TIME SERIES FORECASTING

ON

DIFFERENT WINES OF ABC ESTATE WINES

https://www.foodnetwork.com/healthyeats/holidays/2016/12/which-sparkling-wines-are-

Name- Debsmita Chakraborty

Batch-July 'C"

TABLE OF CONTENTS

Graphs and Plots:

SUMMARY

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, we are tasked to analyse and forecast Wine Sales in the 20th century.

Data being of Sparkling wine and Rose wine.

INTRODUCTION

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304

	YearMonth	Sparkling
184	1995-05	1670
185	1995-06	1688
186	1995-07	2031

Fig a) Head and Tail-Sparkling

The data for Sparkling has:

- Data ranges from January 1980-July 1995. There are 187 rows in the data and 2 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   YearMonth   187 non-null    object
 1   Sparkling   187 non-null    int64
dtypes: int64(1), object(1)
memory usage: 3.0+ KB
```

```
(187, 2)
```

- The data has one object type column which is the Year Month and the other is the Sales column which is int64. The memory occupied by the data set is 3 KB. The data in 0-2 represents the head and the 184-186 is the tail of the data.
- No null values in the data:

```
YearMonth    0
Sparkling    0
dtype: int64
```

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

	YearMonth	Rose
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

Fig a) Head and Tail-Rose

The data for Rose has:

- Data ranges from January 1980-July 1995. There are 187 rows in the data and 2 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   YearMonth   187 non-null    object
 1   Rose        185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.0+ KB
```

```
(187, 2)
```

- The data has one object type column which is the Year Month and the other is the Sales column which is int64. The memory occupied by the data set is 3 KB.
  The data in 0-2 represents the head and the 184-186 is the tail of the data.
- There are two null values as below in the data for years mentioned:

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

YearMo	Rose
1994-07	NA
1994-08	NA

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

- The data has two wines namely Sparkling and Rose as mentioned in the Introduction.
- Sparkling as shown above in fig a has no missing data but Rose has.
- This can also mean that Sparkling has been preferred throughout years but Rose was not a preference in July-August 1994.
- Let's look at the data individually:

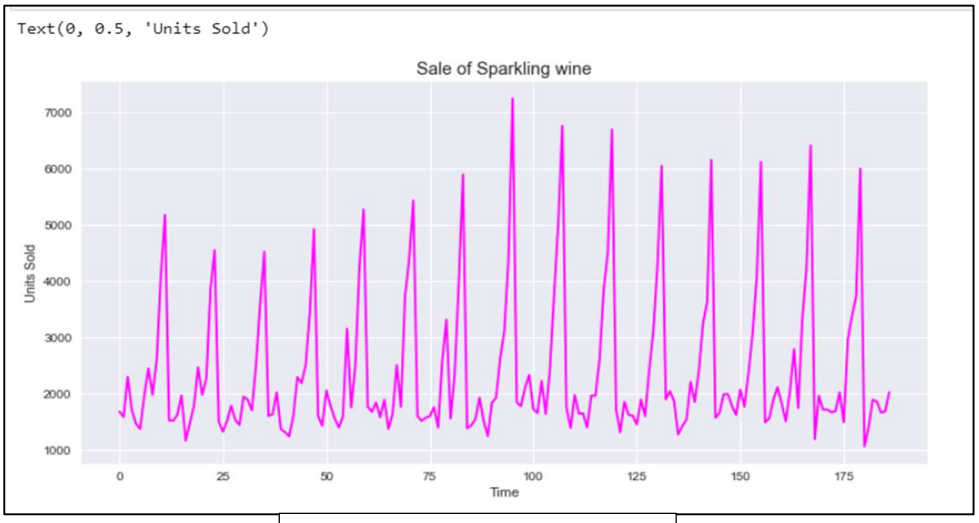


Fig 1.a Trend Seasonality - Sparkling

The data shows a trend though not consistent with significant seasonality as well. The trend is on the lower side initially and one's a peak in of sales of 7000 post which its sales decline but not much.

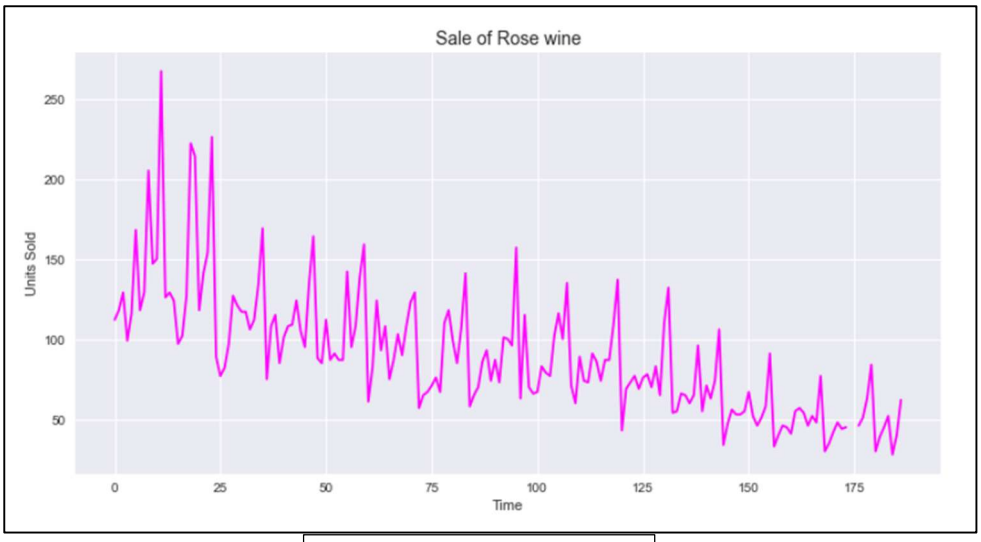


Fig 1.b Trend Seasonality - Rose

The data shows a declining trend with significant seasonality as well. The trend is on the lower side throughout. The sales are declining from 267 to 28 continuously.

- **The data for Rose has missing data so it has been imputed with Linear Method, the data being imputed in daily wise mean to get more accurate Prediction since we have monthly data of all years.**

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

- We will first see the exploratory Data Analysis when Linear Method and combining two models were not done.

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

	count	mean	std	min	25%	50%	75%	max
Rose	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

For Sparkling we see the minimum and maximum has a lot of variances hence it can be interpreted to have outliers. The Description for Rose also has a lot of variances between minimum and maximum values which means a lot of outliers will exist.

- The outliers are as below:

Sparkling: **The yearly plot for outliers:**

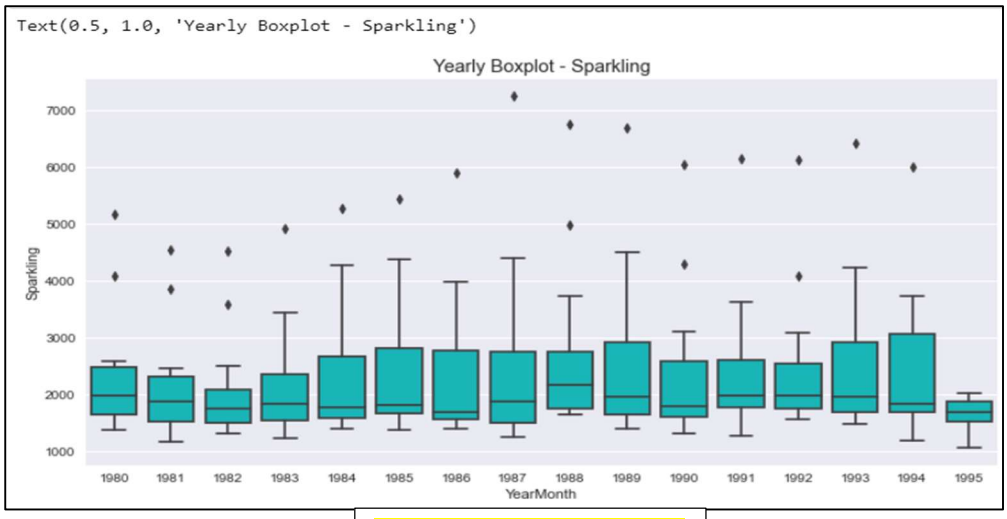


Fig 2.a Yearly box plot-Sparkling

INTERPRETATION-

1. The data shows an outlier in all the years is present.
2. The highest outlier is in the year 1987.
3. No outlier can be seen in 1995.

