TIME SERIES FORECASTING PROJECT

Question 1

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

Answer:

Sparkling.csv and Rose.csv are two given datasets.

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

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

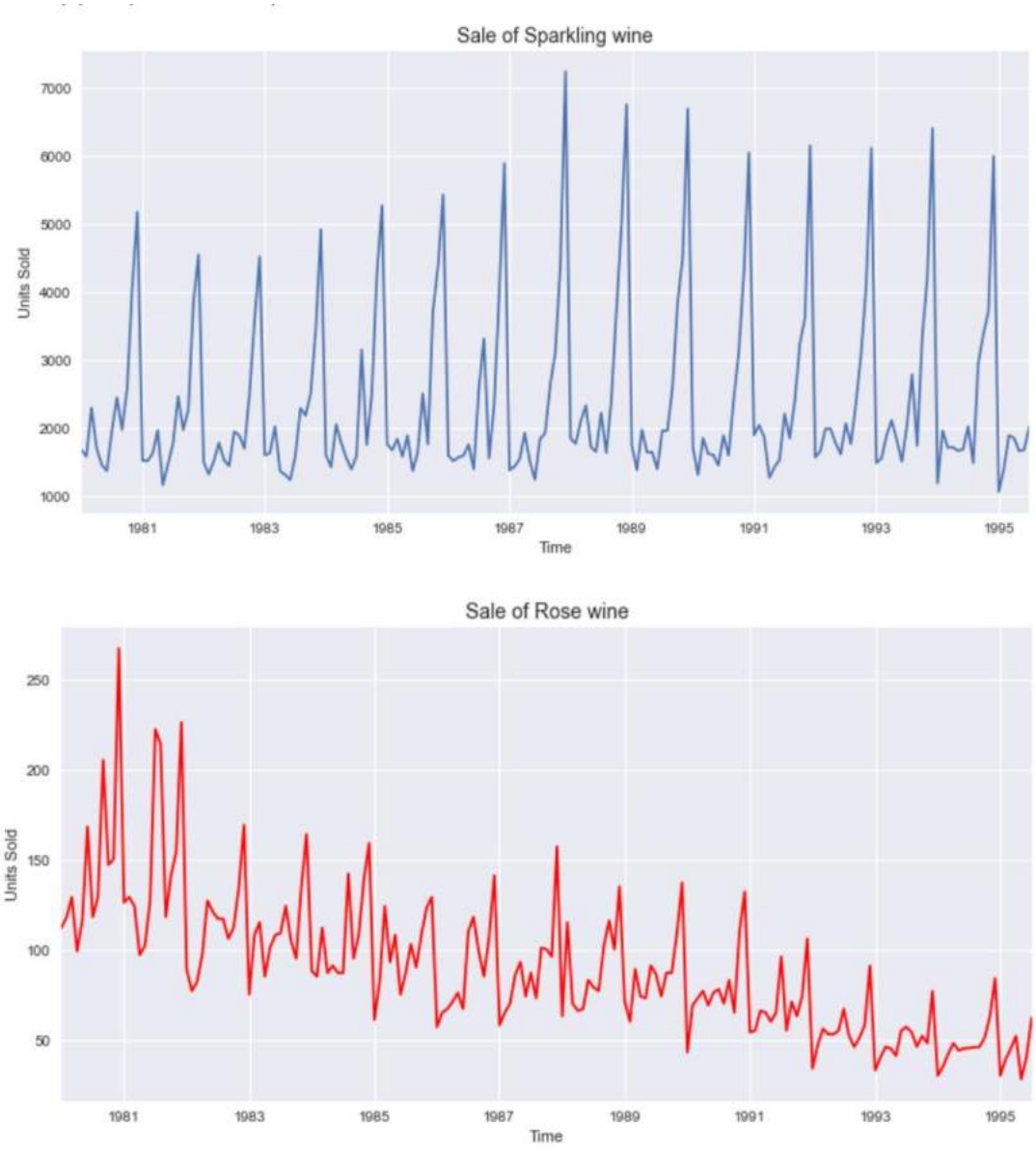
1. The two datasets represent monthly sales of two sets of wines from January 1980 to July 1995.
2. Then a date range has been performed on the "YearMonth" column.
3. To be able to get better comparisons, the two datasets are combined into a single data frame as shown below.

```
df = pd.DataFrame({'YearMonth':date,
                   'Sparkling':df_spark.Sparkling,
                   'Rose':df_rose.Rose})
df.set_index('YearMonth',inplace=True)
```

:

YearMonth	Sparkling	Rose
1980-01-31	1686	112.0
1980-02-29	1591	118.0
1980-03-31	2304	129.0
1980-04-30	1712	99.0
1980-05-31	1471	116.0

4. There are no missing values in Sparkling Dataset.
5. There are two missing values in Rose series dataset for two months in 1994 and they have been imputed using linear method.
6. After plotting the data, we can clearly observe that both of the datasets show significant seasonality.



7. There is consistency seen in Sparkling wine and there is a downward slope trend observed in Rose wine data.

Question 2:

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

Answer:

Descriptive Analysis of both the datasets:

1. The average number of Sparkling and Rose wine units sold are 2402 units and 90 units respectively.
2. Maximum number of wine units sold by Sparkling and Rose units are 7242 and 267 units respectively each month on the given duration.

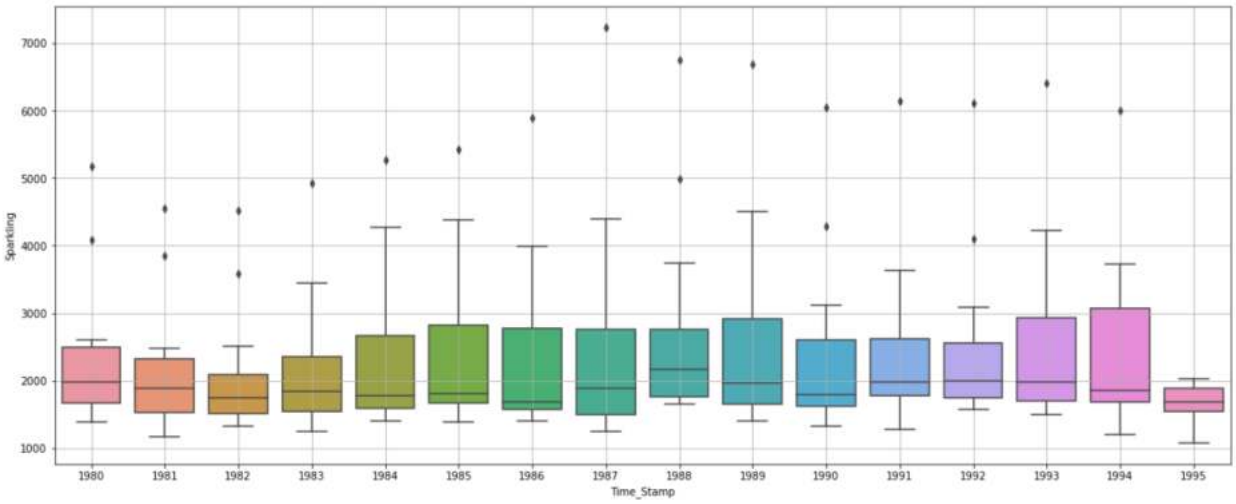
	Sparkling	Rose
count	187.000000	185.000000
mean	2402.417112	90.394595
std	1295.111540	39.175344
min	1070.000000	28.000000
25%	1605.000000	63.000000
50%	1874.000000	86.000000
75%	2549.000000	112.000000
max	7242.000000	267.000000

	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.898502	39.256767	28.0	62.5	85.0	111.0	267.0

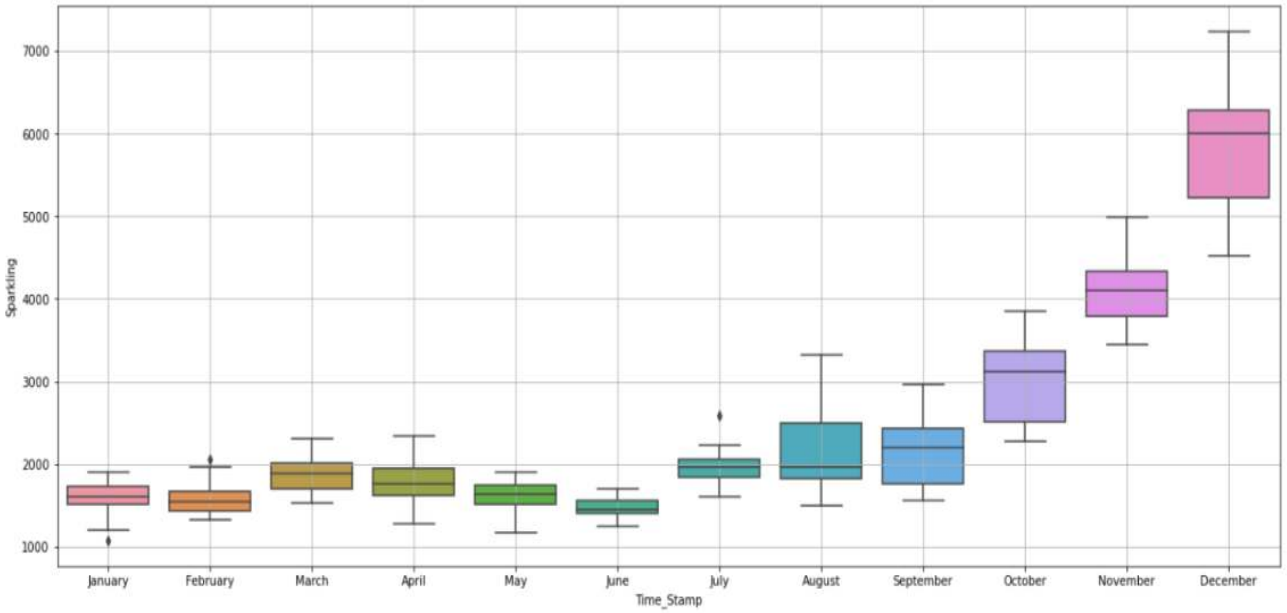
EDA: Sparkling Wine

1. The box plots show that there are outliers present in the dataset and they represent the seasonal change in sales over the period of years.

Yearly Box plot of the sparkling wine dataset



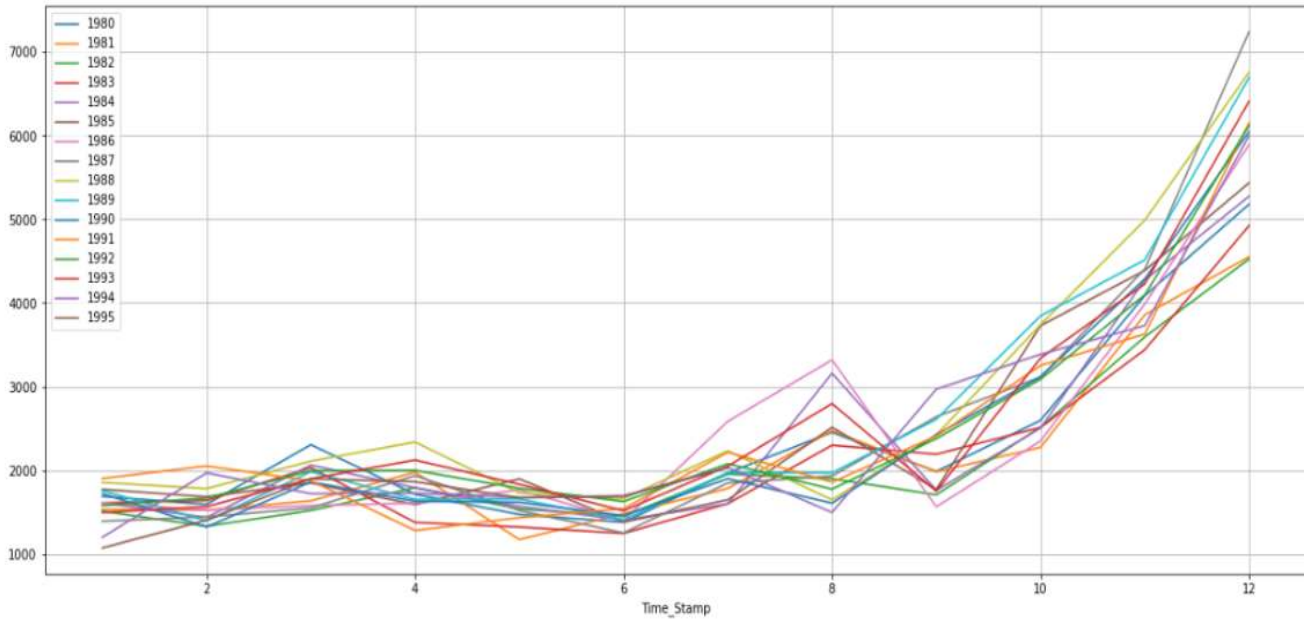
Monthly box plot of Sparkling wine dataset

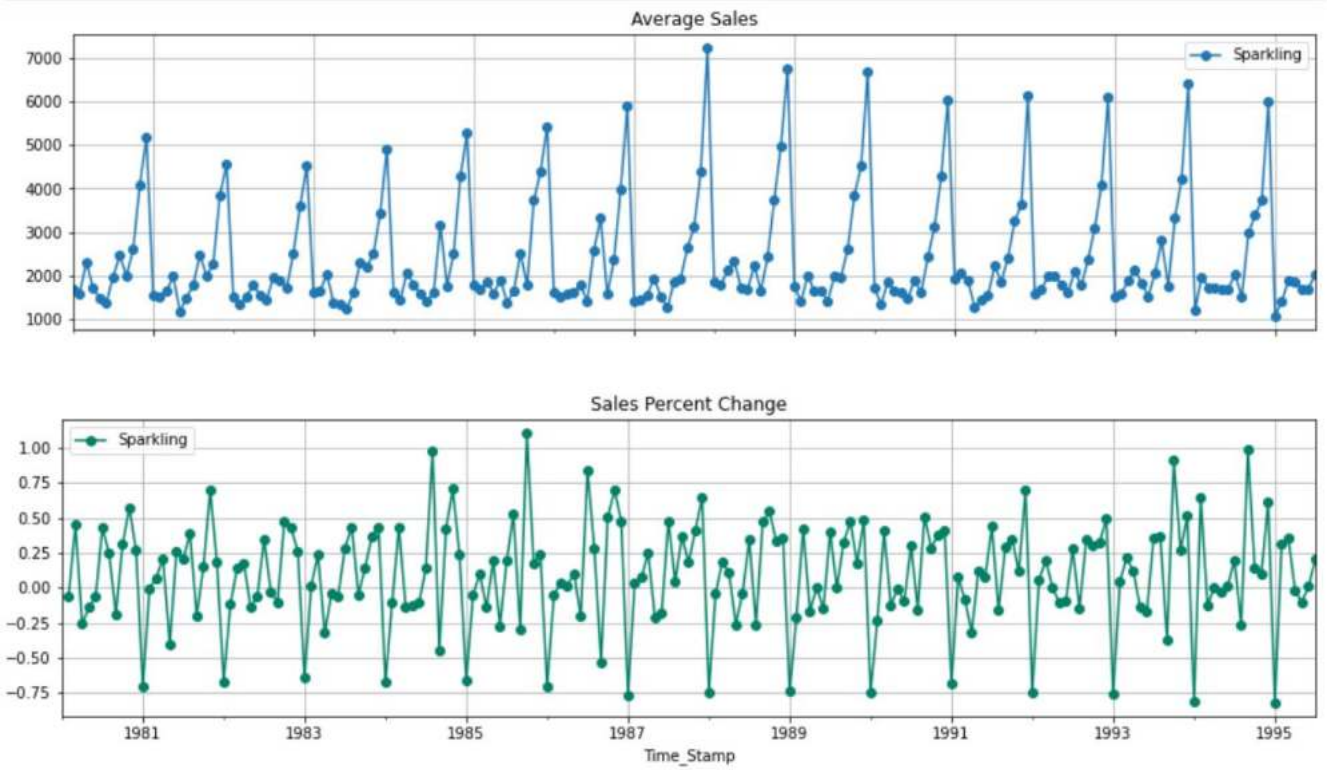


2. Another observation that can be made is that the monthly sales in December is particularly high due to the holidays/festive season.

Monthly sales across the years for Sparkling wine dataset

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



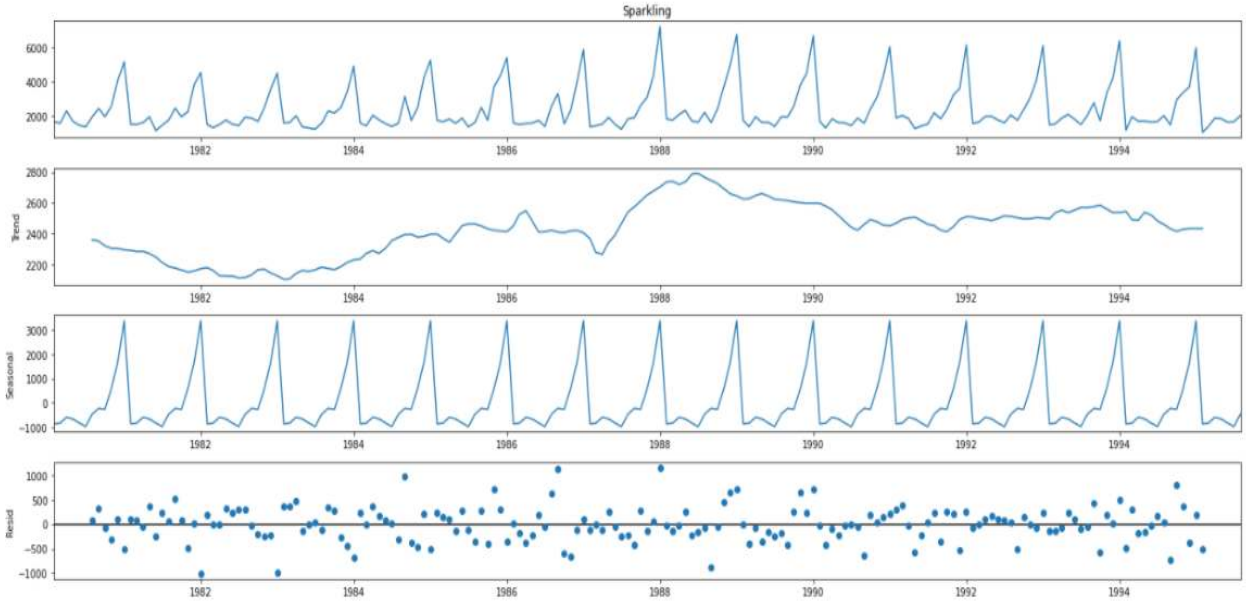

Average Sales
Sparkling
7000
6000
5000
4000
3000
2000
1000
Sales Percent Change
Sparkling
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
1981
1983
1985
1987
1989
1991
1993
1995
Time_Stamp

Decomposition of sparkling dataset plots:

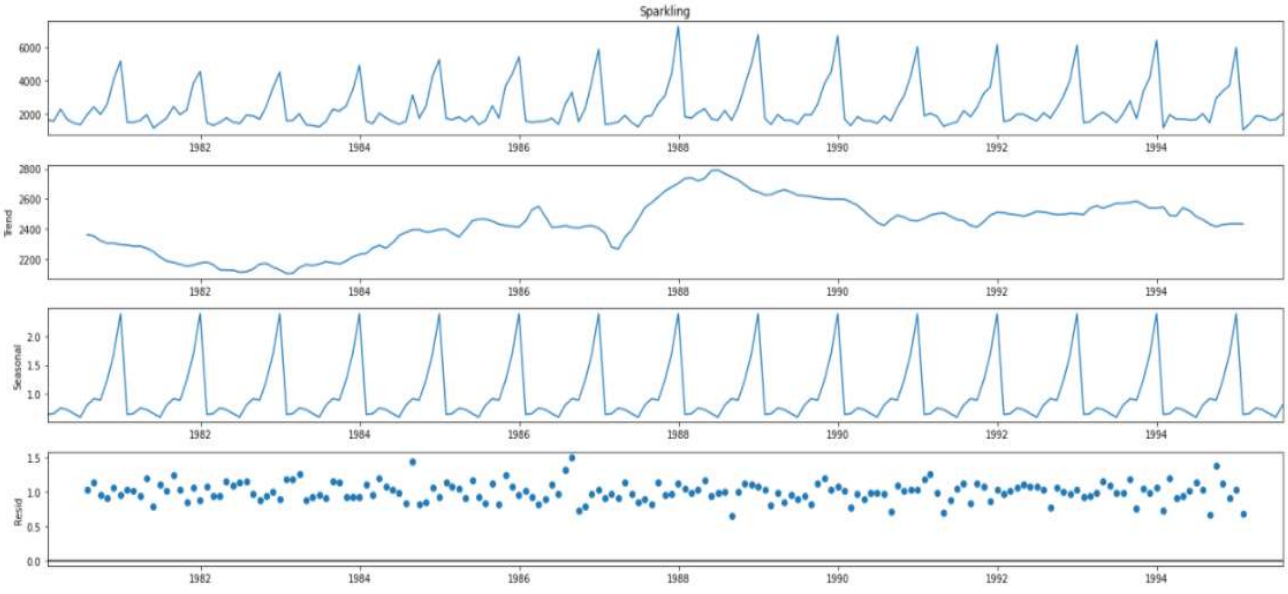
If the seasonality and residual components are independent of the trend, then you have an additive series. If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then you have a multiplicative series

