TIME SERIES FORECASTING

PGP DSBA FEB_A 2021

07/10/2021

HARSH ALKESH PANDYA

Problem:

For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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 analyses and forecast Wine Sales in the 20th century.

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

INDEX

TABLE INDEX

FIGURE INDEX

Q1.) 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.

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

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

(Tab 1. Data Table)

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

Sparkling Dataset-

```
DatetimeIndex(['1980-01-01', '1980-02-01', '1980-03-01', '1980-04-01',
               '1980-05-01', '1980-06-01', '1980-07-01', '1980-08-01',
               '1980-09-01', '1980-10-01',
               ...
               '1994-10-01', '1994-11-01', '1994-12-01', '1995-01-01',
               '1995-02-01', '1995-03-01', '1995-04-01', '1995-05-01',
               '1995-06-01', '1995-07-01'],
              dtype='datetime64[ns]', name='YearMonth', length=187, freq=None)
```

(Tab 2. Sparkling Date Data)

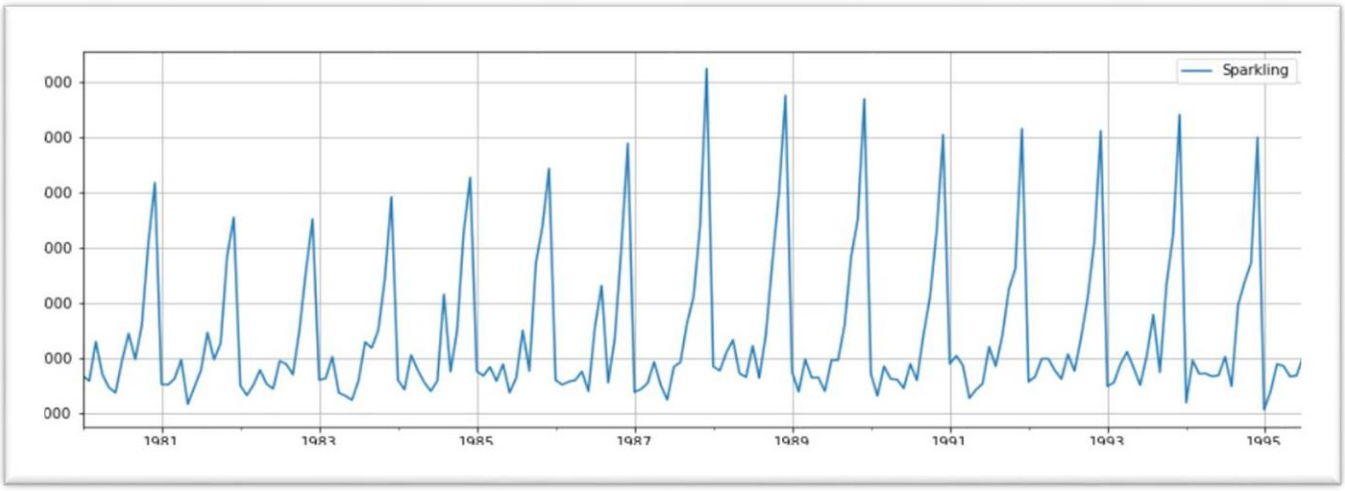
Rose Dataset-

```
atetimeIndex(['1980-01-01', '1980-02-01', '1980-03-01', '1980-04-01',
               '1980-05-01', '1980-06-01', '1980-07-01', '1980-08-01',
               '1980-09-01', '1980-10-01',
               ...
               '1994-10-01', '1994-11-01', '1994-12-01', '1995-01-01',
               '1995-02-01', '1995-03-01', '1995-04-01', '1995-05-01',
               '1995-06-01', '1995-07-01'],
              dtype='datetime64[ns]', name='YearMonth', length=187, freq=None
```

(Tab 3. Rose Date Data)

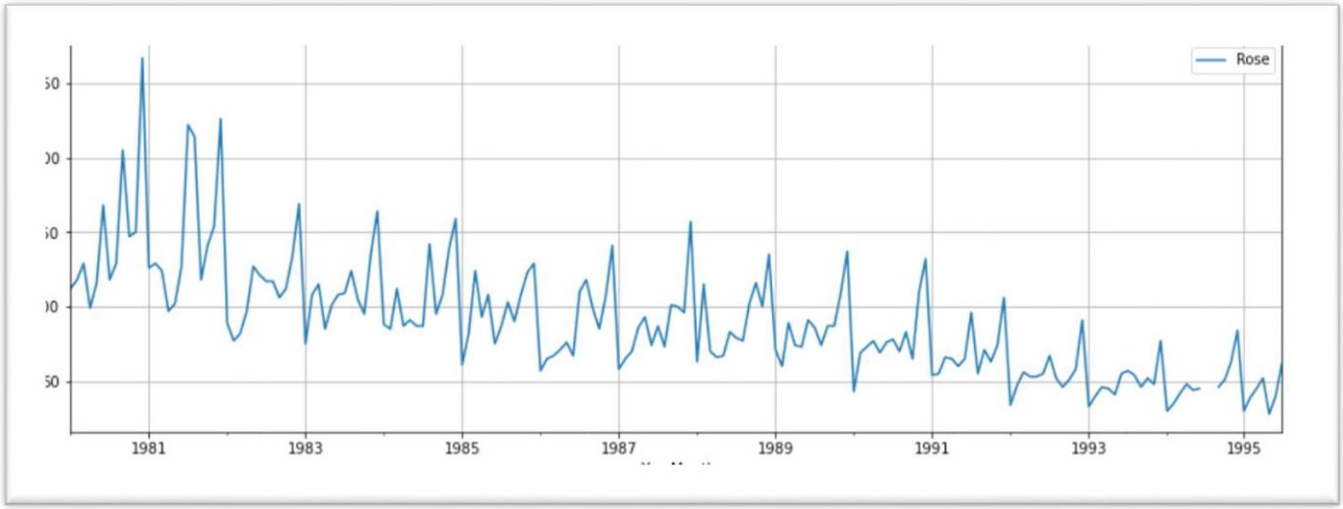
- Both the given datasets of the respective type of wines are combined to a single data frame, for the sake of comparability of the time series components and forecast.
- The Rose time-series got values missing for two months in 1994, which are imputed using interpolation (linear method)
- Rose data after interpolation for year 1994 is given below as well as the plot.
- Both the datasets show significant seasonality. While sale of Rose shows evident downward trend, Sparkling doesn't show any consistent trend but has upward and downward slopes during the period.

Sales of Sparkling Wine-



(Fig 1. Sales of Sparkling Wine-Year)

Sales of Rose Wine-



(Fig 2. Sales of Rose Wine-Year)

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

Q2) Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

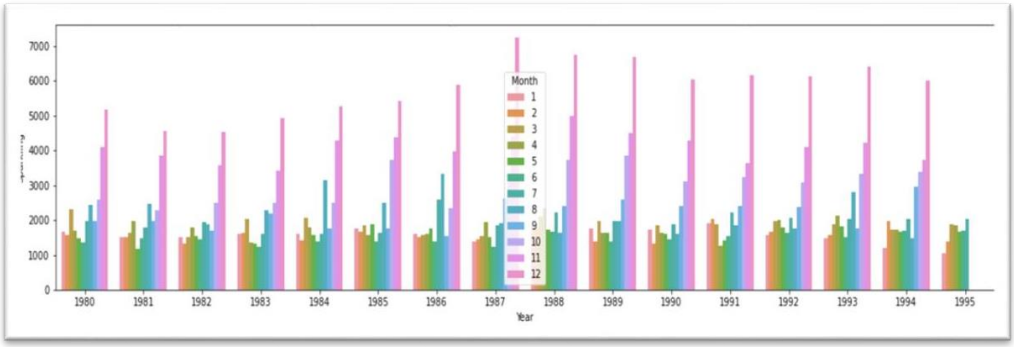
Sparkling-

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

	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000
50%	1874.000
75%	2549.000
max	7242.000

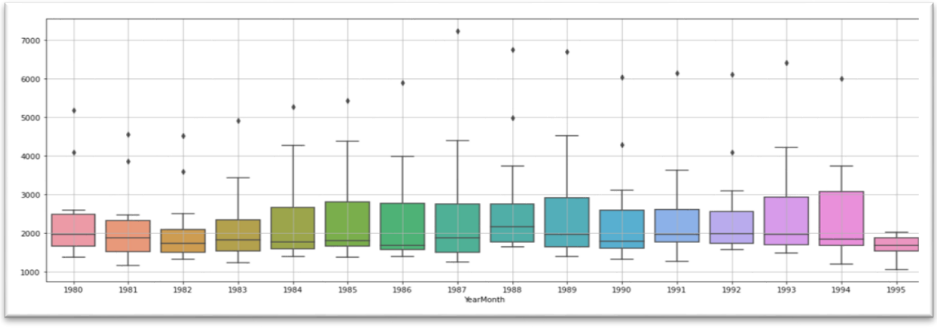
(Tab 4. Sparkling Discribe)

- 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.



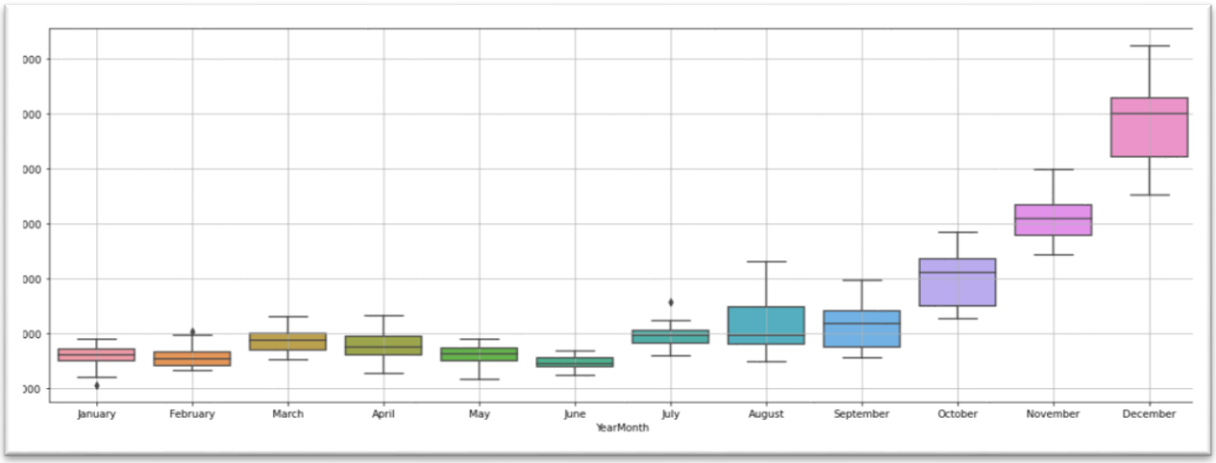
(Fig 3. Yearly bar plot)

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



(Fig 4. Yearly box plot for Sparkling)

- The monthly boxplot 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.



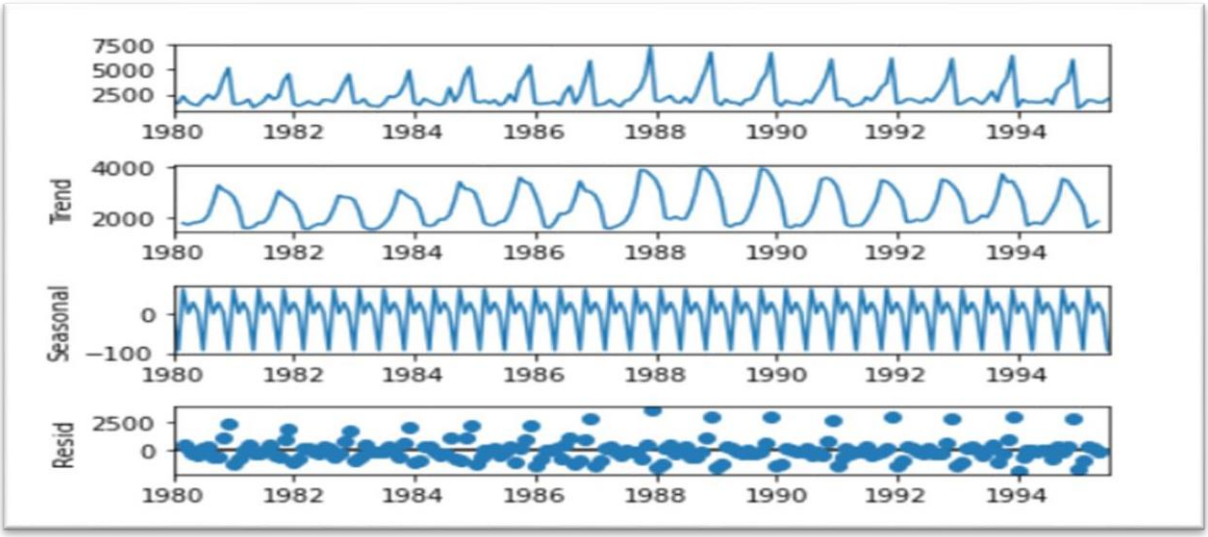
(Fig 5. Monthly box plot for Sparkling)

- The monthly plot for Sparkling shows means and variation of units sold each month over the years. Sale in seasonal months shows a higher variation than in the lean months.
- Sale in December with a mean few point below 6000, varies from 7400 to 4500 units over the years. Whereas sale in November varies from 3500 units to 5000 units and sale in October varies from 2500 to 4000 units.
- The lean months from January till September show more or less consistent are sale around 2000 units.
- The plot of monthly sale over the years also shows the seasonality component of the time-series, with October November and December selling exponentially higher volumes
- The highest volume of Sparkling wines was sold in December,1987 and the least of December sale was in 1981. Post 1987 December sales is around an average 6500 units, which was around 5000 in early 80'.