ROSE: **The yearly plot for outliers:**

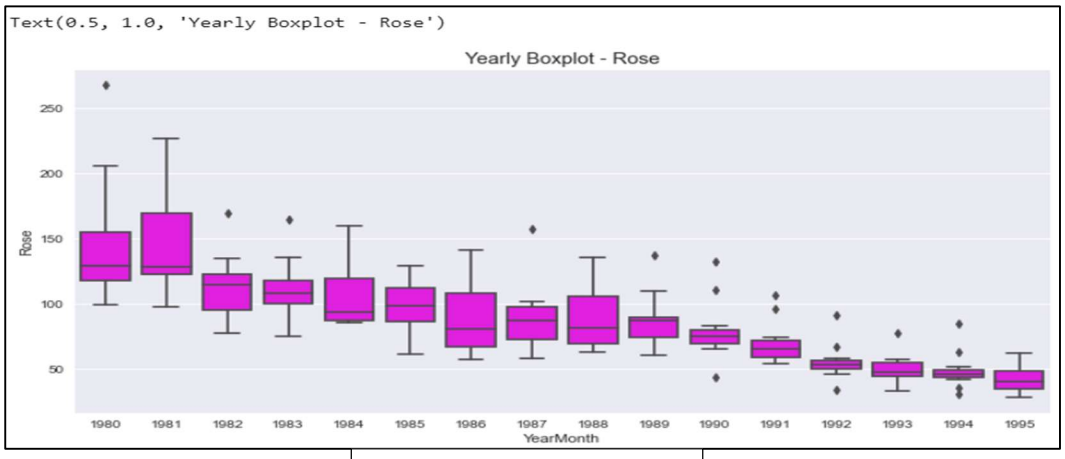


Fig 2.a Yearly box plot-Rose

INTERPRETATION-

1. The data shows an outlier in some of the years is present.
2. The highest outlier is in the year 1980.
3. No outlier can be seen in 1981, 1984,1985,1986,1988 and 1995.

Sparkling: **The Monthly plot for outliers:**

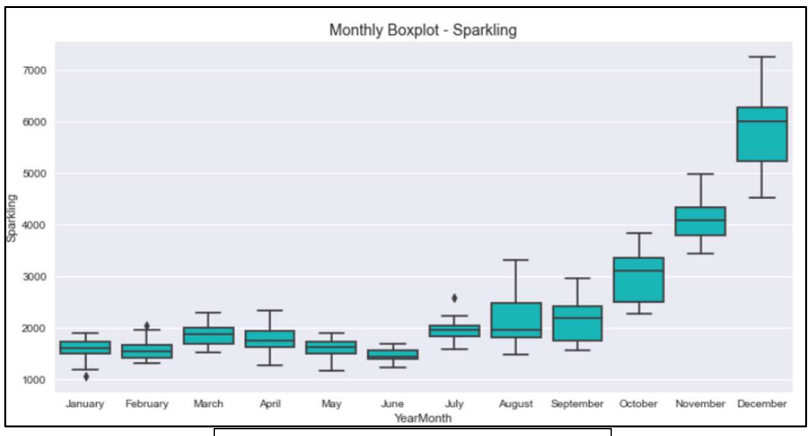


Fig 2.b Monthly box plot-Sparkling

INTERPRETATION-

1. We can see that we have a downward outlier in the month of January. February and July seem to have one outlier. Other months have no outliers.

Rose: **The Monthly plot for outliers:**

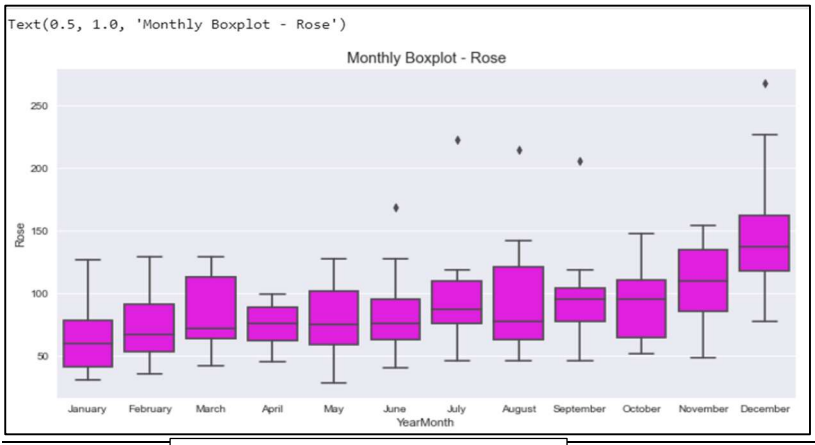


Fig 2.b Monthly box plot-Rose

INTERPRETATION-

1. We can see that we have an outlier in the months ranging from June to September and in December. Other months have no outliers.
2. The month of December has comparatively more sales.

- The distribution of Sparkling wine and Rose wine is as below:

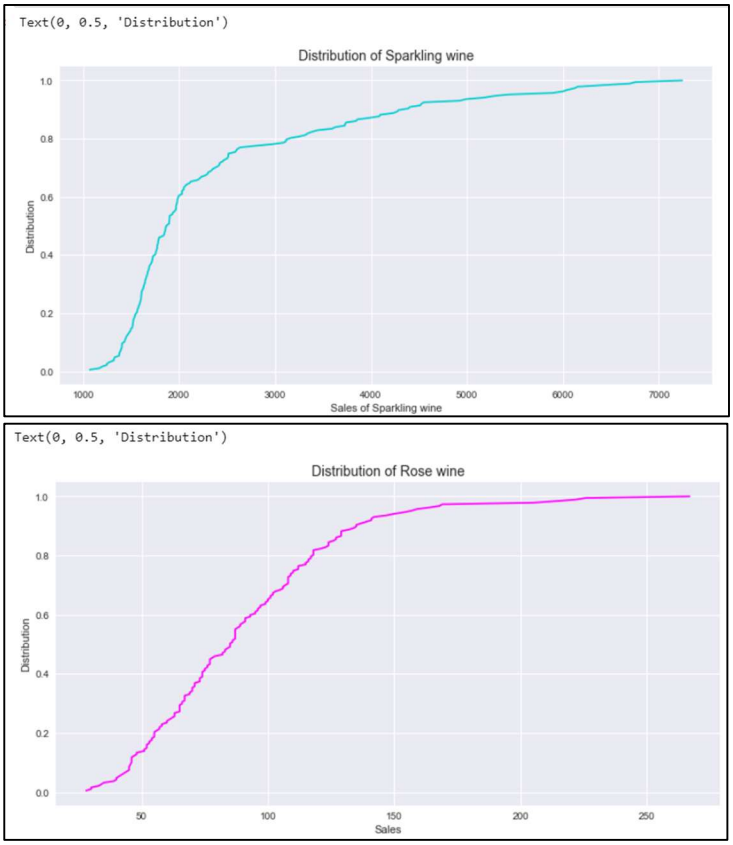


Fig 2.c Data Distribution-Sparkling & Rose

INTERPRETATION:

- The distribution shows in Rose wine starting from 28 and having a constant sale from 100-267 units. The sales seem to rise in small units so the graph looks flattened post 100 units of sales.
- For the distribution of Sparkling wine, we see the sales is rising upward throughout after 100 before which it did see a small fall in sales. The smallest sales are for units 1070 and the highest sales is for 7242 units.
-

	Sparkling	Rose
count	187.000000	187.000000
mean	2402.417112	89.914497
std	1295.111540	39.238259
min	1070.000000	28.000000
25%	1605.000000	62.500000
50%	1874.000000	85.000000
75%	2549.000000	111.000000
max	7242.000000	267.000000

The data shows the minimum count of Sparkling being more as well s maximum sales is more for Sparkling. Rose doesn't seem to outshine Sparkling in any of the mean, standard deviation or maximum value.

We can again conclude that Sparkling is most preferred at any given season or period.

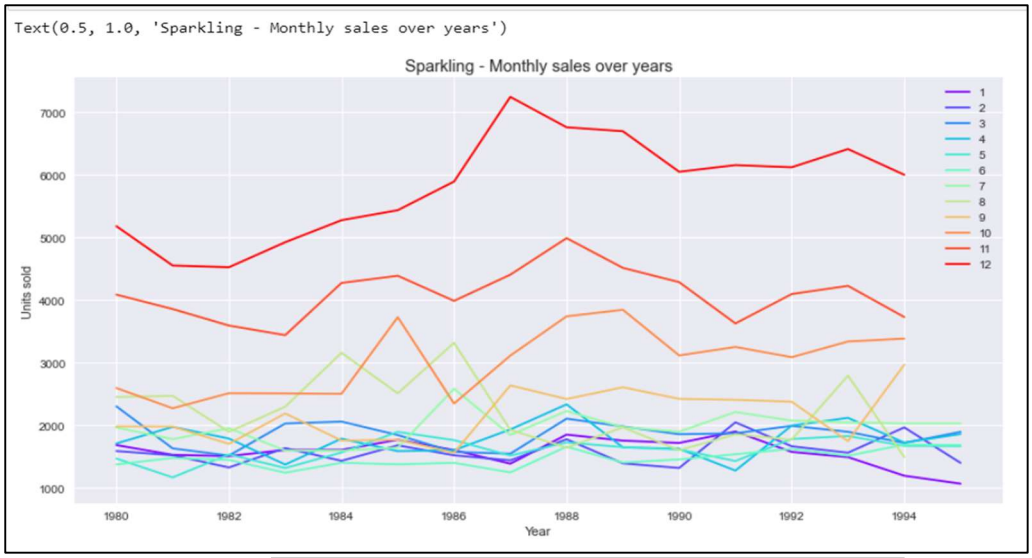


Fig 2.d Data Distribution over years as per months-Sparkling

INTERPRETATION:

1. The monthly sales are maximum in December for Sparkling wines followed by November and October. We can say that sales more is October to December for Sparkling wines.
2. The sales being the lowest is for January. The sales are also low from all months ranging January to May. However, January is the least sales month.
3. January 1995 and May 1981 are the months-years of the least sales.
4. The sales for October to December have the highest sales of above 7000 as the graph represents for the years in between 1986 to 1988.

