Time Series Forecasting
Business Report

Great learning

BR Rahul

Problem:

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

Please do perform the following questions on each of these two data sets separately.

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 all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models 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.

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 .

- Both the given datasets of the respective type of wines is combined to a single data frame, for the sake of comparability of the timeseries components and forecast .

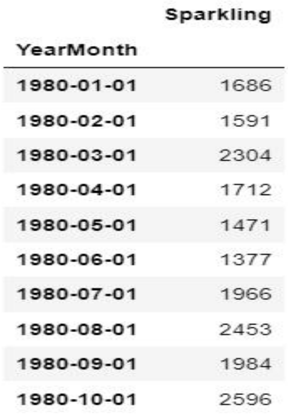
- The Rose time-series got values missing for two months in 1994, which are imputed using interpolation (linear method) .

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

- Both the datasets shows 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 .

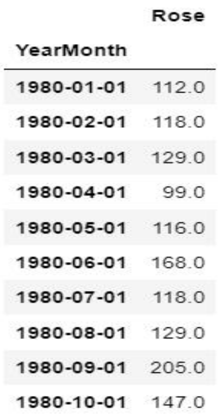
- While Sparkling wine has been consistently favoured over the years by customers, the demand for Rose had been fell out-of-favour over the years .

After reading data of Sparkling wine data



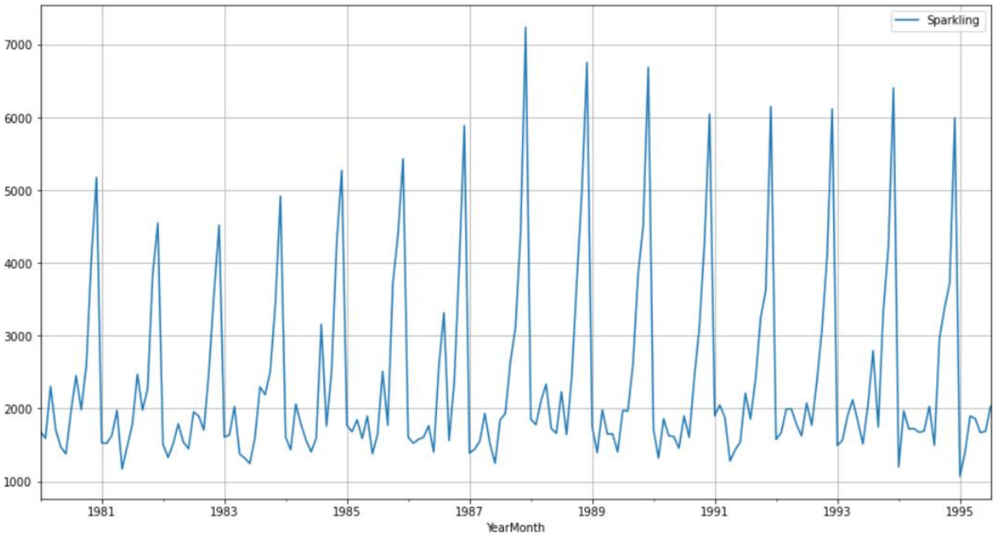
	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471
1980-06-01	1377
1980-07-01	1966
1980-08-01	2453
1980-09-01	1984
1980-10-01	2596

After reading data of Rose wine data

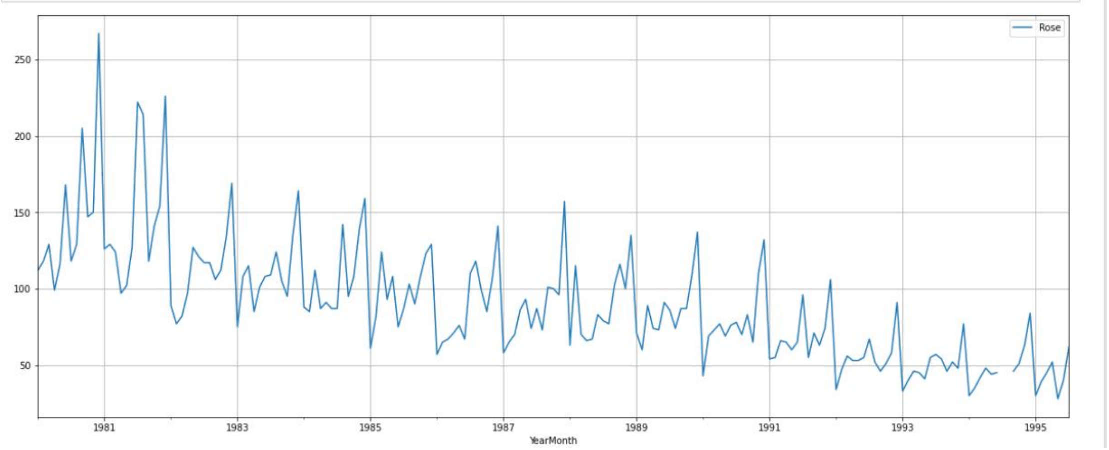


	Rose
YearMonth	
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0
1980-06-01	168.0
1980-07-01	118.0
1980-08-01	129.0
1980-09-01	205.0
1980-10-01	147.0

Sale of Sparkling Wine

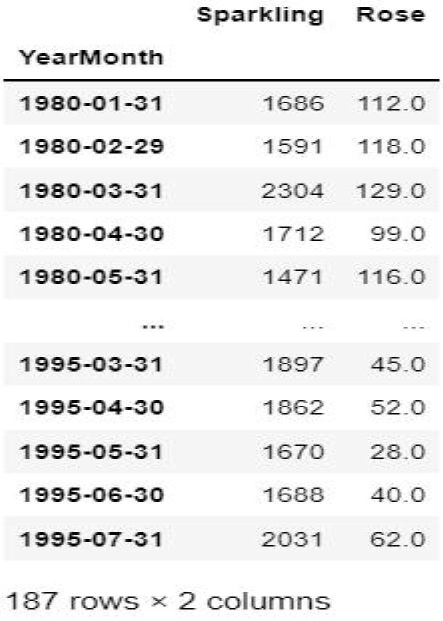


Sale of Rose Wine



Y-axis : Units sold
X-axis : YearMonth

After Creating combined DataFrame :



	Sparkling	Rose
YearMonth		
1980-01-31	1686	112.0
1980-02-29	1591	118.0
1980-03-31	2304	129.0
1980-04-30	1712	99.0
1980-05-31	1471	116.0
...	...	...
1995-03-31	1897	45.0
1995-04-30	1862	52.0
1995-05-31	1670	28.0
1995-06-30	1688	40.0
1995-07-31	2031	62.0

187 rows × 2 columns

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

Exploratory Data Analysis:

Sparkling

The descriptive summary 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.

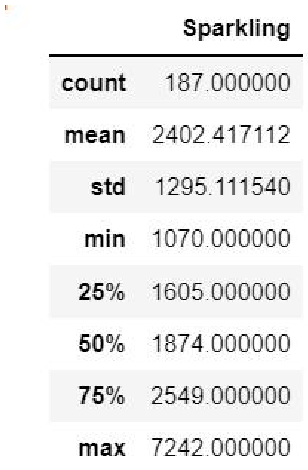
• The Empirical CDF plot shows that, in 80% of months, at least 3000 units of Sparkling wine were sold .

- 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.

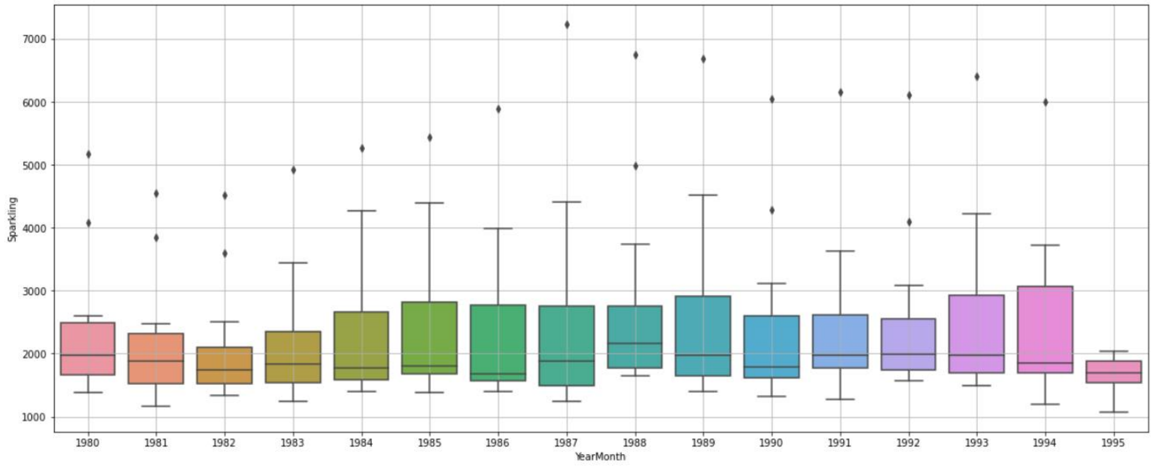
- The monthly-box-plot shows a clear seasonality during the festive seasonal months of October, November and December, which peaks in December. The sale tanks in the month of June.

Descriptive Statistics for Sparkling wine:

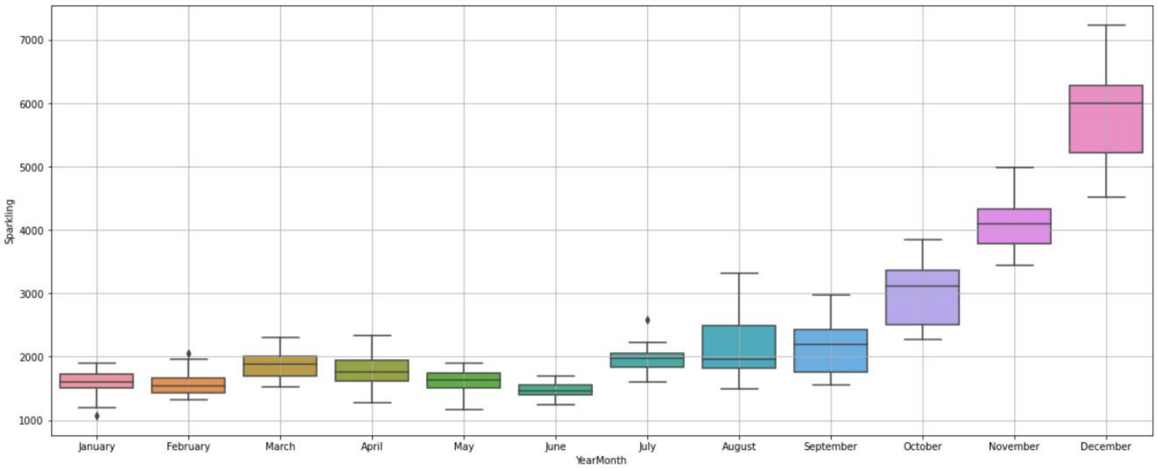


	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

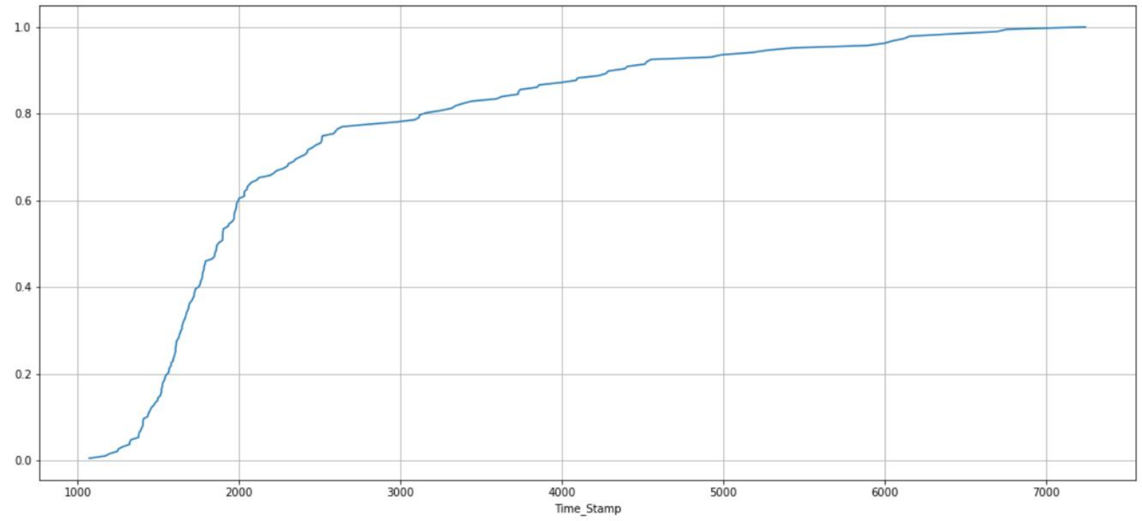
Yearly Boxplot - Sparkling



Monthly Boxplot - Sparkling



Distribution of Sparkling :

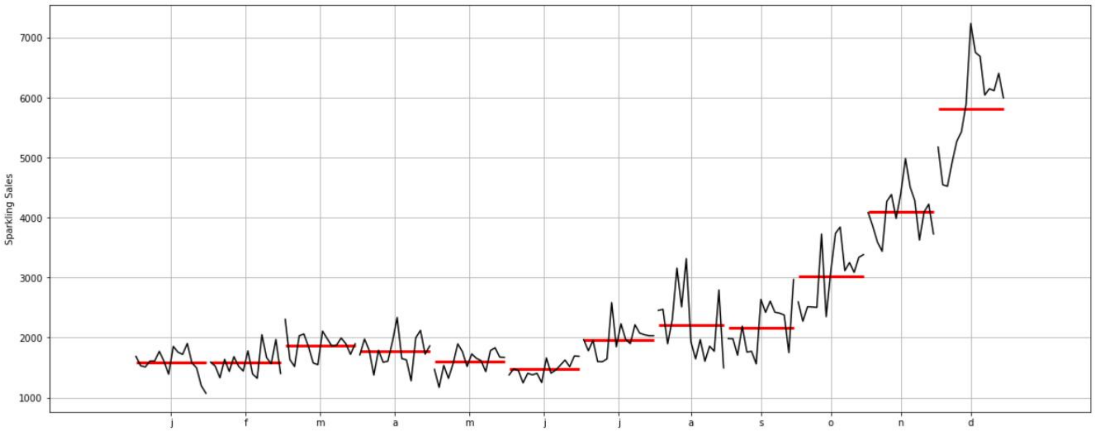


The monthly plot for Sparkling shows mean and variation of units sold each month over the years. Sale in seasonal months shows a higher variation than in the lean months.

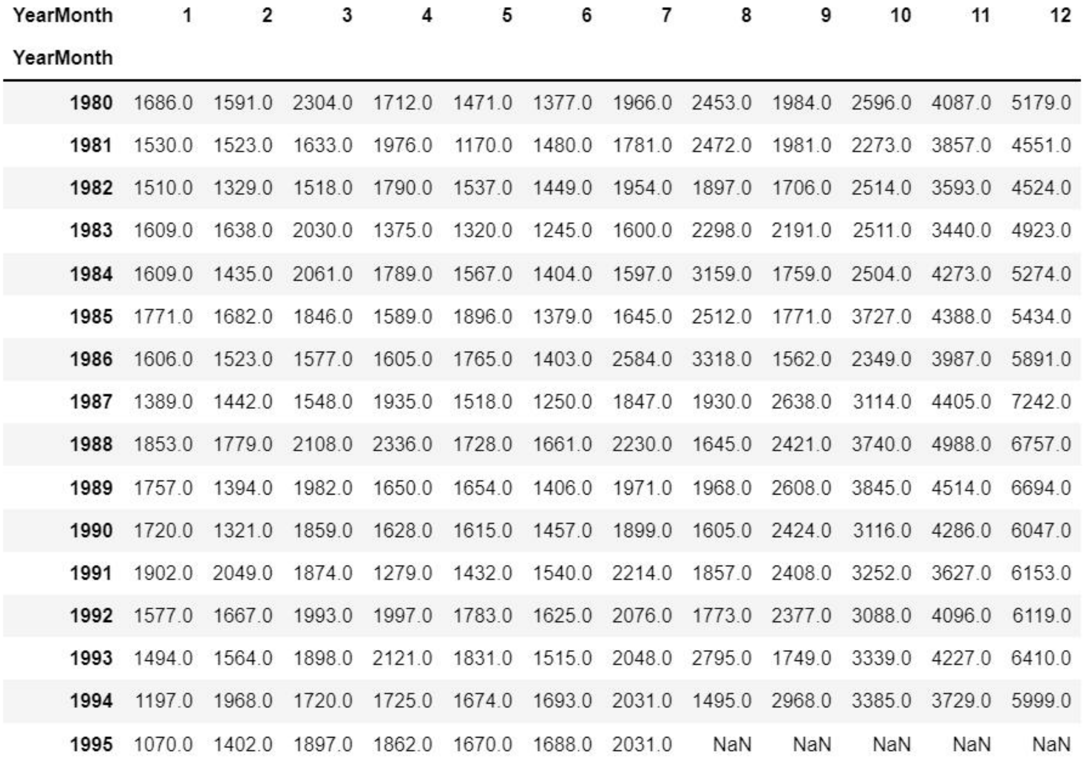
• Sale in December with a mean few points below 6000, varies from 7400 to 4500 units over the years. whereas sale in November varies from 3500 units to 5000 units and sale in October varies from 2500 to 4000 units .

