TIME FORECAST PROJECT

NAME – AMRITA JENA

BATCH – May_B 2021

COURSE – PGPDSBA

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

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

Monthly sales of two type 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 on the data as index

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

```
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')
```

Table-**<u>1- Time series reading for Sparkling and Rose</u>**

• The Sparkling time series has got no null values but Rose time-series has values missing for two months in 1994, which are imputed using interpolation (linear method)

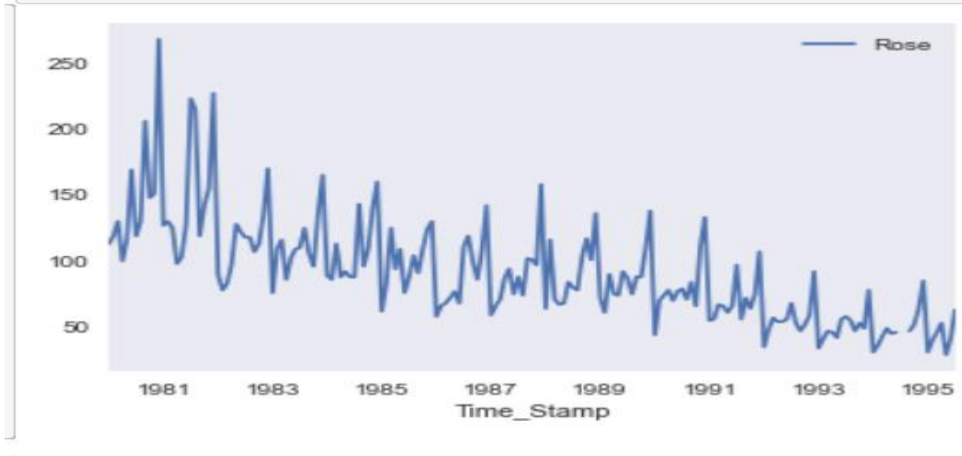
• Rose data after interpolation for year 1994 is given below as well as the plot

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

Time_Stamp	Rose
1994-07-31	NaN
1994-08-31	NaN

Time_Stamp	Rose
1994-07-31	45.333333
1994-08-31	45.666667

Table-2- Managing missing value in Rose data set. The missing values were imputed using linear interpolation mthod



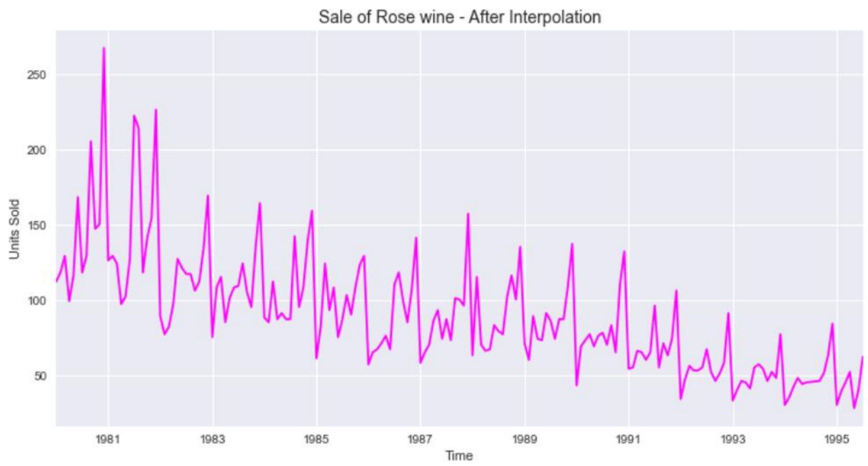


Figure-1- Time series plot before and after interpolation of Rose data set

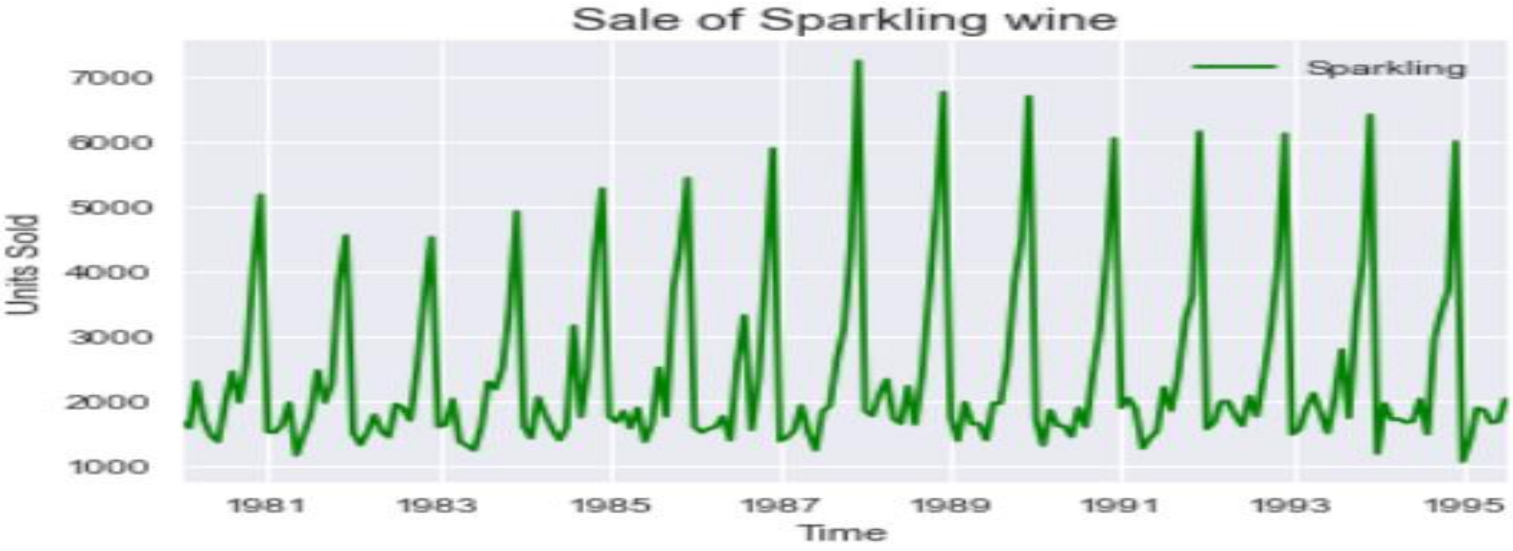


Figure-2- Time series plot of Sparkling data set

- Significant seasonality can be observed in both the datasets.
- While sale of Rose shows 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 had been fell out-of-favour over the year

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

SPARKLING:

- The descriptive analysis of the data shows that on an average 2402 units of Sparkling wines were sold each month on the given period of time. 50% of months 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

	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-3-Descriptive analysis

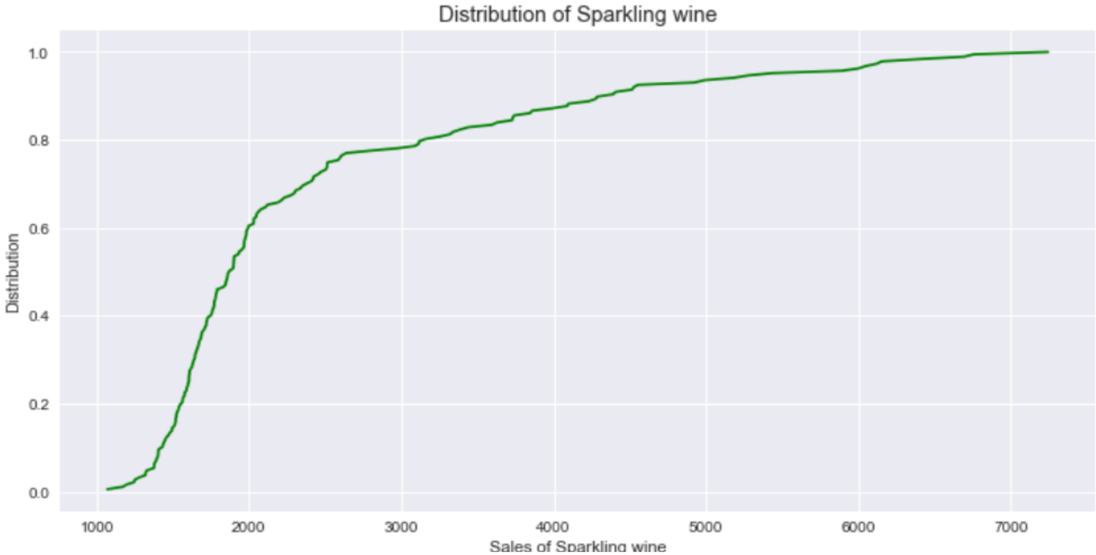
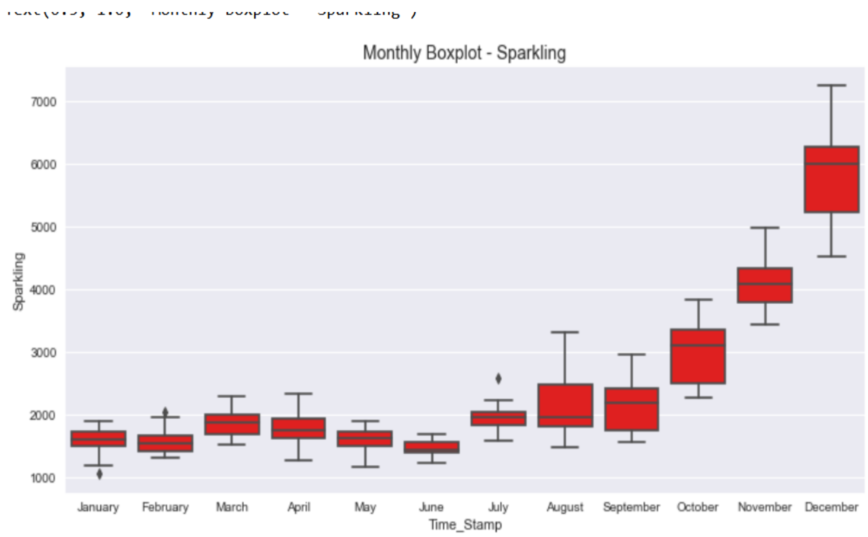
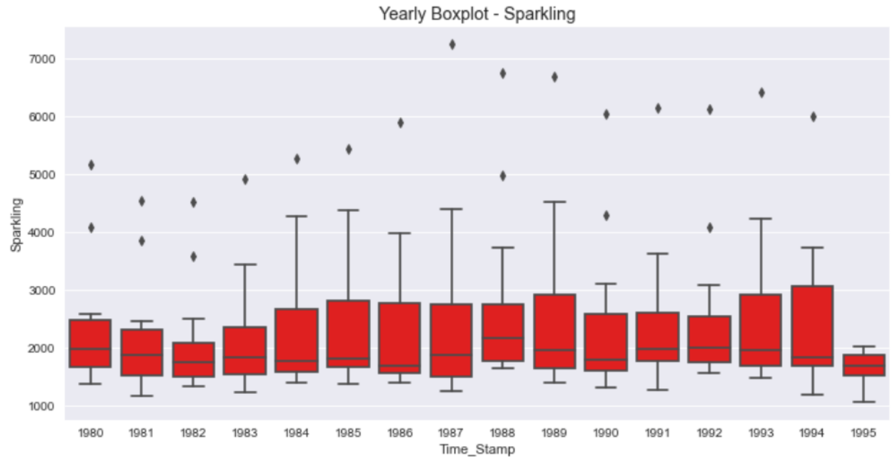


Figure-3-Emperical CDF PLOT





<u>FIGURE-4- Yearly-boxplot and Monthly-box-plot of Sparkling Sales</u>

The yearly-boxplot, 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.

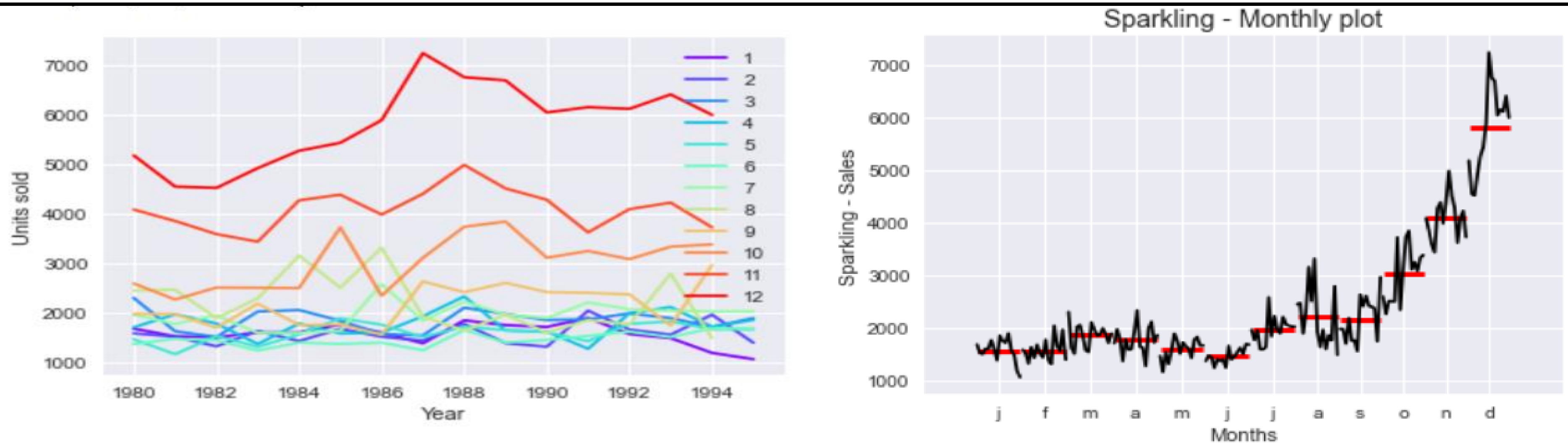


FIGURE-5- Yearly and Monthly plot of Sparkling Sales

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

Sale in December with a mean few point below 6000, varies from 7400 to 4500 units over the years. Whereas sale in November varies from 3500 units to 5000 units and sale in October varies from 2500 to 4000 units

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

The plot of monthly sale 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 sale was in 1981. Post 1987 December sales is around an average 6500 units, which was around 5000 in 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 is seen to be consistent across the years, compared to the rest of the months.

