The linear regression on the Rose dataset shows an apparent downward trend consistent with the observed time-series.

- The RMSE and MAPE of the forecast is given above. The model leaves a 23% error in the forecast against the test set.

- The model has successfully captured the trend of both the series but does not reflect the seasonality.

NAÏVE FORECASTING

- In a naive model, 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 the same as today, therefore the prediction for day after tomorrow is also today.

Sparkling

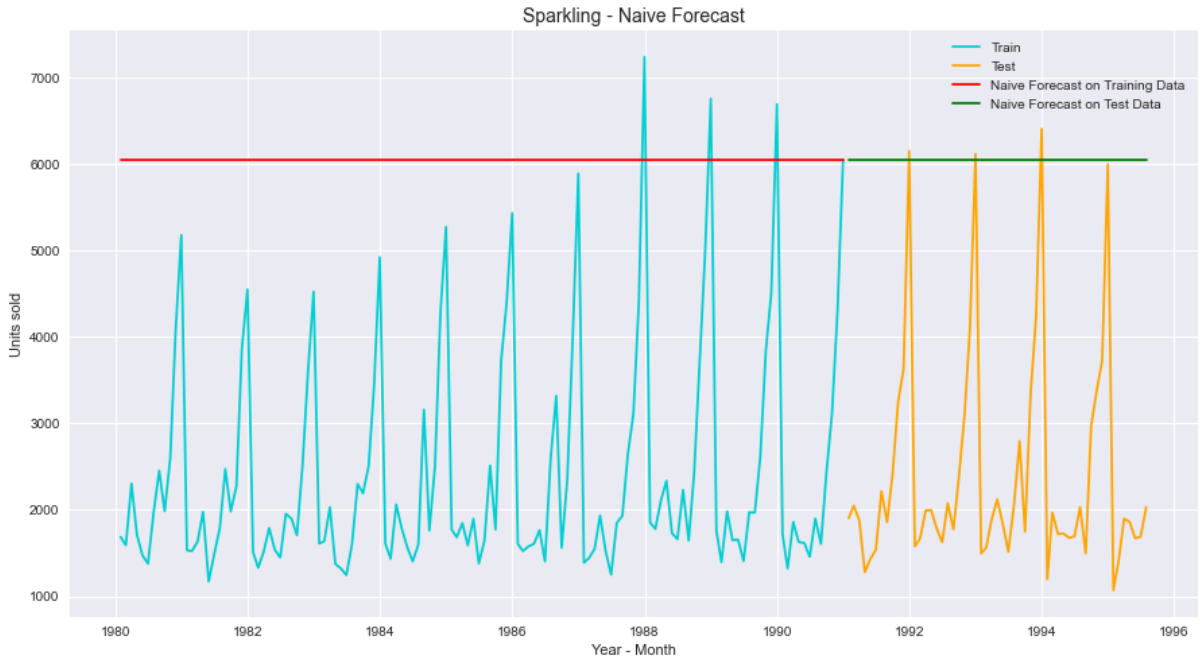


Fig:34

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87

Table:37

Observations:

- The model has taken the last value from the test set and fitted it on the rest of the train time period and used the same value to forecast the test set.

- The performance metrics above show a very poor fitment and high % of error.

Rose

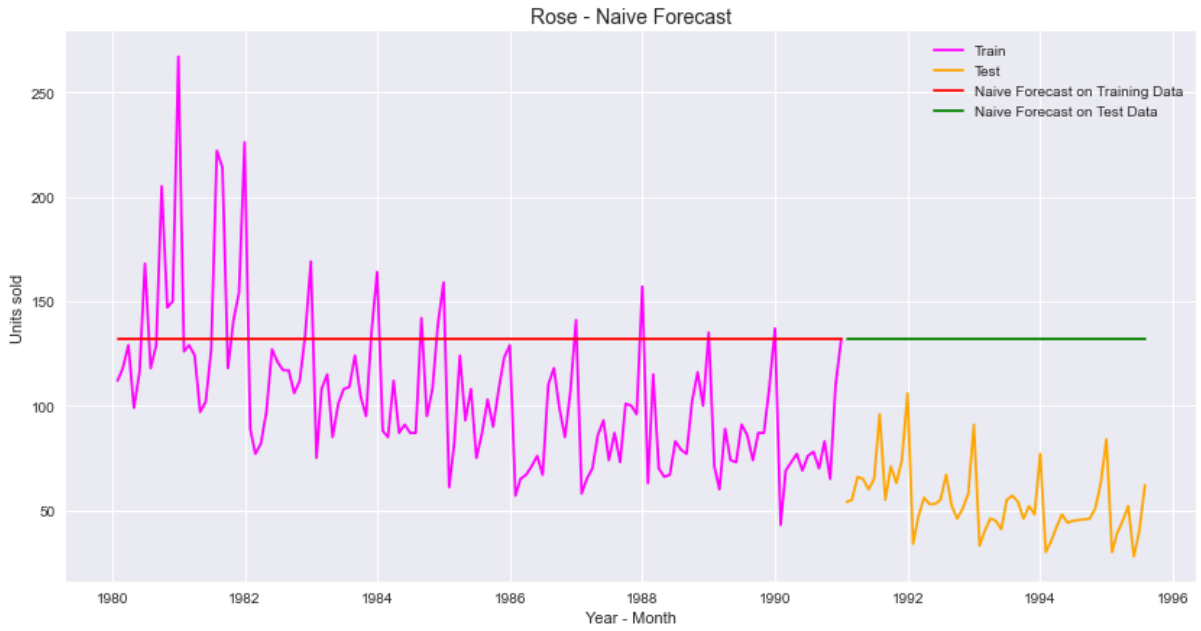


Fig:35

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.10

Table:38

Observations:

- The model does not capture the trend or seasonality of the given dataset.

SIMPLE AVERAGE FORECASTING

- In the Simple Average model, the forecast is done using the mean of the time-series variable from the training set

Sparkling

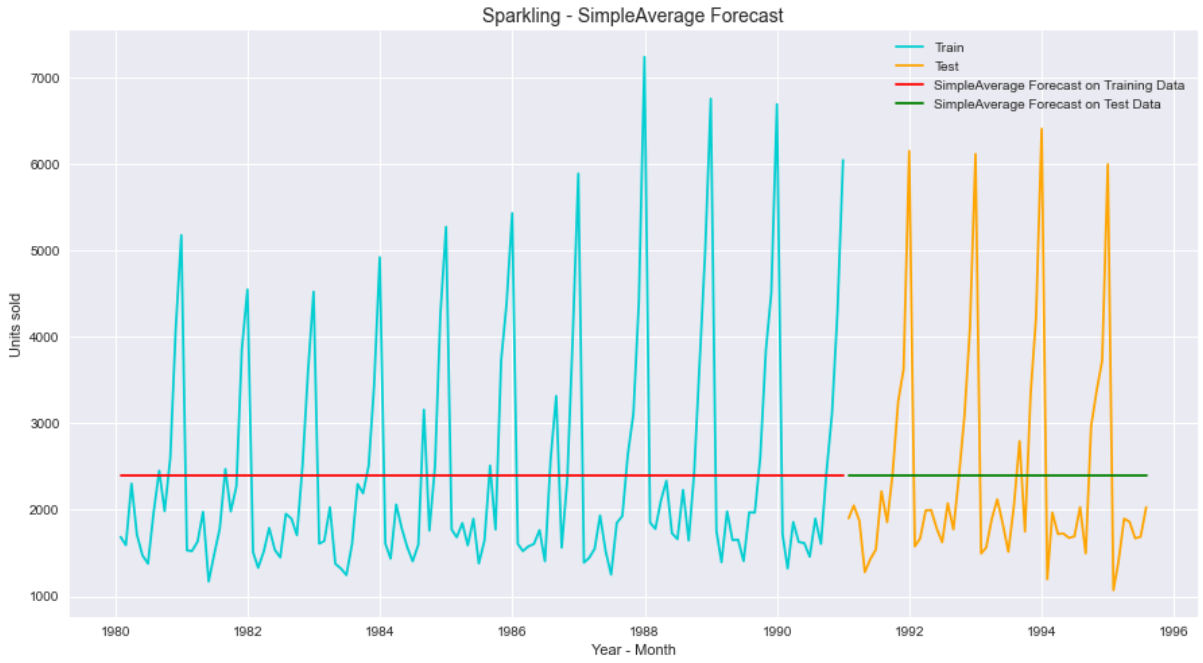


Fig:36

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90

Table:39

Observations:

- The model is not capable of either forecasting nor able to capture the trend and seasonality present in the dataset.

Rose

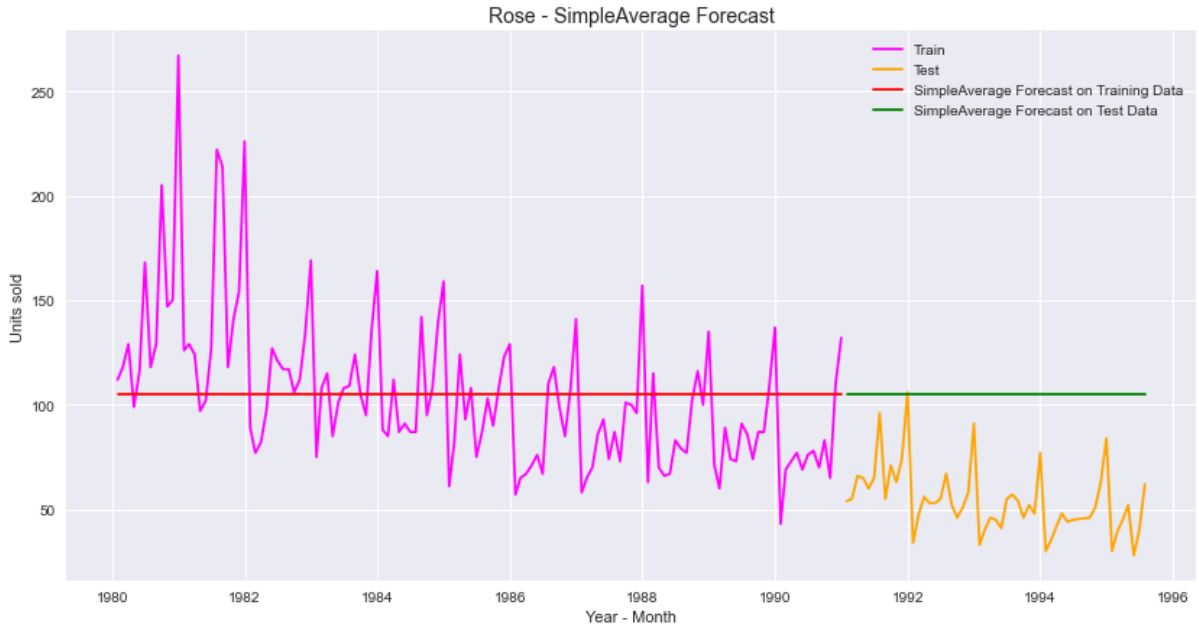


Fig:37

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93

Table:40

Observations:

- For the Rose dataset, the model forecast is almost 100% error in test data.
- Due to the downward trend, the performance in the train data set is better than the test dataset.

MOVING AVERAGE

- For the moving average model, we will calculate rolling means (or trailing moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error).

- The moving average models are built for trailing 2 points, 4 points, 6 points and 9 points.

- In moving average forecasts, the values can be fitted with a delay of n number of points.

- The Root Mean Squared Error and Mean Absolute Percentage Error of the test set are given below.

Sparkling

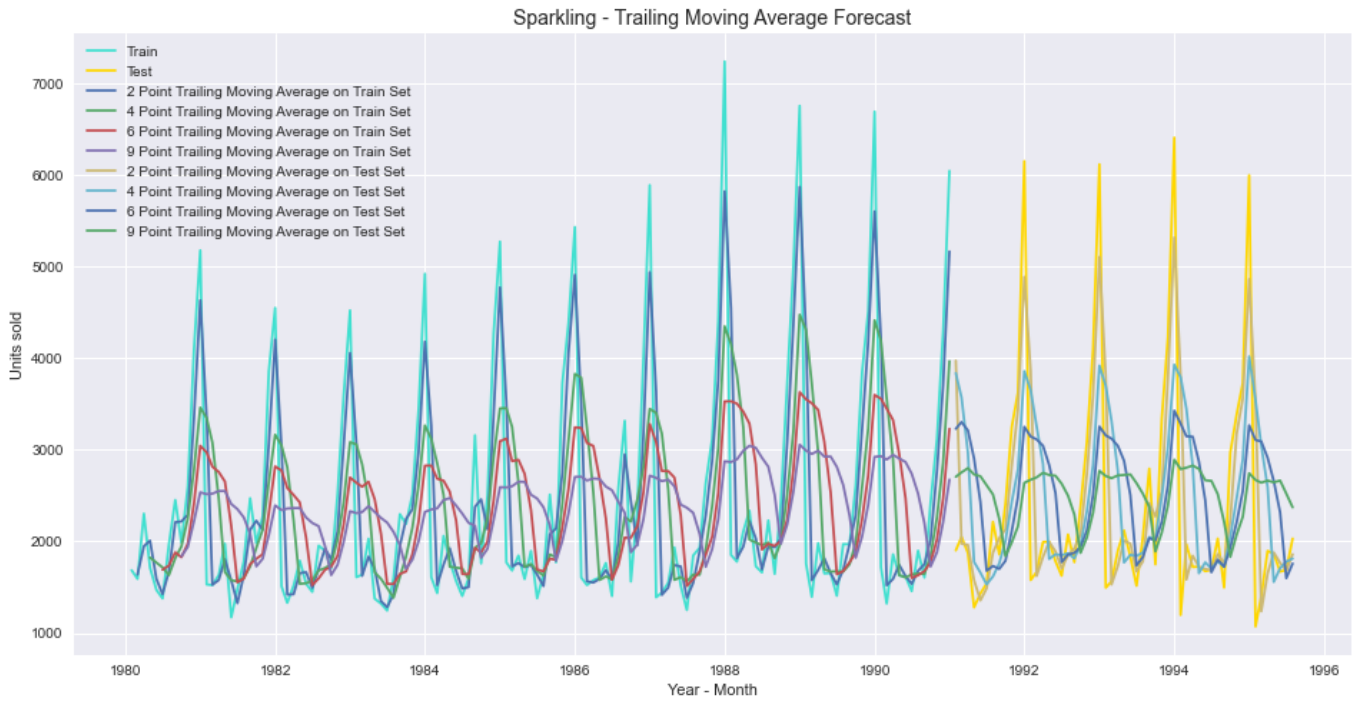


Fig:38

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86

Table:41

Observations:

- Moving Average is done only on Test Data.
- For the Sparkling dataset, the accuracy is found to be higher with the lower rolling point averages. The best interval of moving average from the model is 2-point TMA.

Rose

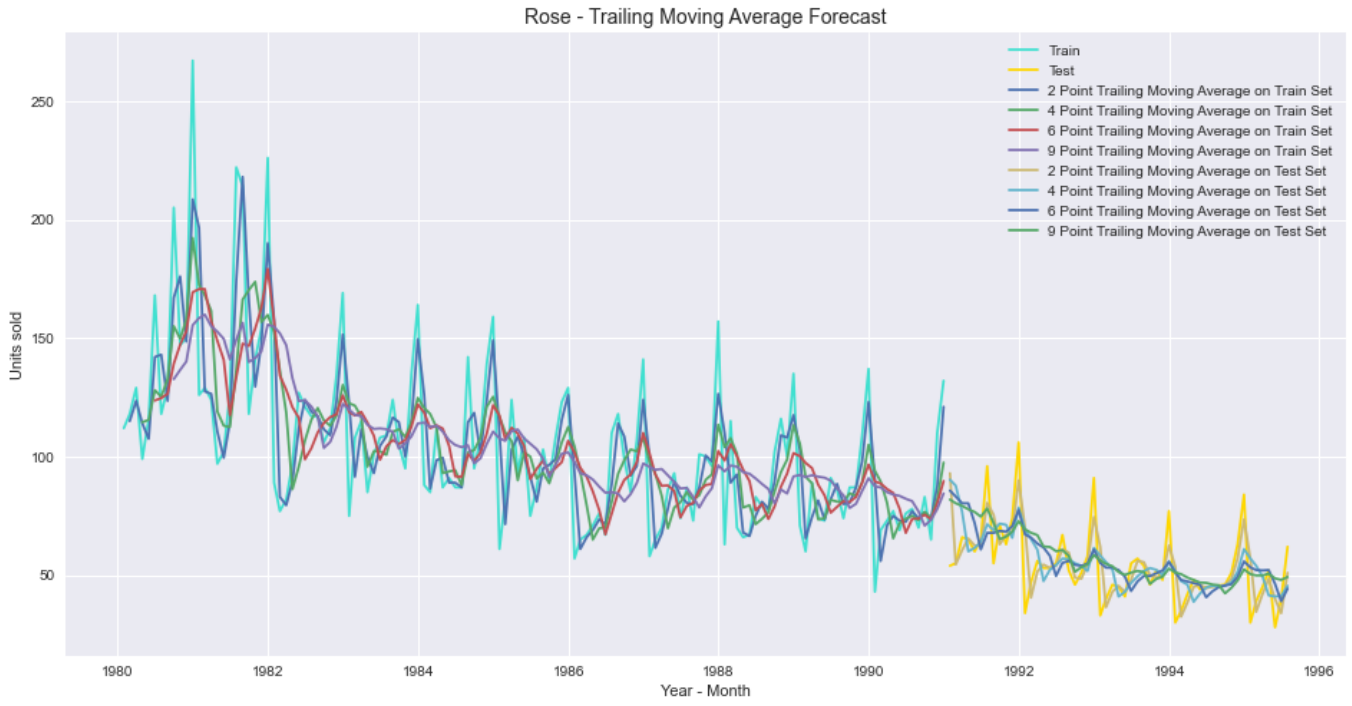


Fig:39

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01

Table:42

Observations:

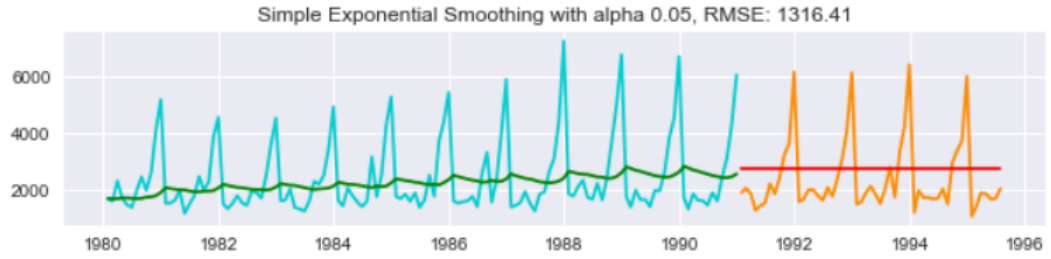
- Moving Average is done only on Test Data.
- For the Rose dataset also, the accuracy is found to be higher with the lower rolling point averages. The best interval of moving average from the model is 2-point TMA.

SIMPLE EXPONENTIAL SMOOTHING

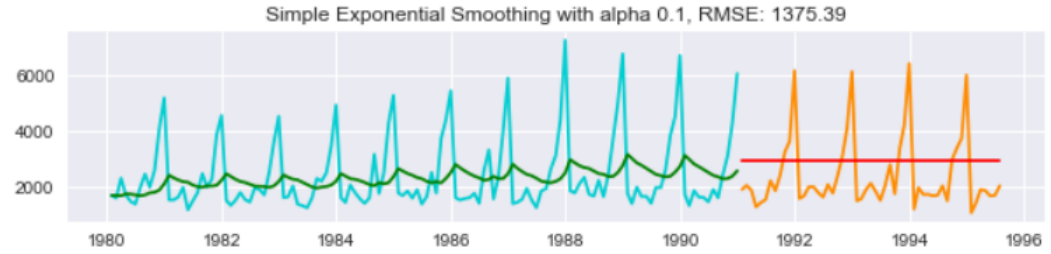
- Simple Exponential Smoothing is applied if the time-series has neither a trend or seasonality, which is not the case with the given data.
- The forecasting using smoothing levels of alpha between 0 and 1 are as below, where the smoothing levels are passed manually.
- For alpha values closer to 1, forecasts follow the actual observation closely and closer to 0, forecasts are farther from actual and line gets smoothened.