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

Sparkling Monthly plot



The Figure below Shows the Sparkling - Monthly sales over the years



YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

The Dataframe 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.

Rose :

The descriptive summary 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 .

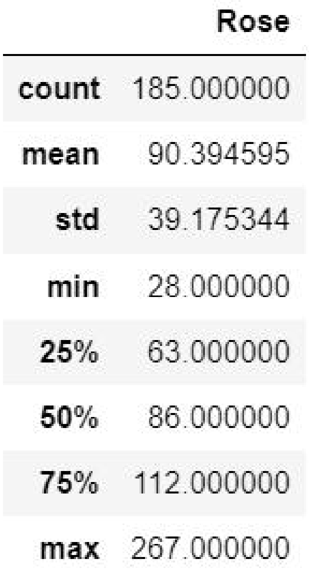
- The Empirical CDF plot shows that, in 80% of months, at least 120 units of Rose wine were sold .

- 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 upperbound 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.

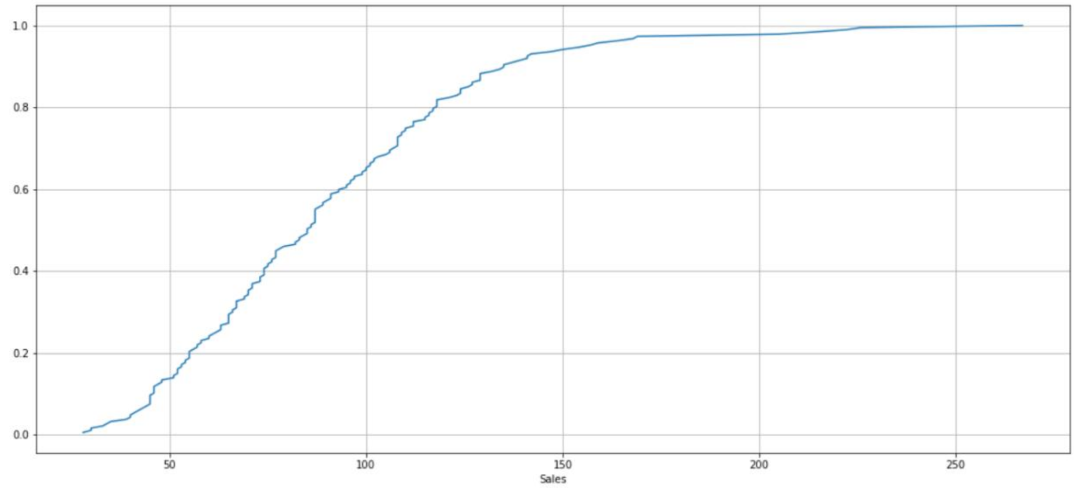
- Average sale in December is around 140 units, November is around 110 units and October is around 90 units .

Descriptive Statistics for Rose Wine :

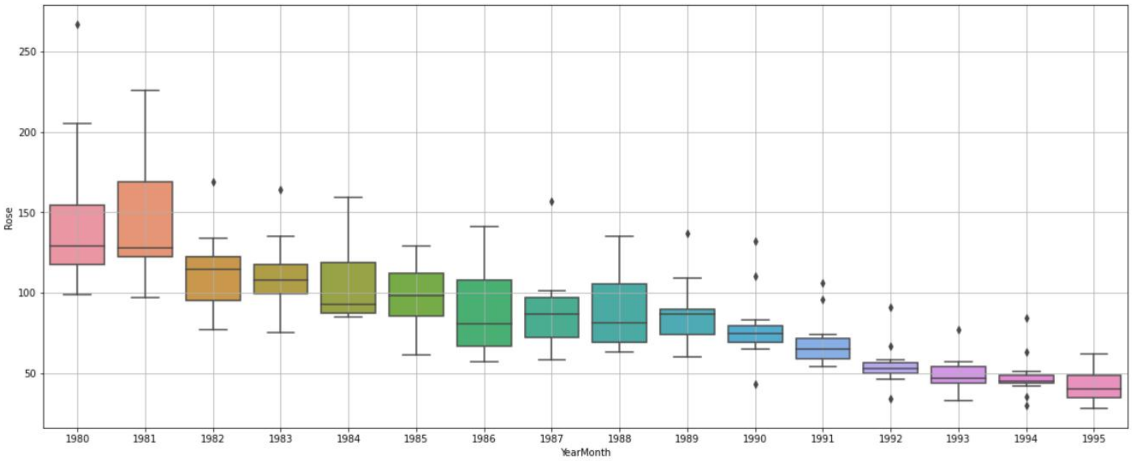


	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

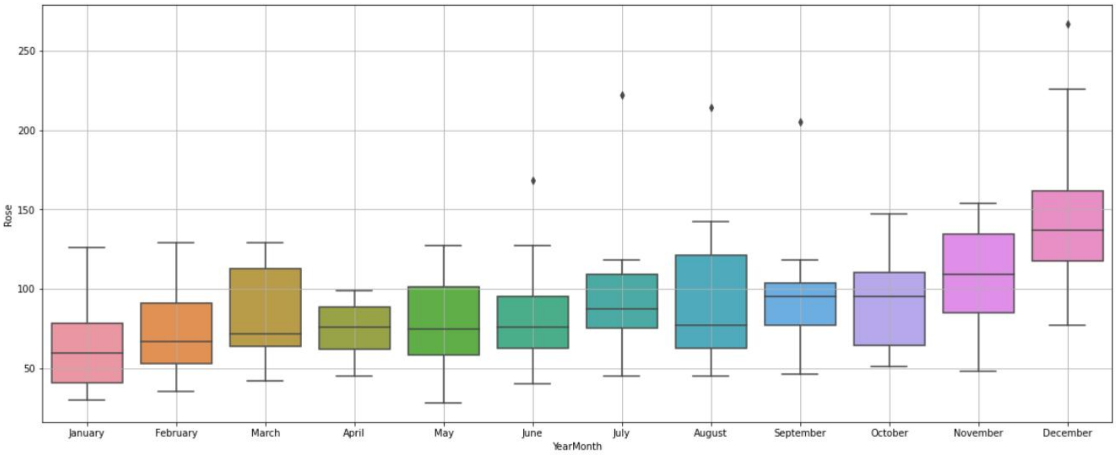
Distribution of Rose wine



Yearly Boxplot - Rose



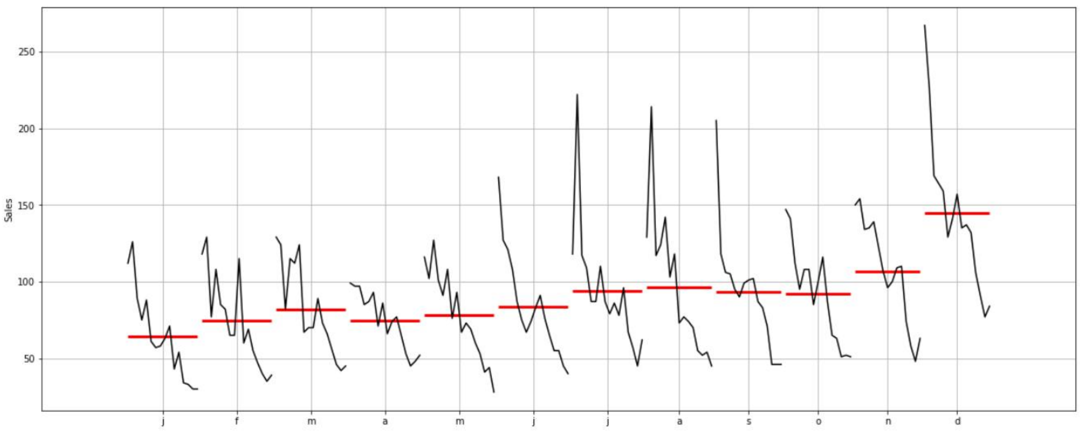
Monthly Boxplot - Rose



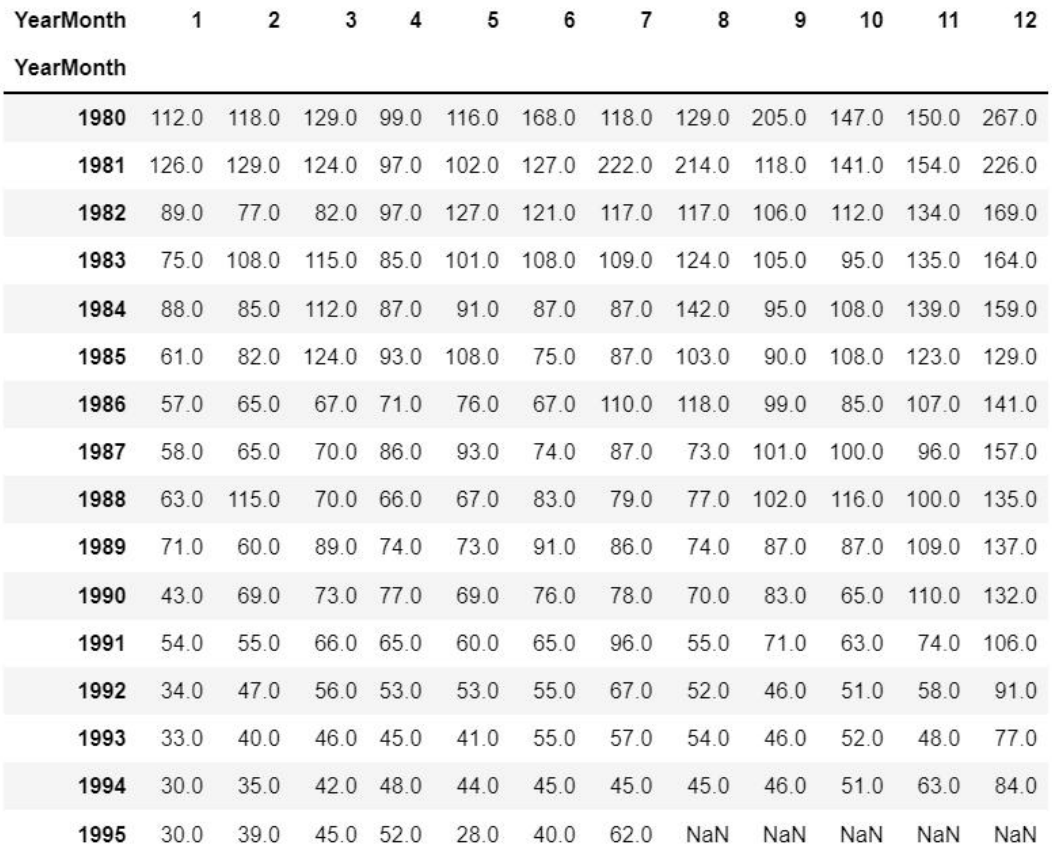
The monthly plot for Rose shows mean and variation of units sold each month over the years. Sale in months such as July, August, September and December shows a higher variation than the rest .

• Sale in December with a mean few points below 100, varies from 75 to 270 units over the years. Whereas the average sale is less than or closer to 100 units (above 50) for the rest of the year.

Rose - Monthly Plot



Rose - Monthly over the years



YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.0	45.0	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

The DataFrame 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 were sold in December,1980 and the least of December sale was in 1993. Though December sale picked after 1983, it consistently dipped after 1987.

Time Series Decomposition:

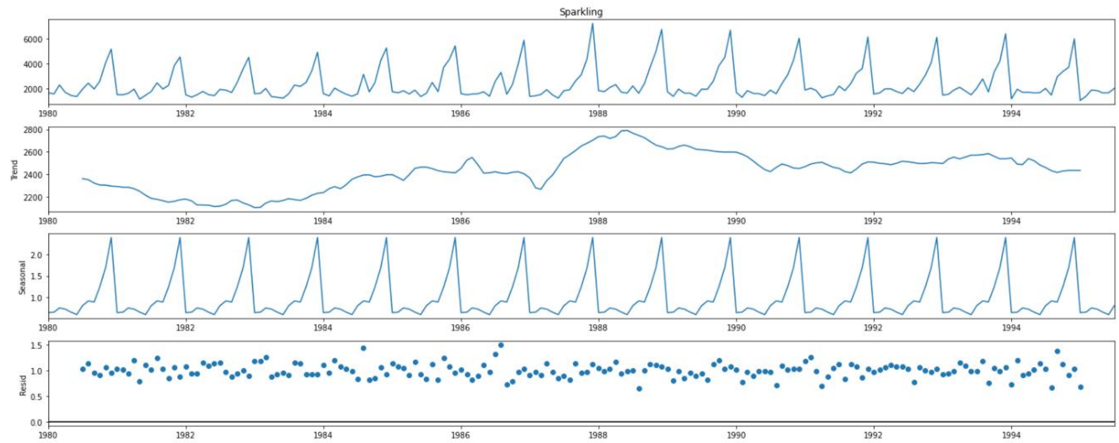
Sparkling :
The decomposition plots of Sparkling wine sales is given here .

- 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 of time-series .

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

Multiplicative Model :

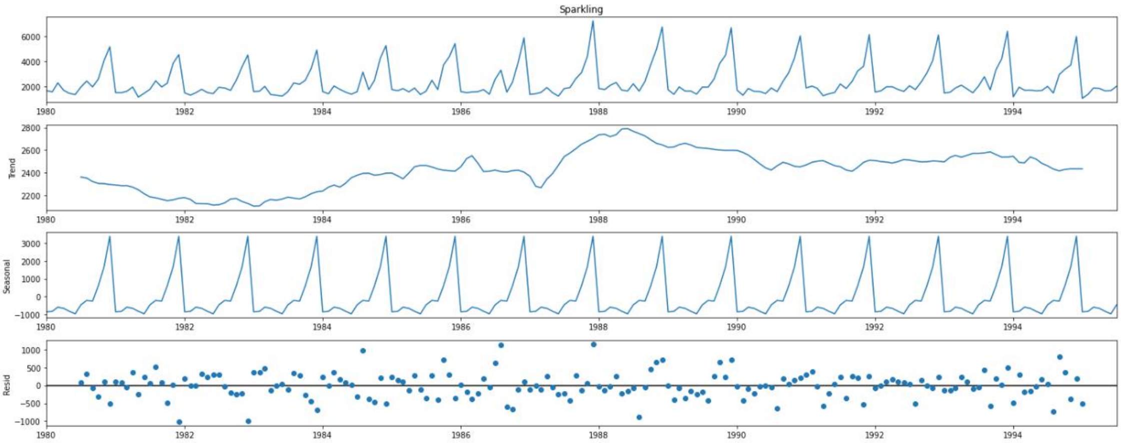


The residual shows a pattern of high variability across the period of time-series, which is more or less consistent in both additive and multiplicative decompositions .

- 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.

Additive Model :



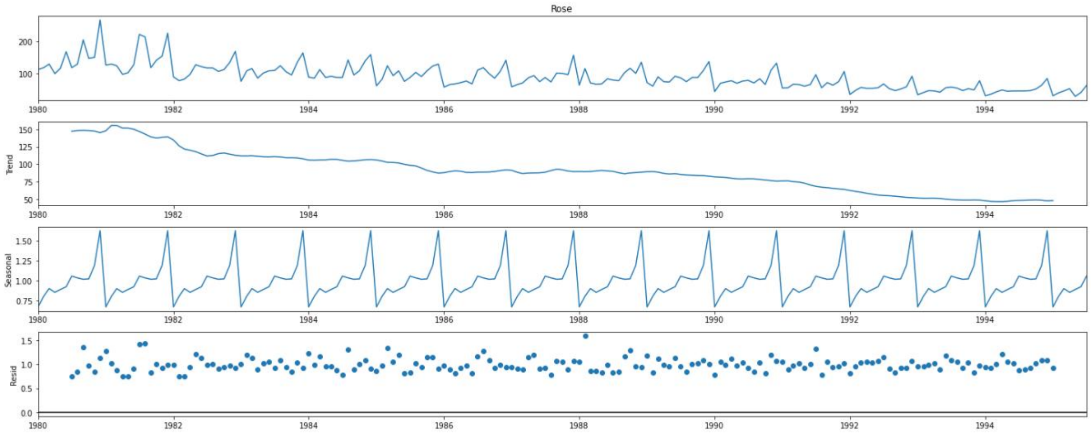
Rose :

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 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% .

Multiplicative Model :



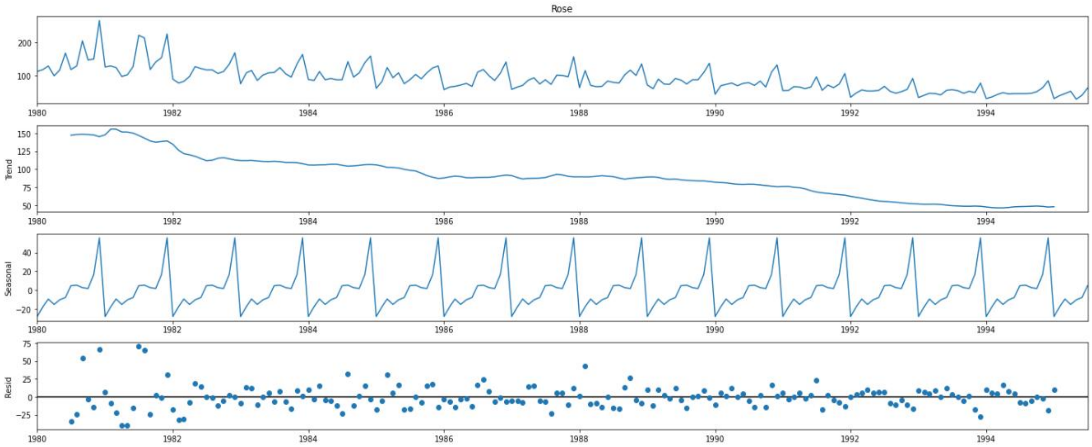
The residuals shows a pattern of high variability across the period of time-series, which is more or less consistent in both additive and multiplicative decompositions .