ROSE

	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

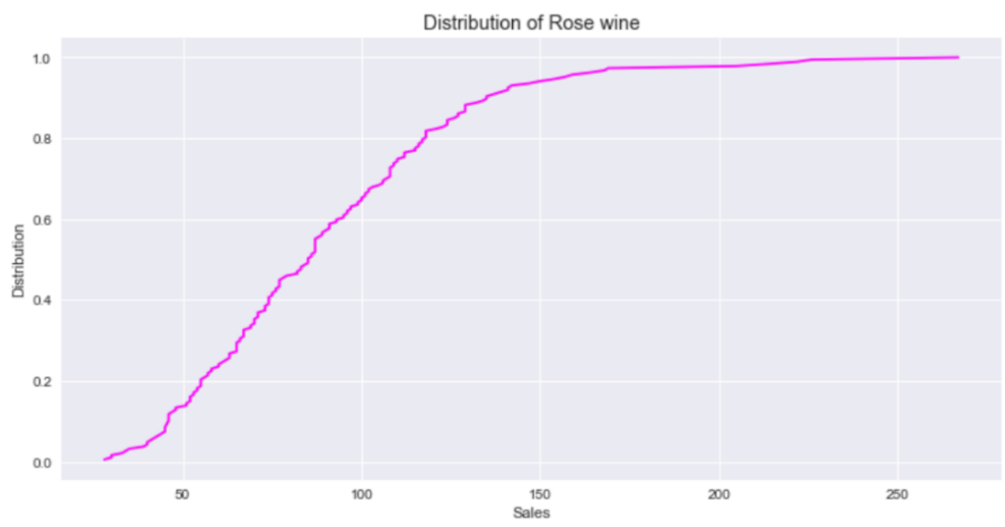
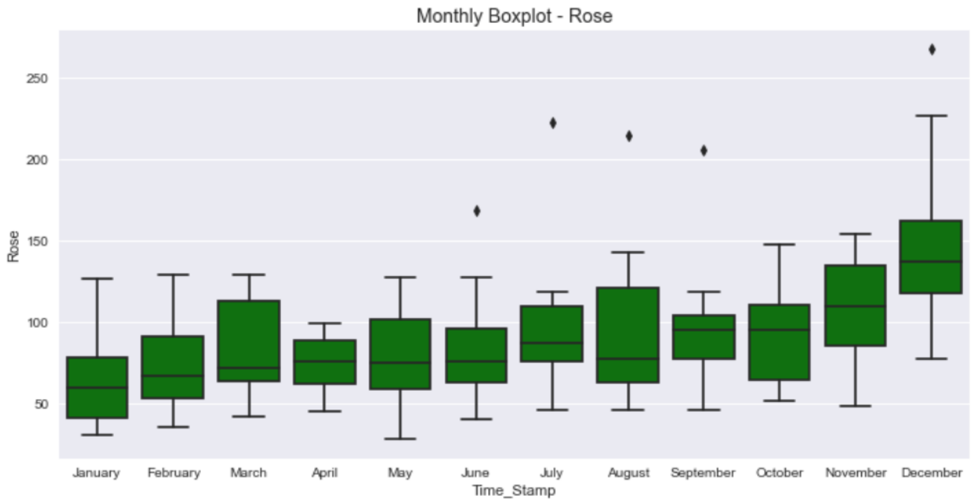


FIGURE-6- Descriptive analysis of ROSE dataset

The descriptive analysis of the data shows that on an average 90 units of Rose wines were sold each month on the given period of time. 50% of months sales varied from 63 units to 112 units.

Maximum sale reported in a month is 267 units and minimum of 28 units.

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



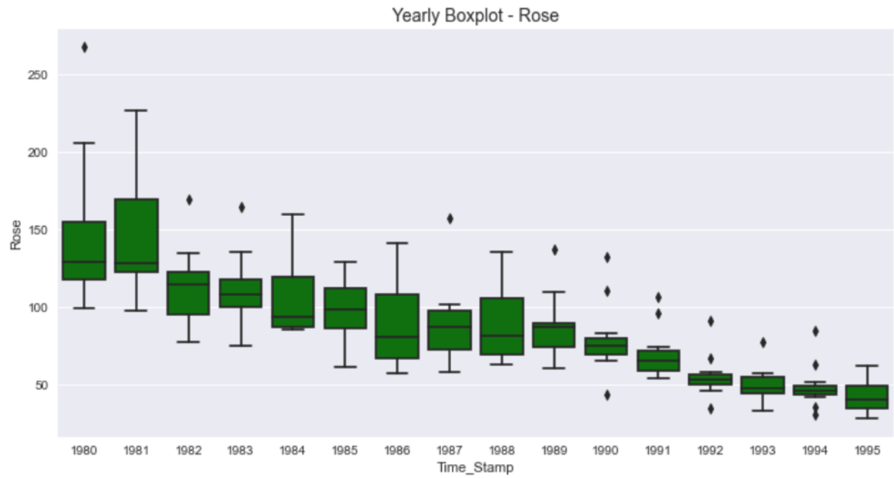


FIGURE-6- Yearly and Monthly Boxplot of ROSE dataset

The yearly-boxplot, shows that the average sale of Rose wine moving according to the downward trend in sales over the years.

The outliers over 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

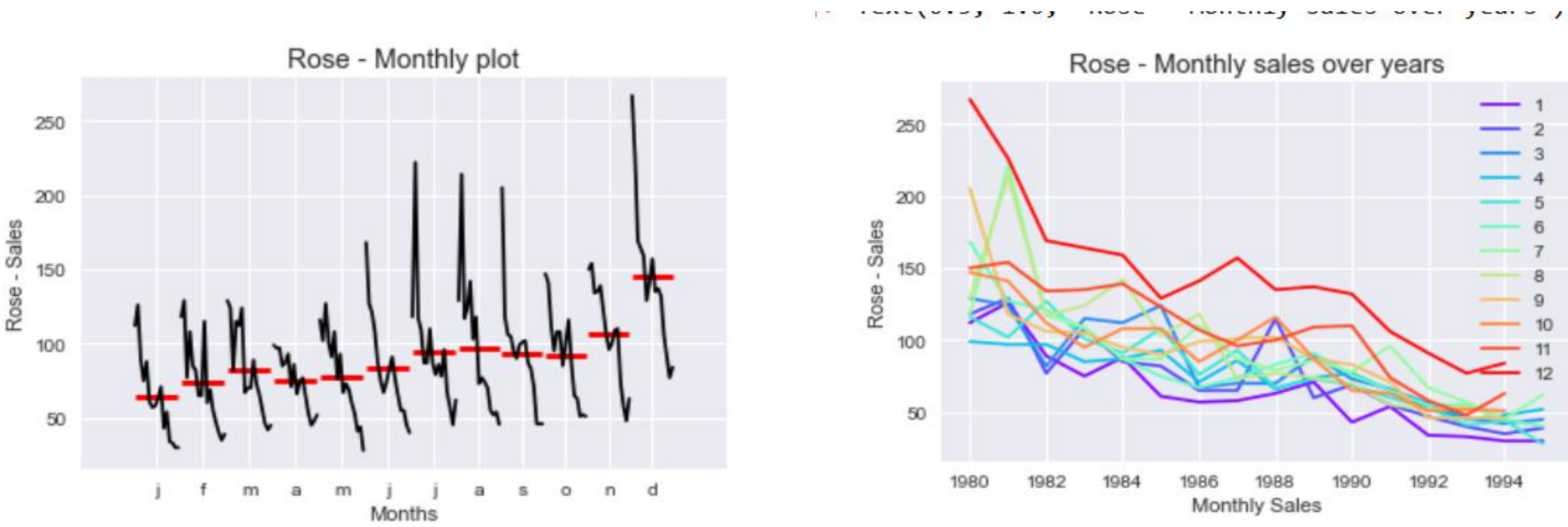


FIGURE-6- Yearly and Monthly sales Plot of ROSE dataset

The monthly plot for Rose shows a mean and variation of units sold each month over the years.

Sale in months such as July, August, September and December show a higher variation than the rest

Sale in December with a mean few points below 100, varies 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

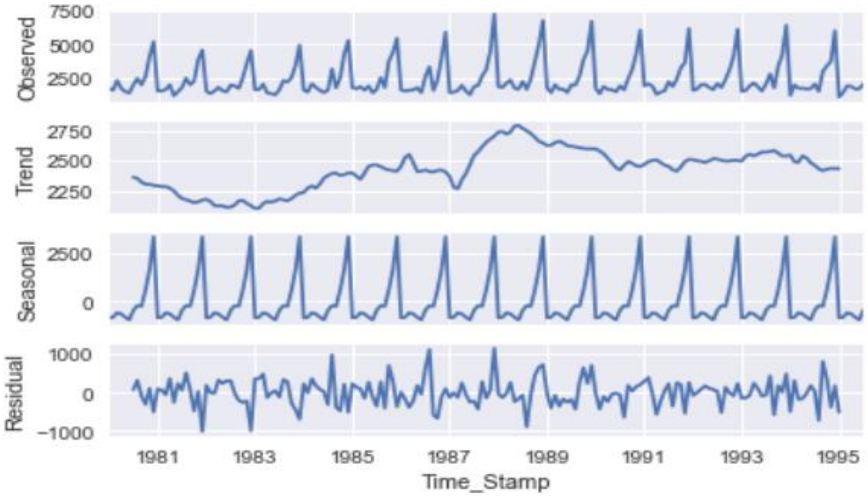
The plot of monthly sale over the years also shows the seasonality component of the time-series, with November and December selling exponentially higher volumes than other months.

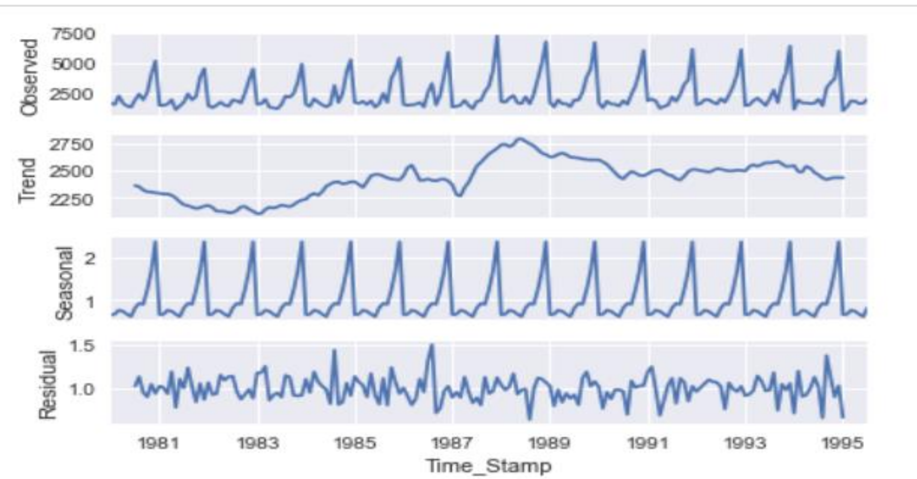
The highest volume of Rose wines was sold in December,1980 and the least of December sale was in 1993.

Though December sale picked after 1983, it consistently dipped after 1987.

Decomposition- Sparkling

Additive Decomposition   Multiplicative Decomposition





The plot of the trend component does not show a consistent trend, but an intermediary period shows an upward slope which gets consistent on the late half of time-series

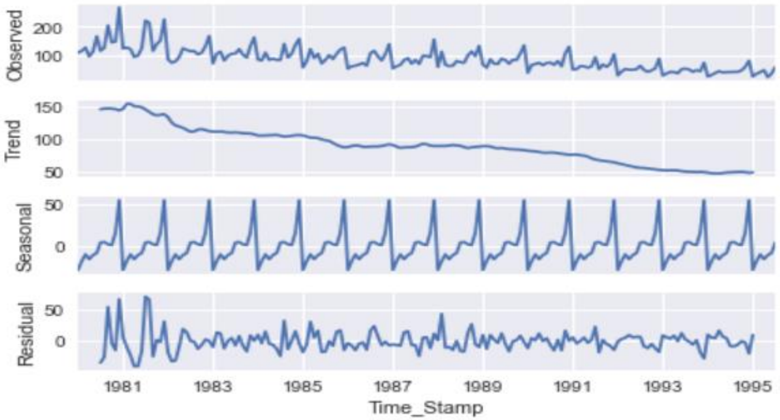
The additive model shows the seasonality with a variance of 3000 units and the multiplicative model shows a variance of 30%

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.

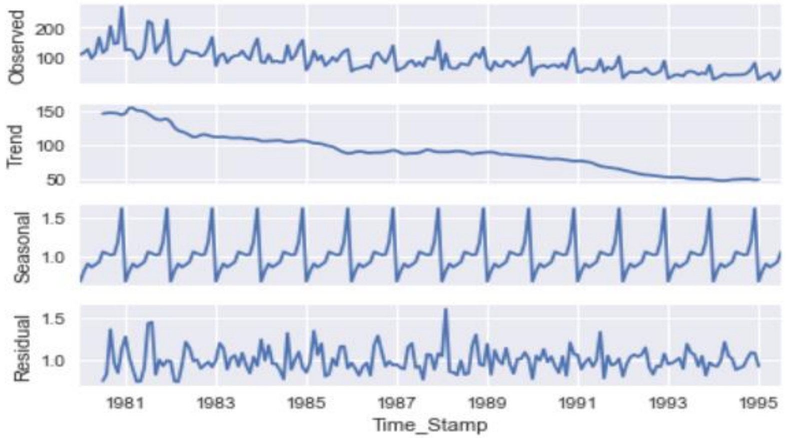
The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 10%

Decomposition- Rose

Additive Decomposition



Multiplicative Decomposition



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 observed plot at same time period • The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 15%

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

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

Sparkling-

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

```
Last few rows of Test Data
```

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

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

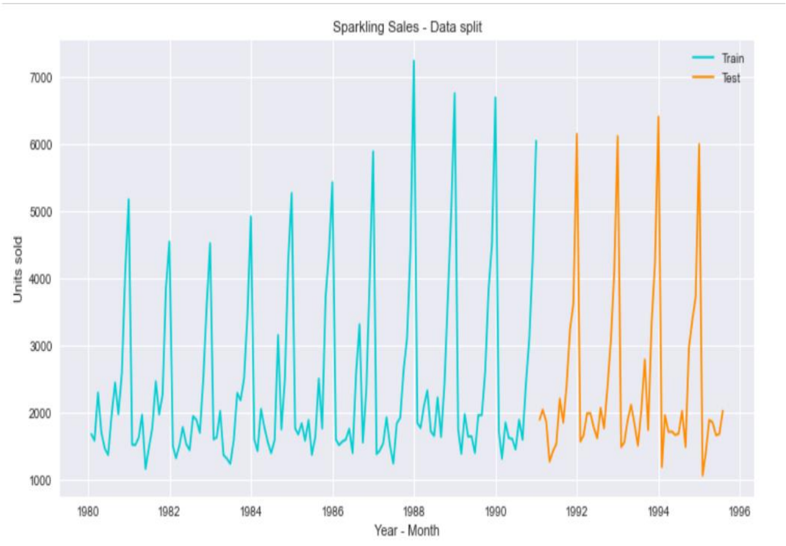
Last few rows of Training Data

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

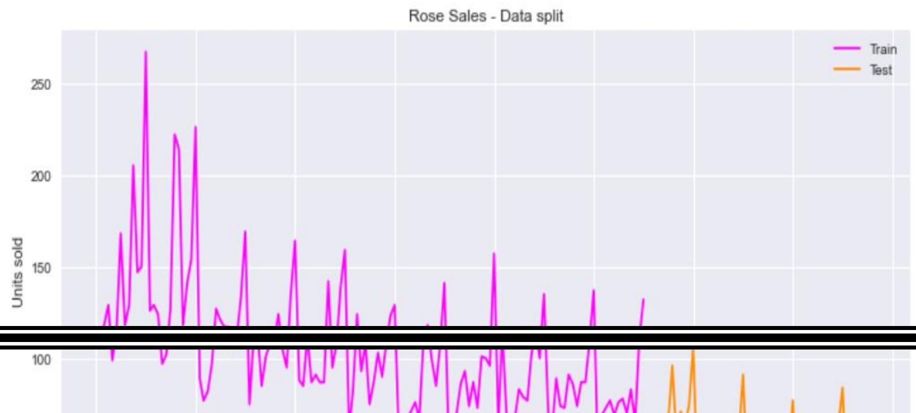
First few rows of Test Data

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

There are total of 132 rows in train data set and 55 rows in test data set.



