Mandira Roy

Time series

Data science

OVERVIEW

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

Data set for the Problem: Sparkling.csv and Rose.csv

GOALS

1. Read the data as an appropriate Time Series data and plot the data.
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.
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.
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.

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.
8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
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.
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

OBSERVATION

1.Reading 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 .The Rose time-series got values missing for two months in 1994, which are imputed using interpolation (forward fill) .Rose data after interpolation for year 1994 is given below as well as the plot .Both the datasets show significant seasonality. While sale of Rose shows evident downward trend, Sparkling doesn't shows 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 years .

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

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

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

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

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

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

	Rose
count	187.000000
mean	89.909091
std	39.244440
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

	Sparkling
count	187.00
mean	2402.42
std	1295.11
min	1070.00
25%	1605.00
50%	1874.00
75%	2549.00
max	7242.00

```
dfr.isnull().valu
```

```
Rose
False    187
dtype: int64
```

```
dfr.isnull().valu
```

```
Rose
False    187
dtype: int64
```

```
dfs.isnull().valu
```

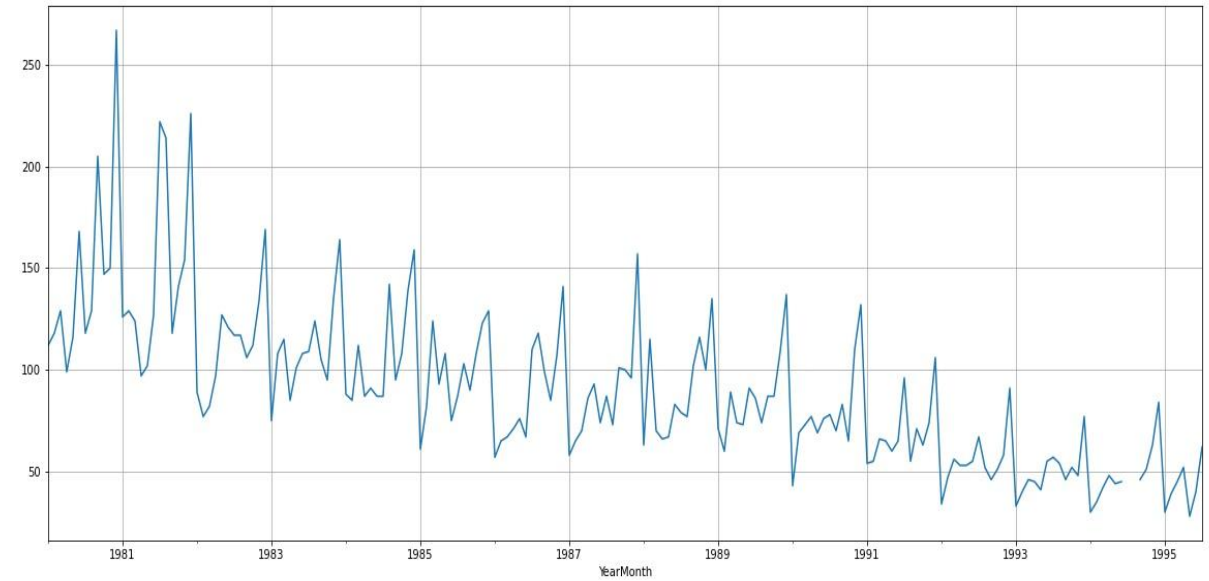
```
Sparkling
False        187
dtype: int64
```

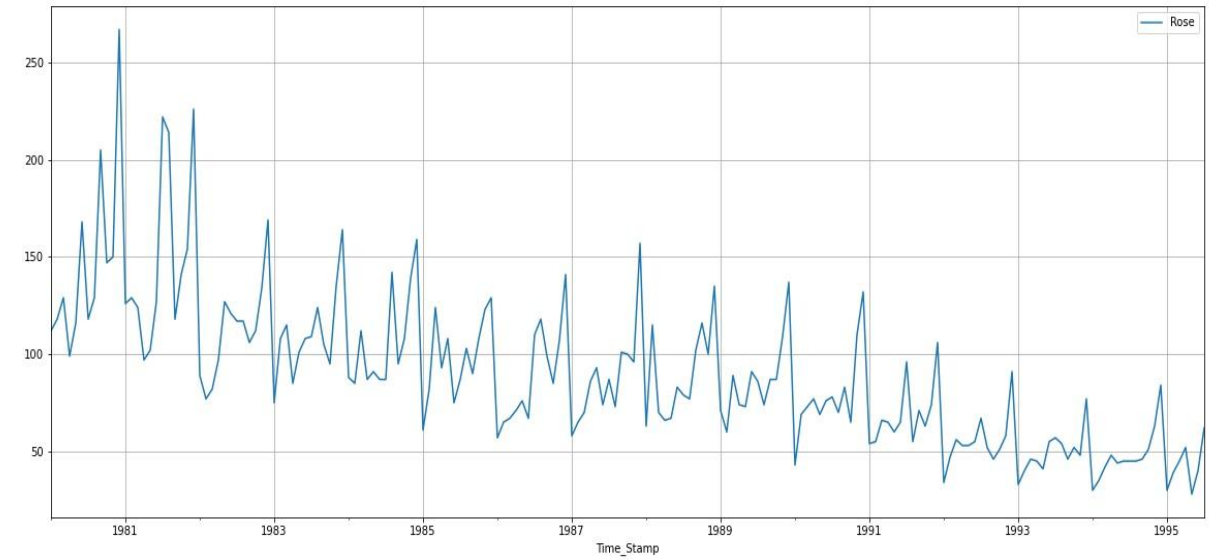
```
dfs.isnull().valu
```

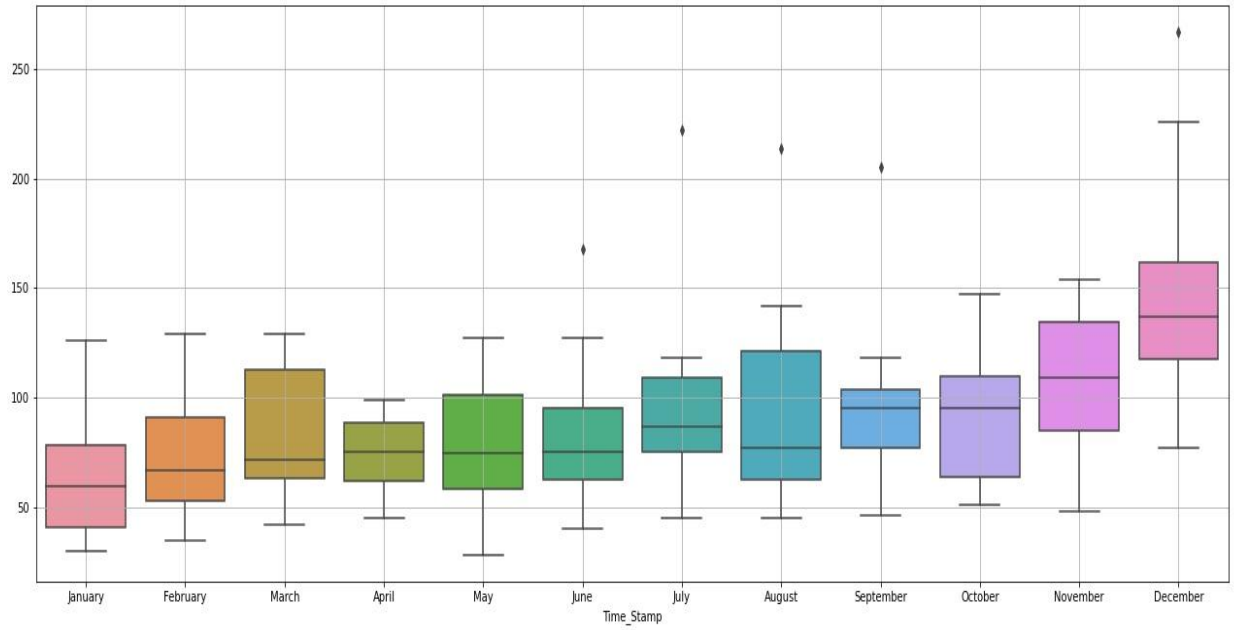
```
Sparkling
False        187
dtype: int64
```

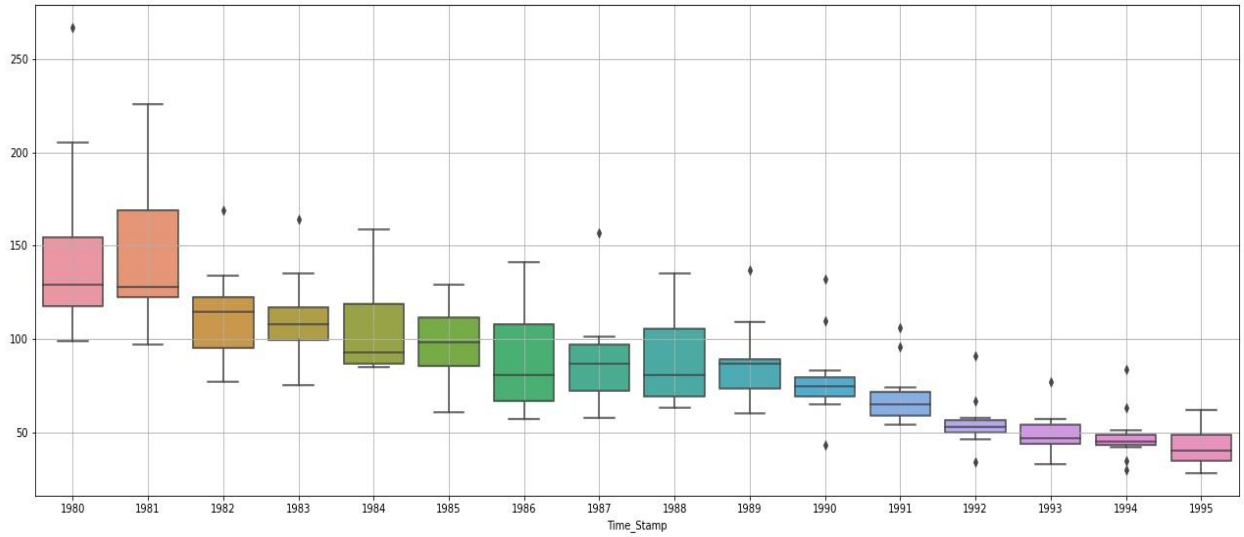
2.EXPLORATORY DATA ANALYSIS:

For the rose data set.


250
200
150
100
50
1981
1983
1985
1987
1989
1991
1993
1995
YearMonth


Rose
250
200
150
100
50
1981
1983
1985
1987
1989
1991
1993
1995
Time_Stamp

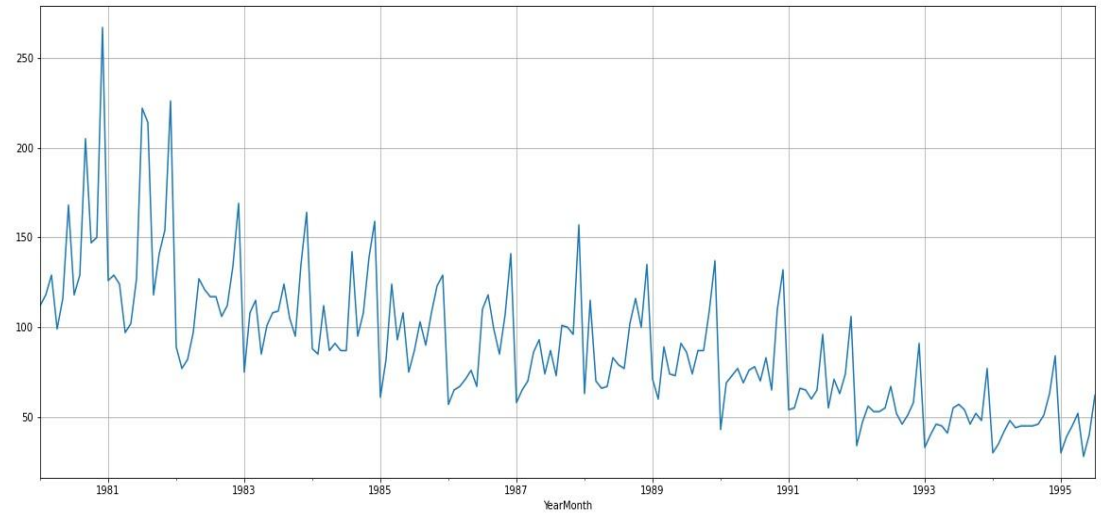

250
200
150
100
50
January
February
March
April
May
June
July
August
September
October
November
December
Time_Stamp

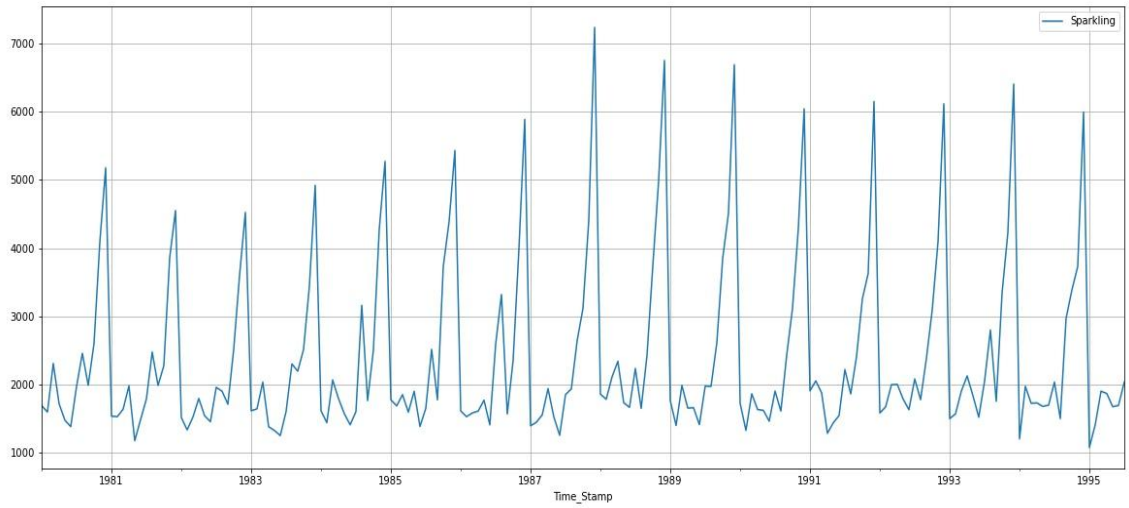


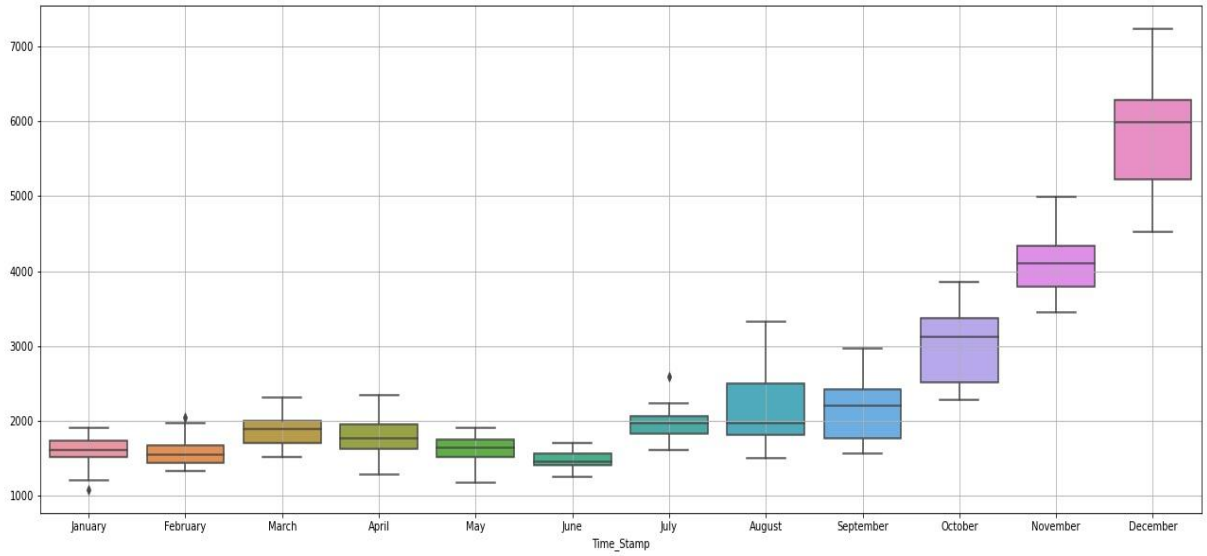
The descriptive summary of the data shows that on an average 90 units of Rose wines were sold each month in the given period of time. 50% of monthly sales varied from 63 units to 112 units. Maximum sale reported in a month is 267 units and minimum of 28 units .The Empirical CDF plot shows that, in 80% of months, at least 120 units of Rose wine were sold .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