• 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.

Additive Model :

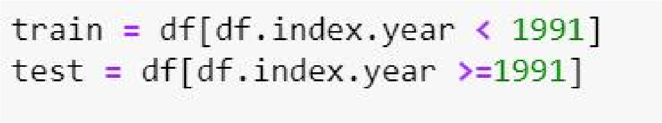


Additive model are the one where the variance of data doesn’t change over the different values of the time series. The systematic component is the arithmetic sum of the individual effects of the predictors. Additive model is linear and the trend line here is a straight line and the seasonality has the same frequency and amplitude.

In Multiplicative Model as the data increases, so does the seasonal pattern or the variance increases. Here the trend and seasonal components are multiplied and then added to the error component. Multiplicative model is non linear , such as such quadratic or exponential and the trend is a curved line and seasonality has an increasing or decreasing frequency and amplitude over time.

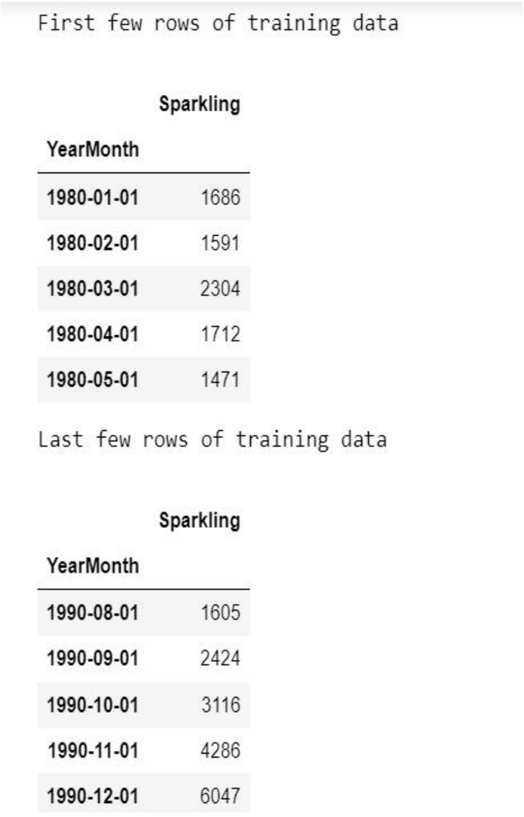
3 . Split the data into training and test. The test data should start in 1991.

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



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

First and the last few rows of Training Data:



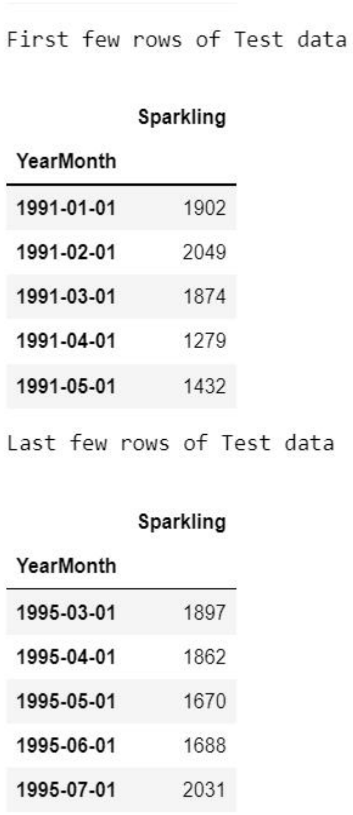
```
First few rows of training data
```

	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

```
Last few rows of training data
```

	Sparkling
YearMonth	
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

First and last few rows of Test Data :



```
First few rows of Test data
```

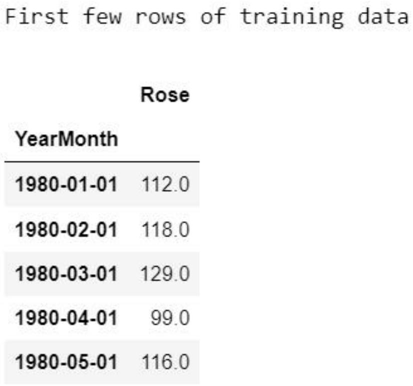
	Sparkling
YearMonth	
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

```
Last few rows of Test data
```

	Sparkling
YearMonth	
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

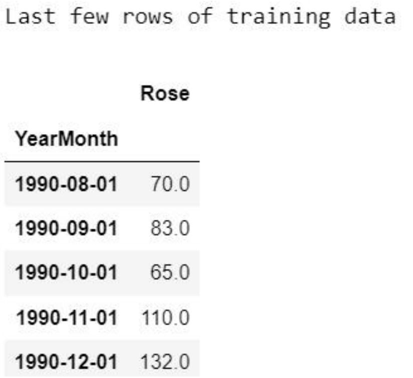
Rose :

First and the last few rows of Training Data:



```
First few rows of training data
```

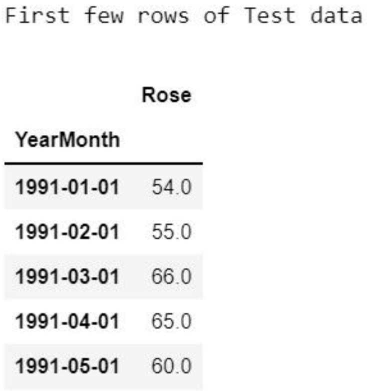
YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0



```
Last few rows of training data
```

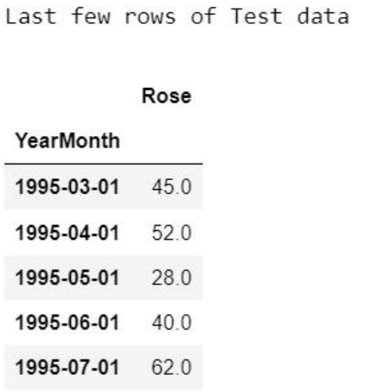
YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

First and the last few rows of Test Data:



```
First few rows of Test data
```

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0



```
Last few rows of Test data
```

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

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

Sparkling :

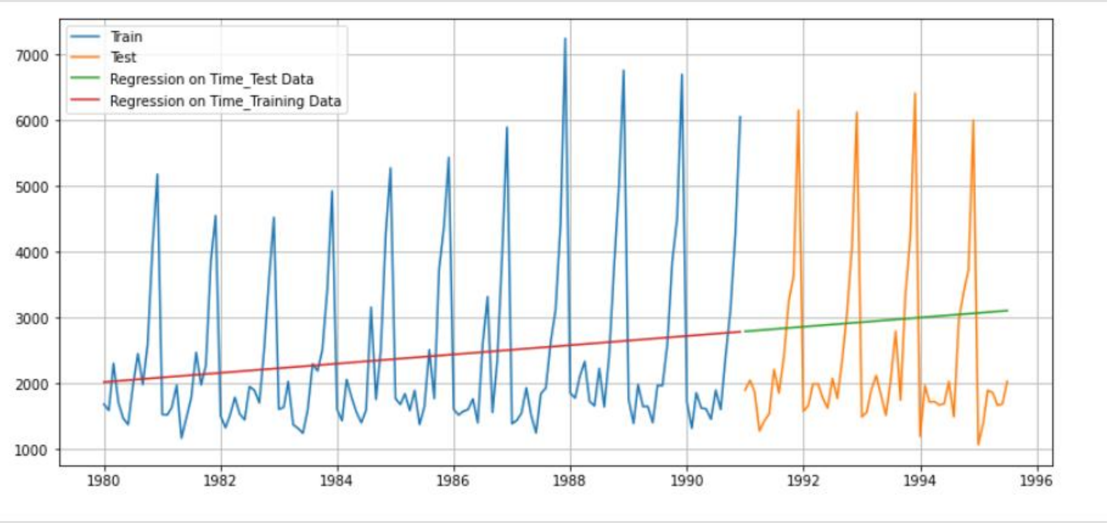
Linear Regression :
To regress the sale of Sparkling and Rose wines, numerical time instance order for both training and test set were generated and the values added to the respective datasets .

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

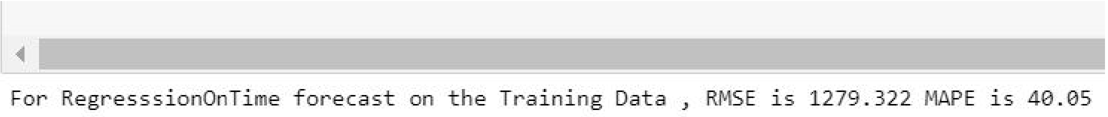
The linear regression on the Rose dataset shows an apparent downward trend as consistent with the observed time-series

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

Linear regression model for Sparkling Dataset

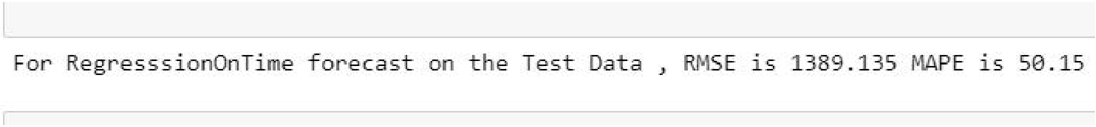


RMSE and MAPE for training Data(Sparkling)



```
For RegresssionOnTime forecast on the Training Data , RMSE is 1279.322 MAPE is 40.05
```

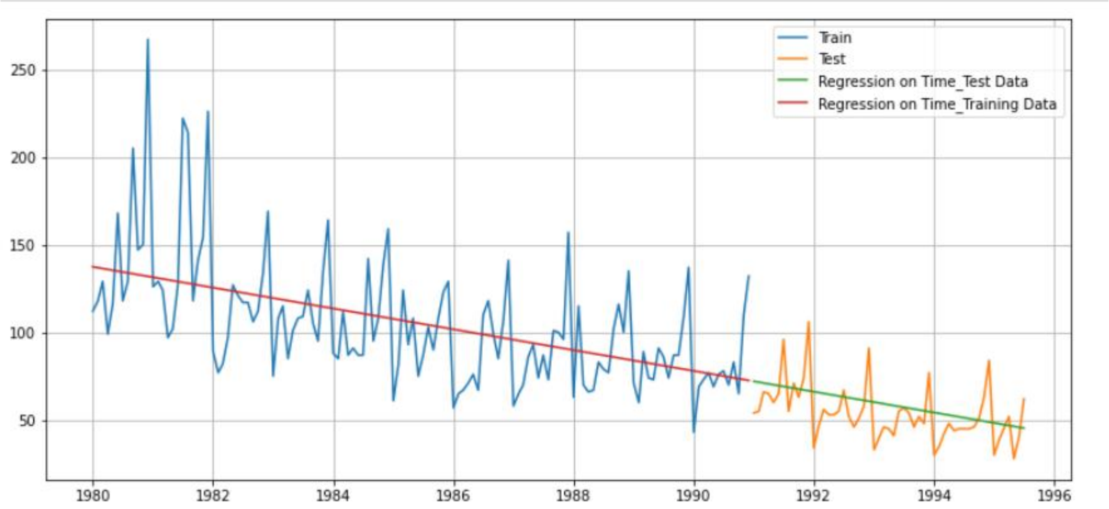
RMSE and MAPE for testing Data(Sparkling)



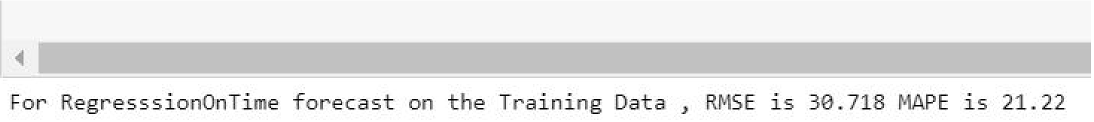
```
For RegresssionOnTime forecast on the Test Data , RMSE is 1389.135 MAPE is 50.15
```

Rose :

Linear regression model for Rose Dataset

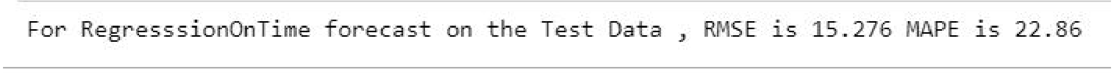


RMSE and MAPE for Training Data (Rose)



```
For RegresssionOnTime forecast on the Training Data , RMSE is 30.718 MAPE is 21.22
```

RMSE and MAPE for testing Data (Rose)



```
For RegresssionOnTime forecast on the Test Data , RMSE is 15.276 MAPE is 22.86
```

Naive Forecasting :

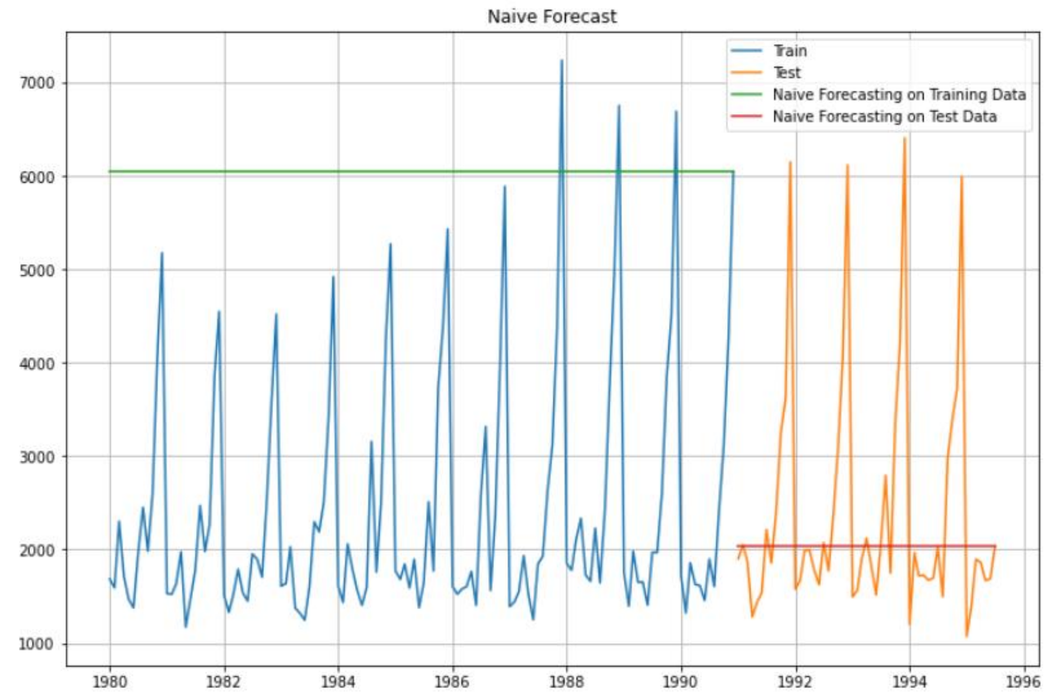
In naive model, the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

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 datasets.