Sparkling

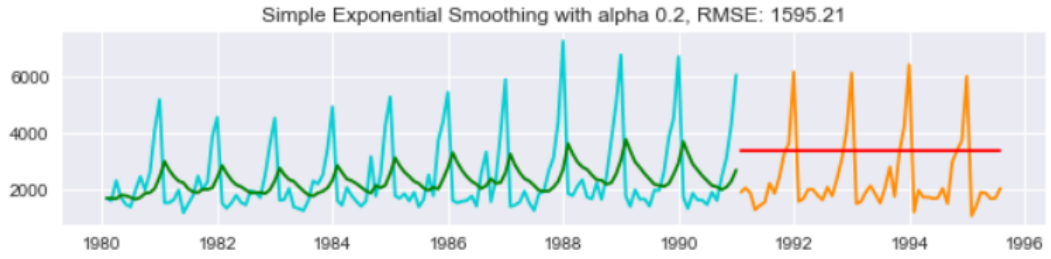
```
Test: For alpha = 0.05,  RMSE is 1316.4117 MAPE is 45.50
For smoothing level = 0.05,  Initial level 1686.00
```



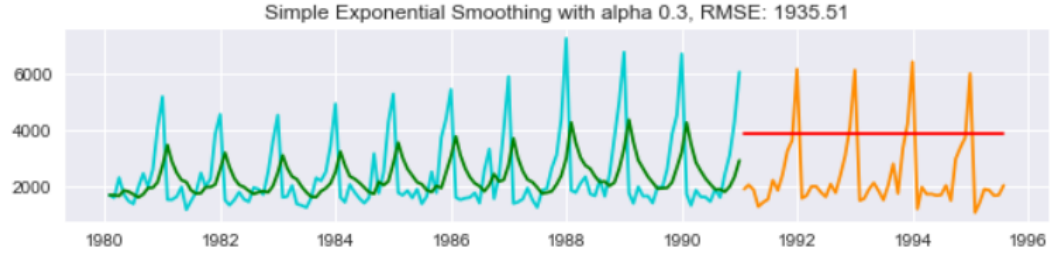

```
Test: For alpha = 0.10,  RMSE is 1375.3934 MAPE is 49.53
For smoothing level = 0.10,  Initial level 1686.00
```



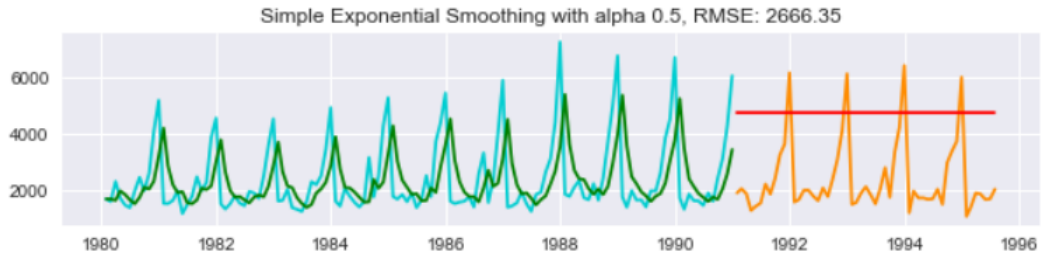

```
Test: For alpha = 0.20,  RMSE is 1595.2068 MAPE is 60.46
For smoothing level = 0.20,  Initial level 1686.00
```




```
Test: For alpha = 0.30,  RMSE is 1935.5071 MAPE is 75.66
For smoothing level = 0.30,  Initial level 1686.00
```



```
Test: For alpha = 0.50,  RMSE is 2666.3514 MAPE is 106.27
For smoothing level = 0.50,  Initial level 1686.00
```




```
Test: For alpha = 0.99,  RMSE is 3847.5490 MAPE is 152.21
For smoothing level = 0.99,  Initial level 1686.00
```

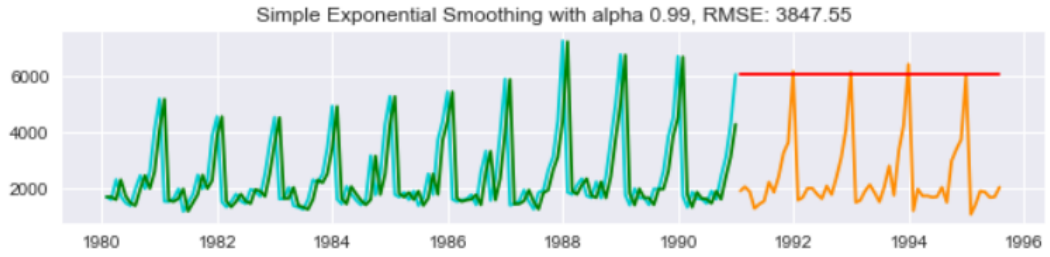


Fig:40

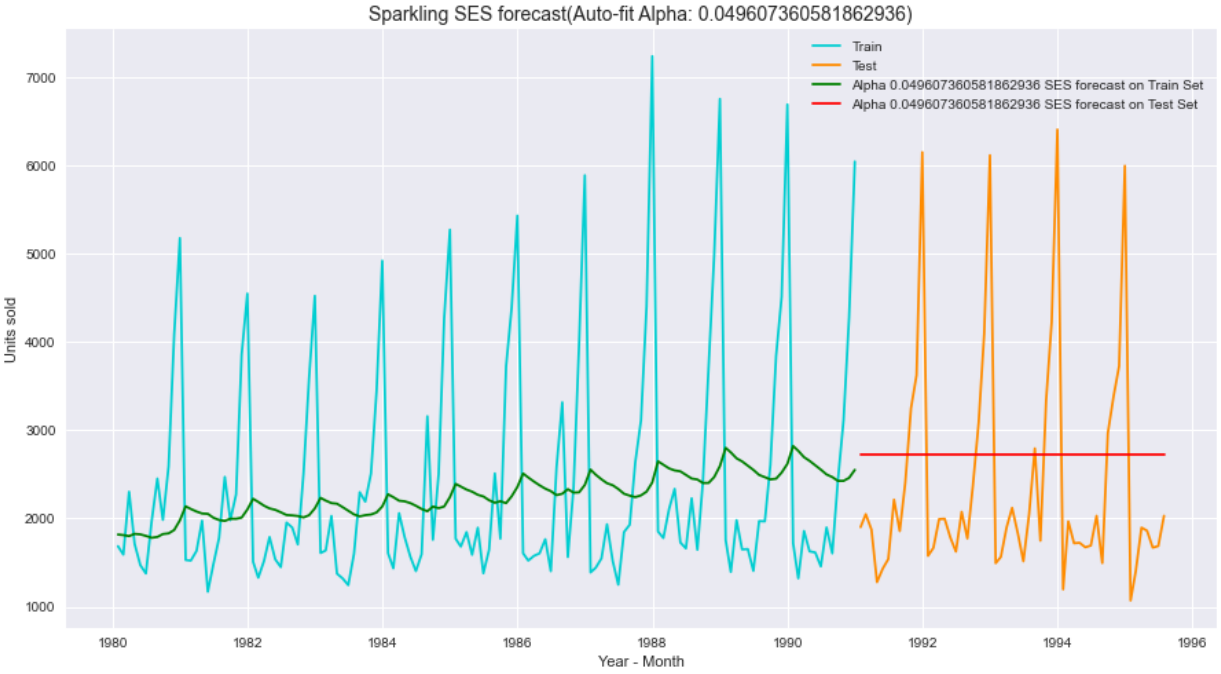


Fig:41

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.049607360581862936	1316.035487	45.47

Table:43

Observations:

- For Sparkling, the test RMSE is found to be higher, which is the same as in the Simple average forecast.

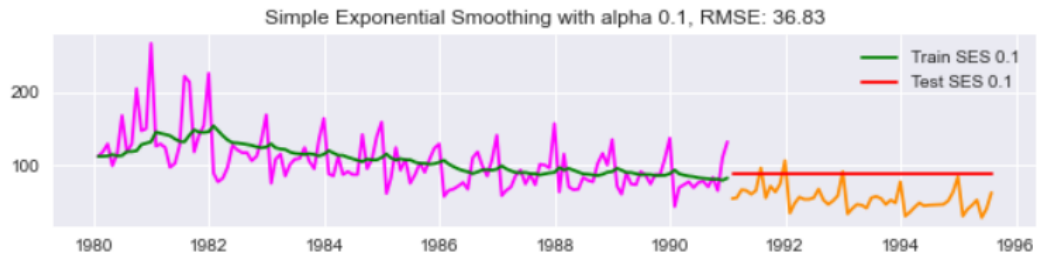
- On the second iteration, the model was running without passing a value for alpha and used parameters 'optimized=True, use_brute=True'.

- The autofit model picked 0.049 as the smoothing parameter and returned consistent RMSE values in train and test datasets, which is higher in accuracy than in the first iteration.

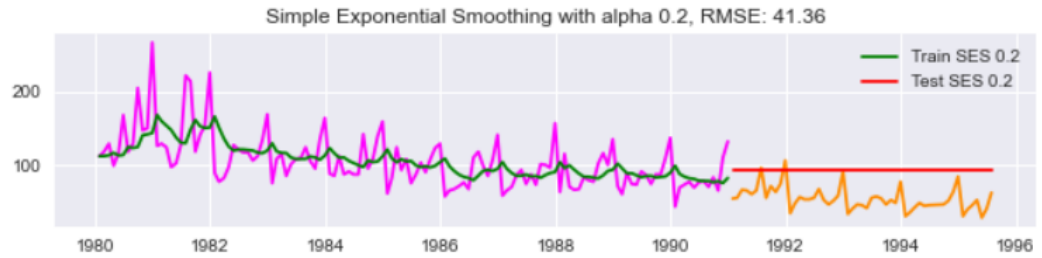
- As the smoothing level is 0.049, we got a completely smoothed out forecast with an initial value 1316 applied across the series.

Rose

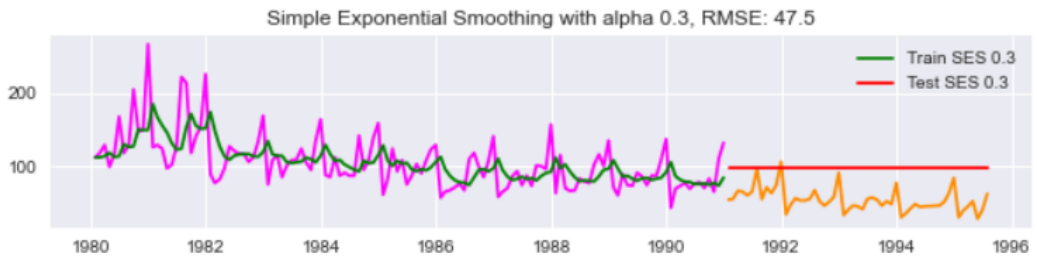
```
Test: For alpha = 0.10,  RMSE is 36.8278 MAPE is 63.94
For smoothing level = 0.10,  Initial level 112.00
```



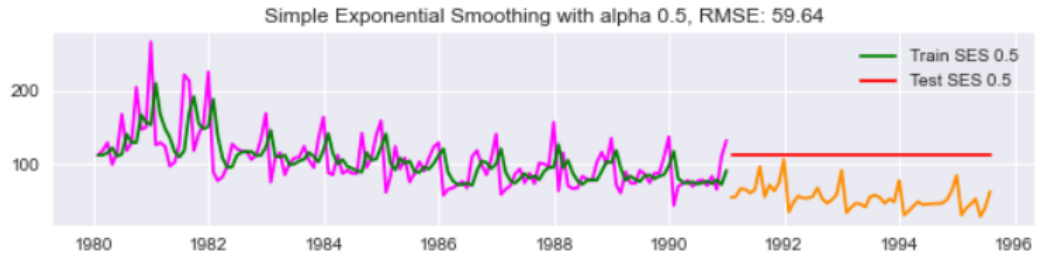

```
Test: For alpha = 0.20,  RMSE is 41.3617 MAPE is 72.21
For smoothing level = 0.20,  Initial level 112.00
```




```
Test: For alpha = 0.30,  RMSE is 47.5046 MAPE is 83.71
For smoothing level = 0.30,  Initial level 112.00
```




```
Test: For alpha = 0.50,  RMSE is 59.6416 MAPE is 106.81
For smoothing level = 0.50,  Initial level 112.00
```



```
Test: For alpha = 0.99,  RMSE is 79.4985 MAPE is 144.69
For smoothing level = 0.99,  Initial level 112.00
```

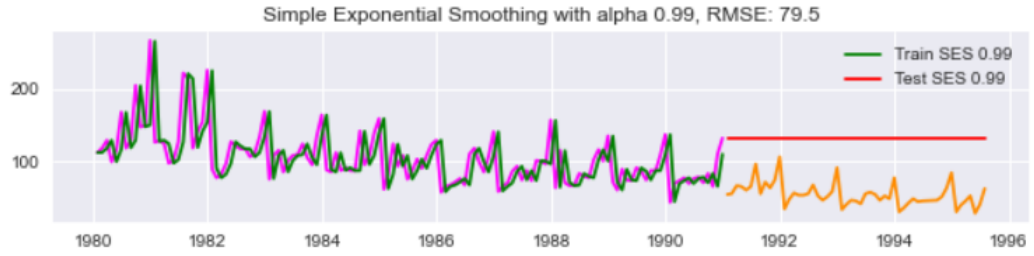


Fig:42

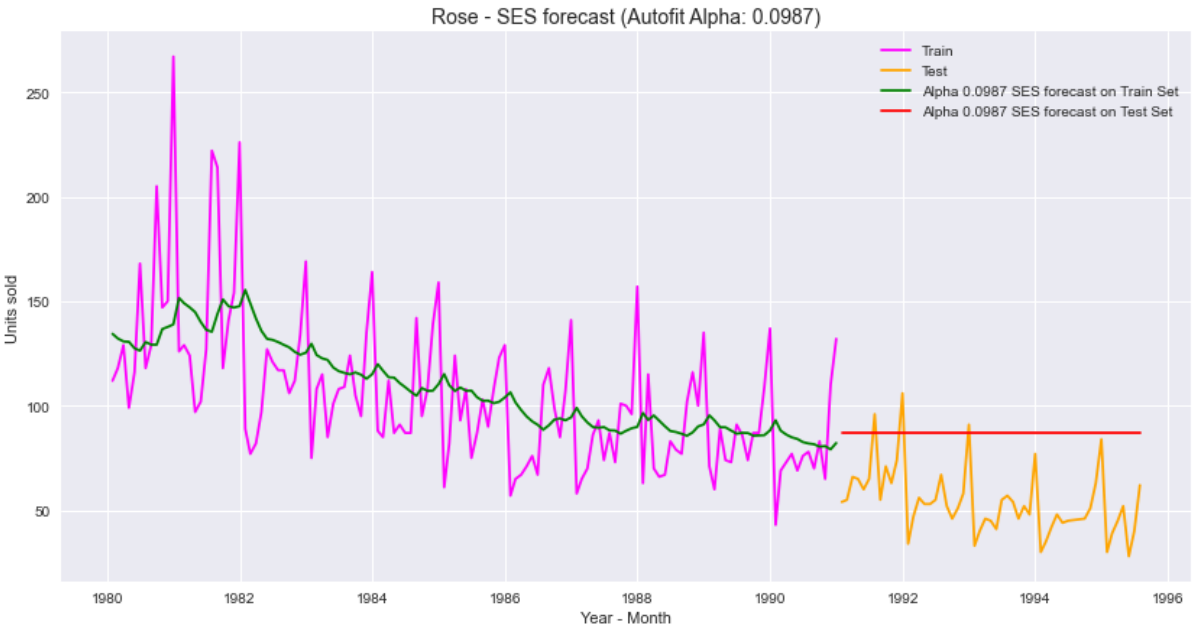


Fig:43

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
SES Alpha 0.01	36.796004	63.88

Table:44

Observations:

- On the second iteration, the model was running without passing a value for alpha and used parameters 'optimized=True, use_brute=True'.

- The autofit model picked 0.098 as the smoothing parameter and retuned consistent RMSE values in train and test datasets, which is consistent with alpha 0.1 in first iteration.

DOUBLE EXPONENTIAL SMOOTHING (Holt's Model)

- The Double Exponential Smoothing model is applicable when data have trends, but no seasonality. Sparkling data contain a slight trend component and very significant seasonality.

- In the first iteration, smoothing level (alpha) and trend (beta) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.1 and beta 0.1.

- On the second iteration the model was allowed to choose the optimized values using parameters 'optimized=True, use_brute=True'.

Sparkling

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	1363.47	44.26	1779.42	67.23
1	0.1	0.2	1398.19	45.61	2601.54	95.50
10	0.2	0.1	1412.03	46.62	3611.77	135.41
2	0.1	0.3	1431.37	46.90	4288.43	155.25
20	0.3	0.1	1428.27	46.92	5908.19	223.50

Table:45

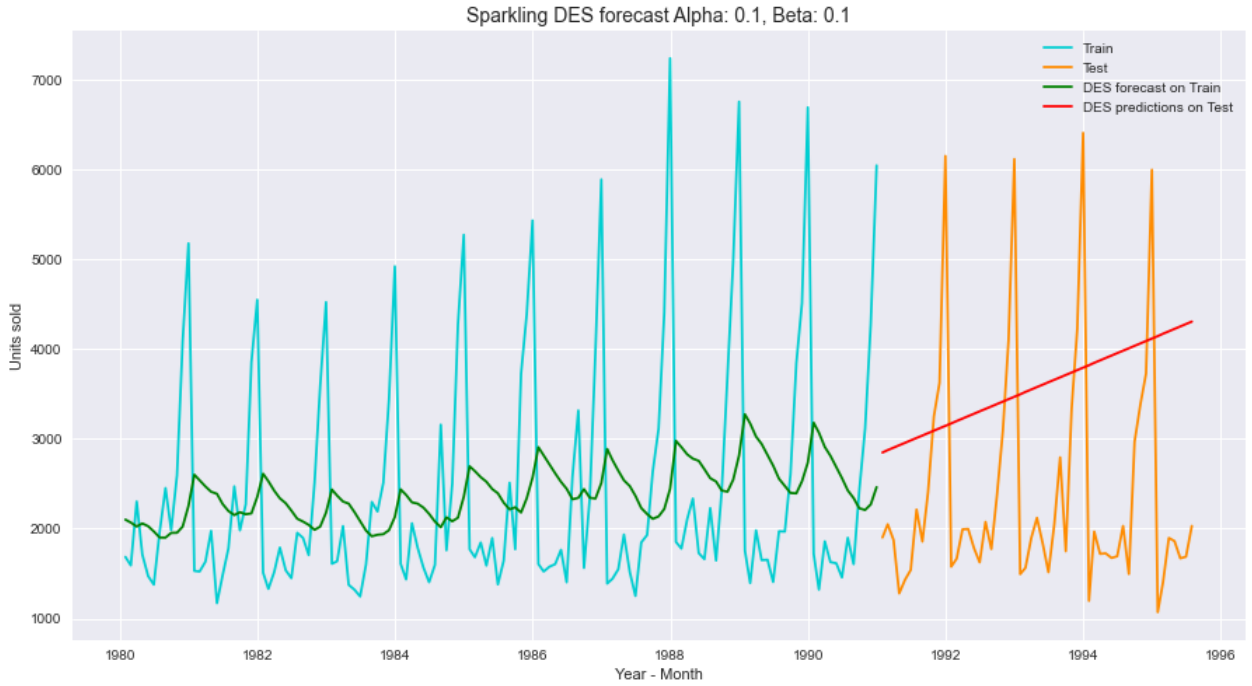


Fig:44

Trying Auto-fit by the model

```
model_DES_autofit.params
```

```
{'smoothing_level': 0.6885714285714285,
 'smoothing_trend': 9.999999999999999e-05,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1686.0,
 'initial_trend': -95.0,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove bias': False}
```

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100000	0.1000	1363.47000	44.26	1779.420000	67.23
100	0.688571	0.0001	1349.65046	39.23	2007.238526	68.23
1	0.100000	0.2000	1398.19000	45.61	2601.540000	95.50
10	0.200000	0.1000	1412.03000	46.62	3611.770000	135.41
2	0.100000	0.3000	1431.37000	46.90	4288.430000	155.25

Table:46

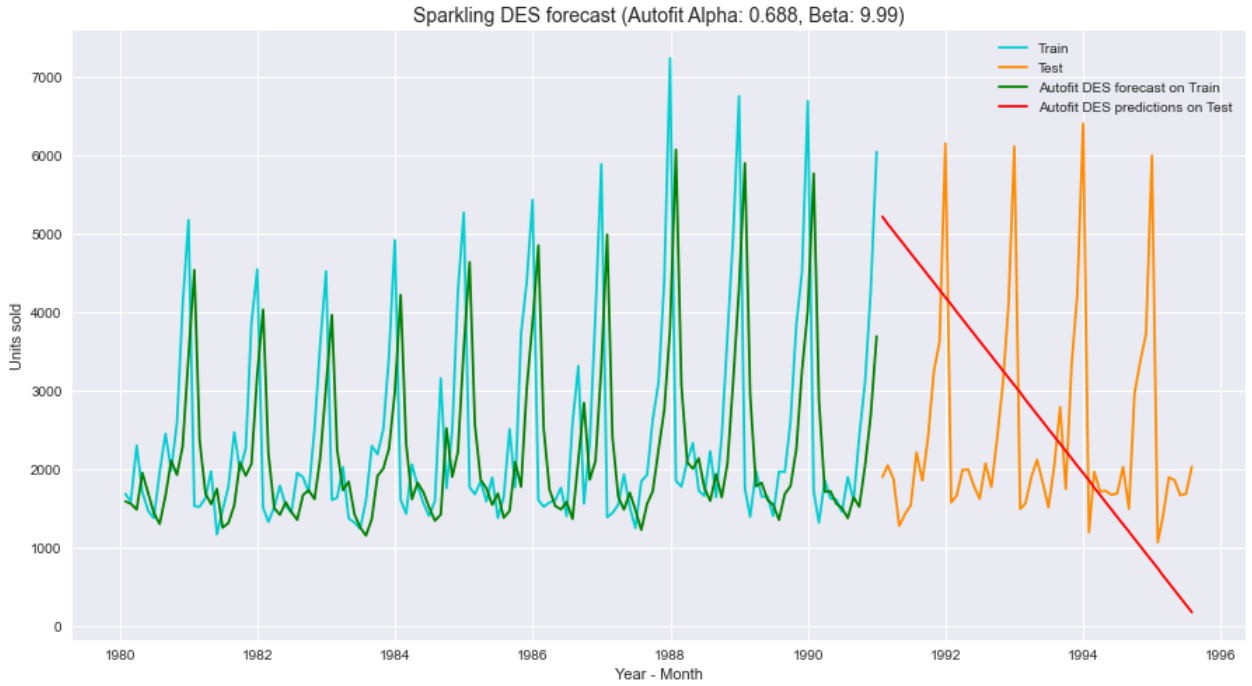


Fig:45

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.049607360581862936	1316.035487	45.47
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23

Table:47

Observations:

- The autofit model returned higher accuracy in the train dataset, but fared poorly in tests, compared with the values in manual iteration.

- The model evaluation parameters of the top three models from manual iteration and the autofit models are as given above.

- The best model chosen as the final one is with alpha 0.1 and beta 0.1.