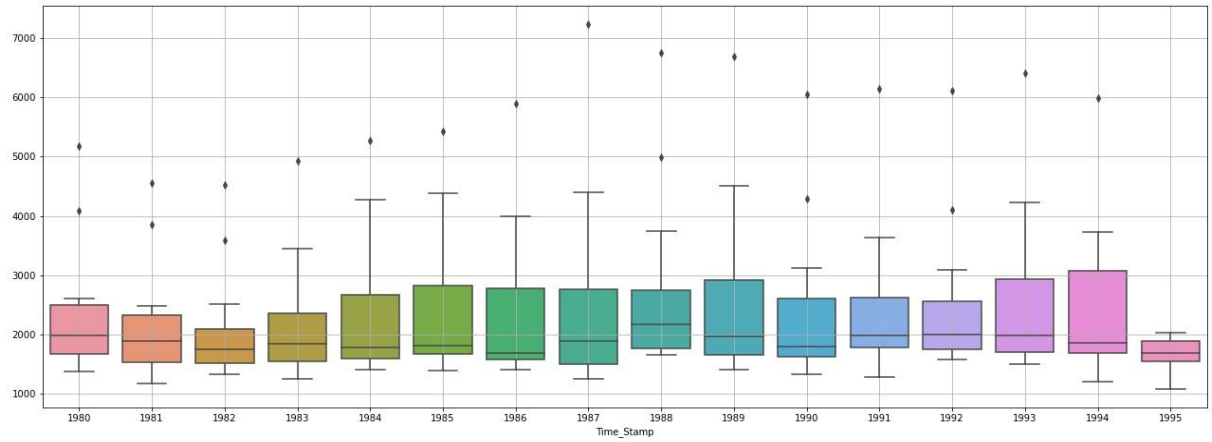
The monthly-box-plot shows a clear seasonality during the seasonal months of November and December. Though the sale tanks in the month of January, it picks up in the due course of the year. Average sale in December is around 140 units, November is around 110 units and October is around 90 units The monthly plot for Rose shows mean and variation of units sold each month over the years. Sale in months such as July, August, September and December shows a higher variation than the rest Sale in December with a mean few points below 100, varies from 75 to 270 units

For Sparkling dataset:

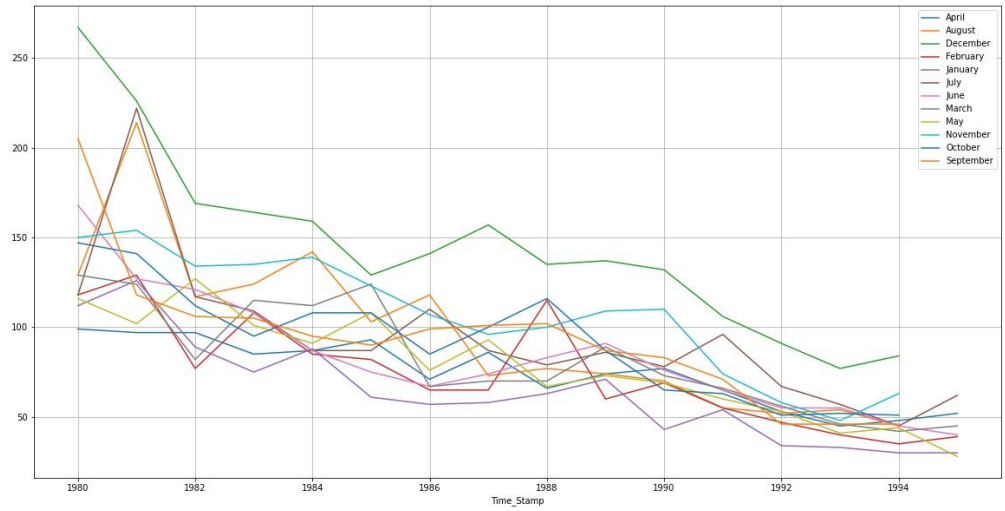

250
200
150
100
50
1981
1983
1985
1987
1989
1991
1993
1995
YearMonth

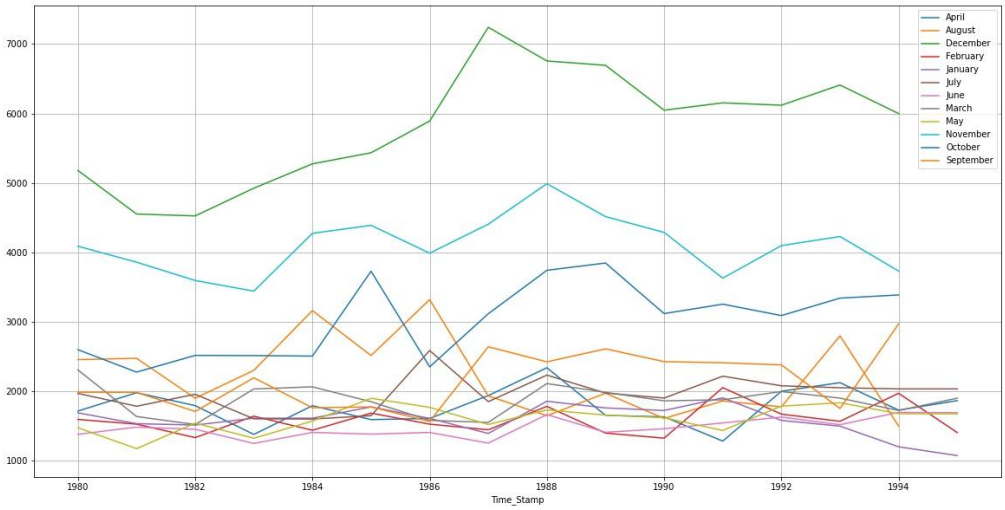

Sparkling
7000
6000
5000
4000
3000
2000
1000
1981
1983
1985
1987
1989
1991
1993
1995
Time_Stamp





The monthly plot for Sparkling shows mean 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 points 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 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 were 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


April
August
December
February
January
July
June
March
May
November
October
September
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
Time_Stamp


April
August
December
February
January
July
June
March
May
November
October
September
7000
6000
5000
4000
3000
2000
1000
1980
1982
1984
1986
1988
1990
1992
1994
Time_Stamp

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

ıe_Stamp	April	August	December	February	January	July	June	March	May	November	October	September
ıe_Stamp												
1980	1712.0	2453.0	5179.0	1591.0	1686.0	1966.0	1377.0	2304.0	1471.0	4087.0	2596.0	1984.0
1981	1976.0	2472.0	4551.0	1523.0	1530.0	1781.0	1480.0	1633.0	1170.0	3857.0	2273.0	1981.0
1982	1790.0	1897.0	4524.0	1329.0	1510.0	1954.0	1449.0	1518.0	1537.0	3593.0	2514.0	1706.0
1983	1375.0	2298.0	4923.0	1638.0	1609.0	1600.0	1245.0	2030.0	1320.0	3440.0	2511.0	2191.0
1984	1789.0	3159.0	5274.0	1435.0	1609.0	1597.0	1404.0	2061.0	1567.0	4273.0	2504.0	1759.0
1985	1589.0	2512.0	5434.0	1682.0	1771.0	1645.0	1379.0	1846.0	1896.0	4388.0	3727.0	1771.0
1986	1605.0	3318.0	5891.0	1523.0	1606.0	2584.0	1403.0	1577.0	1765.0	3987.0	2349.0	1562.0
1987	1935.0	1930.0	7242.0	1442.0	1389.0	1847.0	1250.0	1548.0	1518.0	4405.0	3114.0	2638.0
1988	2336.0	1645.0	6757.0	1779.0	1853.0	2230.0	1661.0	2108.0	1728.0	4988.0	3740.0	2421.0
1989	1650.0	1968.0	6694.0	1394.0	1757.0	1971.0	1406.0	1982.0	1654.0	4514.0	3845.0	2608.0
1990	1628.0	1605.0	6047.0	1321.0	1720.0	1899.0	1457.0	1859.0	1615.0	4286.0	3116.0	2424.0
1991	1279.0	1857.0	6153.0	2049.0	1902.0	2214.0	1540.0	1874.0	1432.0	3627.0	3252.0	2408.0
1992	1997.0	1773.0	6119.0	1667.0	1577.0	2076.0	1625.0	1993.0	1783.0	4096.0	3088.0	2377.0
1993	2121.0	2795.0	6410.0	1564.0	1494.0	2048.0	1515.0	1898.0	1831.0	4227.0	3339.0	1749.0
1994	1725.0	1495.0	5999.0	1968.0	1197.0	2031.0	1693.0	1720.0	1674.0	3729.0	3385.0	2968.0
1995	1862.0	NaN	NaN	1402.0	1070.0	2031.0	1688.0	1897.0	1670.0	NaN	NaN	NaN

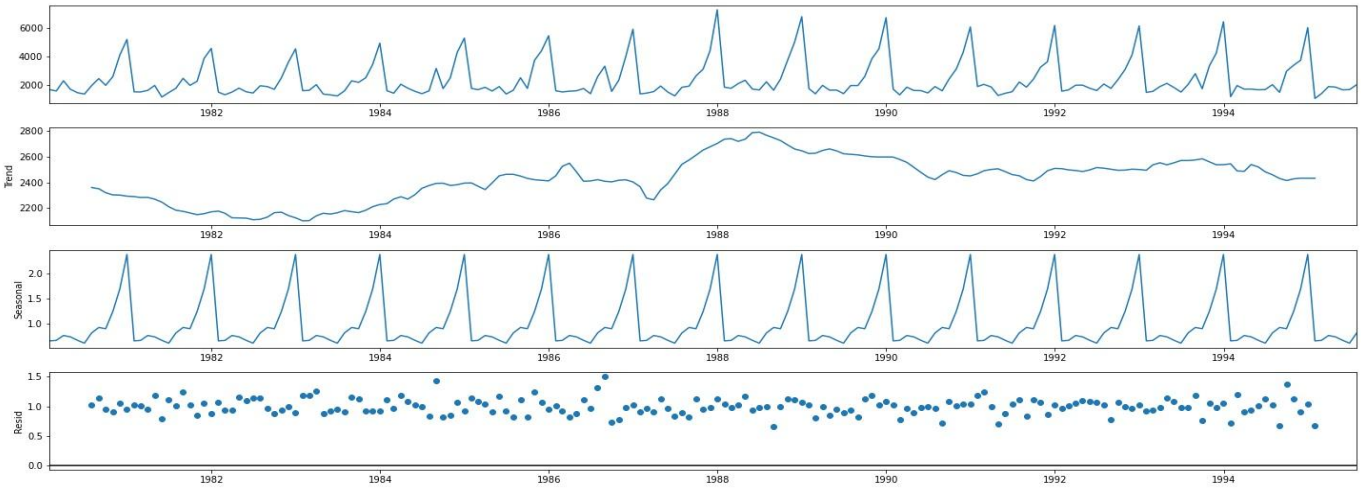
DECOMPOSITION OF DATA:

Sparkling:

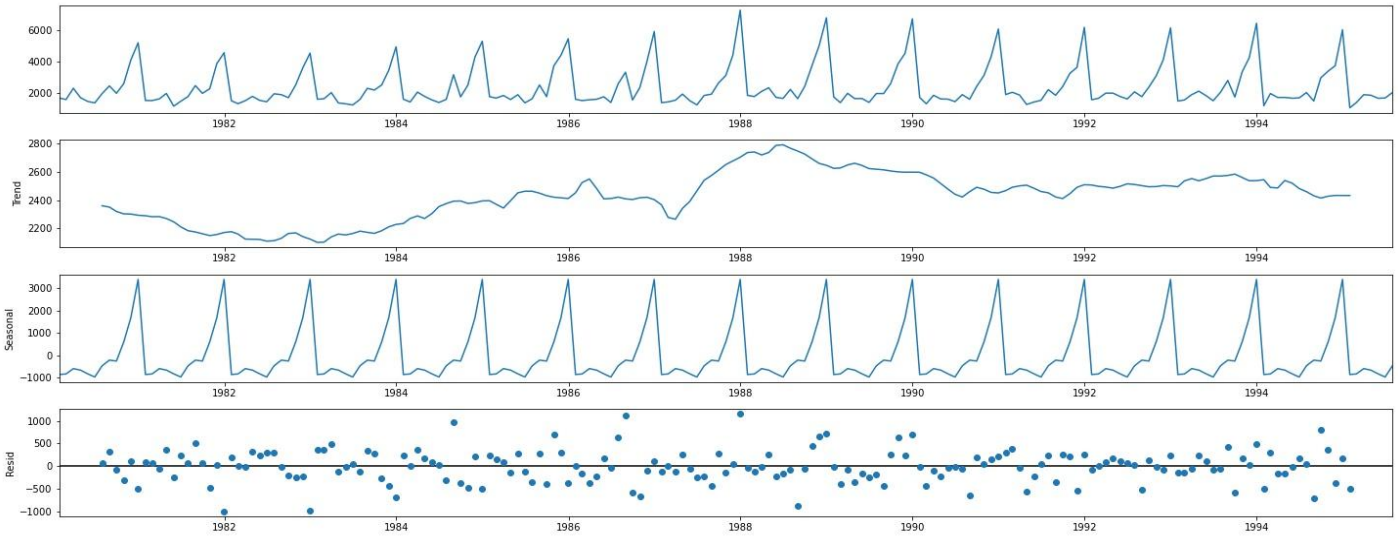
The decomposition plots of Sparkling wine sales is given below. As the altitude of the seasonal peaks in the observed plot is changing according to the change in trend, the time-series is assumed to be ‘multiplicative’ .The plot of the trend component does not show a consistent trend, but an intermediary period shows an upward slope which gets consistent on the late half.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% .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.