Sparkling decomposition-

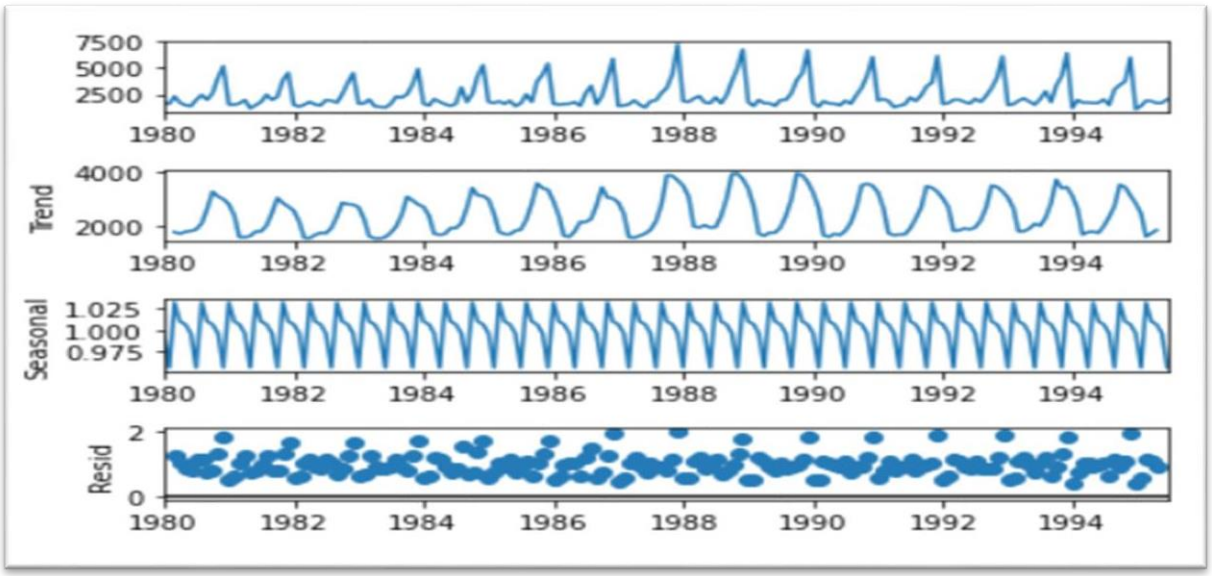
- The decomposition plots of Sparkling wine sales are given here-
- As the altitude of the seasonal peaks in the observed plot is changing according to the change in trend, the time-series is assumed to be 'multiplicative.
- The plot of the trend component does not show a consistent trend, but an intermediary 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.

Additive-



(Fig 6. Additive)

Multiplicative-



(Fig 7. Multiplicative)

- The residual shows a pattern of high variability across the period of time-series, which is 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.
- Sparkling dataset doesn't show a visible trend however it shows seasonality, also if observed from additive decomposition the residual is catching some pattern. Multiplicative decomposition on the other hand seems to dictate on the series as the scale of the residual plot had decreased considerably Monthly bar plots showed that the sales are higher towards the last months than the earlier.

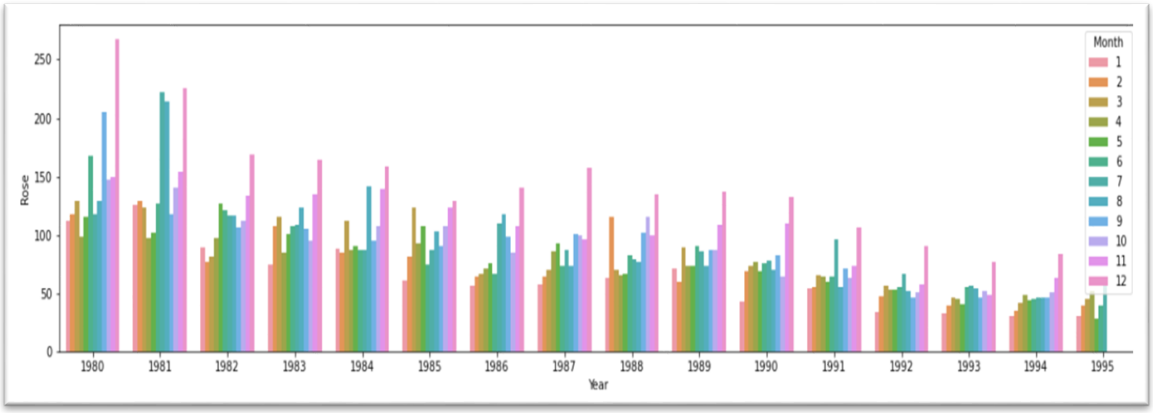
ROSE-

- 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 yearly boxplot, shows that the average sale of Rose wine moving according to the downward trend

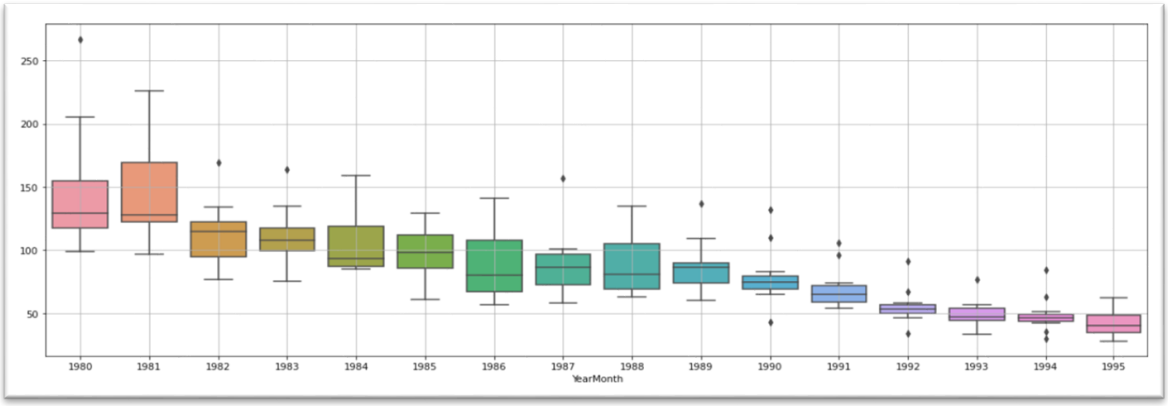
	Rose
count	185.000
mean	90.395
std	39.175
min	28.000
25%	63.000
50%	86.000
75%	112.000
max	267.000

(Tab 5. Rose Describe)

in sales over the years. The outliers over upper bound in the yearly- boxplot most probably represent the seasonal sale during the seasonal months.

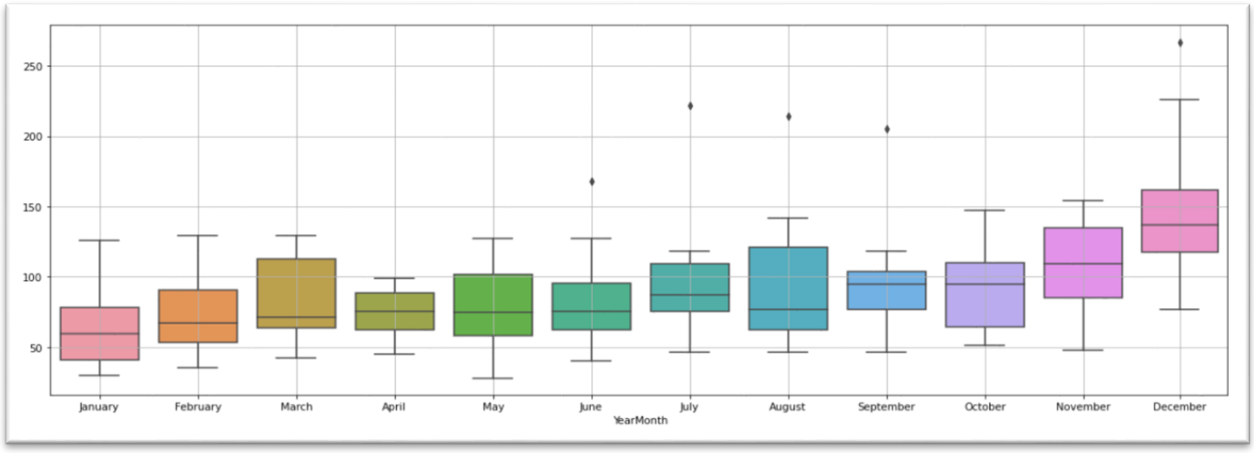


(Fig 8. Yearly bar plot for Rose)



(Fig 9. Yearly box plot for Rose)

- The monthly boxplot 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.



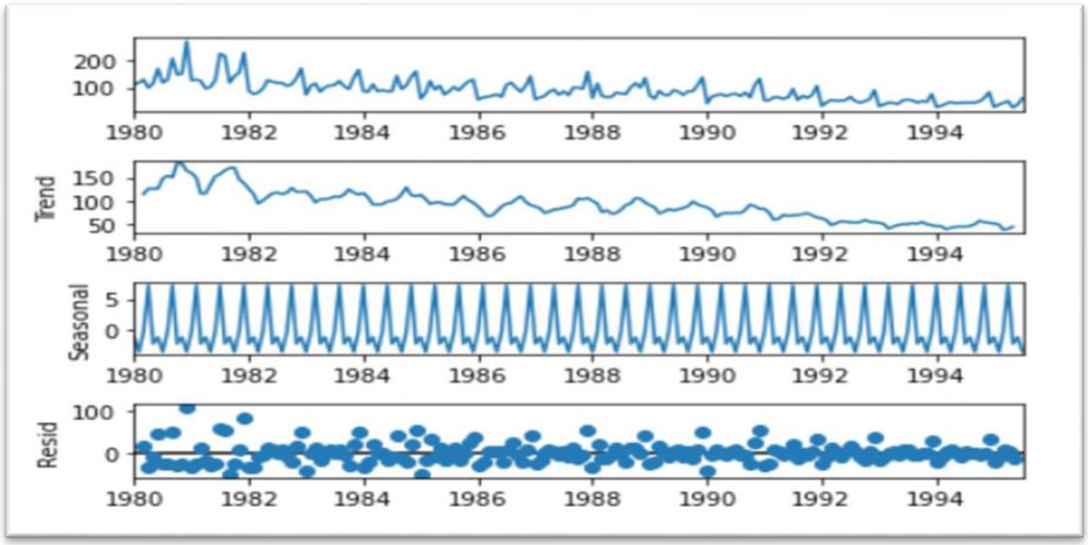
(Fig 10. Monthly box plot for Rose)

- The monthly plot for Rose shows means and variation of units sold each month over the years. Sale in months such as July, August, September, and December show a higher variation than the rest.
- Sale in December with a mean few point below 100 varies from 75 to 270 units over the years. Whereas the average sale is less than or closer to 100 units (above 50) for the rest of the year.
- The plot of monthly sale over the years also shows the seasonality component of the time-series, with November and December selling exponentially higher volumes than other months.
- The highest volume of Rose wines was sold in December 1980 and the least of December sale was in 1993. Though December sale picked after 1983, it consistently dipped after 1987.