Rose

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	32.026565	22.78	37.056911	64.02
1	0.1	0.2	32.685228	23.63	48.806921	83.29
10	0.2	0.1	32.796403	23.06	65.731352	113.20
2	0.1	0.3	32.925494	24.23	78.209401	131.33
20	0.3	0.1	33.528397	23.47	98.653063	170.12

Table:48

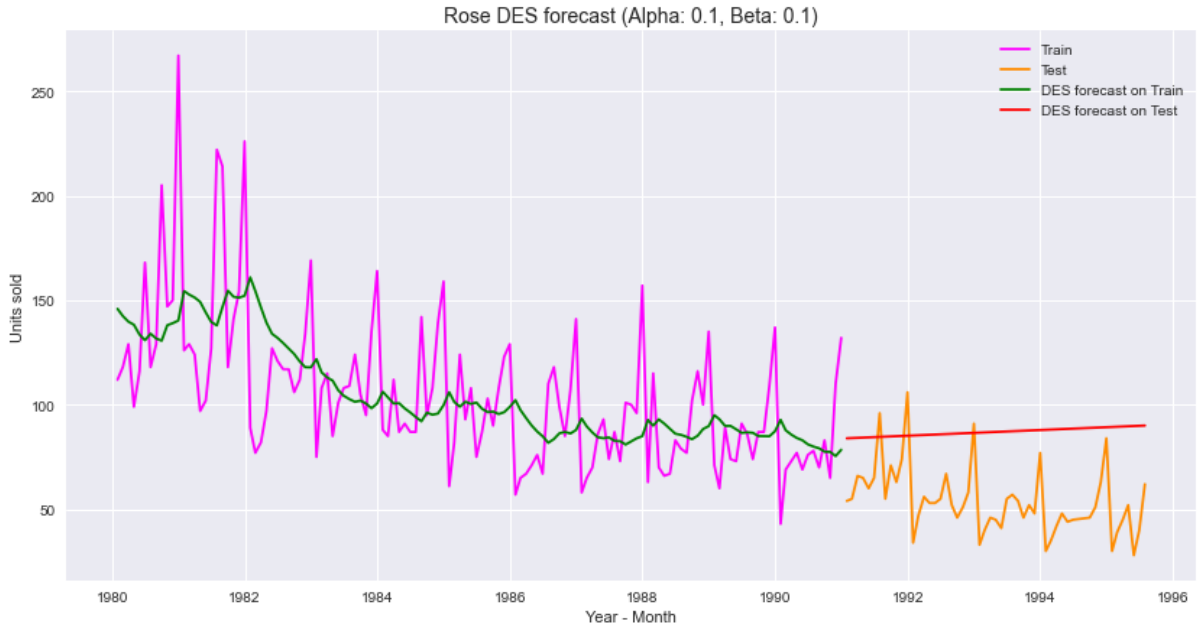


Fig:46

Trying Auto-fit by the model

```
model_DES_rose_autofit.params
```

```
{'smoothing_level': 0.017549790270679714,
 'smoothing_trend': 3.236153800377395e-05,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 138.82081494774005,
 'initial_trend': -0.492580228245491,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
100	0.01755	0.000032	30.890794	21.61	15.706968	24.12
0	0.10000	0.100000	32.026565	22.78	37.056911	64.02
1	0.10000	0.200000	32.685228	23.63	48.806921	83.29
10	0.20000	0.100000	32.796403	23.06	65.731352	113.20
2	0.10000	0.300000	32.925494	24.23	78.209401	131.33

Table:49

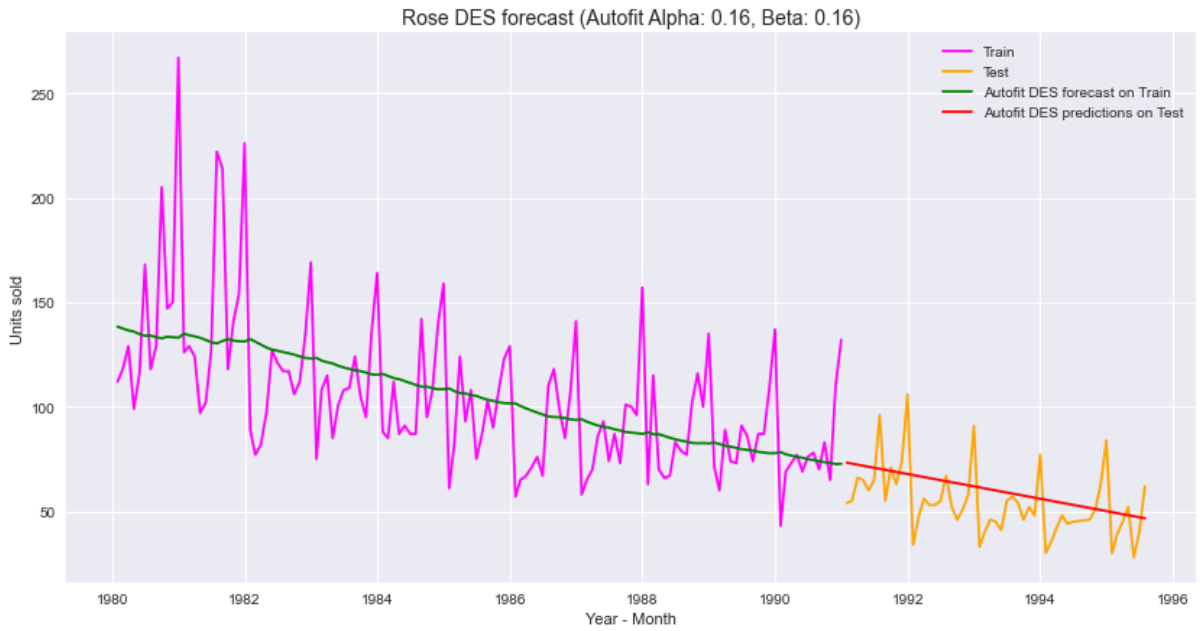


Fig:47

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.16, Beta 0.16	15.706968	24.12
DES Alpha 0.10, Beta 0.10	37.056911	64.02

Table:50

Observations:

- The autofit model returned higher accuracy in the train dataset, in comparison with the best models from iteration 1, but fared behind in the test accuracy scores.
- The model evaluation parameters of the best models are given as above.
- The best model chosen as the final one is the one with alpha 0.1 and beta 0.1.

TRIPLE EXPONENTIAL SMOOTHING(Holt Winter's Model)

- The Triple Exponential Smoothing model (Holt-Winter's Model) is applicable when data has both trend and seasonality. Sparkling data contain slight trends and significant seasonality.

Sparkling

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
211	0.3	0.2	0.2	377.777799	11.26	314.906684	10.10
301	0.4	0.1	0.2	374.619861	11.24	315.935531	10.45
300	0.4	0.1	0.1	370.612639	11.03	318.103496	10.01
402	0.5	0.1	0.3	390.175608	11.54	325.544937	9.99
30	0.1	0.4	0.1	400.768319	11.48	331.257015	10.54

Table:51

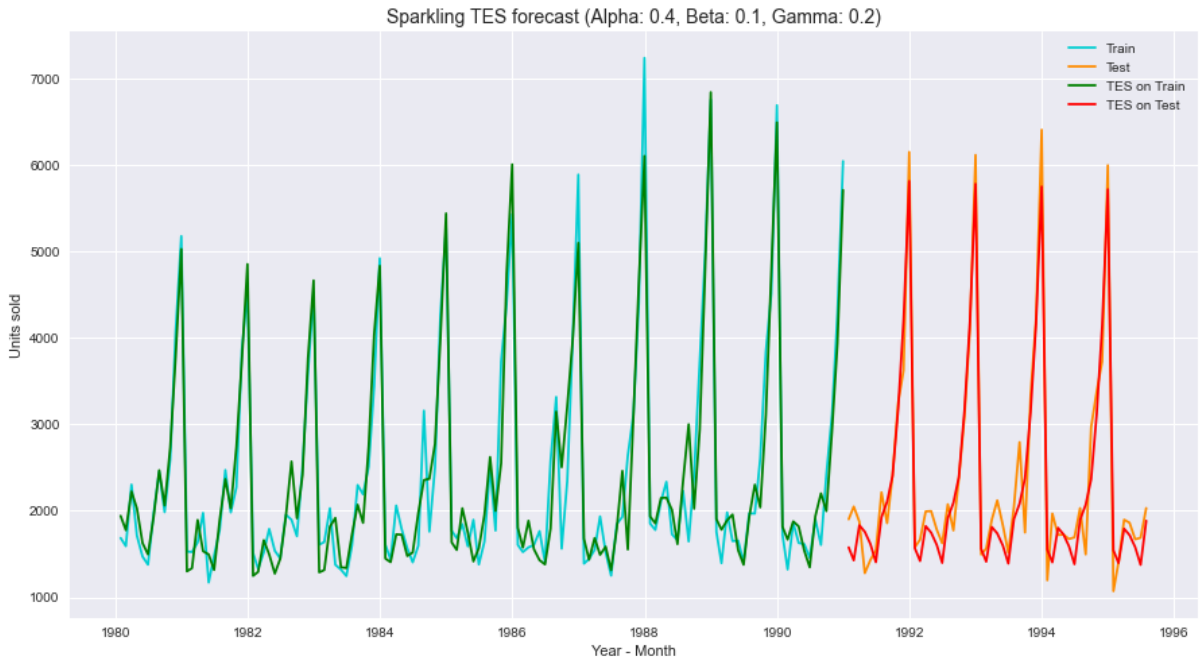


Fig:48

Trying Auto-fit by the model

```
model_TES_autofit.params
```

```
{'smoothing_level': 0.11133818361298699,
 'smoothing_trend': 0.049505131019509915,
 'smoothing_seasonal': 0.3620795793580111,
 'damping_trend': nan,
 'initial_level': 2356.4967888704355,
 'initial_trend': -10.187944726007238,
 'initial_seasons': array([0.71296382, 0.68242226, 0.90755008, 0.80515228, 0.65597218,
        0.65414505, 0.88617935, 1.13345121, 0.92046306, 1.21337874,
        1.87340336, 2.37811768]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
211	0.3	0.2	0.2	377.777799	11.26	314.906684	10.10
301	0.4	0.1	0.2	374.619861	11.24	315.935531	10.45
300	0.4	0.1	0.1	370.612639	11.03	318.103496	10.01
402	0.5	0.1	0.3	390.175608	11.54	325.544937	9.99
30	0.1	0.4	0.1	400.768319	11.48	331.257015	10.54

Table:52

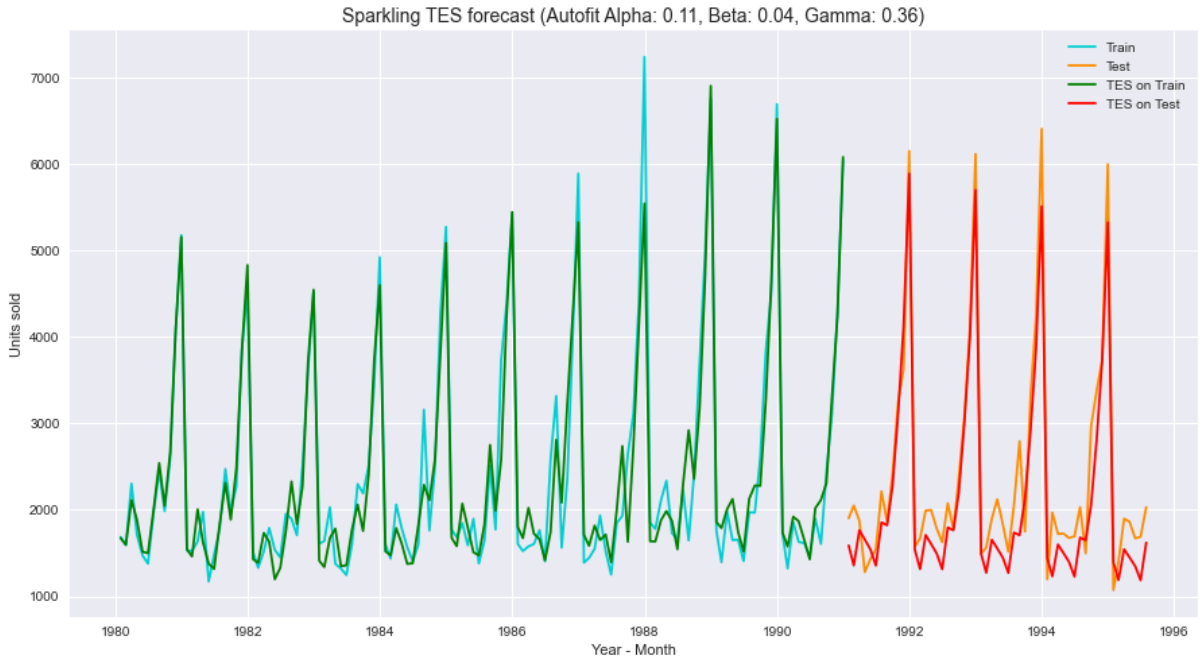


Fig:49

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.049607360581862936	1316.035487	45.47
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.935531	10.45
TES Alpha 0.15, Beta 0.00, Gamma 0.37	404.286809	13.93

Table:53

Observations

- In first iteration, smoothing level (alpha), trend (beta) and seasonality (gamma) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.4, beta 0.1 and gamma 0.2.
- On the second iteration the model was allowed to choose the optimized values using parameters 'optimized=True, use_brute=True'.

- The autofit model returned similar results as in iteration 1 for both train and test datasets.

- The model evaluation parameters of the best models are given as above, including one from the autofit iteration.

- The best model chosen as the final one is the one with alpha 0.4, beta 0.1 and gamma 0.2.

Rose

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
10	0.1	0.2	0.1	19.651464	14.31	9.171615	13.19
11	0.1	0.2	0.2	20.140683	14.66	9.493835	13.68
151	0.2	0.6	0.2	22.793871	17.02	9.682585	13.71
142	0.2	0.5	0.3	23.300524	17.35	9.885717	14.21
12	0.1	0.2	0.3	20.725703	14.88	9.896169	14.16

Table:54

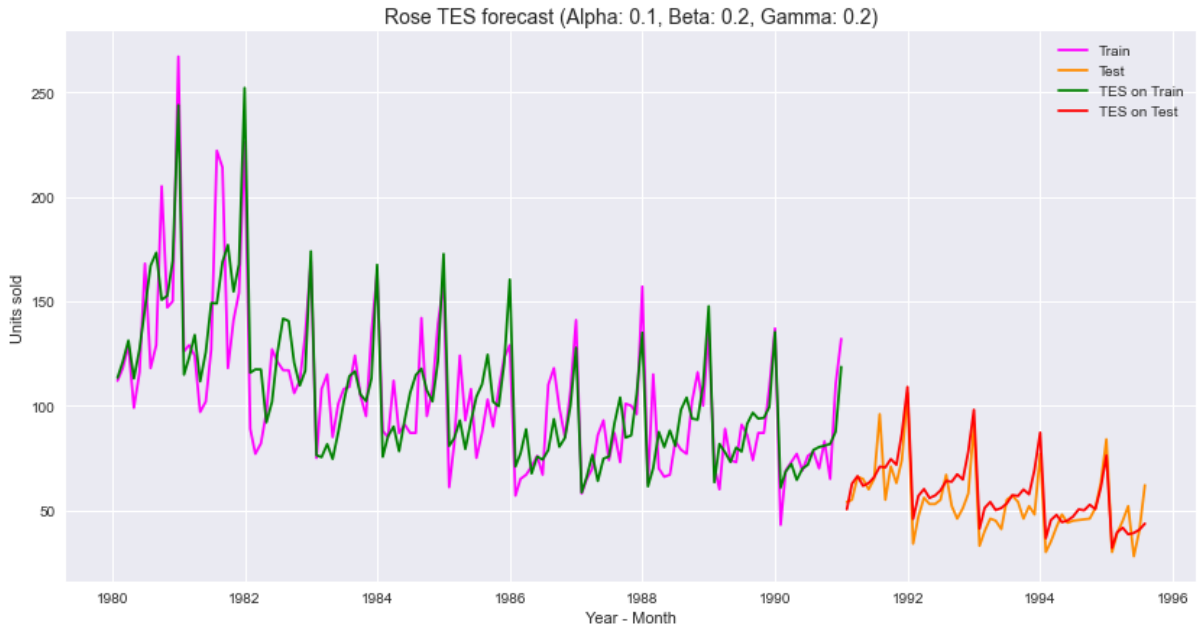


Fig:50

Trying Auto-fit by the model

```
model_TES_autofit.params
```

```
{'smoothing_level': 0.0715106306609405,
 'smoothing_trend': 0.04529179757535142,
 'smoothing_seasonal': 7.244325029450242e-05,
 'damping_trend': nan,
 'initial_level': 130.40839142502193,
 'initial_trend': -0.77985743179386,
 'initial_seasons': array([0.86218996, 0.977675  , 1.0687727 , 0.93403881, 1.050625  ,
        1.14410977, 1.25836944, 1.33937772, 1.26778766, 1.24131254,
        1.44724625, 1.99553681]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
10	0.1	0.2	0.1	19.651464	14.31	9.171615	13.19
11	0.1	0.2	0.2	20.140683	14.66	9.493835	13.68
151	0.2	0.6	0.2	22.793871	17.02	9.682585	13.71
142	0.2	0.5	0.3	23.300524	17.35	9.885717	14.21
12	0.1	0.2	0.3	20.725703	14.88	9.896169	14.16

Table:55

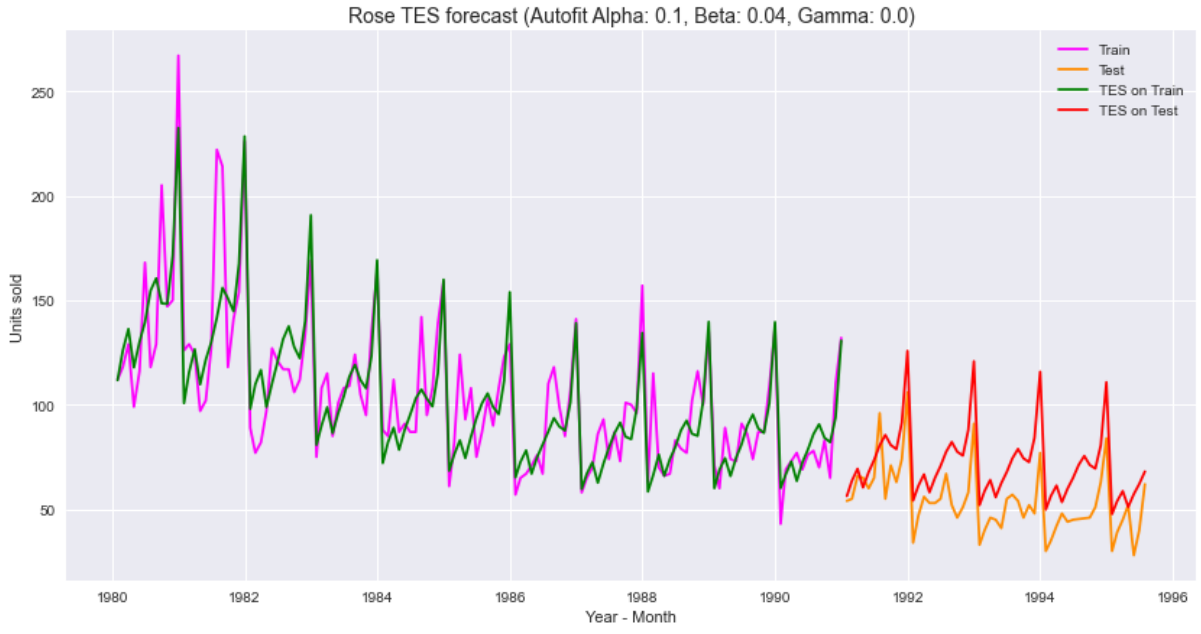


Fig:51

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.16, Beta 0.16	15.706968	24.12
DES Alpha 0.10, Beta 0.10	37.056911	64.02
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.493835	13.68
TES Alpha 0.11, Beta 0.05, Gamma 0.00	20.156483	33.63

Table:56

Observations:

- In first iteration, smoothing level (alpha), trend (beta) and seasonality (gamma) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.1, beta 0.2 and gamma 0.2.

- On the second iteration the model was allowed to choose the optimized values using parameters ‘optimized=True, use_brute=True’.

- The autofit model returned similar results as in iteration 1 for both train and test datasets.

- The model evaluation parameters of the best models are given as above, including one from the autofit iteration.

- The best model chosen as the final one is the one with alpha 0.1, beta 0.2 and gamma 0.2.

MODEL COMPARISON

Sparkling

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.935531	10.45
TES Alpha 0.15, Beta 0.00, Gamma 0.37	404.286809	13.93
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
6 point TMA	1283.927428	43.86
SES Alpha 0.049607360581862936	1316.035487	45.47
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23
NaiveModel	3864.279352	152.87

Table:57

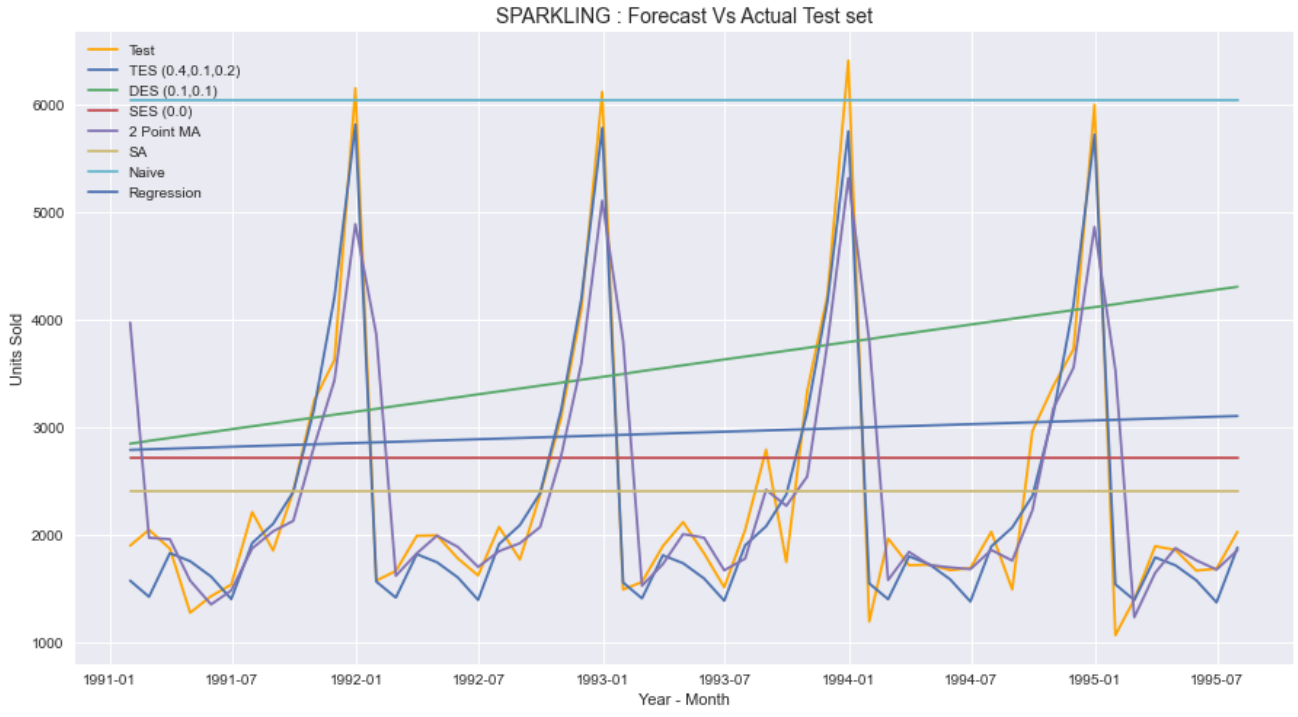


Fig:52

Observations:

- The accuracy of the time-series forecast models built in the previous sections of this report is as below, sorted by RMSE in test data.
- The plot of the forecasts fitted on to the test data is given as well.
- From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data.
- The 2-point trailing moving average model is also found to have fit well with a slight lag in the test dataset.

Rose

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.493835	13.68
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
DES Alpha 0.16, Beta 0.16	15.706968	24.12
TES Alpha 0.11, Beta 0.05, Gamma 0.00	20.156483	33.63
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.10, Beta 0.10	37.056911	64.02
SimpleAverage	53.460350	94.93
NaiveModel	79.718559	145.10

Table:58

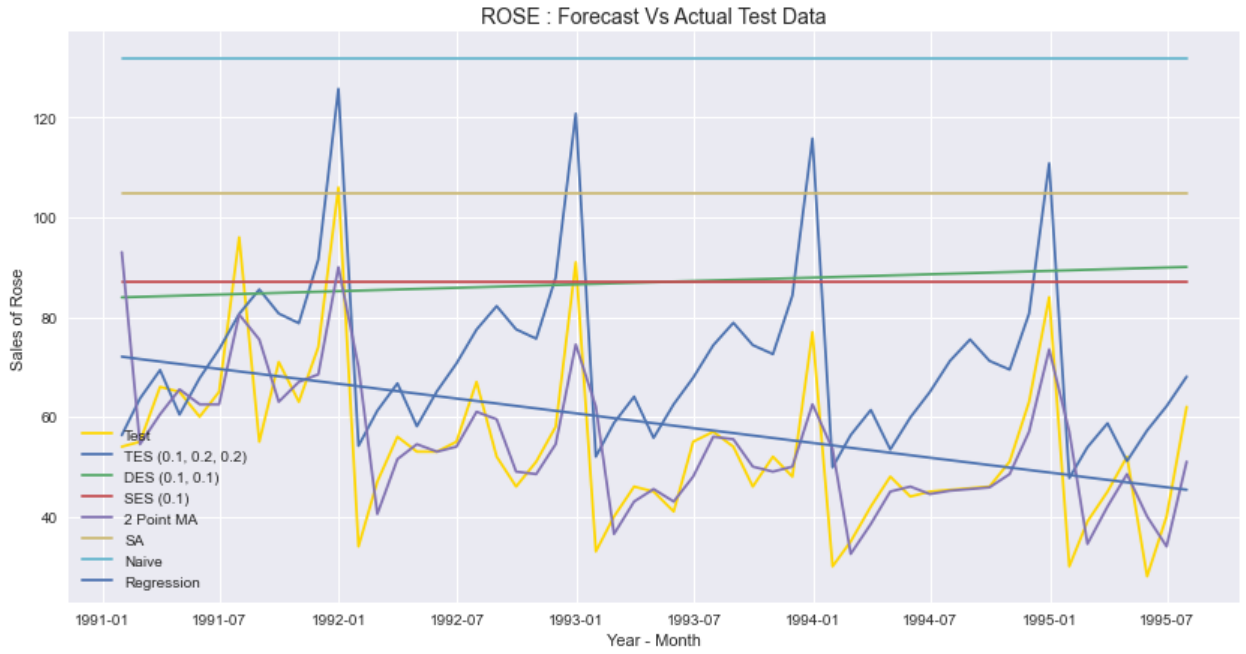


Fig:53

Observations:

- The accuracy of the time-series forecast models built in the previous sections of this report is as below, sorted by RMSE in test data.

- The plot of the forecasts fitted on to the test data is given as well.

- From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data.

- The 2-point trailing moving average model is also found to have fit well with a slight lag in the test dataset.

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and 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

Original Series

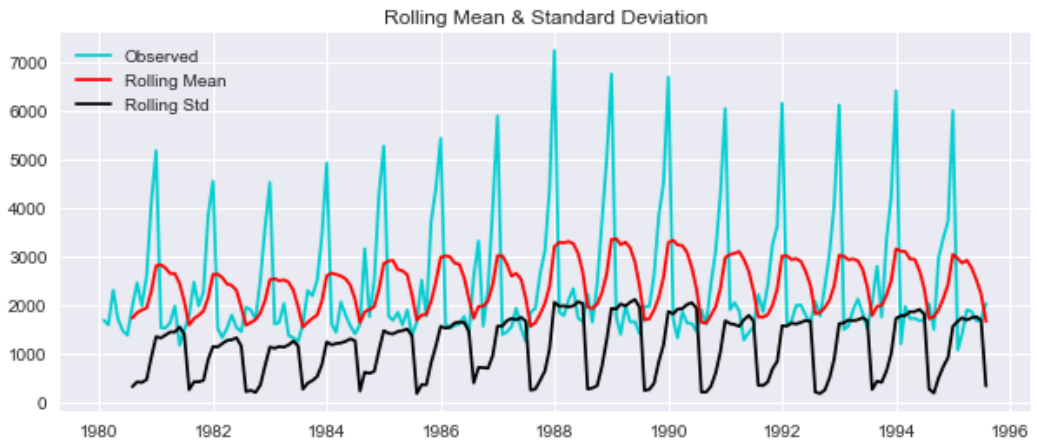


Fig:54