multiplicative:

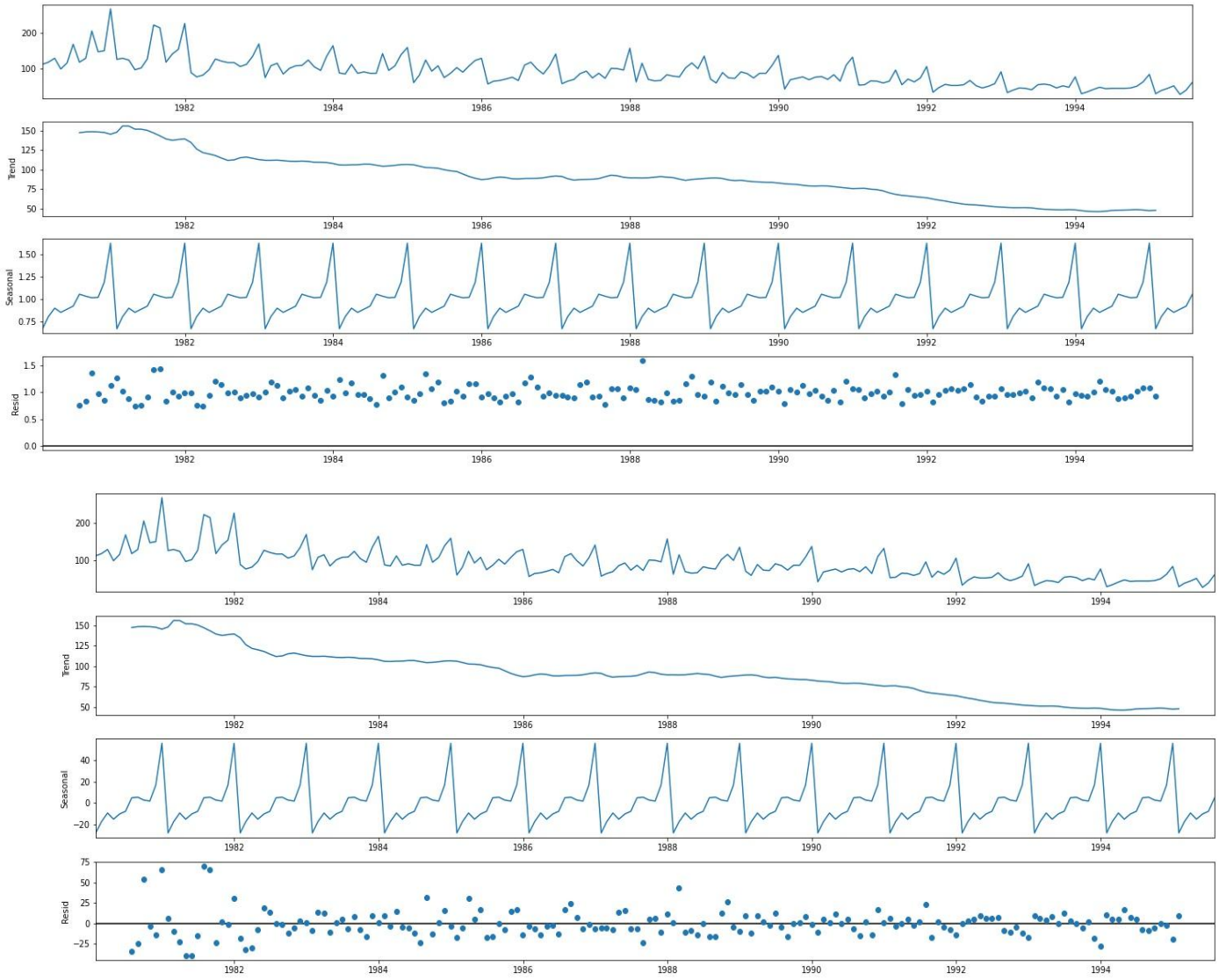


Additive:

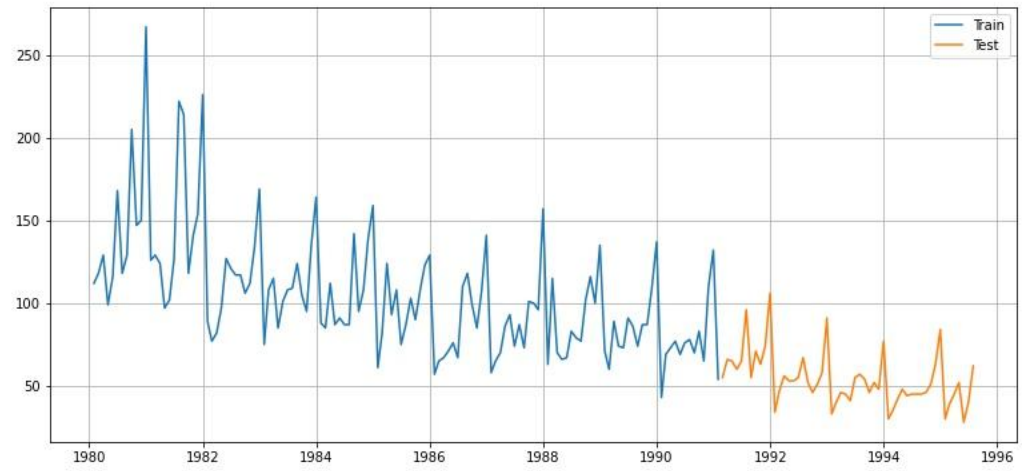


ROSE :

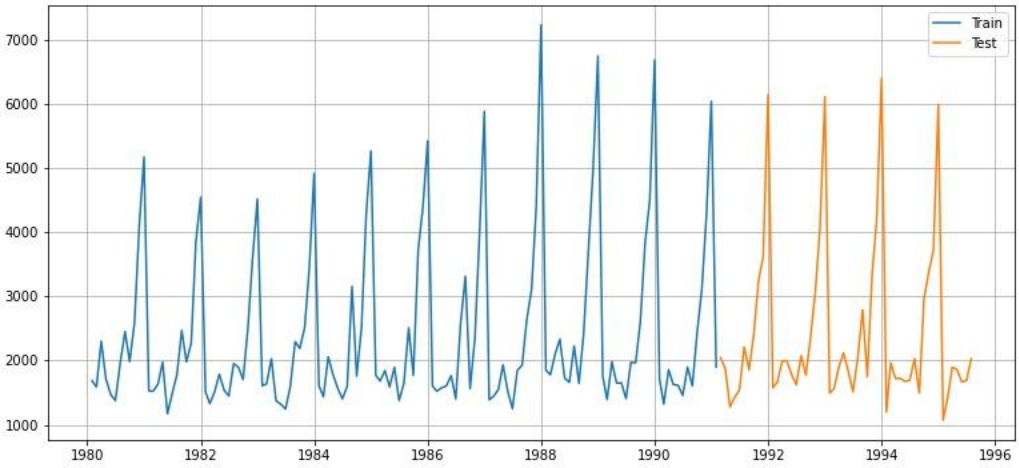
The observed plot of the decomposition diagrams shows visible annual seasonality and a downward trend.The early period of the plot shows higher variation than in the later periods . The residuals shows a pattern of high variability across the period of time-series,which is more or less consistent in both additive and multiplicative decompositions .The 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% .As the seasonality peaks are consistently reducing its altitude in consistent with trend, the series can be treated as multiplicative in model building .



3. SPLIT THE DATA INTO TRAIN AND TEST. -ROSE



SPARKLING:

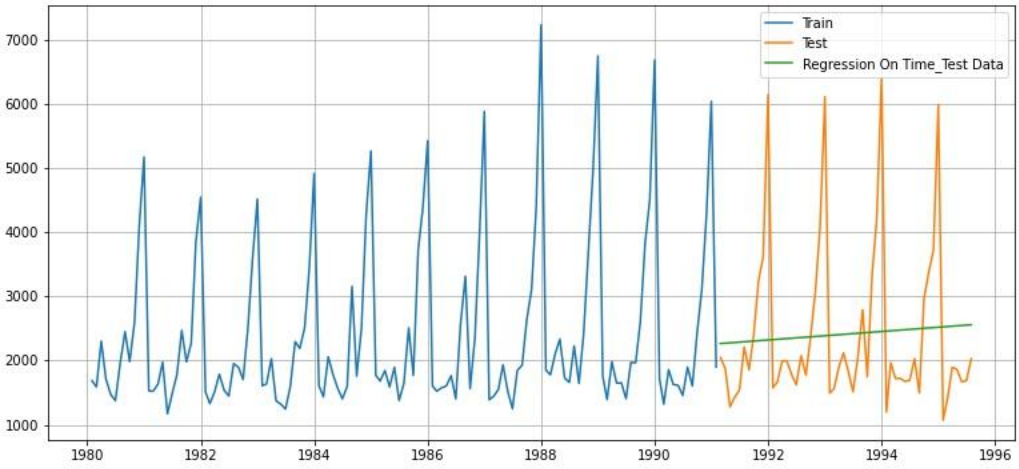


The train and test datasets are created with year 1991 as starting year for test data, using the index.year property of time series index .The plots of Sparkling and Rose time-series as train and test are given here

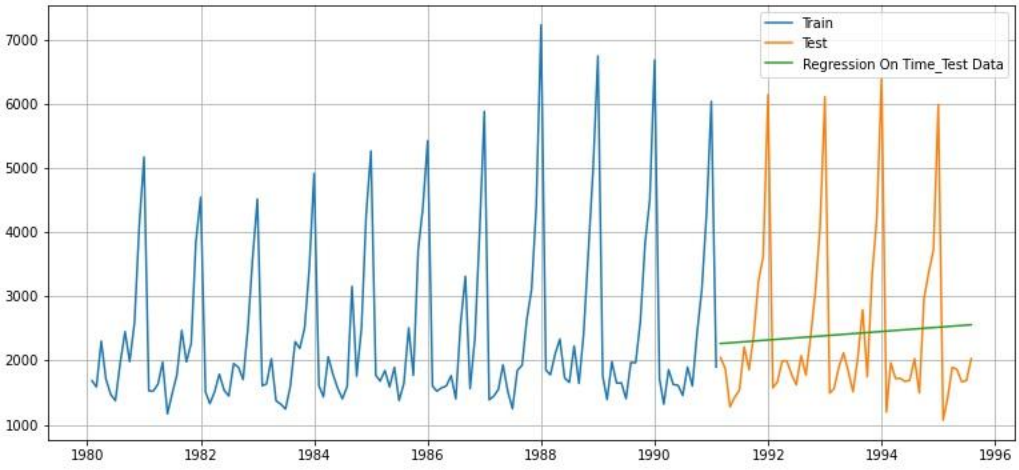
.

4.BUILD DIFFERENT MODELS.

Linear regression model:



Rose- The linear regression on the Rose dataset shows an apparent downward trend as consistent with the observed time-series .The RMSE of the forecast is given above. The model leaves a 23% error in forecast against the test set.



Sparkling: The linear regression plots show a gradual upward trend in the forecast of Sparkling wine,consistent with the observed trend which was not visually apparent . The RMSE values for Train and Test data sets are as above. 50% of the forecast is erroneous.

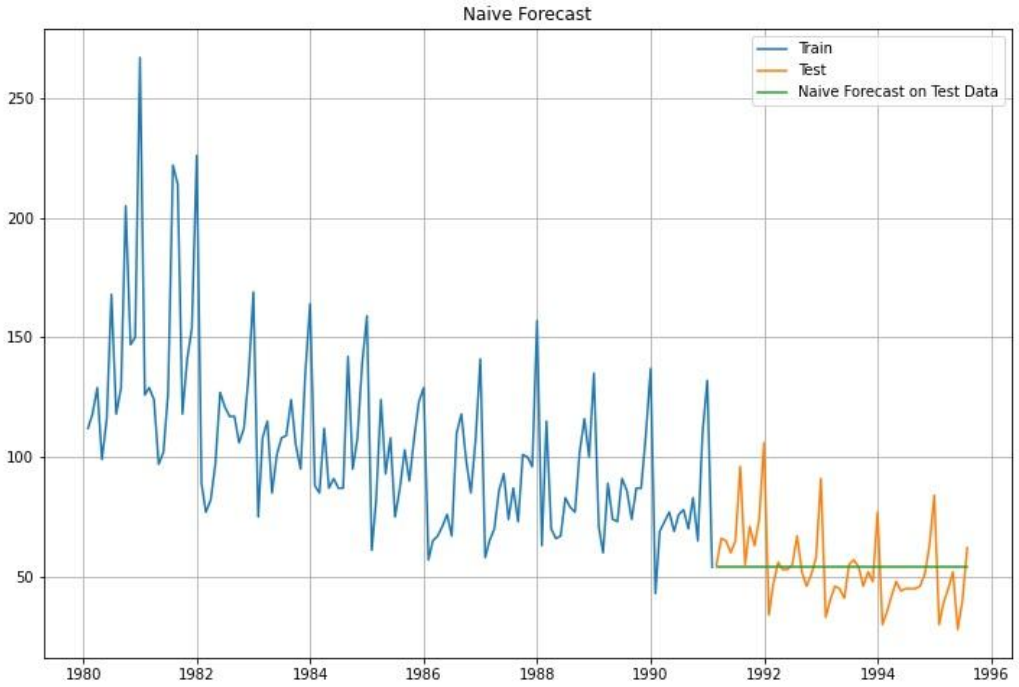
The model has successfully captured the trend of both the series, but does not reflects the seasonality

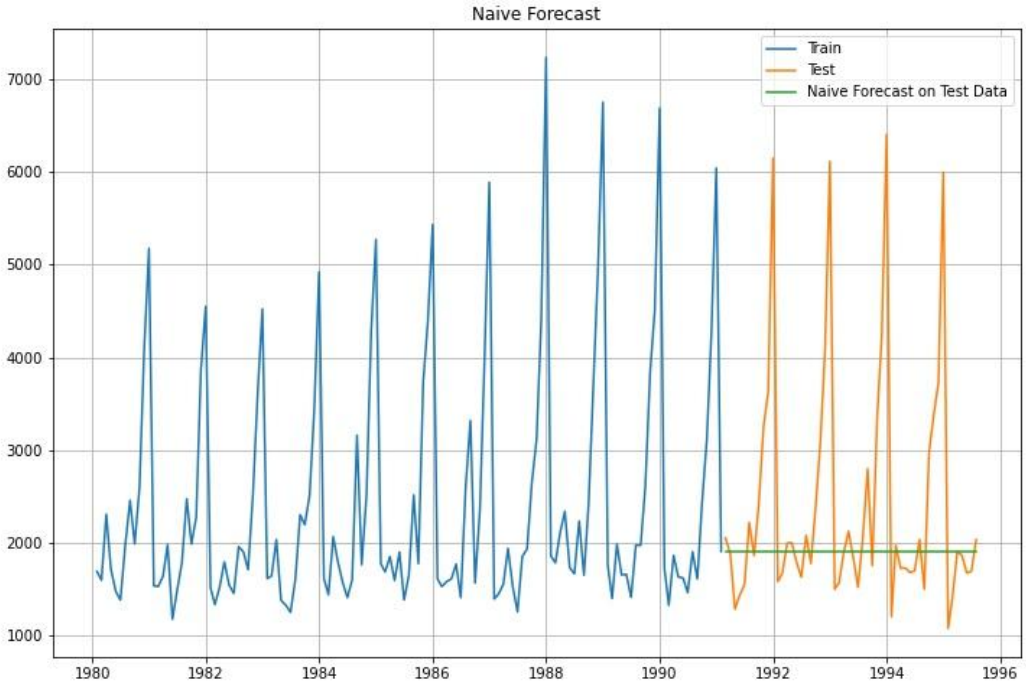
	Rose Test RMSE	Sparkling Test RMSE
Rose RegressionOnTime	51.554113	NaN
Sparkling RegressionOnTime	NaN	1286.31005

NAIVE MODEL:

In a naive model, the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is the same as today, therefore the prediction for day after tomorrow is also today. 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 .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 data sets.The performance metrics above shows a very poor fitment and high percentage of error .