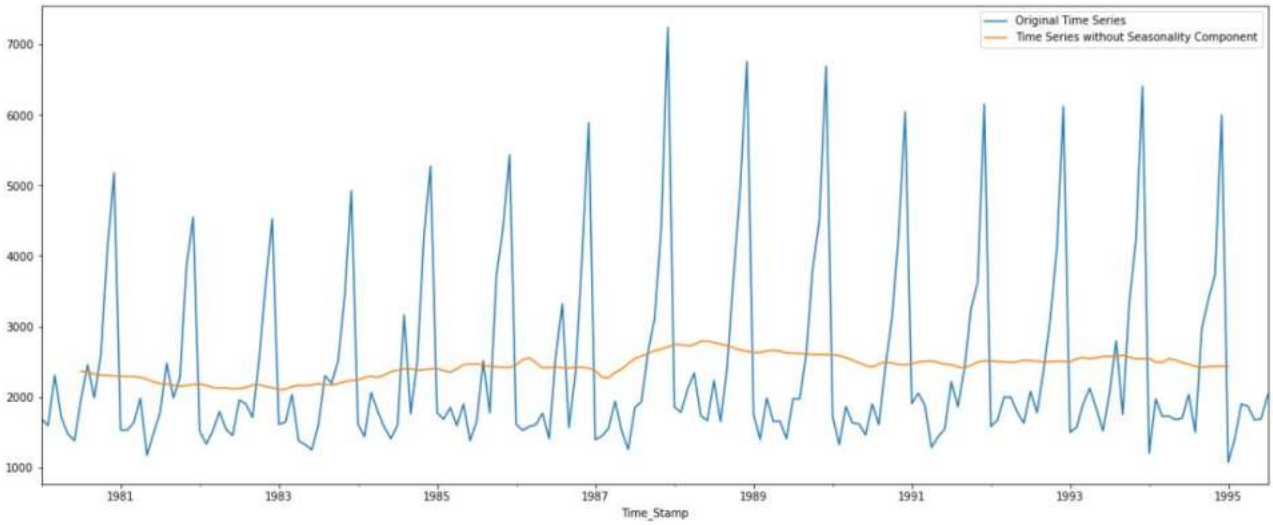
3. No consistent trend in the below plots

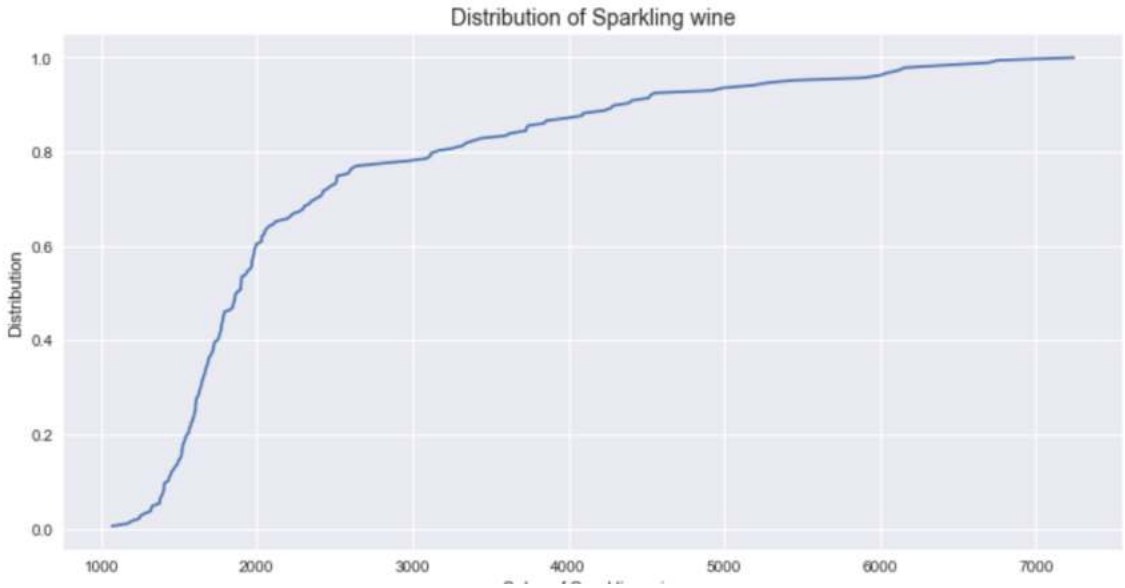
Additive Model



Multiplicative model



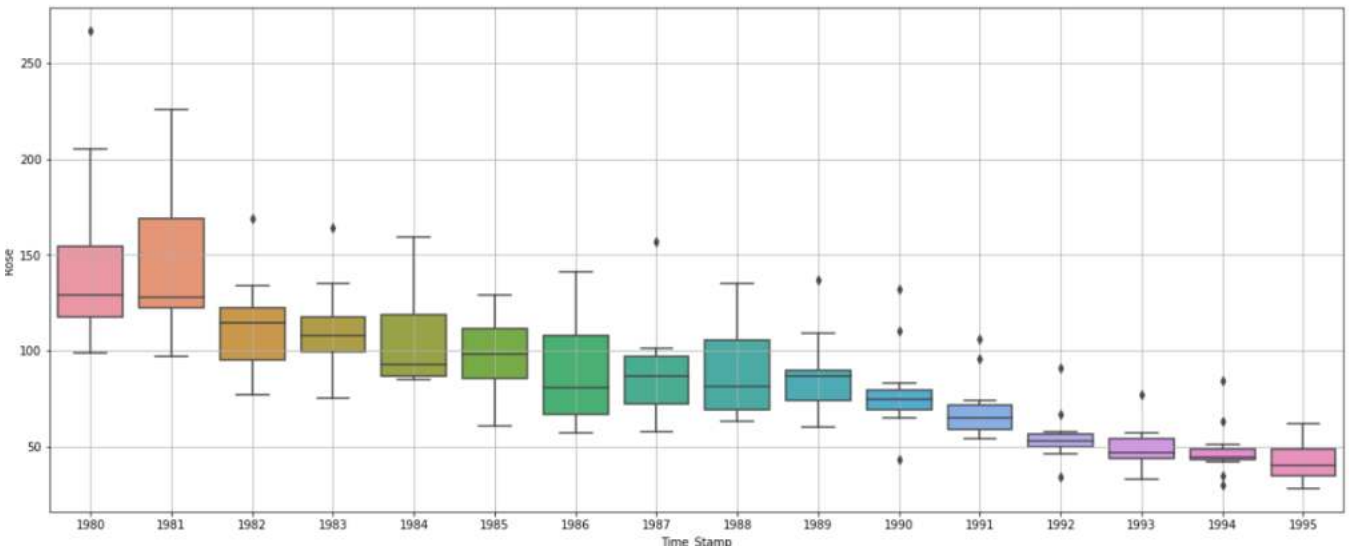

Original Time Series
Time Series without Seasonality Component
7000
6000
5000
4000
3000
2000
1000
1981
1983
1985
1987
1989
1991
1993
1995
Time_Stamp


Distribution of Sparkling wine
Distribution
1.0
0.8
0.6
0.4
0.2
0.0
1000
2000
3000
4000
5000
6000
7000

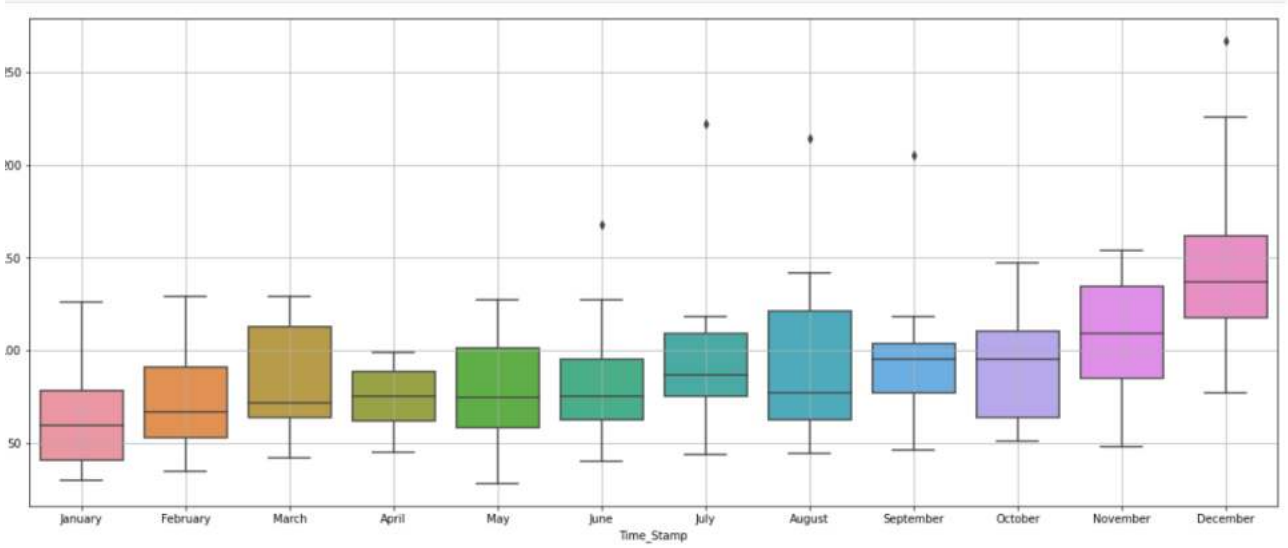
ROSE WINE DATASET: EDA

1. The same kind of seasonality trend shown in Sparkling dataset, is seen in Rose wine boxplots. The sales are exponentially higher in the months of November to December and then decrease in the following month of January.

Yearly Box plot of rose wine dataset

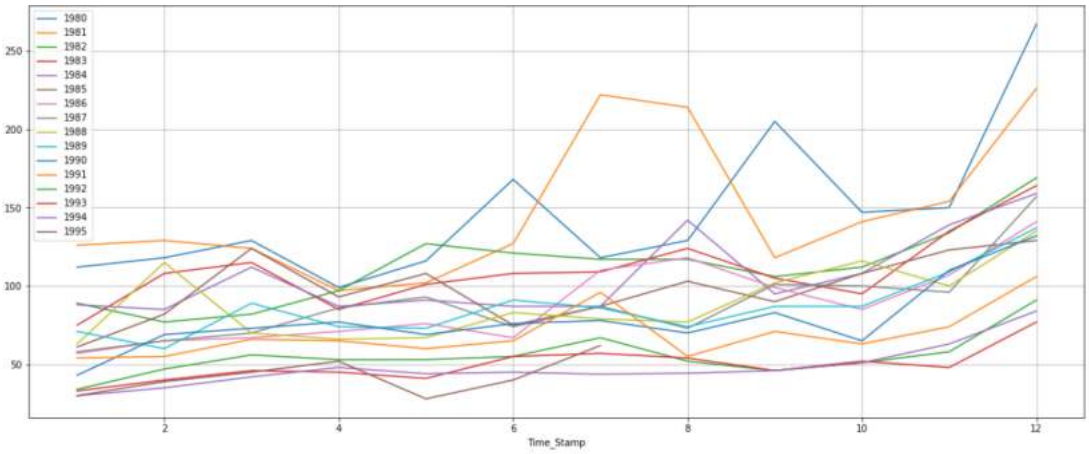


Monthly plot: Rose wine



Monthly sales across years

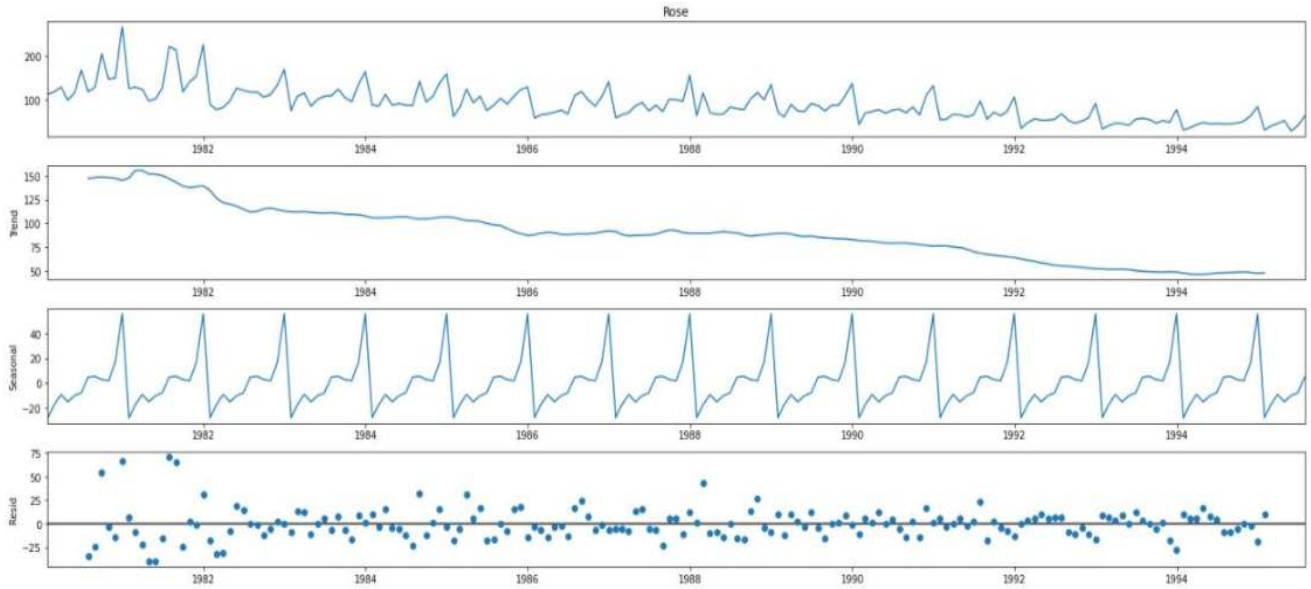
Time_Stamp	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Time_Stamp																
1	112.0	126.0	89.0	75.0	88.0	61.0	57.0	58.0	63.0	71.0	43.0	54.0	34.0	33.0	30.000000	30.0
2	118.0	129.0	77.0	108.0	85.0	82.0	65.0	65.0	115.0	60.0	69.0	55.0	47.0	40.0	35.000000	39.0
3	129.0	124.0	82.0	115.0	112.0	124.0	67.0	70.0	70.0	89.0	73.0	66.0	56.0	46.0	42.000000	45.0
4	99.0	97.0	97.0	85.0	87.0	93.0	71.0	86.0	66.0	74.0	77.0	65.0	53.0	45.0	48.000000	52.0
5	116.0	102.0	127.0	101.0	91.0	108.0	76.0	93.0	67.0	73.0	69.0	60.0	53.0	41.0	44.000000	28.0
6	168.0	127.0	121.0	108.0	87.0	75.0	67.0	74.0	83.0	91.0	76.0	65.0	55.0	55.0	45.000000	40.0
7	118.0	222.0	117.0	109.0	87.0	87.0	110.0	87.0	79.0	86.0	78.0	96.0	67.0	57.0	43.693064	62.0
8	129.0	214.0	117.0	124.0	142.0	103.0	118.0	73.0	77.0	74.0	70.0	55.0	52.0	54.0	44.326877	NaN
9	205.0	118.0	106.0	105.0	95.0	90.0	99.0	101.0	102.0	87.0	83.0	71.0	46.0	46.0	46.000000	NaN
10	147.0	141.0	112.0	95.0	108.0	108.0	85.0	100.0	116.0	87.0	65.0	63.0	51.0	52.0	51.000000	NaN
11	150.0	154.0	134.0	135.0	139.0	123.0	107.0	96.0	100.0	109.0	110.0	74.0	58.0	48.0	63.000000	NaN
12	267.0	226.0	169.0	164.0	159.0	129.0	141.0	157.0	135.0	137.0	132.0	106.0	91.0	77.0	84.000000	NaN



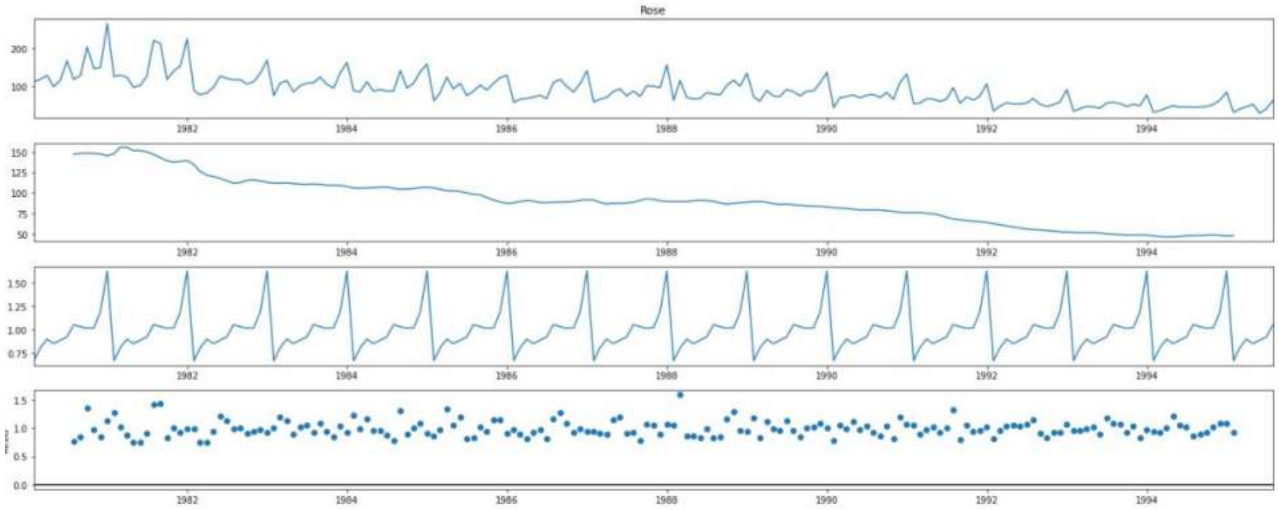
Time series decomposition plots of rose datasets

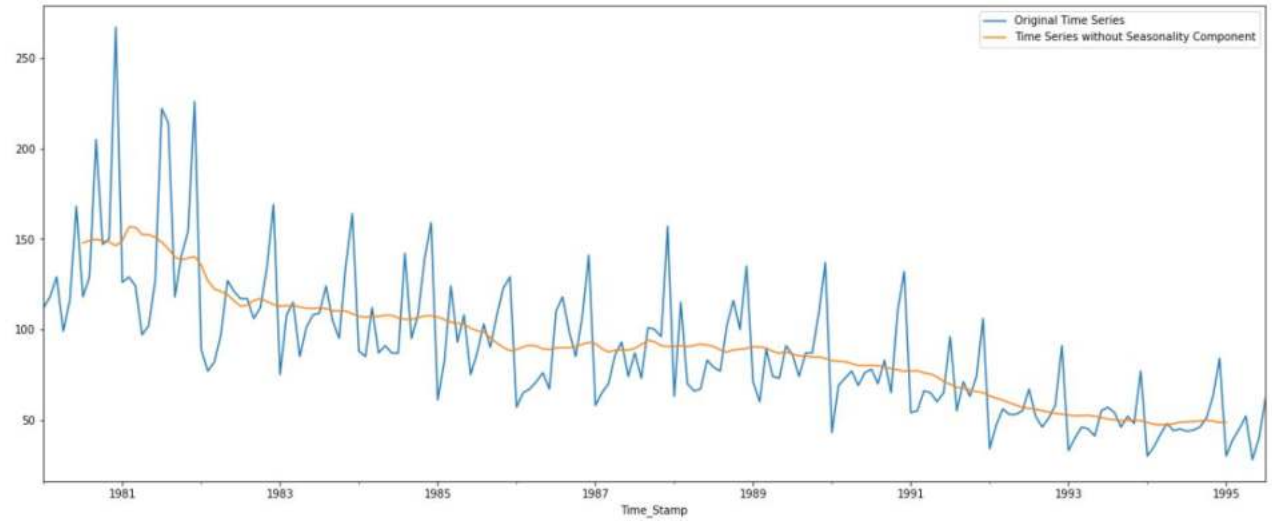
2. Trend component is shown decreasing indicating a downward trend and seasonality components are seen to be consistent.
3. High variability is demonstrated in both additive and multiplicative decompositions.