```
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
```

Table:59

ADF on original series

- P-Value > alpha .05
- Test statistic > Critical values
- Fail to reject the null hypothesis
- The series is non-stationary

Observations:

- The Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determines the presence of unit root in the series to understand if the series is stationary or not.

- **Null Hypothesis:** The series has a unit root, that is, the series is non-stationary.
- **Alternate Hypothesis:** The series has no unit root, that is, the series is stationary.

- If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that the series is stationary.

- The ADF test on the original Sparkling series returned the below values, where the p-value is greater than alpha .05 so we fail to reject the null hypothesis.

Differenced Series

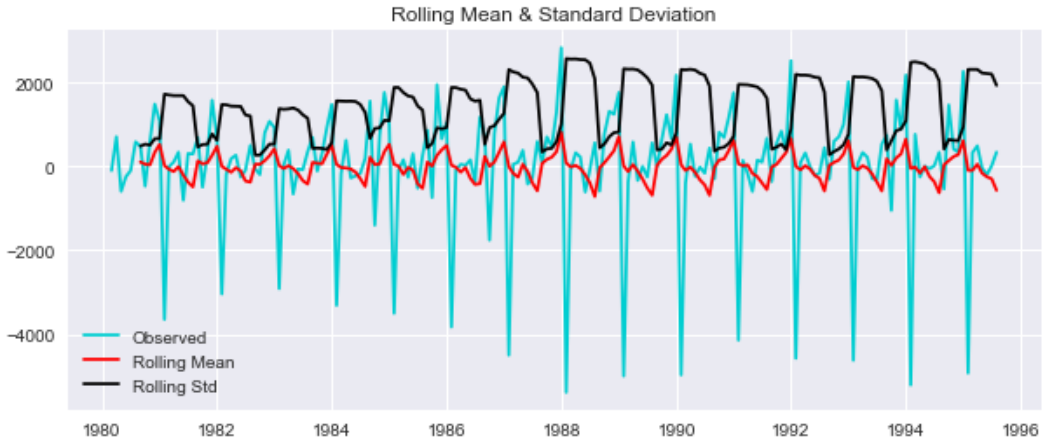


Fig:55

```
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
```

Table:60

ADF on differenced series

- P-Value < alpha .05
- Test statistic < Critical values
- Reject the null hypothesis
- The series is stationary

Observations:

- Differentiating order, one is applied to the Sparkling series as above and tested for stationarity. At an order of differencing 1, the series is found to be stationary as above.

- The rolling mean and standard deviation are also plotted to understand the component of seasonality and to ascertain if it is multiplicative or additive in character.

- The altitude of rolling mean and std dev is seen changing according to change in slope, which indicates multiplicity.

- The ADF test is also done in this exercise with logarithmic transformation of the train data and differencing of seasonal order (12), to understand if removing the multiplicity of the seasonal component will have an impact on the accuracy of the model.

Rose

Original Series

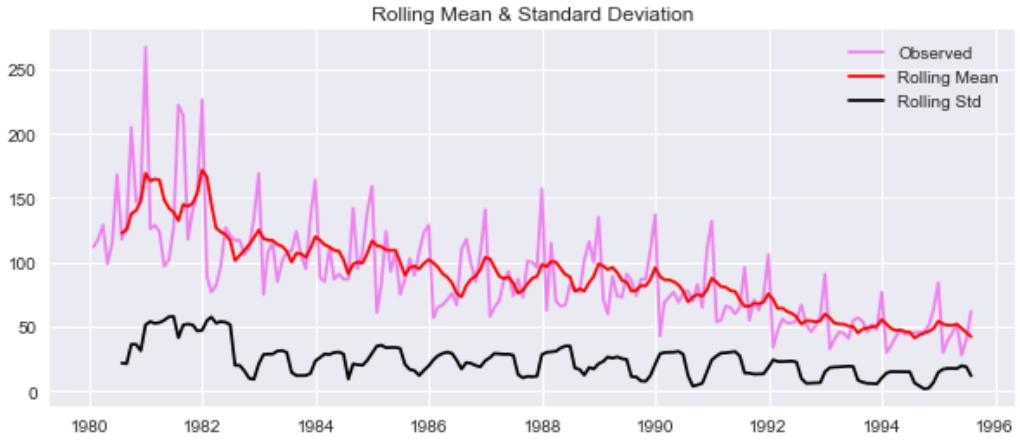


Fig:56

Results of Dickey-Fuller Test:	
Test Statistic	-1.876719
p-value	0.343091
#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	

Table:61

ADF on original series

- P-Value > alpha .05
- Test statistic > Critical values
- Fail to reject the null hypothesis
- The series is non-stationary

Observations:

- The Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determines the presence of unit root in the series to understand if the series is stationary or not.

- **Null Hypothesis:** The series has a unit root, that is, the series is non-stationary.

- **Alternate Hypothesis:** The series has no unit root, that is, the series is stationary.

- If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that the series is stationary.

- The ADF test on the original Rose series returned the below values, where the p-value is greater than alpha .05 so we fail to reject the null hypothesis.

Differenced Series

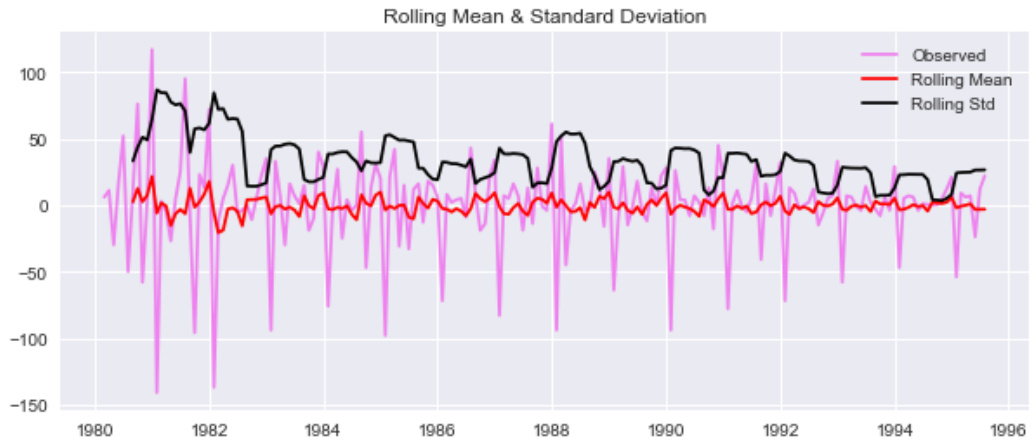


Fig:57