	Rose Test RMSE	Sparkling Test RMSE
Rose RegressionOnTime	51.554113	NaN
Sparkling RegressionOnTime	NaN	1286.310050
NaiveModel	15.915867	NaN
NaiveModel	NaN	1381.177135



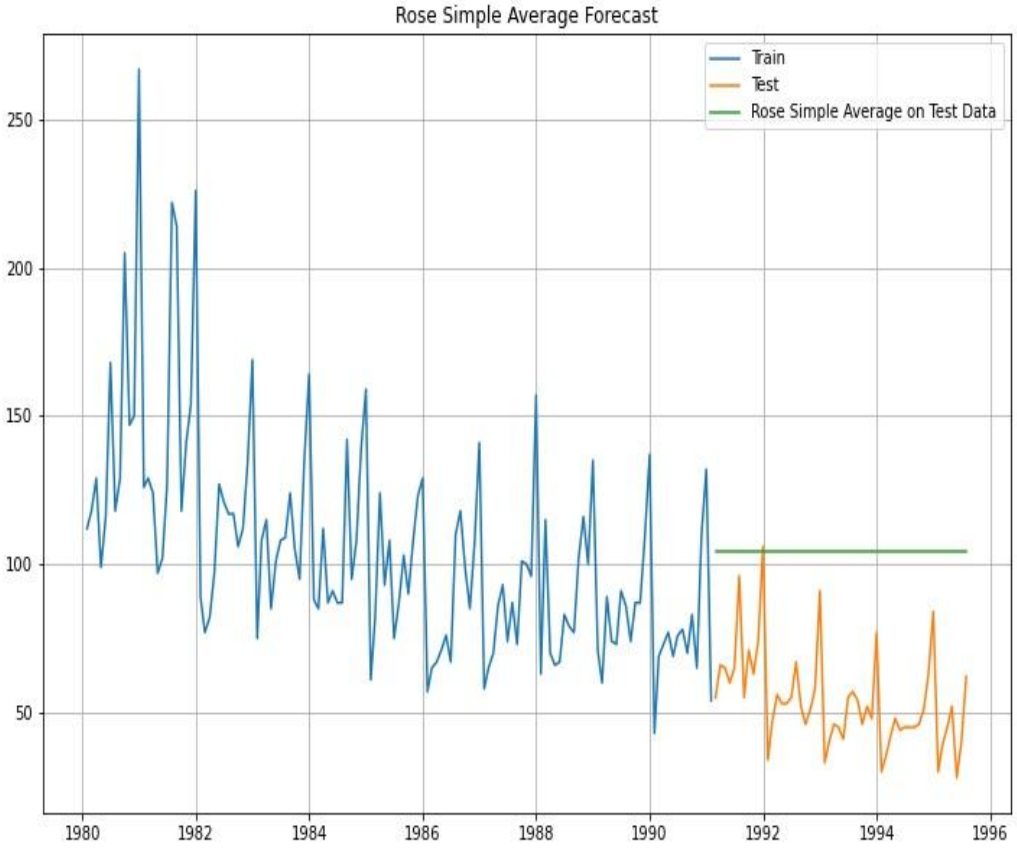


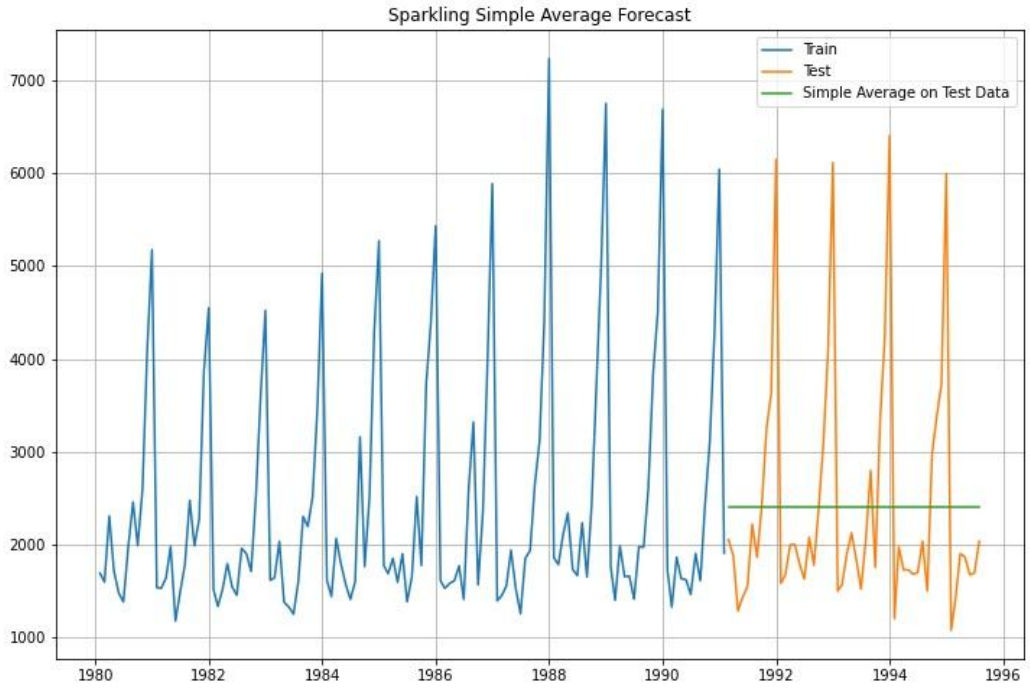
SIMPLE AVERAGE FORECASTING:

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 Rose dataset, 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

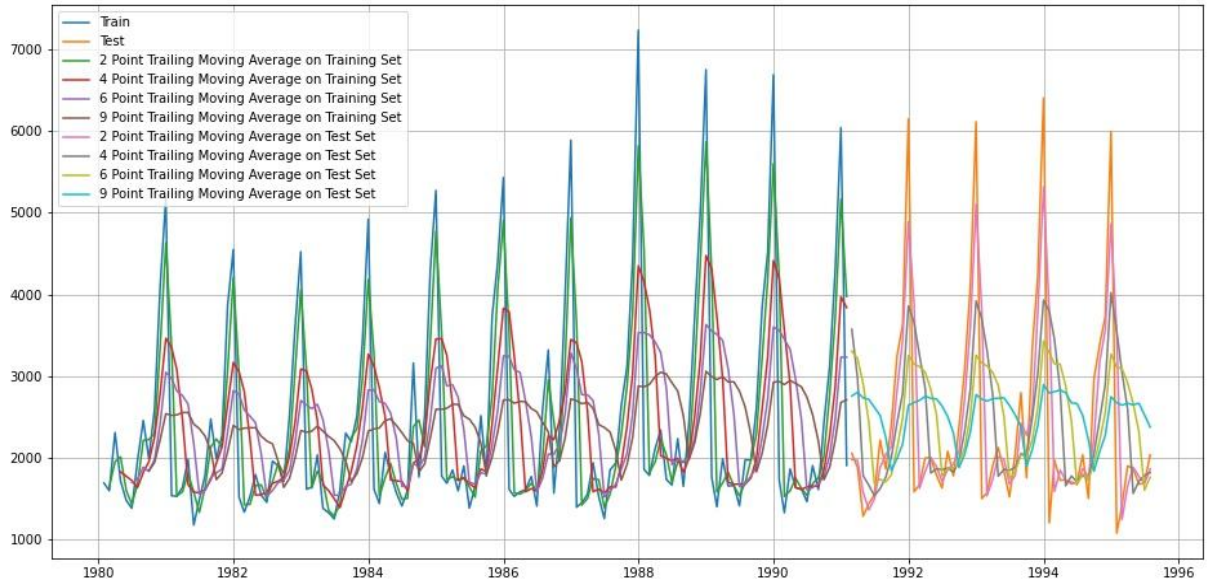
For Sparkling the RMSE is consistent in both test and train datasets.

	Rose Test RMSE	Sparkling Test RMSE
Rose RegressionOnTime	51.554113	NaN
Sparkling RegressionOnTime	NaN	1286.310050
NaiveModel	15.915867	NaN
NaiveModel	NaN	1381.177135
SimpleAverageModel	2346.228164	NaN
SimpleAverageModel	NaN	1285.039964

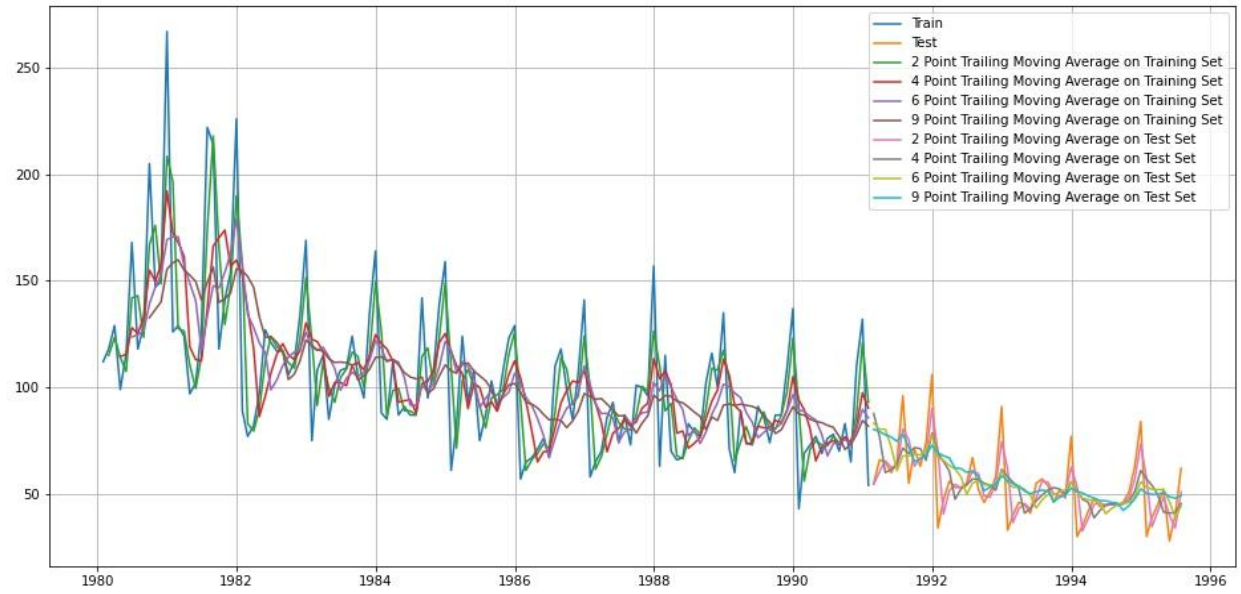




TRAILING MOVING AVERAGE:



Sparkling dataset

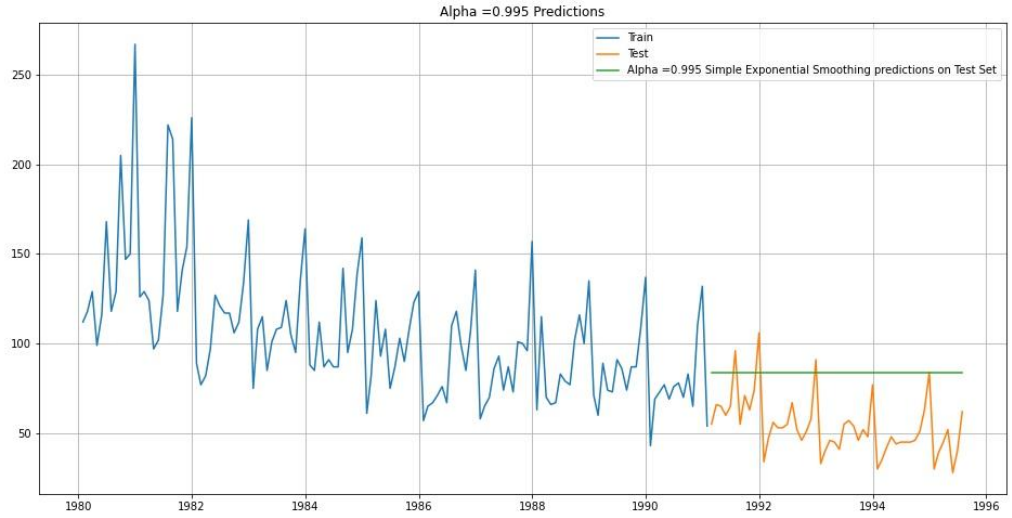


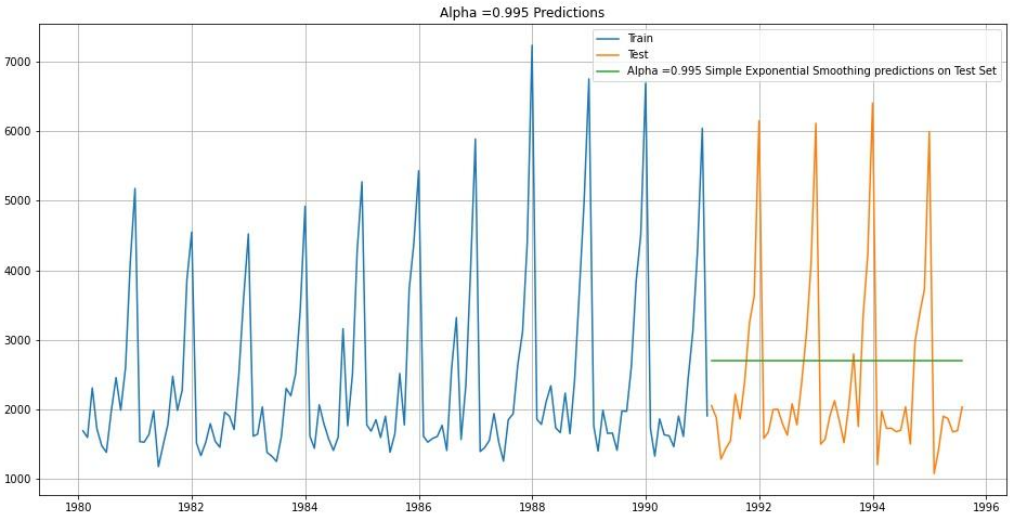
Rose dataset

For the moving average model, we are going to 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 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 Root Mean Squared Error of the test set are given below. The best interval of moving average from the model is 2 points. 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 Root Mean Squared Error of the test set are given below The best interval of moving average from the model is 2 point.