Rose-



First few rows of Training Data

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

Last few rows of Training Data

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

First few rows of Test Data

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

Last few rows of Test Data

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

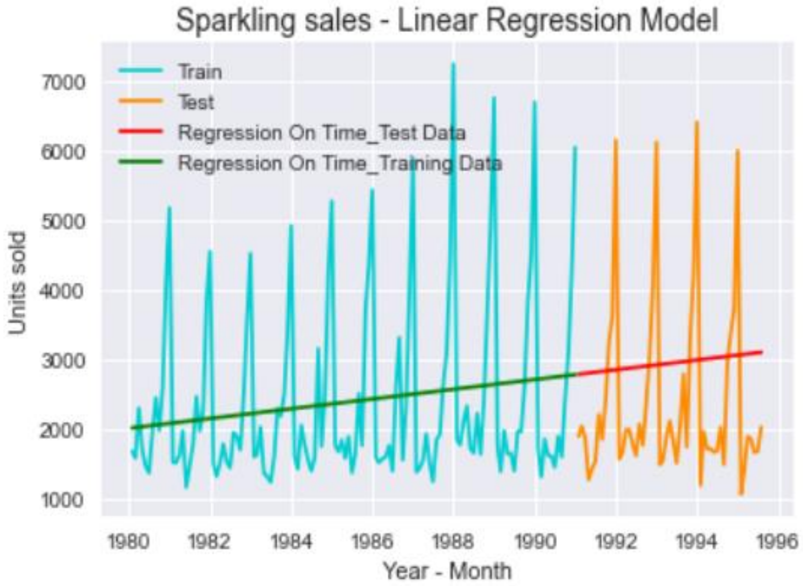
There are total of 132 rows in train data set and 55 rows in test data set.

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

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

Model-1-Linear Regression

Sparkling-



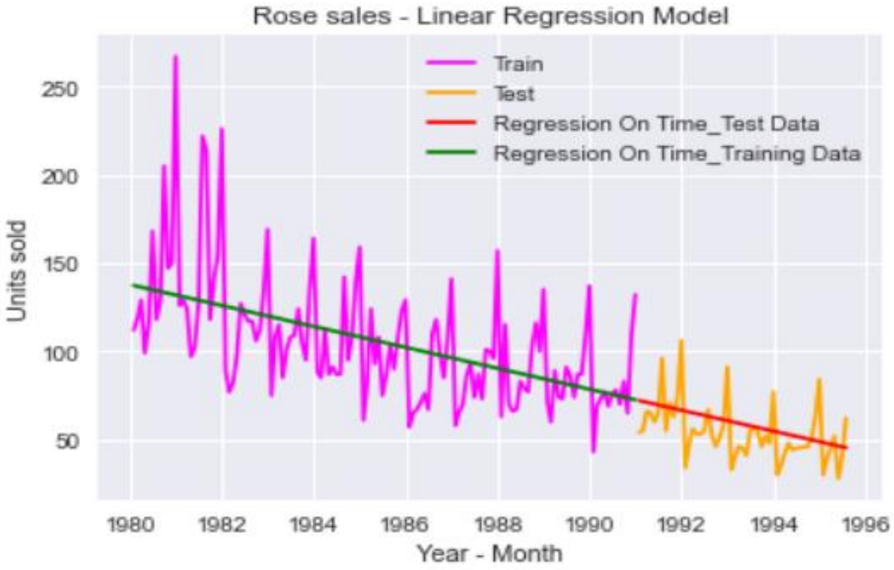
For RegressionOnTime forecast on the Sparkling Training Data: RMSE is 1279.322 and MAPE is 40.05

For RegressionOnTime forecast on the Sparkling Testing Data: RMSE is 1389.135 and MAPE is 50.15

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

The RMSE and MAPE values for Train and Test data sets are as above.

ROSE-



For RegressionOnTime forecast on the Rose Training Data: RMSE is 30.718 and MAPE is 21.22

For RegressionOnTime forecast on the Rose testing Data: RMSE is 15.269 and MAPE is 22.82

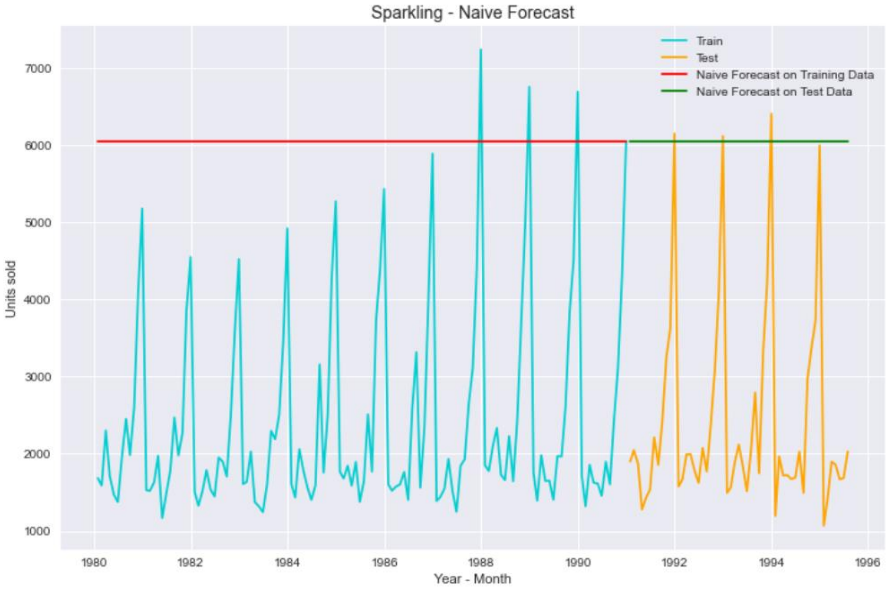
The linear regression on the Rose dataset shows an apparent downward trend as consistent with the observed time-series

The model leaves a 23% error in forecast against test set

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

Model 2: Naive Forecast

SPARKLING -



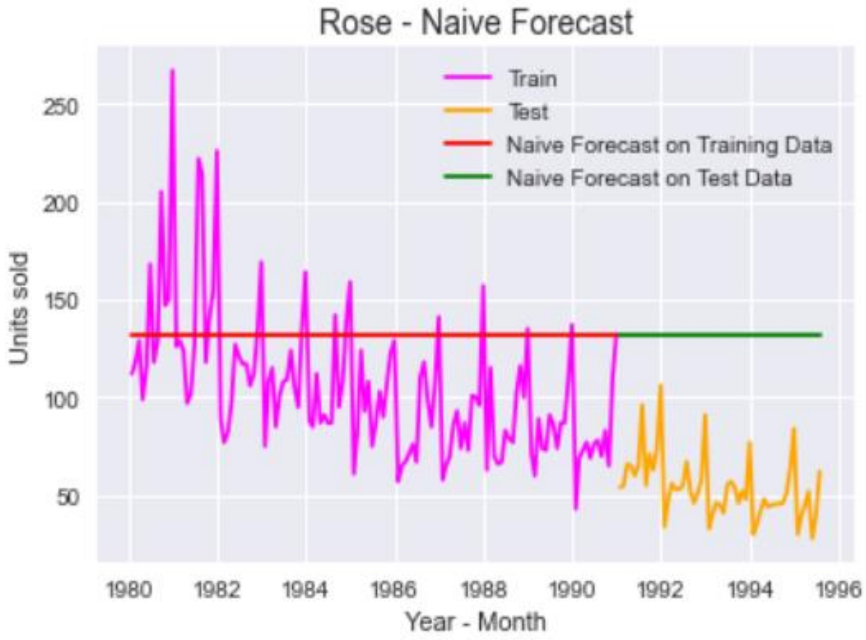
For Naive forecast on the Sparkling Training Data: RMSE is 3867.701 and MAPE is 153.17

For Naive forecast on the Sparkling Testing Data: RMSE is 3864.279 and MAPE is 152.87

In naive model, the prediction for tomorrow is the same as today and the prediction or day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

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 metrices above is very poor and contains high percentage of error.

ROSE-



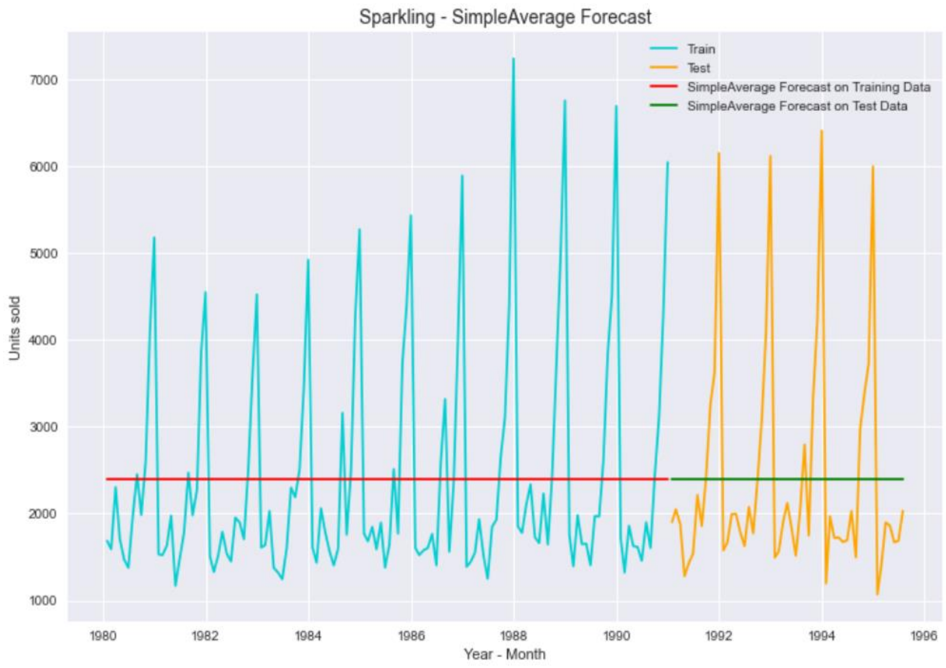
For Naive forecast on the Rose Training Data: RMSE is 45.064 and MAPE is 36.38

For Naive forecast on the Rose Testing Data: RMSE is 79.719 and MAPE is 145.10

As Rose data set has a downward trend the percentage of error in train is lesser and is very high in test

The model does not capture the trend nor seasonality of the given datasets

MODEL-3-SIMPLE AVERAGE FORECASTING



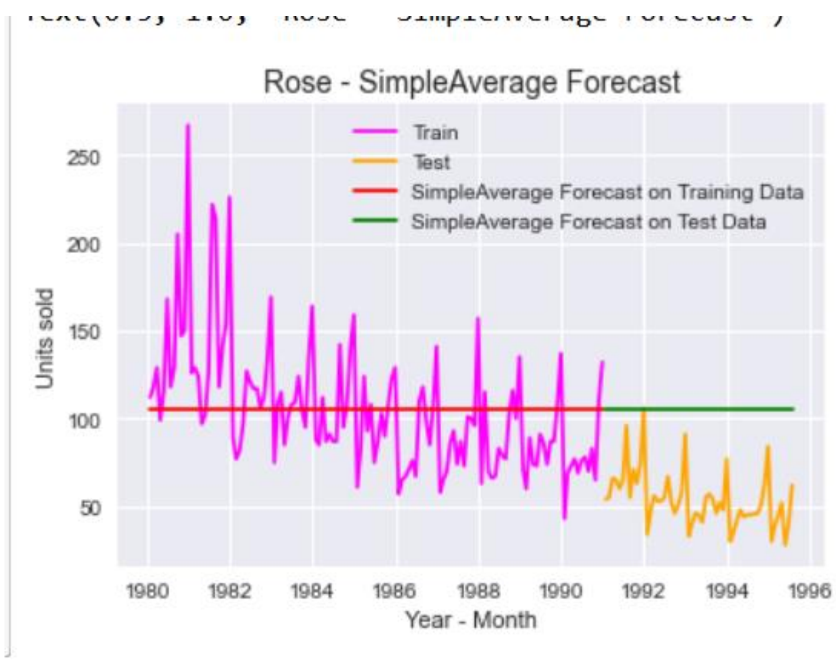
For Simple Average forecast on the Sparkling Training Data: RMSE is 1298.484 and MAPE is 40.36

For Simple Average forecast on the Sparkling Testing Data: RMSE is 1275.082 and MAPE is 38.90

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

The model is not capable of either forecasting nor able to capture the trend and seasonality present in the dataset • For Sparkling the RMSE and MAPE is consistent in both test and train datasets

ROSE



For Simple Average forecast on the Rose Training Data: RMSE is 36.034 and MAPE is 25.39

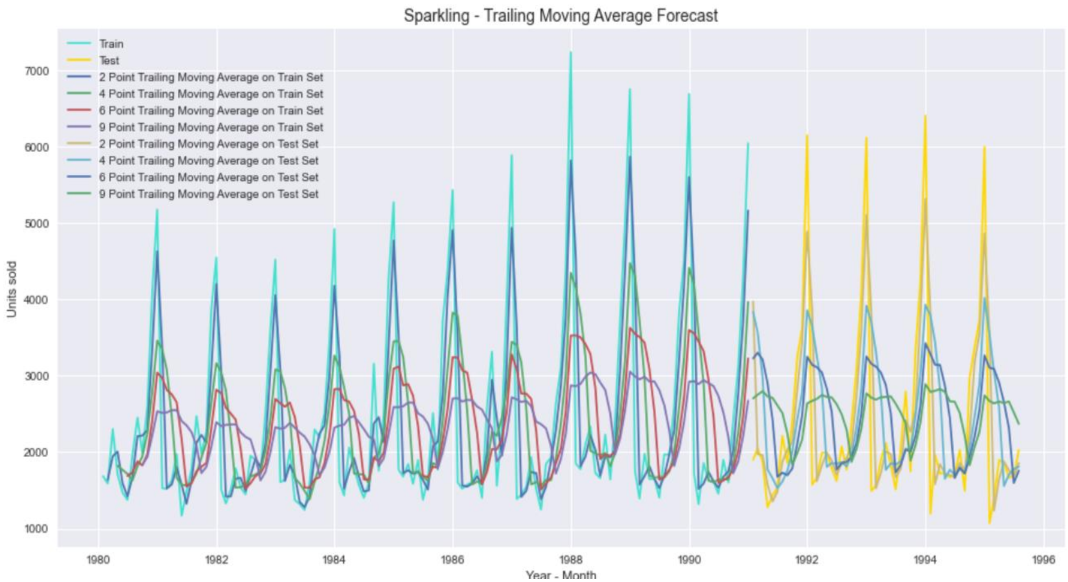
For Simple Average forecast on the Rose Testing Data: RMSE is 53.461 and MAPE is 94.93

We could see here the model forecast is almost 100% error in test data and 25% in train

Due to the downward trend the performance in train data set is better than the test dataset

MODEL-4-MOVING AVERAGE-

SPARKLING



	Test RMSE	Test MAPE
2 point MA	813.400684	19.70
4 point MA	1156.589694	35.96
6 point MA	1283.927428	43.86
9 point MA	1346.278315	46.86

For the moving average model, we will calculate rolling means (or trailing moving averages) for different intervals.

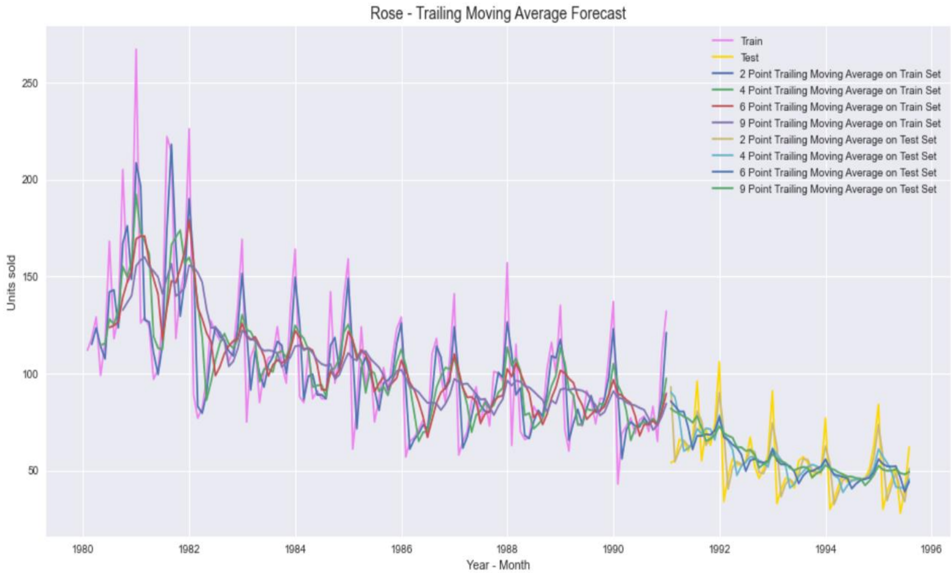
The best interval can be determined by the maximum accuracy (or the minimum error),

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

For Sparkling dataset, the accuracy is found to be higher with the lower rolling point averages

In moving average forecasts the values can be fitted with a delay of n number of points. The best interval of moving average from the model is 2 point.