```
Results of Dickey-Fuller Test:
Test Statistic                -8.044395e+00
p-value                        1.810868e-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
```

Table:62

ADF on differenced series

- P-Value < alpha .05
- Test statistic < Critical values
- Reject the null hypothesis
- The series is stationary

Observations:

- Differentiation of order is applied to the Rose series as above and tested for stationarity.

- At an order of differencing 1, the series is found to be stationary as above.

- The rolling mean and standard deviation are also plotted to understand the component of seasonality and to ascertain if it is multiplicative or additive in character.

- The plot of rolling mean and standard deviation indicates that the seasonality is multiplicative as the altitude of the plot varies with respect to trend.

- The ADF test is also done in this exercise with logarithmic transformation of the train data and differencing of seasonal order (12), to understand if removing the multiplicity of the seasonal component will have an impact on the accuracy of the model.

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.

Auto-SARIMA Sparkling

- As the Sparkling series of data contain a seasonality component we will be building the SARIMA model, rather than ARIMA.

- Two iterations of automated SARIMA models were attempted in this exercise, one with original data and another with log transformation of the data, as an element of multiplicity in seasonality is suspected.

	param	seasonal	AIC
183	(2, 1, 3)	(1, 1, 3, 6)	1538.974356
247	(3, 1, 3)	(1, 1, 3, 6)	1543.202992
191	(2, 1, 3)	(3, 1, 3, 6)	1543.416154
251	(3, 1, 3)	(2, 1, 3, 6)	1543.586534
55	(0, 1, 3)	(1, 1, 3, 6)	1543.929506

Table:63

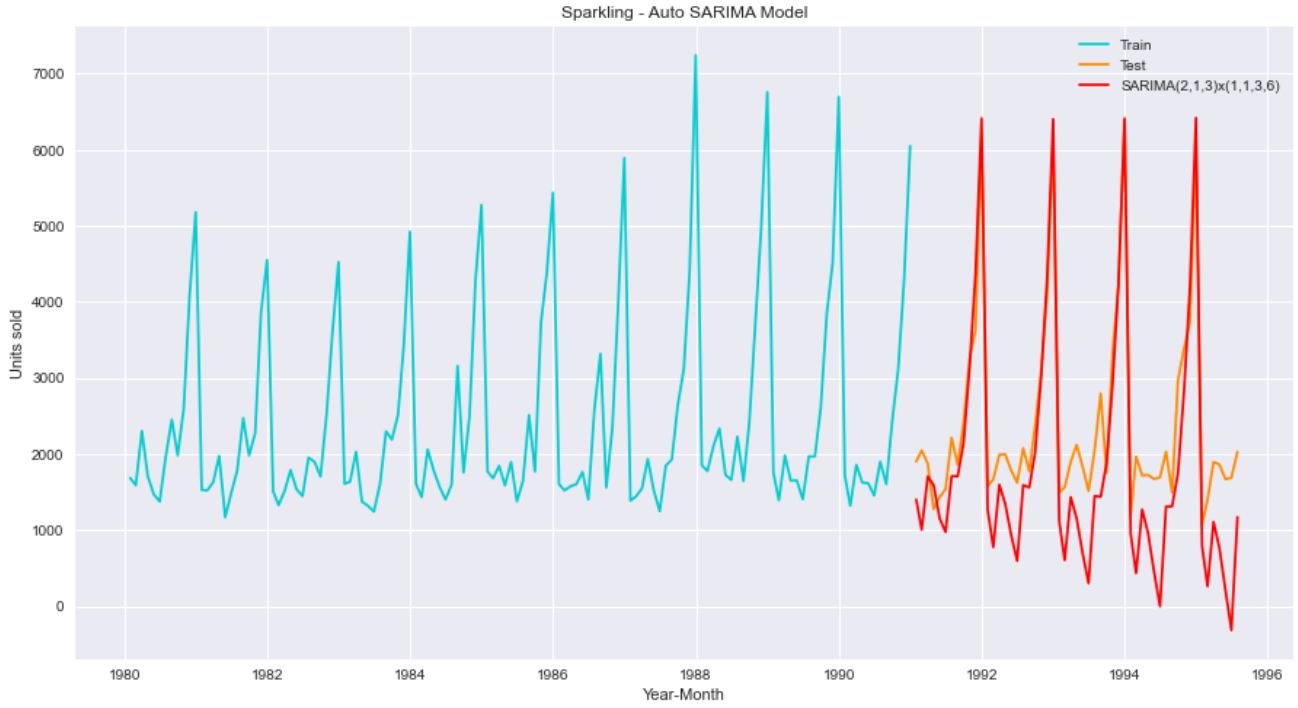


Fig:58

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.9
2 point TMA	813.400684	19.7
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.049607360581862936	1316.035487	45.47
DES Alpha 0.1,Beta 0.1	1779.42	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.935531	10.45
TES Alpha 0.15, Beta 0.00, Gamma 0.37	404.286809	13.93
Auto SARIMA(2,1,3)x(1,1,3,6)	789.939859	26.75

Table:64

Observations:

- The model built with original data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model.

- The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.

- The top three models with lowest AIC values are as given. As per the AIC criteria, the optimum values for the final SARIMA model selected is (2, 1, 3)x(1, 1, 3, 6).

- From the below model summary, it can be inferred that AR(1), AR(2), MA(1), MA(2), MA(3) terms have the highest absolute weightage.

- From the p-values it can be inferred that terms AR(1), AR(2), MA(1), MA(2), MA(3) and seasonal AR(1) are significant terms, as their values are below 0.05.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(1, 1, 3, 6)	Log Likelihood	-759.487
Date:	Tue, 01 Feb 2022	AIC	1538.974
Time:	19:47:00	BIC	1565.322
Sample:	0	HQIC	1549.646
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.7451	0.067	-26.143	0.000	-1.876	-1.614
ar.L2	-0.7880	0.072	-11.019	0.000	-0.928	-0.648
ma.L1	1.0751	0.300	3.586	0.000	0.488	1.663
ma.L2	-0.7722	0.122	-6.322	0.000	-1.012	-0.533
ma.L3	-0.8979	0.267	-3.364	0.001	-1.421	-0.375
ar.S.L6	-1.0251	0.008	-132.819	0.000	-1.040	-1.010
ma.S.L6	0.3660	0.275	1.333	0.182	-0.172	0.904
ma.S.L12	-0.7117	0.182	-3.904	0.000	-1.069	-0.354
ma.S.L18	0.1210	0.154	0.786	0.432	-0.181	0.423
sigma2	1.126e+05	4.7e-06	2.4e+10	0.000	1.13e+05	1.13e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	12.06
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	1.38	Skew:	0.41
Prob(H) (two-sided):	0.36	Kurtosis:	4.46

Table:65

Auto-SARIMA Sparkling Diagnostic Plot

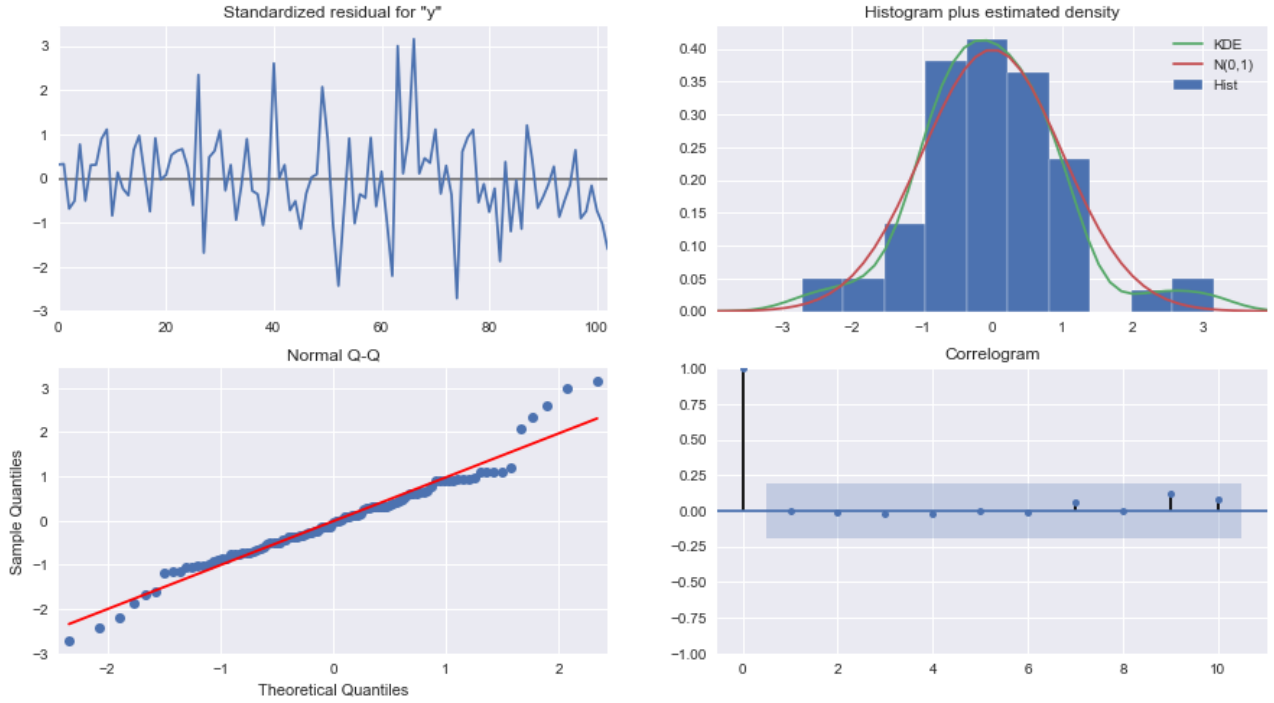


Fig:59

Observations:

- The Normal Q -Q plot also shows that the quantiles come from a normal distribution as the points form roughly a straight line.

- The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence.

Auto-SARIMA on Log-Sparkling

	param	seasonal	AIC
25	(0, 1, 1)	(1, 0, 1, 12)	-284.472032
79	(1, 1, 1)	(1, 0, 1, 12)	-282.517330
43	(0, 1, 2)	(1, 0, 1, 12)	-281.567996
97	(1, 1, 2)	(1, 0, 1, 12)	-279.611701
133	(2, 1, 1)	(1, 0, 1, 12)	-278.288232

Table:66

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.9
2 point TMA	813.400684	19.7
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.049607360581862936	1316.035487	45.47
DES Alpha 0.1,Beta 0.1	1779.42	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.935531	10.45
TES Alpha 0.15, Beta 0.00, Gamma 0.37	404.286809	13.93
Auto SARIMA(2,1,3)x(1,1,3,6)	789.939859	26.75
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.799059	11.19

Table:67

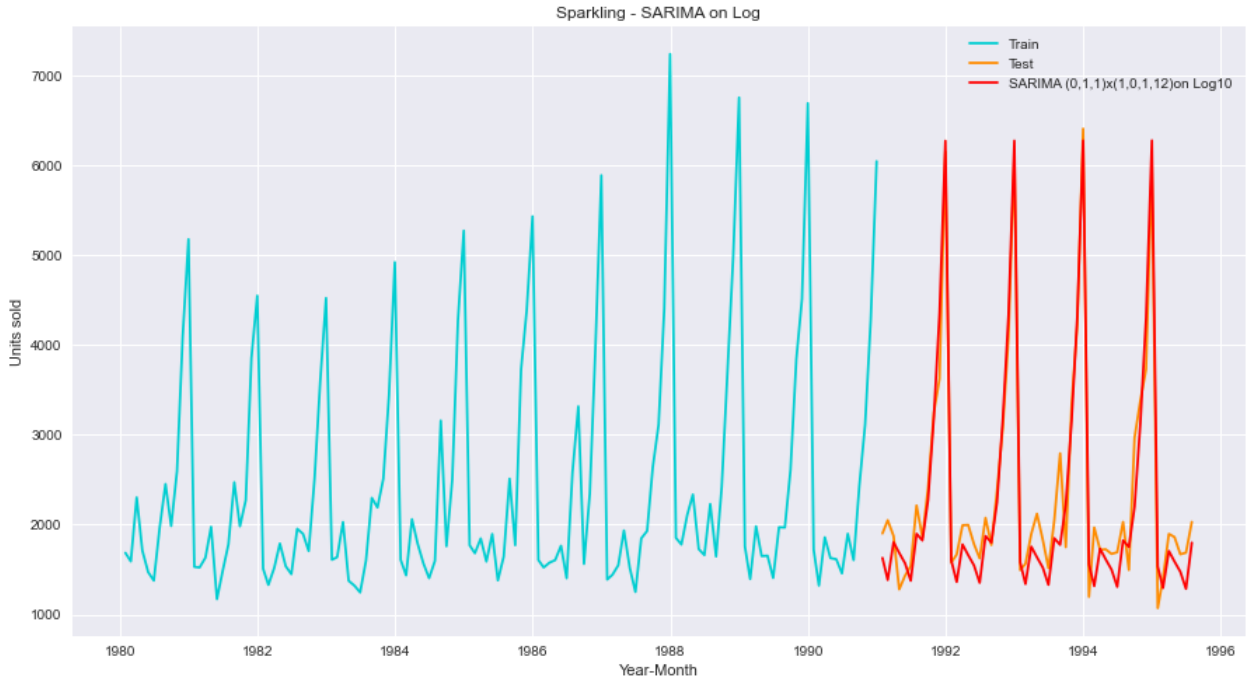


Fig:60

Observations:

- Two iterations of automated SARIMA models were attempted in this exercise, one with original data and another with log transformation of the data, as the seasonality got apparent multiplicity.

- To handle multiplicity of seasonality, the data was log transformed to make it additive.

- The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.

- The top three models with lowest AIC values are as given here and the final selected one is (0, 1, 1)x(1, 0, 1, 12).

- The model built with log transformed data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model.

- From the below model summary, it can be inferred that all AR and MA terms have the highest weightage, also from the p-values it can be inferred that these terms are significant as the values are below .05.

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(0, 1, 1)x(1, 0, 1, 12)	Log Likelihood	146.236
Date:	Tue, 01 Feb 2022	AIC	-284.472
Time:	20:00:59	BIC	-273.423
Sample:	01-31-1980	HQIC	-279.986
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8966	0.045	-19.863	0.000	-0.985	-0.808
ar.S.L12	1.0112	0.020	49.871	0.000	0.971	1.051
ma.S.L12	-0.6489	0.075	-8.629	0.000	-0.796	-0.502
sigma2	0.0045	0.001	7.842	0.000	0.003	0.006

Ljung-Box (L1) (Q):	0.11	Jarque-Bera (JB):	5.26
Prob(Q):	0.74	Prob(JB):	0.07
Heteroskedasticity (H):	1.43	Skew:	-0.00
Prob(H) (two-sided):	0.27	Kurtosis:	4.04

Table:68

Auto-SARIMA on Log-Sparkling Diagnostic Plot

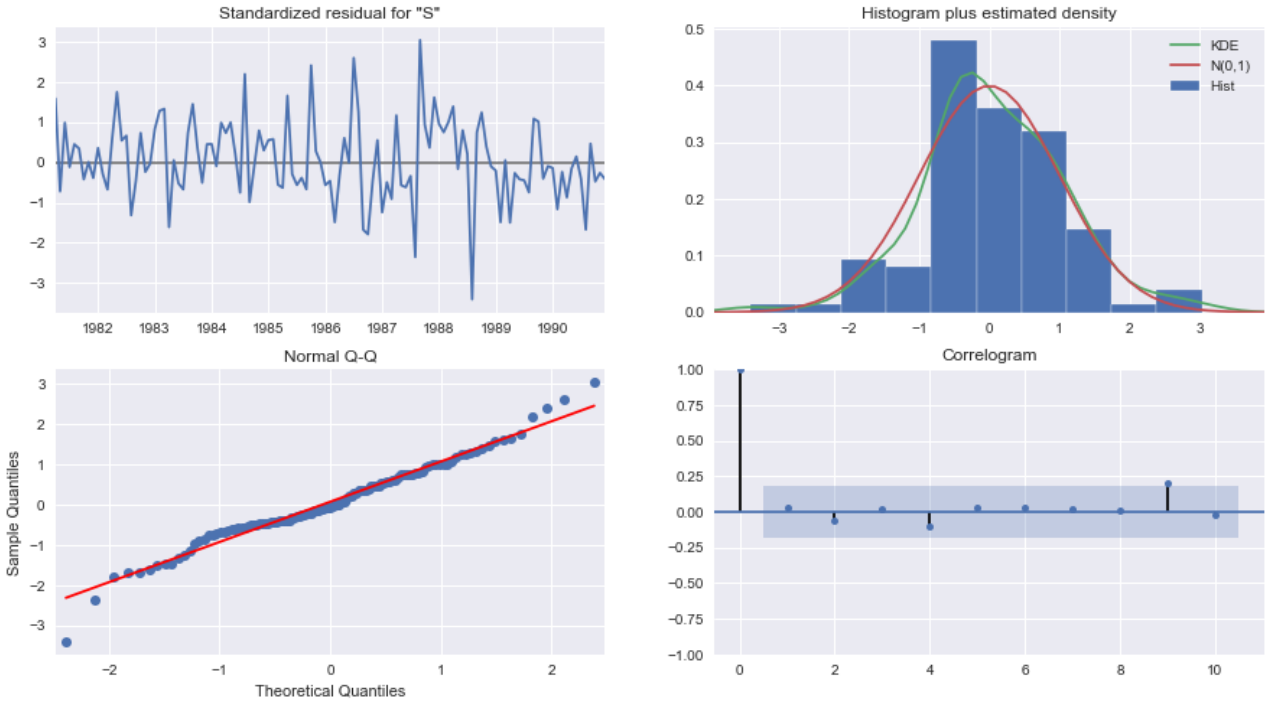


Fig:61

Observations:

- The Auto-SARIMA diagnostic plot also shows that the quantiles come from a normal distribution as the points form roughly a straight line and the correlogram shows the autocorrelation of the residuals as there are no significant lags above the confidence.

Auto-SARIMA Rose

	param	seasonal	AIC
187	(2, 1, 3)	(2, 1, 3, 6)	889.189499
251	(3, 1, 3)	(2, 1, 3, 6)	891.128678
255	(3, 1, 3)	(3, 1, 3, 6)	893.125602
183	(2, 1, 3)	(1, 1, 3, 6)	894.757072
63	(0, 1, 3)	(3, 1, 3, 6)	894.905688

Table:69

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.1
SimpleAverage	53.46035	94.93
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.16, Beta 0.16	15.706968	24.12
DES Alpha 0.10, Beta 0.10	37.056911	64.02
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.493835	13.68
TES Alpha 0.11, Beta 0.05, Gamma 0.00	20.156483	33.63
Auto SARIMA(2,1,3)x(2,1,3,6)	16.726881	27.6

Table:70

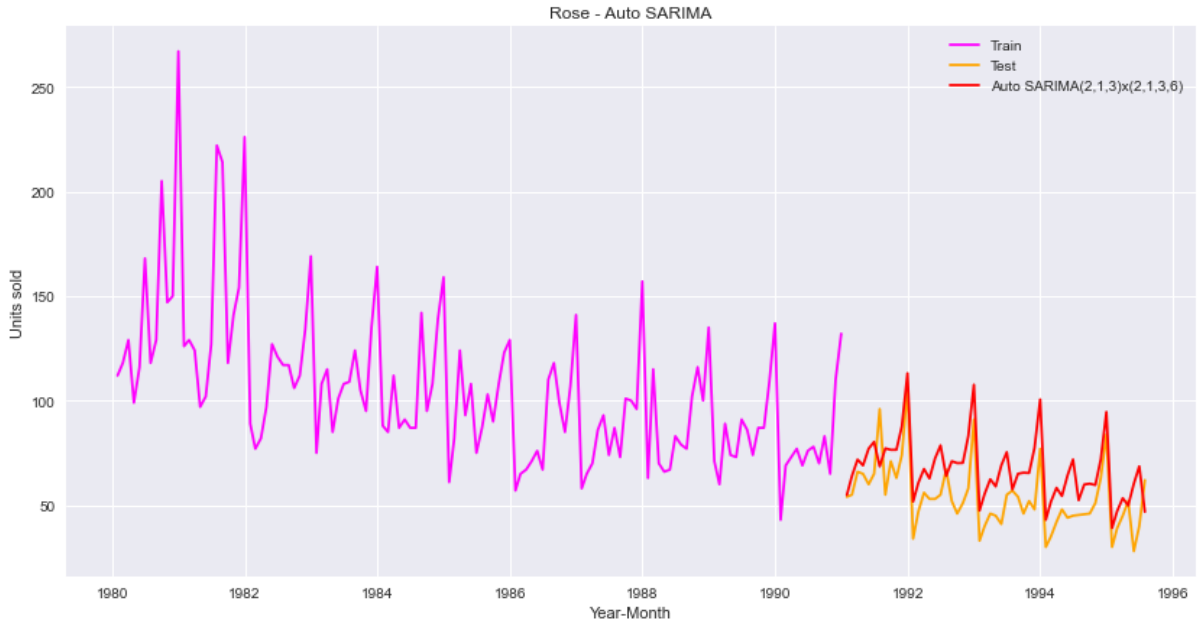


Fig:62

Observations:

- As the Rose series of data contain a seasonality component we will be building the SARIMA model, rather than ARIMA.

- Two iterations of automated SARIMA models were attempted in this exercise, one with original data and another with log transformation of the data, as the seasonality got apparent multiplicity.

- The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.

- The top three models with lowest AIC values are as given here and the final selected one is (2, 1, 3)x(2, 1, 3, 6).

- From the below model summary, it can be inferred that AR(1), AR(2), seasonal AR(1), seasonal AR(2) terms has the highest weightage, also from the p-values it can be inferred that these terms are significant as the values are below .05.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 1, 3, 6)	Log Likelihood	-433.595
Date:	Wed, 02 Feb 2022	AIC	889.189
Time:	21:30:21	BIC	918.172
Sample:	0	HQIC	900.928
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5746	0.023	25.057	0.000	0.530	0.620
ar.L2	-0.9163	0.021	-43.652	0.000	-0.957	-0.875
ma.L1	-1.4571	1081.429	-0.001	0.999	-2121.019	2118.105
ma.L2	1.5182	4419.040	0.000	1.000	-8659.640	8662.677
ma.L3	-0.8408	2951.088	-0.000	1.000	-5784.867	5783.186
ar.S.L6	-0.4344	0.106	-4.089	0.000	-0.643	-0.226
ar.S.L12	0.4838	0.102	4.743	0.000	0.284	0.684
ma.S.L6	-1.6631	15.322	-0.109	0.914	-31.694	28.368
ma.S.L12	-1.0697	40.603	-0.026	0.979	-80.651	78.511
ma.S.L18	1.5899	24.199	0.066	0.948	-45.840	49.020
sigma2	68.6080	2.41e+05	0.000	1.000	-4.72e+05	4.72e+05

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	5.80
Prob(Q):	0.86	Prob(JB):	0.05
Heteroskedasticity (H):	0.45	Skew:	0.56
Prob(H) (two-sided):	0.02	Kurtosis:	3.29

Table:71

Auto-SARIMA Rose Diagnostic Plot

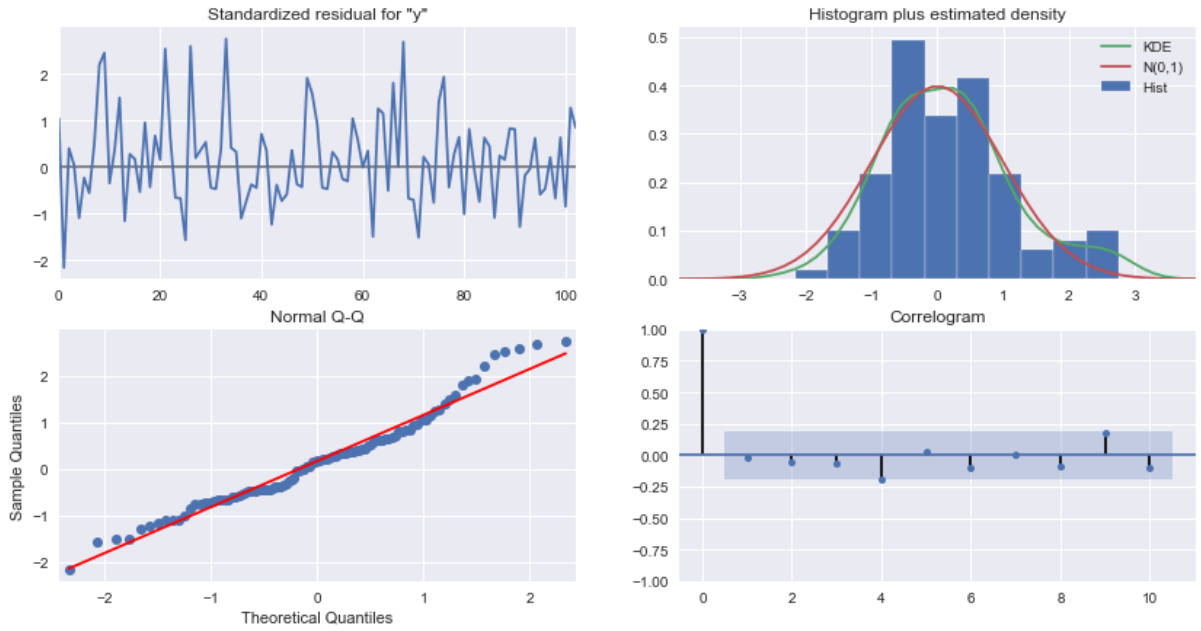


Fig:63

Observations:

- The diagnostic plot of the model was derived, and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.

- The Normal Q -Q plot also shows that the quantiles come from a normal distribution as the points form roughly a straight line.

- The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index.

Auto-SARIMA on Log-Rose

	param	seasonal	AIC
115	(1, 0, 0)	(1, 0, 1, 12)	-257.620750
7	(0, 0, 0)	(1, 0, 1, 12)	-256.170281
133	(1, 0, 1)	(1, 0, 1, 12)	-255.482084
25	(0, 0, 1)	(1, 0, 1, 12)	-254.978845
223	(2, 0, 0)	(1, 0, 1, 12)	-253.620649

Table:72

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.1
SimpleAverage	53.46035	94.93
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.16, Beta 0.16	15.706968	24.12
DES Alpha 0.10, Beta 0.10	37.056911	64.02
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.493835	13.68
TES Alpha 0.11, Beta 0.05, Gamma 0.00	20.156483	33.63
Auto SARIMA(2,1,3)x(2,1,3,6)	16.726881	27.6
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.590947	21.92

Table:73

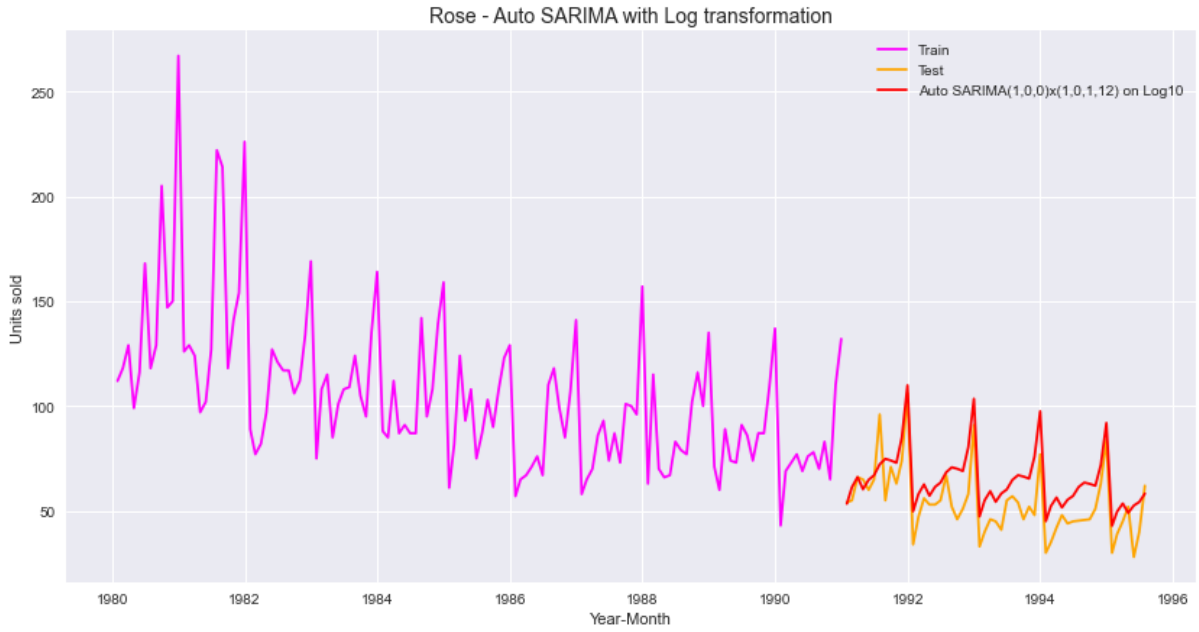


Fig:64

Observations:

- Two iterations of automated SARIMA models were attempted in this exercise, one with original data and another with log transformation of the data, as the seasonality got apparent multiplicity.

- The model built with log transformed data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model.

- To handle multiplicity of seasonality, the data was log transformed to make it additive.

- The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.

- The top three models with lowest AIC values are as given here and the final selected one is (1, 0, 0)x(1, 0, 1, 12).

- The model built with log transformed data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model.

- From the below model summary, it can be inferred that AR and seasonal AR terms have the highest weightage, also from the p-values it can be inferred that these terms are significant as the values are below .05.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(1, 0, 0)x(1, 0, [1], 12)	Log Likelihood	132.810
Date:	Wed, 02 Feb 2022	AIC	-257.621
Time:	21:52:30	BIC	-246.504
Sample:	01-31-1980	HQIC	-253.107
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1689	0.078	2.179	0.029	0.017	0.321
ar.S.L12	0.9872	0.001	751.593	0.000	0.985	0.990
ma.S.L12	-0.9408	0.349	-2.697	0.007	-1.624	-0.257
sigma2	0.0052	0.002	2.899	0.004	0.002	0.009

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	4.00
Prob(Q):	0.89	Prob(JB):	0.14
Heteroskedasticity (H):	0.86	Skew:	0.40
Prob(H) (two-sided):	0.64	Kurtosis:	3.40

Table:74

Auto-SARIMA on Log-Sparkling Diagnostic Plot

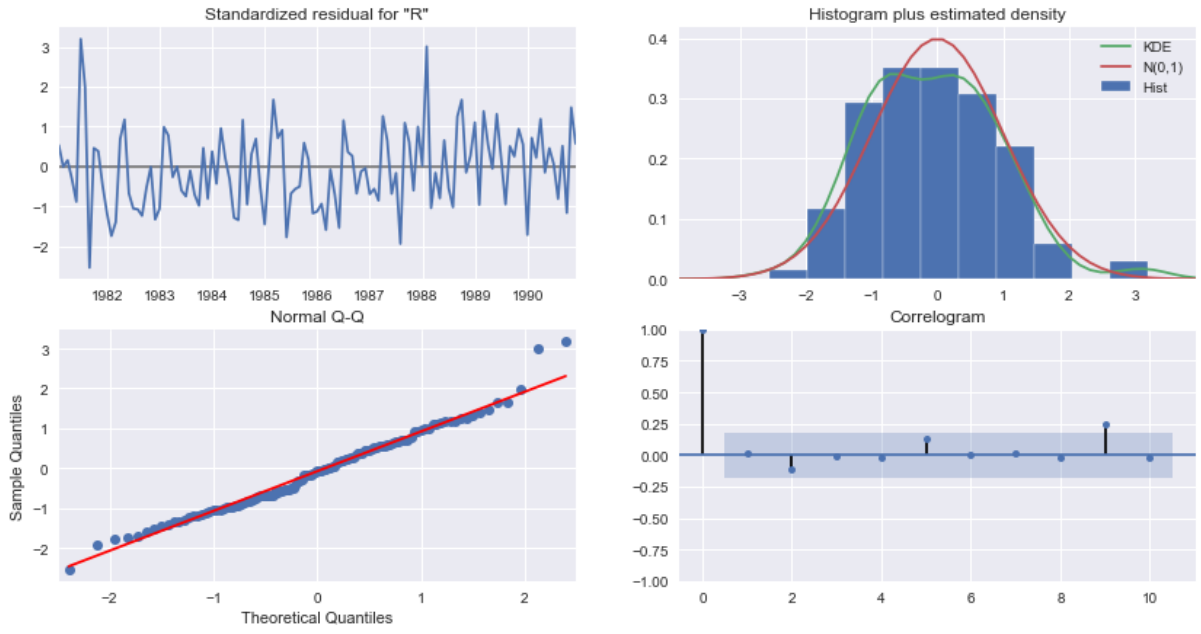


Fig:65

Observations:

- The diagnostic plot of the model was derived, and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.

- The Normal Q -Q plot also shows that the quantiles come from a normal distribution as the points form roughly a straight line.

- The correlogram shows the autocorrelation of the residuals and there is 1 point significant above the confidence index.

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.

Manual-SARIMA Sparkling

Fig:66

Observations:

- From the plots above an apparent slight trend still exists after differencing the seasonal order of 12. With a further differentiation of order one, no trend is present.

ACF & PACF Plots

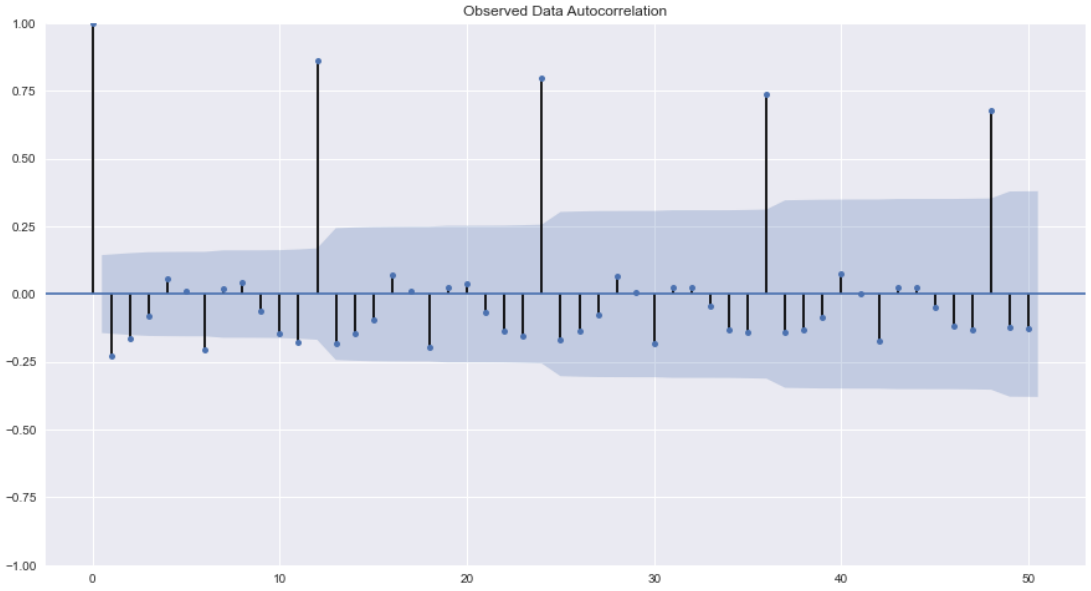


Fig:67

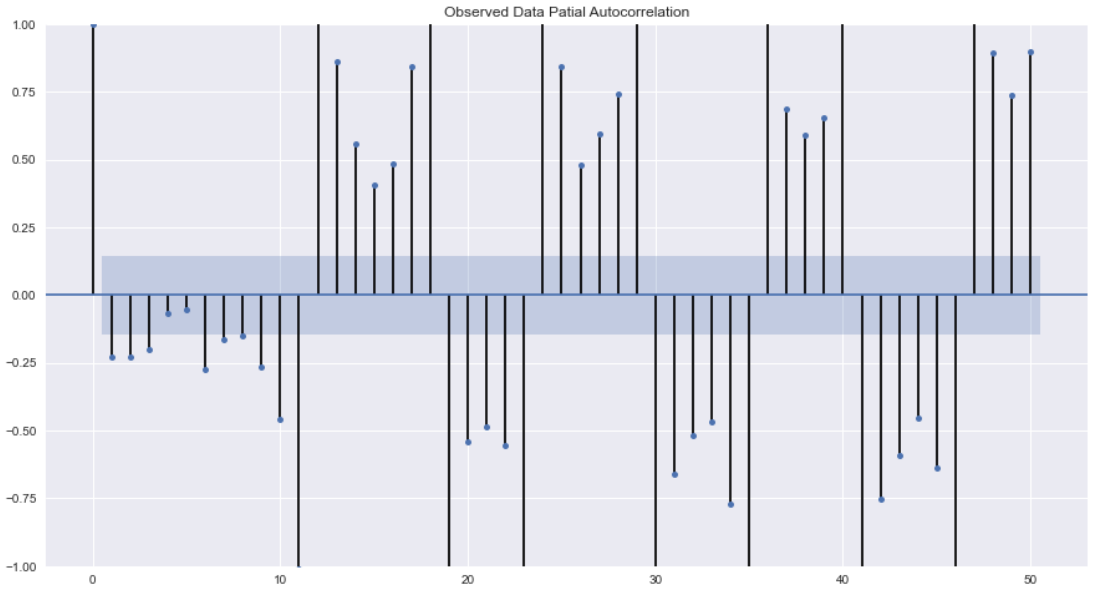


Fig:68

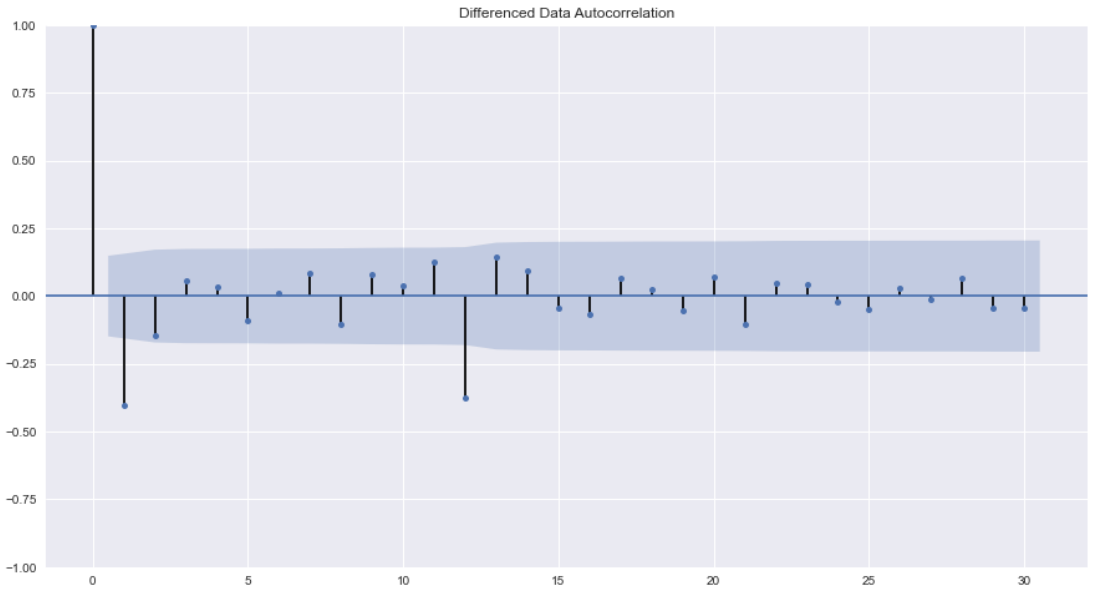

Differenced Data Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30

Fig:69

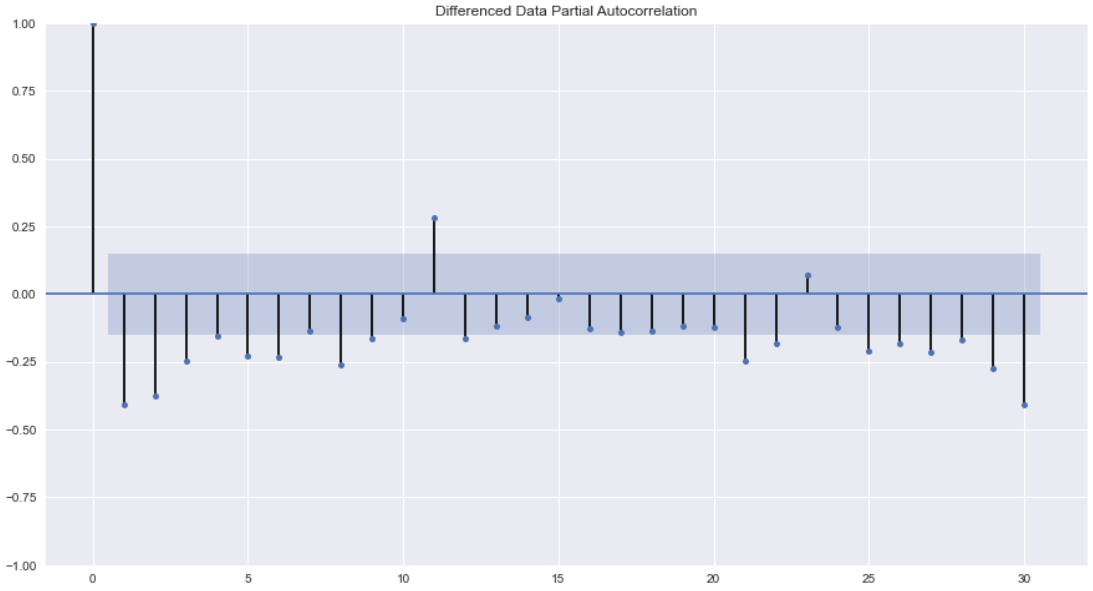

Differenced Data Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30

Fig:70

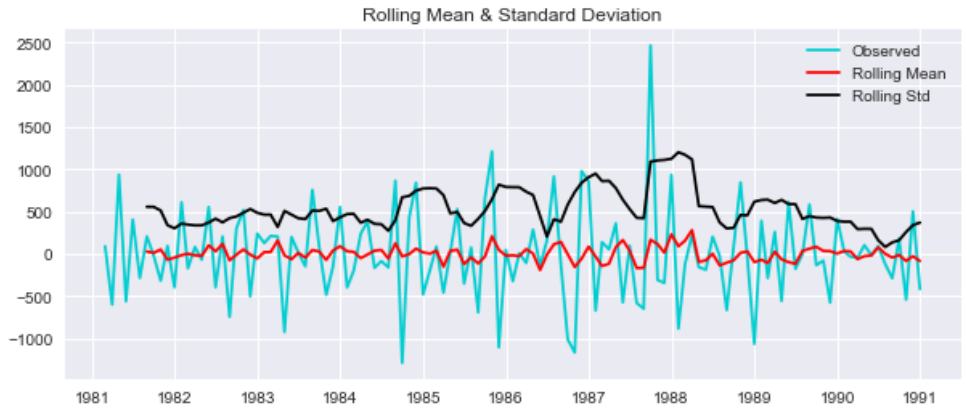


Fig:71

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.342905
p-value                          0.013066
#Lags Used                      10.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
dtype: float64
```