Rose decomposition-

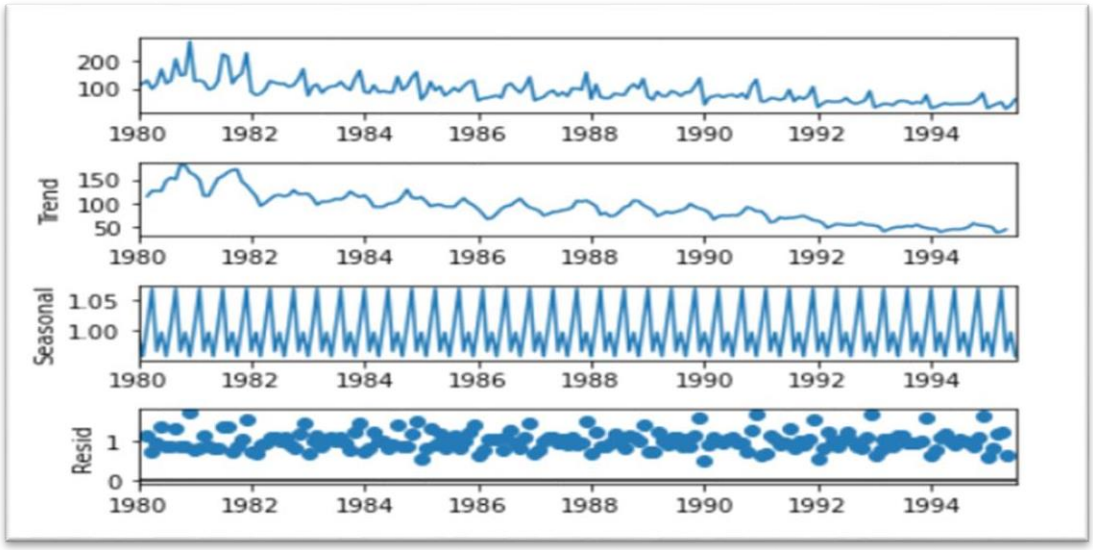
- The observed plot of the decomposition diagrams shows visible annual seasonality and a downward trend. The early period of the plot shows higher variation than in the later periods.
- The trend diagram shows a downward trend overall. Exponential dips can be seen between 1981 and 1983 and later from 1991 to 1993.
- Seasonal components are quite visible and consistent in both the observed and seasonal charts of the diagrams. The additive chart shows variance in seasonality from -20 to 50 units and the multiplicative model shows variance of 16%.

Additive-



(Fig 11. Additive for Rose)

Multiplicative-



(Fig 12. Multiplicative for Rose)

- The residuals show a pattern of high variability across the period of time-series, which is more or less consistent are 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.
- Rose datasets show a clear decreasing trend as well as seasonality, multiplicative decomposition dictates the series the the noise is reduced considerably in it also the seasonal patterns increase and decrease in the size across difference years the sales tend to go up during the July-August and also during end of the year.

Q3) 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.
- The plots of Sparkling as train and test are given here-

YearMonth	Sparkling	Year	Month
1990-08-01	1605	1990	8
1990-09-01	2424	1990	9
1990-10-01	3116	1990	10
1990-11-01	4286	1990	11
1990-12-01	6047	1990	12

Train Data: (132, 3)

(Tab 6. Sparkling Train Data)

YearMonth	Sparkling	Year	Month
1991-01-01	1902	1991	1
1991-02-01	2049	1991	2
1991-03-01	1874	1991	3
1991-04-01	1279	1991	4
1991-05-01	1432	1991	5

Test Data: (55, 3)

(Tab 7. Sparkling Test Data)

- The plots Rose time-series as train and test are given here-

YearMonth	Rose	Year	Month
1990-08-01	70.0	1990	8
1990-09-01	83.0	1990	9
1990-10-01	65.0	1990	10
1990-11-01	110.0	1990	11
1990-12-01	132.0	1990	12

Train Data: (132, 3)

(Tab 8. Rose Train Data)

YearMonth	Rose	Year	Month
1991-01-01	54.0	1991	1
1991-02-01	55.0	1991	2
1991-03-01	66.0	1991	3
1991-04-01	65.0	1991	4
1991-05-01	60.0	1991	5

Test Data: (55, 3)

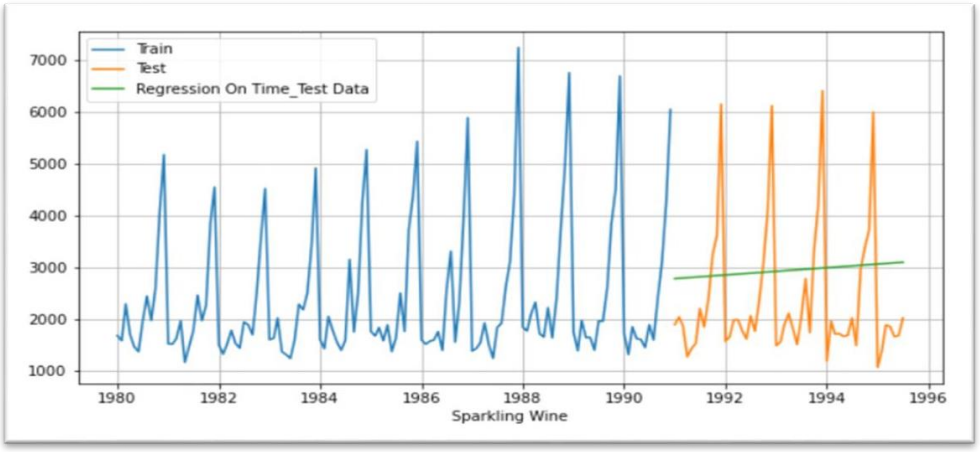
(Tab 9. Rose Test Data)

Q4) Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.

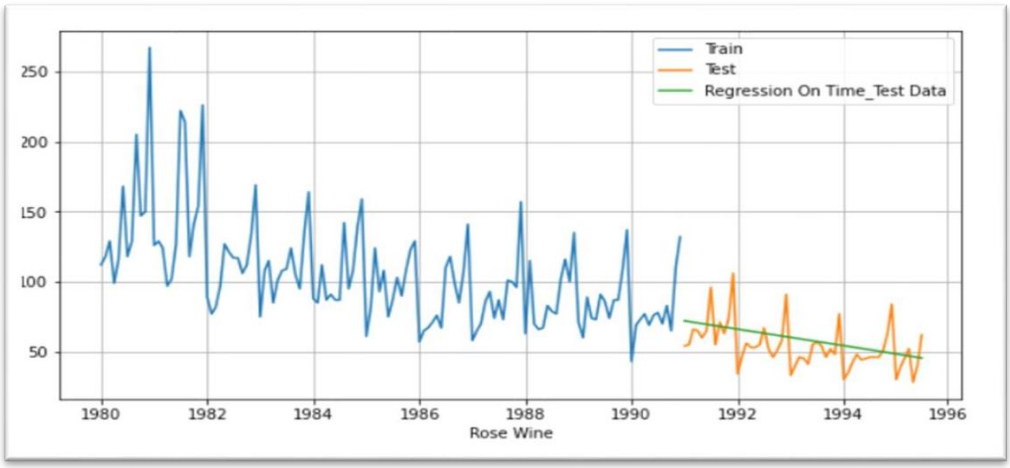
Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Model 1- LINEAR REGRESSION

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



(Fig 13. Linear Regression for Sparkling)



(Fig 14. Linear Regression for Rose)

- The RMSE values for Test data sets are different.

	Test_Spark RMSE
Regression	1389.135175

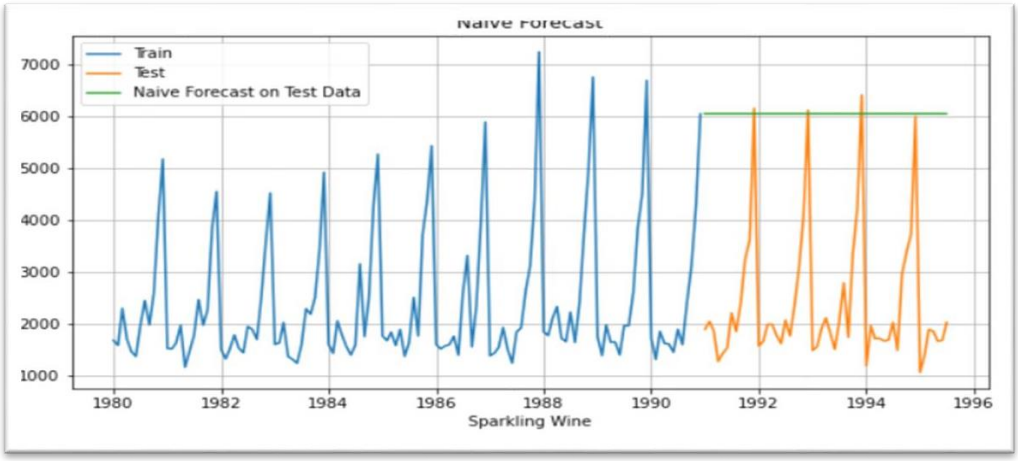
	Test_Rose RMSE
Regression	15.262509

(Tab 10. Regression for Sparkling & Rose)

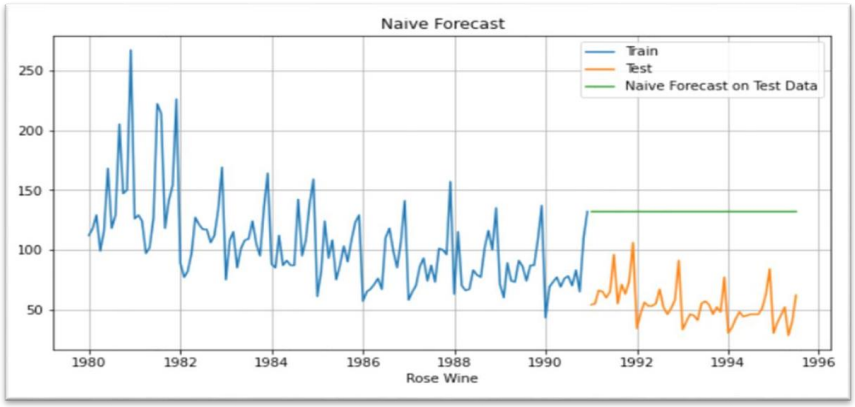
- The linear regression on the Rose dataset shows an apparent downward trend as consistent with the observed time-series.
- The model has successfully captured the trend of both the series but does not reflects the seasonality.

Model 2- NAÏVE FORECASTING

- In the Simple Average model, the forecast is done using the mean of the time-series Rose 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.



(Fig 15. Naïve Forecasting for Sparkling)



(Fig 16. Naïve Forecasting for Rose)

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

	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352

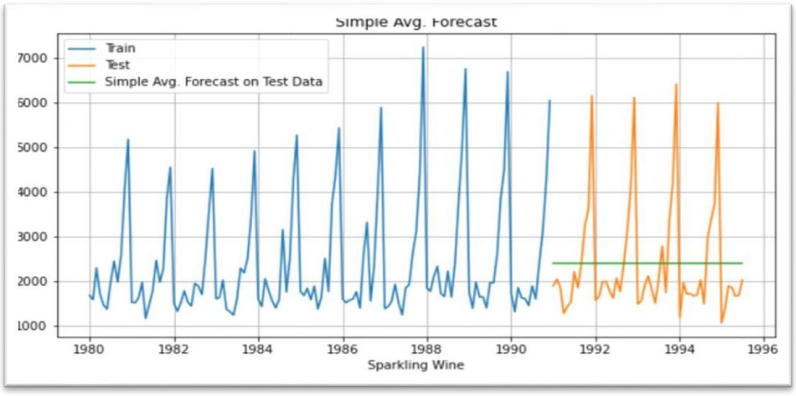
	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093

(Tab 11. Naïve test for Sparkling & Rose)

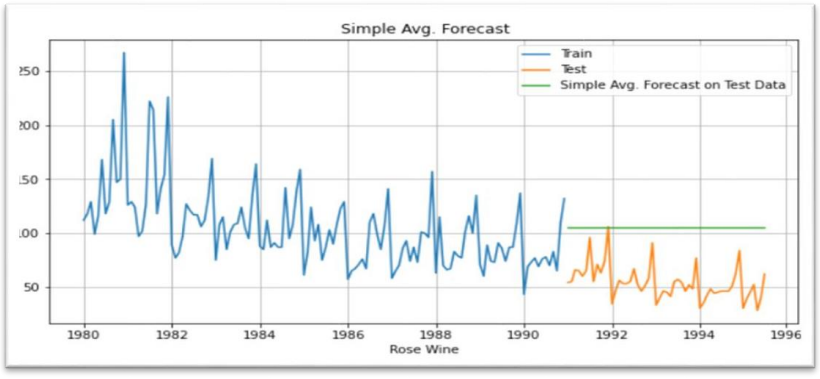
- The model is not capable of either forecasting nor able to capture the trend and seasonality present in the dataset.
- For rose the RMSE is consistent in both test and train datasets.