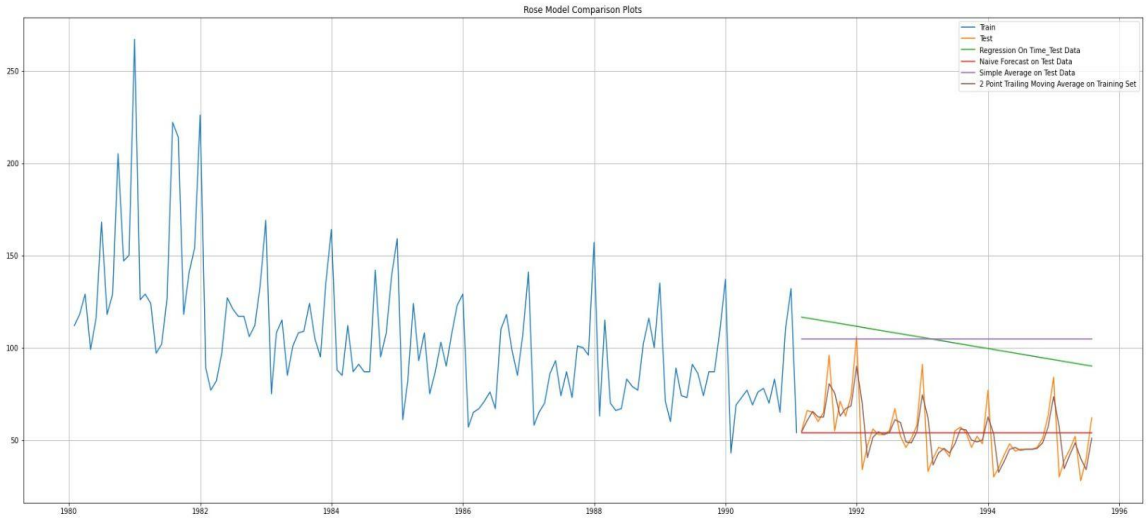
SIMPLE EXPONENTIAL SMOOTHING.
Simple Exponential Smoothing is applied if the time-series has neither a trend nor seasonality, which is not the case with the given data .The forecasting using smoothing levels of alpha between 0 and 1 are as below, where the smoothing levels are passed manually.For alpha value closer to 1, forecasts follows 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 .On the second iteration, the model was ran without passing a value for alpha and used parameters '*optimized=True, use_brute=True'*. The autofit model picked 0.0 as the smoothing parameter

and return 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 smoothed out forecast with an initial value 2403.79 applied across the series.



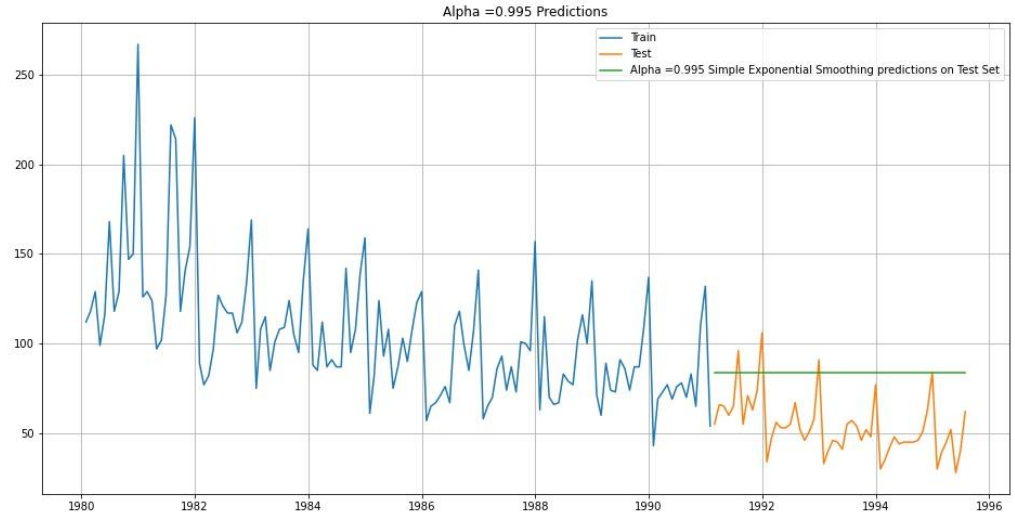


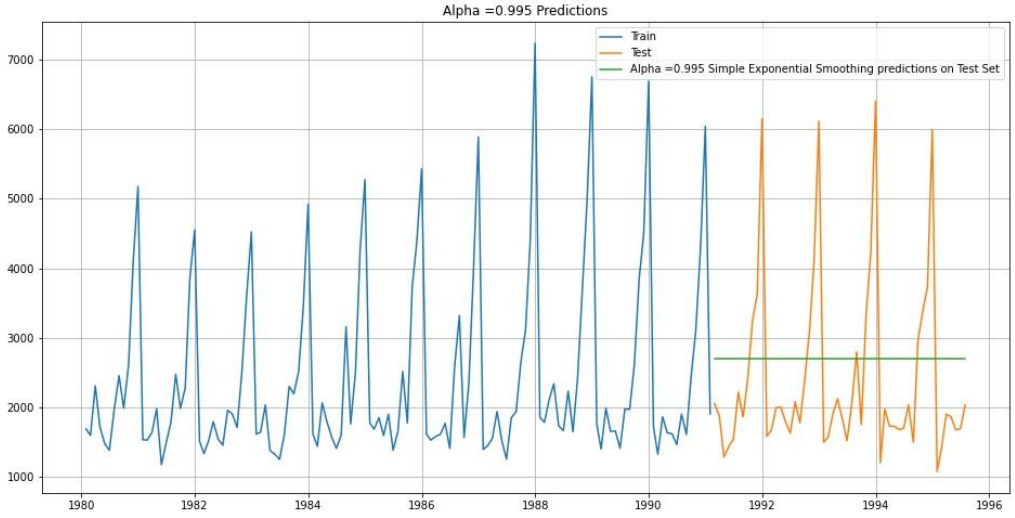
COMPARE DIFFERENT MODEL:

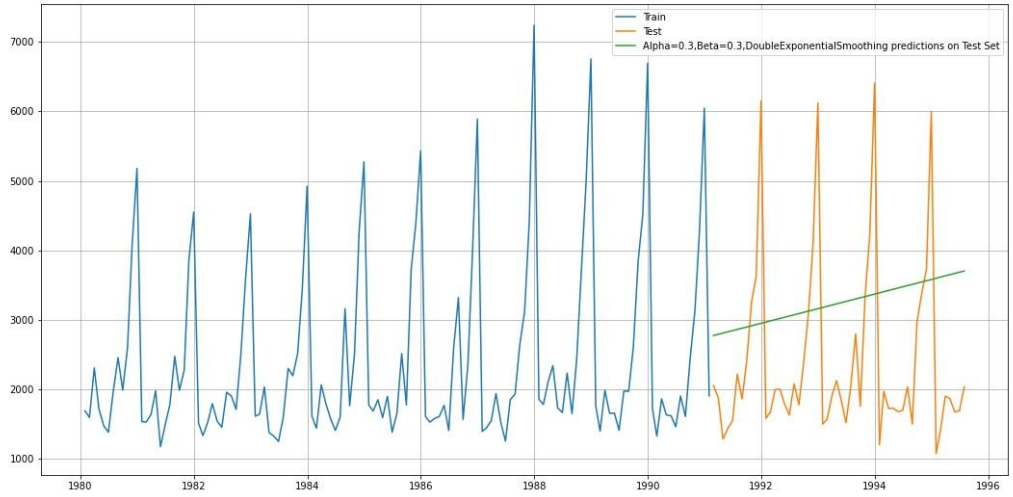


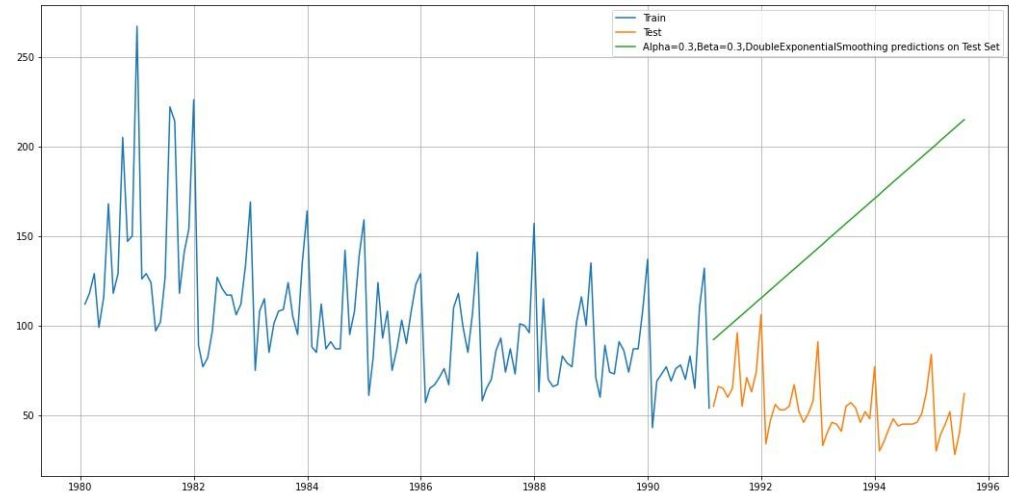
The accuracy of the time-series forecast models build in the previous sections of this report is as below, sorted by RMSE in test data .The plot of the forecasts fitted on to the test data is given as well .From the comparison of accuracy values and the plot it can be inferred that Triple Exponential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

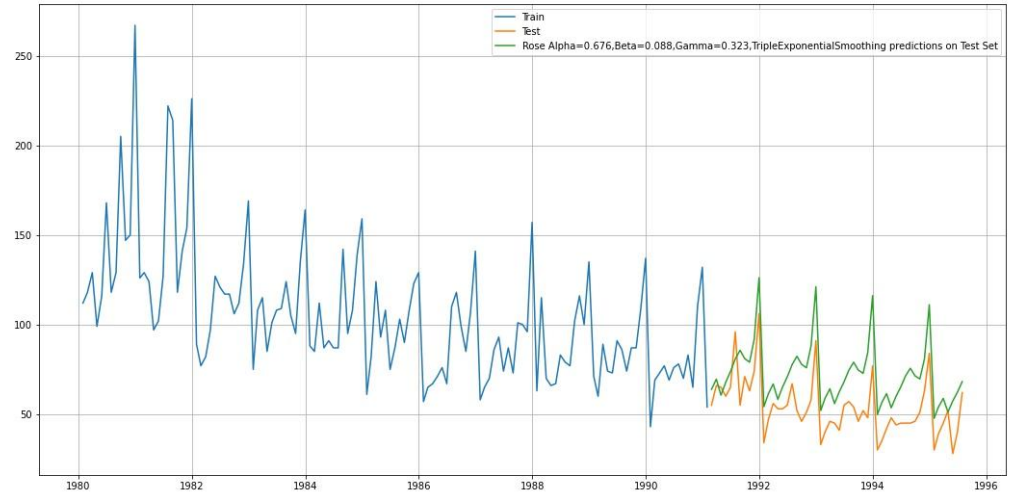
SIMPLE EXPONENTIAL SMOOTHING, DOUBLE EXPONENTIAL SMOOTHING,TRIPLE EXPONENTIAL SMOOTHING:

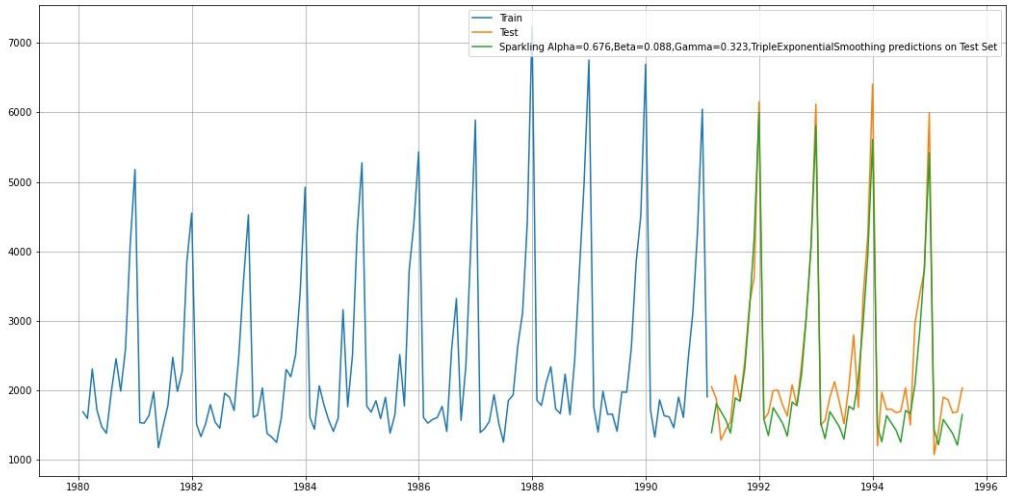

Alpha =0.995 Predictions
Train
Test
Alpha =0.995 Simple Exponential Smoothing predictions on Test Set
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
1996


Alpha =0.995 Predictions
Train
Test
Alpha =0.995 Simple Exponential Smoothing predictions on Test Set
7000
6000
5000
4000
3000
2000
1000
1980
1982
1984
1986
1988
1990
1992
1994
1996


Train
Test
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing predictions on Test Set
7000
6000
5000
4000
3000
2000
1000
1980
1982
1984
1986
1988
1990
1992
1994
1996


Train
Test
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing predictions on Test Set
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
1996





Simple Exponential Smoothing is applied if the time-series has neither a trend or seasonality, which is not the case with the given data .The forecasting using smoothing levels of alpha between 0 and 1 are as below, where the smoothing levels are passed manually . For alpha 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. On the second iteration, the model was run without passing a value for alpha and used parameters *'optimized=True, use_brute=True'* . The autofit model picked 0.0 as the smoothing parameter and return 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 smoothed out forecast with an initial value 2403.79 applied across the series.