Table:75

Observations:

- From the ACF plot of the observed/ train data, it can be inferred that at seasonal intervals of 12, the plot is not quickly tapering off. So a seasonal differencing of 12 has to be taken.
- Here we have taken alpha = 0.05 and the seasonal period as 12.
- From the PACF plot it can be seen that till 3rd lag is significant before cut-off, so AR term ‘p = 3’ is chosen. At seasonal lag of 12, it almost cuts off, so seasonal AR ‘P = 1’.
- From the ACF plot it can be seen that lag 1 is significant before it cuts off, so MA term ‘q = 1’ is selected and at seasonal lag of 12, a significant lag is apparent, so keep seasonal MA term ‘Q = 1’ initially.
- An ADF test needs to be done to check the stationarity after the above differencing. With a p-value below alpha 0.05 and test statistics below critical values, it can be confirmed that the data is stationary.

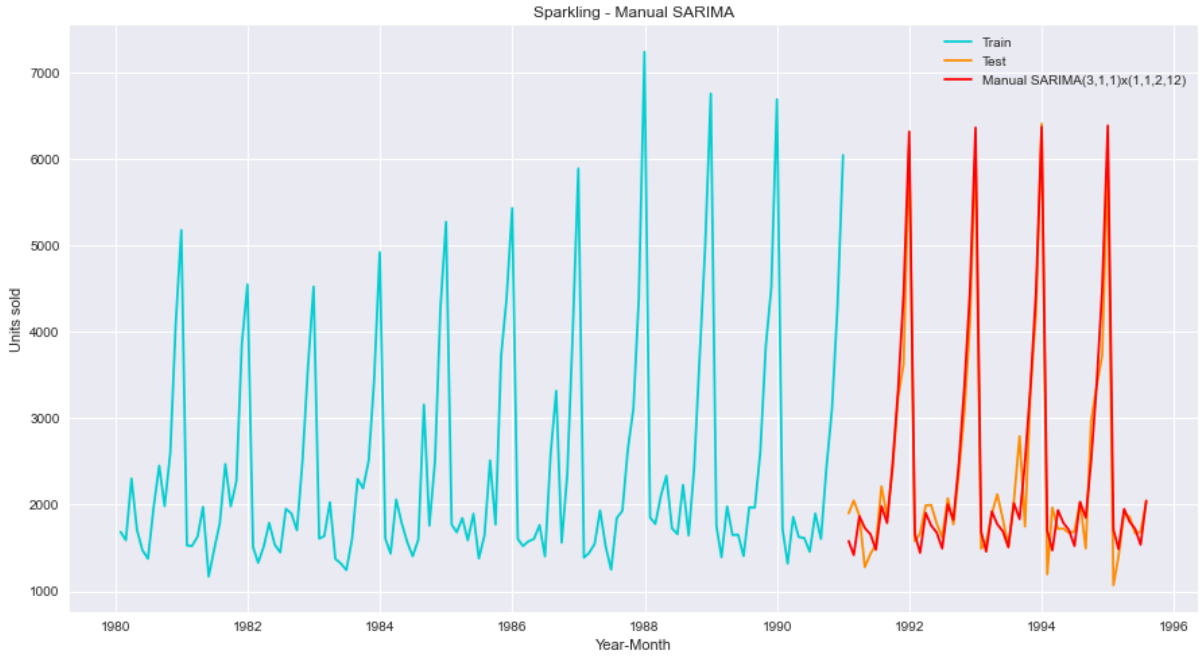


Fig:72

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(1, 1, [1, 2], 12)	Log Likelihood	-693.697
Date:	Tue, 01 Feb 2022	AIC	1403.394
Time:	20:17:05	BIC	1423.654
Sample:	0	HQIC	1411.574
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2229	0.130	1.713	0.087	-0.032	0.478
ar.L2	-0.0798	0.131	-0.607	0.544	-0.337	0.178
ar.L3	0.0921	0.122	0.756	0.450	-0.147	0.331
ma.L1	-1.0241	0.094	-10.925	0.000	-1.208	-0.840
ar.S.L12	-0.1992	0.866	-0.230	0.818	-1.897	1.499
ma.S.L12	-0.2109	0.881	-0.239	0.811	-1.938	1.516
ma.S.L24	-0.1299	0.381	-0.341	0.733	-0.877	0.617
sigma2	1.654e+05	2.62e+04	6.302	0.000	1.14e+05	2.17e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	19.66
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	0.81	Skew:	0.69
Prob(H) (two-sided):	0.56	Kurtosis:	4.78

Table:76

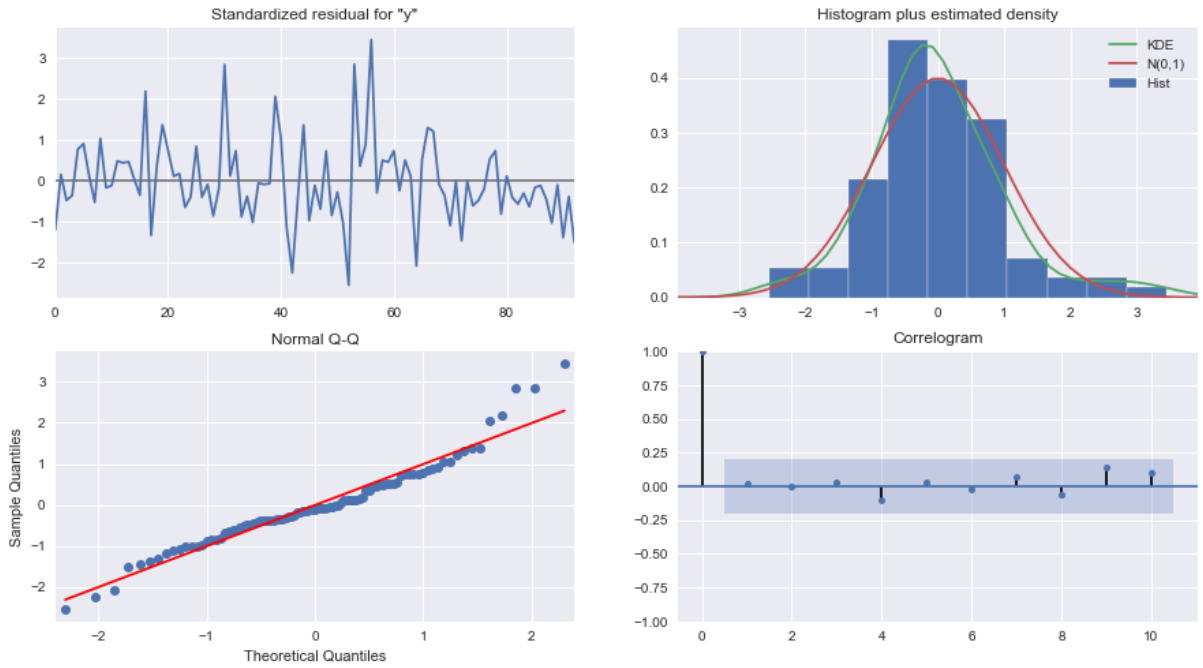


Fig:73

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.9
2 point TMA	813.400684	19.7
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.049607360581862936	1316.035487	45.47
DES Alpha 0.1,Beta 0.1	1779.42	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.935531	10.45
TES Alpha 0.15, Beta 0.00, Gamma 0.37	404.286809	13.93
Auto SARIMA(3,1,3)x(3,1,0,12)	789.939859	26.75
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.799059	11.19
Manual SARIMA(3,1,1)x(1,1,2,12)	324.10651	9.48

Table:77

Observations:

- The seasonal MA term ‘Q’ was later optimized to 2, by validating model performance, as the data might be under-differenced.
- The final selected term for the SARIMA model is (3, 1, 1)x(0, 1, 2, 12).
- The diagnostic plot for the model is as below, which clearly shows a normal distribution of residuals, where more values are around zero.
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the points form roughly a straight line.
- The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index.
- The model summary indicates that only the MA(1) term used in the model is significant in terms of p-values.
- From the multiple iterations of SARIMA models, above is the comparison of the models in terms of their accuracy attributes of RMSE and MAPE.

Manual-SARIMA Rose

Log transformation of the data for Rose data is done to handle the multiplicity of seasonality.

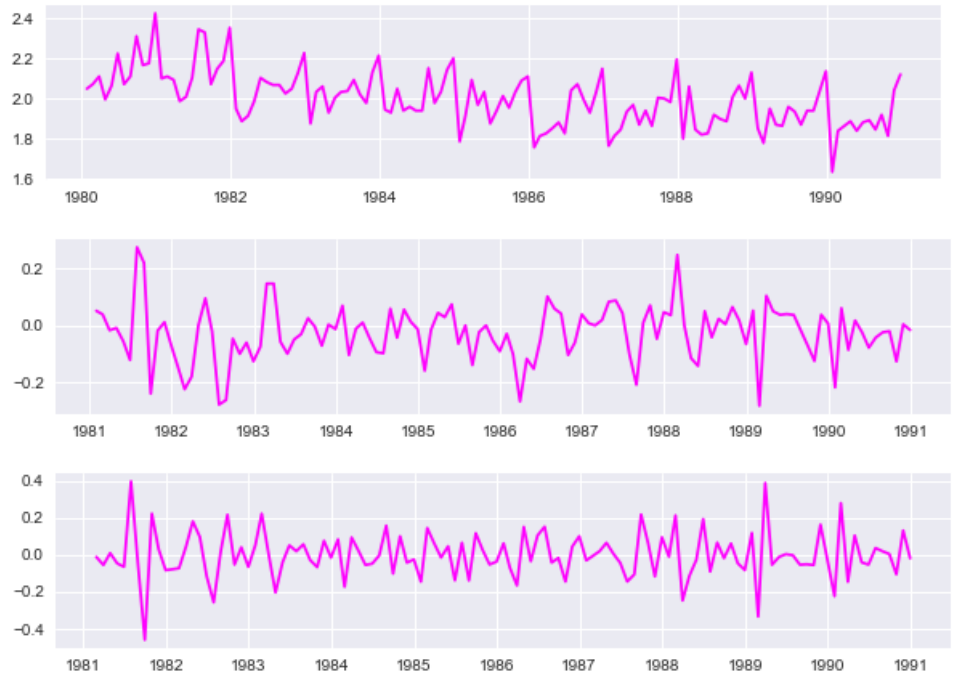


Fig:74

Observations:

- From the plots below it can be seen that a slight trend is still existing after differencing the seasonal order of 12. With a further differentiation of order one, no trend is present.