Model 3- SIMPLE AVERAGE FORECASTING

- For this simple average method, we will forecast by using the average of the training values.
- The model is not capable of either forecasting nor able to capture the trend and seasonality present in the dataset.



(Fig 17. Simple Average Forecasting for Sparkling)



(Fig 18. Simple Average Forecasting for Rose)

- For Rose dataset, the model forecast is almost 100% error in test data and 25% in train.

	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352
SimpleAvg	1275.081804

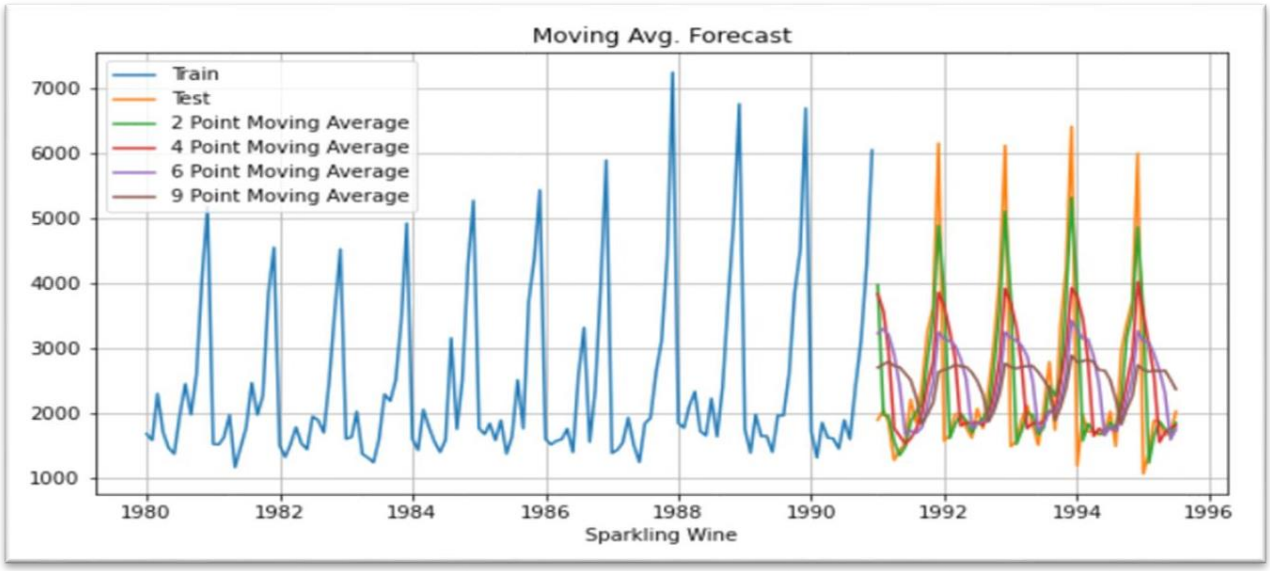
	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093
SimpleAvg	53.440426

(Tab 12. Simple Avg for Sparkling & Rose)

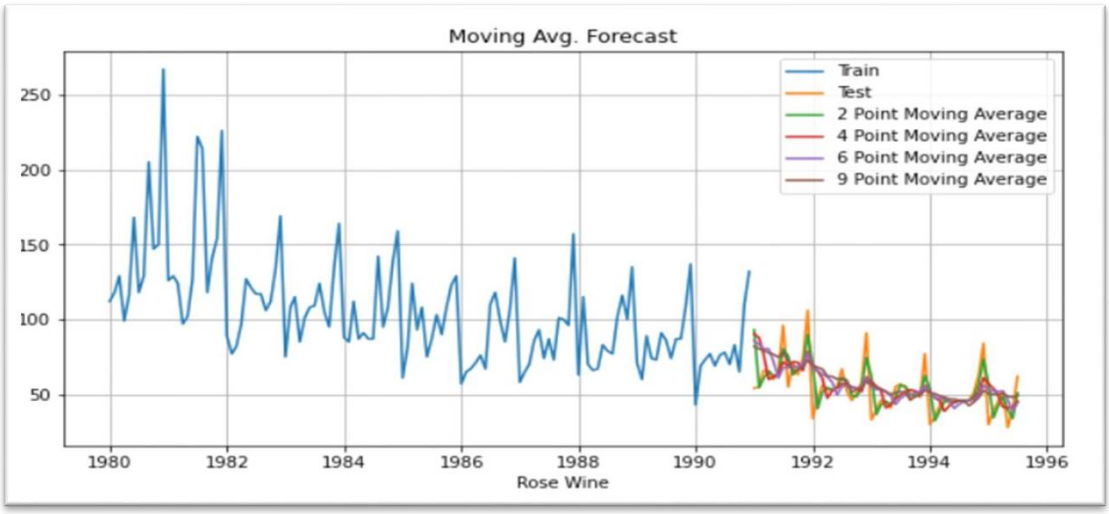
- Due to the downward trend the performance in train data set is better than the test dataset.

Model 4- MOVING AVERAGE (MA)

- 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 and Mean Absolute Percentage Error of the test set are given below.
- The best interval of moving average from the model is 2 point.
- For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the minimum error. The below plot shows the forecast for different rolling means:



(Fig 19. Moving Average Forecasting for Sparkling)



(Fig 20. Moving Average Forecasting for Rose)

- 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 and Mean Absolute Percentage Error of the test set are given below.
- The best interval of moving average from the model is 2 point.

	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352
SimpleAvg	1275.081804
MovingAvg2	813.400684
MovingAvg4	1156.589694
MovingAvg6	1283.927428
MovingAvg9	1346.278315

	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093
SimpleAvg	53.440426
MovingAvg2	11.529409
MovingAvg4	14.448930
MovingAvg6	14.560046
MovingAvg9	14.724503

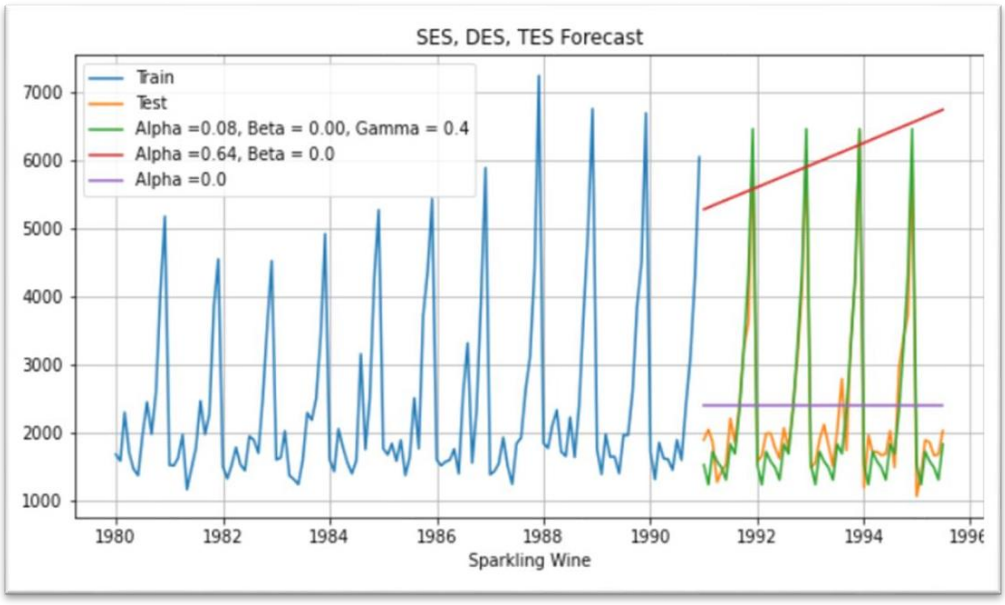
(Tab 13. Moving Average for Sparkling & Rose)

Model 4- EXPONENTIAL SMOOTHING (Simple, Double & Triple)

- Exponential smoothing methods consist of flattening time series data.
- Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous period's data with exponentially declining influence on the older observations.
- Simple Exponential Smoothing (SES): The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES).
- This method is suitable for forecasting data with no clear trend or seasonal pattern.

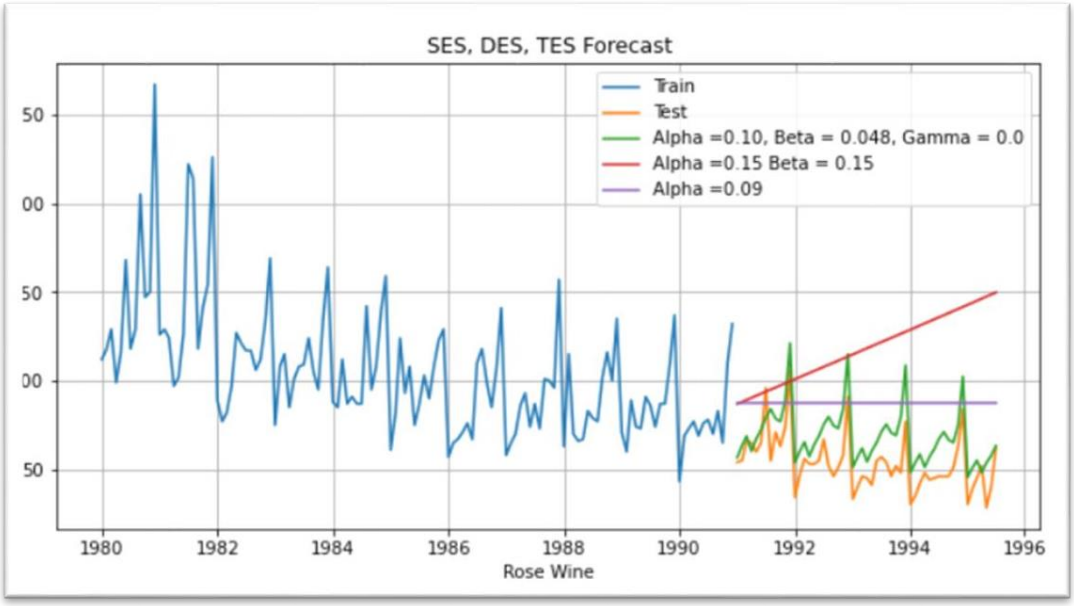
- In Single ES, the forecast at time (t + 1) is given by Winters,1960 ¦t+1=αYt+(1−α)¦t

  Parameter α is called the smoothing constant and its value lies between 0 and 1.
- Since the model uses only one smoothing constant, it is called Single Exponential Smoothing. Sparkling data doesn't show visible trend however it shows seasonality, Rose data on the other hand shows both trend and seasonality, and all the Exponential models will still be built on both the datasets.
- Double Exponential Smoothing (DES): One of the drawbacks of the simple exponential smoothing is that the model does not do well in the presence of the trend.



(Fig 21. Exponential Smoothing for Sparkling)

- Applicable when data has Trend but no seasonality.
- Two separate components are considered: Level and Trend. Level is the local mean. One smoothing parameter α corresponds to the level series A second smoothing parameter β corresponds to the trend series.



(Fig 22. Exponential Smoothing for Rose)

- Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short-term average value or level and the other for capturing the trend. Intercept or Level equation, ŷt is given by: ŷt=αyt+(1−α) ŷt Trend
  Equation is given by Tt=β(ŷt−ŷt−1) +(1−β) Tt−1 Here, α and β are the smoothing constants for level and trend, respectively, 0 < 1 and 0 < β < 1.
- The forecast at time t + 1 is given by Ft+1=ŷt+Tt Ft+n=ŷt+nTt 14
- Though our Sparkling data doesn't seem to have a visible trend we are still going to build this model for the project.