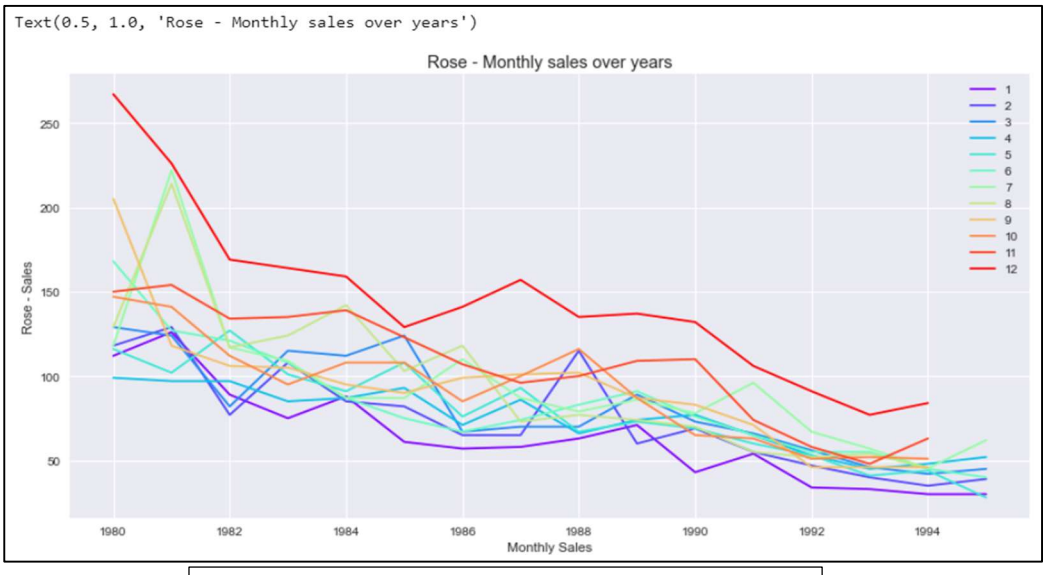


Fig 2.d Data Distribution over years as per months-Rose

INTERPRETATION:

1. The monthly sales are maximum in December for Rose wines followed by July and August. We can say that seasonal season is October to December for Sparkling wines.
2. The sales being the lowest in May 1995. The sales are also low from all months ranging January to May. However, January is the least sales month.
3. May 1995 and January 1995 are the months-years of the least sales.

4. The sales for October to December have the highest sales of above 267 as the graph represents for the years in between 1980 to 1982.

Checking the data together for Sparkling and Rose:

Fig 2.e Data Distribution over years as per months-Rose/Sparkling

INFERENCE:

- The data for Rose wine can be seen lower in sales throughout compared to Sparkling.
- The numbers also suggest the first five top sales month and unit wise that Rose is not much preferred like Sparkling is:

Year Month	Rose
1980-12	267
1981-12	226
1981-07	222
1981-08	214
1980-09	205

Year Month	Sparkling
1987-12	7242
1988-12	6757
1989-12	6694
1993-12	6410
1991-12	6153

Decomposing the Time Series is often done to help improve understanding of the time series, but it can also be used to improve forecast accuracy. We can usually identify an additive or multiplicative time series from its variation. If the magnitude of the seasonal component changes with time, then the series is multiplicative. Otherwise, the series is additive. Here we will perform both.

Additive Decomposition and Multiplicative Decomposition

Sparkling:

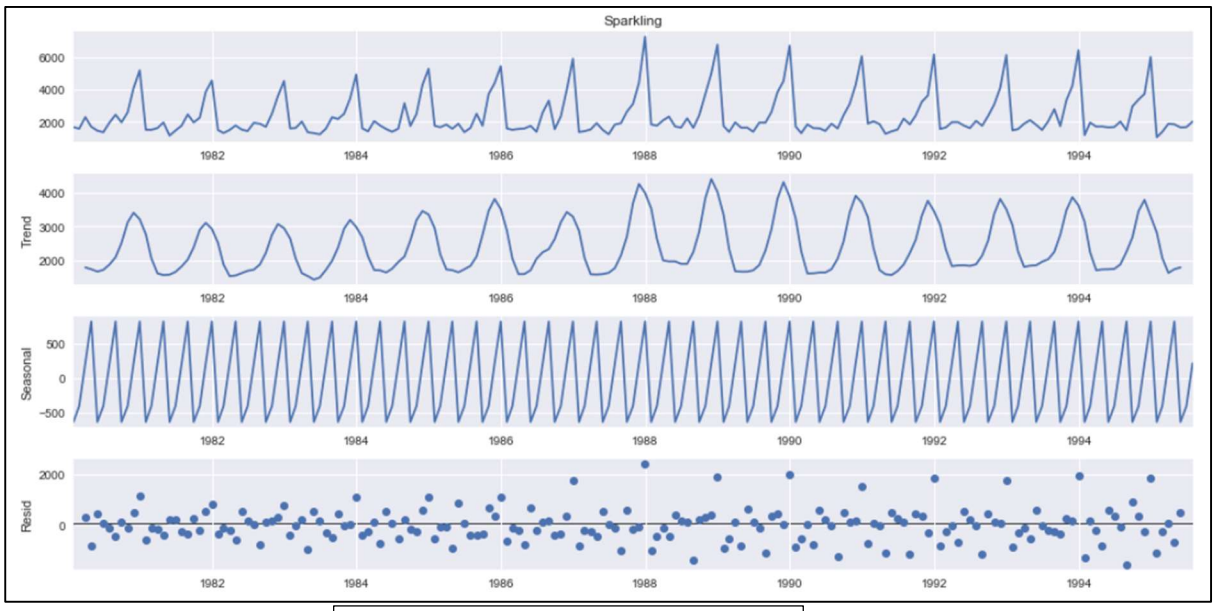


Fig 2.f Additive Decomposition-Sparkling

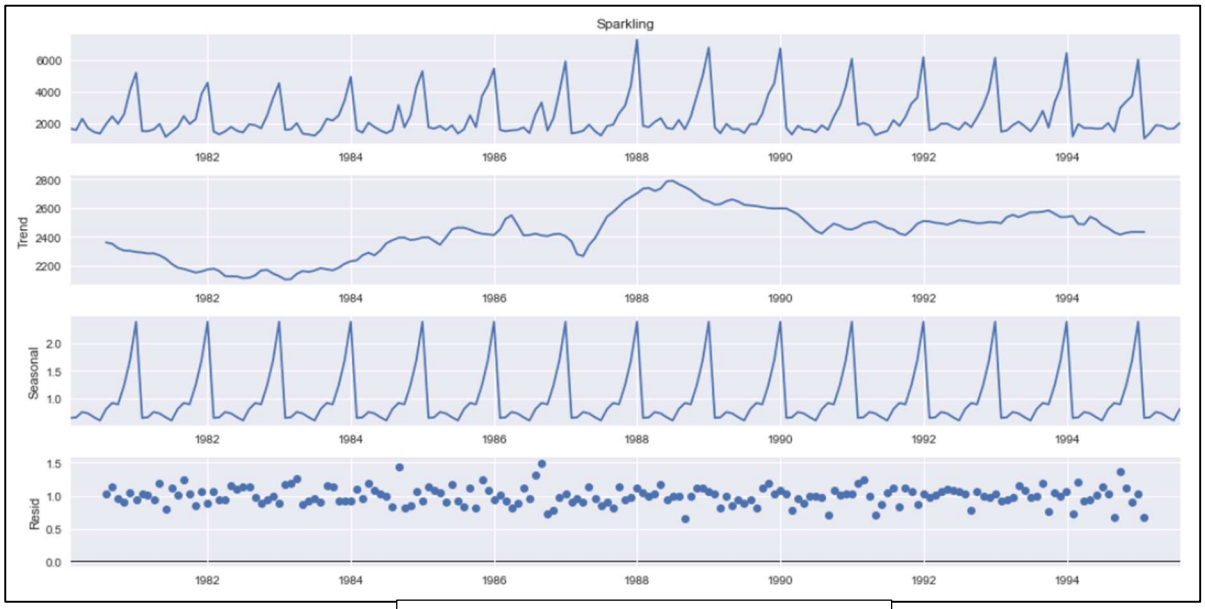


Fig 2.g Multiplicative Decomposition-Sparkling

INFERENCE:

- In additive model we can see the trend is consistent rising and falling quite consistently over all periods which is not likewise in Multiplicative model. The multiplicative model has a trend wherein we see a rise then a drop with a consistent rise until 1988 which has the highest peak of the model. Post 1990 we see a consistency in the data till 1994.
- The trend for both the models can be seen quite same which makes the data non-volatile since it doesn't depend on seasonality.
- Additive residual is ranging from -1 to 2000 which makes us sure that multiplicative model is the best suit for the data as residual is less 0.5 to 1.5.

- More the residual more the errors in the model and lesser the residual better the model with non-volatile data. Volatility hampers seasonality which is why additive model shows frequent high and low in the decomposition under seasonality section.