Simple Exponential Smoothing is usually applied if the time-series has neither a trend or seasonality, which is not the case with the given data .The forecasting using smoothing

levels or alpha between 0 and 1 are as below, where the values were passed manually .For alpha value closer to 1, forecasts follows 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‘*optimized=True, use_brute=True’* .The autofit model picked 0.098 as the smoothing parameter and return consistent RMSE values in train and test datasets, which is consistent with alpha 0.1 in first iteration.

The Double Exponential Smoothing model 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 chose the optimized values using parameters ‘*optimized=True, use_brute=True’* . The autofit model returned 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.The Double Exponential Smoothing models is applicable when data has trend, but no .The autofit model returned 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 the final one is the one with alpha 0.1 and beta 0.1.
The Triple Exponential Smoothing model (Holt-Winters 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

The autofit model returned higher accuracy in train dataset, much higher than the values from iteration 1, but fared 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 0.2, beta 0.1 and gamma 0.2.The Triple Exponential Smoothing models (Holt-Winters Model) is applicable when data has both trend and seasonality. Rose data contain both trend and seasonality significantly

In the 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. The autofit model returned higher accuracy in the train dataset, much higher than the values from iteration 1, but fared 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 0.1

5.CHECKING STATIONARITY

Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determines the presence of a 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 below values, where p-value is

ADF on differenced series

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

- Differencing of order one is applied on the Sparkling series as above and tested for stationarity. At an order of differencing 1, the series is found to be stationary as above
- The rolling mean and standard deviation is also plotted to understand the component of seasonality and to ascertain if its multiplicative or additive in character
- The altitude of rolling mean and std dev is seen changing according to change in slope,

ADF on original series

• P-Value > alpha .05

• Test statistic > Critical values • Fail to reject the null hypothesis • The series is non-stationary which indicates multiplicity

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

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 below values, where p-value is

ADF on differenced series

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

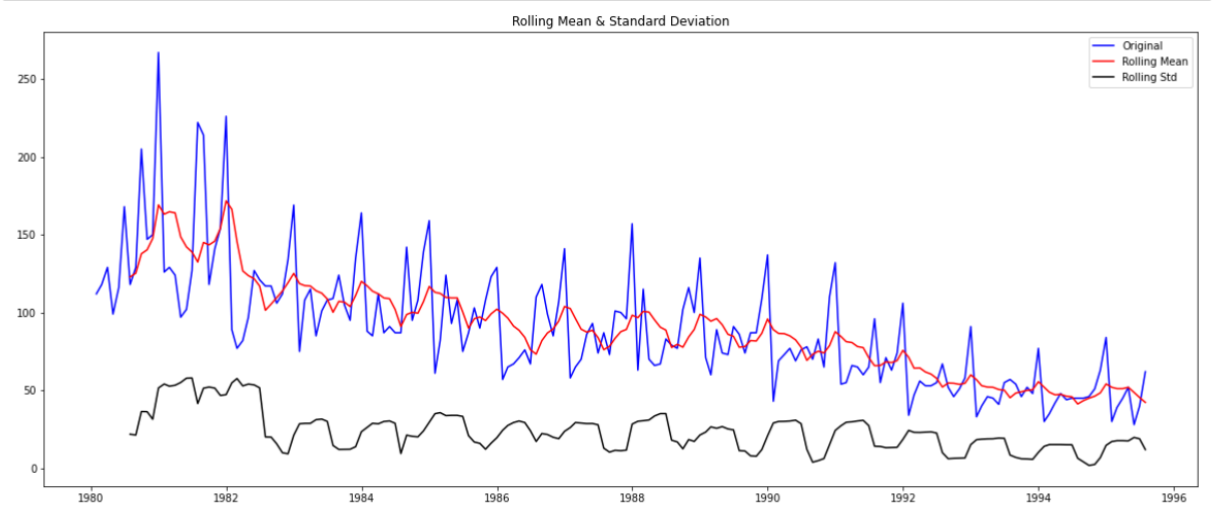
- Differencing of order one is applied on the Sparkling series as above and tested for stationarity
- At an order of differencing 1, the series is found to be stationary as above
- The rolling mean and standard deviation is also plotted to understand the component of

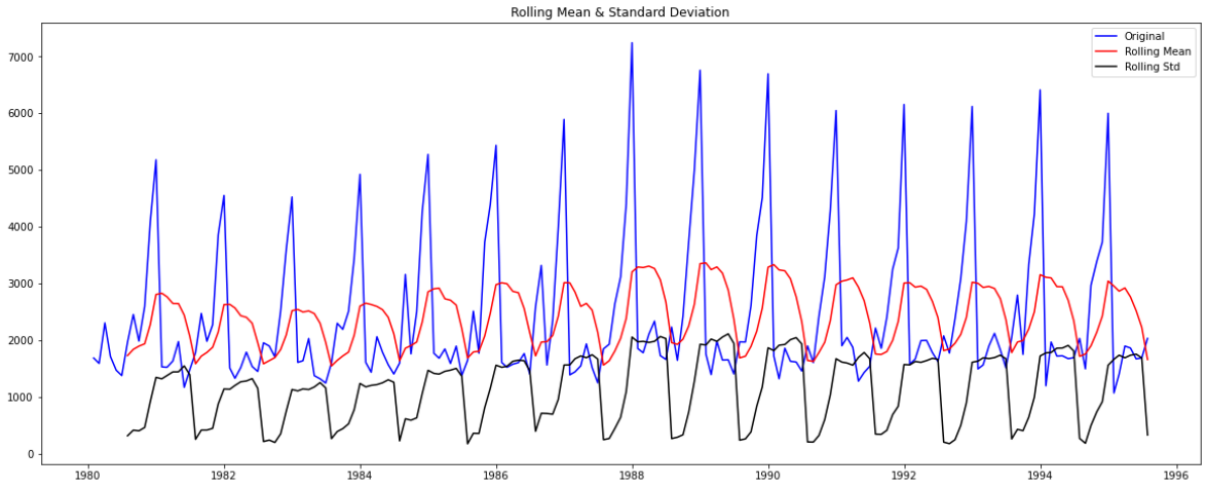
seasonality and to ascertain if its multiplicative or additive in character

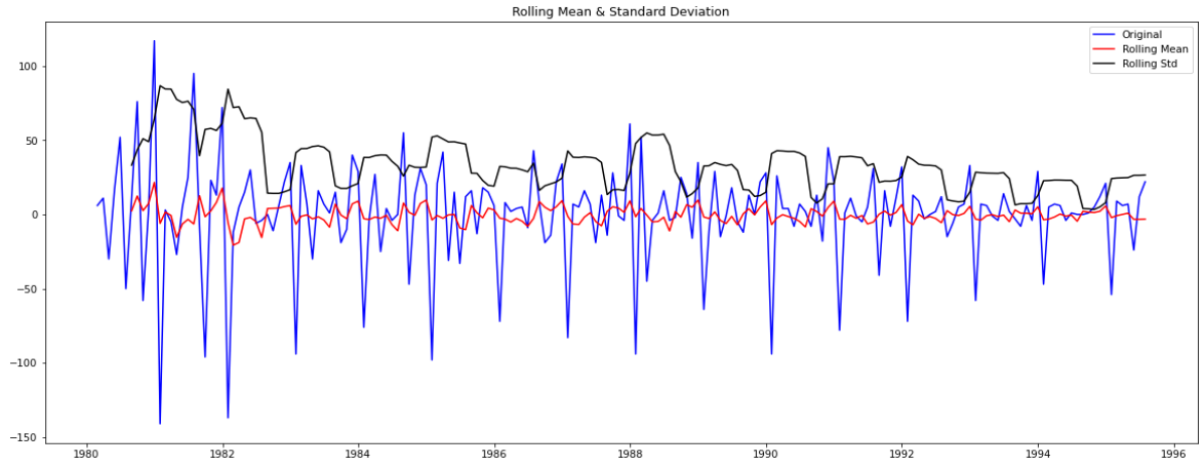
ADF on original series

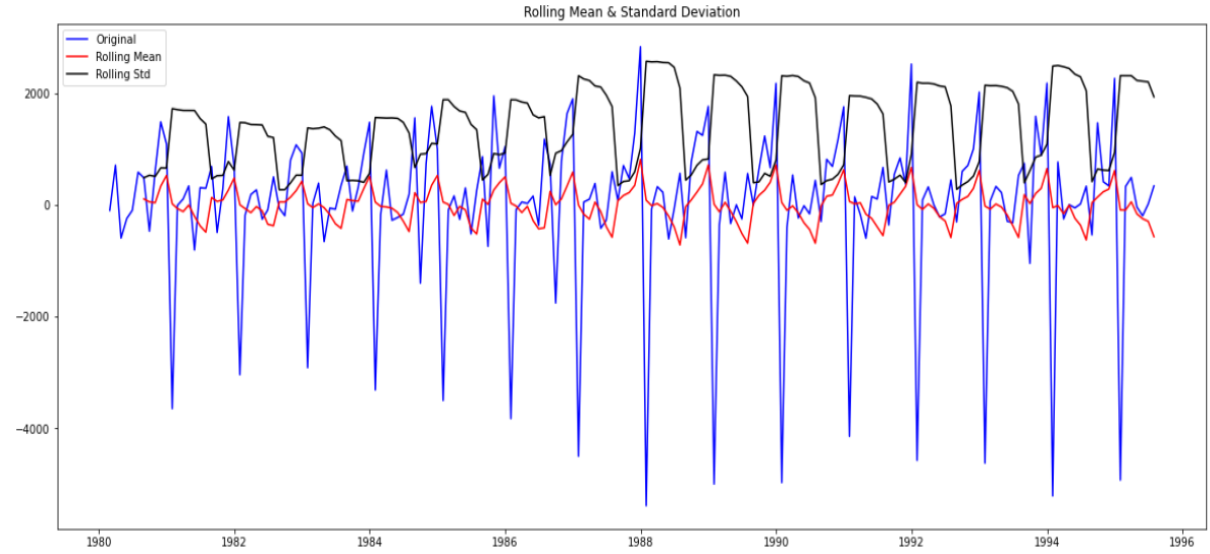
- P-Value > alpha .05
- Test statistic > Critical values • Fail to reject the null hypothesis
- The plot of rolling mean and standard deviation indicates that the seasonality is multiplicative as the altitude of plot varies with respect to trend

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 the model.










```
Results of Dickey-Fuller Test:
Test Statistic                  -1.874856
p-value                          0.343981
#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


<Figure size 1440x576 with 0 Axes>
```

```
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


<Figure size 1440x576 with 0 Axes>
```

```
Results of Dickey-Fuller Test:
Test Statistic                -8.044139e+00
p-value                        1.813580e-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


<Figure size 1440x576 with 0 Axes>
```

```
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


<Figure size 1440x576 with 0 Axes>
```

6.AUTO ARIMAX/SARIMAX MODEL

SPARKLING:

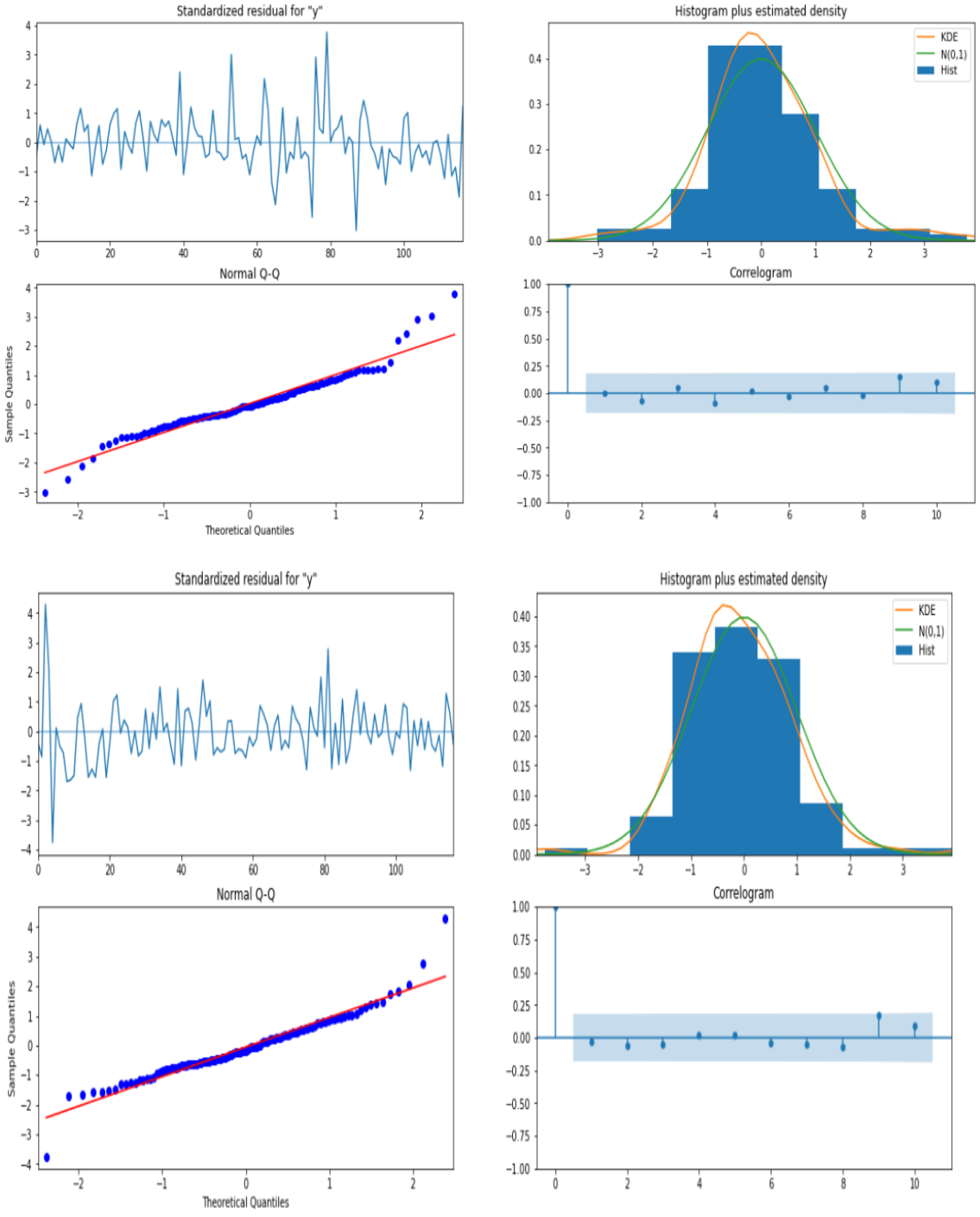
As the Sparkling series of data contain seasonality component we will be building SARIMA model, rather than ARIMA.Two iterations of automated SARIMA models were attempted in this exercise, one with original data and another with log transformation of the data, as an element of multiplicity in seasonality is suspected .The model built with original data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model , The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values .The top three models with lowest AIC values are as given. As per the AIC criteria, the optimum values for 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 .The RMSE values of the automated SARIMA models built are given here .The diagnostics plot of the selected model is given in the next slide.