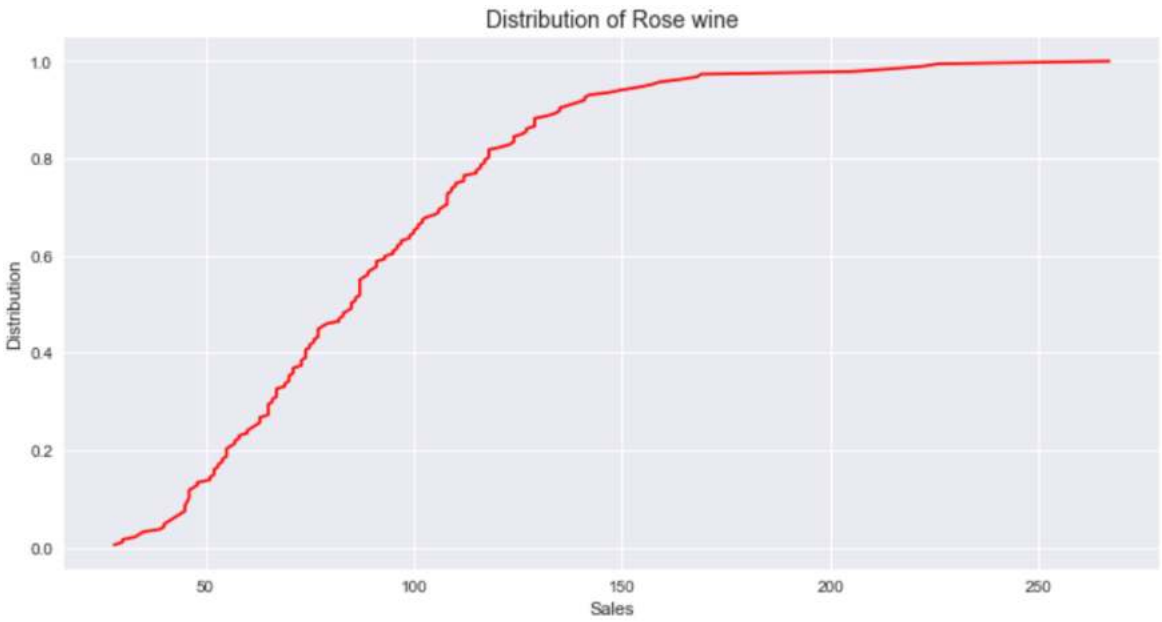
Additive Model



Multiplicative model






Distribution of Rose wine
Distribution
1.0
0.8
0.6
0.4
0.2
0.0
50
100
150
200
250
Sales

Q3. Split the data into training and test. The test data should start in 1991.

- Sparkling and Rose wines datasets are split into training and testing datasets and the test data is split from the year 1991.

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

Last and few rows of the training and testing datasets of both Sparkling and RoseWine sales

First few rows of Training Data

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

Last few rows of Training Data

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

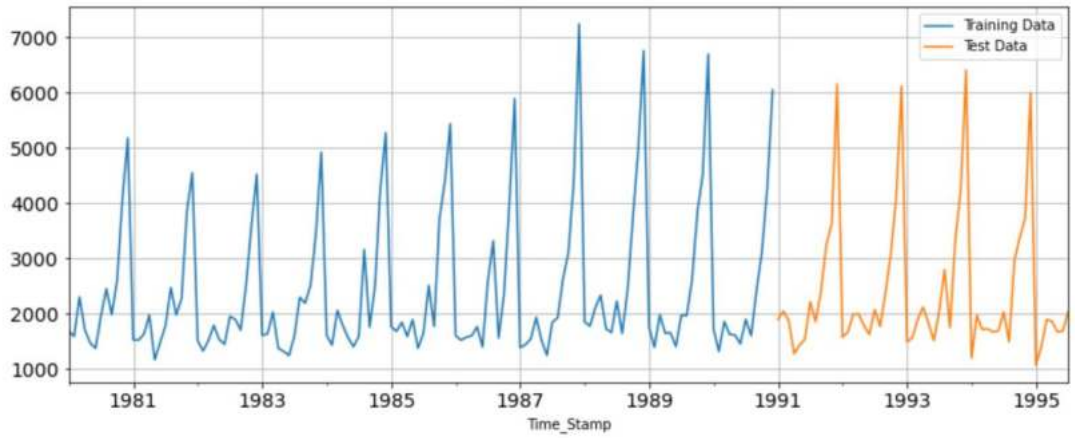
Last few rows of Training Data

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

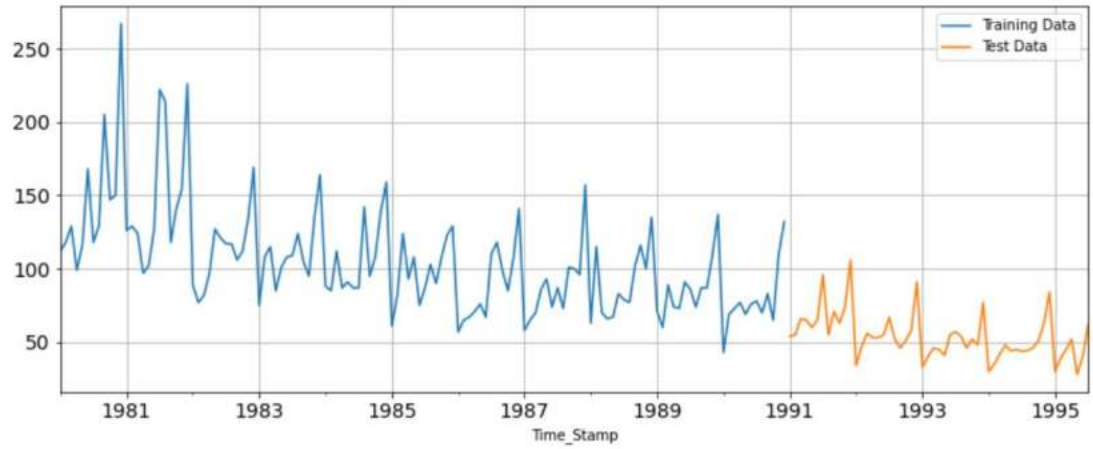
First few rows of Test Data

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

Sparkling Wine sales Data split plot



Rose Wine sales Data split plot



Q4. Build various 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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE

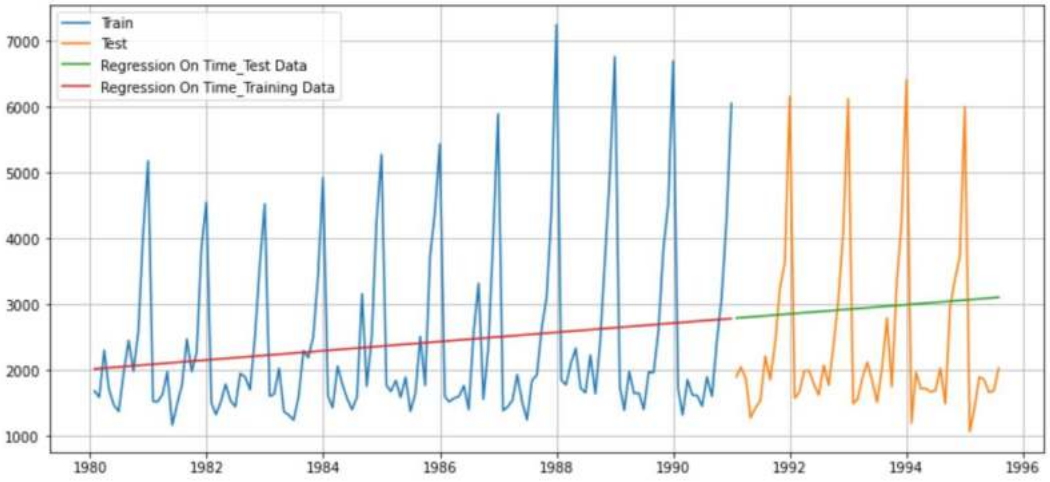
MODEL 1: Linear Regression

- Linear Regression Model is applied on both training and testing sets of both the datasets.

```
lr.fit(LinearRegression_train[['time']],LinearRegression_train['Sparkling'].values)
```

- For sparkling dataset, we can observe that there is an upward trend shown in Linear regression plot

Sparkling wine: Linear Regression plot



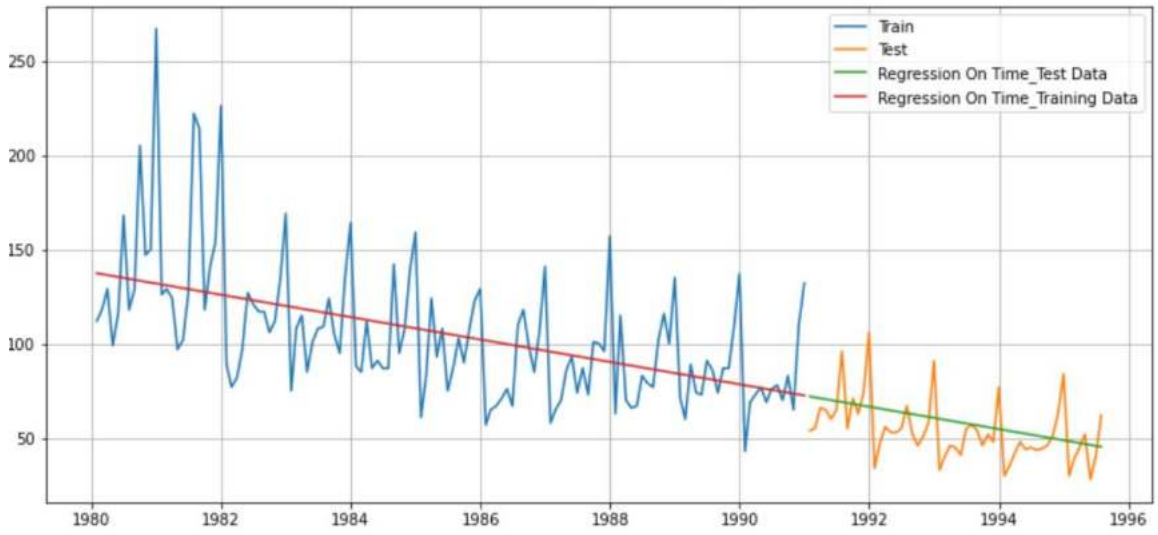
- The test RMSE and MAPE values of the Sparkling Test dataset are 1389 and 50 respectively.
- MAPE value of the test dataset suggests that model gives an error rate of 50% .

Sparkling wine: RMSE and MAPE values on test set

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15

- In Linear regression plot of rose dataset, there is a gradual downward trend observed

Rose wine: Linear Regression plot



- The RMSE values of Linear regression model rose sales test set are shown below.

Sparkling wine: RMSE and MAPE values on test set

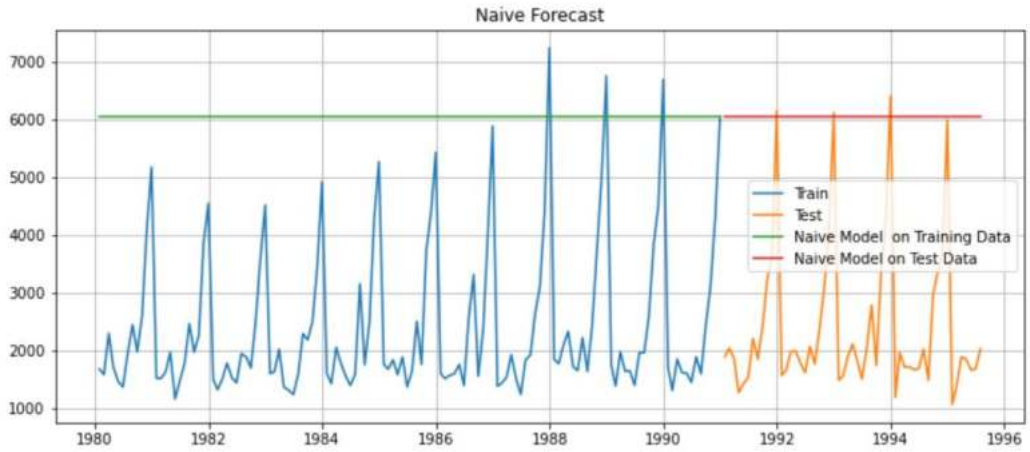
	Test RMSE	Test MAPE
RegressionOnTime	51.486843	91.84

MODEL 2: Naïve Forecast

- Naïve Forecast Model is applied on both training and testing sets of both the datasets.

```
NaiveModel_train = train.copy()
NaiveModel_test = test.copy()
```

Sparkling wine: Naïve Forecast plot



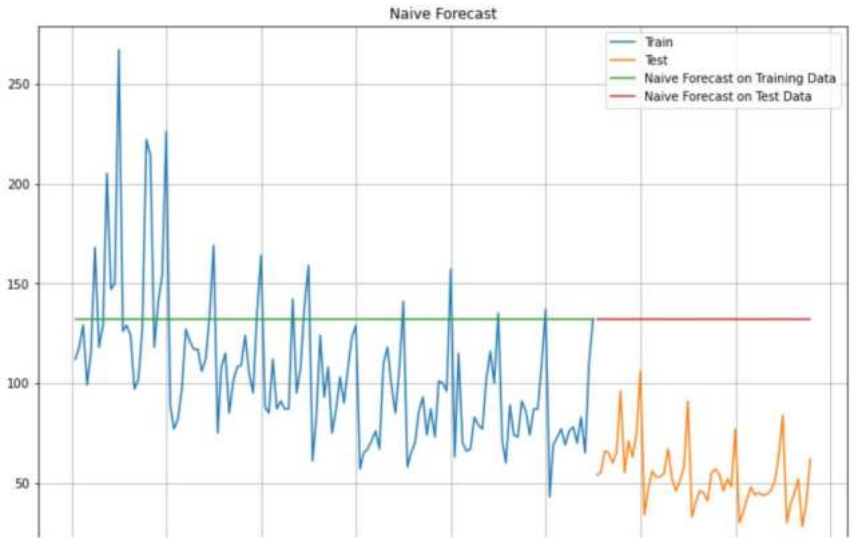
- For naïve model, Error percentage is too high for this model, hence it is considered as a poor fit for this dataset.

Sparkling wine: RMSE and MAPE values on test set

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87

- The naïve forecast model plot for rose dataset are shown below:

Rose wine: Naïve Forecast plot



- For naïve model, Error percentage is too high for this model, hence it is considered as a poor fit for this dataset.

Rose wine: RMSE and MAPE values on test set

	Test RMSE	Test MAPE
RegressionOnTime	51.486843	91.84
NaiveModel	79.778066	145.35

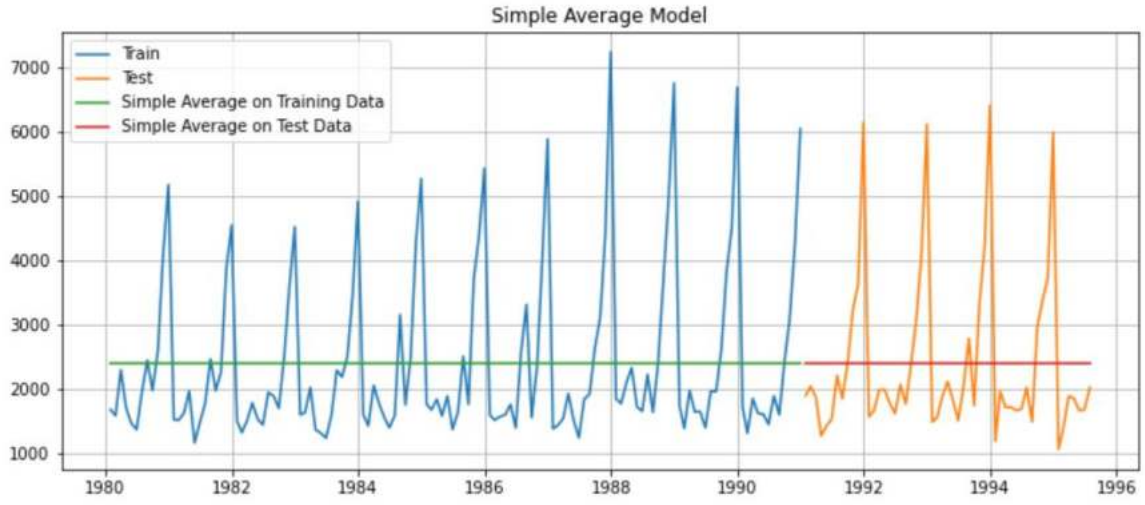
MODEL 3: Simple Average Model

- Simple average Model is applied on both training and testing sets of both the datasets and it is also applied on using the mean of the time series variable.

```
SimpleAverage_train = train.copy()
SimpleAverage_test = test.copy()
```

- Model did not capture the trend or seasonality components.

Sparkling wine: Simple average model plot



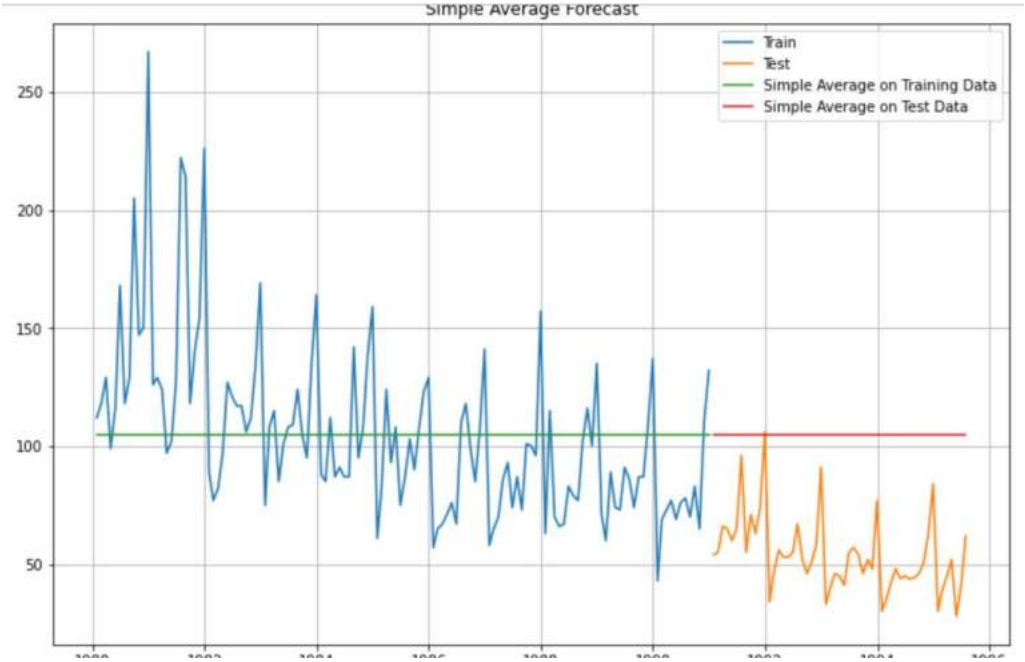
- Error percentage for this model is around 39%.

Sparkling wine: RMSE and MAPE values on test set

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

- The Simple Average model plot for rose dataset are shown below:

Rose wine: Simple Average Forecast plot



- For simple average model, Error percentage is too high for this model, hence it is considered as a poor fit for this dataset.