	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352
SimpleAvg	1275.081804
MovingAvg2	813.400684
MovingAvg4	1156.589694
MovingAvg6	1283.927428
MovingAvg9	1346.278315
SES	1275.081839
DES	3851.331290
TES	362.719971

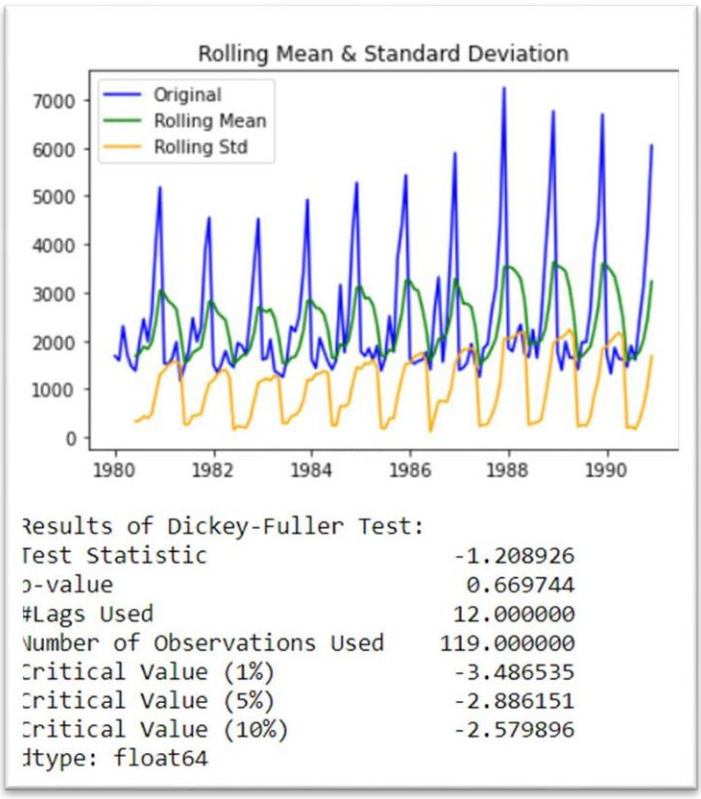
	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093
SimpleAvg	53.440426
MovingAvg2	11.529409
MovingAvg4	14.448930
MovingAvg6	14.560046
MovingAvg9	14.724503
SES	36.775787
DES	70.549148
TES	17.345537

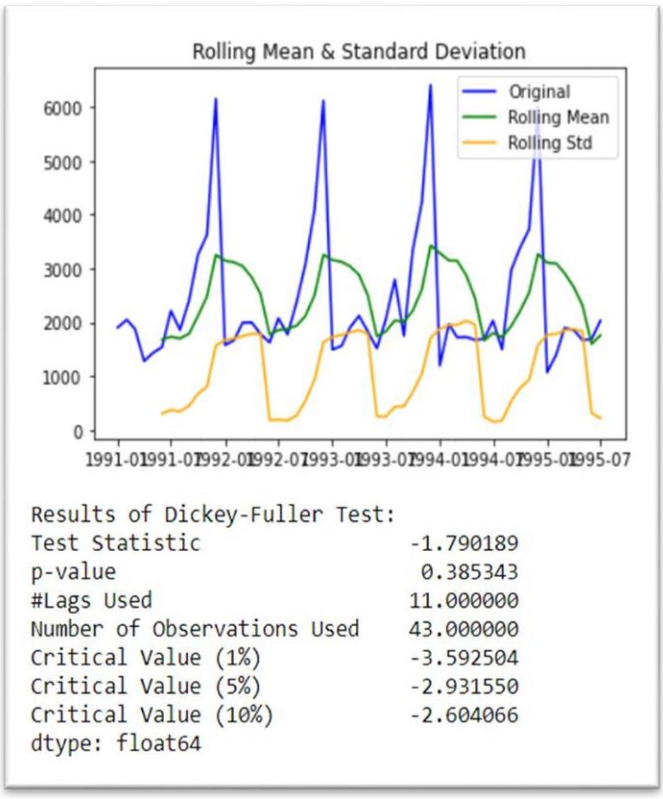
(Tab 14. Exponential Smoothing for Sparkling & Rose)

- Rose data has a clear trend from the plot above Inference Here; we see that the Double Exponential Smoothing model has picked up the trend component as well.
- Our data has seasonality too so we will include one more smoothing parameter for seasonality which is gamma.
- We will use ETS (A, A, A) Holt Winter's linear method with additive trend and seasonality for Sparkling data and ETS (A, A, M) Holt Winter's linear method with additive trend and multiplicative seasonality for Rose wine data.
- We will call it Triple Exponential Smoothing (TES).

Q5) Check for the stationary 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 stationary and comment.
Note: Stationary should be checked at alpha = 0.05.

Sparkling Train and Test set-



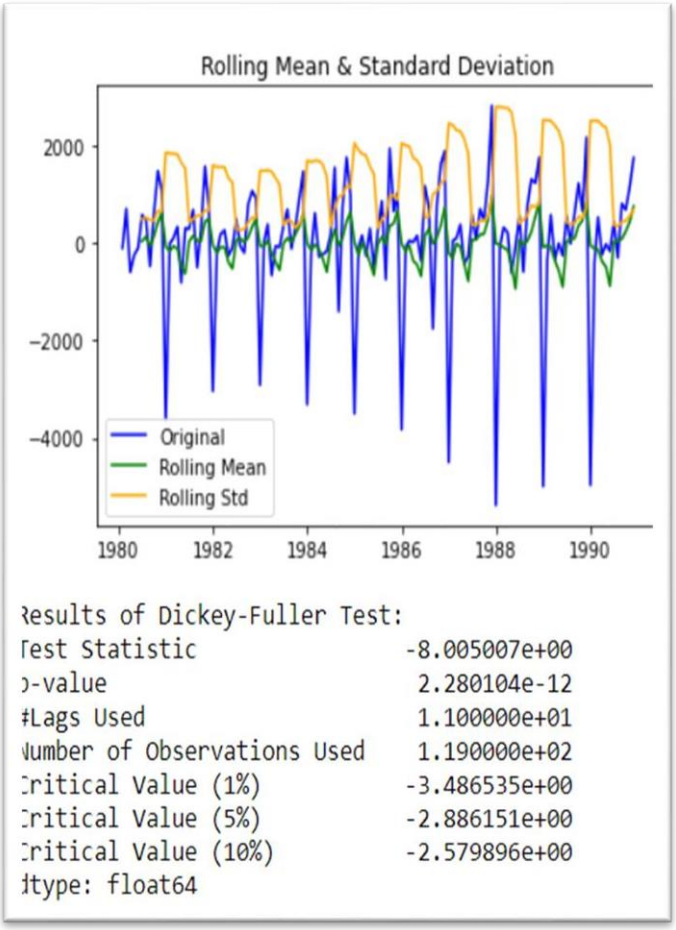


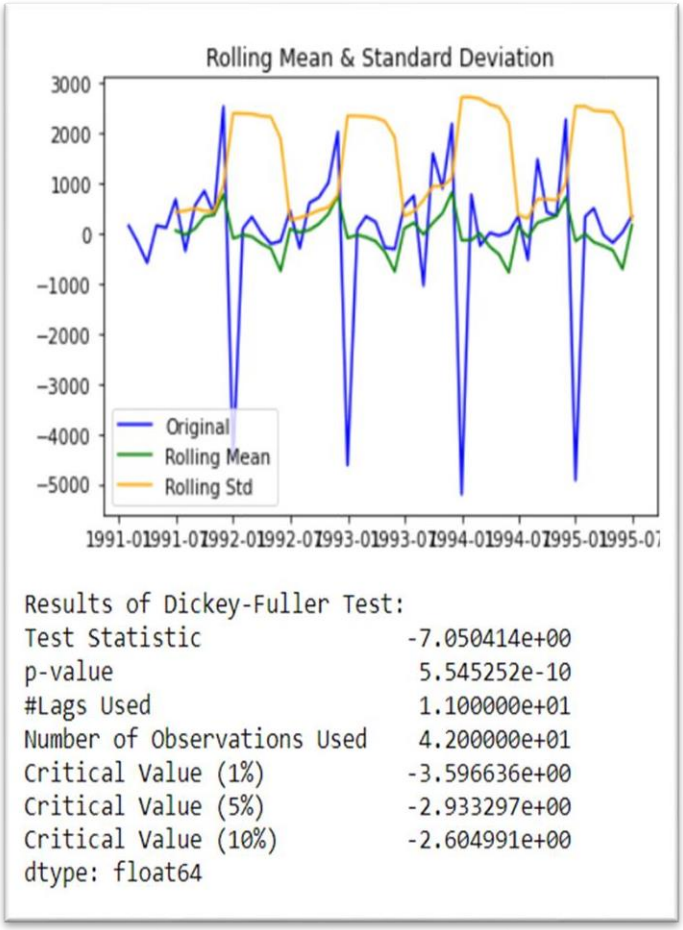
(Fig 23. Sparkling Train and Test Set)

Since then,

- Null Hypothesis H0: The series is non-stationary
- Alternate Hypothesis H1: The series is stationary
- We cannot reject the null as the p values is greater than 0.05 (significance level) from the Augmented Dickey Fuller test above for both Train and Test of Sparkling Wine dataset.
- We can correct the non-stationary by using multiple methods like taking differences at various levels, using logged transformed series etc.
- Here we will take difference of level 1 of the original series.

Differentiated Sparkling Train and Test set-





(Fig 24. Differentiated Sparkling Train and Test Set)

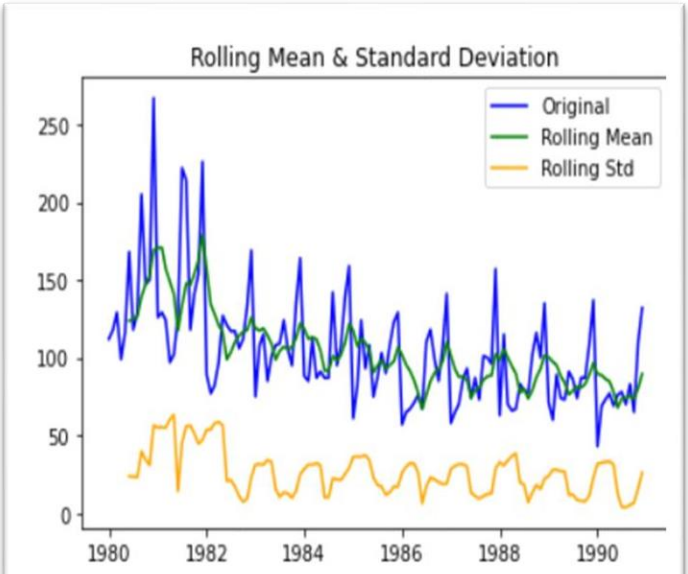
- We can now see that the p –value < than 0.05 so we can reject the null-hypothesis and accept the alternate. So, we say the series is stationary.

Rose Train and Test set-

Since then,

- Null Hypothesis H0: The series is non-stationary
- Alternate Hypothesis H1: The series is stationary
- we cannot reject the null as the p values is greater than 0.05 (significance level) from the Augmented Dickey Fuller test above Train set of Rose Wine dataset, on the contrary we can reject the null as the p values is less than 0.05 (significance level) from the Augmented Dickey Fuller test above Test set of Rose Wine dataset.

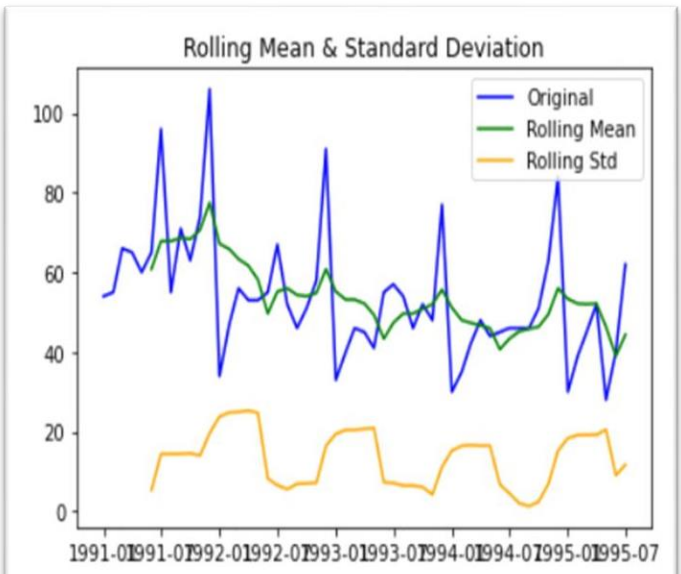
- We can correct the non-stationary by using multiple methods like taking differences at various level, using logged transformed series etc.
- Here we will take difference of level 1 of the original train series and we will use the train dataset as is.