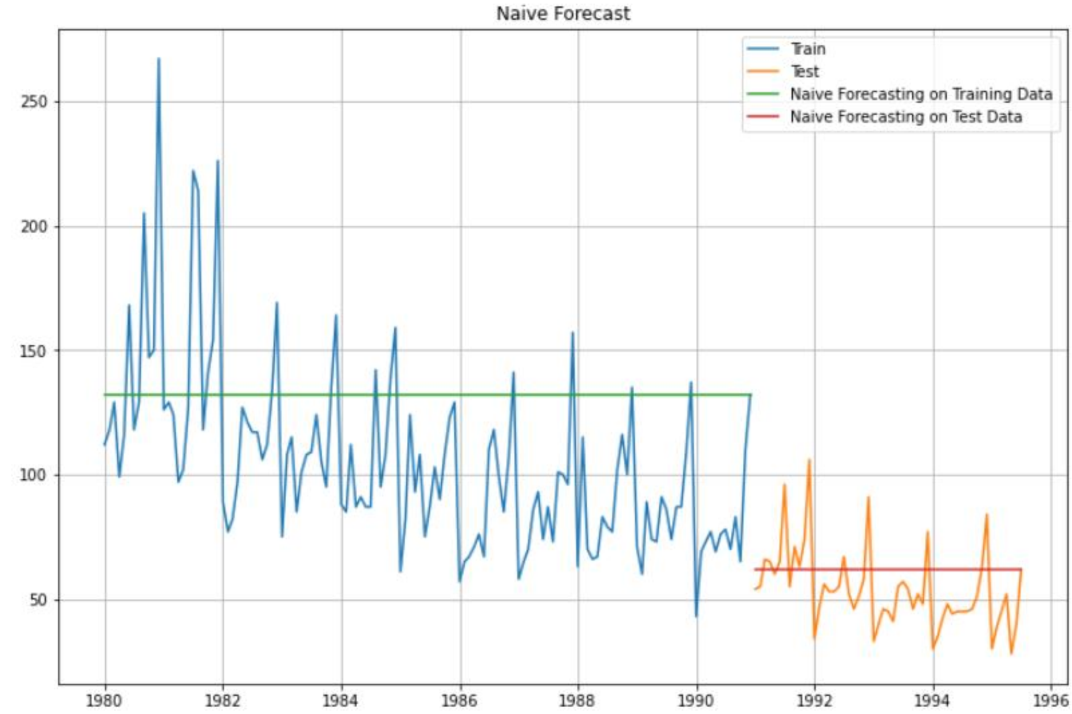
Naive forecasting for Sparkling:



For NaiveModel forecast on the Testing Data , RMSE is 1327.156 MAPE is 32.90

For NaiveModel forecast on the Training Data , RMSE is 3867.701 MAPE is 153.17

Rose :
Naive forecasting for Rose



For NaiveModel forecast on the Training Data , RMSE is 45.064 MAPE is 36.38

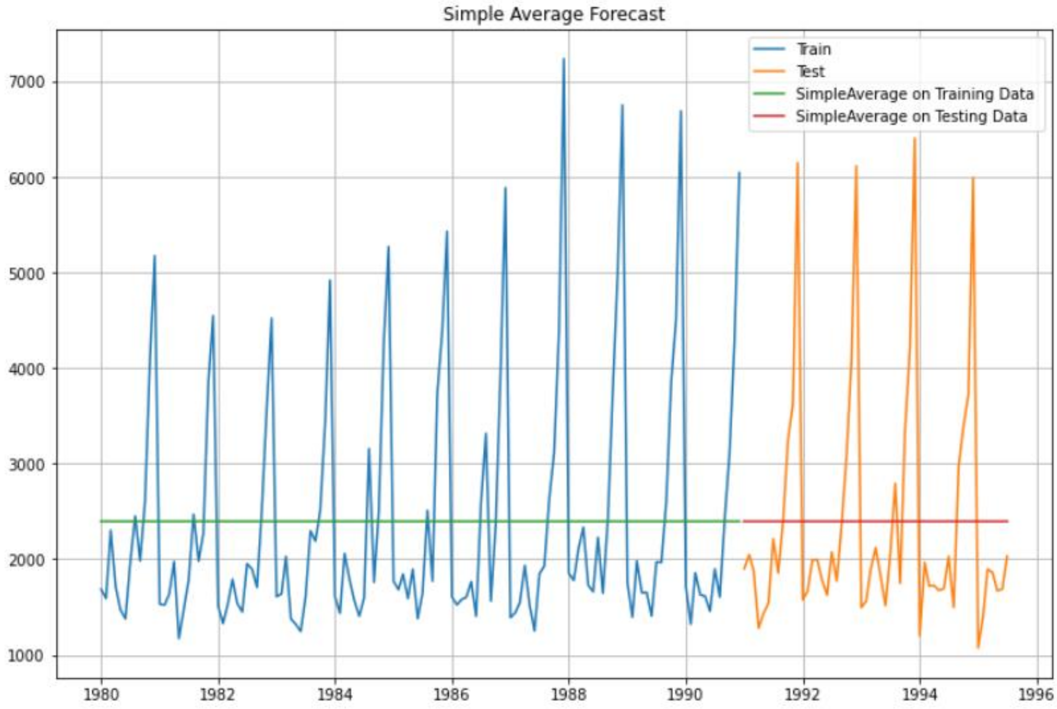
For NaiveModel forecast on the Testing Data , RMSE is 17.757 MAPE is 27.46

Simple Average Foreacasting:

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 .

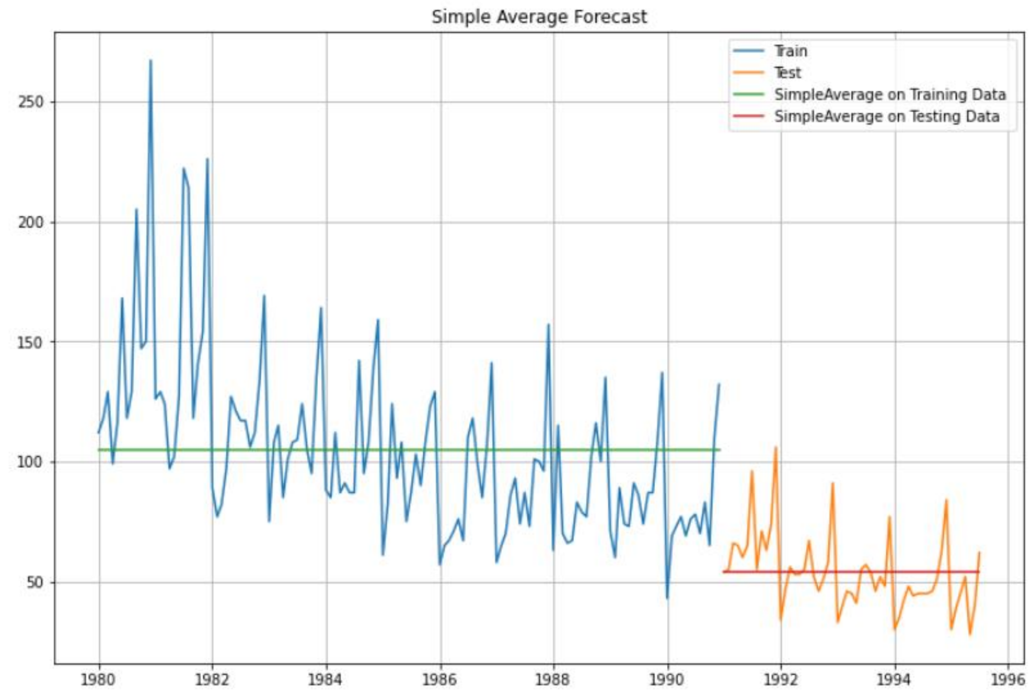
Simple average forecast of Sparkling :



For SimpleAverage forecast on the Training Data , RMSE is 1298.484 MAPE is 40.36

For SimpleAverage forecast on the Training Data , RMSE is 1275.073 MAPE is 38.81

Simple average forecast of Rose :



For SimpleAverage forecast on the Training Data , RMSE is 36.034 MAPE is 25.39

For SimpleAverage forecast on the Training Data , RMSE is 15.770 MAPE is 21.41

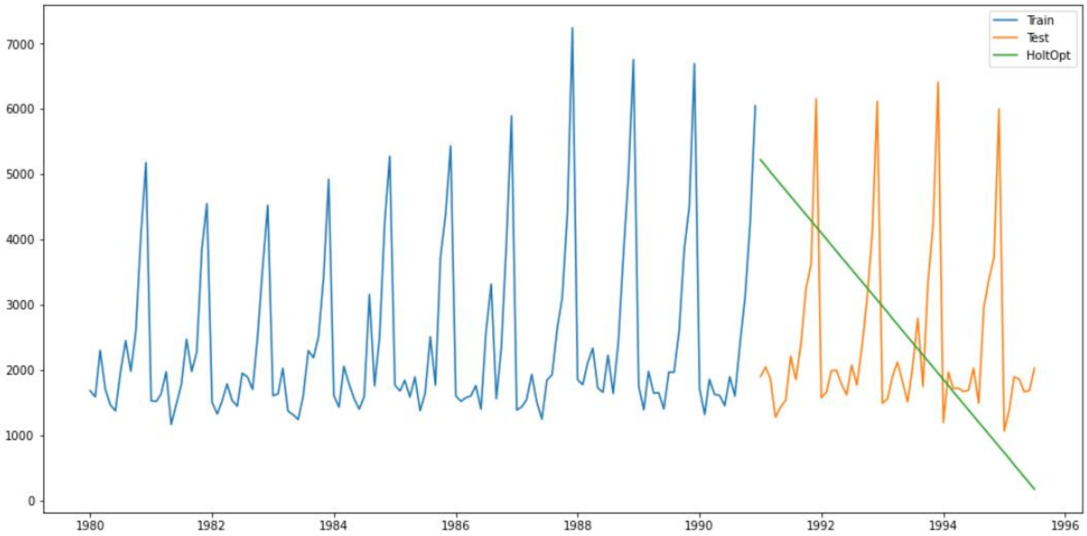
Double Exponential Smoothing: Sparkling

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

Double Exponential Smoothing



For alpha = 0.69, RMSE is 2007.2385 MAPE is 68.23

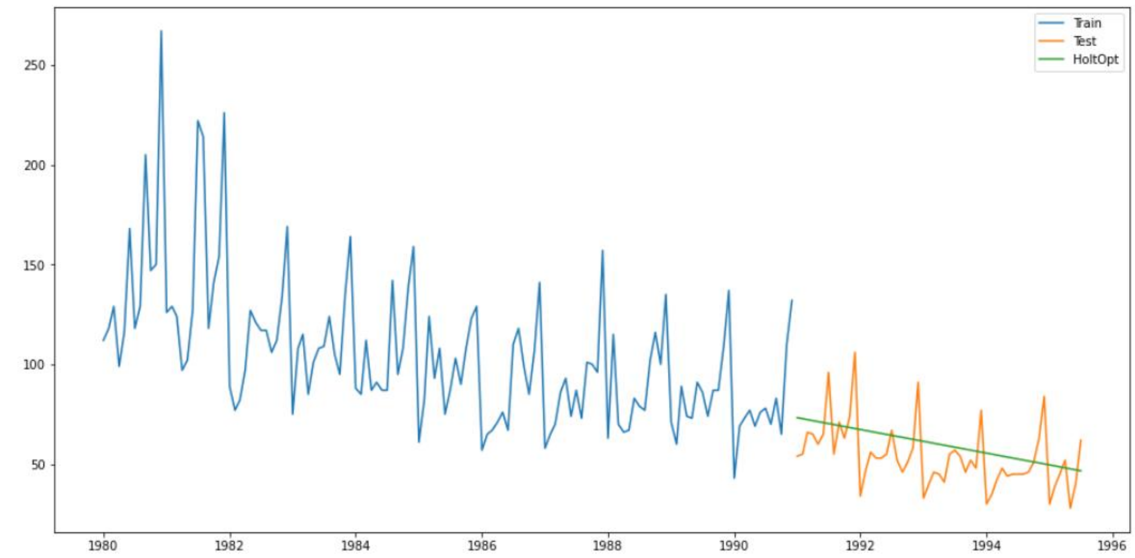
Rose:

The Double Exponential Smoothing models is applicable when data has trend, but no seasonality. 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 chose the optimized values using parameters '*optimized=True, use_brute=True*' .

Double Exponential Smoothing



For alpha = 0.02, RMSE is 15.7151 MAPE is 24.16

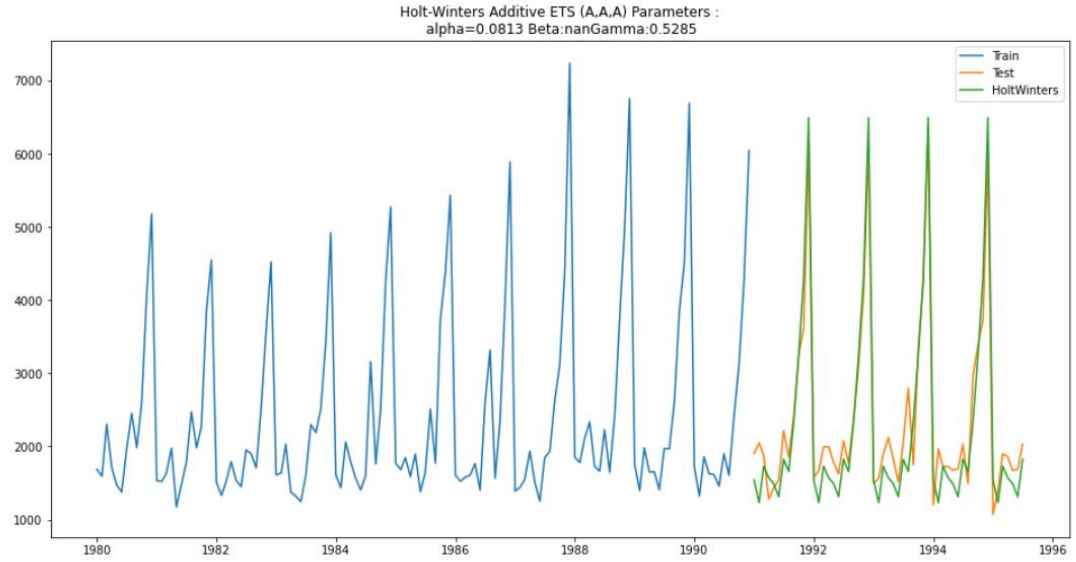
Triple Exponential Smoothing

Sparkling

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.*



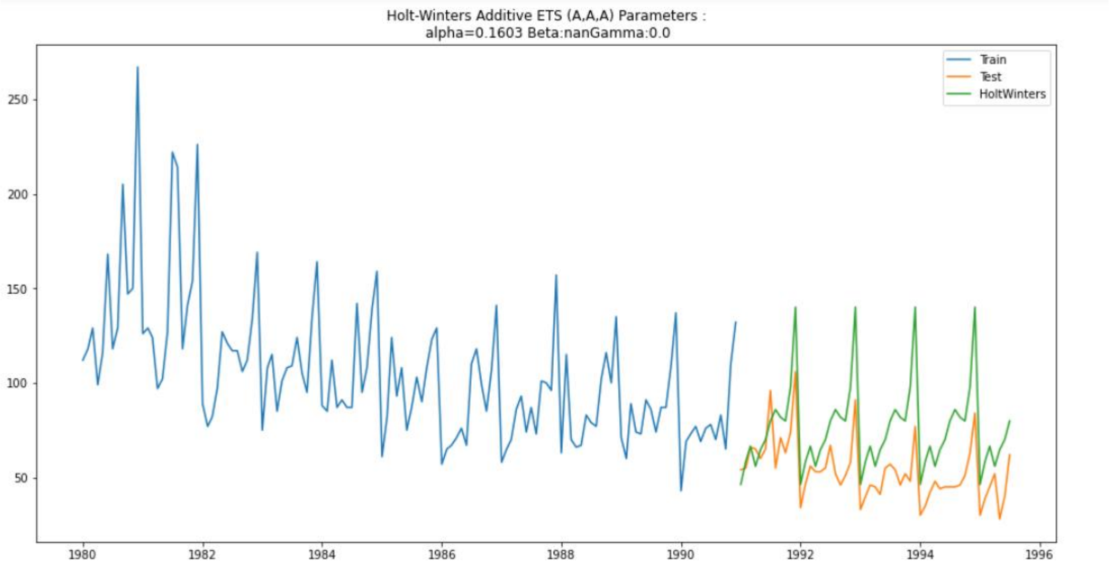
The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but faired poorly in accuracy in test .

- 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.

For alpha = 0.08, RMSE is 369.9635 MAPE is 12.35

Rose:



For alpha = 0.16, RMSE is 26.7093 MAPE is 42.41

Simple Exponential Smoothing
Sparkling :

Simple Exponential Smoothing is applied if the time-series has neither a trend nor seasonality, which is not the case with the given data .

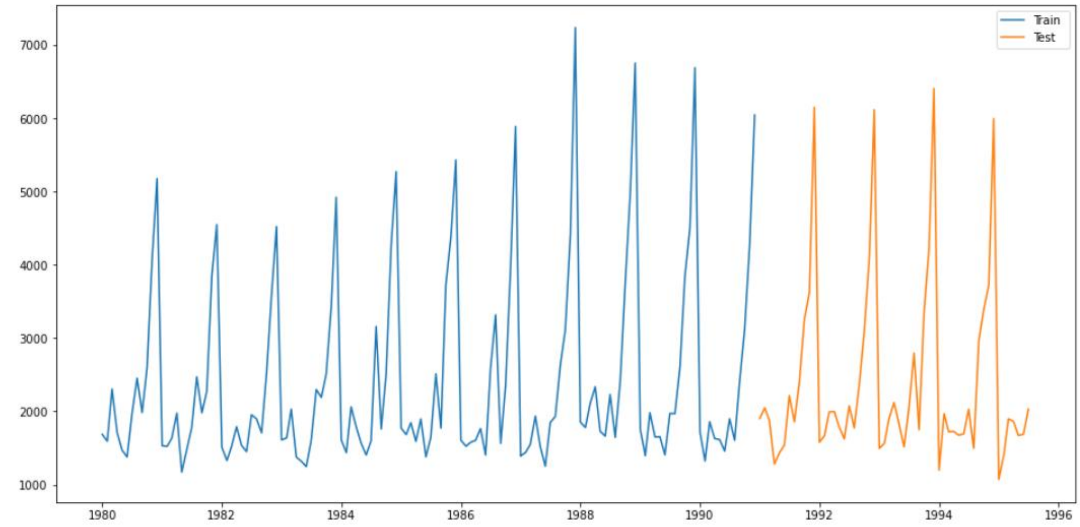
- The forecasting using smoothing levels of alpha between 0 and 1 are as below, where the smoothing levels are passed manually .

- 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.