Rose wine: RMSE and MAPE values on test set

	Test RMSE	Test MAPE
RegressionOnTime	51.486843	91.84
NaiveModel	79.778066	145.35
SimpleAverageModel	53.521557	95.13

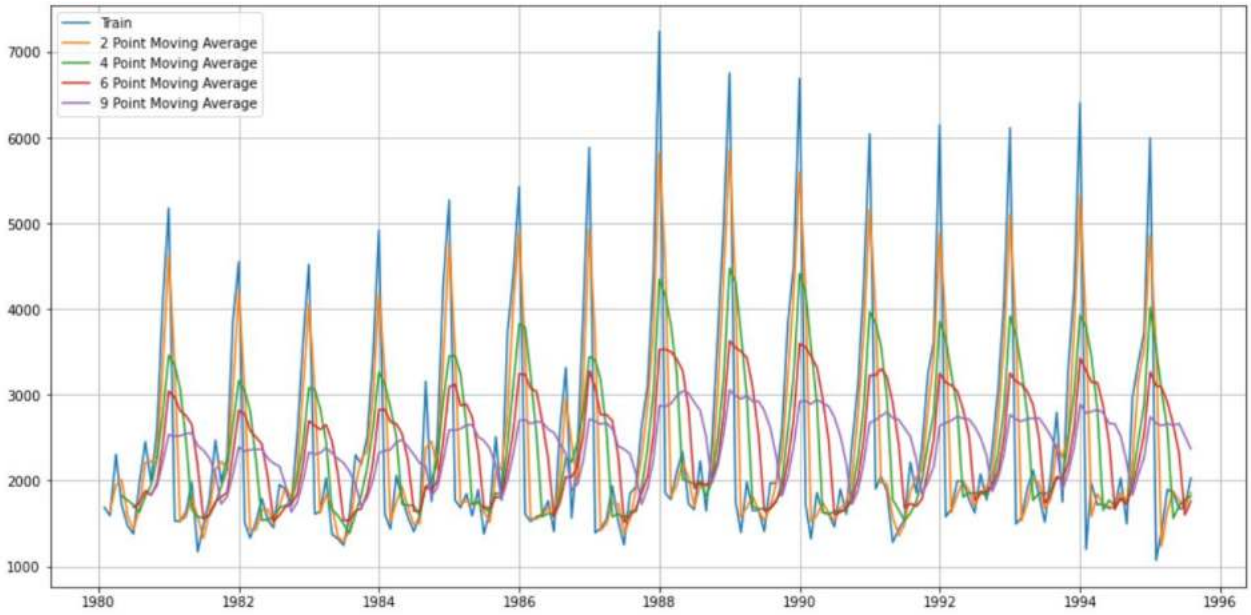
MODEL 4: Moving Average Model

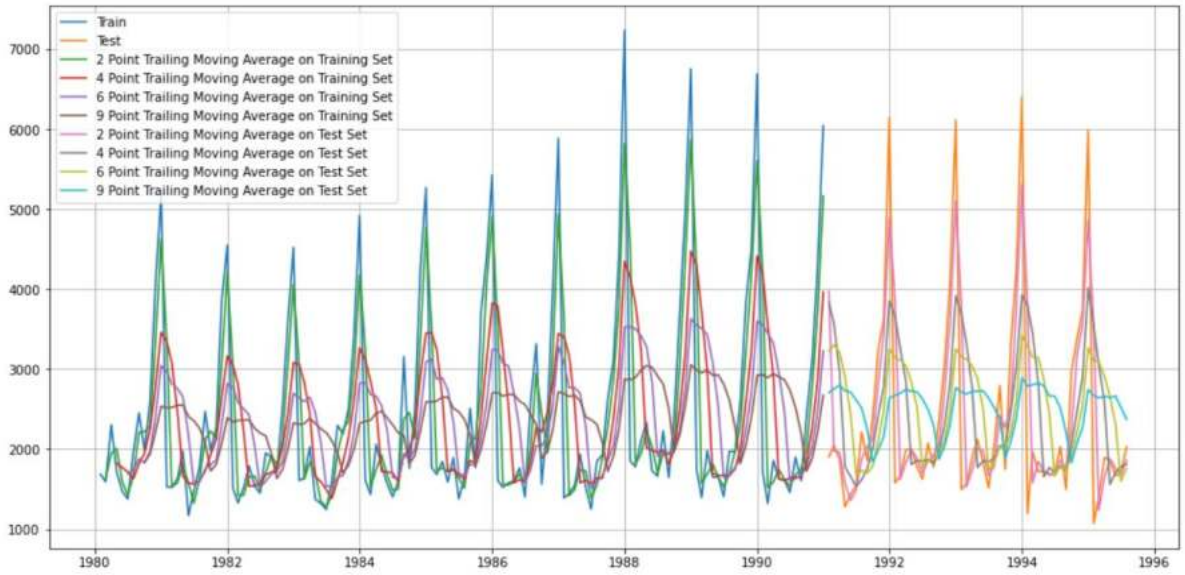
- In moving average model, we will calculate rolling means for 2-point, 4-point, 6-point and 9-point intervals.

Trailing moving averages

```
MovingAverage['Trailing_2'] = MovingAverage['Sparkling'].rolling(2).mean()
MovingAverage['Trailing_4'] = MovingAverage['Sparkling'].rolling(4).mean()
MovingAverage['Trailing_6'] = MovingAverage['Sparkling'].rolling(6).mean()
MovingAverage['Trailing_9'] = MovingAverage['Sparkling'].rolling(9).mean()
```

Sparkling wine: Moving average model plot

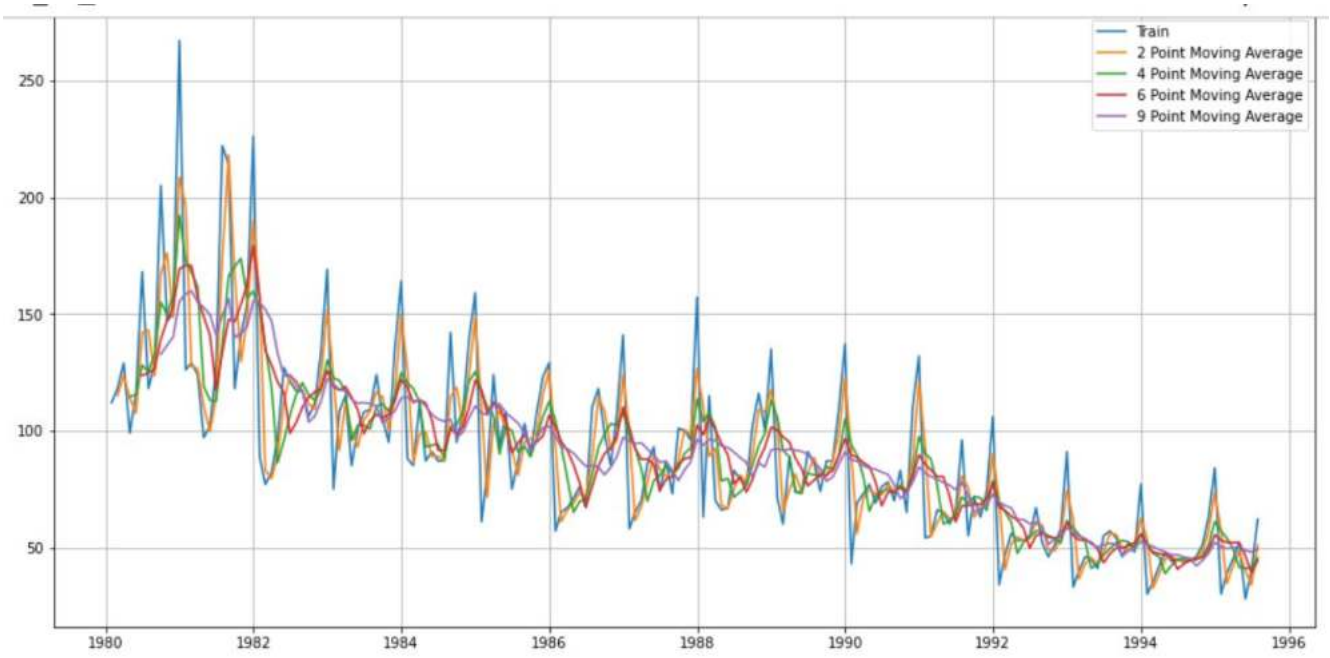


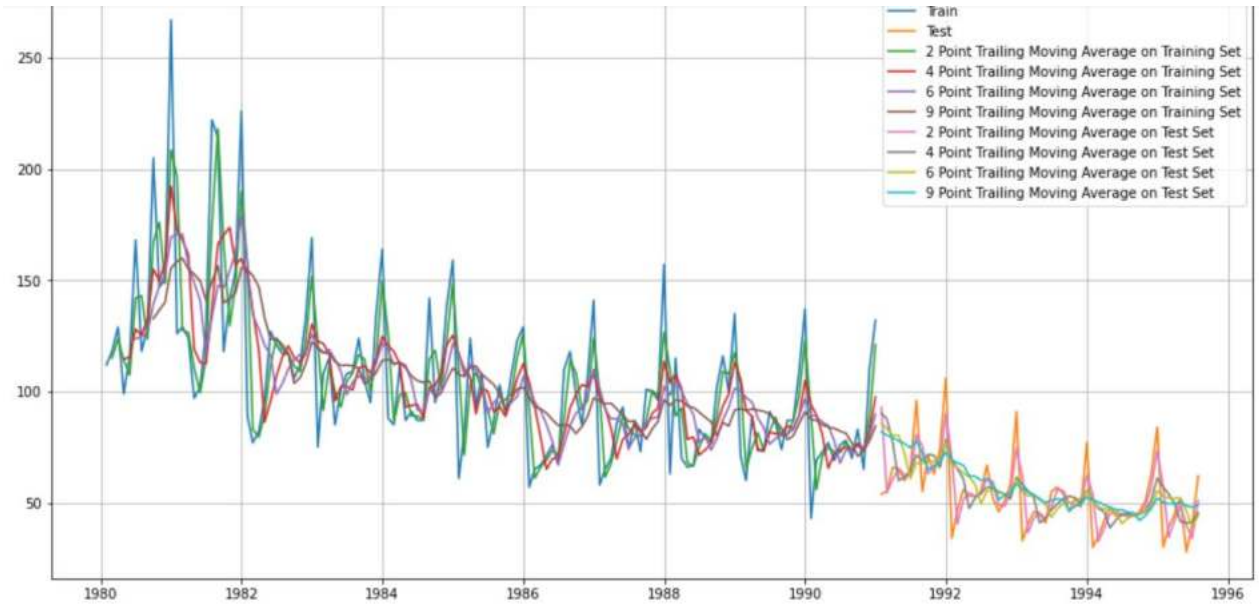


- The test RMSE and MAPE values of all the trailing points of different intervals are shown below. The minimum error percentage is shown in 2-point; therefore, the best interval is 2-point.

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	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

Rose wine: Moving average model plot





- The test RMSE and MAPE values of all the trailing points of different intervals are shown below. The minimum error percentage is shown in 2-point; therefore, the best interval is 2-point.

	Test RMSE	Test MAPE
RegressionOnTime	51.486843	91.84
NaiveModel	79.778066	145.35
SimpleAverageModel	53.521557	95.13
2pointTrailingMovingAverage	11.530180	13.60
4pointTrailingMovingAverage	14.462330	19.59
6pointTrailingMovingAverage	14.586916	20.83
9pointTrailingMovingAverage	14.740112	21.13

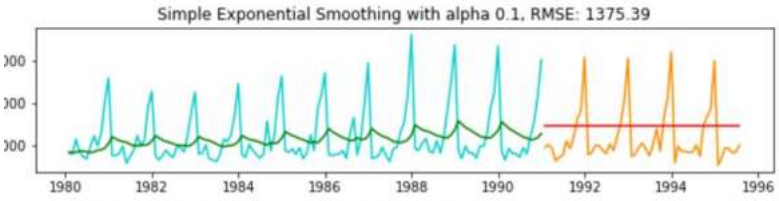
MODEL 5: SIMPLE EXPONENTIAL SMOOTHING

Sparkling Dataset

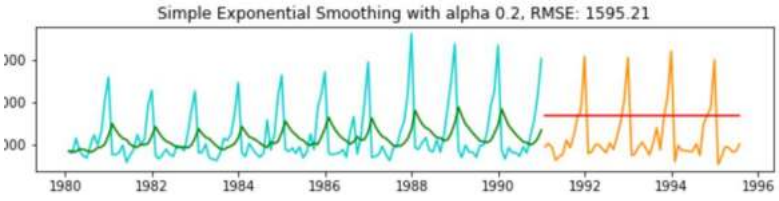
Simple exponential smoothing is applied to the dataset with different smoothing alpha levels between 0 and 1

```
alpha_list = [0.1, 0.2, 0.3, 0.5, 0.6,0.7,0.8, 0.9, 1.00]
pred_train_SES  = train.copy()
pred_test_SES  = test.copy() # Have a copy of the test dataset
```

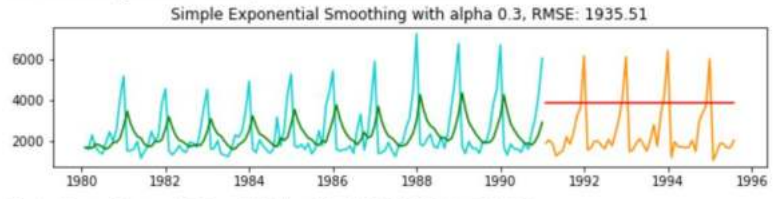
```
st: For alpha = 0.10,  RMSE is 1375.3934 MAPE is 49.53
r smoothing level = 0.10,  Initial level 1686.00
```



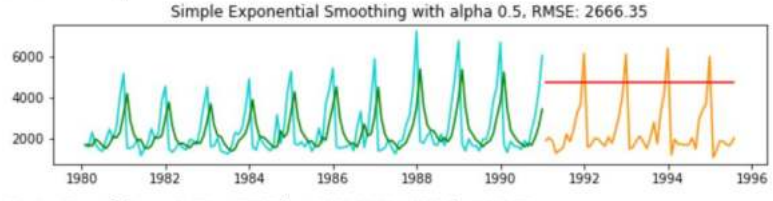

```
st: For alpha = 0.20,  RMSE is 1595.2068 MAPE is 60.46
r smoothing level = 0.20,  Initial level 1686.00
```




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

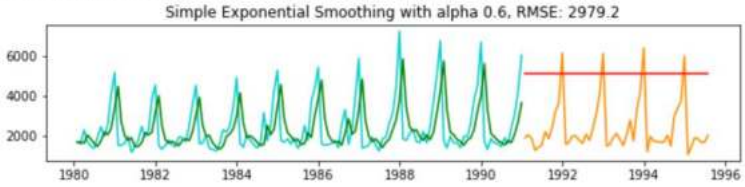



```
Test: For alpha = 0.50,  RMSE is 2666.3514 MAPE is 106.27
For smoothing level = 0.50,  Initial level 1686.00
```

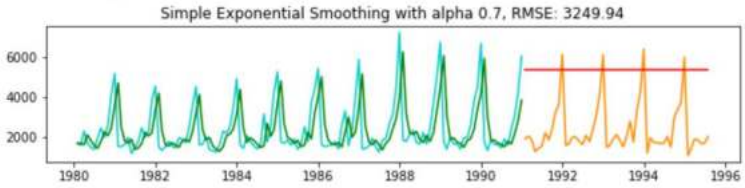



```
Test: For alpha = 0.60,  RMSE is 2979.2044 MAPE is 118.77
```

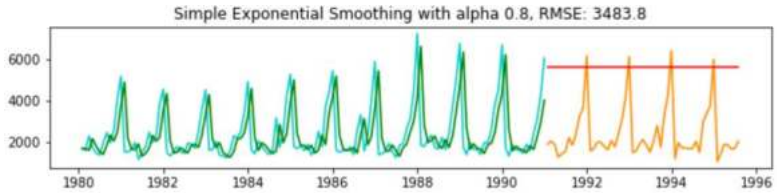
For smoothing level = 0.60, Initial level 1686.00



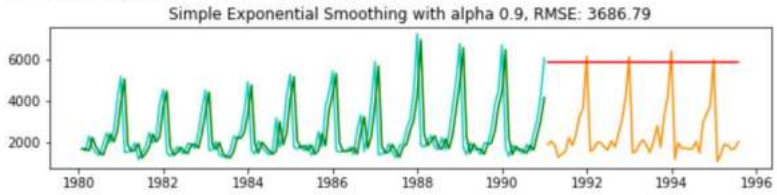
Test: For alpha = 0.70, RMSE is 3249.9441 MAPE is 129.34
For smoothing level = 0.70, Initial level 1686.00



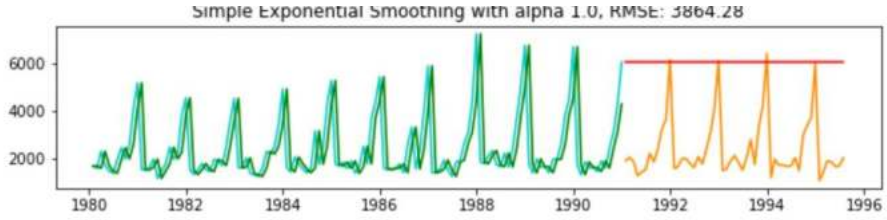
Test: For alpha = 0.80, RMSE is 3483.8010 MAPE is 138.34
For smoothing level = 0.80, Initial level 1686.00



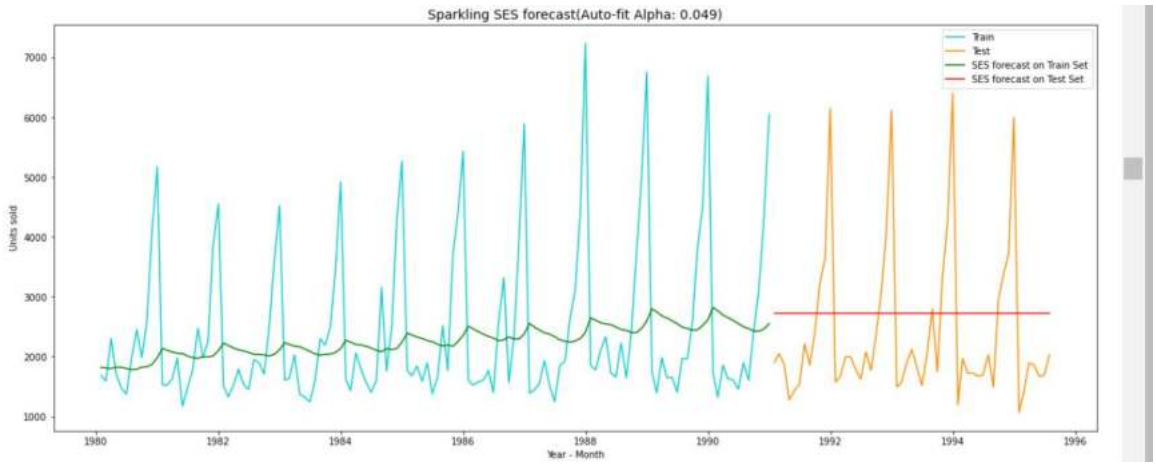
Test: For alpha = 0.90, RMSE is 3686.7943 MAPE is 146.08
For smoothing level = 0.90, Initial level 1686.00



Test: For alpha = 1.00, RMSE is 3864.2794 MAPE is 152.87
For smoothing level = 1.00, Initial level 1686.00



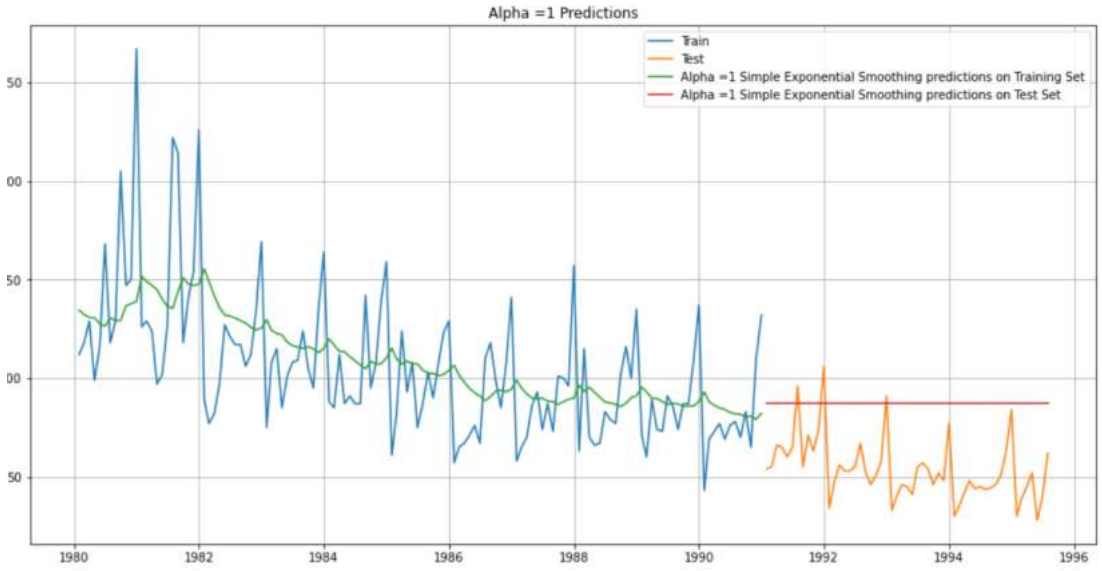
- Then autofit model is run and the chosen smoothing parameter is 0.49. The values that were found in this method were higher than the result values found in the other smoothing parameter values
- RMSE and MAPE values were descreasing in respect to increase in the alpha values.