Rose:

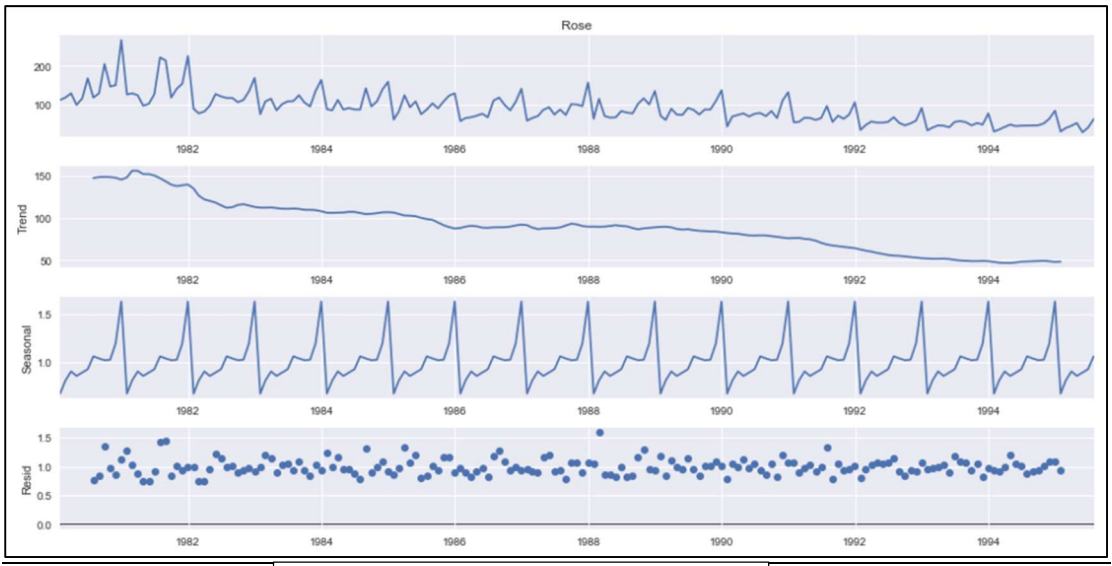


Fig 2.f Multiplicative Decomposition-ROSE

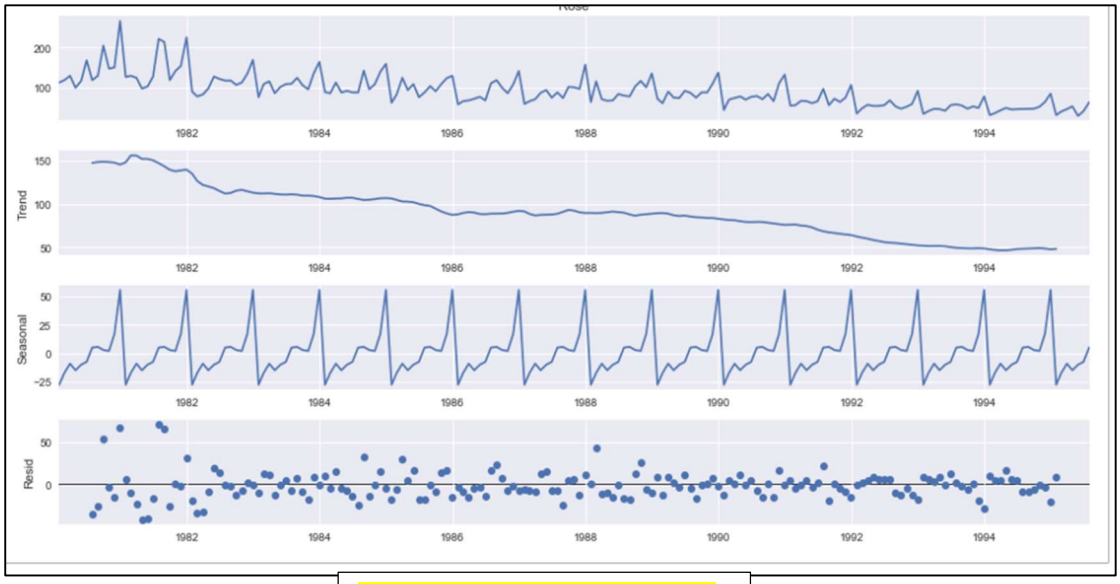


Fig 2.f Additive Decomposition-ROSE

INFERENCE:

- In additive model we can see the trend we see the trend is consistently falling which is also the same in multiplicative model. Exponential dips is seen between 1981-1983 and later 1991-1993.
- The trend for both the models can be seen quite same which makes the data non-volatile since it doesn't depend on seasonality. The variance in seasonality for additive model is 25-50 and that for multiplicative is 16%
- Additive residual is ranging from 0 to 50 which makes us sure that multiplicative model is the best suit for the data as residual is less 0.5 to 1.5.

- More the residual more the errors in the model and lesser the residual better the model with non-volatile data. Volatility hampers seasonality which is why additive model shows frequent high and low in the decomposition under seasonality section.
- Also, to add, if seasonality peaks are consistently reducing altitude with trend, we will conclude saying for ROSE multiplicative model is the best.

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

The train and test have been split as asked in the question 1991 wise. This means 70% train and 20% test.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 132 entries, 1980-01-31 to 1990-12-31
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  132 non-null    int64
 1   Rose       132 non-null    float64
dtypes: float64(1), int64(1)
memory usage: 3.1 KB
```

Fig 3. a Train DATA information

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 55 entries, 1991-01-31 to 1995-07-31
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  55 non-null     int64
 1   Rose       55 non-null     float64
dtypes: float64(1), int64(1)
memory usage: 1.3 KB
```

Fig 3. a Test DATA information

The data suggests no null value in train or test data. Also the Sales is an integer data type and the year is a float data type.

The train data has 55 columns and 2 columns. The test data has 55 rows and 2 columns:

```
(132, 2)
(55, 2)
```

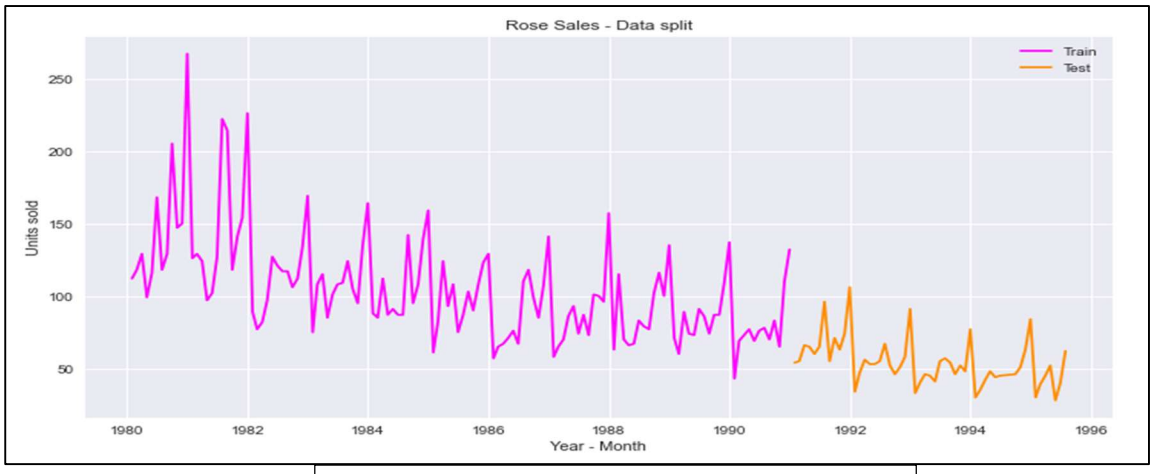


Fig 3. b Rose graphical representation of Train and Test

Inference: The Data for Train and Test for ROSE wine looks like above wherein we have data up to 1991 in Pink taken for TRAIN and Orange post 1991 as Test.

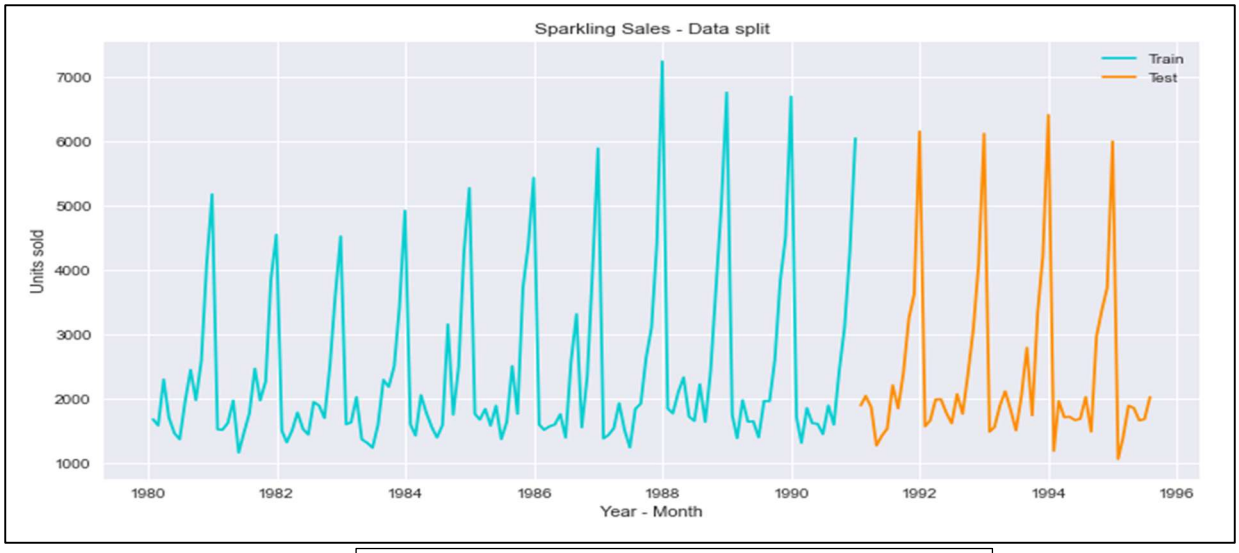


Fig 3. c Sparkling graphical representation of Train and Test

Inference: The Data for Train and Test for Sparkling wine looks like above wherein we have data up to 1991 in blue taken for TRAIN and Orange post 1991 as Test.