Results of Dickey-Fuller Test:

Test Statistic	-2.164250
p-value	0.219476
#Lags Used	13.000000
Number of Observations Used	118.000000
Critical Value (1%)	-3.487022
Critical Value (5%)	-2.886363
Critical Value (10%)	-2.580009

dtype: float64



Results of Dickey-Fuller Test:

Test Statistic	-4.464772
p-value	0.000228
#Lags Used	11.000000
Number of Observations Used	43.000000
Critical Value (1%)	-3.592504
Critical Value (5%)	-2.931550
Critical Value (10%)	-2.604066

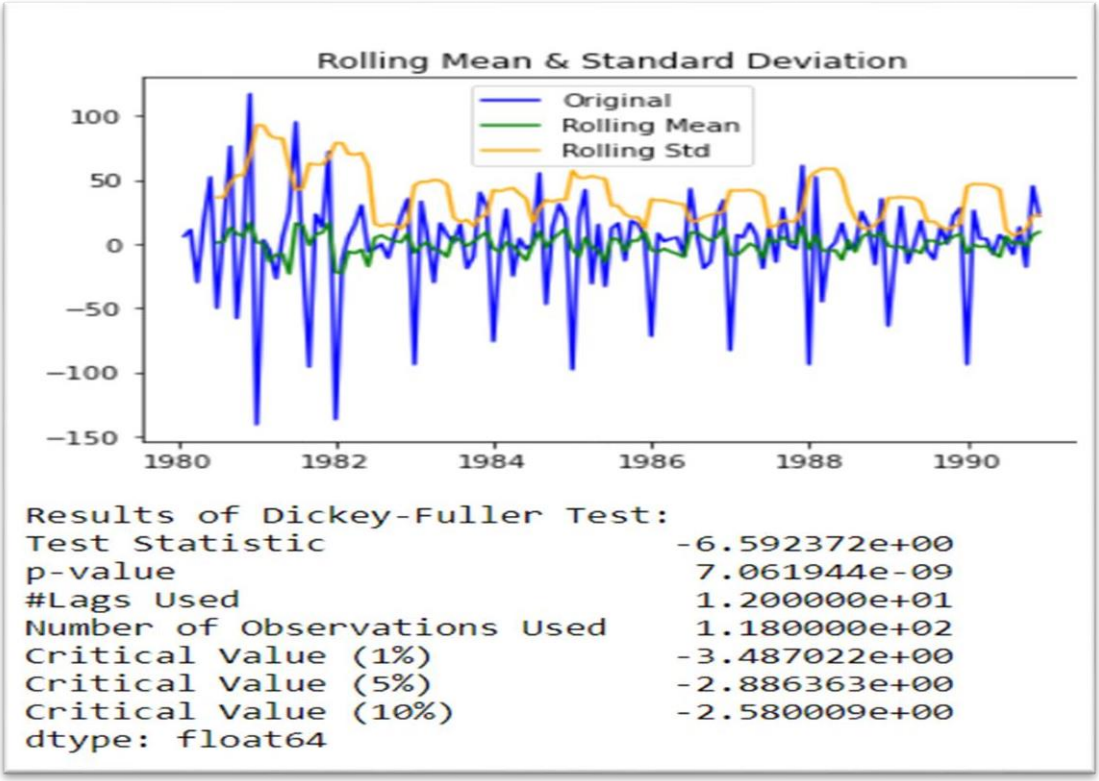
dtype: float64

(Fig 25. Rose Train and Test Set)

Differentiated Rose Train and Test set-

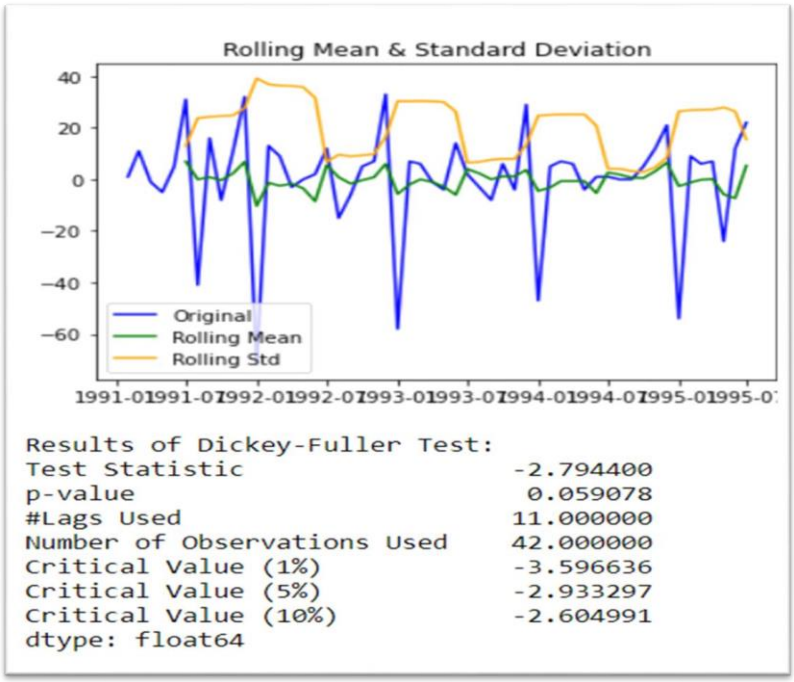
- Null Hypothesis: The series has a unit root that is series is non-stationary.
- Alternate Hypothesis: The series has no unit root that is series is stationary.
- If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that 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.

Train set-



(Fig 26. Rose Train Set)

Test set-



(Fig 27. Rose Test Set)

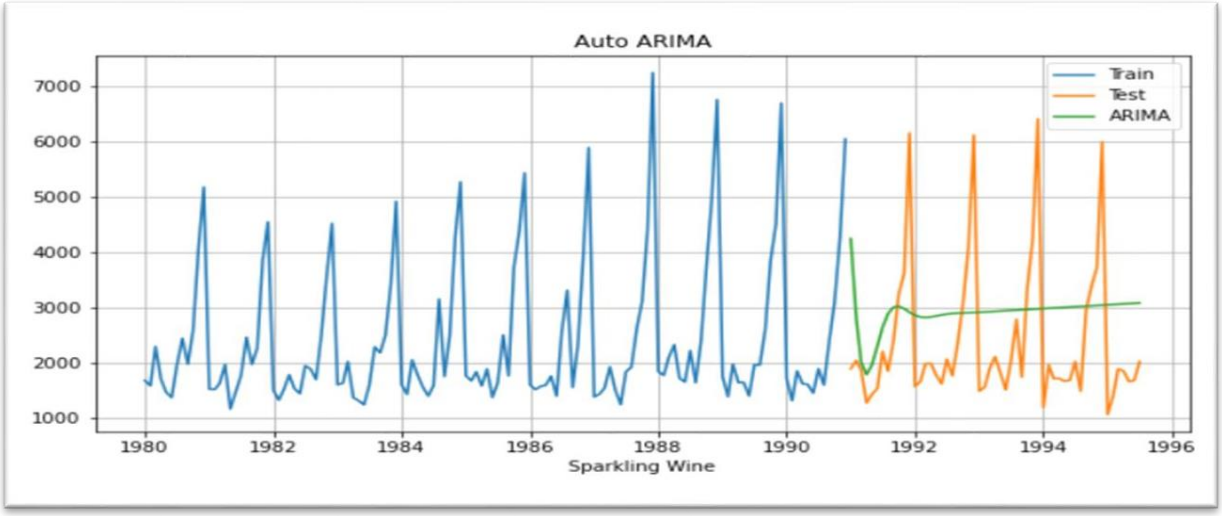
Q6) Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

- 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.

	param	AIC_Sparkling
8	(2, 1, 2)	2210.623720
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906437
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

(Tab 15. AIC Sparkling)

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



(Fig 28. AIC for Sparkling)

- 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 quintiles come from a normal distribution as the point's 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.
- From the below model summary, it can be inferred that AR (1), MA (1), MA (3), MA (2) terms has the highest absolute weightage.
- From the p-values it can be inferred that terms AR (1), AR (2), MA (1), MA (2), MA (3) and seasonal AR (1) are significant terms, as their values are below 0.05

Sparkling Data:

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.312
Method:	css-mle	S.D. of innovations	1013.589
Date:	Sun, 23 May 2021	AIC	2210.624
Time:	16:22:47	BIC	2227.875
Sample:	02-01-1980 - 12-01-1990	HQIC	2217.634

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5845	0.519	10.767	0.000	4.568	6.601
ar.L1.D.Sparkling	1.2699	0.075	17.041	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5602	0.074	-7.618	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-1.9960	0.043	-46.887	0.000	-2.079	-1.913
ma.L2.D.Sparkling	0.9960	0.043	23.336	0.000	0.912	1.080

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1334	-0.7074j	1.3361	-0.0888
AR.2	1.1334	+0.7074j	1.3361	0.0888
MA.1	1.0003	+0.0000j	1.0003	0.0000
MA.2	1.0037	+0.0000j	1.0037	0.0000

(Tab 16. ARIMA Model Results for Sparkling)

- 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 which is selected as the final model.

	param	AIC_Rose
2	(0, 1, 2)	1276.835372
5	(1, 1, 2)	1277.359224
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

(Tab 17. AIC for Rose)

- To handle multiplicity of seasonality, the data was log transformed to make it additive.
- The optimal parameters for (p, d, and q) x (P, D, Q) was selected in accordance with the lowest A kaki 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).

Rose Data:

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	130
Model:	ARIMA(0, 1, 2)	Log Likelihood	-636.749
Method:	css-mle	S.D. of innovations	30.399
Date:	Sun, 23 May 2021	AIC	1281.497
Time:	16:22:58	BIC	1292.967
Sample:	03-01-1980 - 12-01-1990	HQIC	1286.158

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0091	0.005	2.010	0.044	0.000	0.018
ma.L1.D.Rose	-1.9983	0.038	-51.949	0.000	-2.074	-1.923
ma.L2.D.Rose	0.9983	0.038	25.932	0.000	0.923	1.074

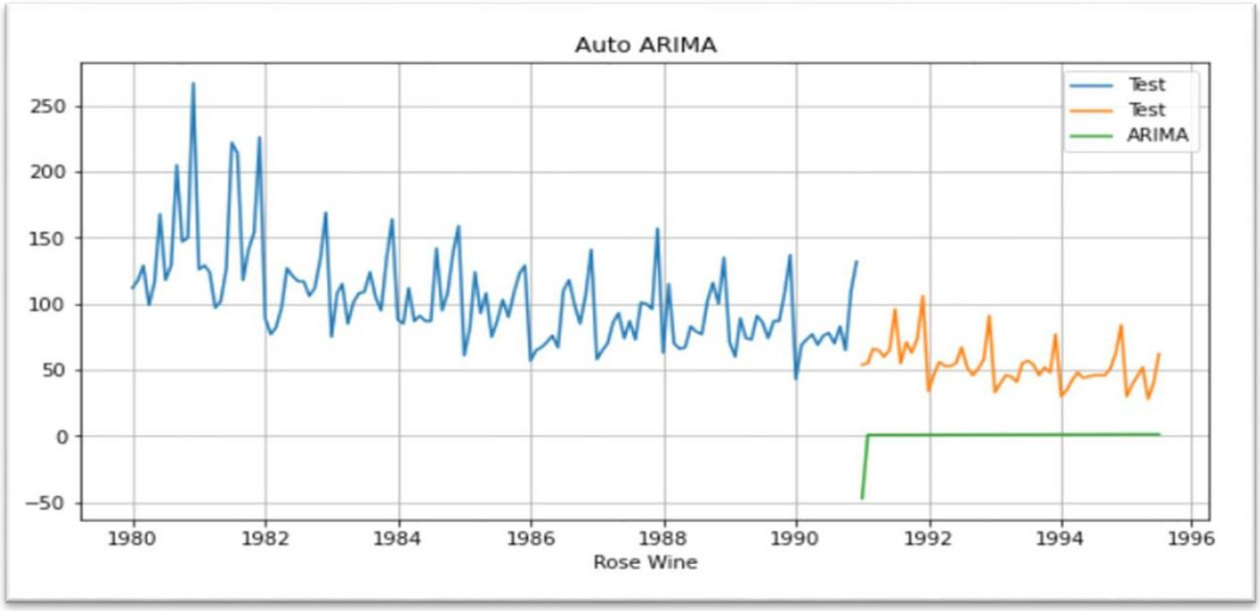
Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.0002	+0.0000j	1.0002	0.0000
MA.2	1.0015	+0.0000j	1.0015	0.0000

(Tab 18. ARIMA Model Results for Rose)

- 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 quintiles come from a normal distribution as the point's 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.