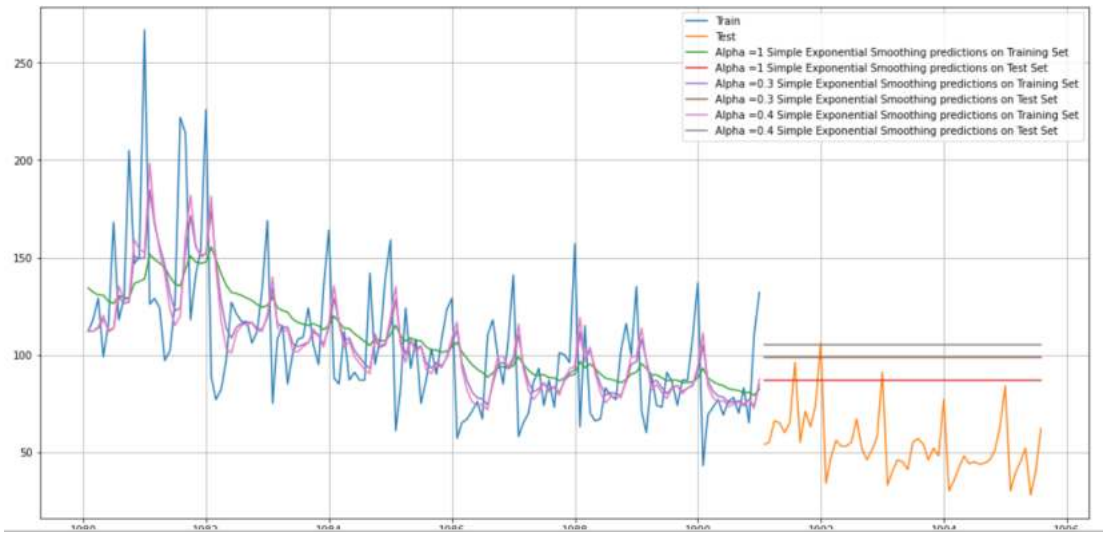


	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
Alpha= 0.49,SimpleExponentialSmoothing	1316.035487	45.47

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
Alpha= 0.49,SimpleExponentialSmoothing	1316.035487	45.47
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	75.66
Alpha=0.4,SimpleExponentialSmoothing	2311.919615	91.55

ROSE WINE DATASET:





Alpha=1,SimpleExponentialSmoothing	36.858571	64.05
Alpha=0.3,SimpleExponentialSmoothing	47.566302	83.89
Alpha=0.4,SimpleExponentialSmoothing	53.828369	95.70

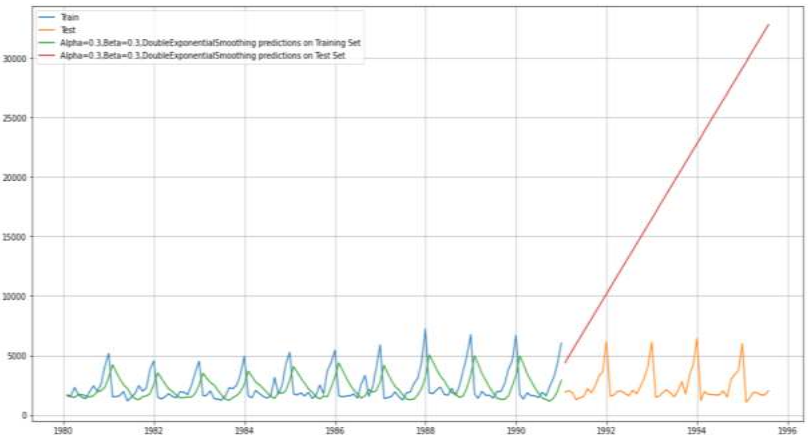
MODEL 6: Double exponential smoothing

The Double Exponential Smoothing has two parameters 'smoothing level' and 'smoothing_slope' parameter which are optimized using inbuilt hyperparameter 'optimized' and also optimized iteratively based on Test RMSE values

Sparkling Dataset

Alpha and beta values are fitted properly on the model

0	0.3	0.3	1592.292788	53.75	18259.110704	675.28
1	0.3	0.4	1682.573828	57.13	26069.841401	960.18
2	0.3	0.5	1771.710791	59.92	34401.512440	1264.55
3	0.3	0.6	1848.576510	62.06	42162.748095	1548.94
4	0.3	0.7	1899.949006	62.87	47832.397419	1757.91
...	...	...	...	...	...	...
59	1.0	0.6	1753.402326	45.75	49327.087977	1824.42
60	1.0	0.7	1825.187155	47.97	52655.765663	1945.23
61	1.0	0.8	1902.013709	50.11	55442.273880	2046.36
62	1.0	0.9	1985.368445	52.03	57823.177011	2132.75
63	1.0	1.0	2077.672157	53.81	59877.076519	2207.28

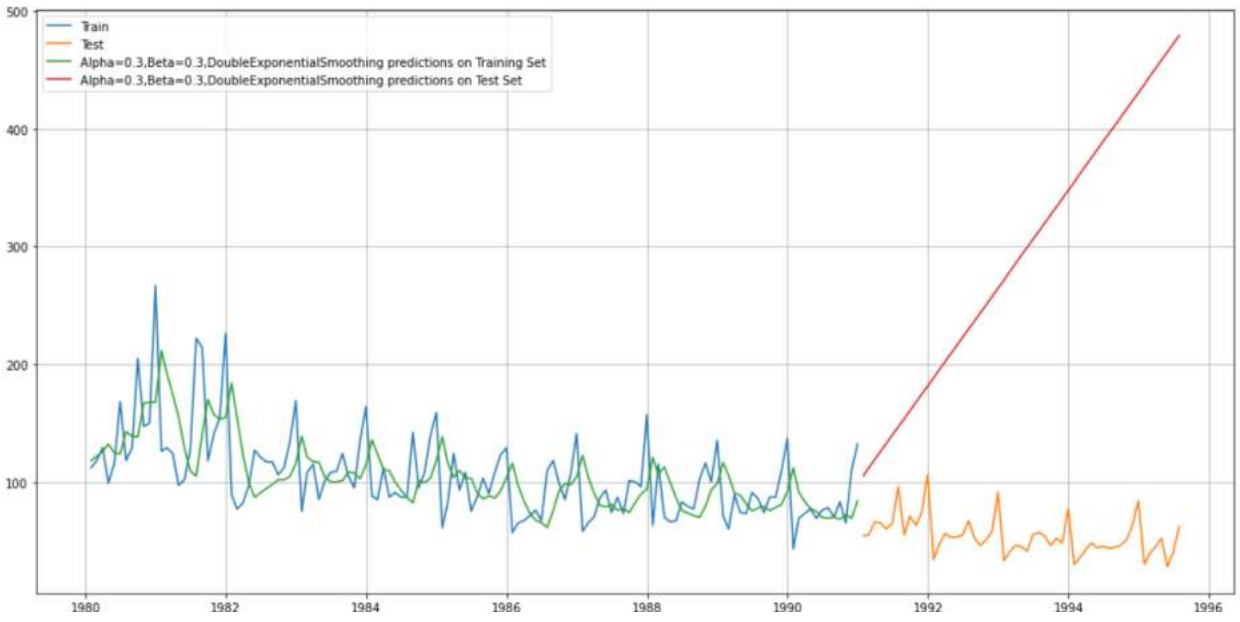


Alpha and beta values with 0.3 is found to be the best among all. The RMSE and MAPE values are mentioned below.

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	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
Alpha= 0.49,SimpleExponentialSmoothing	1316.035487	45.47
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	75.66
Alpha=0.4,SimpleExponentialSmoothing	2311.919615	91.55
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	675.28

Rose Dataset

- The test RMSE and MAPE values of rose dataset are 265.6 and 443 approximately. They are shown in the below table.



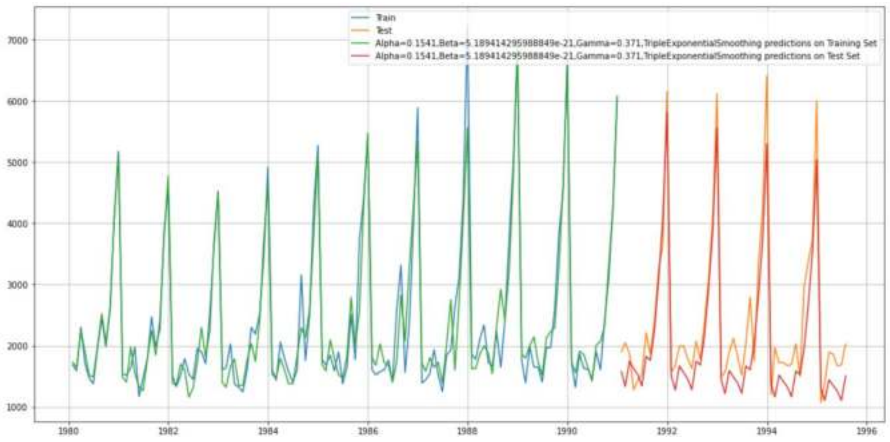
	Test RMSE	Test MAPE
RegressionOnTime	15.291460	22.94
NaiveModel	79.778066	145.35
SimpleAverageModel	53.521557	95.13
2pointTrailingMovingAverage	11.530180	13.60
4pointTrailingMovingAverage	14.462330	19.59
6pointTrailingMovingAverage	14.586916	20.83
9pointTrailingMovingAverage	14.740112	21.13
Alpha=1,SimpleExponentialSmoothing	36.858571	64.05
Alpha=0.3,SimpleExponentialSmoothing	47.566302	83.89
Alpha=0.4,SimpleExponentialSmoothing	53.828369	95.70
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.639892	443.04

Model 7: Triple exponential Smoothing Model

Sparkling Dataset

- Auto fit model returned with the following smoothing parameters for this dataset

```
{'smoothing_level': 0.111108139467838,
 'smoothing_trend': 0.06172875597197263,
 'smoothing_seasonal': 0.3950479631147446,
 'damping_trend': nan,
 'initial_level': 1639.9340657558994,
 'initial_trend': -12.22494561218149,
 'initial_seasons': array([1.06402008, 1.02352078, 1.40671876, 1.20165543, 0.97593   ,
       0.97100155, 1.31897446, 1.69588922, 1.3895294 , 1.81476396,
       2.85150039, 3.62470528]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```



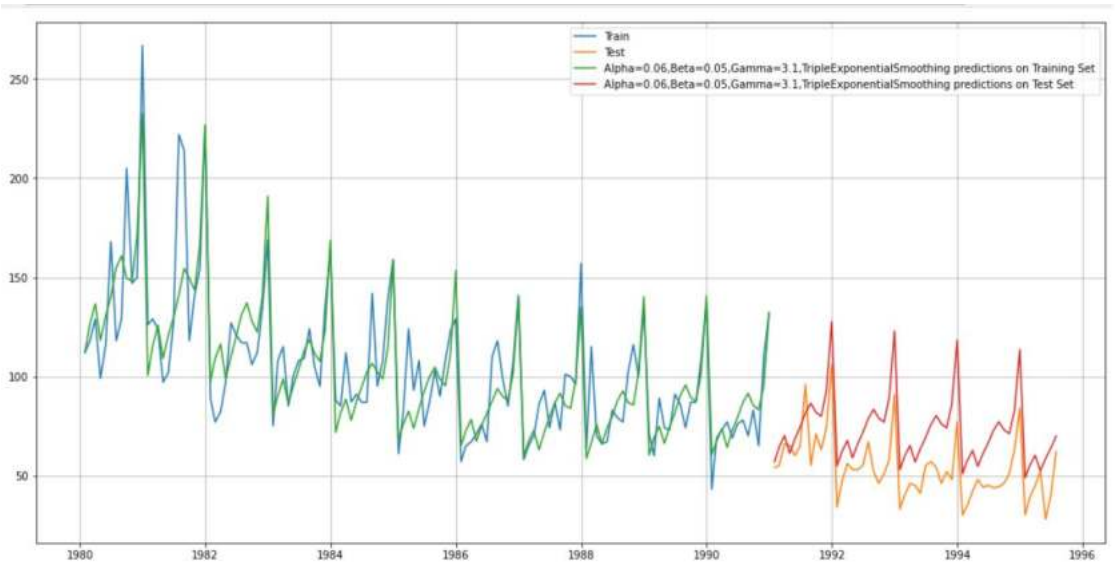
	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
Alpha= 0.49,SimpleExponentialSmoothing	1316.035487	45.47
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	75.66
Alpha=0.4,SimpleExponentialSmoothing	2311.919615	91.55
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	675.28
Alpha: 0.1,Beta:0.06 and Gamma: 0.4,TripleExponentialSmoothing	469.767970	16.40

The model chosen has the following parameters : alpha = 0.1, beta = 0.06, gamma = 0.4 and RMSE score of 21.45

Rose Dataset

- Auto fit model returned with the following smoothing parameters for this dataset

```
{'smoothing_level': 0.06280372101991354,
 'smoothing_trend': 0.05568813542468586,
 'smoothing_seasonal': 3.115268099923303e-06,
 'damping_trend': nan,
 'initial_level': 59.29348777160217,
 'initial_trend': -0.3727281817131398,
 'initial_seasons': array([1.90362937, 2.16072498, 2.36051494, 2.06290154, 2.31908662,
       2.52928387, 2.77978951, 2.95507465, 2.80602228, 2.74429704,
       3.19876277, 4.41211721]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```



	Test RMSE	Test MAPE
RegressionOnTime	15.291460	22.94
NaiveModel	79.778066	145.35
SimpleAverageModel	53.521557	95.13
2pointTrailingMovingAverage	11.530180	13.60
4pointTrailingMovingAverage	14.462330	19.59
6pointTrailingMovingAverage	14.586916	20.83
9pointTrailingMovingAverage	14.740112	21.13
Alpha=1,SimpleExponentialSmoothing	36.858571	64.05
Alpha=0.3,SimpleExponentialSmoothing	47.566302	83.89
Alpha=0.4,SimpleExponentialSmoothing	53.828369	95.70
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.639892	443.04
Alpha = 0.06 ,Beta = 0.05,Gamma = 3.1,TripleExponentialSmoothing	21.458971	35.96

The model chosen has the following parameters : alpha = 0.06, beta = 0.05, gamma = 3.1 and RMSE score of 21.45

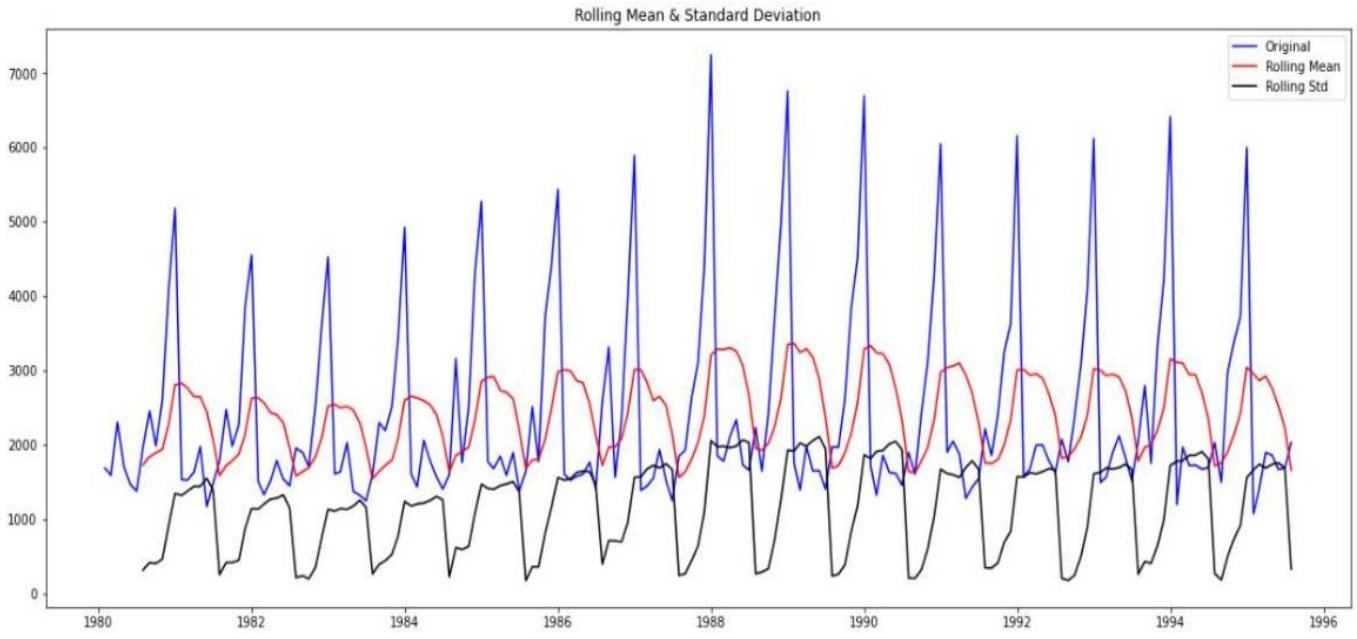
Q5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

SPARKLING DATASET

- For models such as ARIMA and SARIMA to be able to predict better values, the time series must be stationary. As mentioned in the question, we have to perform appropriate steps to make it stationary in case the time series are not stationary.
- Dickey Fuller test is the statistical test used here to check the stationarity of time series.

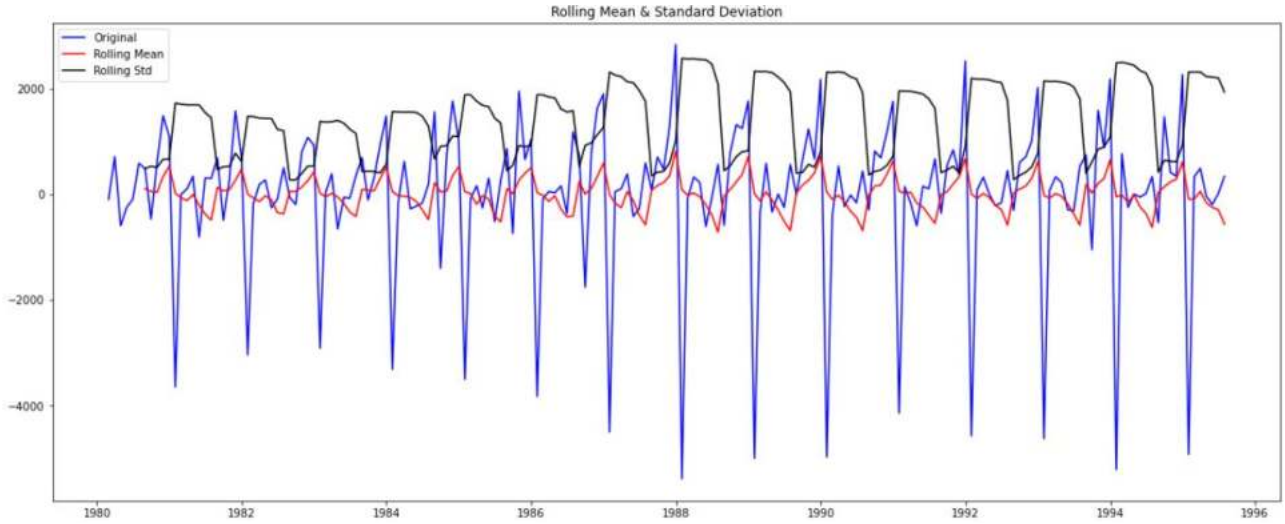
Hypothesis statements for the augmented Dickey Fuller test are mentioned below:

1. **Null Hypothesis:** The Time series is non-stationary.
2. **Alternate Hypothesis:** The Time series is stationary.



```
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
```

- From the above tests, the time series clearly non stationary, therefore differencing is done to the time series



- As seen below, p-value in differenced series is less than 0.05, therefore we can reject the null hypothesis and the time series is stationary.