ROSE-



	Test RMSE	Test MAPE
2 point TMA	11.529278	13.54
4 point TMA	14.451403	19.49
6 point TMA	14.566327	20.82
9 point TMA	14.727630	21.01

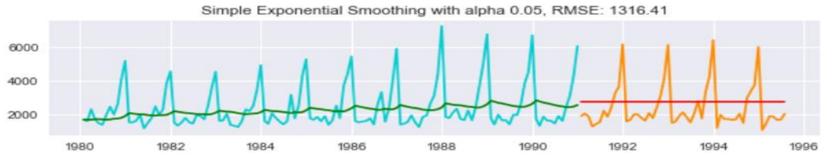
For Rose dataset the accuracy is found to be higher with the lower rolling point averages

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

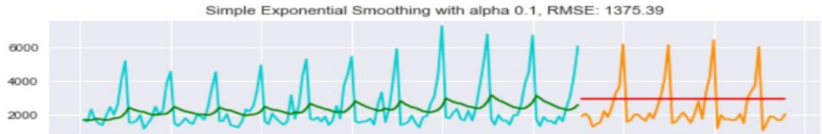
The best interval of moving average from the model is 2 points as the error percentage is less than other points.

<u>MODEL-5-SIMPLE EXPONENTIAL SMOOTHING</u> - Sparkling

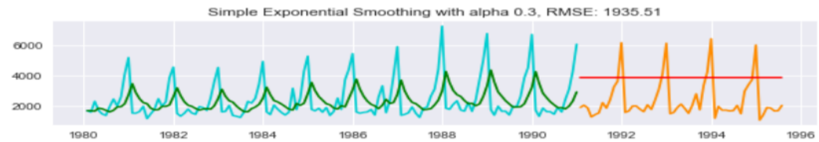
```
Test: For alpha = 0.05,  RMSE is 1316.4117 MAPE is 45.50
For smoothing level = 0.05,  Initial level 1686.00
```



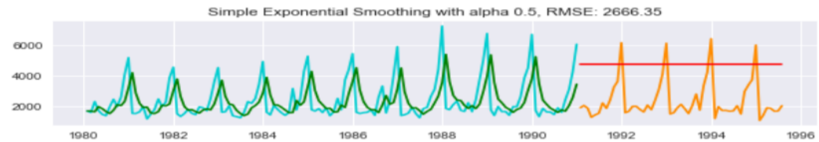

```
Test: For alpha = 0.10,  RMSE is 1375.3934 MAPE is 49.53
For smoothing level = 0.10,  Initial level 1686.00
```



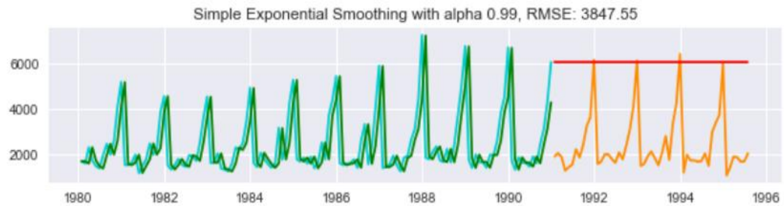

```
Test: For alpha = 0.30,  RMSE is 1935.5071 MAPE is 75.66
For smoothing level = 0.30,  Initial level 1686.00
```

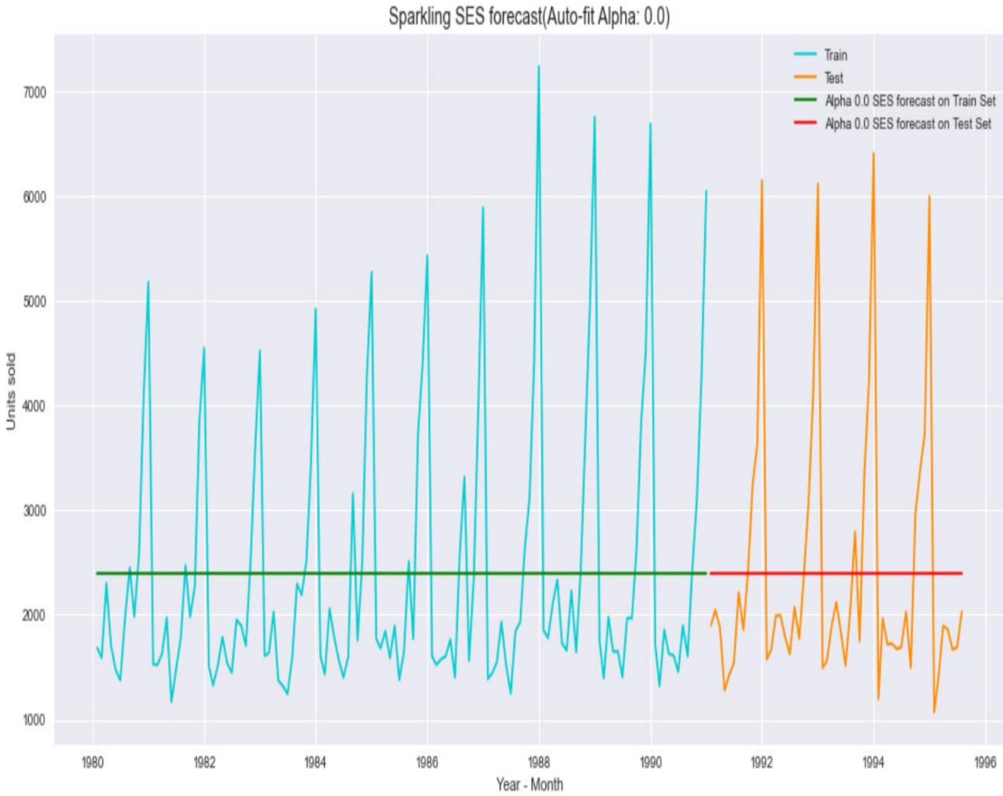



```
Test: For alpha = 0.50,  RMSE is 2666.3514 MAPE is 106.27
For smoothing level = 0.50,  Initial level 1686.00
```




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





```
For SES forecast on the Sparkling Training Data:  RMSE is 1298.484 and MAPE is 40.36

For SES forecast on the Sparkling Testing Data:  RMSE is 1275.082 and MAPE is 38.90
```

Simple Exponential Smoothing is applied if the time-series has neither a trend nor 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

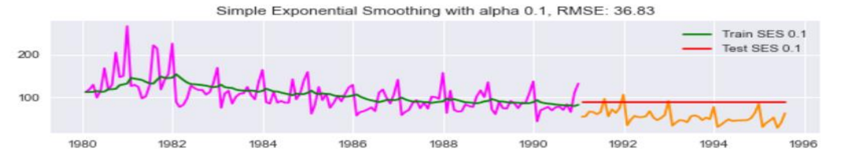
For alpha value closer to 1, forecasts follow the actual observation closely and closer to 0, forecasts are farther from actual and line gets smoothened
For Sparkling, test RMSE is found to be higher for values closer to zero, which is same as in Simple average forecast.

The autofit model picked 0.0 as the smoothing parameter and retuned consistent RMSE values in train and test datasets, which is higher in accuracy than in first iteration

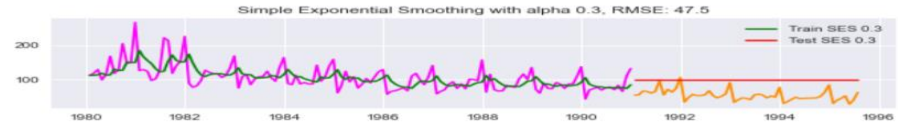
As the smoothing level is 0.0, we got a completely smoothened out forecast with an initial value 2403.79 applied across the series

ROSE -

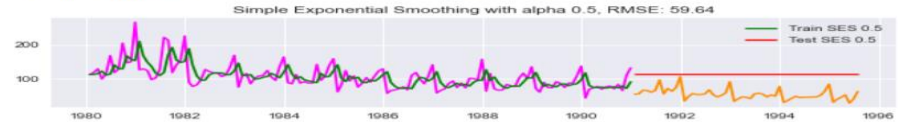
```
Test: For alpha = 0.10,  RMSE is 36.8280 MAPE is 63.94
For smoothing level = 0.10,  Initial level 112.00
```



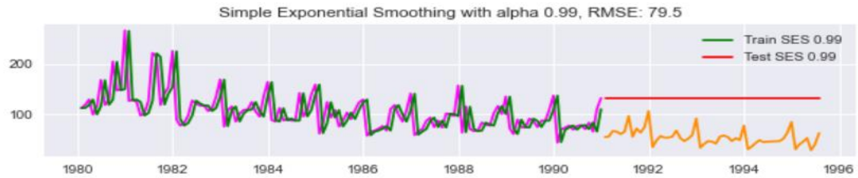

```
Test: For alpha = 0.30,  RMSE is 47.5048 MAPE is 83.71
For smoothing level = 0.30,  Initial level 112.00
```

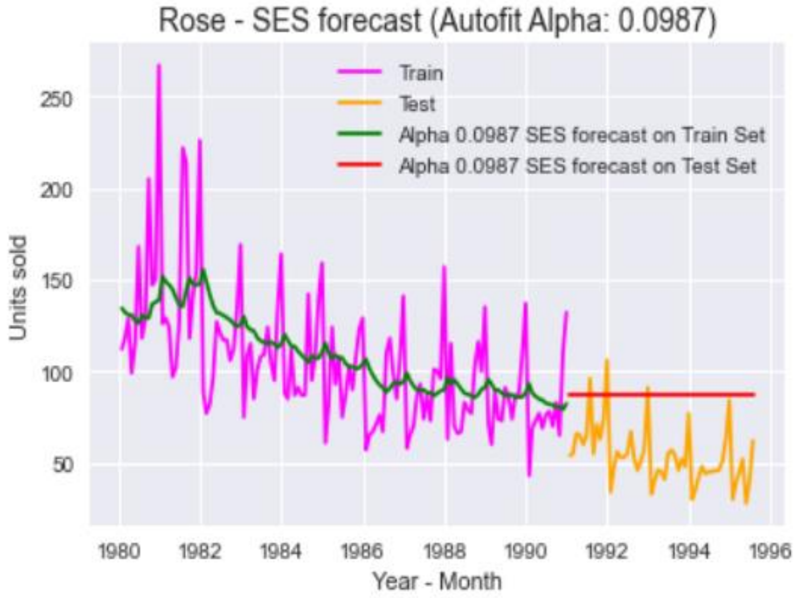



```
Test: For alpha = 0.50,  RMSE is 59.6418 MAPE is 106.81
For smoothing level = 0.50,  Initial level 112.00
```




```
Test: For alpha = 0.99,  RMSE is 79.4987 MAPE is 144.69
For smoothing level = 0.99,  Initial level 112.00
```





For SES forecast on the Rose Training Data: RMSE is 31.501 and MAPE is 22.73

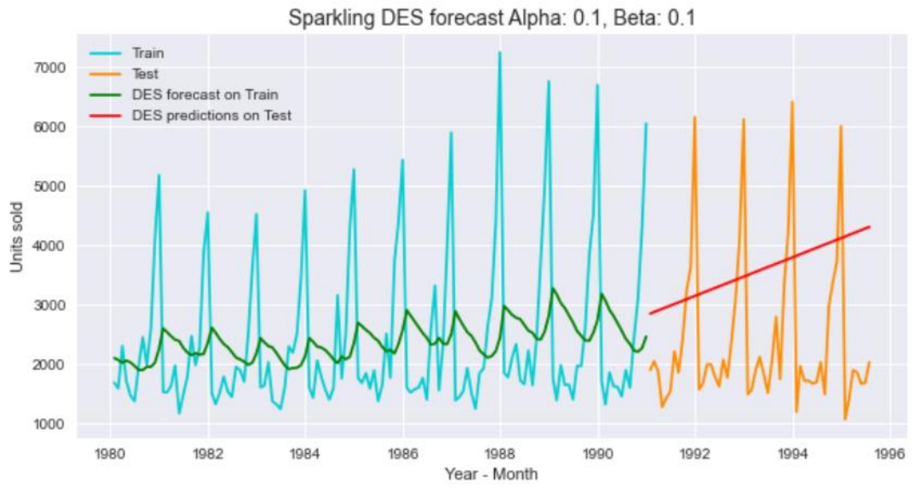
For SES forecast on the Rose Testing Data: RMSE is 36.796 and MAPE is 63.88

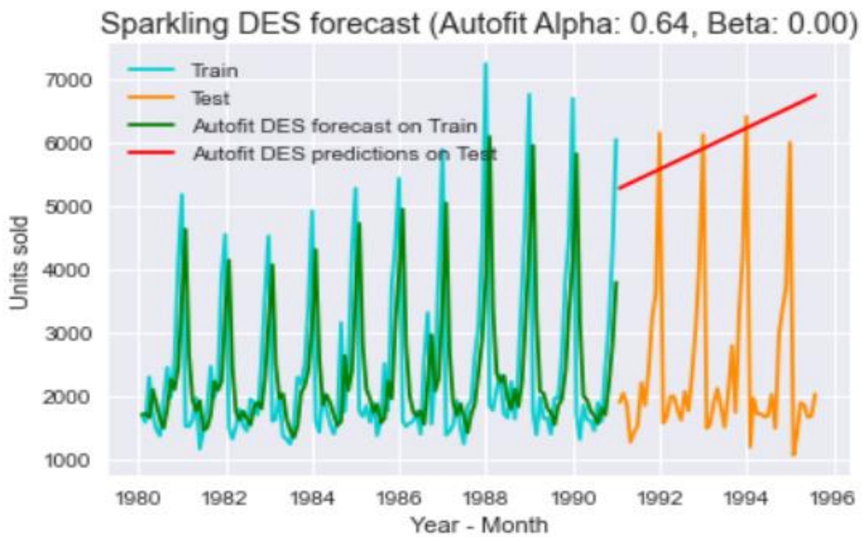
The forecasting using smoothing levels or alpha between 0 and 1 are as below, where the values were passed manually.