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

Exponential smoothing is a time series forecasting method for univariate data. Exponential smoothing is usually used to make short term forecasts, as longer-term forecasts using this technique can be quite unreliable. Simple (single) exponential smoothing uses a weighted moving average with exponentially decreasing weights.

LINEAR REGRESSION:

We use linear regression for time series analysis, it is used for predicting the result for time series as its trends. For example, here, we have a dataset of time series with the help of linear regression we can predict the sales with the time.

First few rows of Training Data

YearMonth	Sparkling	Rose	time
1980-01-31	1686	112.0	1
1980-02-29	1591	118.0	2
1980-03-31	2304	129.0	3
1980-04-30	1712	99.0	4
1980-05-31	1471	116.0	5

Last few rows of Training Data

YearMonth	Sparkling	Rose	time
1990-08-31	1605	70.0	128
1990-09-30	2424	83.0	129
1990-10-31	3116	65.0	130
1990-11-30	4286	110.0	131
1990-12-31	6047	132.0	132

First few rows of Test Data

YearMonth	Sparkling	Rose	time
1991-01-31	1902	54.0	133
1991-02-28	2049	55.0	134
1991-03-31	1874	66.0	135
1991-04-30	1279	65.0	136
1991-05-31	1432	60.0	137

Last few rows of Test Data

YearMonth	Sparkling	Rose	time
1995-03-31	1897	45.0	183
1995-04-30	1862	52.0	184
1995-05-31	1670	28.0	185
1995-06-30	1688	40.0	186
1995-07-31	2031	62.0	187

Fig 4. a. Training and Test Data for Linear regression

INFERENCE:

- We see that the training data set has dataset until 132 rows and the columns are for year-month and sales. The time column has been added for prediction.
- The Test Set has columns from 133-187 both for Rose and Sparkling wine.

Plot for LINEAR REGRESSION:

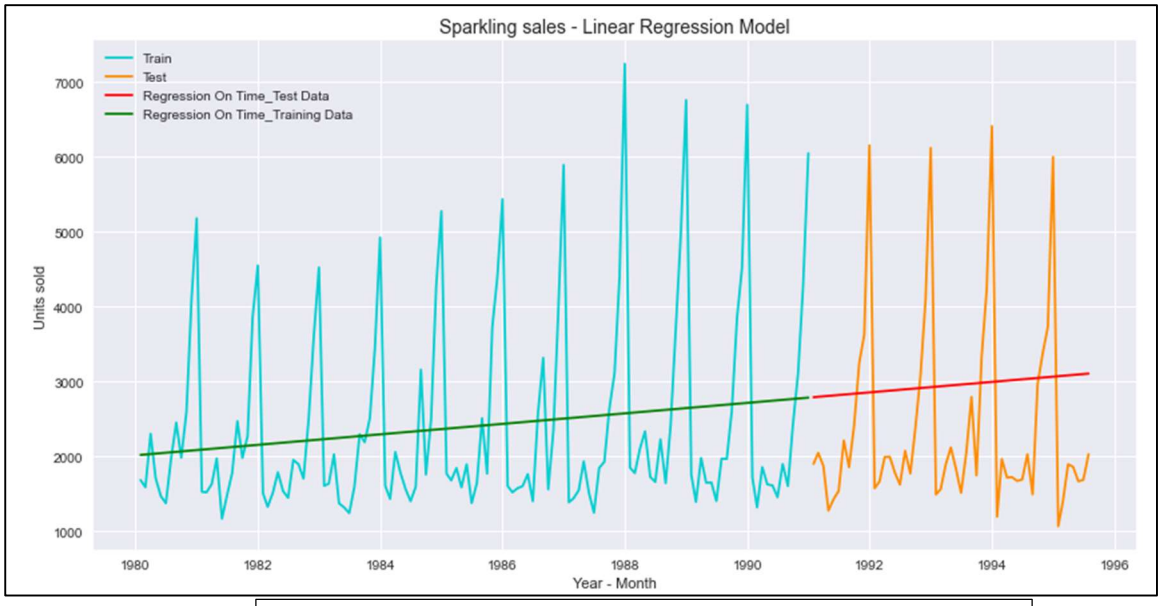


Fig 4. b. Training and Test Data for Linear regression- SPARKLING WINE

INFERENCE:

- We see that the forecast shows us an upward rising sale.
- The train and test data if seen in the image about the trend is consistent so we can expect the sales to be rising upward in future.

```
For RegressionOnTime forecast on the Sparkling Training Data:  RMSE is 1279.322 and MAPE is 40.05
For RegressionOnTime forecast on the Sparkling Testing Data:  RMSE is 1389.135 and MAPE is 50.15
```

- The RMSE value for Training data is 1279.322 which is too high and indicates the data to have errors and not have correct forecasting. The value for MAPE is 40.05 which means it is reasonable a forecasting but not great or accurate.
- The RMSE value of testing is higher with a MAPE of 50% error the forecasting is leaving us with.

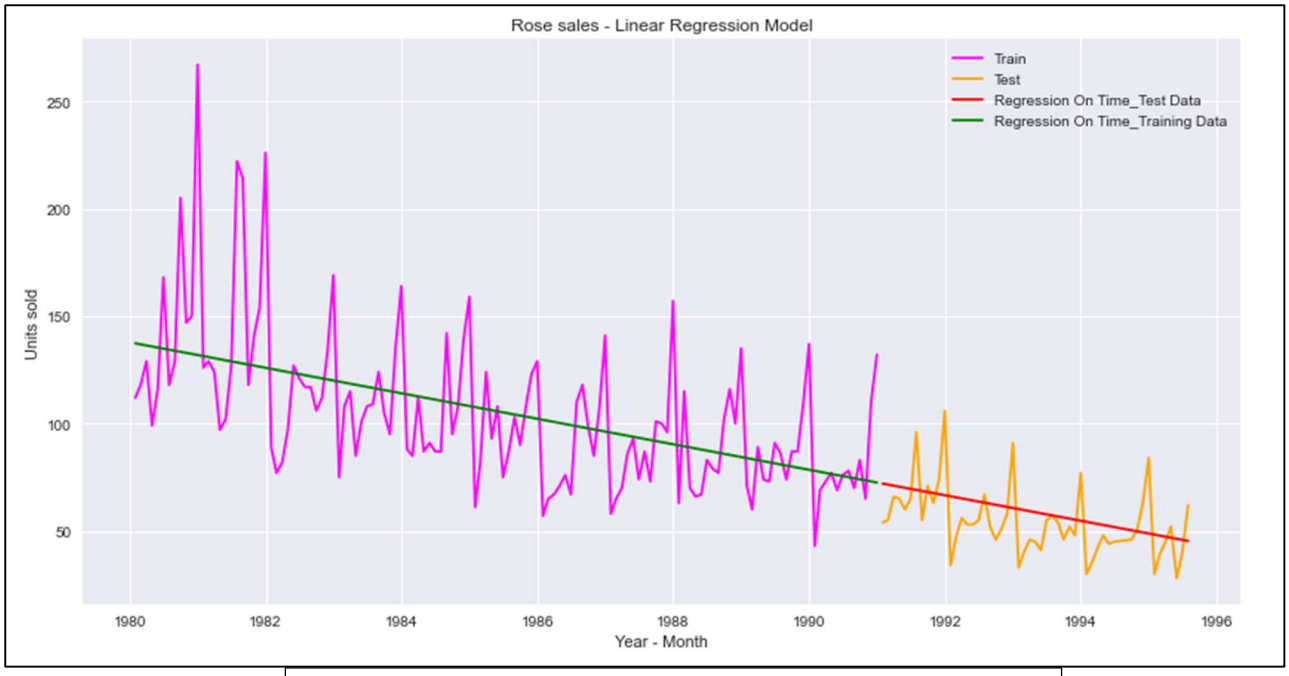


Fig 4. c. Training and Test Data for Linear regression- ROSE WINE

INFERENCE:

- We see that the forecast shows us a downward declining sale.
- The train and test data if seen in the image about the trend is inconsistent so we can expect the sales to go further down in future in the company under ROSE wine sales.

```
For RegressionOnTime forecast on the Rose Training Data:  RMSE is 30.718 and MAPE is 21.22
For RegressionOnTime forecast on the Rose testing Data:  RMSE is 15.269 and MAPE is 22.82
```

- The RMSE value for Training data is 30.718 which is too high and indicates the data to have errors and not have correct forecasting. The value for MAPE is 21.22 which means it is reasonable a forecasting but not great or accurate.
- The RMSE value for Testing is lower than training confirming the forecasting leaving us an error of 22.82% or 23% error in prediction.

Comparing linear regression results for both wines that we are selling ROSE will decline in future and the error percentage in forecast of Sparkling is more.