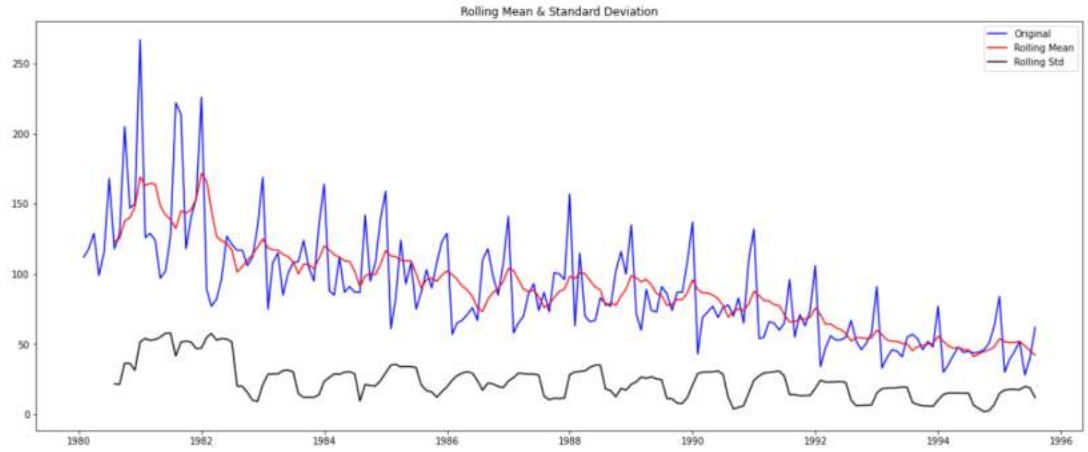
```
Results of Dickey-Fuller Test:
Test Statistic                -45.050301
p-value                         0.000000
#Lags Used                     10.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)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

ROSE DATASET

- For models such as ARIMA and SARIMA to be able to predict better values, the time series must be stationary. As mentioned in the question, we have to perform appropriate steps to make it stationary in case the time series are not stationary.
- Dickey Fuller test is the statistical test used here to check the stationarity of time series.

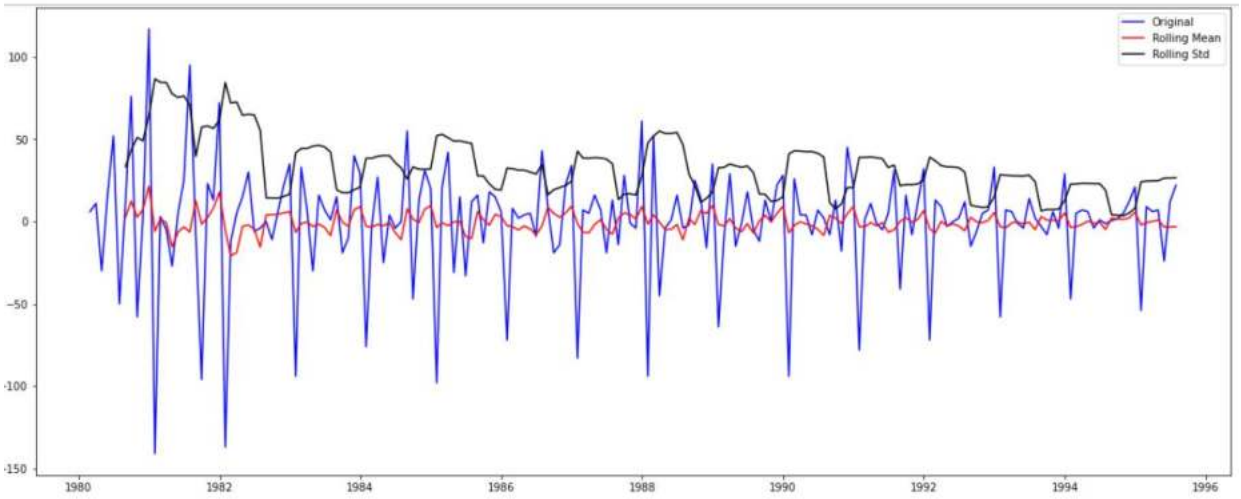
Hypothesis statements for the augmented Dickey Fuller test are mentioned below:

1. **Null Hypothesis:** The Time series is non-stationary.
2. **Alternate Hypothesis:** The Time series is stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.873273
p-value                          0.344737
#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
```

- From the above tests, the time series clearly non stationary, therefore differencing is done to the time series and the plot after differencing of order 1 is done to the time series is shown below



- As seen below, p-value in differenced series is less than 0.05, therefore we can reject the null hypothesis and the time series is stationary.

```
Results of Dickey-Fuller Test:
Test Statistic                -8.043385e+00
p-value                        1.821604e-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
```

Q6. 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 ARIMA

SPARKLING DATASET

- The optimal parameters for (p, d, q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.
- For the ARIMA model the minimum AIC value is observed for ARIMA(2,1,2) model.

	param	AIC
8	(2, 1, 2)	2210.618988
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906437
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

- The ARIMA model (2,1,2) is applied and the summary of the model results can be found below:

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.309
Method:	css-mle	S.D. of innovations	1012.929
Date:	Fri, 26 Feb 2021	AIC	2210.619
Time:	19:09:03	BIC	2227.870
Sample:	02-29-1980 - 12-31-1990	HQIC	2217.629

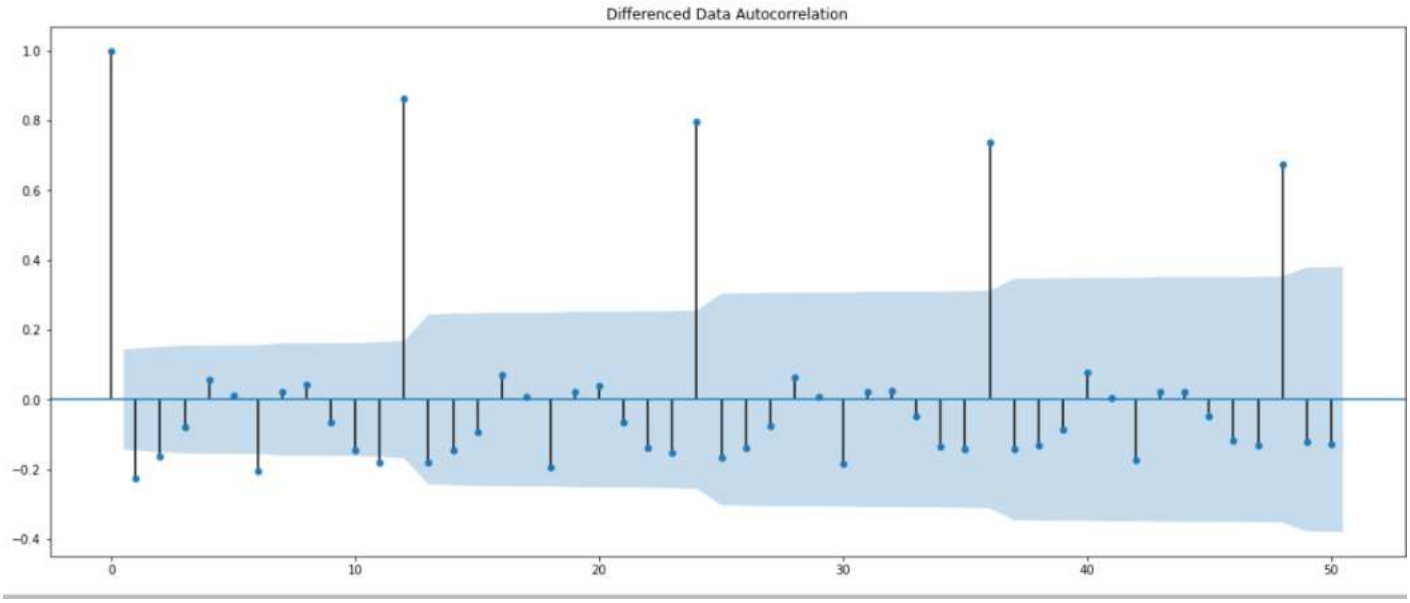
	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5854	0.517	10.806	0.000	4.572	6.598
ar.L1.D.Sparkling	1.2699	0.075	17.046	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5602	0.074	-7.618	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-1.9974	0.042	-47.112	0.000	-2.080	-1.914
ma.L2.D.Sparkling	0.9974	0.042	23.497	0.000	0.914	1.081

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1334	-0.7074j	1.3361	-0.0888
AR.2	1.1334	+0.7074j	1.3361	0.0888
MA.1	1.0006	+0.0000j	1.0006	0.0000

- The **RMSE score** of auto ARIMA is **<u>1374.48</u>**

	RMSE
ARIMA(2,1,2)	1374.484105



ROSE DATASET

- The optimal parameters for (p, d, q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.
- For the ARIMA model the minimum AIC value is observed for ARIMA(0,1,2) model.

	param	AIC
2	(0, 1, 2)	1276.835372
5	(1, 1, 2)	1277.359223
4	(1, 1, 1)	1277.775749
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

- The ARIMA model (0,1,2) is applied and the summary of the model results can be found below:

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(0, 1, 2)	Log Likelihood	-634.418
Method:	css-mle	S.D. of innovations	30.167
Date:	Sat, 27 Feb 2021	AIC	1276.835
Time:	13:21:54	BIC	1288.336
Sample:	02-29-1980 - 12-31-1990	HQIC	1281.509

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4885	0.085	-5.742	0.000	-0.655	-0.322
ma.L1.D.Rose	-0.7601	0.101	-7.499	0.000	-0.959	-0.561
ma.L2.D.Rose	-0.2398	0.095	-2.518	0.012	-0.427	-0.053

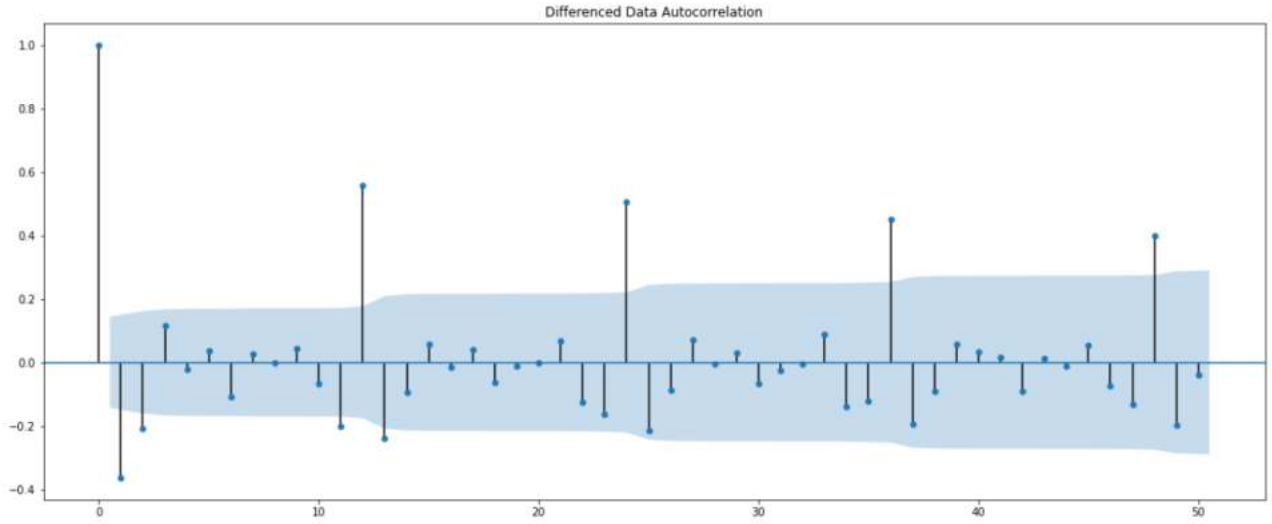
Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-4.1695	+0.0000j	4.1695	0.5000

- The **RMSE score** of auto ARIMA is **15.64**

RMSE

ARIMA(0,1,2) 15.640546



AUTO SARIMA

SPARKLING DATASET

- The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.
- For the SARIMA model the minimum AIC value is observed for SARIMA (1,1,2) x (2,0,2,6) model.

```
SARIMA_AIC.sort_values(by=['AIC']).head()
```

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.678698
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888814
17	(0, 1, 1)	(2, 0, 2, 6)	1741.703707
44	(1, 1, 1)	(2, 0, 2, 6)	1743.330541
71	(2, 1, 1)	(2, 0, 2, 6)	1744.040751

- The ARIMA model (1,1,2) x (2,0,2,6) is applied and the summary of the model results can be found below:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-855.839
Date:	Fri, 26 Feb 2021	AIC	1727.679
Time:	19:09:37	BIC	1749.707
Sample:	0	HQIC	1736.621
	- 132		
Covariance Type:	opg		

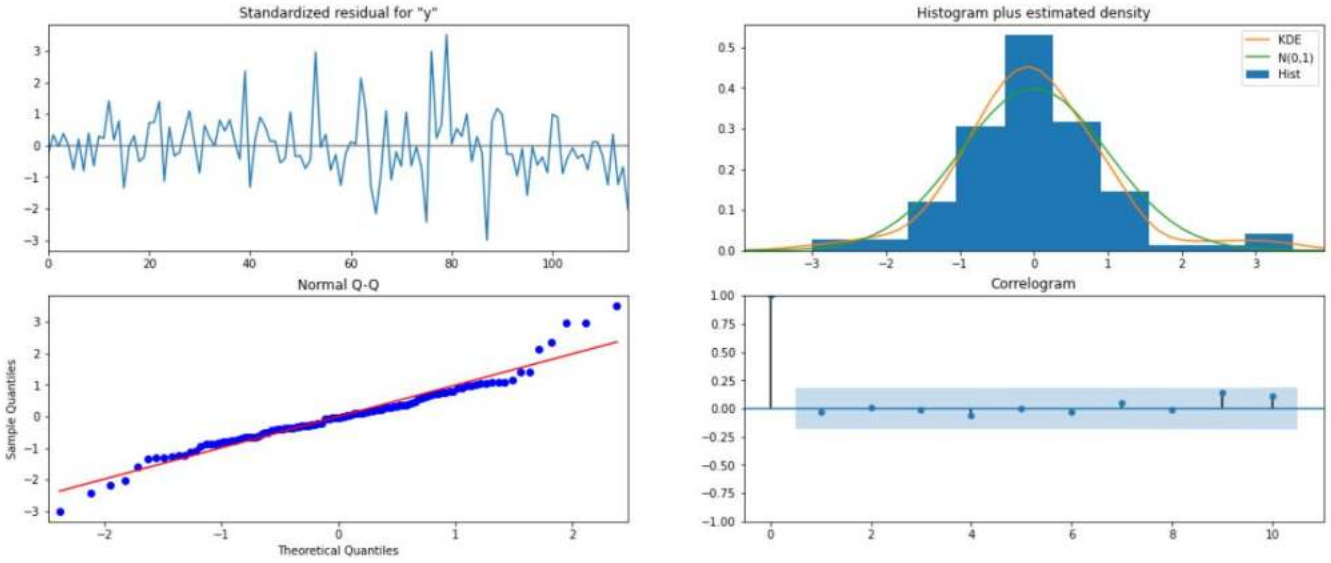
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6449	0.286	-2.256	0.024	-1.205	-0.085
ma.L1	-0.1068	0.250	-0.428	0.669	-0.596	0.383
ma.L2	-0.7006	0.202	-3.471	0.001	-1.096	-0.305
ar.S.L6	-0.0045	0.027	-0.165	0.869	-0.057	0.049
ar.S.L12	1.0361	0.018	56.082	0.000	1.000	1.072
ma.S.L6	0.0676	0.152	0.444	0.657	-0.231	0.366
ma.S.L12	-0.6123	0.093	-6.590	0.000	-0.794	-0.430
sigma2	1.448e+05	1.71e+04	8.466	0.000	1.11e+05	1.78e+05

Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	25.24
Prob(Q):	0.77	Prob(JB):	0.00
Heteroskedasticity (H):	2.63	Skew:	0.47
Prob(H) (two-sided):	0.00	Kurtosis:	5.09

- The **RMSE score** of auto ARIMA is **626.9**

	RMSE
ARIMA(2,1,2)	1374.484105
SARIMA(0,1,2)(2,0,2,6)	626.931138

- Diagnostic plot is shown below



ROSE DATASET

- The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.
- For the ARIMA model the minimum AIC value is observed for SARIMA(1,1,2) x(2,0,2,6) model.

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655818
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.220571
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

- The SARIMA(1,1,2) x(2,0,2,6) is applied and the summary of the model results can be found below:

SARIMAX Results

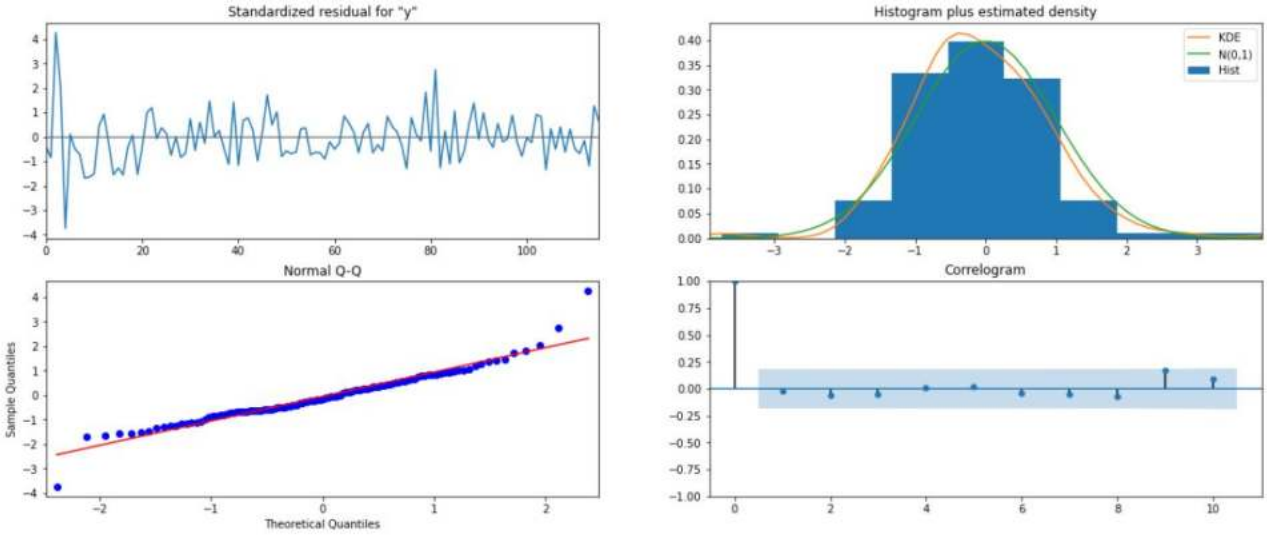
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-512.828
Date:	Sat, 27 Feb 2021	AIC	1041.656
Time:	15:20:14	BIC	1063.685
Sample:	0	HQIC	1050.598
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5939	0.152	-3.899	0.000	-0.892	-0.295
ma.L1	-0.1954	799.740	-0.000	1.000	-1567.656	1567.265
ma.L2	-0.8046	643.516	-0.001	0.999	-1262.072	1260.463
ar.S.L6	-0.0626	0.035	-1.764	0.078	-0.132	0.007
ar.S.L12	0.8451	0.039	21.883	0.000	0.769	0.921
ma.S.L6	0.2226	532.437	0.000	1.000	-1043.334	1043.780
ma.S.L12	-0.7774	413.876	-0.002	0.999	-811.960	810.406
sigma2	335.1994	3.09e+05	0.001	0.999	-6.04e+05	6.05e+05

Ljung-Box (L1) (Q):	0.07	Jarque-Bera (JB):	56.68
Prob(Q):	0.78	Prob(JB):	0.00
Heteroskedasticity (H):	0.47	Skew:	0.52
Prob(H) (two-sided):	0.02	Kurtosis:	6.26

- The **RMSE score** of auto ARIMA is **26.2**

	RMSE
ARIMA(0,1,2)	15.640546
SARIMA(1,1,2)(2,0,2,6)	26.209519

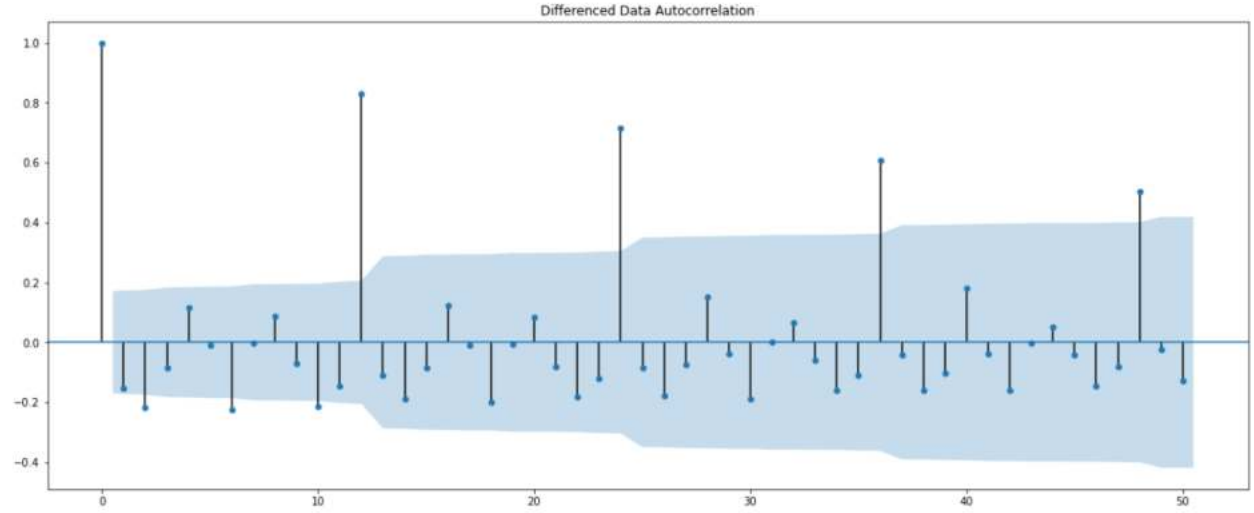


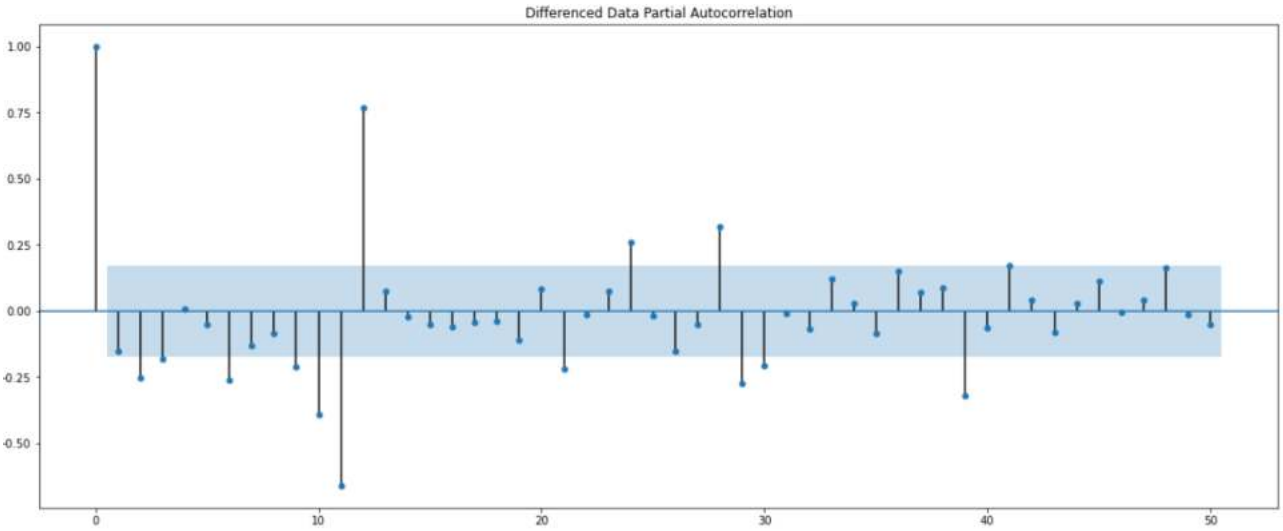
<u>NOTE: There has been no change in RMSE value by using the seasonal parameters and RMSE of ARIMA model for both the datasets is found to be the lowest.</u>

Q7. 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.