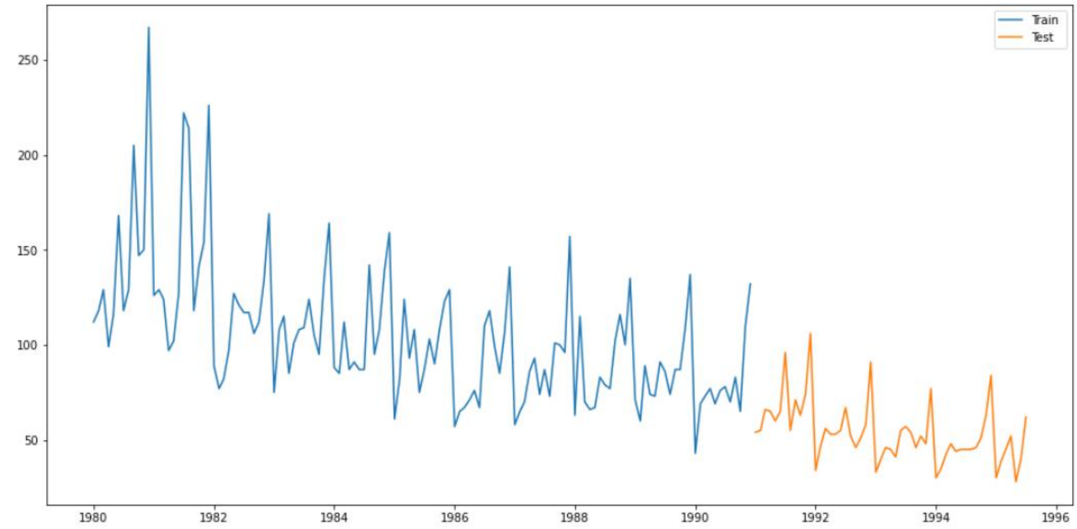
On the second iteration, the model was ran without passing a value for alpha and used parameters ‘*optimized=True, use_brute=True’* .

- The autofit model picked 0.0 as the smoothing parameter and 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.



For alpha = 0.50,RMSE is 2666.3514 MAPE is 106.27 (Test data)

Rose :



For alpha = 0.50,RMSE is 59.6619 MAPE is 106.88 (Test Data)

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.

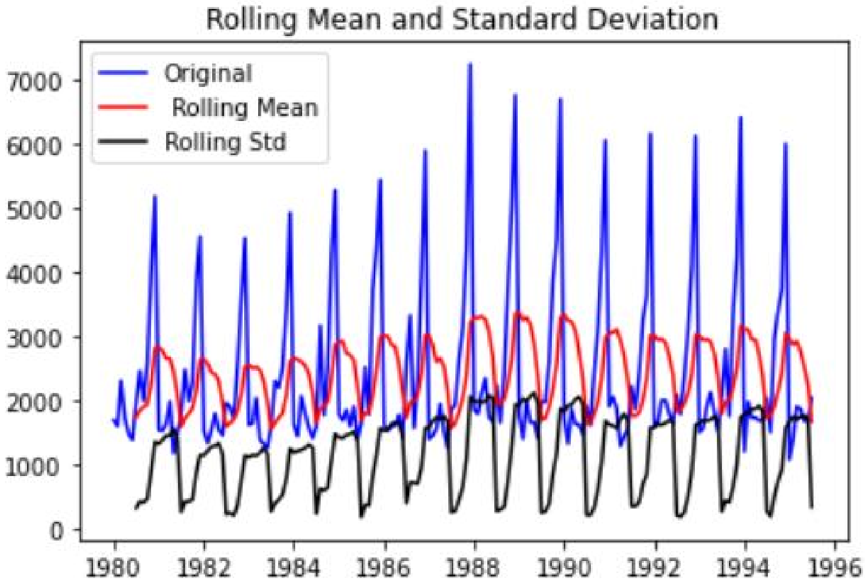
Sparkling :

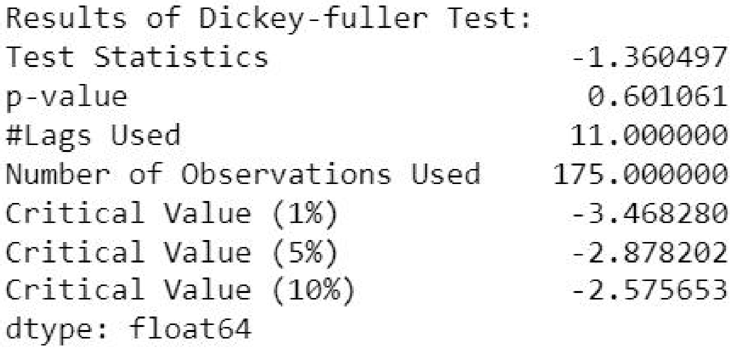
Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. This test determine's the presence of unit root in the series to understand if the series is stationary or not .

• **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 below values, where p-value is greater than alpha .05 so we fail to reject the null hypothesis.

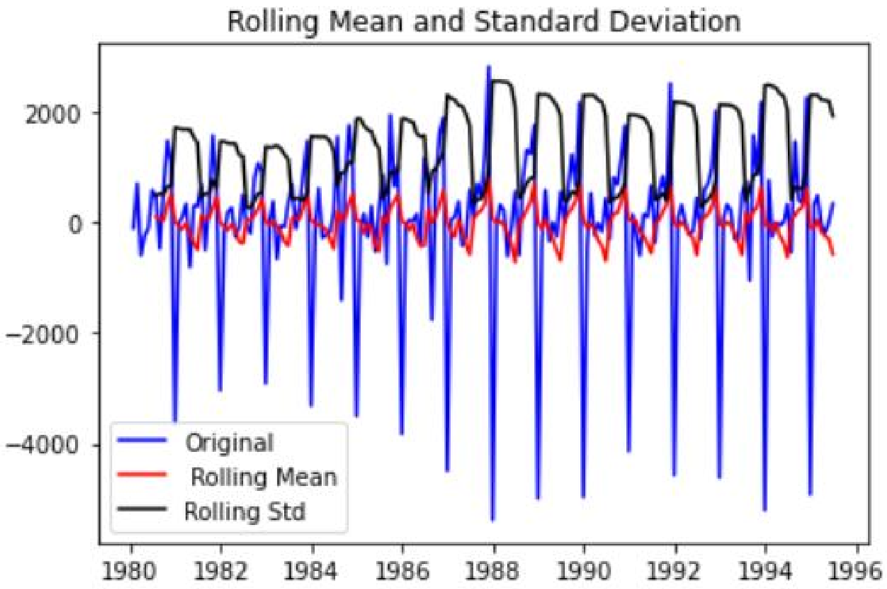


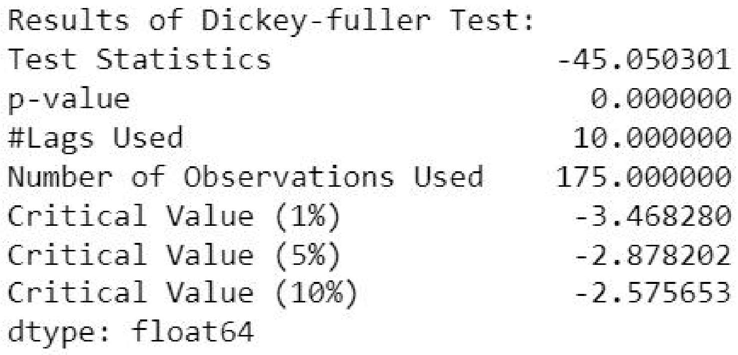


```
Results of Dickey-fuller Test:
Test Statistics                 -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value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

ADF on original series

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





```
Results of Dickey-fuller Test:
Test Statistics                 -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
```

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 deviation is seen changing according to change in slope, which indicates multiplicity .

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

ADF on differenced series

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

Rose :

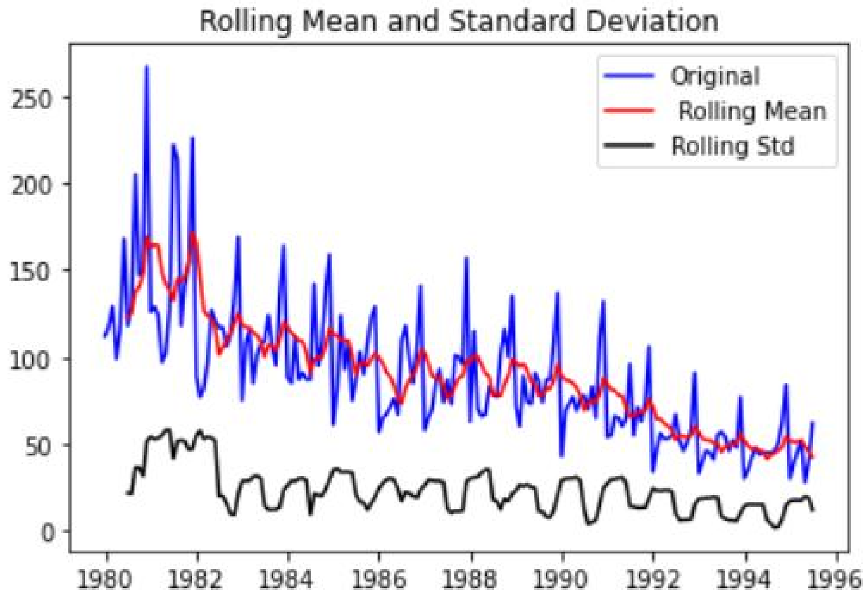
Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determine the presence of unit root in the series to understand if the series is stationary or not.

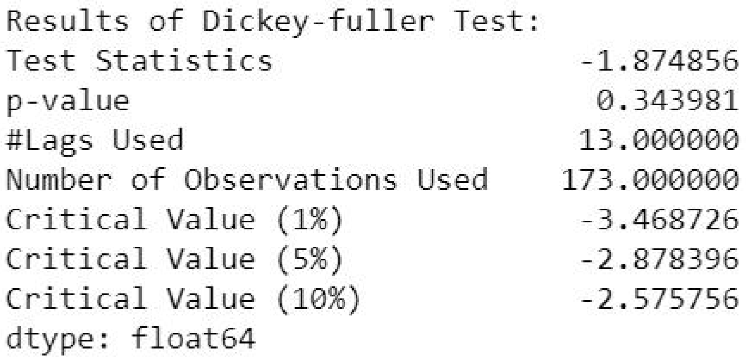
• **Null Hypothesis**: The series has a unit root, that is series is non-stationary.

• **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 greater than alpha .05 so we fail to reject the null hypothesis.





```
Results of Dickey-fuller Test:
Test Statistics                  -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
```

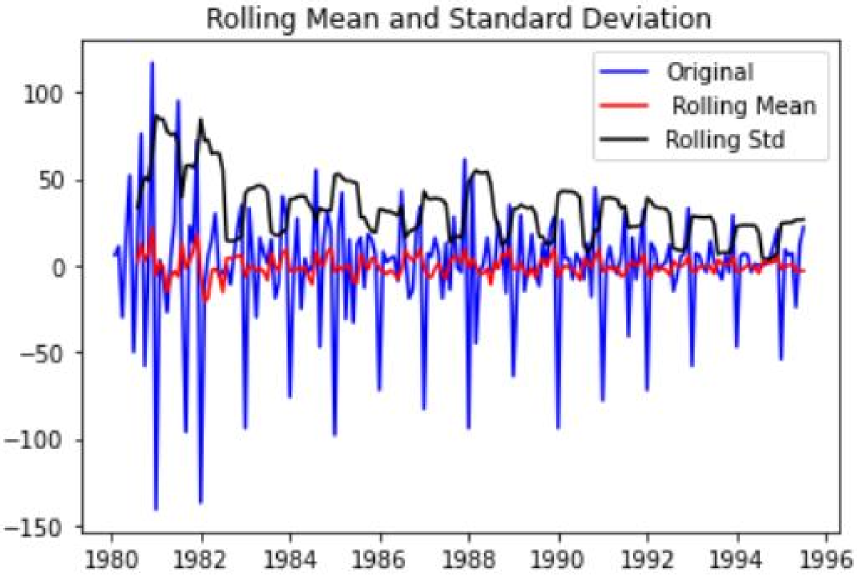
ADF on original series

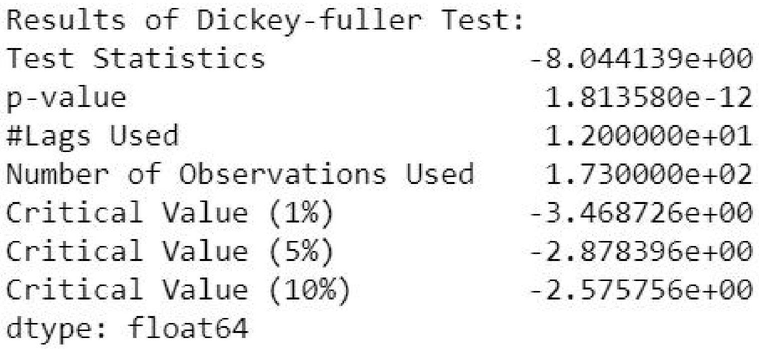
• P-Value > alpha .05

• Test statistic > Critical values

• Fail to reject the null hypothesis

• The series is non-stationary





```
Results of Dickey-fuller Test:
Test Statistics                 -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
```

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 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.

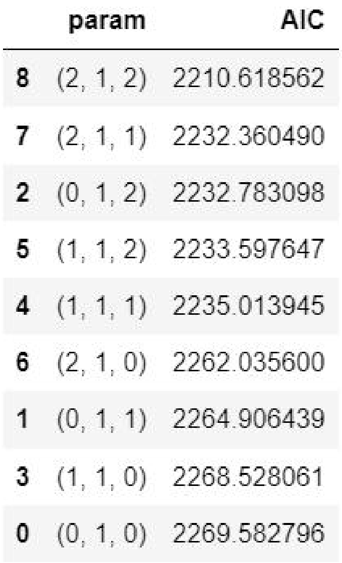
ADF on differenced series

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

6 . Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

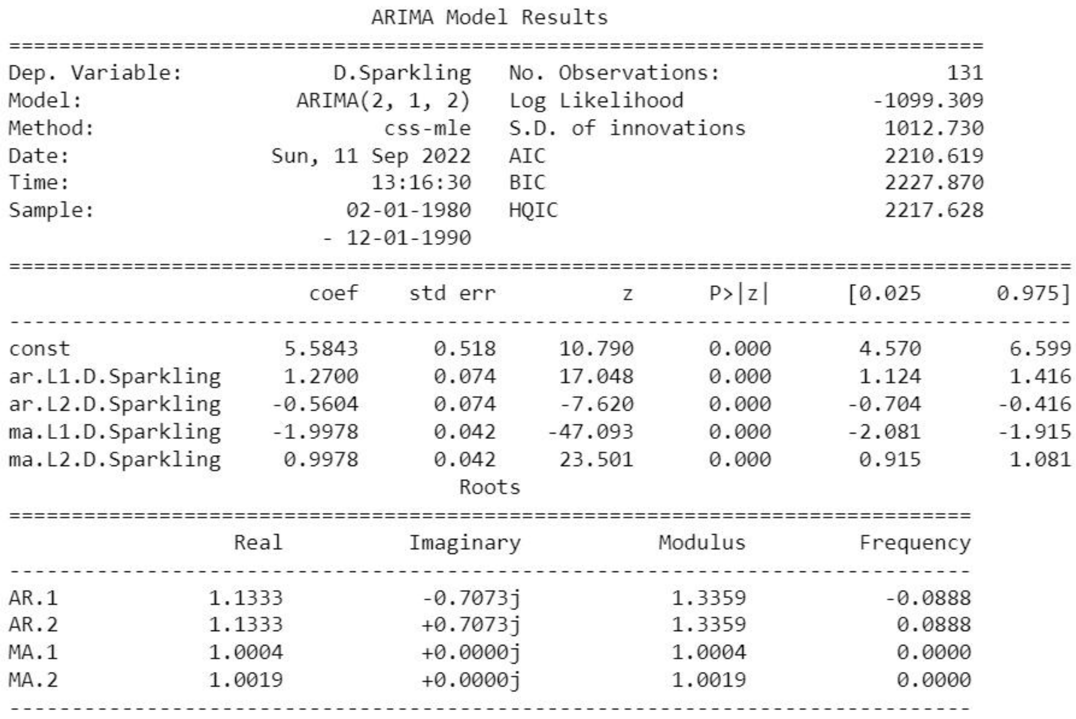
ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. A statistical model is autoregressive if it predicts future values based on past values.

ARIMA on Sparkling Wine :



	param	AIC
8	(2, 1, 2)	2210.618562
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906439
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

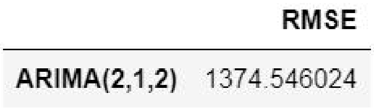
Results of running automated ARIMA model on Sparkling Wine Dataset:



```
                              ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.309
Method:                       css-mle   S.D. of innovations           1012.730
Date:                Sun, 11 Sep 2022   AIC                           2210.619
Time:                        13:16:30   BIC                           2227.870
Sample:                    02-01-1980   HQIC                          2217.628
                         - 12-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5843      0.518     10.790      0.000       4.570       6.599
ar.L1.D.Sparkling     1.2700      0.074     17.048      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5604      0.074     -7.620      0.000      -0.704      -0.416
ma.L1.D.Sparkling    -1.9978      0.042    -47.093      0.000      -2.081      -1.915
ma.L2.D.Sparkling     0.9978      0.042     23.501      0.000       0.915       1.081
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1333          -0.7073j            1.3359           -0.0888
AR.2            1.1333          +0.7073j            1.3359            0.0888
MA.1            1.0004          +0.0000j            1.0004            0.0000
MA.2            1.0019          +0.0000j            1.0019            0.0000
-----------------------------------------------------------------------------
```

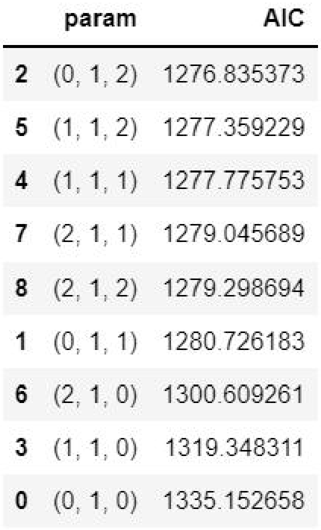
The Lowest AIC for Sparkling Wine Dataset is 2210.6 for p,d,q values of 2,1,2 respectively .

The RMSE for Sparkling Wine Dataset:



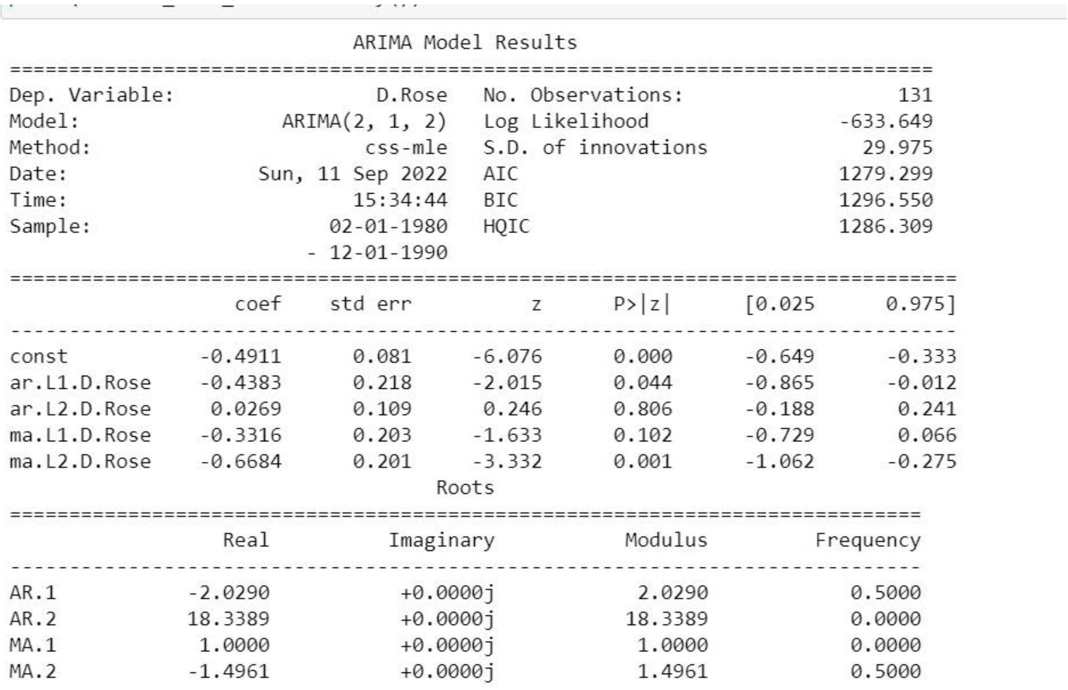
	RMSE
ARIMA(2,1,2)	1374.546024

ARIMA on Rose Wine Dataset :



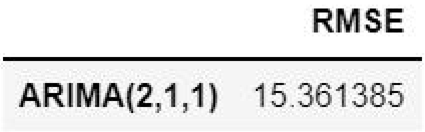
	param	AIC
2	(0, 1, 2)	1276.835373
5	(1, 1, 2)	1277.359229
4	(1, 1, 1)	1277.775753
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

Results of running automated ARIMA model on Rose Wine Dataset:



```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -633.649
Method:                       css-mle   S.D. of innovations             29.975
Date:                Sun, 11 Sep 2022   AIC                           1279.299
Time:                        15:34:44   BIC                           1296.550
Sample:                    02-01-1980   HQIC                          1286.309
                         - 12-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4911      0.081     -6.076      0.000      -0.649      -0.333
ar.L1.D.Rose    -0.4383      0.218     -2.015      0.044      -0.865      -0.012
ar.L2.D.Rose     0.0269      0.109      0.246      0.806      -0.188       0.241
ma.L1.D.Rose    -0.3316      0.203     -1.633      0.102      -0.729       0.066
ma.L2.D.Rose    -0.6684      0.201     -3.332      0.001      -1.062      -0.275
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -2.0290           +0.0000j            2.0290            0.5000
AR.2           18.3389           +0.0000j           18.3389            0.0000
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.4961           +0.0000j            1.4961            0.5000
```

The RMSE for Rose Wine Dataset:



	RMSE
ARIMA(2,1,1)	15.361385

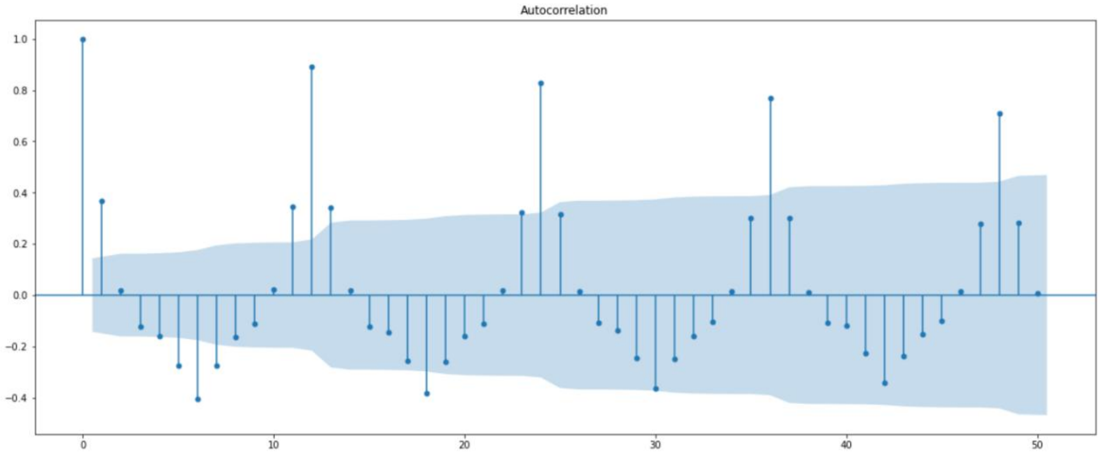
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

From the ACF plot of the observed/ train data, it can be inferred that at seasonal interval of 6, the plot is not quickly tapering off. So, a seasonal differencing of 6 or it’s multiple has to be taken.

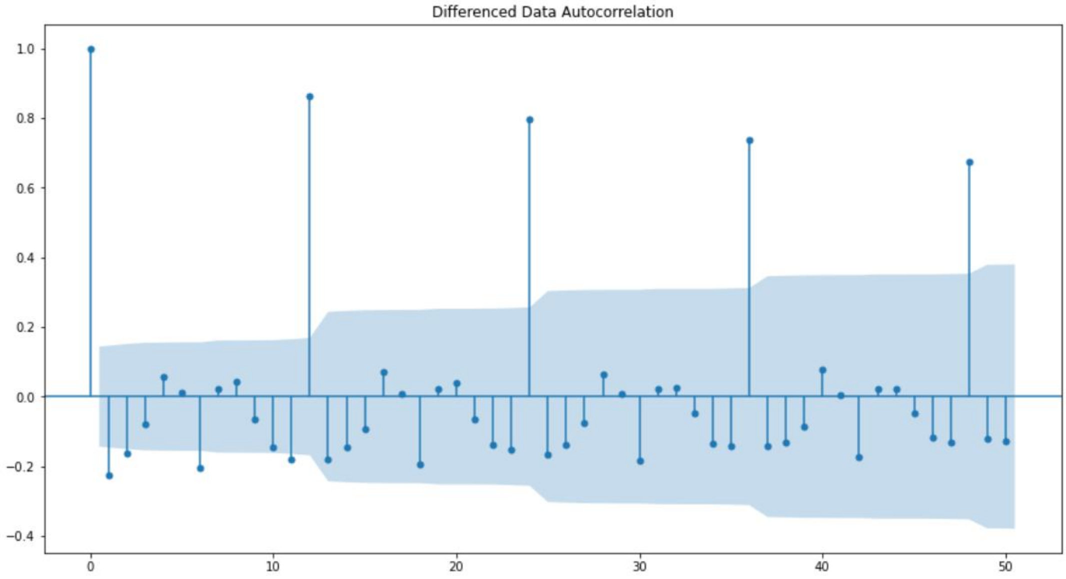
From the plots below an apparent slight trend is still existing after differencing of seasonal order of 6 or its multiple. With a further differencing of order one, no trend is present.

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.

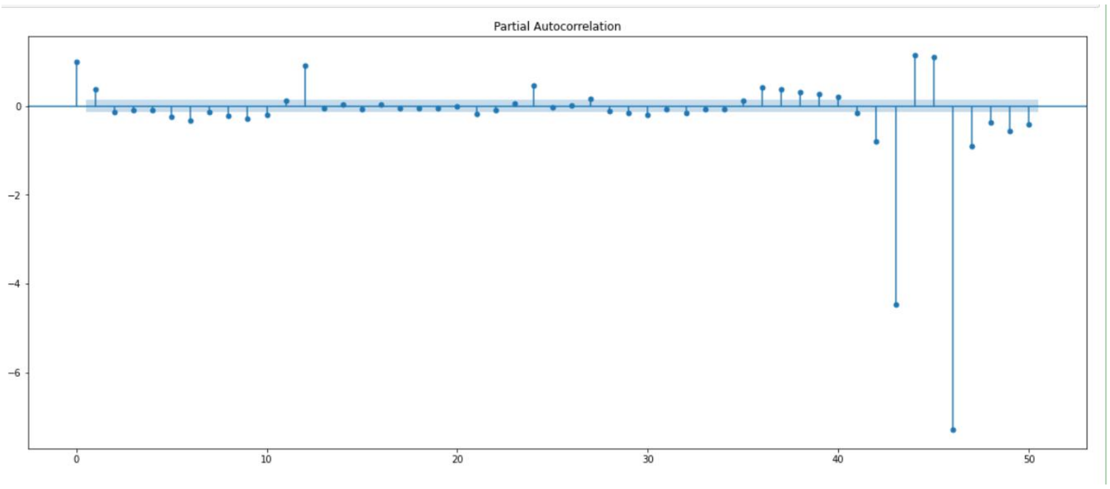
Autocorrelation



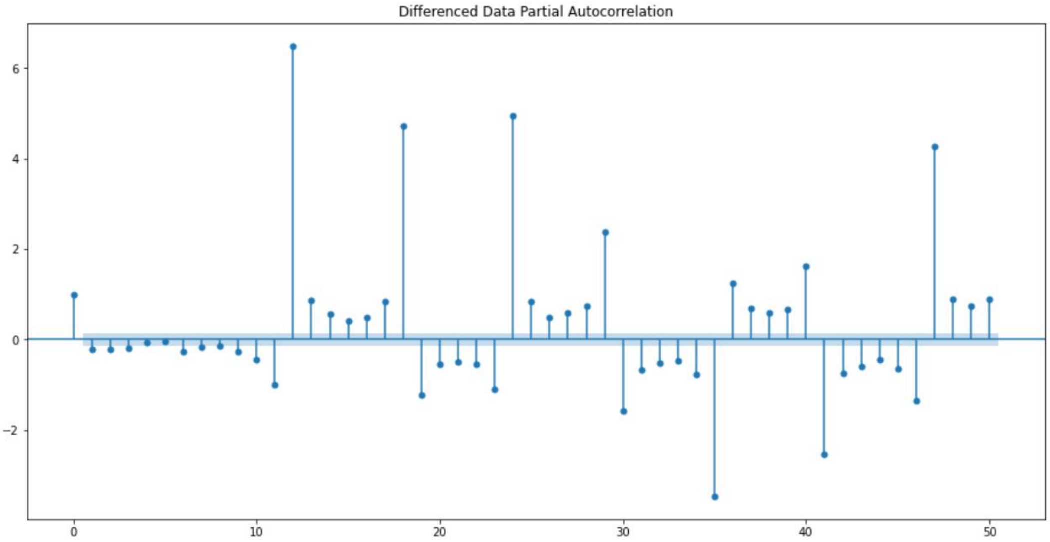
Differenced Data Auto Correlation



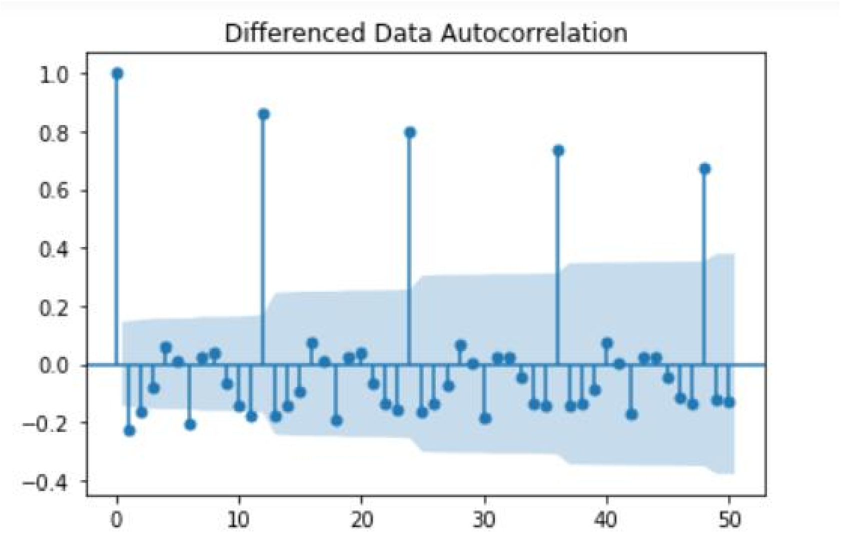
Partial Autocorrelation

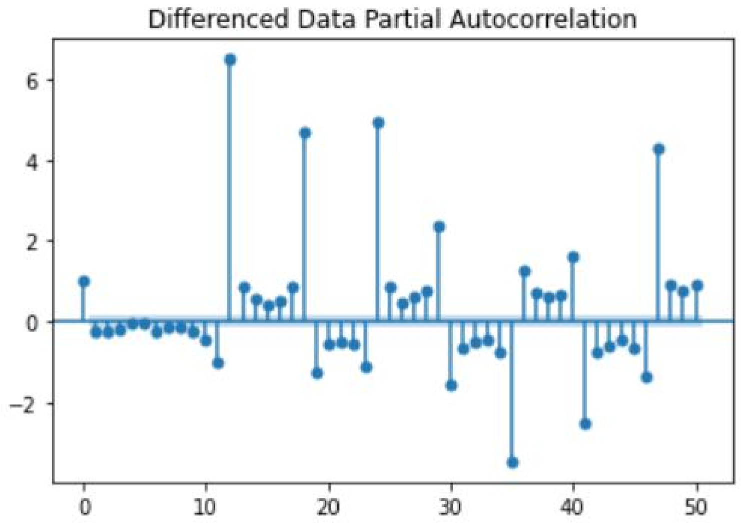


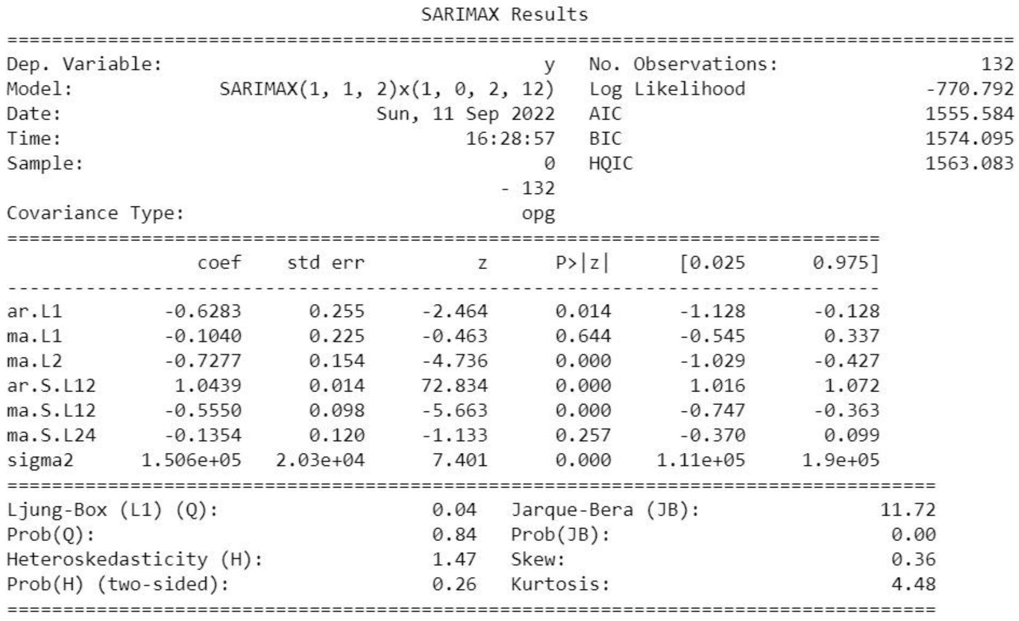
Differenced Data Partial Autocorrelation:



Differenced Data AutoCorrelation and Differenced data Partial Autocorrelation







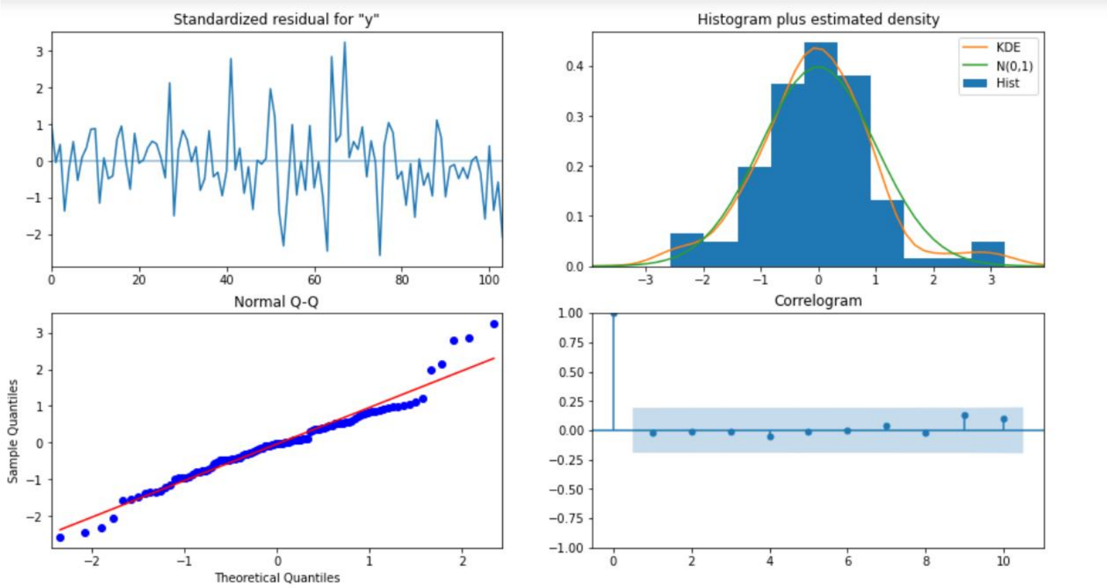
```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                 132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                             Sun, 11 Sep 2022   AIC                           1555.584
Time:                                     16:28:57   BIC                           1574.095
Sample:                                          0   HQIC                          1563.083
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6283      0.255     -2.464      0.014      -1.128      -0.128
ma.L1         -0.1040      0.225     -0.463      0.644      -0.545       0.337
ma.L2         -0.7277      0.154     -4.736      0.000      -1.029      -0.427
ar.S.L12       1.0439      0.014     72.834      0.000       1.016       1.072
ma.S.L12      -0.5550      0.098     -5.663      0.000      -0.747      -0.363
ma.S.L24      -0.1354      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.401      0.000    1.11e+05     1.9e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                11.72
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               1.47   Skew:                             0.36
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.48
===================================================================================
```

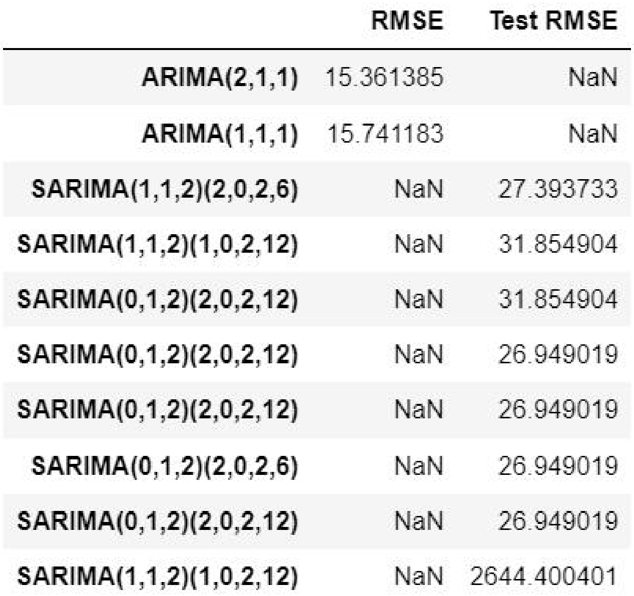
The above figure shows SARIMA model (1,1,2) X (1,0,2,12) with AIC 1555.584

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 .

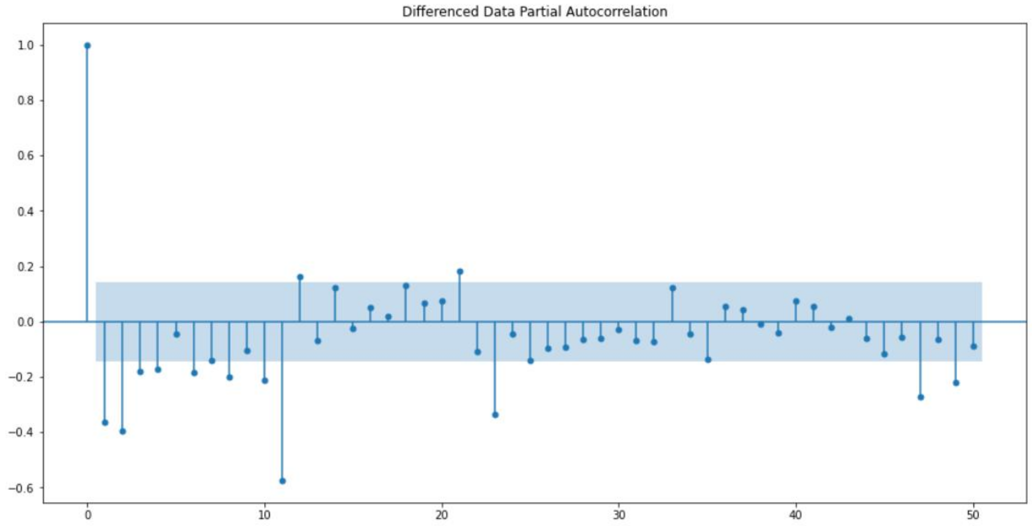


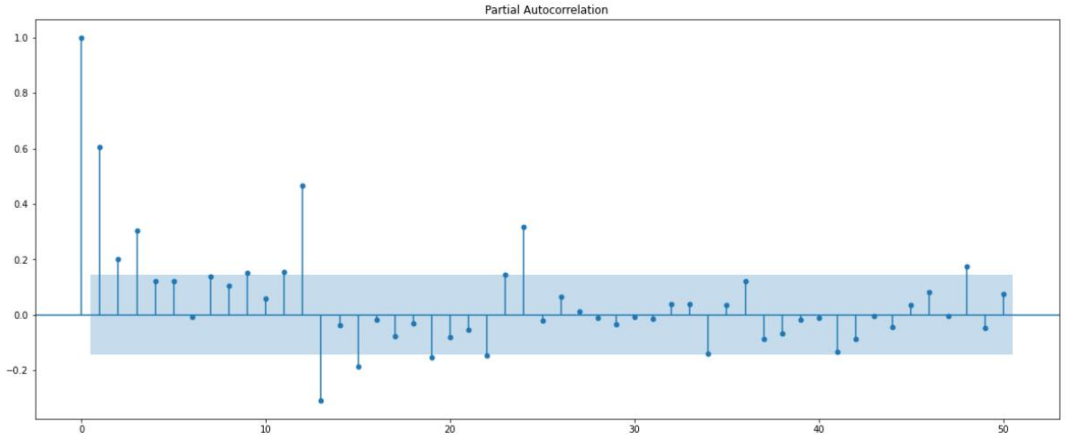


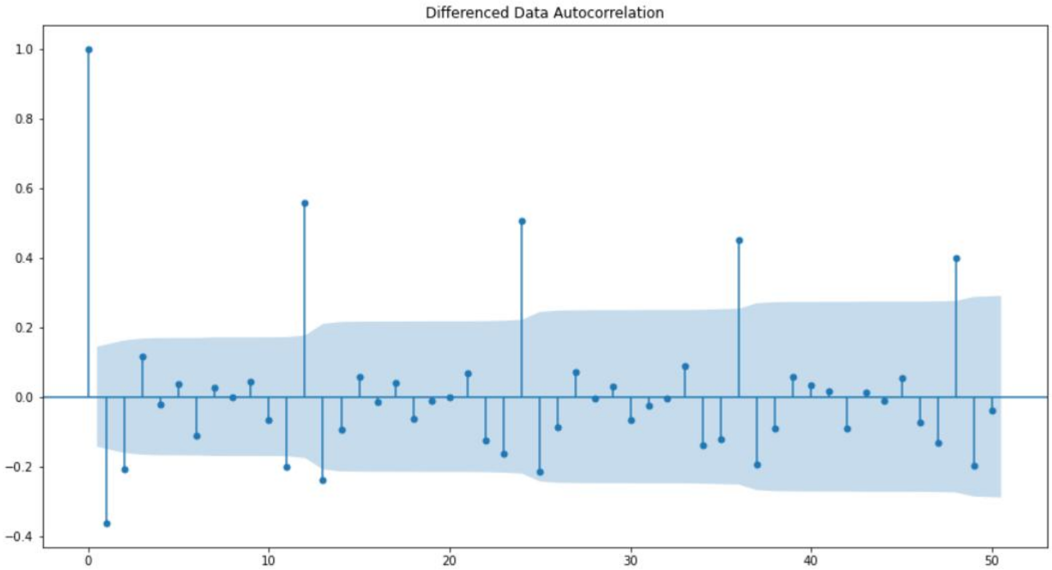
	RMSE	Test RMSE
ARIMA(2,1,1)	15.361385	NaN
ARIMA(1,1,1)	15.741183	NaN
SARIMA(1,1,2)(2,0,2,6)	NaN	27.393733
SARIMA(1,1,2)(1,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,6)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(1,1,2)(1,0,2,12)	NaN	2644.400401

Rose :

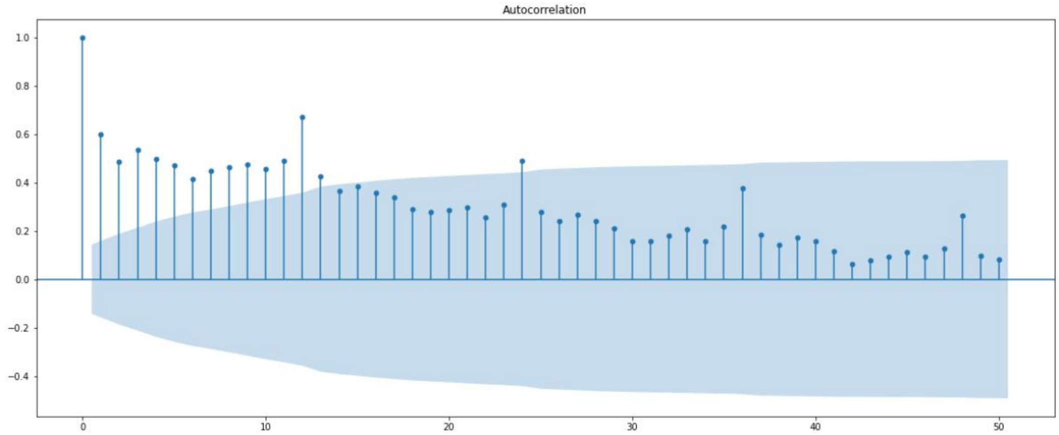
Here we have taken alpha = 0.05
ACF and PACF plots of the seasonal-differenced + one order differenced data is created to find the values for (p,d,q)x(P,D,Q).



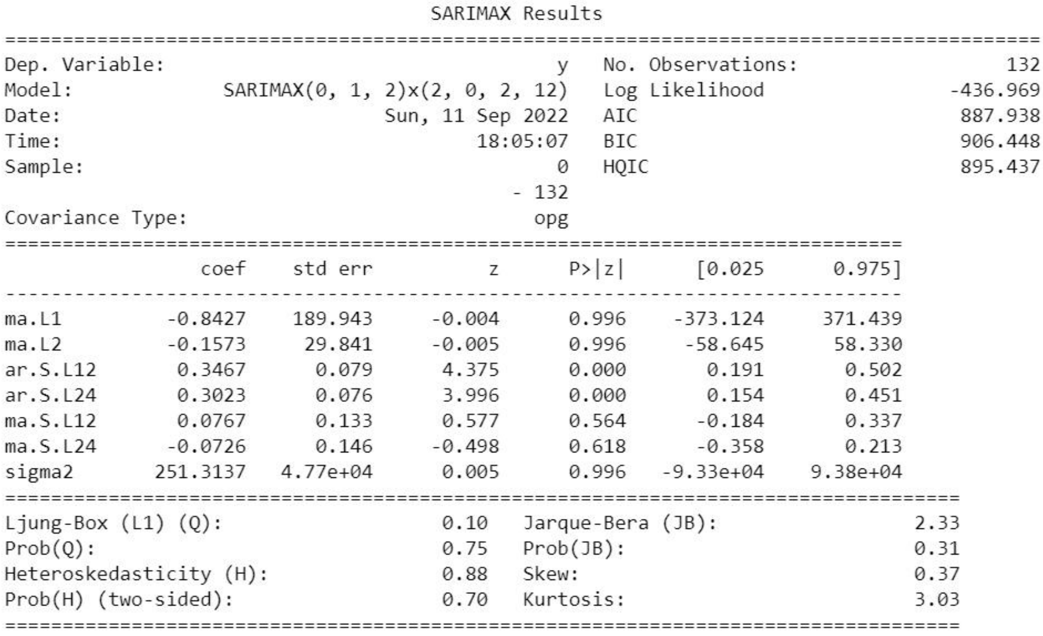




Differenced Data Autocorrelation



Autocorrelation



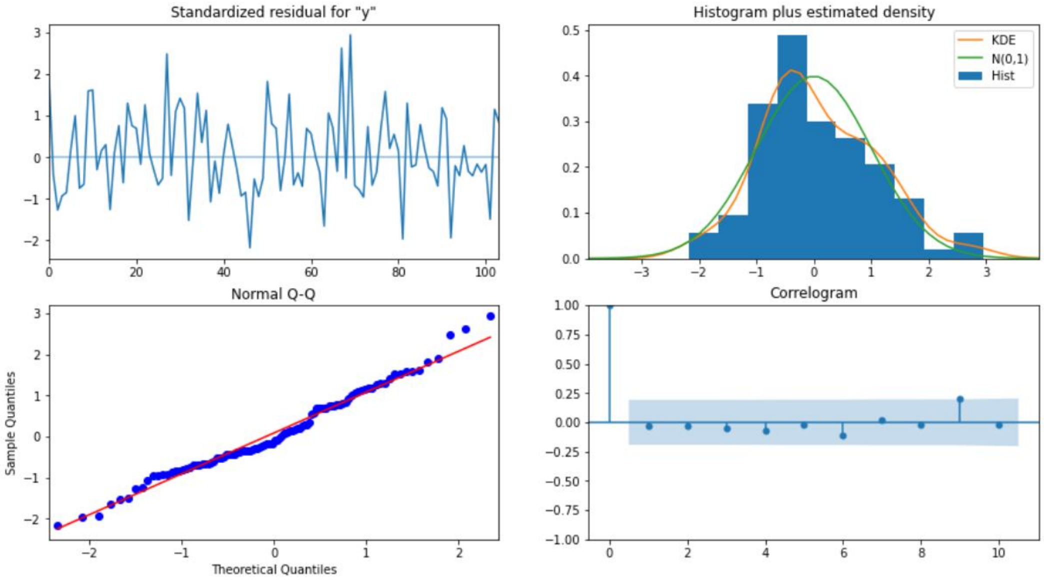
```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Sun, 11 Sep 2022   AIC                            887.938
Time:                                    18:05:07   BIC                            906.448
Sample:                                         0   HQIC                           895.437
                                            - 132                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.943     -0.004      0.996    -373.124     371.439
ma.L2         -0.1573     29.841     -0.005      0.996     -58.645      58.330
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3137   4.77e+04      0.005      0.996   -9.33e+04    9.38e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.75   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================
```

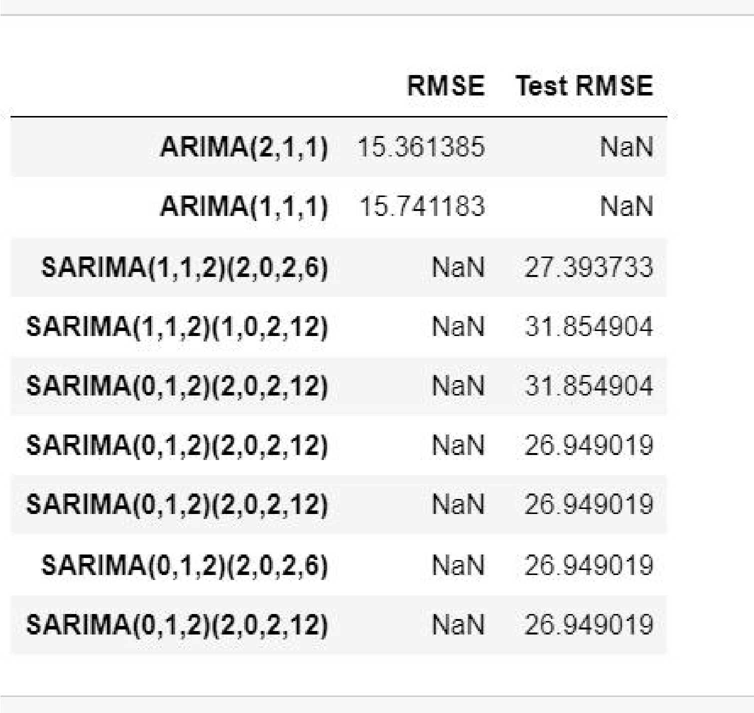
The final SARIMA model (0,1,2) x (2,0,2,12)

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 .





	RMSE	Test RMSE
ARIMA(2,1,1)	15.361385	NaN
ARIMA(1,1,1)	15.741183	NaN
SARIMA(1,1,2)(2,0,2,6)	NaN	27.393733
SARIMA(1,1,2)(1,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,6)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019

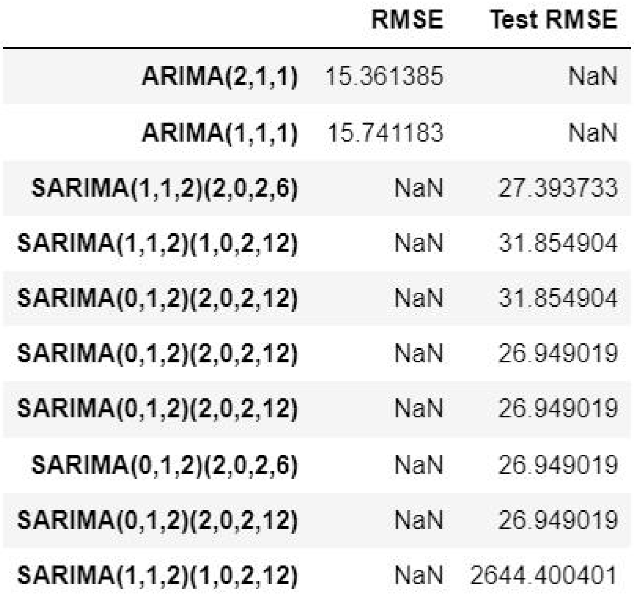
8 . Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

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 SARIMA and Triple Exponential Smoothing are found to be comparable in terms of performance and fitment with the test data.



	RMSE	Test RMSE
ARIMA(2,1,1)	15.361385	NaN
ARIMA(1,1,1)	15.741183	NaN
SARIMA(1,1,2)(2,0,2,6)	NaN	27.393733
SARIMA(1,1,2)(1,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,6)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(1,1,2)(1,0,2,12)	NaN	2644.400401

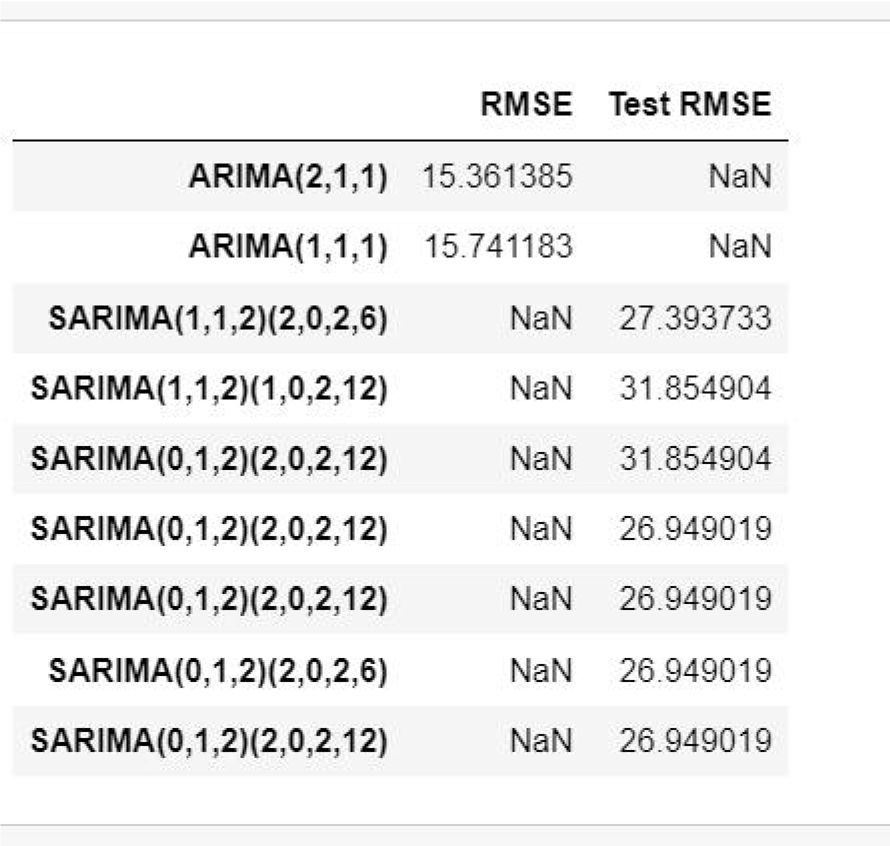
Rose:

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 .

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.



	RMSE	Test RMSE
ARIMA(2,1,1)	15.361385	NaN
ARIMA(1,1,1)	15.741183	NaN
SARIMA(1,1,2)(2,0,2,6)	NaN	27.393733
SARIMA(1,1,2)(1,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,6)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019

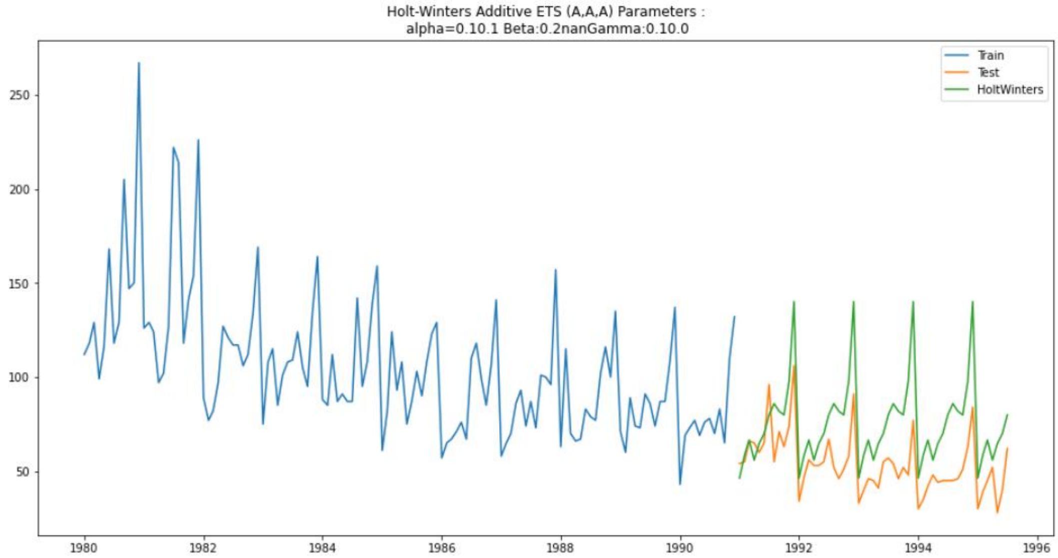
9.Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Till now, the forecast has been done on the test data, now let's do the forecast on the full data using the same parameters and taking the model that gives least RMSE.

The two models to be built on the whole data are the following:

1. Alpha=0.1, Beta=0.2, Gamma=0.1, Triple Exponential Smoothing
2. Manual SARIMA (0,1,2) x (2,0,2,12)

1. Alpha=0.1, Beta=0.2, Gamma=0.1, Triple Exponential Smoothing



Model showing 12 month prediction using TES

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.

2. Manual SARIMA (0,1,2) x (2,0,2,12)

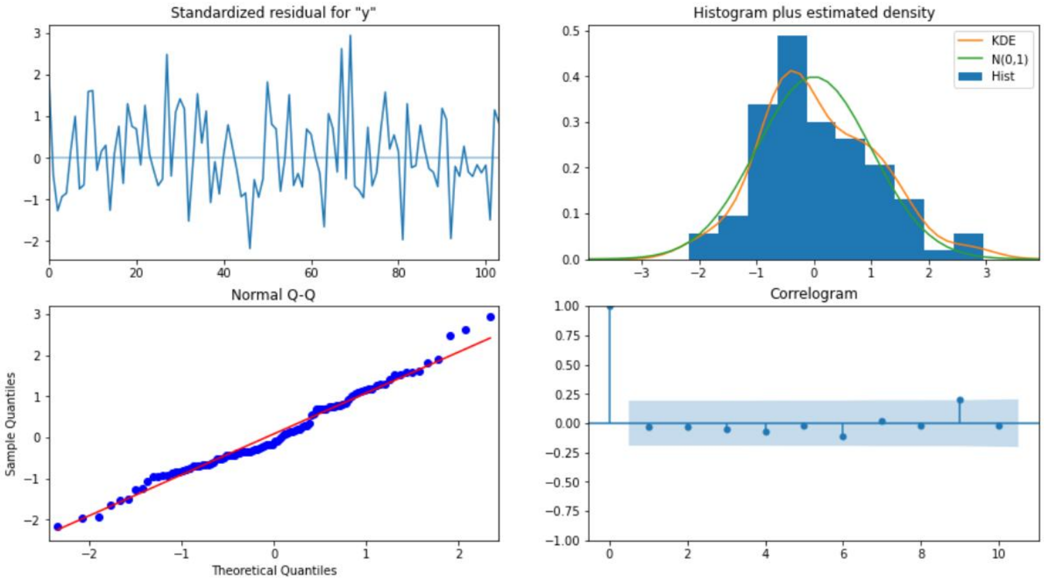
The diagnostics plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in normal QQ plot.

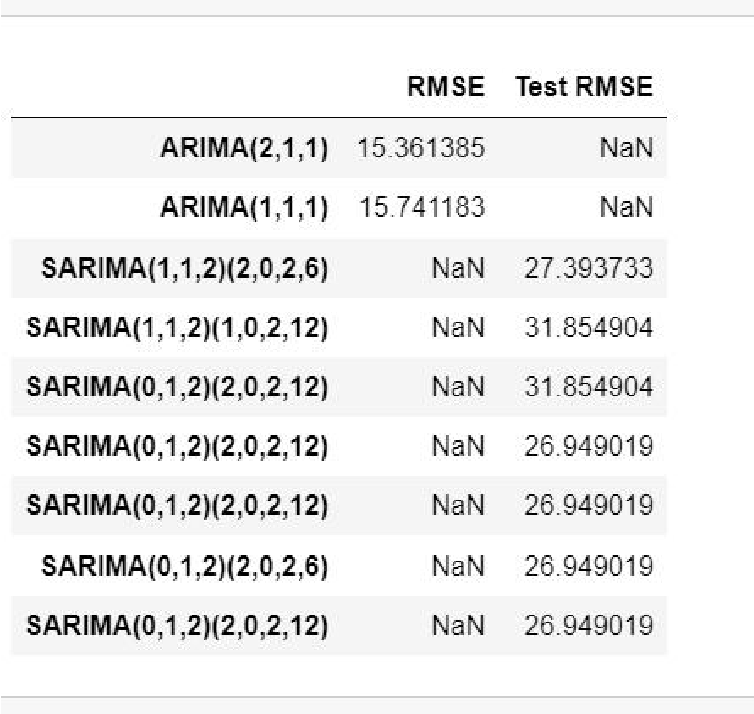
The model summary also provides valuable insights in the model. From the snapshot of summary below it can be understood that MA (1) term has the highest weightage. The p-values indicates that the term MA (1) is the most significant term.

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.

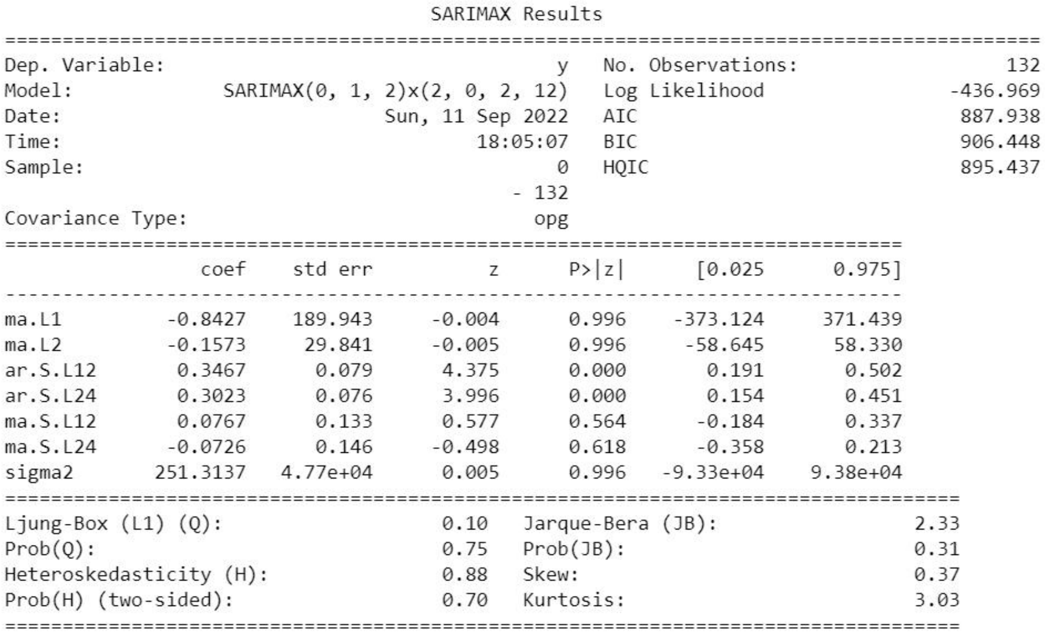
Prediction on the Rose time-series is on a wider confidence band than sparkling

12 month forecast using SARIMA model For Rose wine





	RMSE	Test RMSE
ARIMA(2,1,1)	15.361385	NaN
ARIMA(1,1,1)	15.741183	NaN
SARIMA(1,1,2)(2,0,2,6)	NaN	27.393733
SARIMA(1,1,2)(1,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	31.854904
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,6)	NaN	26.949019
SARIMA(0,1,2)(2,0,2,12)	NaN	26.949019