ROSE:

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) .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 .The RMSE values

of the automated SARIMA models built are given here .The diagnostics plot of the selected model is given in the next slide.From the below model summary it can be inferred that seasonal AR(2) term has the highest weightage, followed by seasonal MA(2) .From the p-values it can be inferred that all the AR and MA terms are significant as the values are below .05.

7.



SARIMAX Results

Dep. Variable:	y	No. Observations:	133
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-516.783
Date:	Thu, 09 Sep 2021	AIC	1049.566
Time:	00:59:46	BIC	1071.663
Sample:	0	HQIC	1058.537
	- 133		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5956	0.148	-4.027	0.000	-0.885	-0.306
ma.L1	-0.1915	531.305	-0.000	1.000	-1041.531	1041.148
ma.L2	-0.8085	429.521	-0.002	0.998	-842.654	841.037
ar.S.L6	-0.0614	0.035	-1.733	0.083	-0.131	0.008
ar.S.L12	0.8480	0.039	21.984	0.000	0.772	0.924
ma.S.L6	0.2248	280.384	0.001	0.999	-549.318	549.768
ma.S.L12	-0.7751	217.384	-0.004	0.997	-426.840	425.290
sigma2	333.0217	1.92e+05	0.002	0.999	-3.76e+05	3.76e+05

Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	59.38
Prob(Q):	0.76	Prob(JB):	0.00
Heteroskedasticity (H):	0.47	Skew:	0.53
Prob(H) (two-sided):	0.02	Kurtosis:	6.33

8.

	Rose Test RMSE	Sparkling Test RMSE
Rose RegressionOnTime	51.554113	NaN
Sparkling RegressionOnTime	NaN	1286.310050
NaiveModel	15.915867	NaN
NaiveModel	NaN	1381.177135
SimpleAverageModel	2346.228164	NaN
SimpleAverageModel	NaN	1285.039964
Alpha=0.995,SimpleExponentialSmoothing	33.970592	NaN
Alpha=0.995,SimpleExponentialSmoothing	NaN	2656.047647
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	20.504660	NaN
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	NaN	382.857943
Alpha=0.1,Beta=0.1,Gamma=0.1,TripleExponentialSmoothing	10.100519	NaN
Alpha=0.2.0,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	NaN	309.038218
SARIMA(1,1,2)(2,0,2,6)	25.811718	NaN
SARIMA(0,1,2)(2,0,2,6)	NaN	494.216771

Taking consideration of the rmse test we can conclude that for both rose and sparkling data the triple exponential model is working the best.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.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.
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.The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data

Two point trailing moving average is found to be having the best fitment against the test data, though with a lag of 2 and falling short at times . Both SARIMA and TES forecasts are a bit higher than the actuals at any given point in time.

9.

FUTURE PREDICTION:

sparkling:

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

• TES model *alpha: 0.2, 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

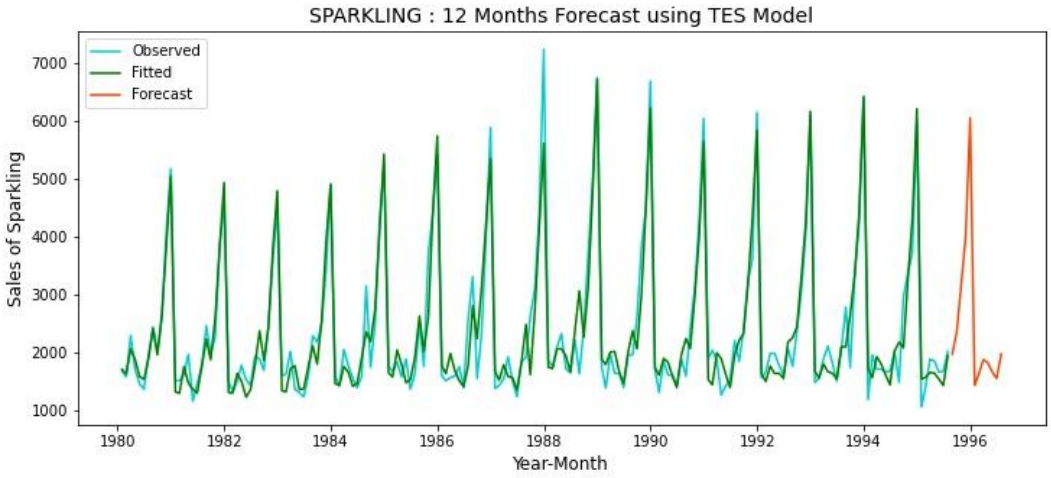
• The 12 month prediction of the TES model is as below

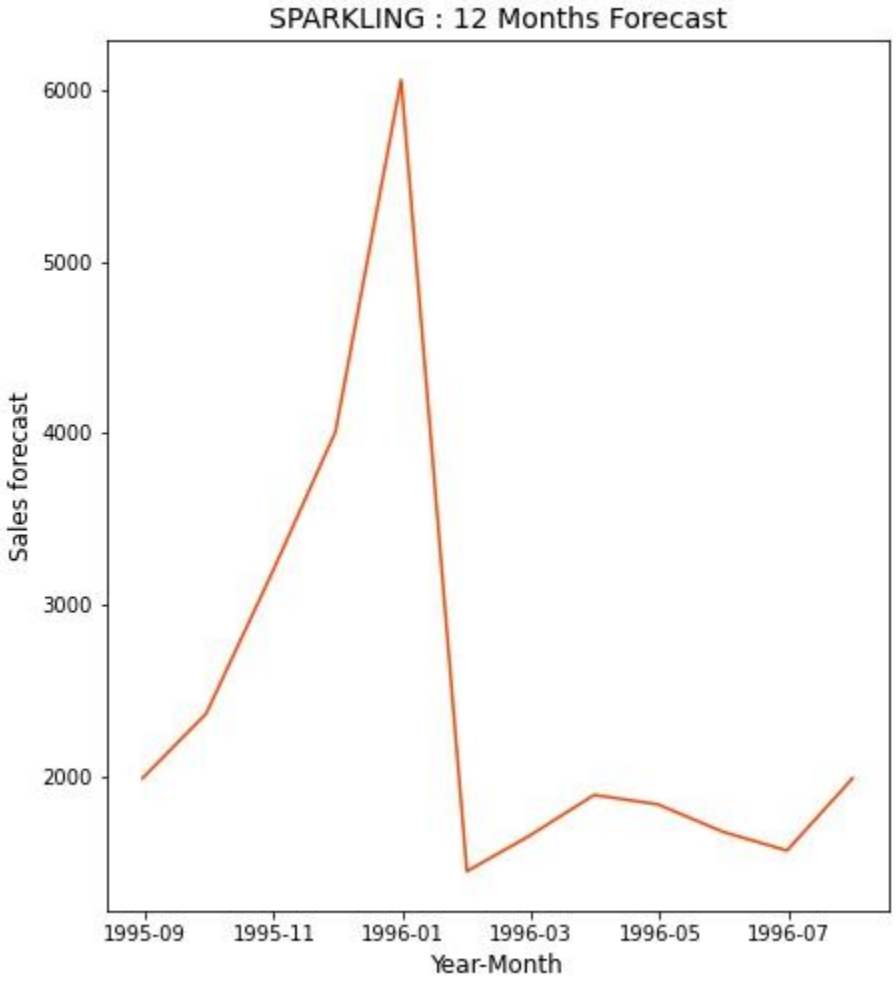
• The SARIMA model is built with parameters (0, 1, 2)x(2, 0, 2, 6), is found to be the most optimal SARIMA model

• The 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

• The SARIMA model is seen to have better fitment with the most recent observed data and shows high variations in the farthest periods of observations, which explains the high RMSE value.

• The RMSE values of the two models are as below





Model for 12 month prediction of the sparkling data.

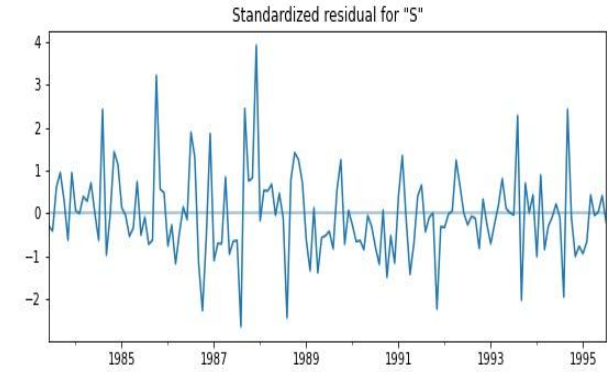
SARIMAX Results

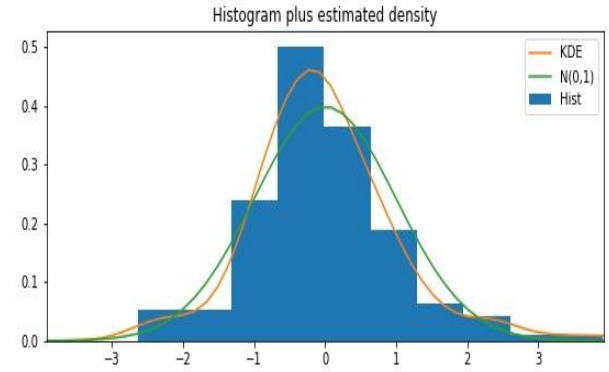
Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(3, 1, 3)x(1, 1, [1, 2], 12)	Log Likelihood	-1078.437
Date:	Thu, 09 Sep 2021	AIC	2176.875
Time:	08:48:42	BIC	2206.711
Sample:	01-31-1980 - 07-31-1995	HQIC	2188.998
Covariance Type:	opg		

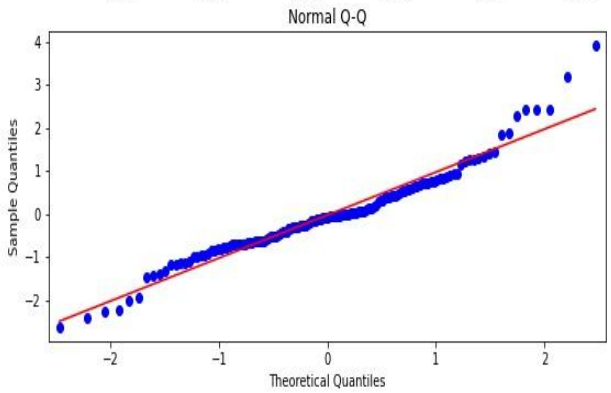
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4230	0.086	-4.914	0.000	-0.592	-0.254
ar.L2	-0.9094	0.053	-17.290	0.000	-1.013	-0.806
ar.L3	0.1424	0.087	1.637	0.102	-0.028	0.313
ma.L1	-0.4113	0.078	-5.270	0.000	-0.564	-0.258
ma.L2	0.4622	0.083	5.574	0.000	0.300	0.625
ma.L3	-0.9673	0.104	-9.307	0.000	-1.171	-0.764
ar.S.L12	-0.0698	0.710	-0.098	0.922	-1.461	1.322
ma.S.L12	-0.4551	0.722	-0.630	0.528	-1.870	0.960
ma.S.L24	-0.0808	0.397	-0.204	0.839	-0.859	0.697
sigma2	1.461e+05	1.06e-06	1.37e+11	0.000	1.46e+05	1.46e+05

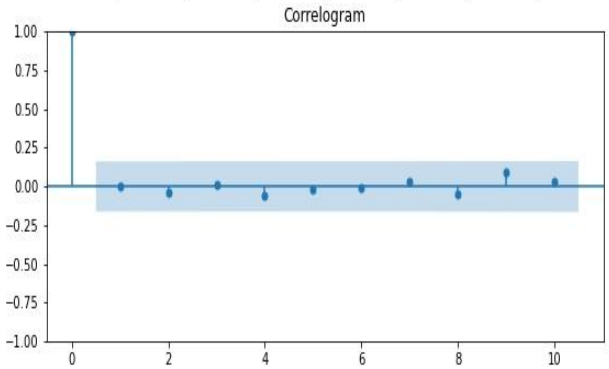
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	35.58
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	0.72	Skew:	0.66
Prob(H) (two-sided):	0.26	Kurtosis:	5.03

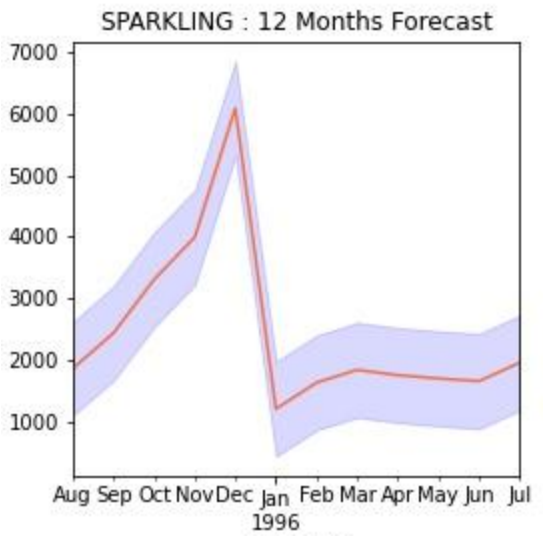
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 4.16e+27. Standard errors may be unstable

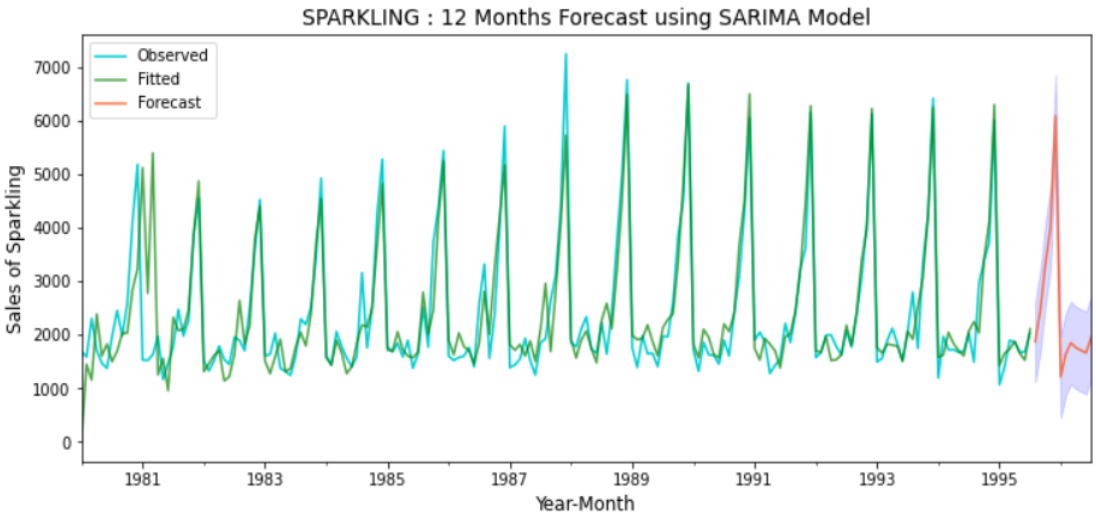












	Sparkling
1995-08-31	1873.39
1995-09-30	2444.94
1995-10-31	3312.60
1995-11-30	3994.61
1995-12-31	6084.08
1996-01-31	1215.91
1996-02-29	1640.57
1996-03-31	1847.28
1996-04-30	1761.99
1996-05-31	1708.30
1996-06-30	1664.02
1996-07-31	1961.28

12 month prediction of the sparkling data.