Plot for NAÏVE FORECAST:

A naive forecast involves using the previous observation directly as the forecast without any change. It is often called the persistence forecast as the prior observation is persisted. This simple approach can be adjusted slightly for seasonal data.

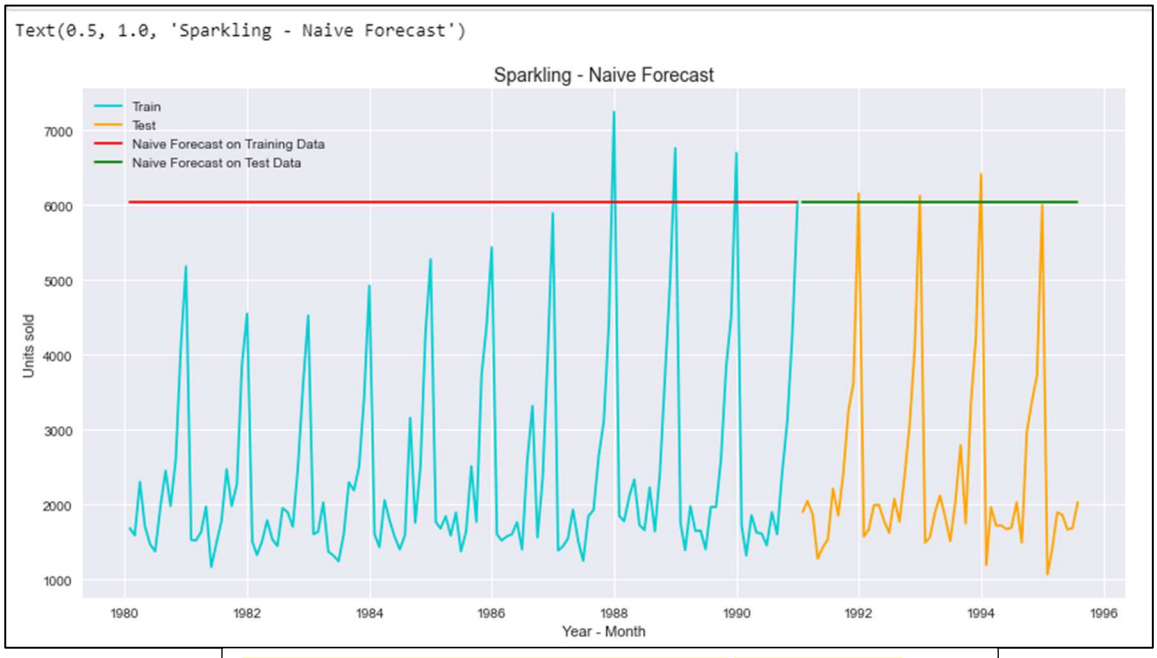


Fig 4. d. Training and Test Data for NAÏVE MODEL- SPARKLING WINE

INFERENCE:

We see that here the test and train data does have a trend. The trend in test is more than train since the volume taken there is less so the trend will impact the forecasting.

- The forecasting is a straight line because in NAÏVE we take today's data for tomorrow and tomorrows data for the day after. Hence this will cause the data to look more or less the same.

```
For Naive forecast on the Sparkling Training Data:  RMSE is 3867.701 and MAPE is 153.17
For Naive forecast on the Sparkling Testing Data:  RMSE is 3864.279 and MAPE is 152.87
```

- The predictions under NAÏVE is completely vague and inaccurate with very high scores for RMSE model. The percentage of error or MAPE is also high which indicates the error in forecasting under training is 153% and under test it is 152.87 %.

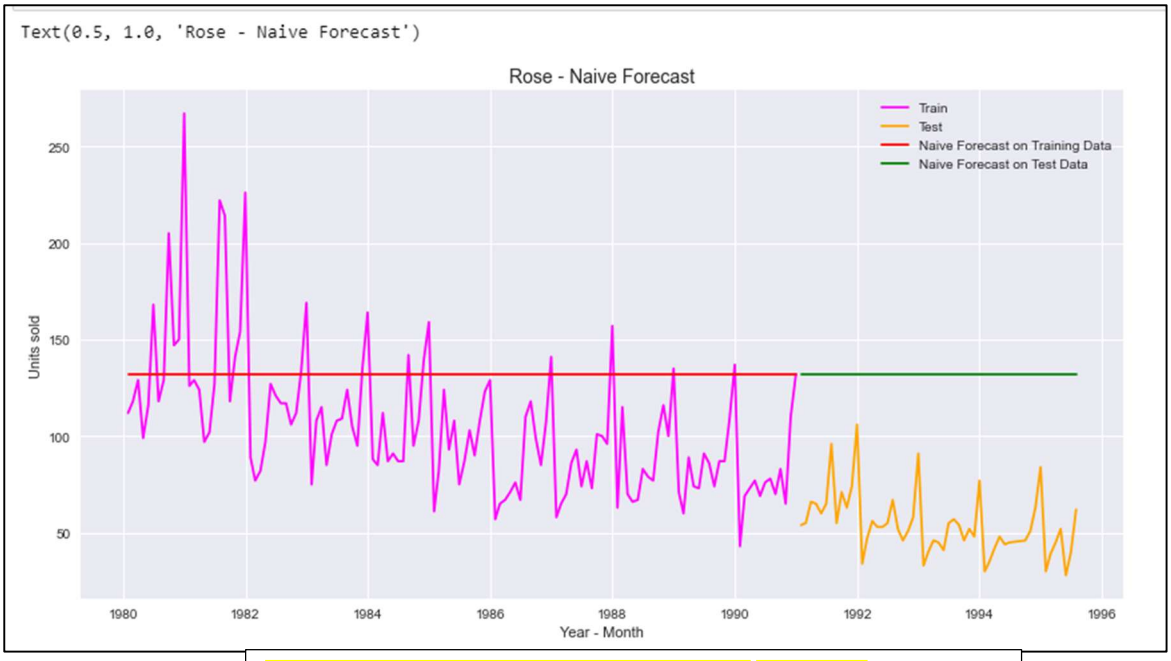


Fig 4. e. Training and Test Data for NAÏVE MODEL- ROSE WINE

INFERENCE:

We see that here the test and train data does have a trend. The trend in test is more than train since the volume taken there is less so the trend will impact the forecasting.

- The forecasting is a straight line because in NAÏVE we take today's data for tomorrow and tomorrows data for the day after. Hence this will cause the data to look more or less the same.

```
For Naive forecast on the Rose Training Data:  RMSE is 45.064 and MAPE is 36.38
For Naive forecast on the Rose Testing Data:  RMSE is 79.719 and MAPE is 145.10
```

- The predictions under NAÏVE is completely vague and inaccurate with very high scores for RMSE model. The percentage of error or MAPE % for Training data can be considered as a reasonable forecasting since scores is below 50%. However, for MAPE in testing we have high inaccuracy or error percentage of 145%.

Comparing NAÏVE results for both wines that we are selling: The NAÏVE model is not suppose to be used or is an unfit for the data we have been provided with. It predicts the forecasting for both ROSE and SPARKLING with high error percentage making this test a failure to be used for our data.

Plot for SIMPLE AVERAGE:

Such forecasting technique which forecasts the expected value equal to the average of all previously observed points is called Simple Average technique. We take all the values previously known, calculate the average and take it as the next value.

TRAIN

```
YearMonth
1980-01-31    2403.780303
1980-02-29    2403.780303
1980-03-31    2403.780303
1980-04-30    2403.780303
1980-05-31    2403.780303
Name: spark_mean_forecast, dtype: float64
```

TEST

```
YearMonth
1991-01-31    2403.780303
1991-02-28    2403.780303
1991-03-31    2403.780303
1991-04-30    2403.780303
1991-05-31    2403.780303
Name: spark_mean_forecast, dtype: float64
```

Fig 4. f. Training and Test Data for Simple average- Sparkling

The above train and test set is how our data looks post taking average of the data available. Since this simple average method is about mean of the values available with us, we see all the forecasting samples of the same value.