For alpha value closer to 1, forecasts follow the actual observation closely and closer to 0, forecasts are farther from actual and line gets smoothened the test RMSE is found to be higher for values closer to zero. 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

MODEL-6-DOUBLE EXPONENTIAL SMOOTHING

SPARKLING-





	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100000	0.1	1363.47000	44.26	1779.4300	67.23
1	0.100000	0.2	1401.76000	45.65	2599.7900	95.44
10	0.200000	0.1	1412.03000	46.62	3611.7700	135.41
100	0.647811	0.0	1337.48427	39.11	3851.1715	152.07

The Double Exponential Smoothing models is applicable when data has trend, but no seasonality. Sparkling data contain slight trend component and very significant seasonality

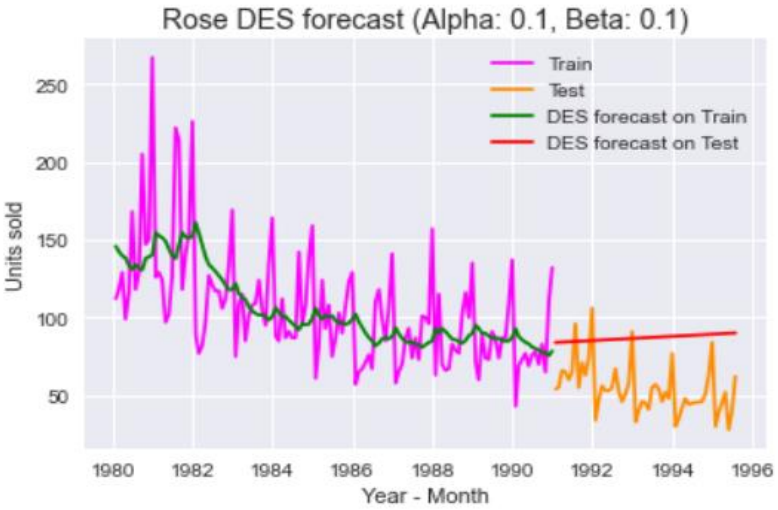
In first iteration, smoothing level (alpha) and trend (beta) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.1 and beta 0.1

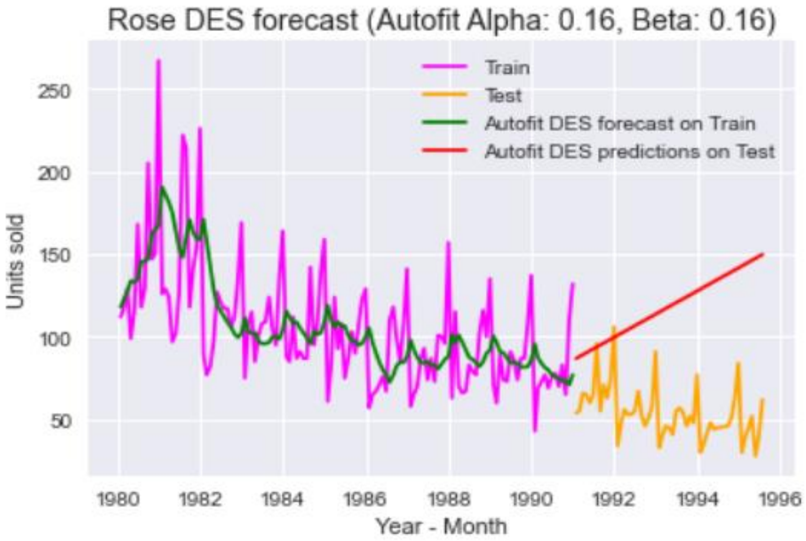
On the second iteration the model was allowed to choose the optimized values using parameters ‘optimized=True, use_brute=True’

The autofit model retuned higher accuracy in train dataset, but fared poorly in test, compared with the values in manual iteration

The model evaluation parameters of top three models from manual iteration and the autofit models are as given above • The best model chosen as final one is with alpha 0.1 and beta 0.1

Rose –





	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100000	0.100000	32.026565	22.78	37.057142	64.02
1	0.100000	0.200000	33.450729	24.45	48.688648	83.09
10	0.200000	0.100000	32.796403	23.06	65.731602	113.20
100	0.157895	0.157895	33.074575	23.99	70.572452	120.25

Rose data contain significant trend component and seasonality

In 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

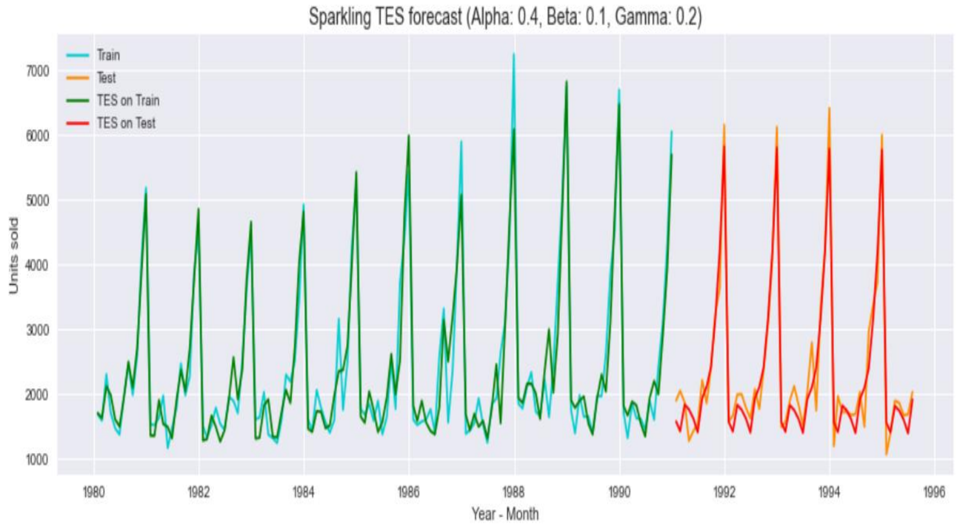
On the second iteration the model was allowed to choose the optimized values using parameters 'optimized=True, use_brute=True'

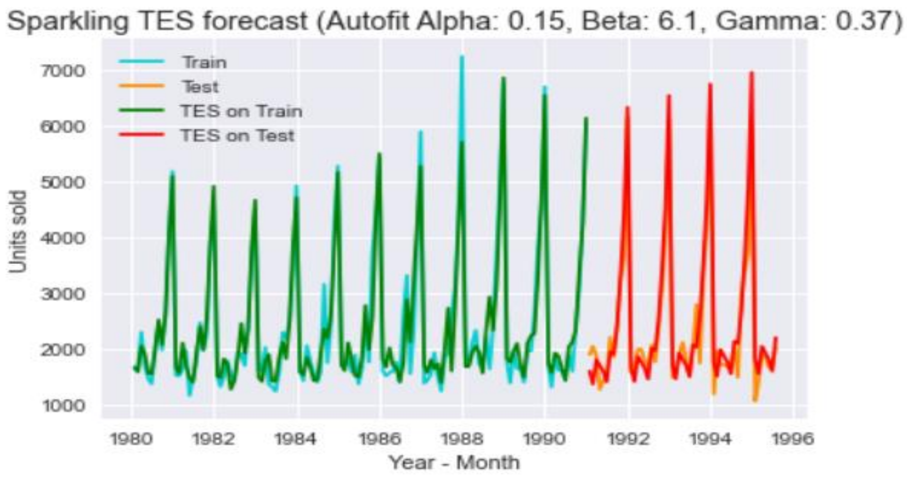
The autofit model retuned higher accuracy in train dataset, on par with the best models from iteration 1, but faired behind in the test accuracy scores the model evaluation parameters of the best models are given as above.

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

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

SPARKLING -





	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
301	0.4	0.1	0.2	373.281410	11.05	312.211095	10.20
211	0.3	0.2	0.2	377.346884	11.23	315.195004	10.07
300	0.4	0.1	0.1	370.807398	11.06	318.281180	10.00
402	0.5	0.1	0.3	390.181794	11.54	325.690492	9.99
403	0.5	0.1	0.4	401.059753	11.55	343.321993	11.07

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

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

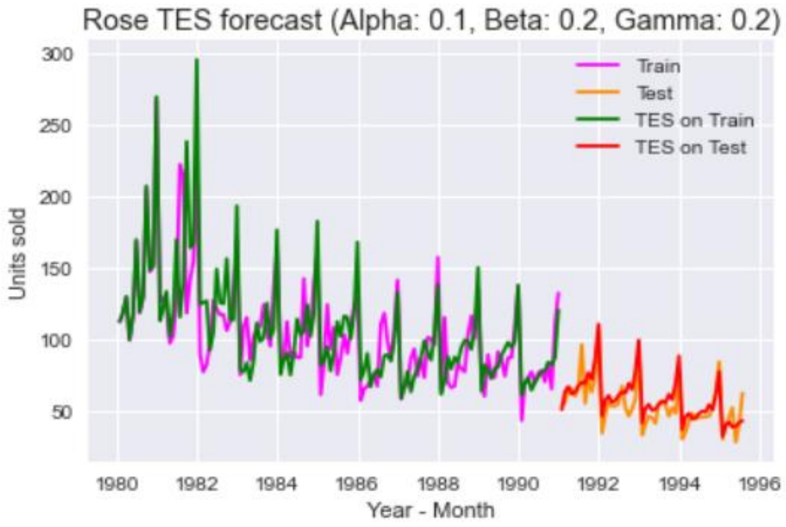
On the second iteration the model was allowed to chose the optimized values using parameters 'optimized=True, use_brute=True'

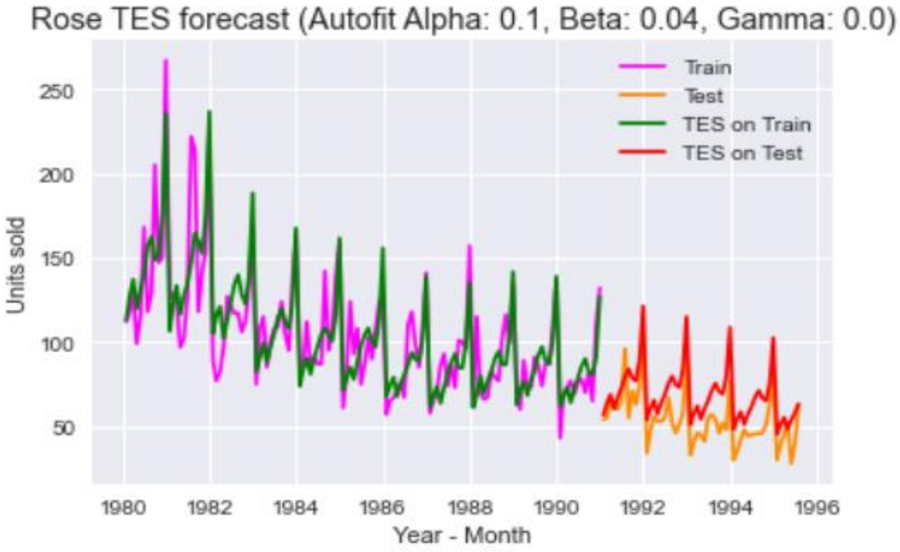
The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but faired poorly in accuracy in test

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

The best model chosen as 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
11	0.1	0.2	0.2	24.365597	15.36	9.640687	13.96
12	0.1	0.2	0.3	23.969166	15.13	9.935740	14.21
10	0.1	0.2	0.1	25.529854	16.06	9.943539	14.39
142	0.2	0.5	0.3	27.631767	17.87	10.026210	14.34

Rose data contain both trend and seasonality significantly

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 chose the optimized values using parameters 'optimized=True, use_brute=True'

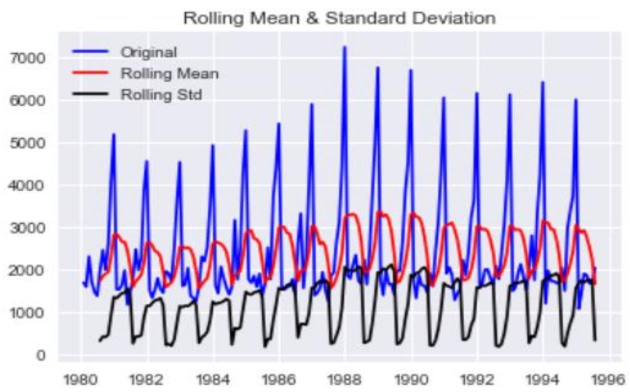
The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but faired poorly in accuracy in test

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

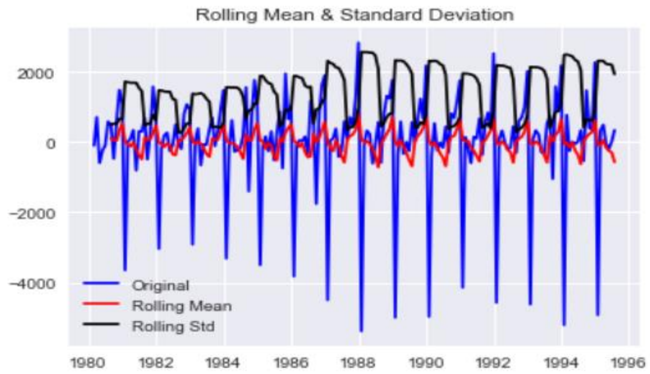
The best model chosen as final one is the one with alpha 0.1, beta 0.2 and gamma 0.2

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

SPARKLING




```
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
```




```
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
```

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

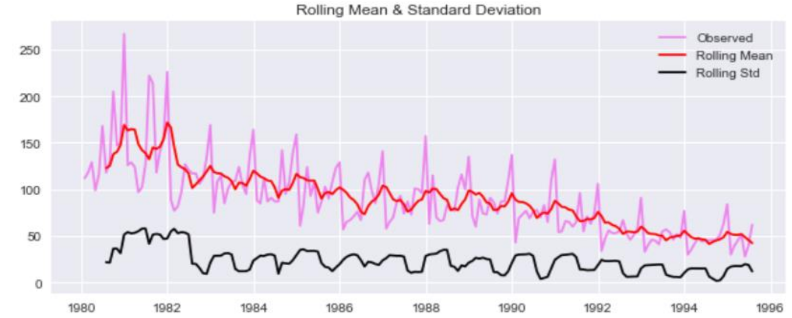
Null Hypothesis: The series has a unit root, that is series is non-stationary

Alternate Hypothesis: The series has no unit root, that is series is stationary

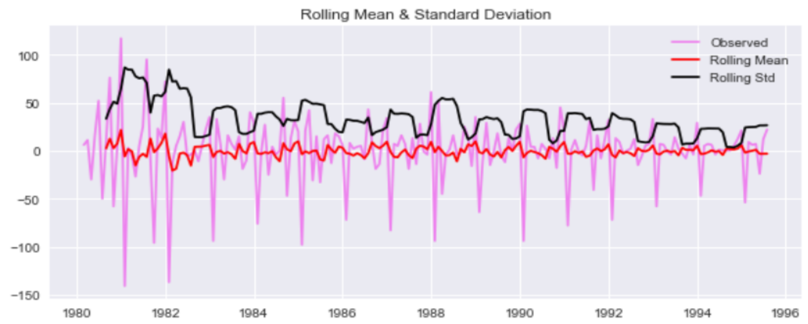
If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that the series is stationary

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

<u>ROSE-</u>




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




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

Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series.

The test determine 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 series is non-stationary

Alternate Hypothesis: The series has no unit root, that is 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 retuned the printed values, where p-value is greater than alpha .05 so we fail to reject the null hypothesis.

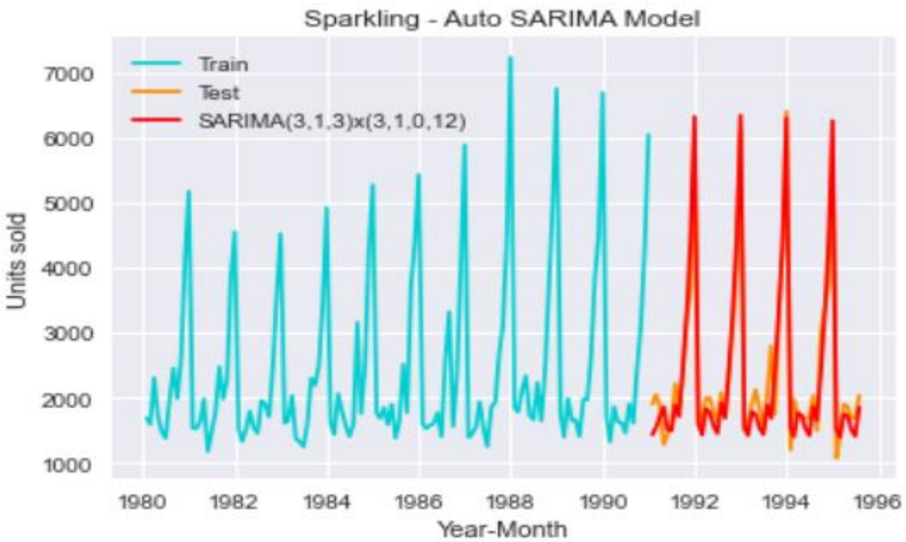
At an order of differencing 1, the series is found to be stationary.