After observing both the below plots, ARIMA model (2,1,2) is chosen.

- MA (q) term is selected as 2 due to the significant lags present till lag-2 from acf plot and the same is observed from PACF plot, therefore AR(q) term is chosen as 2.





- The summary results of the model is shown below:

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.309
Method:	css-mle	S.D. of innovations	1012.929
Date:	Fri, 26 Feb 2021	AIC	2210.619
Time:	19:10:28	BIC	2227.870
Sample:	02-29-1980 - 12-31-1990	HQIC	2217.629

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5854	0.517	10.806	0.000	4.572	6.598
ar.L1.D.Sparkling	1.2699	0.075	17.046	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5602	0.074	-7.618	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-1.9974	0.042	-47.112	0.000	-2.080	-1.914
ma.L2.D.Sparkling	0.9974	0.042	23.497	0.000	0.914	1.081

Roots

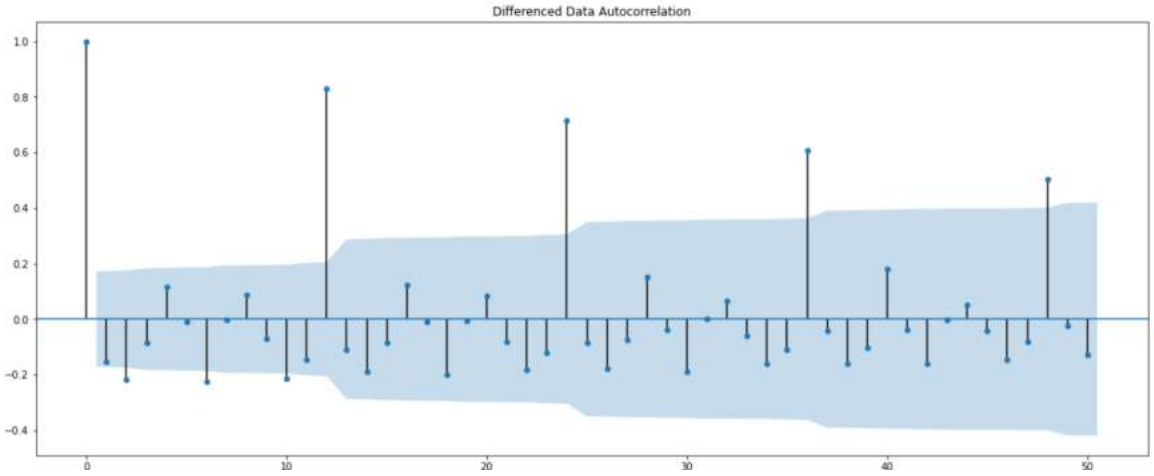
	Real	Imaginary	Modulus	Frequency
AR.1	1.1334	-0.7074j	1.3361	-0.0888
AR.2	1.1334	+0.7074j	1.3361	0.0888
MA.1	1.0006	+0.0000j	1.0006	0.0000
MA.2	1.0020	+0.0000j	1.0020	0.0000

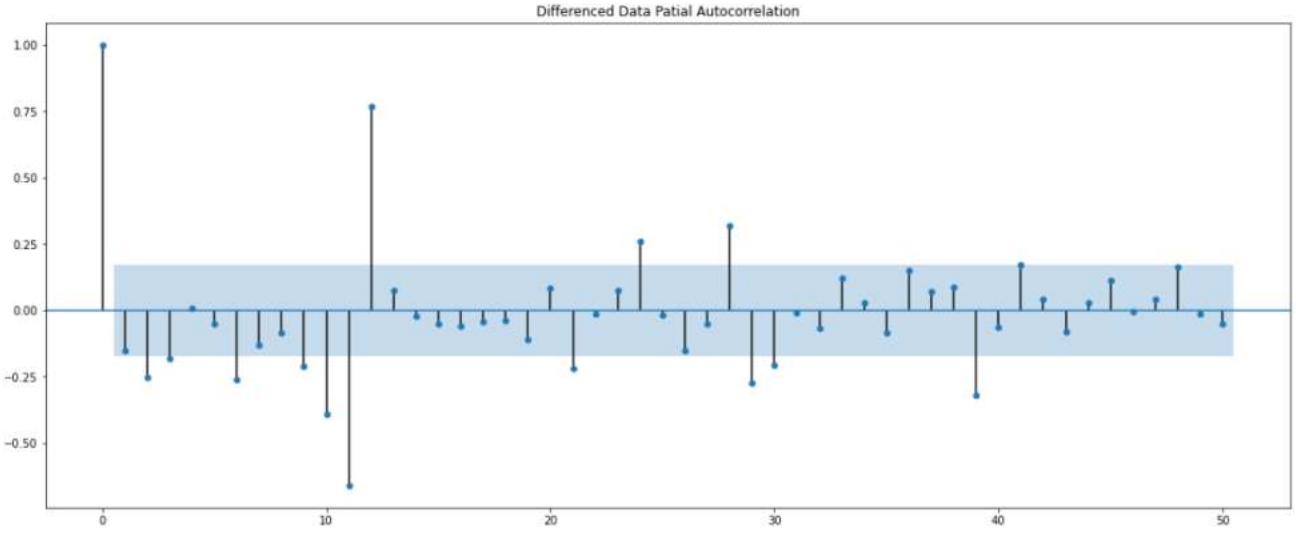
The time series data has high seasonality and using ARIMA model is not deemed to be appropriate for this scenario, therefore we will build manual SARIMA model and evaluate it on test data using RMSE.

- The RMSE value on Test for the ARIMA(2,1,2) model is 1374.

Manual SARIMA model

- From ACF plot, lags we observe that up to lag 2 it is significant. Thus we can consider AR(p) term as 2.
- From PACF plot, we observe 3 significant lags and P term is 3.





- The model (2,1,2) x (3,1,1,12) has been applied.

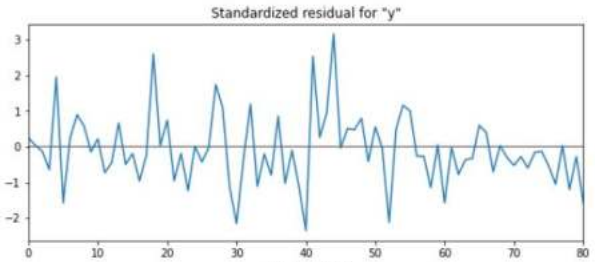
SARIMAX Results

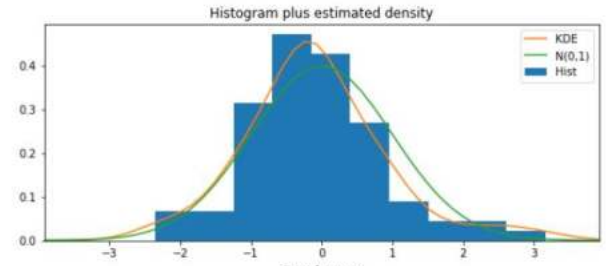
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(3, 1, [1], 12)	Log Likelihood	-606.105
Date:	Sat, 27 Feb 2021	AIC	1230.209
Time:	15:49:58	BIC	1251.759
Sample:	0	HQIC	1238.856
	- 132		
Covariance Type:	opg		

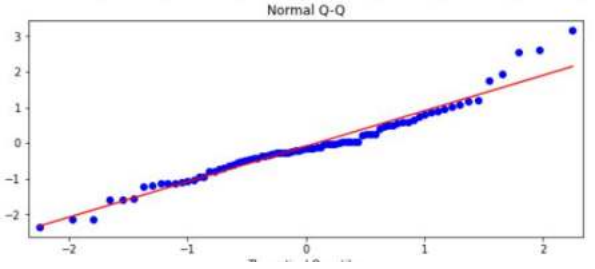
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5701	0.339	-1.682	0.092	-1.234	0.094
ar.L2	0.0334	0.185	0.181	0.856	-0.328	0.395
ma.L1	-0.1571	0.311	-0.506	0.613	-0.766	0.452
ma.L2	-0.7606	0.281	-2.702	0.007	-1.312	-0.209
ar.S.L12	-0.5672	0.967	-0.587	0.557	-2.462	1.327
ar.S.L24	-0.2763	0.372	-0.742	0.458	-1.006	0.453
ar.S.L36	-0.1314	0.176	-0.749	0.454	-0.476	0.213
ma.S.L12	0.1514	0.972	0.156	0.876	-1.754	2.057
sigma2	1.844e+05	3.17e+04	5.823	0.000	1.22e+05	2.46e+05

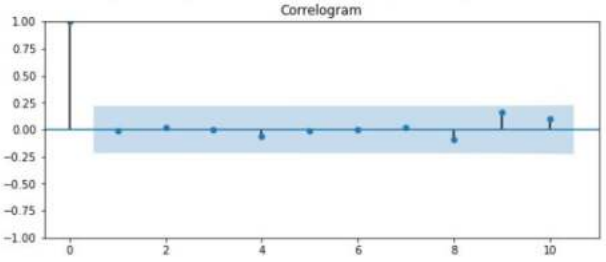
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	10.97
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.71	Skew:	0.62
Prob(H) (two-sided):	0.37	Kurtosis:	4.31

- Diagnostic plots are shown below.







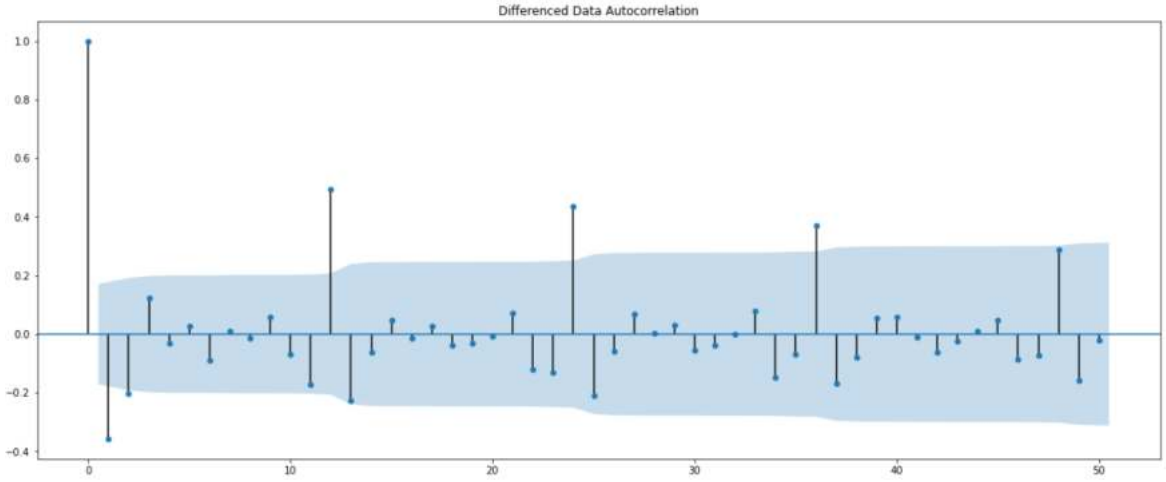


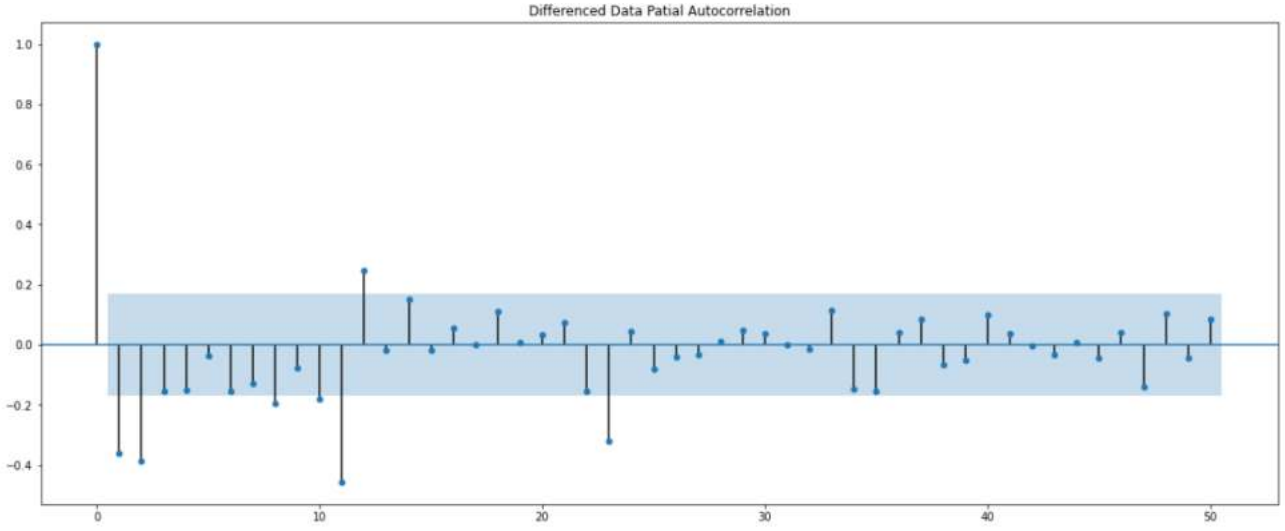
ROSE Dataset

ARIMA Model

After observing both the below plots, ARIMA model (2,1,2) is chosen.

- MA (q) term is selected as 2 due to the significant lags present till lag-2 from acf plot and the same is observed from PACF plot, therefore AR(q) term is chosen as 2.





- The summary results of the model is shown below:

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-633.649
Method:	css-mle	S.D. of innovations	29.975
Date:	Sat, 27 Feb 2021	AIC	1279.299
Time:	15:20:52	BIC	1296.550
Sample:	02-29-1980 - 12-31-1990	HQIC	1286.309

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4911	0.081	-6.076	0.000	-0.649	-0.333
ar.L1.D.Rose	-0.4383	0.218	-2.015	0.044	-0.865	-0.012
ar.L2.D.Rose	0.0269	0.109	0.246	0.806	-0.188	0.241
ma.L1.D.Rose	-0.3316	0.203	-1.633	0.102	-0.729	0.066
ma.L2.D.Rose	-0.6684	0.201	-3.332	0.001	-1.062	-0.275

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-2.0289	+0.0000j	2.0289	0.5000
AR.2	18.3387	+0.0000j	18.3387	0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.4960	+0.0000j	1.4960	0.5000

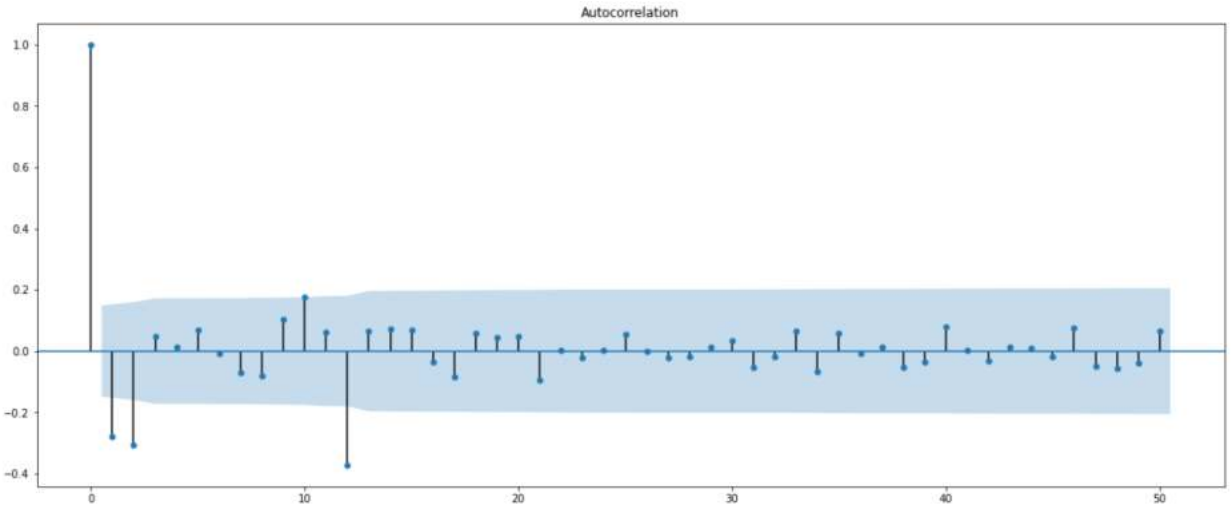
The time series data has high seasonality and using ARIMA model is not deemed to be appropriate for this scenario, therefore we will build manual SARIMA model and evaluate it on test data using RMSE.

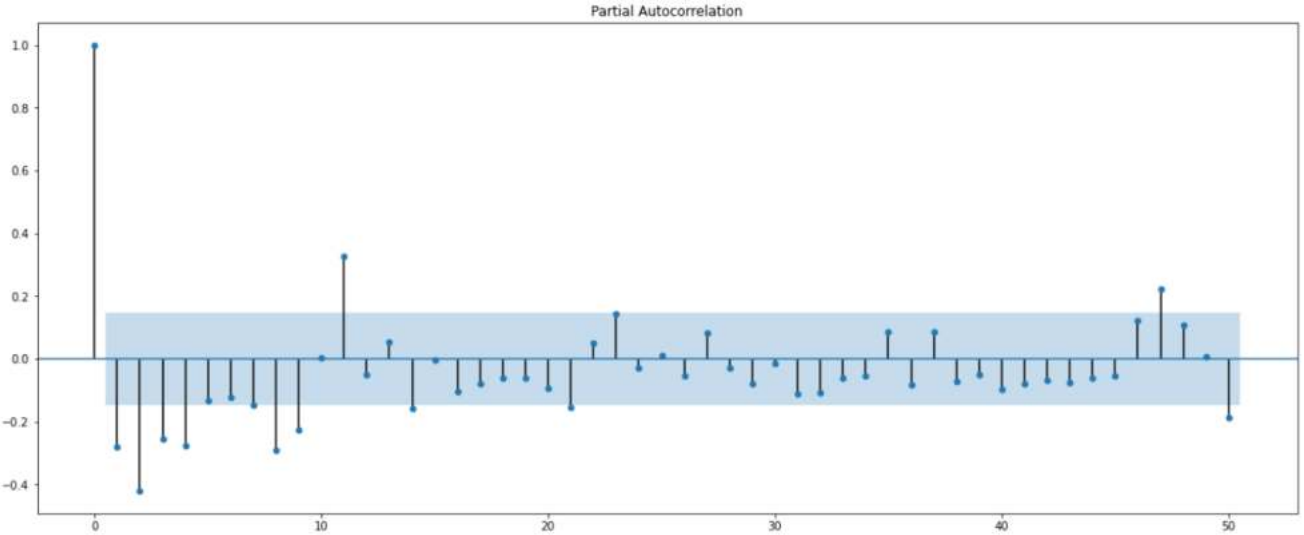
- The RMSE value on Test for the ARIMA(2,1,2) model is 15.37.