(Fig 29. ARIMA for Rose)

	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352
SimpleAvg	1275.081804
MovingAvg2	813.400684
MovingAvg4	1156.589694
MovingAvg6	1283.927428
MovingAvg9	1346.278315
SES	1275.081839
DES	3851.331290
TES	362.719971
Auto ARIMA (2,1,2)	1374.108450

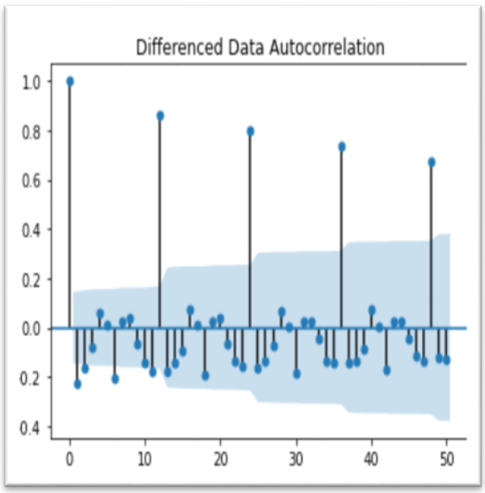
	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093
SimpleAvg	53.440426
MovingAvg2	11.529409
MovingAvg4	14.448930
MovingAvg6	14.560046
MovingAvg9	14.724503
SES	36.775787
DES	70.549148
TES	17.345537
Auto ARIMA (0,1,2)	56.292057

(Tab 19. RMSE for Sparkling & Rose)

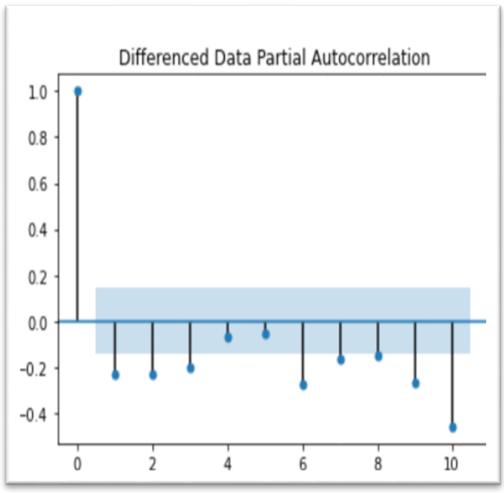
Q7) Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

- AIC for sparkling data is the lowest for the model (3, 1, 2), also we saw the from ACF and PACG plots that the cut off of p and q are at 3 and 2 resp.
- So, we conclude that the auto ARIMAX and the manual ARIMAX models are the same. ARIMA For Rose data let's build a model at the p and q cut off at 4, 2 respectively. Manual ARIMAX Summary on Rose data:
- Here, we have taken alpha=0.05.
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts- off to 0.

Sparkling -



(Fig 30. Differenced Data Autocorrelation)



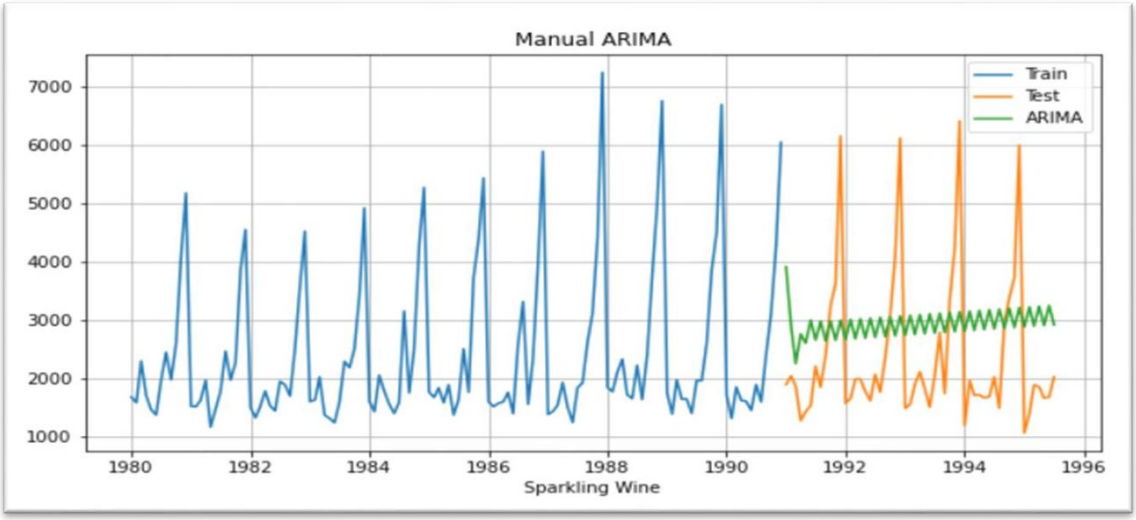
(Fig 31. Differenced Data Partial Autocorrelation)

- By looking at the above plots for Sparkling data, we can say that both the PACF cuts off at 3 and ACF plot cuts-off at lag 2.

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.9847	3.643	1.643	0.100	-1.156	13.125
ar.L1.D.Sparkling	-0.4419	1.06e-05	-4.17e+04	0.000	-0.442	-0.442
ar.L2.D.Sparkling	0.3079	3.64e-05	8457.324	0.000	0.308	0.308
ar.L3.D.Sparkling	-0.2501	3.05e-05	-8208.169	0.000	-0.250	-0.250
ma.L1.D.Sparkling	-0.0007	0.020	-0.037	0.970	-0.039	0.038
ma.L2.D.Sparkling	-0.9993	0.020	-50.921	0.000	-1.038	-0.961
		Roots				

(Tab 20. PACF for Sparkling)

- The model summary indicates that that only MA (1) term used in the model is significant in terms of p-values.
- From the multiple iterations of ARIMA models, below is the comparison of the models in terms of its accuracy attributes of RMSE and MAPE.



(Fig 32. Manual ARIMA for Sparkling)

	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352
SimpleAvg	1275.081804
MovingAvg2	813.400684
MovingAvg4	1156.589694
MovingAvg6	1283.927428
MovingAvg9	1346.278315
SES	1275.081839
DES	3851.331290
TES	362.719971
Auto ARIMA (2,1,2)	1374.108450
Manual ARIMA (3,1,2)	1379.045400

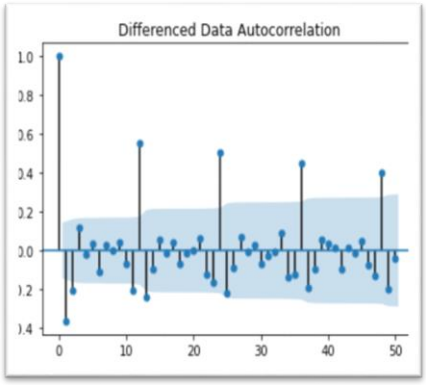
	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093
SimpleAvg	53.440426
MovingAvg2	11.529409
MovingAvg4	14.448930
MovingAvg6	14.560046
MovingAvg9	14.724503
SES	36.775787
DES	70.549148
TES	17.345537
Auto ARIMA (0,1,2)	56.292057
Manual ARIMA (4,1,2)	33.930300

(Tab 21. Manual ARIMA Test for Sparkling & Rose)

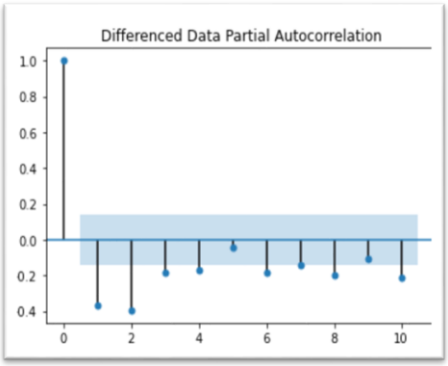
- Here we have taken alpha = 0.05 and seasonal period as 12.
- From the PACF plot it can be seen that till lag 4 is significant before cut-off, so AR term 'p = 4' is chosen. At seasonal lag of 12, it cuts off, so keep seasonal AR 'P = 0'.
- From ACF plot, lag 1 and 2 are significant before it cuts off, so let's keep MA term 'q = 1'
  And at seasonal lag of 12, a significant lag is apparent, so let's keep 'Q = 1'

- By looking at the below plots for Rose data, we can say that PACF cuts off at 4 and ACF plot cuts-off at lag 2.

Rose-



(Fig 33. Autocorrelation for Rose)



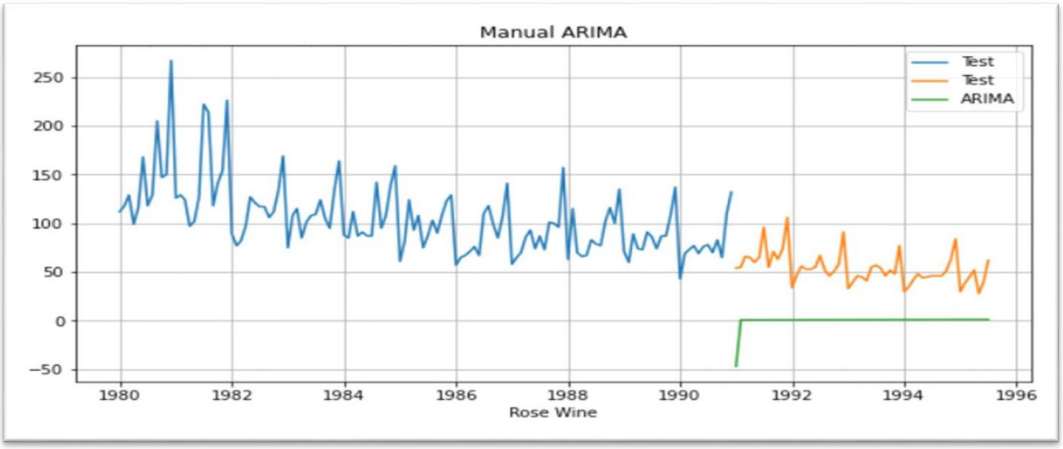
(Fig 34. Partial Autocorrelation for Rose)

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

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1905	0.576	-0.331	0.741	-1.319	0.938
ar.L1.D.Rose	1.1685	0.087	13.391	0.000	0.997	1.340
ar.L2.D.Rose	-0.3562	0.132	-2.693	0.007	-0.616	-0.097
ar.L3.D.Rose	0.1855	0.132	1.402	0.161	-0.074	0.445
ar.L4.D.Rose	-0.2227	0.091	-2.443	0.015	-0.401	-0.044
ma.L1.D.Rose	-1.9506	nan	nan	nan	nan	nan
ma.L2.D.Rose	1.0000	nan	nan	nan	nan	nan

Roots

(Tab 22. Residuals for Rose)



(Fig 35. Manual ARIMA for Rose)

Q8) Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

- 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 for sparkling Wine.
- Triple Exponential Smoothing is found to be the best model, followed by 2 point Moving Average for Rose wine.