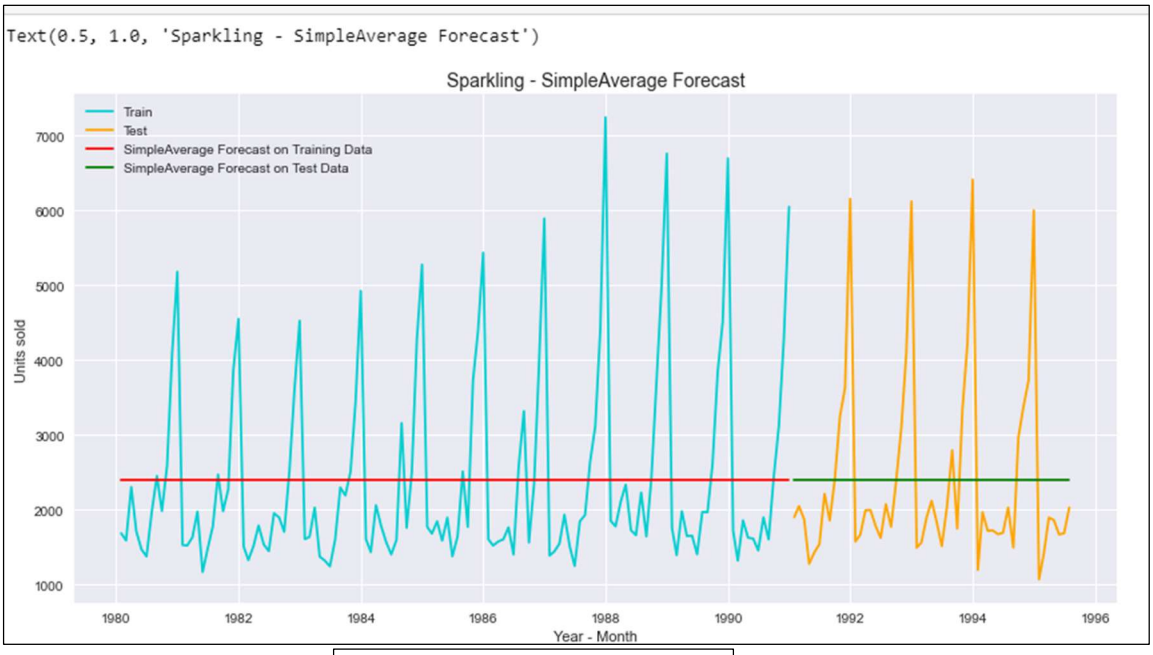


Fig 4. g. Simple average- Sparkling

INFERENCE:

- The data can be seen incapable of capturing any trend or seasonality. The same line throughout is because we have taken out the mean for the amounts.
-

```
For Simple Average forecast on the Sparkling Training Data:  RMSE is 1298.484 and MAPE is 40.36
For Simple Average forecast on the Sparkling Testing Data:  RMSE is 1275.082 and MAPE is 38.90
```

  The RMSE scores for both training and testing is high which shows Simple Average cannot be used to forecast sales of our data. The error % for both training and testing data is below 50% which means its reasonable but cannot be relied on.

TRAIN

```
YearMonth
1980-01-31    104.939394
1980-02-29    104.939394
1980-03-31    104.939394
1980-04-30    104.939394
1980-05-31    104.939394
Name: rose_mean_forecast, dtype: float64
```

TEST

```
YearMonth
1991-01-31    104.939394
1991-02-28    104.939394
1991-03-31    104.939394
1991-04-30    104.939394
1991-05-31    104.939394
Name: rose_mean_forecast, dtype: float64
```

Fig 4. h. Training and Test Data for Simple average- ROSE

The above train and test set is how our data looks post taking average of the data available. Since this simple average method is about mean of the values available with us, we see all the forecasting samples of the same value.

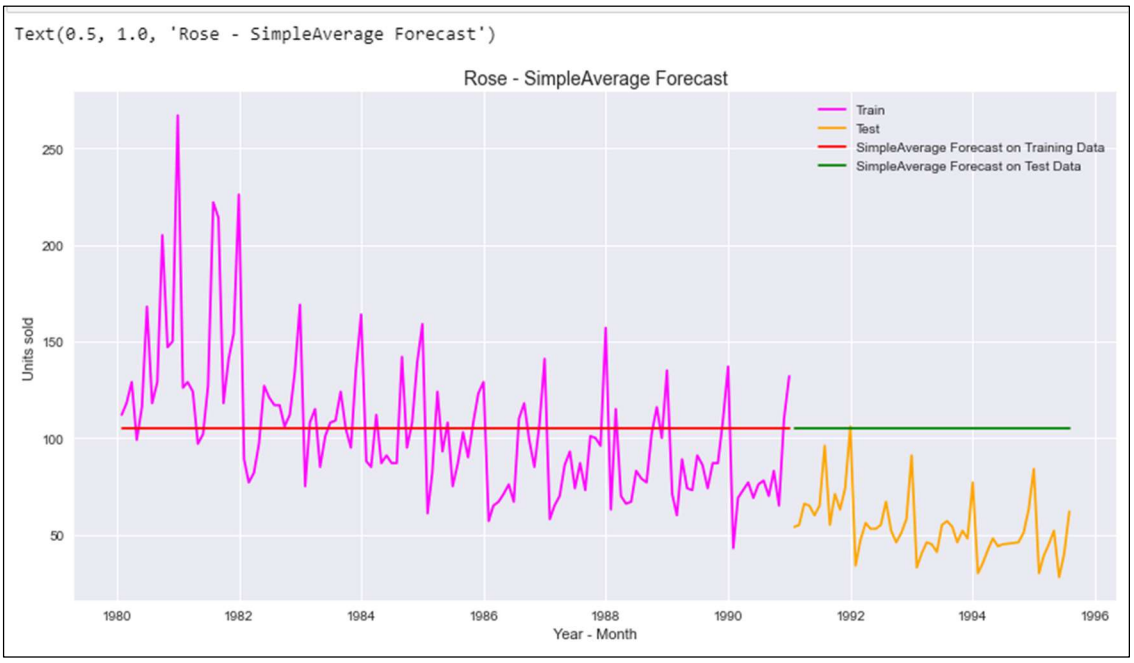


Fig 4. i. Simple average- ROSE

INFERENCE:

- The data can be seen incapable of capturing any trend or seasonality. The same line throughout is because we have taken out the mean for the amounts.
  Training data and testing data is highly inconsistent with the prediction as we seen test is below the mean line.
-

```
For Simple Average forecast on the Rose Training Data:  RMSE is 36.034 and MAPE is 25.39
For Simple Average forecast on the Rose Testing Data:  RMSE is 53.460 and MAPE is 94.93
```

  The RMSE scores for both training and testing is high which shows Simple Average cannot be used to forecast sales of our data. The error % for training is 25% which is considerate but testing data is higher than 50% which means the model cannot be trusted and the forecasting has high inaccuracy in its prediction.

Plot for MOVING AVERAGE:

A moving average is defined as an average of fixed number of items in the time series which move through the series by dropping the top items of the previous averaged group and adding the next in each successive average.

YearMonth	Sparkling	Rose	Spark_Trailing_2	Spark_Trailing_4	Spark_Trailing_6	Spark_Trailing_9	Rose_Trailing_2	Rose_Trailing_4	Rose_Trailing_6	Rose_Trai
1980-01-31	1686	112.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1980-02-29	1591	118.0	1638.5	NaN	NaN	NaN	115.0	NaN	NaN	
1980-03-31	2304	129.0	1947.5	NaN	NaN	NaN	123.5	NaN	NaN	
1980-04-30	1712	99.0	2008.0	1823.25	NaN	NaN	114.0	114.5	NaN	
1980-05-31	1471	116.0	1591.5	1769.50	NaN	NaN	107.5	115.5	NaN	

Fig 4. j. Moving average- ROSE and SPARKLING HEAD

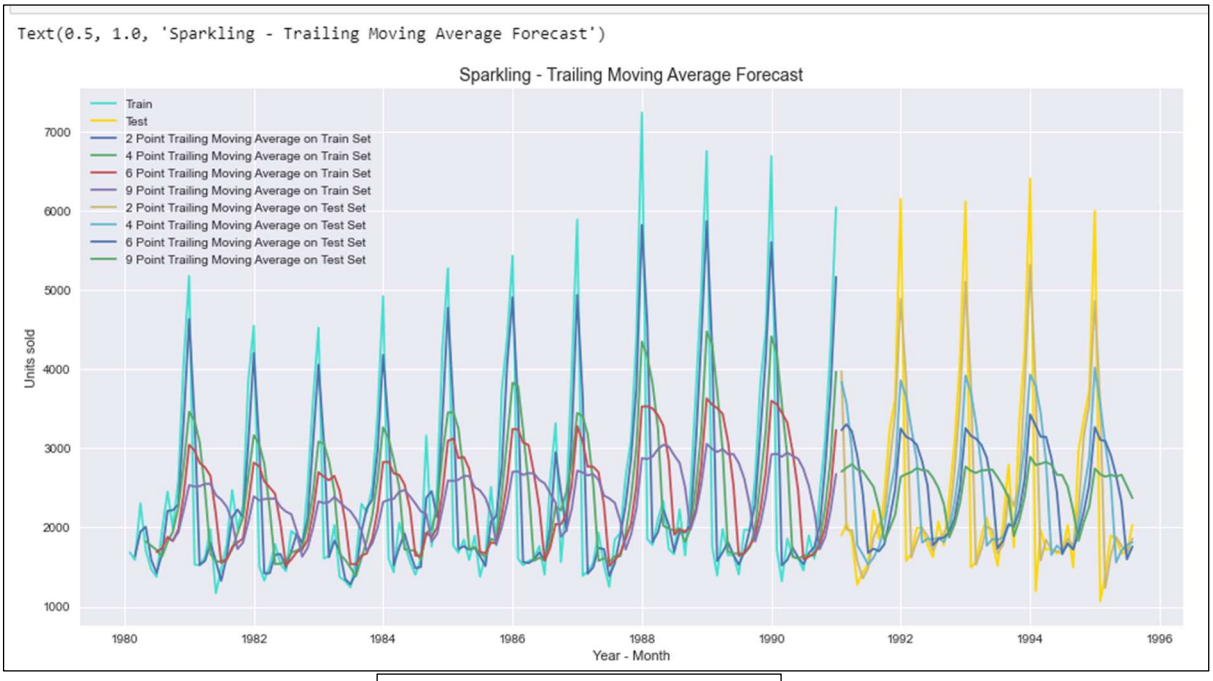


Fig 4. k. Moving average- SPARKLING

INFERENCE:

- Different trailing moving averages is counted here. One with the best scores will be considered apt.
- As indicated in the graph the points are 2,4,6 and 9.
- As the graph also indicates the lower point which is 2 in blue for train and mud yellow for test shows best result and is consistent with the trend.

```
For 2 point Moving Average Model forecast on the Training Data,  rmse_spark is 813.401 mape_spark is 19.70
For 4 point Moving Average Model forecast on the Training Data,  rmse_spark is 1156.590 mape_spark is 35.96
For 6 point Moving Average Model forecast on the Training Data,  rmse_spark is 1283.927 mape_spark is 43.86
For 9 point Moving Average Model forecast on the Training Data,  rmse_spark is 1346.278 mape_spark is 46.86
```

- The best point is point 2 as the percentage of error is 19%. The RMSE score is also lesser than other points at 813.