ARIMA(2,1,2)	15.376555

Manual SARIMA model

- From ACF plot, lags we observe that up to lag 2 it is significant. Thus we can consider AR(p) term as 2.
- From PACF plot, we observe 4 significant lags and P term is 4.





- The model (2,1,2) x (4,1,2,12) has been applied.

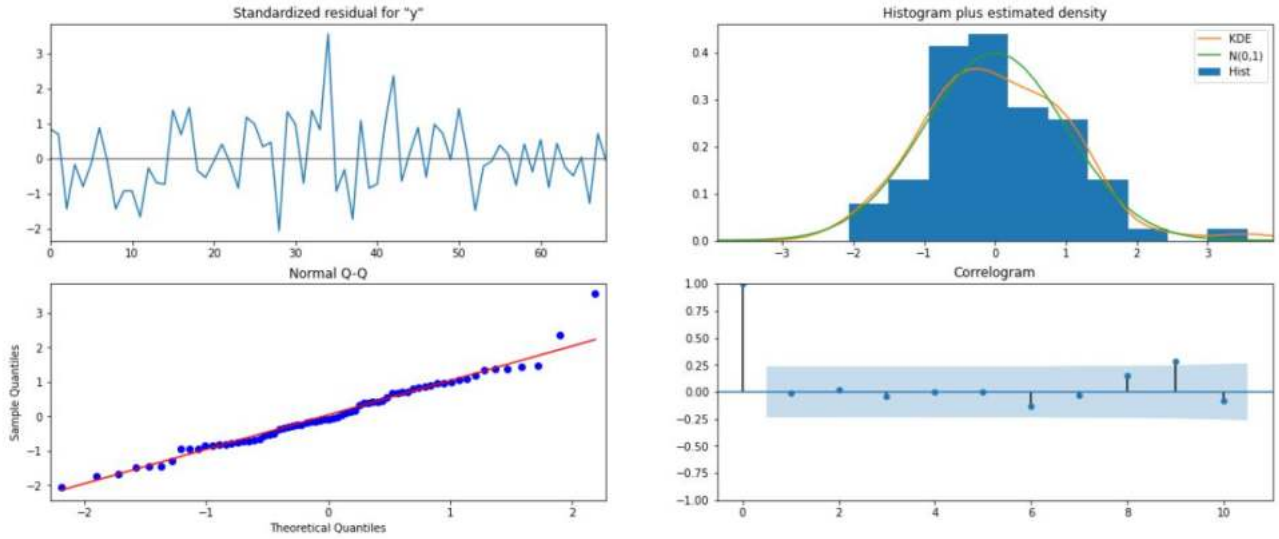
SARIMAX Results

Dep. Variable:		y	No. Observations:		132	
Model:	SARIMAX(2, 1, 2)x(4, 1, 2, 12)		Log Likelihood		-284.472	
Date:	Sat, 27 Feb 2021		AIC		590.945	
Time:	15:21:13		BIC		615.520	
Sample:	0		HQIC		600.695	
	- 132					
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9798	0.224	-4.366	0.000	-1.420	-0.540
ar.L2	-0.1273	0.143	-0.891	0.373	-0.407	0.153
ma.L1	0.0208	0.247	0.084	0.933	-0.463	0.505
ma.L2	-0.8825	0.193	-4.575	0.000	-1.261	-0.504
ar.S.L12	-0.7355	0.198	-3.706	0.000	-1.125	-0.346
ar.S.L24	-0.0736	0.174	-0.423	0.673	-0.415	0.268
ar.S.L36	0.0758	0.088	0.859	0.390	-0.097	0.249
ar.S.L48	-0.0064	0.021	-0.309	0.758	-0.047	0.034
ma.S.L12	-0.3533	0.696	-0.508	0.612	-1.717	1.010
ma.S.L24	-0.9038	0.558	-1.619	0.105	-1.998	0.190
sigma2	144.5357	109.831	1.316	0.188	-70.730	359.801

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	6.01
Prob(Q):	0.91	Prob(JB):	0.05
Heteroskedasticity (H):	0.62	Skew:	0.53
Prob(H) (two-sided):	0.25	Kurtosis:	3.98

- Diagnostic plots are shown below.



Q8.

SPARKLING DATASET

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	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
Alpha= 0.49,SimpleExponentialSmoothing	1316.035487	45.47
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	75.66
Alpha=0.4,SimpleExponentialSmoothing	2311.919615	91.55
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	675.28
Alpha: 0.1,Beta:0.06 and Gamma: 0.4,TripleExponentialSmoothing	469.767970	16.40

	RMSE
ARIMA(2,1,2)	1374.484105
SARIMA(0,1,2)(2,0,2,6)	626.931138
SARIMA(1,1,2)(2,0,2,12)	528.566817
ARIMA(2,1,2)	1374.484105
SARIMA(2,1,2)(3, 1, 1, 12)	337.358722

ROSE DATASET

	Test RMSE	Test MAPE
RegressionOnTime	15.291460	22.94
NaiveModel	79.778066	145.35
SimpleAverageModel	53.521557	95.13
2pointTrailingMovingAverage	11.530180	13.60
4pointTrailingMovingAverage	14.462330	19.59
6pointTrailingMovingAverage	14.586916	20.83
9pointTrailingMovingAverage	14.740112	21.13
Alpha=1,SimpleExponentialSmoothing	36.858571	64.05
Alpha=0.3,SimpleExponentialSmoothing	47.566302	83.89
Alpha=0.4,SimpleExponentialSmoothing	53.828369	95.70
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.639892	443.04
Alpha = 0 0.106,Beta = 0.048,Gamma = 0.0,TripleExponentialSmoothing	21.458971	35.96

:

	RMSE
ARIMA(0,1,2)	15.640546
SARIMA(1,1,2)(2,0,2,6)	26.209519
SARIMA(0,1,2)(2,0,2,12)	26.992038
ARIMA(2,1,2)	15.376555
SARIMA(2,1,2)(4, 1, 2, 12)	17.399706

Q9.

Sparkling dataset

Looking at the dataframes of all the model scores of both the datasets,

SARIMA (2,1,2) x (3,1,1,12) and SARIMA (2,1,2) x (4,1,2,12) models are chosen to be the optimal models for Sparkling and Rose datasets respectively.

We need to take the entire data to build a model using the SARIMA (2,1,2) x (3,1,1,12) and SARIMA (2,1,2) x (4,1,2,12) models and then predict 12 months into the future with a confidence interval of 95%.

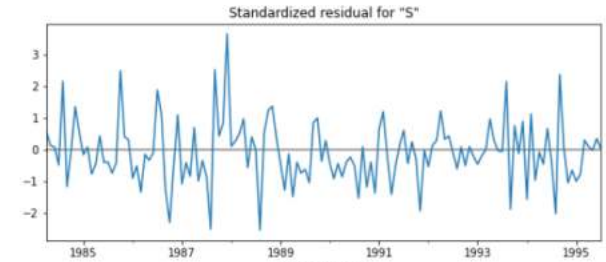
- RMSE of the Model 581.8504815891278

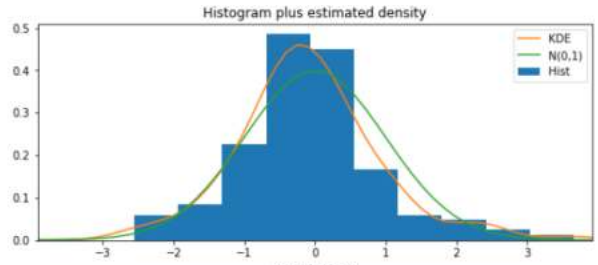
Model Diagnostics

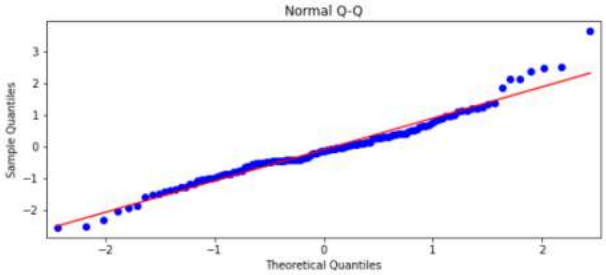
Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(2, 1, 2)x(3, 1, [1], 12)	Log Likelihood	-1006.928
Date:	Sat, 27 Feb 2021	AIC	2031.855
Time:	19:33:29	BIC	2058.069
Sample:	01-31-1980	HQIC	2042.508
	- 07-31-1995		
Covariance Type:	opg		

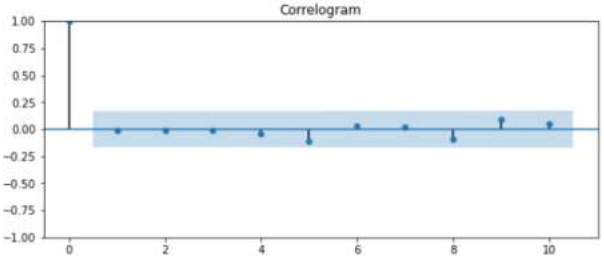
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7443	0.252	-2.951	0.003	-1.239	-0.250
ar.L2	0.0154	0.127	0.121	0.904	-0.234	0.265
ma.L1	-0.0996	0.251	-0.396	0.692	-0.592	0.393
ma.L2	-0.8112	0.246	-3.292	0.001	-1.294	-0.328
ar.S.L12	0.2657	0.080	3.312	0.001	0.108	0.423
ar.S.L24	0.1024	0.100	1.025	0.305	-0.093	0.298
ar.S.L36	0.0919	0.083	1.106	0.269	-0.071	0.255
ma.S.L12	-0.9981	0.119	-8.419	0.000	-1.230	-0.766
sigma2	1.338e+05	8.91e-07	1.5e+11	0.000	1.34e+05	1.34e+05

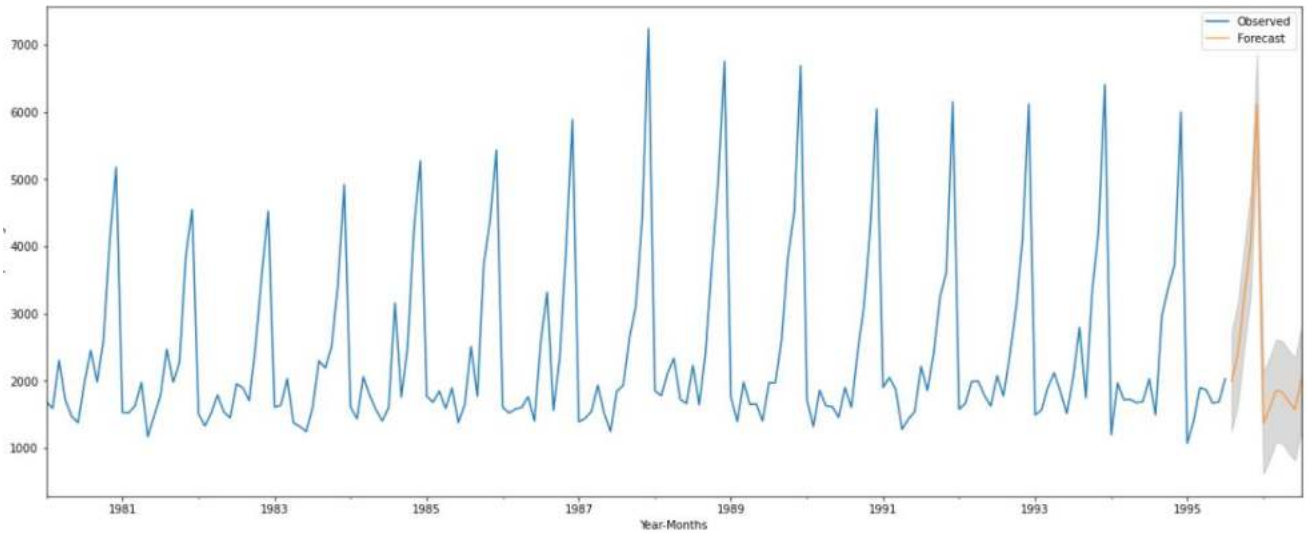
Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	19.75
Prob(Q):	0.86	Prob(JB):	0.00
Heteroskedasticity (H):	0.55	Skew:	0.53
Prob(H) (two-sided):	0.05	Kurtosis:	4.53


Standardized residual for "S"
3
2
1
0
−1
−2
1985
1987
1989
1991
1993
1995


Histogram plus estimated density
KDE
N(0,1)
Hist
0.5
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3


Normal Q-Q
Sample Quantiles
Theoretical Quantiles
3
2
1
0
−1
−2
−2
−1
0
1
2


Correlogram
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10


Observed
Forecast
7000
6000
5000
4000
3000
2000
1000
1981
1983
1985
1987
1989
1991
1993
1995
Year-Months

Rose Dataset

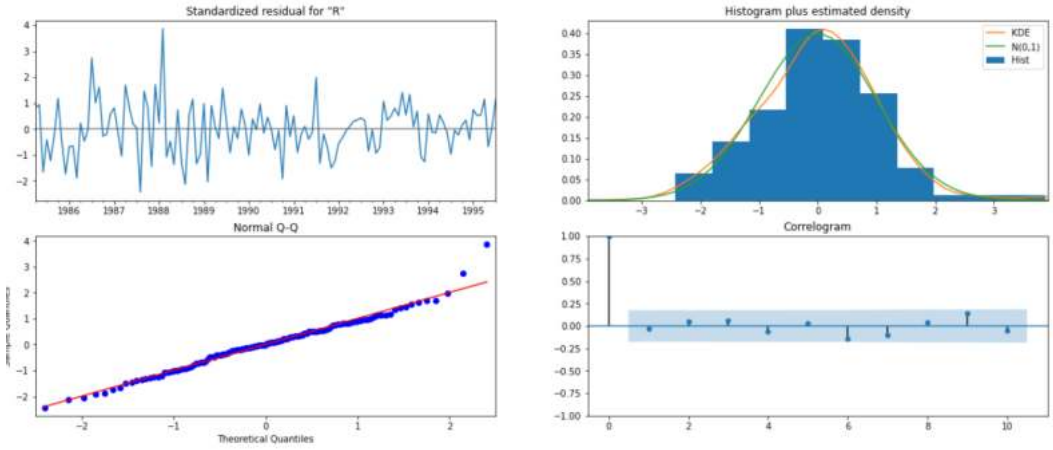
- RMSE of the Model 44.11735879055518

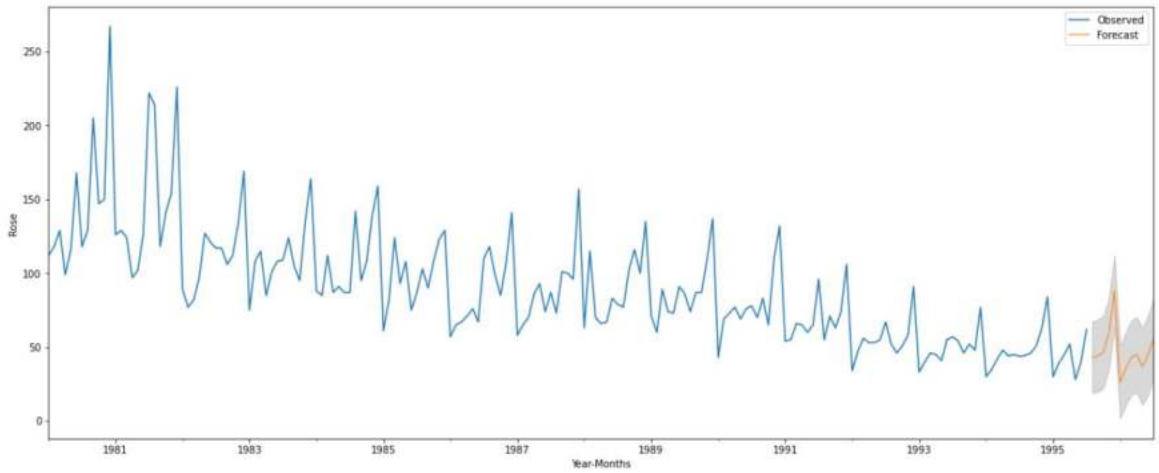
Model Diagnostics

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(2, 1, 2)x(4, 1, 2, 12)	Log Likelihood	-492.673
Date:	Sat, 27 Feb 2021	AIC	1007.346
Time:	19:56:40	BIC	1038.369
Sample:	01-31-1980	HQIC	1019.948
	- 07-31-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.0107	0.106	9.573	0.000	0.804	1.218
ar.L2	-0.1717	0.113	-1.521	0.128	-0.393	0.049
ma.L1	-1.9205	0.092	-20.872	0.000	-2.101	-1.740
ma.L2	0.9669	0.092	10.501	0.000	0.786	1.147
ar.S.L12	-0.7356	0.132	-5.558	0.000	-0.995	-0.476
ar.S.L24	-0.0332	0.154	-0.216	0.829	-0.335	0.269
ar.S.L36	-0.0258	0.096	-0.269	0.788	-0.214	0.162
ar.S.L48	-0.0176	0.032	-0.549	0.583	-0.080	0.045
ma.S.L12	0.0262	0.180	0.145	0.884	-0.327	0.380
ma.S.L24	-0.5248	0.180	-2.919	0.004	-0.877	-0.172
sigma2	154.5880	21.096	7.328	0.000	113.241	195.935

Ljung-Box (L1) (Q):	0.12	Jarque-Bera (JB):	8.49
Prob(Q):	0.73	Prob(JB):	0.01
Heteroskedasticity (H):	0.26	Skew:	0.30
Prob(H) (two-sided):	0.00	Kurtosis:	4.13

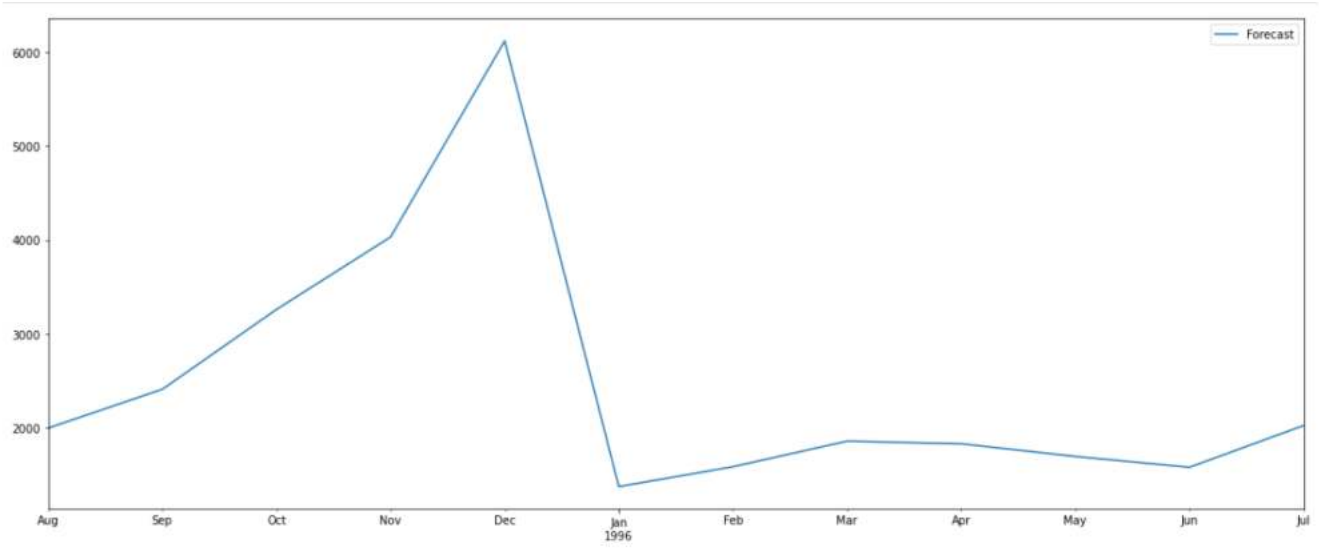



Observed
Forecast
250
200
150
100
50
0
Rose
1981
1983
1985
1987
1989
1991
1993
1995
Year-Months

Q10.

- It is clearly seen that the prediction suggests that there will be sharp increase in the month of December and suggests that month of December has higher sales.
- And right after December, there is a sharp decrease in the sales in the month of January and the sales in the following months after January are almost consistent.
- The wine factories should investigate the factors leading to low sales and take appropriate actions.
- For better model building, few variables that could provide better predictions should be researched and be provided for the analysis.

Sparkling wine dataset plot (1995 – 1996)



Rose wine dataset plot (1995 – 1996)