ACF & PACF Plots

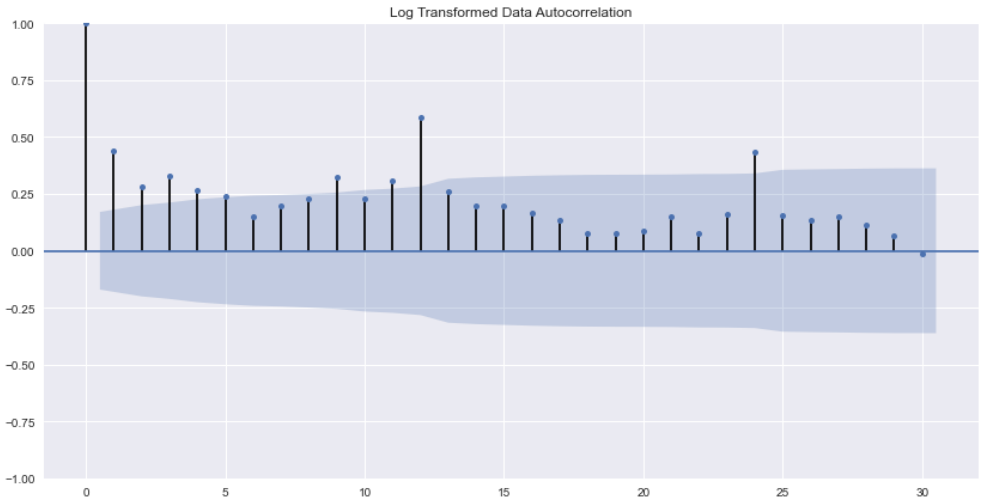


Fig:75

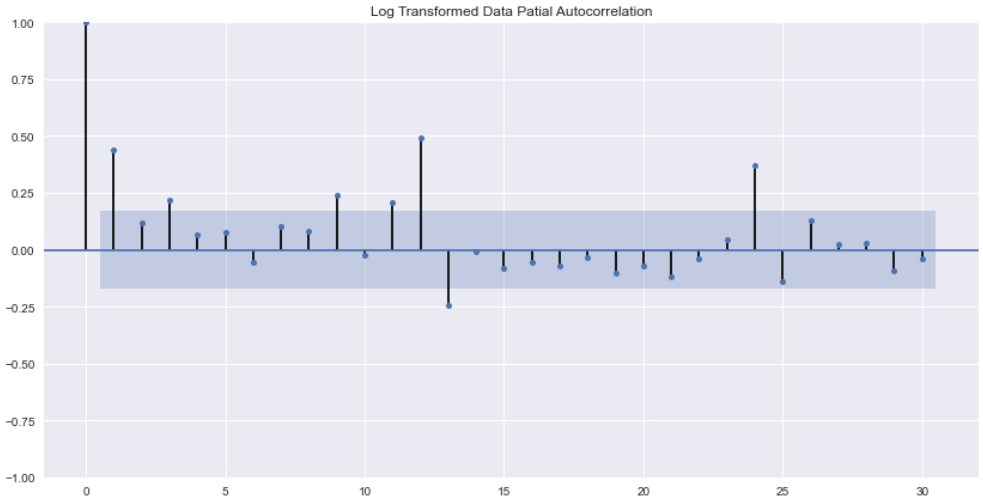


Fig:76

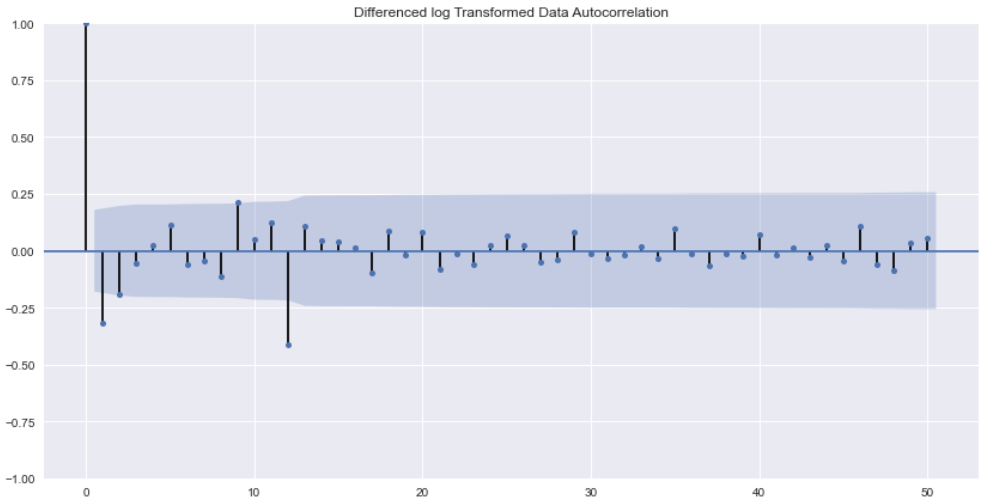


Fig:77

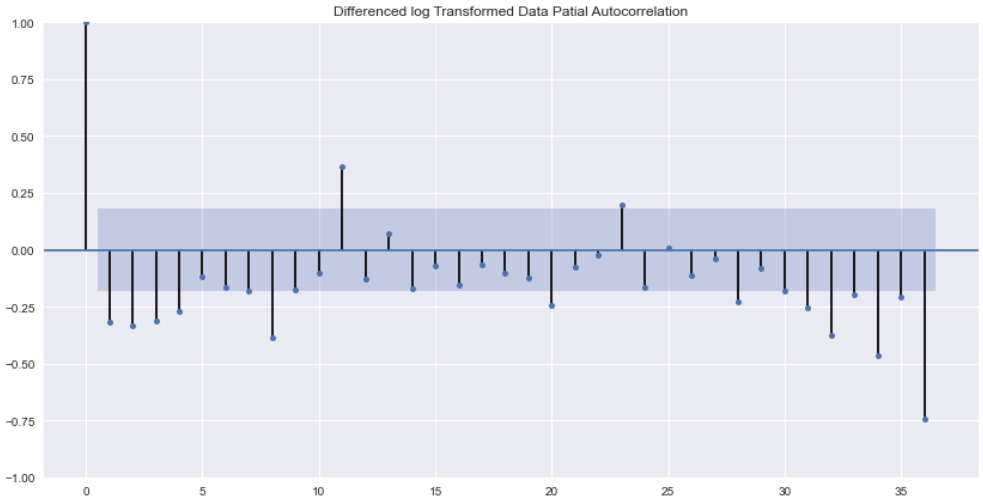


Fig:78

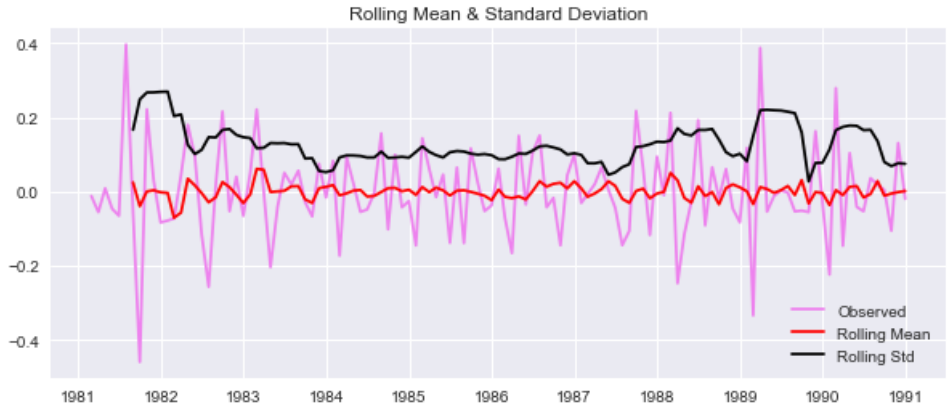


Fig:79

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.910109
p-value                          0.001962
#Lags Used                      11.000000
Number of Observations Used    107.000000
Critical Value (1%)             -3.492996
Critical Value (5%)             -2.888955
Critical Value (10%)            -2.581393
dtype: float64
```

Table:78

Observations:

- From the ACF plot of the log transformed data, it can be seen that at seasonal intervals of 12, the plot is not quickly tapering off. So we need to take a seasonal difference of 12.

- ACF and PACF plots of the seasonal-differenced + one order-differenced data are created to find the values for (p,d,q)x(P,D,Q).

- Here we have taken alpha = 0.05 and the seasonal period as 12.

- From the PACF plot it can be seen that until lag 4 is significant before cut-off, so the AR term ‘p = 4’ is chosen. At seasonal lag of 12, it cuts off, so keep seasonal AR ‘P = 0’.

- From the ACF plot, lags 1 and 2 are significant before it cuts off, so let's keep MA term ‘q = 1’ and at seasonal lag of 12, a significant lag is apparent, so let's keep ‘Q = 1’.

- Have done an ADF test to check the stationarity after the above differencing. With a p value below alpha 0.05 and test statistics below critical values, it can be confirmed that the data is stationary.

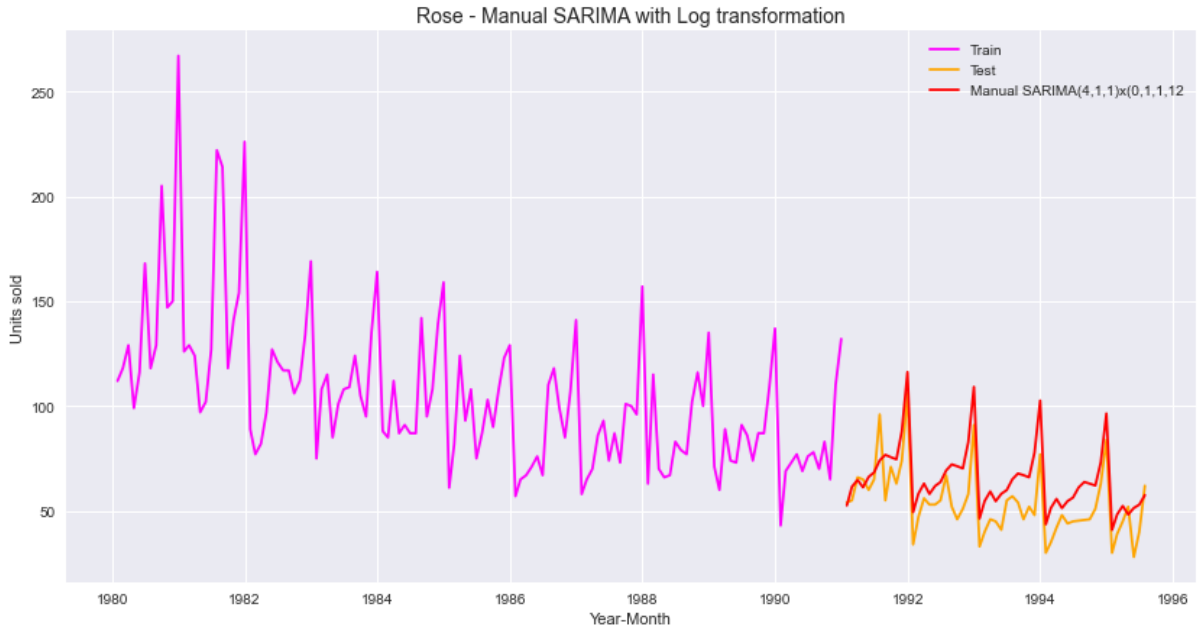


Fig:80

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(4, 1, 1)x(0, 1, 1, 12)	Log Likelihood	128.764
Date:	Wed, 02 Feb 2022	AIC	-243.528
Time:	22:17:58	BIC	-224.950
Sample:	01-31-1980 - 12-31-1990	HQIC	-236.000
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0017	0.118	-0.014	0.989	-0.232	0.229
ar.L2	-0.1553	0.126	-1.230	0.219	-0.403	0.092
ar.L3	-0.1600	0.113	-1.422	0.155	-0.381	0.061
ar.L4	-0.1504	0.121	-1.242	0.214	-0.388	0.087
ma.L1	-0.8434	0.074	-11.406	0.000	-0.988	-0.698
ma.S.L12	-0.9944	4.036	-0.246	0.805	-8.906	6.917
sigma2	0.0041	0.016	0.252	0.801	-0.028	0.036

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	3.83
Prob(Q):	0.91	Prob(JB):	0.15
Heteroskedasticity (H):	1.60	Skew:	0.44
Prob(H) (two-sided):	0.17	Kurtosis:	3.34

Table:79

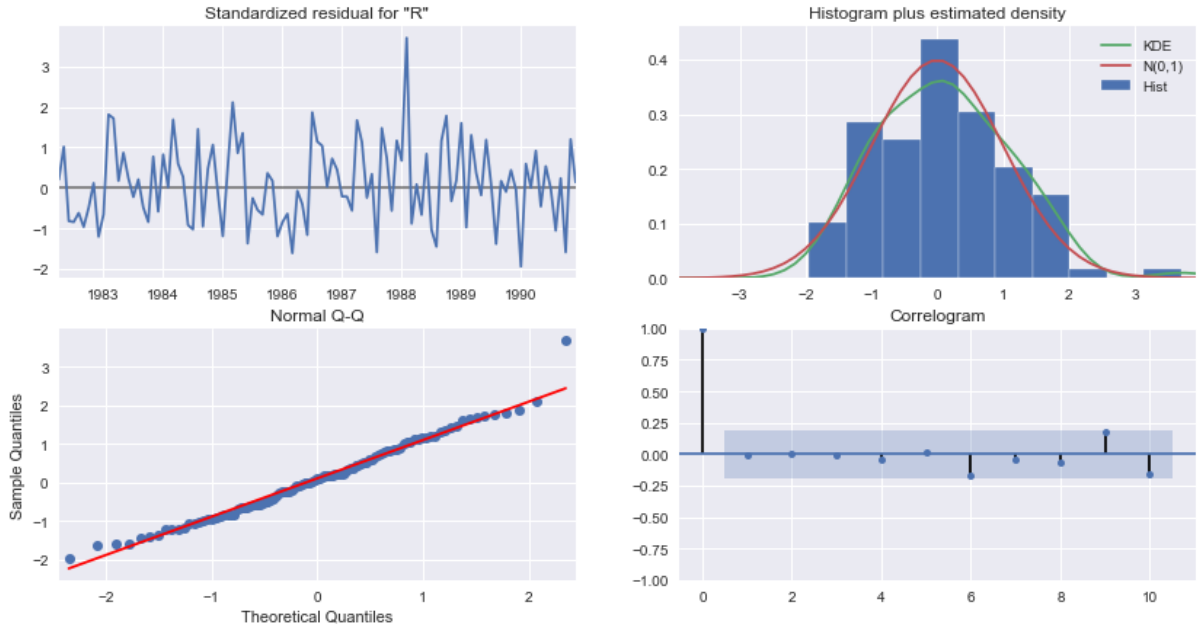


Fig:81

	Test RMSE	Test MAPE
RegressionOnTime	15.268885	22.82
NaiveModel	79.718559	145.1
SimpleAverage	53.46035	94.93
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.16, Beta 0.16	15.706968	24.12
DES Alpha 0.10, Beta 0.10	37.056911	64.02
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.493835	13.68
TES Alpha 0.11, Beta 0.05, Gamma 0.00	20.156483	33.63
Auto SARIMA(2,1,3)x(2,1,3,6)	16.726881	27.6
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.590947	21.92
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.177004	23.1

Table:80

Observations:

- The final selected term for the SARIMA model is (4, 1, 1)x(0, 1, 1, 12), as inferred from the ACF and PACF plots.

- The diagnostic plot for the model is as below, which clearly shows a normal distribution of residuals, where more values are around zero.

- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the points form roughly a straight line.

- The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index.

- The model summary indicates that MA(1) used in the model is significant in terms of p values.

- From the multiple iterations of SARIMA models, above is the comparison of the models in terms of their accuracy attributes of RMSE and MAPE.

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	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.935531	10.45
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106510	9.48
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.799059	11.19
TES Alpha 0.15, Beta 0.00, Gamma 0.37	404.286809	13.93
Auto SARIMA(3,1,3)x(3,1,0,12)	789.939859	26.75
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
6 point TMA	1283.927428	43.86
SES Alpha 0.049607360581862936	1316.035487	45.47
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23
NaiveModel	3864.279352	152.87

Table:81

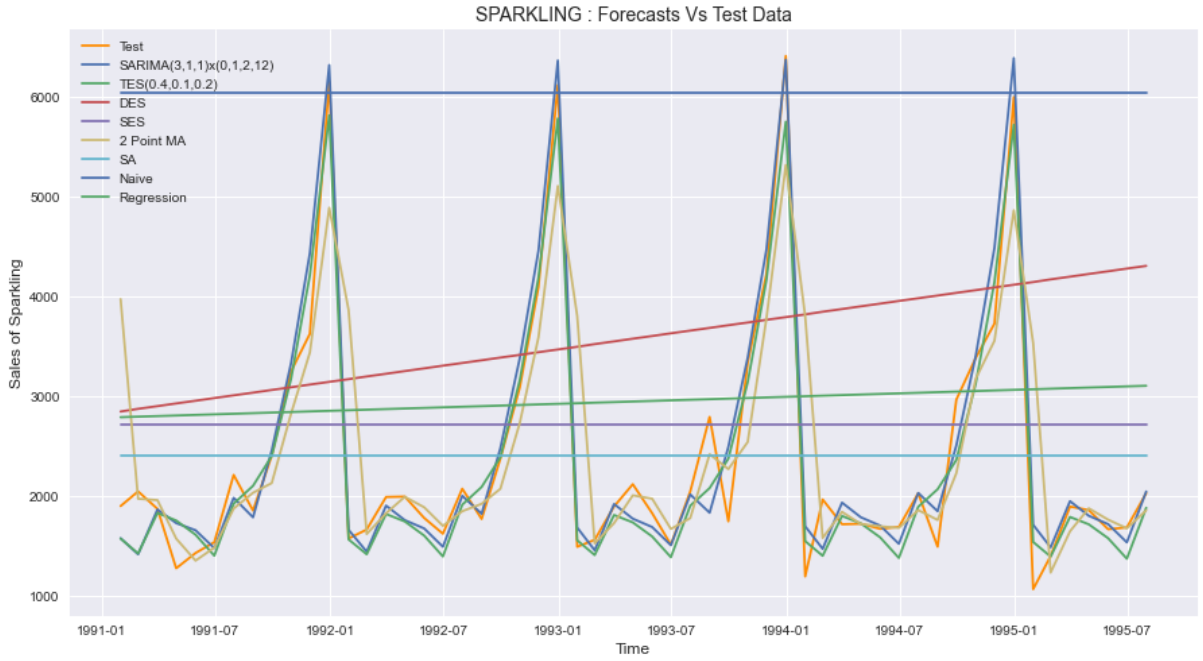


Fig:82

Observations:

- The overall comparison of all the time-series forecast models are listed above in accordance with increasing RMSE against test data or in the order of decreasing accuracy.
- Triple Exponential Smoothing is found to be the best model, followed by SARIMA.
- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data.
- The SARIMA and Triple Exponential Smoothing are found to be comparable in terms of performance and fitment with the test data.

Rose

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.493835	13.68
2 point TMA	11.529278	13.54
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.590947	21.92
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.177004	23.10
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
DES Alpha 0.16, Beta 0.16	15.706968	24.12
Auto SARIMA(2,1,3)x(2,1,3,6)	16.726881	27.60
TES Alpha 0.11, Beta 0.05, Gamma 0.00	20.156483	33.63
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.10, Beta 0.10	37.056911	64.02
SimpleAverage	53.460350	94.93
NaiveModel	79.718559	145.10

Table:82

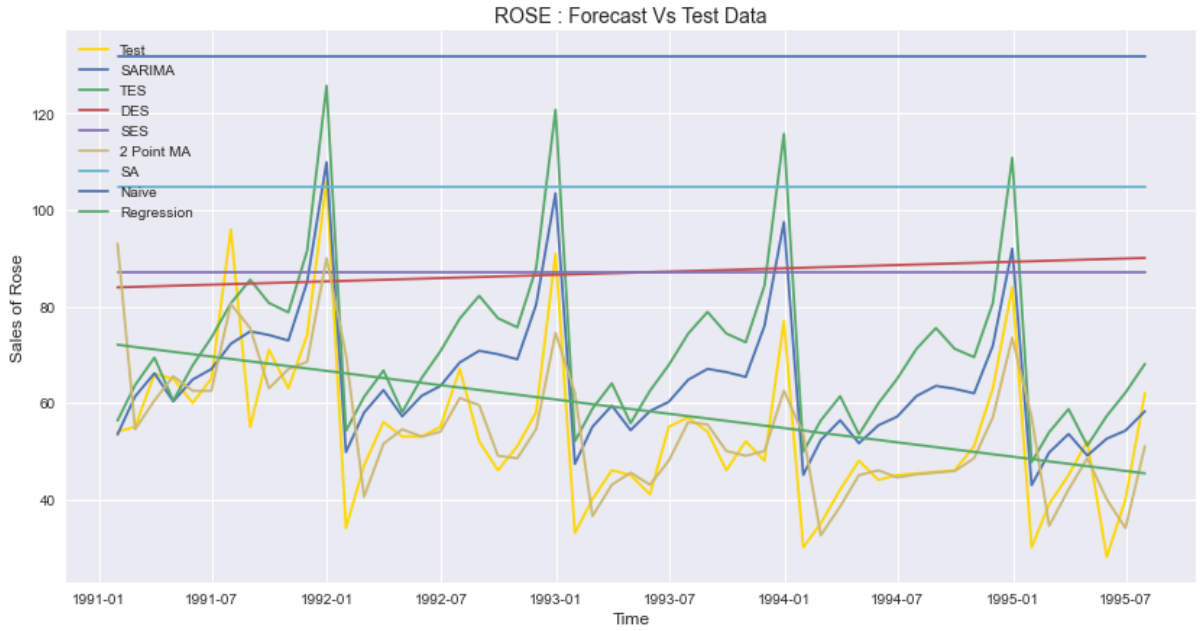


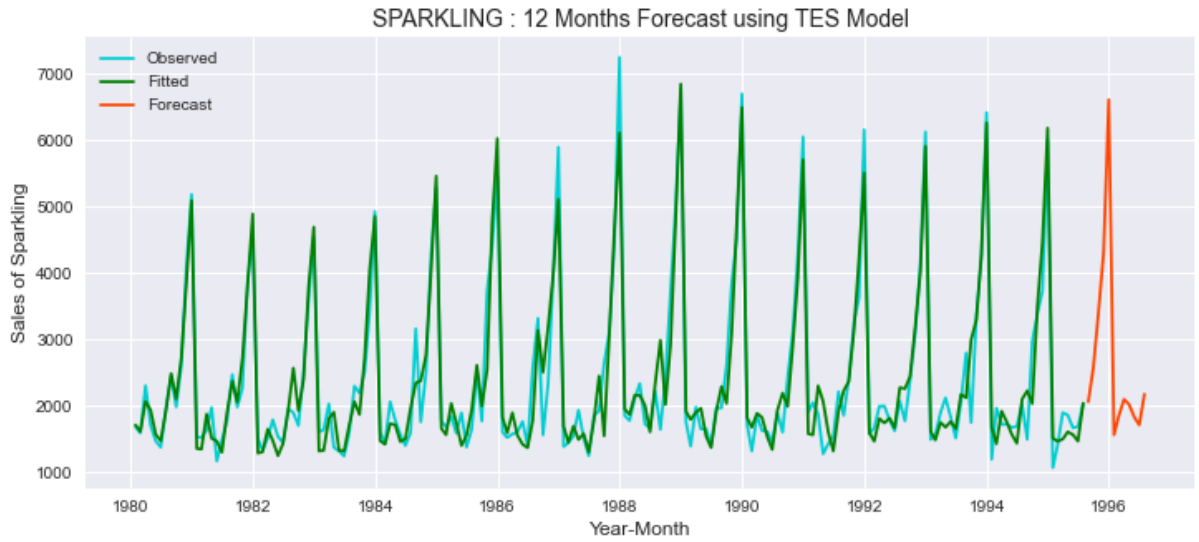
Fig:83

Observations:

- The overall comparison of all the time-series forecast models are listed below with increasing RMSE against test data or in the order of decreasing accuracy.
- Triple Exponential Smoothing is found to be the best model, followed by a 2 point Moving Average.
- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data.
- The 2 point trailing moving average is found to be having the best fitment against the test data, though with a lag of 2 and falling short at times.
- Both SARIMA and TES forecasts are a bit higher than the actuals at any given point in time.

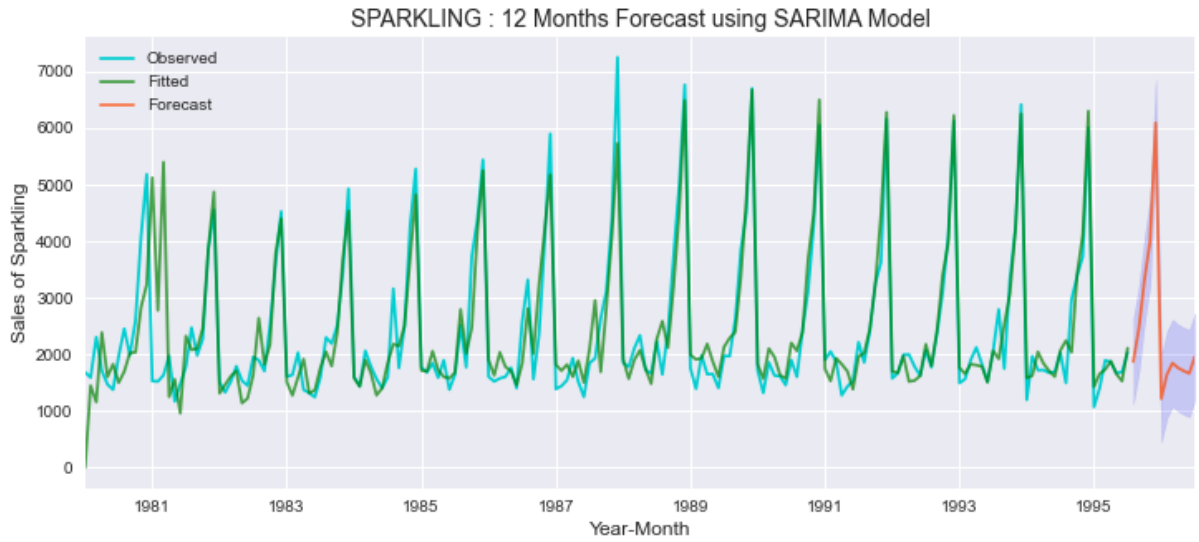
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Sparkling



TES forecast on the Sparkling Full Data: RMSE is 376.775 and MAPE is 11.29

Fig:84



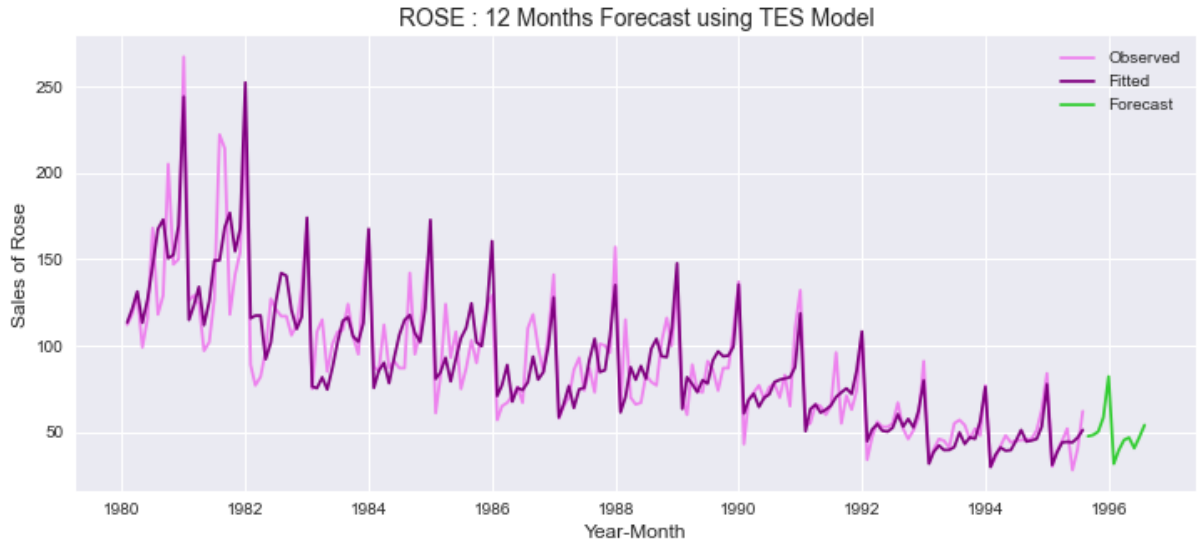