The rolling mean and standard deviation is also plotted to understand the component of seasonality and to ascertain if its multiplicative or additive in character. The plot of rolling mean and standard deviation indicates that the seasonality is multiplicative as the altitude of 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 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.

SPARKLING-



	param	seasonal	AIC
252	(3, 1, 3)	(3, 1, 0, 12)	1213.282570
253	(3, 1, 3)	(3, 1, 1, 12)	1215.213340
220	(3, 1, 1)	(3, 1, 0, 12)	1215.898777

	Test RMSE	Test MAPE
Auto SARIMA(3,1,3)x(3,1,0,12)	324.105224	9.48

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

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

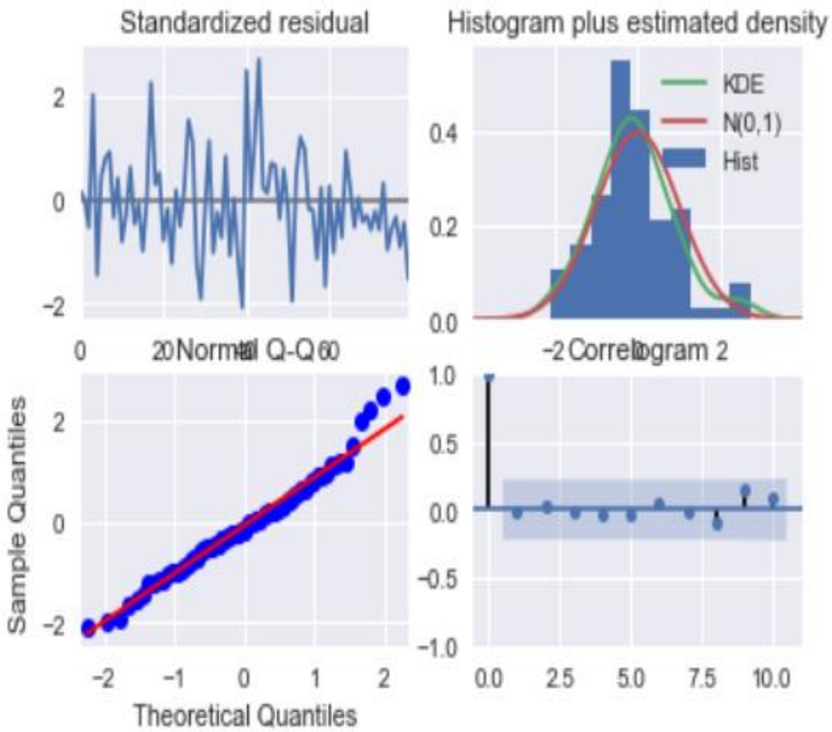
The top three models with lowest AIC values are as given above. As per the AIC criteria, the optimum values for final SARIMA model selected is (3, 1, 3)x(3, 1, 0, 12)

The diagnostics 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 forms roughly a straight line The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index

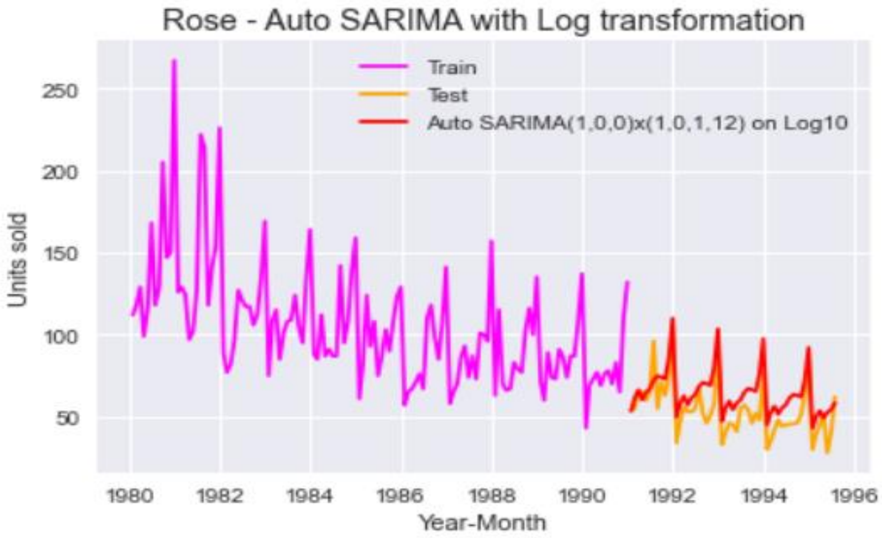
From the below model summary it can be inferred that AR(1), MA(1), MA(3), MA(2) terms has the highest absolute weightage.

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

```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 3)x(3, 1, 0, 12)   Log Likelihood                -596.641
Date:                            Sun, 16 Jan 2022   AIC                           1213.283
Time:                                    18:05:36   BIC                           1237.103
Sample:                                         0   HQIC                          1222.833
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6142      0.176     -9.178      0.000      -1.959      -1.270
ar.L2         -0.6123      0.299     -2.047      0.041      -1.199      -0.026
ar.L3          0.0861      0.161      0.537      0.592      -0.229       0.401
ma.L1          0.9856      0.477      2.066      0.039       0.051       1.920
ma.L2         -0.8739      0.166     -5.268      0.000      -1.199      -0.549
ma.L3         -0.9467      0.494     -1.916      0.055      -1.915       0.022
ar.S.L12      -0.4518      0.142     -3.190      0.001      -0.729      -0.174
ar.S.L24      -0.2341      0.144     -1.622      0.105      -0.517       0.049
ar.S.L36      -0.1008      0.122     -0.829      0.407      -0.339       0.138
sigma2      1.839e+05   9.05e+04      2.031      0.042    6424.357    3.61e+05
===================================================================================
Ljung-Box (Q):                       23.20   Jarque-Bera (JB):                 4.06
Prob(Q):                              0.98   Prob(JB):                         0.13
Heteroskedasticity (H):               0.73   Skew:                             0.48
Prob(H) (two-sided):                  0.42   Kurtosis:                         3.54
===================================================================================
```



ROSE



	param	seasonal	AIC
115	(1, 0, 0)	(1, 0, 1, 12)	-257.620755
7	(0, 0, 0)	(1, 0, 1, 12)	-256.170282
133	(1, 0, 1)	(1, 0, 1, 12)	-255.482061

	Test RMSE	Test MAPE
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.593960	21.93
Auto SARIMA(3,1,1)x(3,1,1,12)	16.823650	25.48

As the Rose series of data contain seasonality component we will be building 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

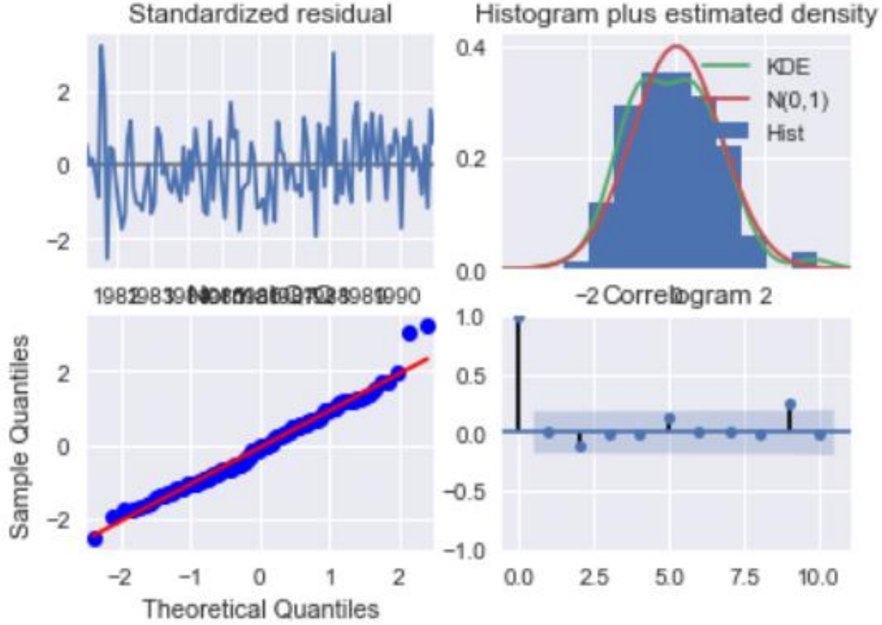
The model built with log transformed data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model

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)

```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(1, 0, 0)x(1, 0, 1, 12)   Log Likelihood                 132.810
Date:                            Sun, 16 Jan 2022   AIC                           -257.621
Time:                                    19:16:04   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          0.1688      0.078      2.178      0.029       0.017       0.321
ar.S.L12       0.9872      0.001    751.506      0.000       0.985       0.990
ma.S.L12      -0.9404      0.347     -2.713      0.007      -1.620      -0.261
sigma2         0.0052      0.002      2.916      0.004       0.002       0.009
===================================================================================
Ljung-Box (Q):                       24.28   Jarque-Bera (JB):                 4.00
Prob(Q):                              0.98   Prob(JB):                         0.14
Heteroskedasticity (H):               0.86   Skew:                             0.40
Prob(H) (two-sided):                  0.64   Kurtosis:                         3.40
===================================================================================
```



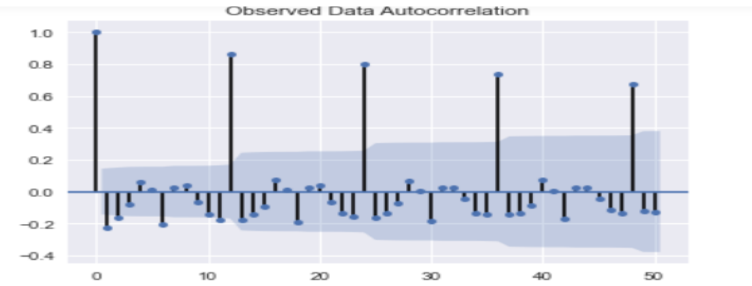
The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution

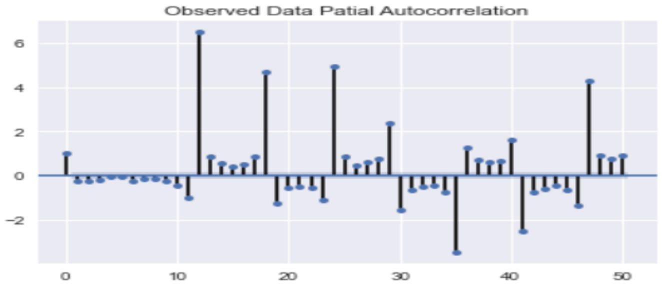
The Normal Q -Q plot also shows that the quantiles come from a normal distribution as the points forms roughly a straight line

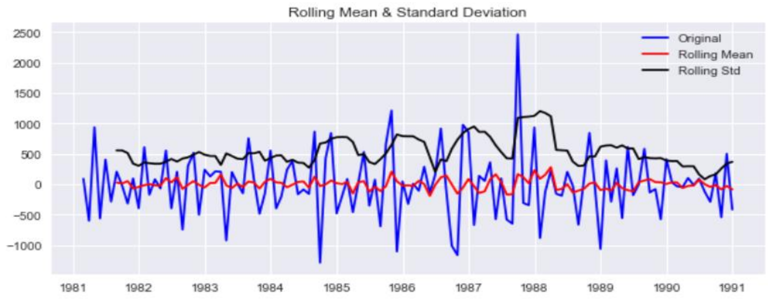
The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index.

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.