FINAL OPTIMIZED MODEL :

• The SARIMA model built on the complete Sparkling timeseries is chosen, as prediction provide confidence interval which give better explain ability 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 have 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 get values higher than alpha 0.05, which fails to reject the null hypothesis that these terms are not significant

Model Diagnostics – SARIMA (0,1,2)x(2,0,2,6)

rose data:

• 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: 0.1, beta: 0.1 and gamma: 0.1 & trend: 'additive', seasonal: 'multiplicative'* is found to be the best model in terms of accuracy scored against the full data

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

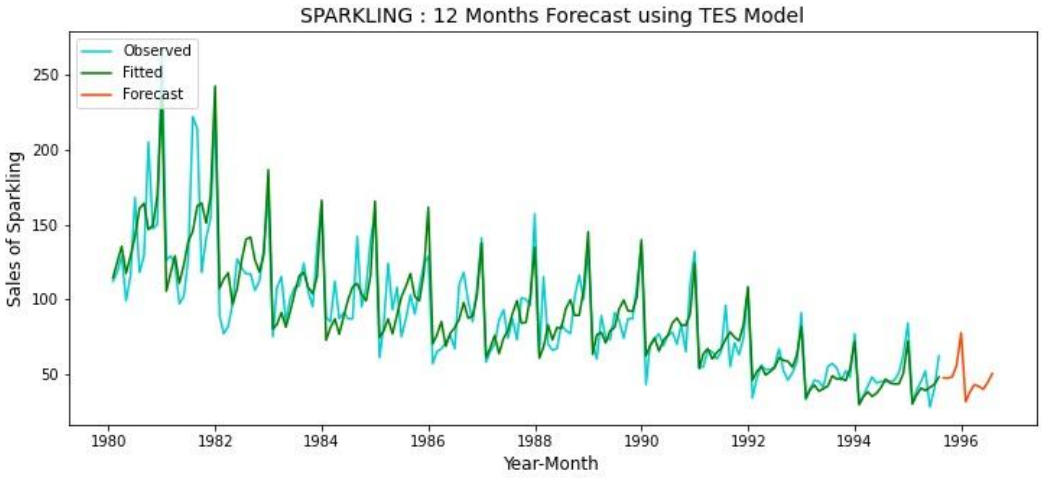
• The 12 month prediction of the TES model is as below

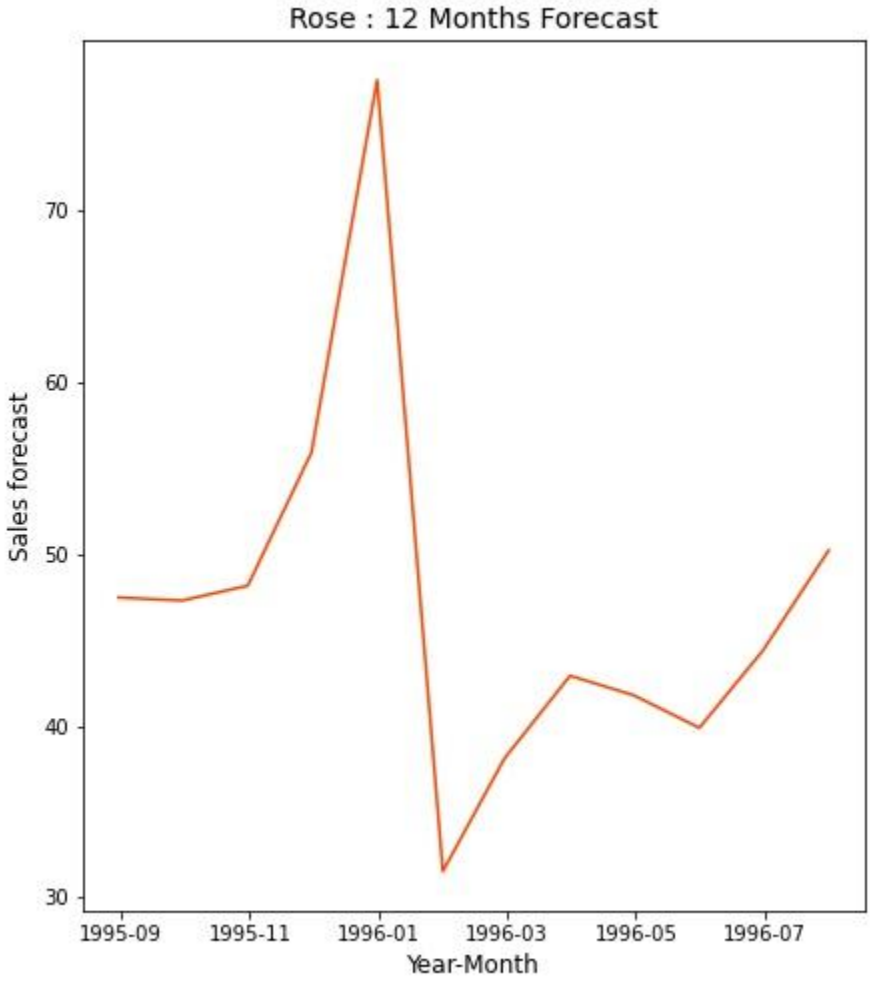
• The SARIMA model is built with parameters (1, 1, 2)x(2, 0, 2, 6), is found to be the most optimal SARIMA model for the complete time-series

• 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 values

• The RMSE and MAPE values of the two models are as below


SPARKLING : 12 Months Forecast using TES Model
Observed
Fitted
Forecast
Sales of Sparkling
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
1996
Year-Month


Rose : 12 Months Forecast
Sales forecast
70
60
50
40
30
1995-09
1995-11
1996-01
1996-03
1996-05
1996-07
Year-Month

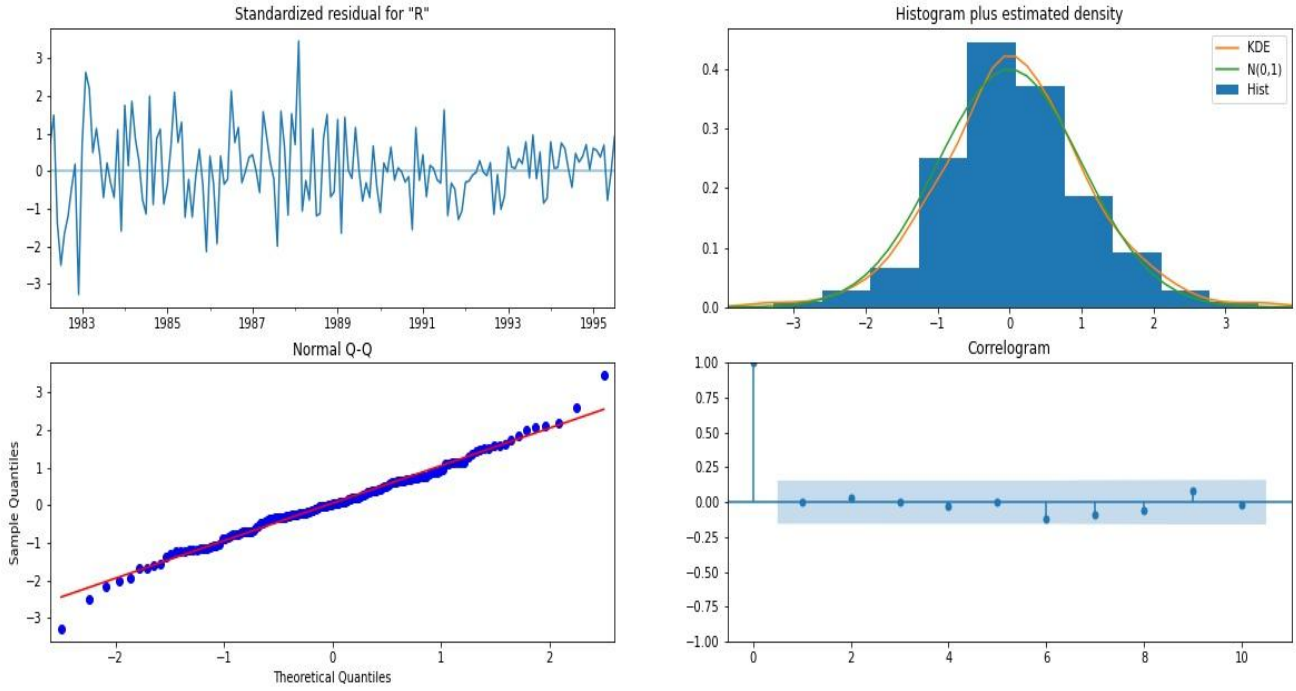
SARIMAX Results

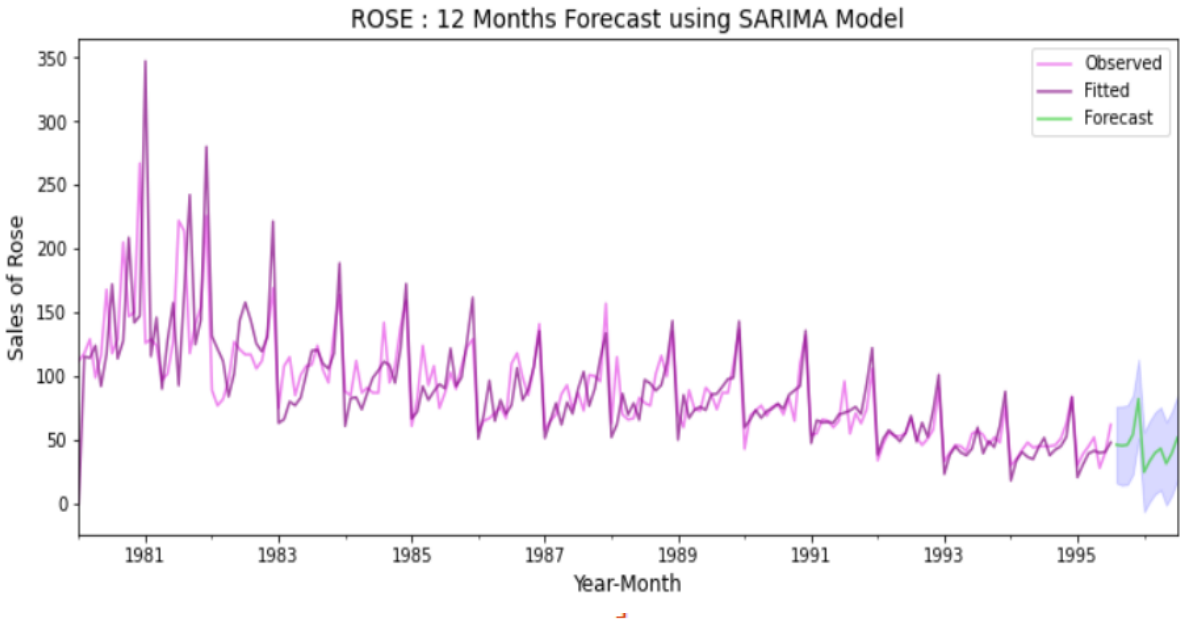
Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(4, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-664.149
Date:	Thu, 09 Sep 2021	AIC	1342.298
Time:	08:48:32	BIC	1363.825
Sample:	01-31-1980	HQIC	1351.039
	- 07-31-1995		
Covariance Type:	opg		

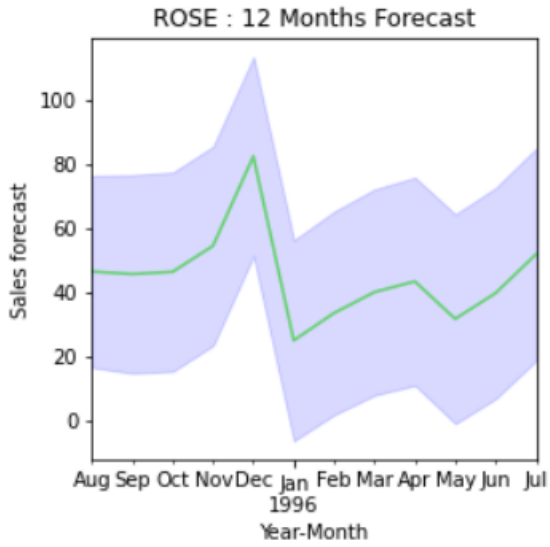
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0916	0.084	1.096	0.273	-0.072	0.255
ar.L2	-0.1078	0.077	-1.393	0.163	-0.259	0.044
ar.L3	-0.1315	0.076	-1.729	0.084	-0.281	0.018
ar.L4	-0.1072	0.078	-1.375	0.169	-0.260	0.046
ma.L1	-0.8271	0.055	-14.904	0.000	-0.936	-0.718
ma.S.L12	-0.5966	0.059	-10.124	0.000	-0.712	-0.481
sigma2	232.4613	24.368	9.540	0.000	184.702	280.221

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	5.29
Prob(Q):	0.93	Prob(JB):	0.07
Heteroskedasticity (H):	0.22	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.89

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







	ROSE
1995-08-31	46.27
1995-09-30	45.49
1995-10-31	46.19
1995-11-30	54.30
1995-12-31	82.21
1996-01-31	24.80
1996-02-29	33.35
1996-03-31	39.86
1996-04-30	43.21
1996-05-31	31.52
1996-06-30	39.55
1996-07-31	51.61

12 Month prediction values.

FINAL MODEL:

• 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 **Model**

Diagnostics – SARIMA (1,1,2)x(2,0,2,6)

10. RECOMMENDATION

	Sparkling
count	12.000000
mean	2459.080833
std	1384.638657
min	1215.910000
25%	1697.230000
50%	1860.335000
75%	2661.855000
max	6084.080000

	ROSE
count	12.000000
mean	44.863333
std	14.468179
min	24.800000
25%	38.000000
50%	44.350000
75%	47.605000
max	82.210000

Sparkling

- The model forecasts sales of **29510** units of Sparkling wine in 12 months into the 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), of which a total of **13,392 units** of sparkling wine is expected to be sold.
- The forecast also indicates that the year-on-year sale of sparkling wine is not showing an upward trend. The winery must adopt innovative marketing skills to improve the sale compared to previous years
- Adding more exogenous variable into the timeseries data can improve forecasts

Rose

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