```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y  No. Observations:                  132
Model:              SARIMAX(0, 1, 2)x(2, 0, 2, 12)  Log Likelihood                -436.969
Date:                             Sun, 11 Sep 2022  AIC                            887.938
Time:                                     18:05:07  BIC                            906.448
Sample:                                          0  HQIC                           895.437
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.943     -0.004      0.996    -373.124     371.439
ma.L2         -0.1573     29.841     -0.005      0.996     -58.645      58.330
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3137   4.77e+04      0.005      0.996   -9.33e+04    9.38e+04
====================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                  2.33
Prob(Q):                              0.75   Prob(JB):                          0.31
Heteroskedasticity (H):               0.88   Skew:                              0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                          3.03
====================================================================================
```

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

Summary :

1. Monthly sales given in the period from January 1980 to July 1995.

2. A total of 187 entries are provided in the dataset.

3. No missing values present in the dataset.

4. A time stamp has been created that is coded as index.

5. The dataset shows significant seasonality but no consistent trend but has upward and downward slopes during the time period.

6. EDA is performed: checking yearly as well as monthly boxplots, monthly plots showing sales across years.

7. Two types of decomposition- additive & multiplicative are performed to describe trend and seasonality factors.

8. Now the data is split into training and test data. It was suggested to split in a way that test data starts from the year 1991.

9. Now RMSE values are calculated over different models like Regression, Naïve, Simple Average, Moving Average, Simple Exponential smoothing, Double Exponential Smoothing and Triple Exponential Smoothing. The model with highest accuracy and lowest RMSE is chosen over other models.

10. Now, to build an ARIMA/SARIMA model, first, the stationarity of the dataset is checked using BoxJenkins methodology of estimating the 'p', 'q', 'P' (if Time Series has a seasonality) and 'Q' (if Time Series has a seasonality) by looking at the PACF and the ACF plots or estimate these parameters by looking at the lowest Akaike Information Criterion.

11.Also, the stationarity of training data is checked because the model is built on the training data.

12.Now the automated and manual ARIMA/SARIMA models are built and RMSE values are calculated. The best model is chosen on account of lowest RMSE and higher accuracy.

13. A table is created that shows the RMSE values in ascending order and a graph is plotted showing the model predictions.

14. The best model is chosen and RMSE is calculated for that full model.

Business Insights :

1. The dataset given show significant seasonality but a downward trend.

2. The demand had been fell out-of-favor over the years.

2. The year 1994 has accounted for the lowest mean sale, this can affect the upcoming years mean sale.

4. May 1995 is showing the lowest mean sale since 1980.

5. The model forecasts sale of 536 units of wine in 12 months into future which is an average sale of 45 units per month.

6. 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.

7. 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.

Recommendation:

1. Apart from higher sale in November and December months, Rose sales will be above average in the summer months of July and August.

2. The company can also work on flavor of Rose Wine to increase sales distribution.

3. The forecast also indicates that the year-on-year sale of wine is not showing an upward trend. The winery should investigate the low demand for Rose wine in market and make corrective actions in marketing and promotions.

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

5. The progress against sales target can be tracked.

6.Avoid overstocks and shortages of wine.

7 . Identify retailers with low reorder or high refusal rates.

8. Compare actual performance to goals and uncover sales opportunities.

9. Monitor product and geographic trends, understand how and where the products were selling