	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352
SimpleAvg	1275.081804
MovingAvg2	813.400684
MovingAvg4	1156.589694
MovingAvg6	1283.927428
MovingAvg9	1346.278315
SES	1275.081839
DES	3851.331290
TES	362.719971
Auto ARIMA (2,1,2)	1374.108450
Manual ARIMA (3,1,2)	1379.045400

	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093
SimpleAvg	53.440426
MovingAvg2	11.529409
MovingAvg4	14.448930
MovingAvg6	14.560046
MovingAvg9	14.724503
SES	36.775787
DES	70.549148
TES	17.345537
Auto ARIMA (0,1,2)	56.292057
Manual ARIMA (4,1,2)	33.930300

(Tab 23. RMSE for Sparkling & Rose)

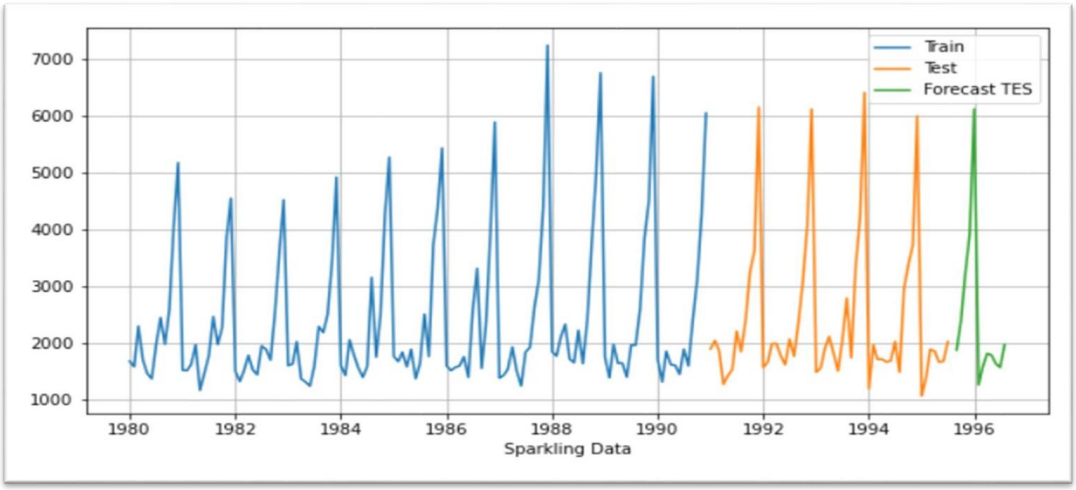
Q9.) Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

- For Sparkling dataset, we see that Triple Exponential smoothing gives the best forecast, so we will move forward with that for forecasting.

Time	Sparkling Forecast	lower CI	upper CI
1995-08-31	1884.976769	1098.923918	2671.029620
1995-09-30	2402.258496	1616.205645	3188.311348
1995-10-31	3245.977232	2459.924381	4032.030084
1995-11-30	3932.213204	3146.160352	4718.266055
1995-12-31	6119.724082	5333.671230	6905.776933
1996-01-31	1266.116913	480.064062	2052.169764
1996-02-29	1583.646638	797.593787	2369.699490
1996-03-31	1821.829048	1035.776197	2607.881900
1996-04-30	1795.729426	1009.676575	2581.782277
1996-05-31	1643.054809	857.001958	2429.107661
1996-06-30	1576.941975	790.889124	2362.994826
1996-07-31	1975.093831	1189.040980	2761.146683

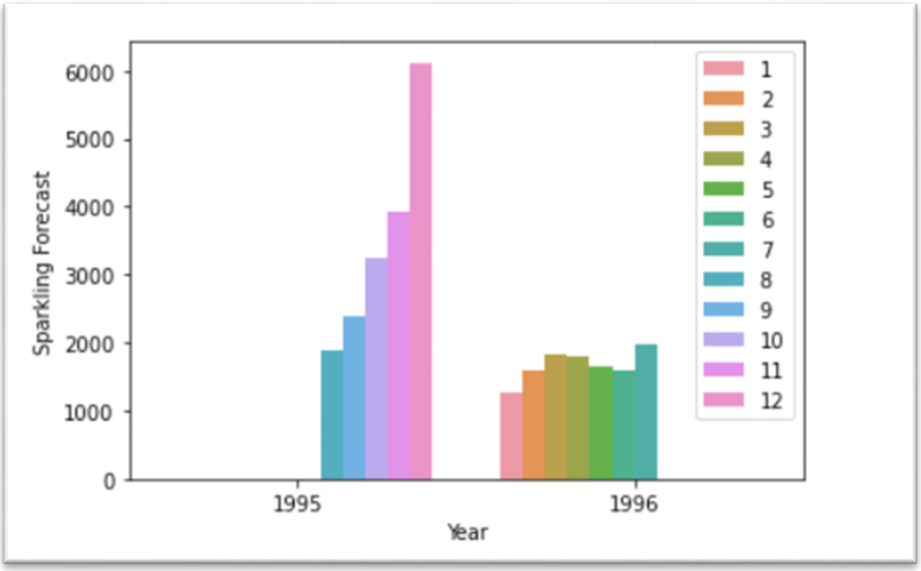
(Tab 24. Sparkling Data Set)

SPARKLING Forecast TES on train and test dataset-



(Fig 36. Sparkling Forecast Train & Test Date Set)

SPARKLING WINE FORECAST-



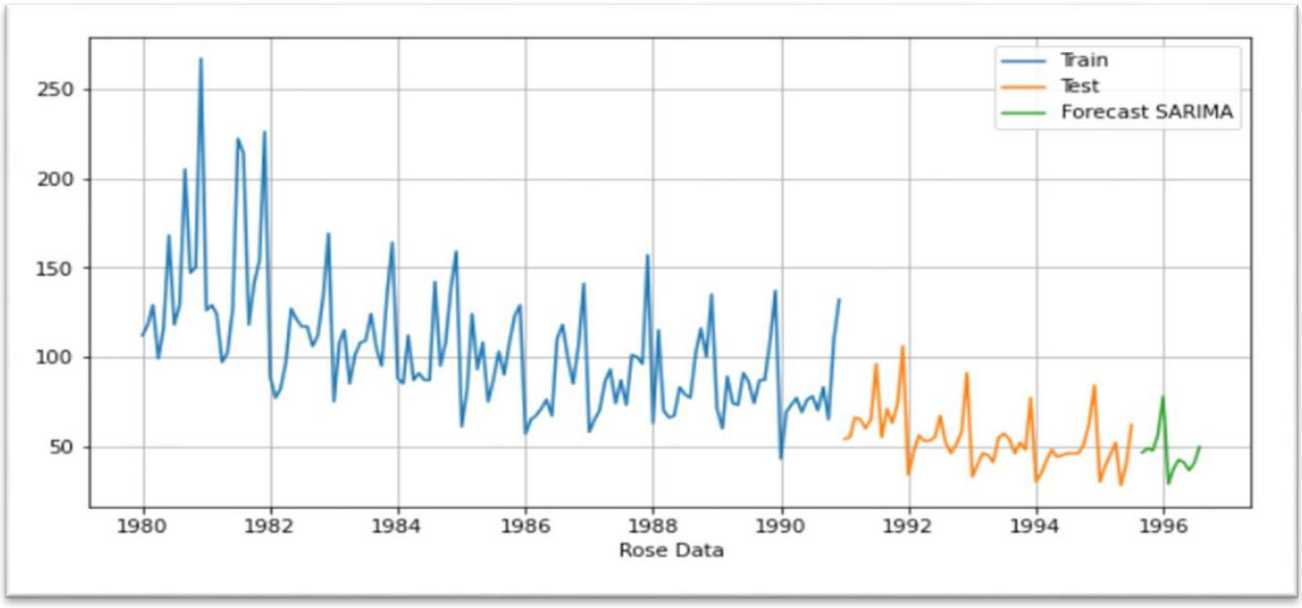
(Fig 37. Sparkling Wine Forecasting)

- For Rose dataset rolling avg. shows the best RMSE, however since the window chosen was very small (2, 4, 6, 9) it was natural it was going to work well on Test set.
- The other model which gave the best RMSE was TES and Manual SARIMAX (4, 1, 2) (3, 0, 2, 12). We will build a final model on the entire Rose dataset using SARIMAX

Time	y	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31		46.412203	11.969608	22.952203	69.872203
1995-09-30		48.790729	12.040279	25.192215	72.389242
1995-10-31		47.507883	12.108770	23.775131	71.240635
1995-11-30		56.268161	12.121389	32.510675	80.025648
1995-12-31		77.865945	12.121866	54.107525	101.624365
1996-01-31		28.705557	12.214734	4.765119	52.645995
1996-02-29		37.188429	12.374990	12.933894	61.442964
1996-03-31		42.400073	12.561983	17.779038	67.021107
1996-04-30		40.940984	12.728262	15.994049	65.887919
1996-05-31		36.439120	12.841667	11.269915	61.608326
1996-06-30		40.435346	12.933318	15.086509	65.784183
1996-07-31		49.549961	13.021629	24.028037	75.071886

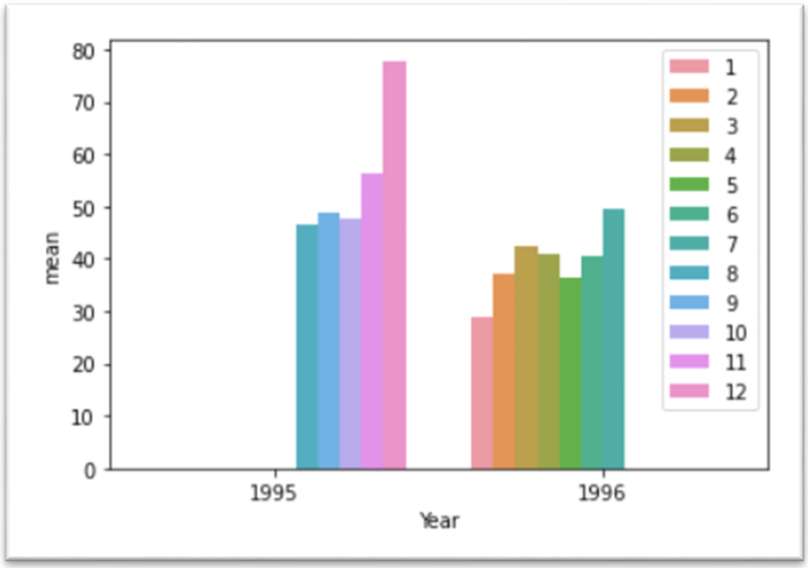
(Tab 25. Manual SARIMAX for Rose)

ROSE Forecast SARIMA on train and test dataset-



(Fig 38. Rose SARIMA Train & Test Data Set)

ROSE WINE FORECAST-



(Fig 39. Rose Wine Forecasting)

Q10) Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

SPARKLING-

Findings:

- 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: 0.2 and gamma: 0.2 & trend: 'additive', seasonal: 'Multiplicative' is found to be the best model in terms of accuracy scored against the full Data.
- The model predicts continuation of the trend in sales and seasonality in yearend 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 (4, 1, 1) x (0, 1, 1, 12), is found to be the most optimal SARIMA model for the complete time-series

Suggestion:

- TES (Triple Exponential Smoothing) has worked the best for the forecast with lowest RMSE on test data 30
- You can see from the above chart that the forecast for next 12 months is slightly over the sales of the previous 12 months however, there isn't a considerable increase.
- Observed from the month wise bar plots previously, we can say that the sales of Sparkling wine tend to go up in last two months probably because it's a holiday season than the rest and its lowest around Jun and July
- ABC can take various measures to increase the sales towards the beginning and mid of the year, it can introduce promotional activities or discounts during the low sales period.
- ABC can tie up with events like concerts; weddings etc. and do some sponsorship to boost sales during the slack.
- 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.

ROSE-

Findings:

- The SARIMA model is chosen as the final model for prediction on Rose dataset, as it provides confidence interval and better explain ability 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 indicate 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.

Suggestion:

- We chose manual SARIMAX model to predict for the Rose wine data. The model was passed the cut offs found through ACF and PACF plots of q and p respectively and seasonality of 12 as the plots showed a patterned significance after 11 lags.
- You can see from the above plot for Rose wine data the forecast for 1996 is more or less same as of for 1995.
- Observed from the monthly bar plot sales shows an increasing trend from August

  Towards December, it's on the lower side beginning of the year.
- ABC can take sought promotional activities and implement some discounts during the first half of the year.
- Unlike Sparkling wine, Rose wine sells very low number of units and the standard deviation is only 14.5. Which means that higher demand does not impact procurement and production
- Apart from higher sale in November and December months, Rose sales will be above average in the summer months of July and August.
- The winery should investigate the low demand for Rose wine in market and make corrective actions in marketing and promotions.

“Wine improves with age. The older I get, the better I like it.”*

Thank You!

***Alcohol content is injurious to health.**
Daily consumption of alcohol in large amounts leads to cardiac problems, viz., arrhythmia (irregular heartbeats), cardiomyopathy (enlarged heart), etc. Heavy alcohol consumption can raise the blood pressure, leading to alcohol-induced hypertension*