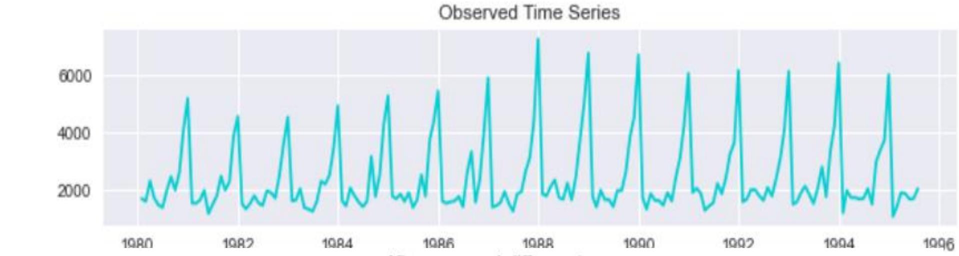
Sparkling-

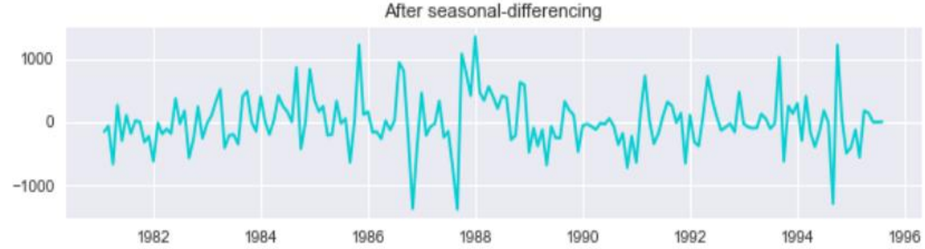


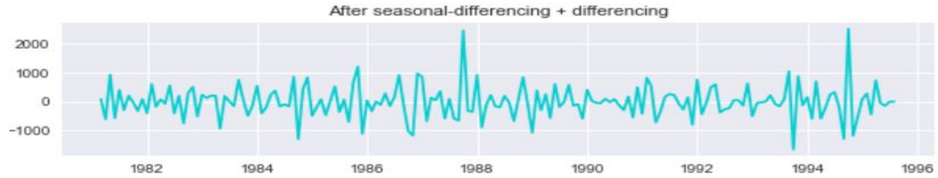





```
Results of Dickey-Fuller Test:
Test Statistic                 -3.342905
p-value                         0.013066
#Lags Used                     10.000000
Number of Observations Used   108.000000
Critical Value (1%)            -3.492401
Critical Value (5%)            -2.888697
Critical Value (10%)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







From the ACF plot of the observed/ train data, it can be inferred that at seasonal interval of 12, the plot is not quickly tapering off. So a seasonal differencing of 12 has to be taken .

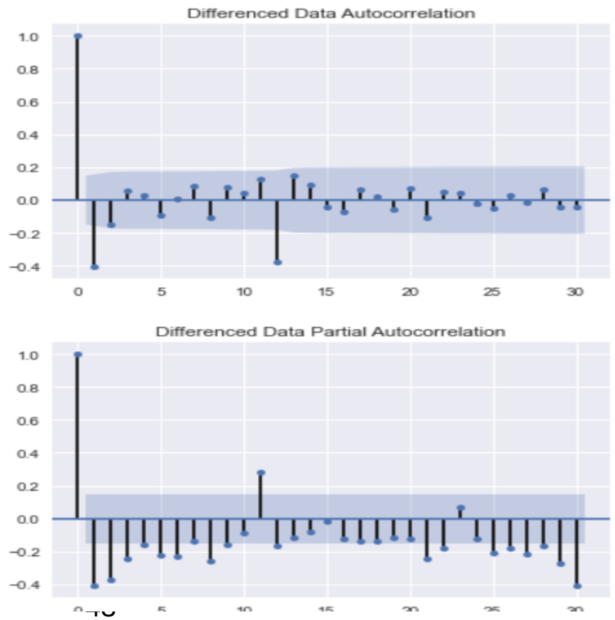
From the plots above an apparent slight trend is still existing after differencing of seasonal order of 12. With a further differencing of order one, no trend is present.

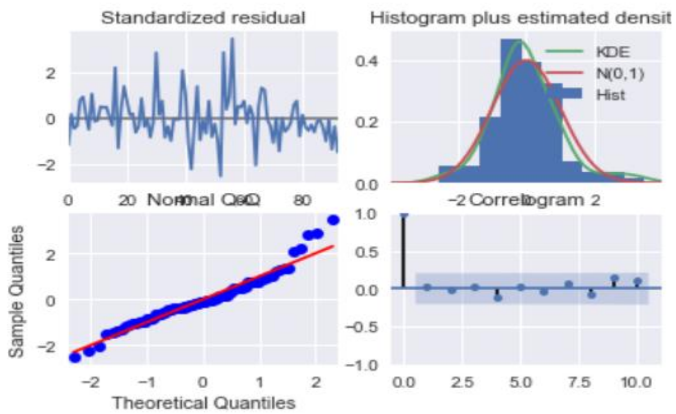
An ADF test need to be done to check the stationarity after the above differencing. With a p-value below alpha 0.05 and test statistic below critical values, it can be confirmed that the data is stationary ACF and PACF plots of the seasonal-differenced + one order differenced data is created to find the values for (p,d,q)x(P,D,Q)

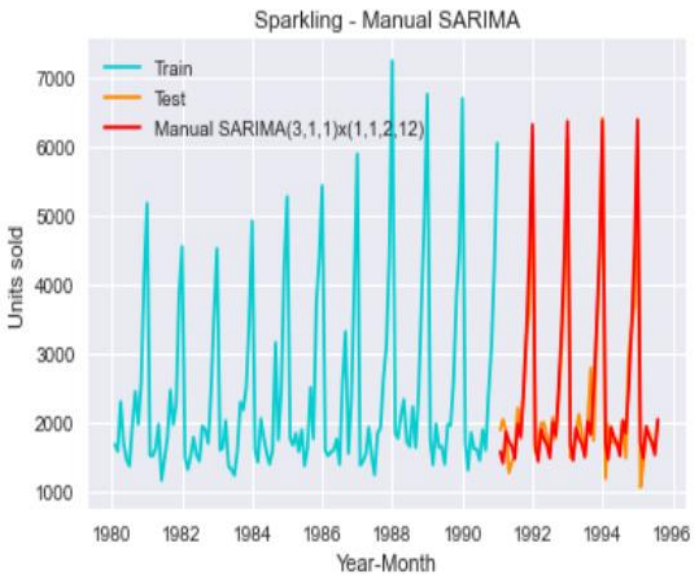
Here we have taken alpha = 0.05 and seasonal period as 12 •

From the PACF plot it can be seen that till 3rd lag its 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 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 kept seasonal MA term ‘Q = 1’ initially







The seasonal MA term 'Q' was later optimized to 2, by validating model performance, as the data might be under-differenced

The final selected terms for SARIMA model is (3, 1, 1)x(1, 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 forms 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 that only MA(1) term used in the model is significant in terms of p-values

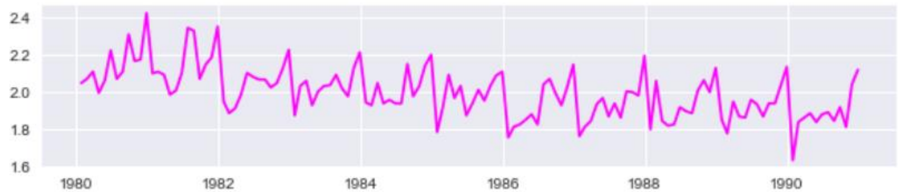
```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                                    y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(1, 1, 2, 12)   Log Likelihood                -693.697
Date:                              Sun, 16 Jan 2022   AIC                           1403.394
Time:                                      18:15:29   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 (Q):                       24.16   Jarque-Bera (JB):                19.66
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               0.81   Skew:                             0.69
Prob(H) (two-sided):                  0.56   Kurtosis:                         4.78
```

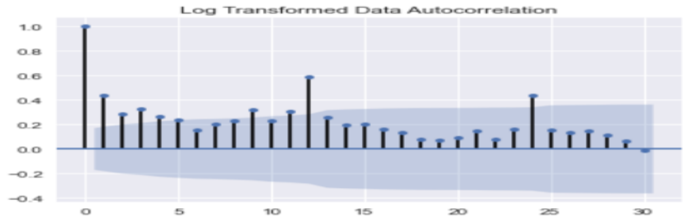
For SARIMA forecast on the Sparkling Testing Data: RMSE is 324.105 and MAPE is 9.48

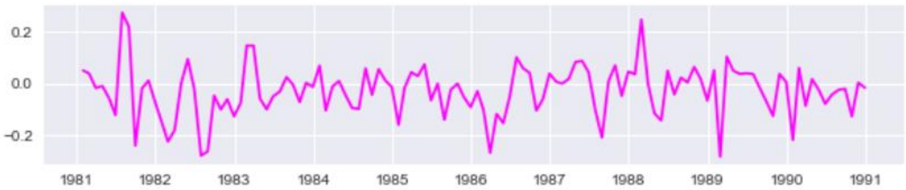
ROSE-

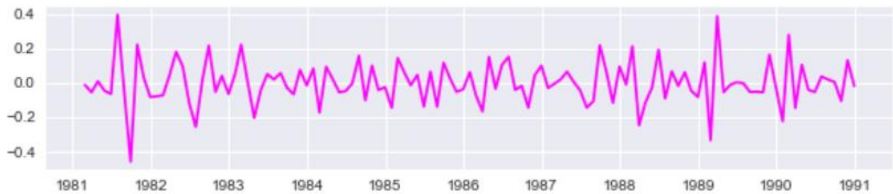
Log transformation of the data is done to handle multiplicity of seasonality • From the ACF plot of the log transformed data, it can be seen that at seasonal interval of 12, the plot is not quickly tapering off.

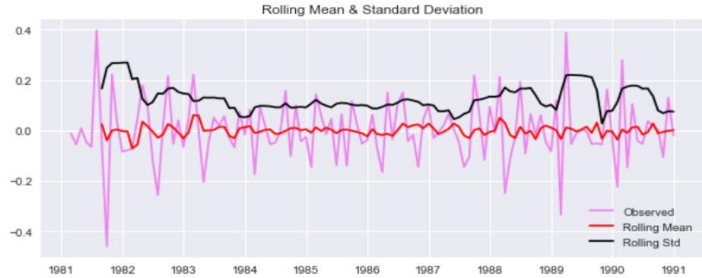
So we need to take a seasonal differencing of 12 .From the plots below it can be seen that a slight trend is still existing after differencing of seasonal order of 12. With a further differencing of order one, no trend is present.












```
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
```

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

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

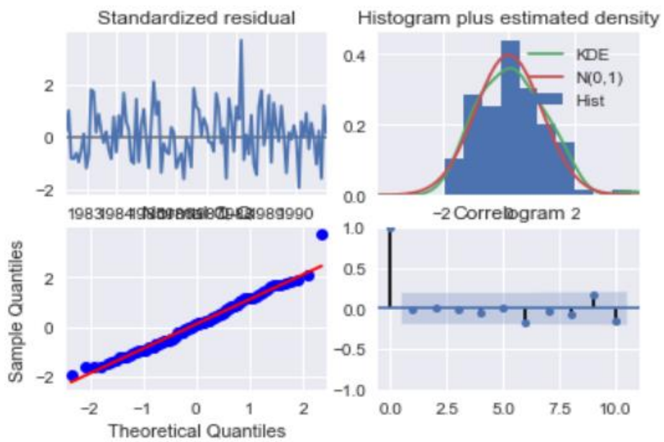
Here we have taken alpha = 0.05 and seasonal period as 12.

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

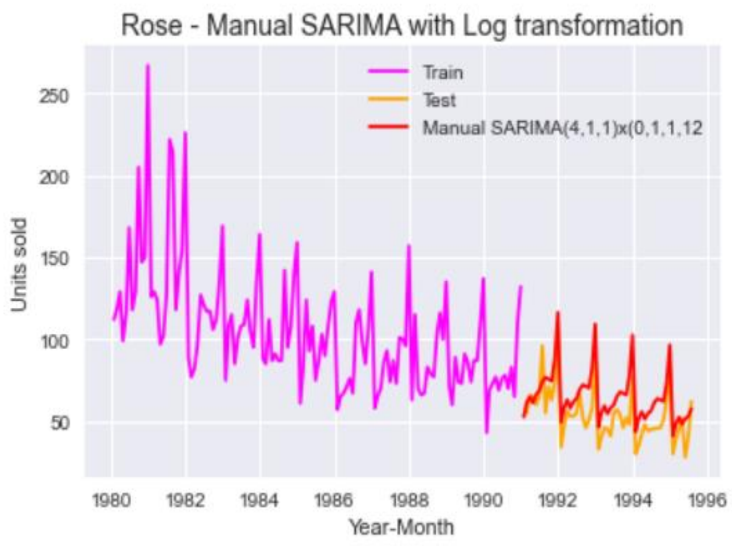
From ACF plot, lag 1 and 2 are significant before it cuts off, so lets keep MA term ‘q = 1’ and at seasonal lag of 12, a significant lag is apparent, so lets keep ‘Q = 1’

The final selected terms for 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 forms roughly a straight line .

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



For SARIMA forecast on the Rose Testing Data: RMSE is 14.180 and MAPE is 23.10




```
                                      Statespace Model Results
==========================================================================================
Dep. Variable:                              Rose   No. Observations:                  132
Model:             SARIMAX(4, 1, 1)x(0, 1, 1, 12)   Log Likelihood                 128.764
Date:                           Sun, 16 Jan 2022   AIC                           -243.528
Time:                                   20:01:43   BIC                           -224.950
Sample:                               01-31-1980   HQIC                          -236.000
                                    - 12-31-1990
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0018      0.118     -0.015      0.988      -0.233       0.229
ar.L2         -0.1549      0.126     -1.226      0.220      -0.403       0.093
ar.L3         -0.1598      0.113     -1.420      0.156      -0.380       0.061
ar.L4         -0.1506      0.121     -1.243      0.214      -0.388       0.087
ma.L1         -0.8434      0.074    -11.397      0.000      -0.988      -0.698
ma.S.L12      -1.0094      2.461     -0.410      0.682      -5.833       3.814
sigma2         0.0041      0.010      0.395      0.693      -0.016       0.024
===================================================================================
Ljung-Box (Q):                       45.57   Jarque-Bera (JB):                  3.84
Prob(Q):                              0.25   Prob(JB):                          0.15
Heteroskedasticity (H):               1.60   Skew:                              0.44
Prob(H) (two-sided):                  0.17   Kurtosis:                          3.34
===================================================================================
```

The model summary indicates that that none of the terms used in the model are significant in terms of p-values.

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.

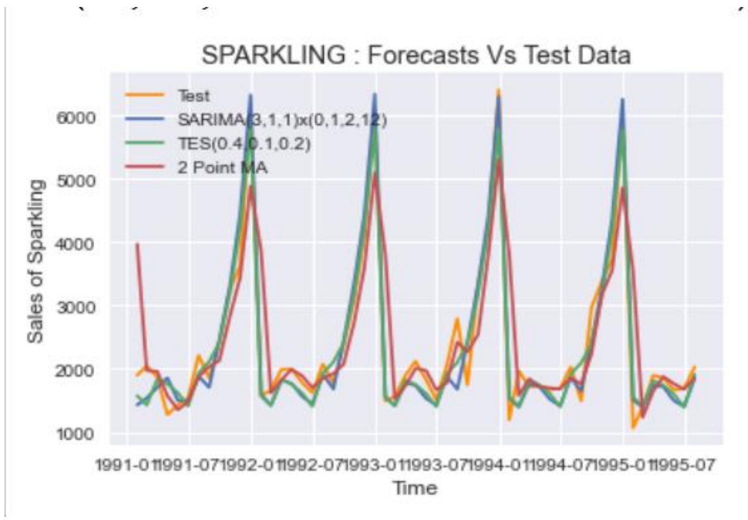
SPARKLING-

The overall comparison of all the time-series forecast models are listed below 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

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	312.211095	10.20
Manual SARIMA(3,1,1)x(1,1,2,12)	324.105224	9.48
Auto SARIMA(3,1,3)x(3,1,0,12)	331.585770	10.33
TES Alpha 0.15, Beta 0.00, Gamma 0.37	384.203001	11.94
2 point MA	813.400684	19.70
4 point MA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
SES Alpha 0.00	1275.081823	38.90
6 point MA	1283.927428	43.86
9 point MA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.430000	67.23
DES Alpha 0.6,Beta 0.0	3851.171500	152.07
NaiveModel	3864.279352	152.87



ROSE-

The overall comparison of all the time-series forecast models are listed below in accordance with increasing RMSE against test data or in the order of decreasing accuracy

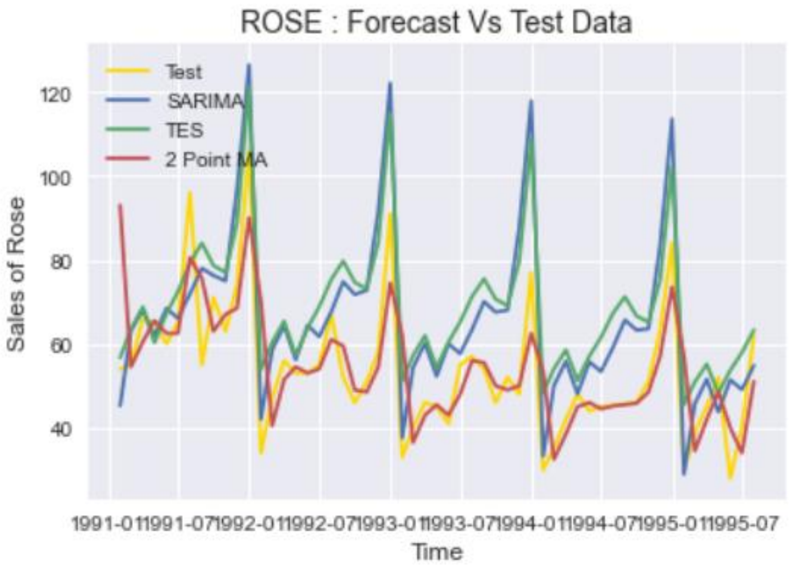
Triple Exponential Smoothing is found to be the best model, followed by 2 point Moving Average

The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data

2 point trailing moving average is found to be having the best fitment against the test data, through 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

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.640687	13.96
2 point TMA	11.529278	13.54
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.593960	21.93
Manual SARIMA(4,1,2)x(0,1,1,12)	13.593960	21.93
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.179837	23.10
4 point TMA	14.451403	19.49
6 point TMA	14.566327	20.82
9 point TMA	14.727630	21.01
RegressionOnTime	15.268955	22.82
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377252	22.16
Auto SARIMA(3,1,1)x(3,1,1,12)	16.823650	25.48
TES Alpha 0.11, Beta 0.05, Gamma 0.00	17.369487	28.88
SES Alpha 0.01	36.796242	63.88
DES Alpha 0.10, Beta 0.10	37.057142	64.02
SimpleAverage	53.460570	94.93
DES Alpha 0.16, Beta 0.16	70.572452	120.25
NaiveModel	79.718773	145.10



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.

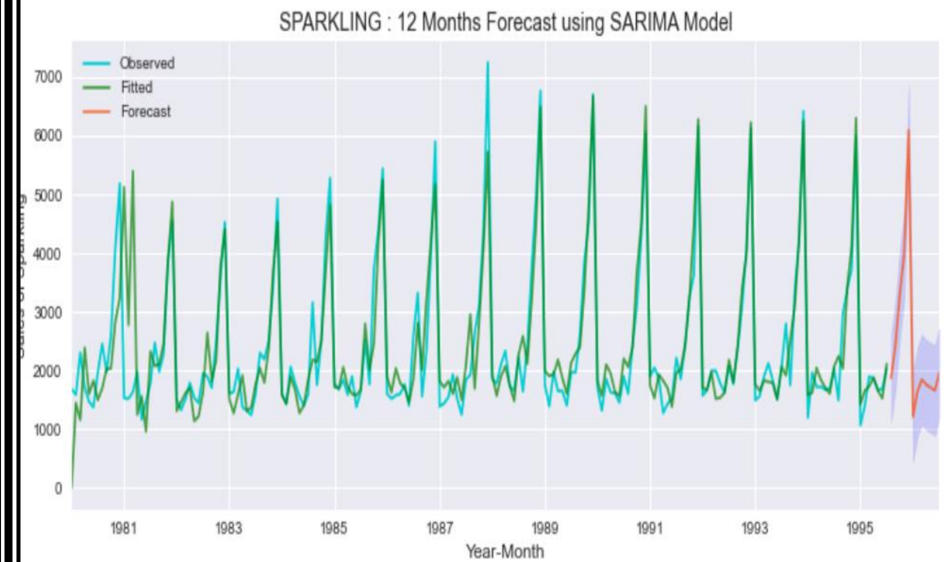
SPARKLING-

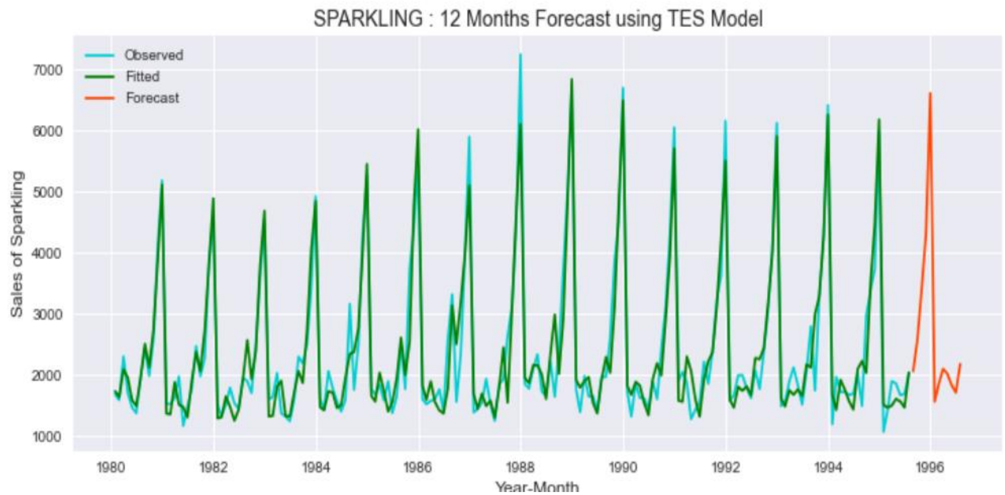
Based on the overall model evaluation and comparison, Triple Exponential Smoothing (Holt Winter's) and SARIMA are selected for final prediction into 12 months in future.

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 below The SARIMA model is built with parameters (3, 1, 3) x (1, 1, 2, 12), is found to be the most optimal SARIMA model





TES forecast on the Sparkling Full Data: RMSE is 376.821 and MAPE is 11.30

For SARIMA forecast on the Sparkling Full Data: RMSE is 591.252 and MAPE is 14.86

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

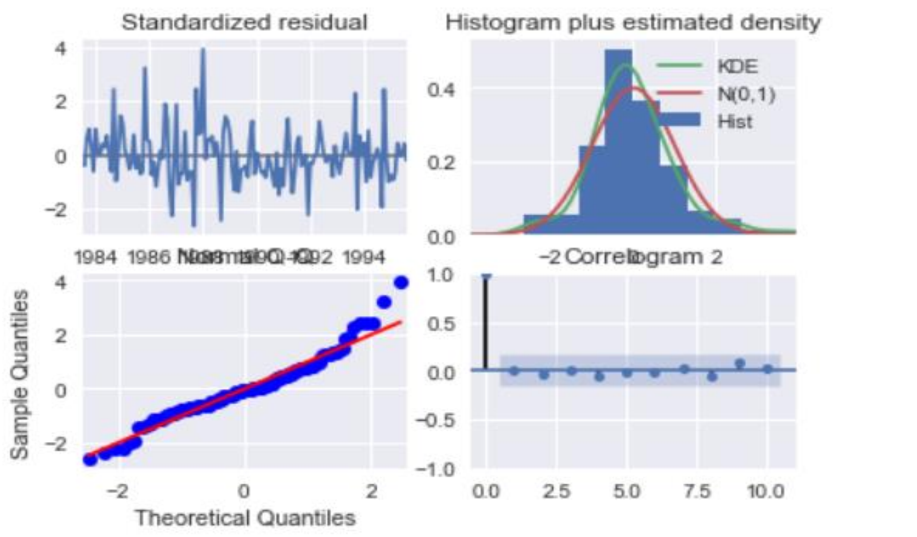
The model predicts an upward trend and continuation of the seasonal surge in sales in the upcoming 12 months. According to the model the seasonal sale will be more than that of the previous year

The 12 month prediction of the TES model is as below • The SARIMA model is built with parameters (3, 1, 3)x(1, 1, 2, 12), is found to be the most optimal SARIMA model

SARIMA model has reflected the trend and seasonality of the series continuing into the future year as well. The seasonal altitude predicted us more conservative than TES model

SARIMA model is seen to have better hold with the most recent observed data and shows high variations in the farthest periods of observations, which explains the high RMSE and MAPE values.