```
For SARIMA forecast on the Sparkling Full Data:  RMSE is 591.263 and MAPE is 14.86
```

Fig:85

Observations:

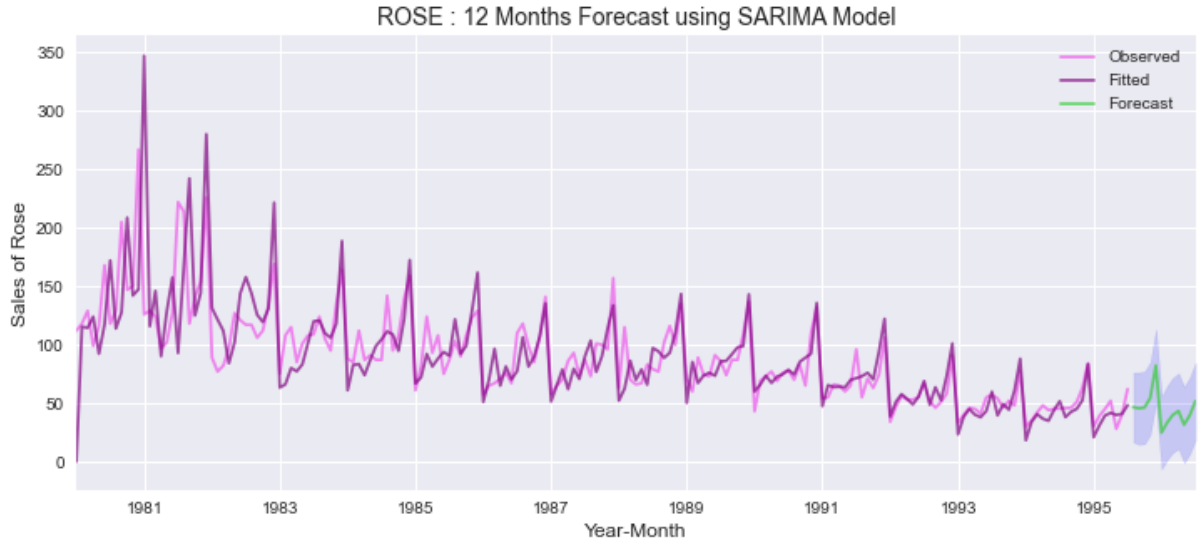
- Based on the overall model evaluation and comparison, Triple Exponential Smoothing (Holt Winter's) and SARIMA were selected for final prediction 12 months into the future.
- TES model alpha: 0.4, beta: 0.1 and gamma: 0.2 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data.
- The model predicts an upward trend and continuation of the seasonal surge in sales in the upcoming 12 months. According to the model, the seasonal sale will be more than that of the previous year.
- The 12-month prediction of the TES model is as above.
- The SARIMA model is built with parameters (3, 1, 3)x(1, 1, 2, 12), and is found to be the most optimal SARIMA model.
- The SARIMA model reflects the trend and seasonality of the series continuing into the future year as well. The seasonal altitude predicted us to be more conservative than the TES model.
- The SARIMA model is seen to have better fitment with the most recent observed data and shows high variations in the farthest periods of observations, which explains the high RMSE and MAPE values.
- The RMSE and MAPE values of the two models are mentioned along with the plots.

Rose



TES forecast on the Rose Full Data: RMSE is 17.404 and MAPE is 13.87

Fig:86



For SARIMA forecast on the Rose Full Data: RMSE is 30.676 and MAPE is 19.40

Fig:87

Observations:

- Based on the overall model evaluation and comparison, Triple Exponential Smoothing (Holt Winter's) and SARIMA were selected for final prediction 12 months into the future.

- TES model alpha: 0.1, beta: 0.2 and gamma: 0.2 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data.

- The model predicts continuation of the trend in sales and seasonality in year end sales. The prediction shows a stabilization of the downward trend, as the sales will be almost the same as the previous observed year.

- The 12 month prediction of the TES model is as above.

- The SARIMA model is built with parameters (4, 1, 1)x(0, 1, 1, 12), and is found to be the most optimal SARIMA model for the complete time-series.

- The SARIMA model has also reflected the trend and seasonality of the series continuing into the future year as well. The SARIMA model is seen to have better fitment with the most recent observed data and shows high variations in the farthest periods of observations, which explains the high RMSE and MAPE values.

- The RMSE and MAPE values of the two models are mentioned above along with plots.

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

Sparkling

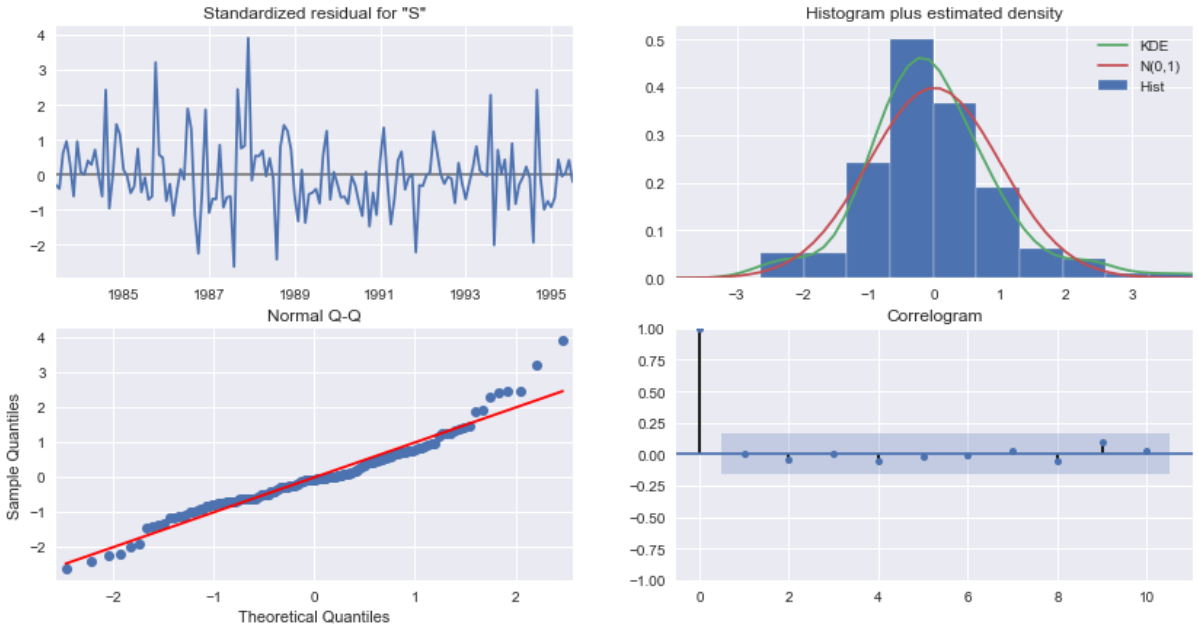


Fig:88

Model:	SARIMAX(3, 1, 3)x(1, 1, [1, 2], 12)	Log Likelihood	-1078.437
Date:	Tue, 01 Feb 2022	AIC	2176.875
Time:	20:31:20	BIC	2206.711
Sample:	01-31-1980 - 07-31-1995	HQIC	2188.998
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4229	0.086	-4.917	0.000	-0.591	-0.254
ar.L2	-0.9093	0.053	-17.302	0.000	-1.012	-0.806
ar.L3	0.1425	0.087	1.640	0.101	-0.028	0.313
ma.L1	-0.4114	0.078	-5.283	0.000	-0.564	-0.259
ma.L2	0.4622	0.083	5.583	0.000	0.300	0.624
ma.L3	-0.9674	0.104	-9.329	0.000	-1.171	-0.764
ar.S.L12	-0.0692	0.708	-0.098	0.922	-1.457	1.318
ma.S.L12	-0.4559	0.720	-0.633	0.527	-1.867	0.955
ma.S.L24	-0.0804	0.396	-0.203	0.839	-0.856	0.696
sigma2	1.46e+05	1.05e-06	1.39e+11	0.000	1.46e+05	1.46e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	35.59
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	0.72	Skew:	0.66
Prob(H) (two-sided):	0.26	Kurtosis:	5.03

Table:83

Observations:

- The SARIMA model built on the complete Sparkling timeseries is chosen, as predictions provide confidence intervals which give better explainability and confidence to the forecasts.

- The diagnostic plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in a normal QQ plot.

- The model summary also provides valuable insights in the model. From the snapshot of the summary below it can be understood that AR(2), MA(3) terms have the highest absolute weightage. The p-values indicate that the terms AR(1), AR(2), MA(1), MA(2) and MA(3) are the most significant terms.

- The rest of the p-values get values higher than alpha 0.05, which fails to reject the null hypothesis that these terms are not significant.

Sparkling 12 Months Forecast

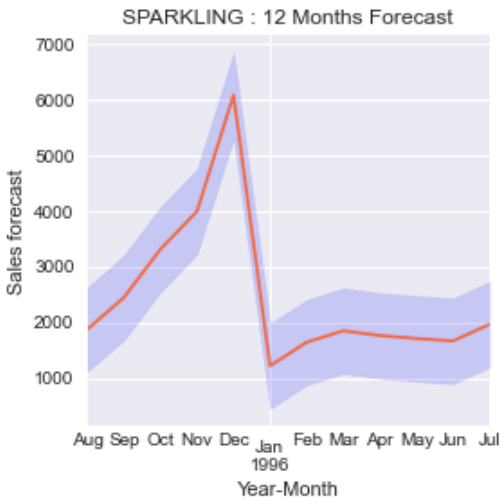


Fig:89

Sparkling Forecasted Values

	Sparkling
1995-08-31	1873.54
1995-09-30	2444.93
1995-10-31	3312.89
1995-11-30	3994.80
1995-12-31	6084.23
1996-01-31	1216.32
1996-02-29	1640.83
1996-03-31	1847.30
1996-04-30	1762.21
1996-05-31	1708.57
1996-06-30	1664.03
1996-07-31	1961.43

Table:84

Sparkling Recommendations:

- The model forecasts sales of 29510 units of Sparkling wine in 12 months. Which is an average sale of 2459 units per month.
- The seasonal sale in December 1995 hit a maximum of 6084 units, before it dropped to the lowest sale in January 1996; at 1216 units.
- The wine company is recommended to ramp up their procurement and production line in accordance with the above forecasts for the third quarter of 1995 (October, November and December), in which a total of 13,392 units of sparkling wine is expected to be sold.
- The forecast also indicates that the year-on-year sale of sparkling wine is not showing an upward trend. The winery must adopt innovative marketing skills to improve sales compared to previous years.
- Adding more exogenous variables into the timeseries data can improve forecasts.

Rose

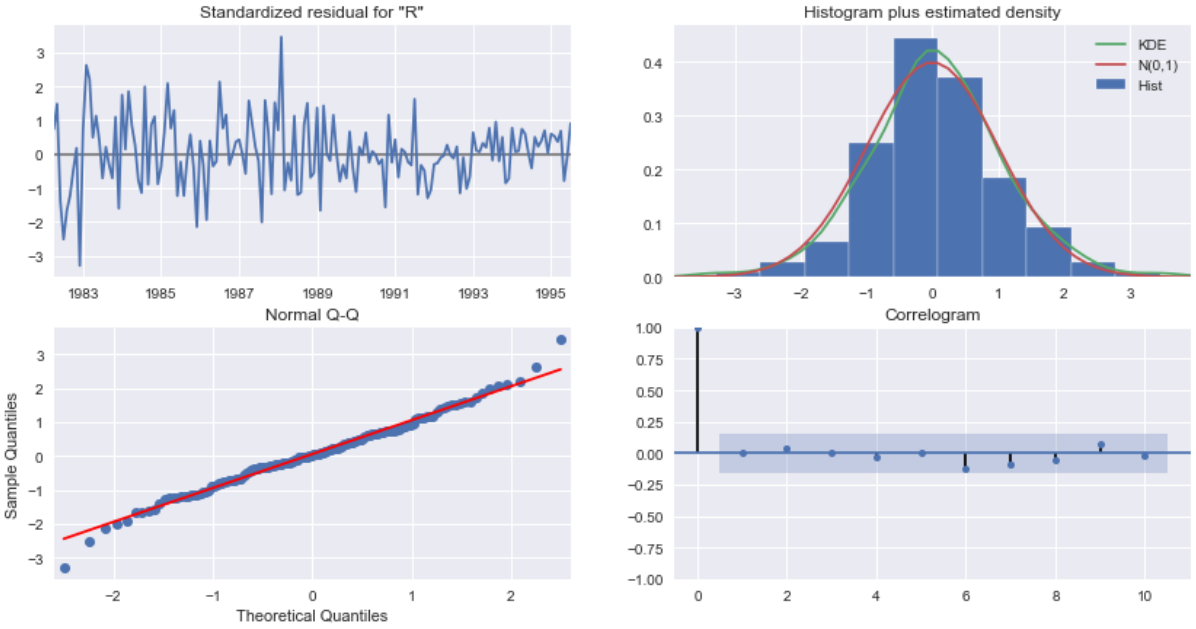


Fig:90

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(4, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-664.135
Date:	Wed, 02 Feb 2022	AIC	1342.270
Time:	22:24:05	BIC	1363.796
Sample:	01-31-1980	HQIC	1351.011
	- 07-31-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0914	0.084	1.093	0.274	-0.072	0.255
ar.L2	-0.1077	0.077	-1.393	0.164	-0.259	0.044
ar.L3	-0.1314	0.076	-1.729	0.084	-0.280	0.018
ar.L4	-0.1071	0.078	-1.375	0.169	-0.260	0.046
ma.L1	-0.8270	0.055	-14.901	0.000	-0.936	-0.718
ma.S.L12	-0.5963	0.059	-10.122	0.000	-0.712	-0.481
sigma2	232.4248	24.359	9.542	0.000	184.682	280.168

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	5.30
Prob(Q):	0.93	Prob(JB):	0.07
Heteroskedasticity (H):	0.22	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.89

Table:85

Observations:

- The SARIMA model is chosen as the final model for prediction on the Rose dataset, as it provides confidence intervals and better explainability of the model.

- The diagnostic plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in a normal QQ plot.

- The model summary also provides valuable insights in the model. From the snapshot of the summary above it can be understood that MA(1) and seasonal MA(1) terms have the highest weightage. The p-values indicate that the terms MA(1) and Seasonal MA(1) are the most significant terms.

- The rest of the p-values get values higher than alpha 0.05, which fails to reject the null hypothesis that these terms are not significant.

- Prediction on the Rose time-series is on a wider confidence band than sparkling.

Rose 12 Months Forecast

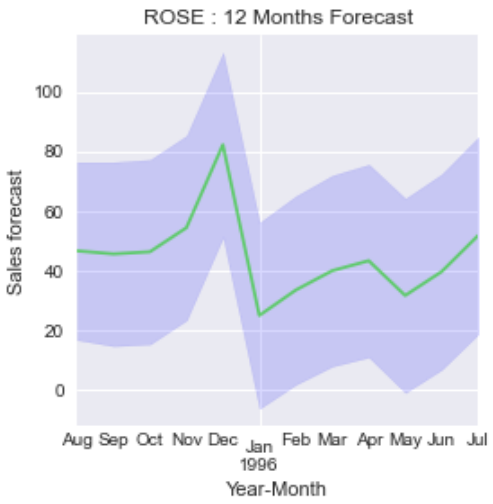


Fig:91

Rose Forecasted Values

	ROSE
1995-08-31	46.54
1995-09-30	45.51
1995-10-31	46.23
1995-11-30	54.32
1995-12-31	82.21
1996-01-31	24.81
1996-02-29	33.35
1996-03-31	39.87
1996-04-30	43.23
1996-05-31	31.53
1996-06-30	39.56
1996-07-31	51.70

Table:86

Rose Recommendations:

- The model forecasts sales of 539 units of Rose wine in 12 months which is an average sale of 45 units per month.
- The seasonal sale in December 1995 reached a maximum of 82 units, before it dropped to the lowest sale in January 1996; at 25 units.
- Unlike Sparkling wine, Rose wine sells a very low number of units and the standard deviation is only 14.5. Which means that higher demand does not impact procurement and production.
- Apart from higher sales in November and December months, Rose sales will be above average in the summer months of July and August.
- The winery should investigate the low demand for Rose wine in the market and make corrective actions in marketing and promotions.