```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                 187
Model:             SARIMAX(3, 1, 3)x(1, 1, 2, 12)   Log Likelihood               -1078.437
Date:                            Sun, 16 Jan 2022   AIC                           2176.875
Time:                                    19:55:03   BIC                           2206.711
Sample:                                01-31-1980   HQIC                          2188.998
                                     - 07-31-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4230      0.086     -4.916      0.000      -0.592      -0.254
ar.L2         -0.9094      0.053    -17.296      0.000      -1.012      -0.806
ar.L3          0.1424      0.087      1.638      0.101      -0.028       0.313
ma.L1         -0.4114      0.078     -5.277      0.000      -0.564      -0.259
ma.L2          0.4622      0.083      5.578      0.000       0.300       0.625
ma.L3         -0.9673      0.104     -9.316      0.000      -1.171      -0.764
ar.S.L12      -0.0697      0.709     -0.098      0.922      -1.459       1.320
ma.S.L12      -0.4553      0.721     -0.632      0.528      -1.868       0.958
ma.S.L24      -0.0809      0.396     -0.204      0.838      -0.858       0.696
sigma2      1.461e+05   1.06e-06   1.38e+11      0.000    1.46e+05    1.46e+05
===================================================================================
Ljung-Box (Q):                       17.11   Jarque-Bera (JB):                35.59
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.72   Skew:                             0.66
Prob(H) (two-sided):                  0.26   Kurtosis:                         5.03
===================================================================================
```



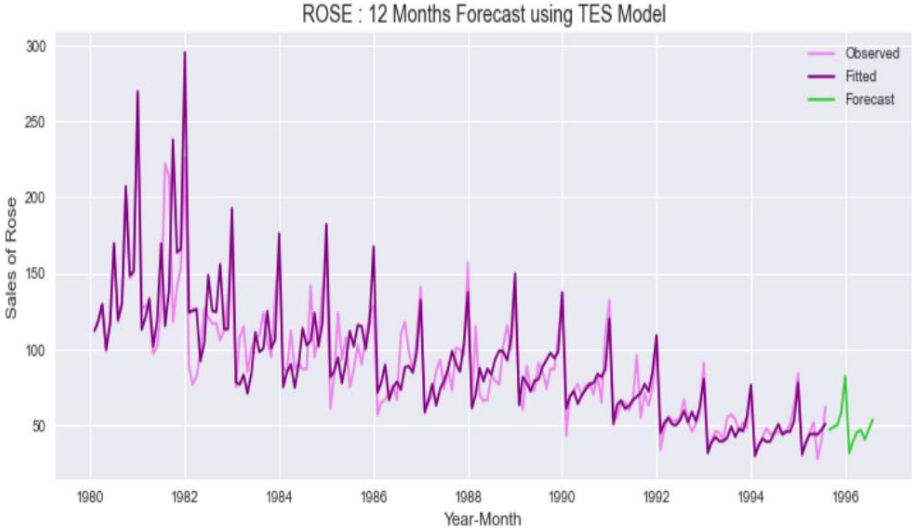
The SARIMA model built on the complete Sparkling timeseries is chosen, as prediction provide confidence interval which give better explainabiity and confidence to the forecasts

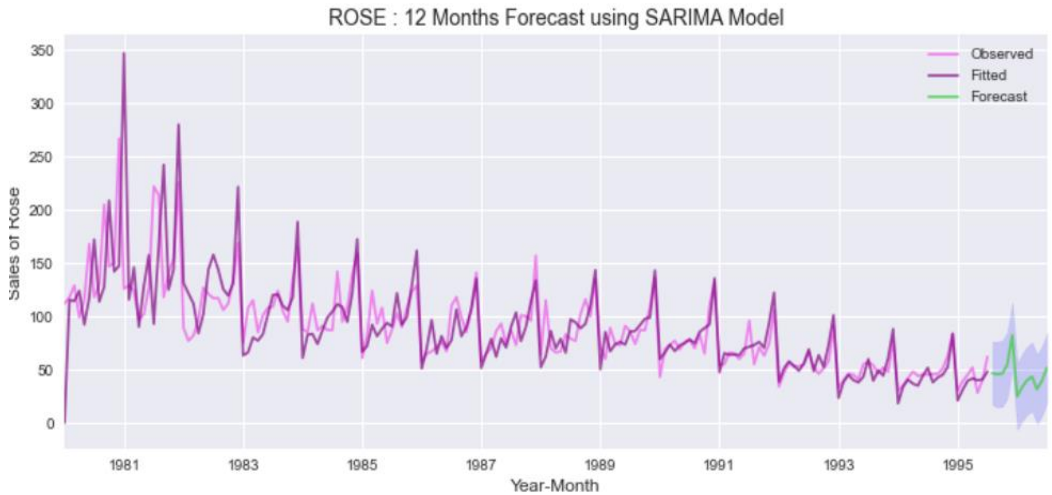
The diagnostics plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in normal QQ plot

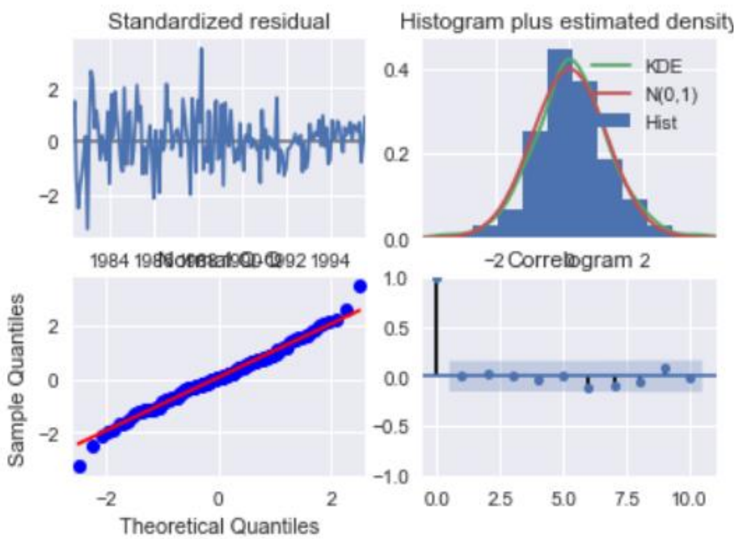
The model summary also provides valuable insights in the model. From the snapshot of summary below it can be understood that AR(2), MA(3) terms has the highest absolute weightage.

The p-values indicates 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 got values higher than alpha 0.05, which fails to reject the hull hypothesis that these terms are not significant

ROSE-








```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  187
Model:             SARIMAX(4, 1, 1)x(0, 1, 1, 12)   Log Likelihood                -664.135
Date:                            Sun, 16 Jan 2022   AIC                           1342.270
Time:                                    18:32:15   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         -0.1315      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       232.4253     24.359      9.542      0.000     184.682     280.169
===================================================================================
Ljung-Box (Q):                       35.39   Jarque-Bera (JB):                 5.30
Prob(Q):                              0.68   Prob(JB):                         0.07
Heteroskedasticity (H):               0.22   Skew:                             0.04
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.89
===================================================================================
```

TES forecast on the Rose Full Data: RMSE is 20.881 and MAPE is 14.48

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

The SARIMA model is chosen as the final model for prediction on Rose dataset, as it provide confidence interval and better explainability of the model
• The diagnostics 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 normal QQ plot

The model summary also provides valuable insights in the model. From the snapshot of summary below it can be understood that MA(1) and seasonal MA(1) term has the highest weightage.

The p-values indicates that the terms MA(1) and Seasonal MA(1) are the most significant terms • The rest of the p-values got values higher than alpha 0.05, which fails to reject the hull hypothesis that these terms are not significant

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

SARIMA model has also reflected the trend and seasonality of the series continuing into the future year as well.

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.

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-

	Sparkling
1995-08-31	1873.38
1995-09-30	2445.07
1995-10-31	3312.75
1995-11-30	3994.67
1995-12-31	6084.21
1996-01-31	1216.28
1996-02-29	1640.61
1996-03-31	1847.32
1996-04-30	1762.20
1996-05-31	1708.40
1996-06-30	1663.95
1996-07-31	1961.45

	Sparkling
count	12.000000
mean	2459.190833
std	1384.633922
min	1216.280000
25%	1697.287500
50%	1860.350000
75%	2661.990000
max	6084.210000

The model forecasts sale of 29510 units of Sparkling wine in 12 months into future. Which is an average sale of 2459 units per month

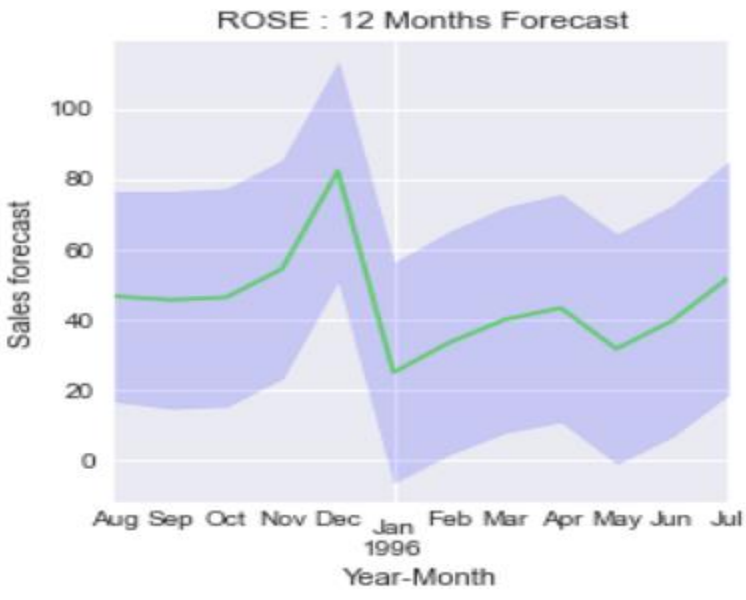
The seasonal sale in December 1995 will hit a maximum of 6084 units, before it drops to the lowest sale in January 1996; at 1216 units.

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), which is 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 the sale compared to previous years

Adding more exogenous variable into the timeseries data can improve forecasts.

ROSE



	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

	ROSE
count	12.000000
mean	44.905000
std	14.473222
min	24.810000
25%	38.007500
50%	44.370000
75%	47.830000
max	82.210000

The model forecasts sale of 539 units of Rose wine in 12 months into future.

Which is an average sale of 45 units per month The seasonal sale in December 1995 will reach a maximum of 82 units, before it drops to the lowest sale in January 1996; at 25 units.

Unlike Sparkling wine, Rose wine sells 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 sale 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 market and make corrective actions in marketing and promotion