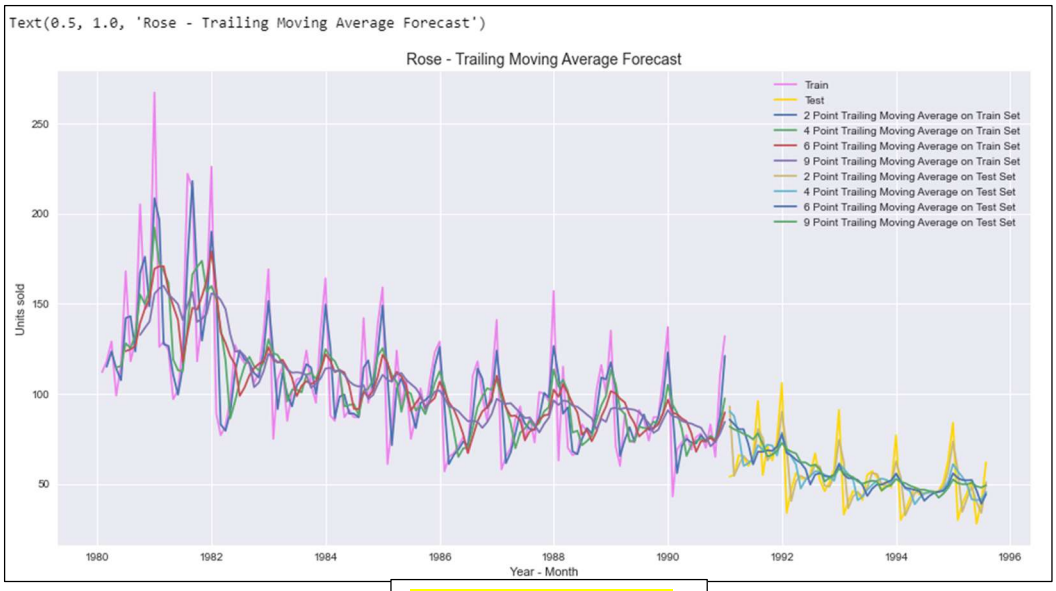


Fig 4. l. Moving average- ROSE

INFERENCE:

- Different trailing moving averages is counted here. One with the best scores will be considered apt.
- As indicated in the graph the points are 2,4,6 and 9.
- As the graph also indicates the lower point which is 2 in blue for train and yellow for test shows best result and is consistent with the trend.

```
For 2 point Moving Average Model forecast on the Training Data,  rmse_rose is 11.529 mape_rose is 13.54
For 4 point Moving Average Model forecast on the Training Data,  rmse_rose is 14.451 mape_rose is 19.49
For 6 point Moving Average Model forecast on the Training Data,  rmse_rose is 14.566 mape_rose is 20.82
For 9 point Moving Average Model forecast on the Training Data,  rmse_rose is 14.728 mape_rose is 21.01
```

- The best point is point 2 as the percentage of error is 14 (13.54) %. The RMSE score is also lesser than other points at 11.529(or 12).

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86

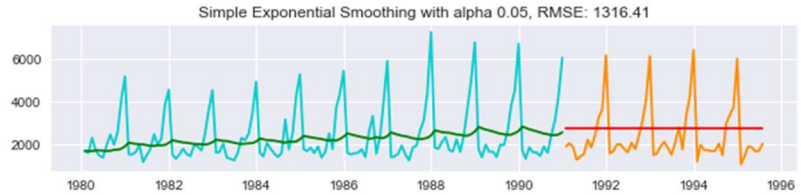
Fig 4. m. Consolidated RMSE and MAPE results

INFERENCE: The scores of points 2 is the best for the moving average model.

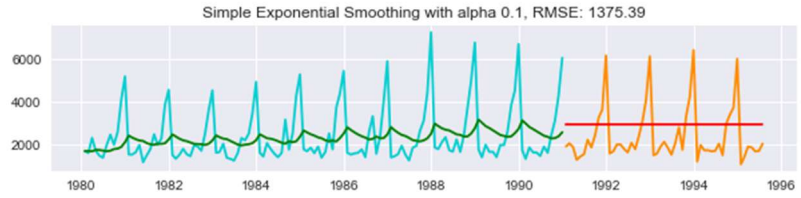
Plot for EXPONENTIAL SMOOTHING:

Single Exponential Smoothing, SES for short, also called Simple Exponential Smoothing, is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha (a), also called the smoothing factor or smoothing coefficient.

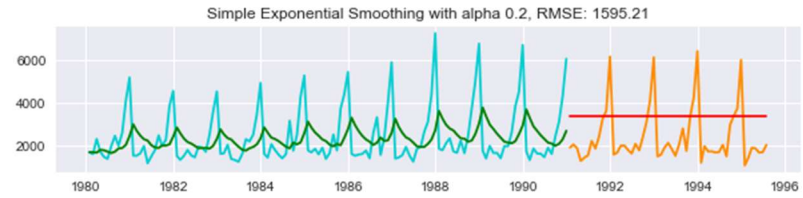
```
Test: For alpha = 0.05,  RMSE is 1316.4117 MAPE is 45.50
For smoothing level = 0.05,  Initial level 1686.00
```



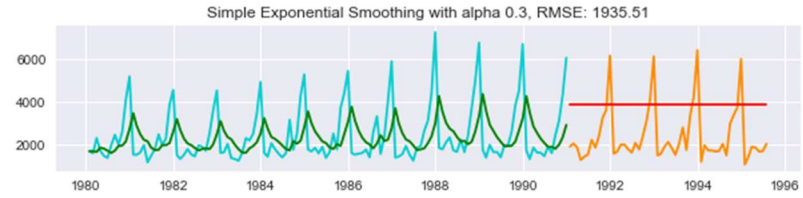

```
Test: For alpha = 0.10,  RMSE is 1375.3934 MAPE is 49.53
For smoothing level = 0.10,  Initial level 1686.00
```



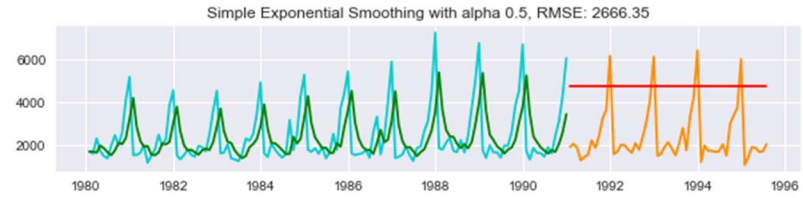

```
Test: For alpha = 0.20,  RMSE is 1595.2068 MAPE is 60.46
For smoothing level = 0.20,  Initial level 1686.00
```




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




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




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

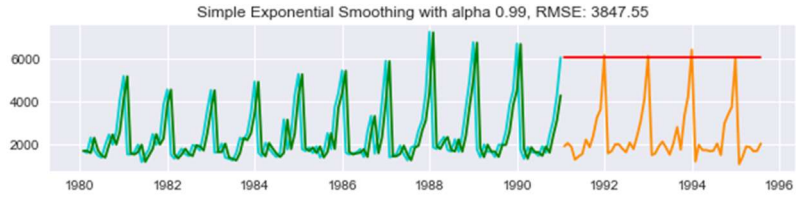


Fig 4. N Sparkling wines

INFERENCE:

- The values are neither having a trend nor a seasonality as per the alpha values.
- The alpha values which are closer to one are better fit as a model and closer to zero should be inconsiderate or rejected.
- The values are as below:

Test: For **alpha = 0.05,** RMSE is 1316.4117 MAPE is 45.50
For smoothing level = 0.05, Initial level 1686.00
Test: For **alpha = 0.10**, RMSE is 1375.3934 MAPE is 49.53
For smoothing level = 0.10, Initial level 1686.00
Test: For **alpha = 0.20,** RMSE is 1595.2068 MAPE is 60.46
For smoothing level = 0.20, Initial level 1686.00
Test: For **alpha = 0.30,** RMSE is 1935.5071 MAPE is 75.66
For smoothing level = 0.30, Initial level 1686.00
Test: For **alpha = 0.50,** RMSE is 2666.3514 MAPE is 106.27
For smoothing level = 0.50, Initial level 1686.00
Test: For **alpha = 0.99**, RMSE is 3847.5490 MAPE is 152.21
For smoothing level = 0.99, Initial level 1686.00

The alpha value that is the best is 0.05 is the best with low RMSE and error % being 45% which is the lowest of all alphas closer to one.

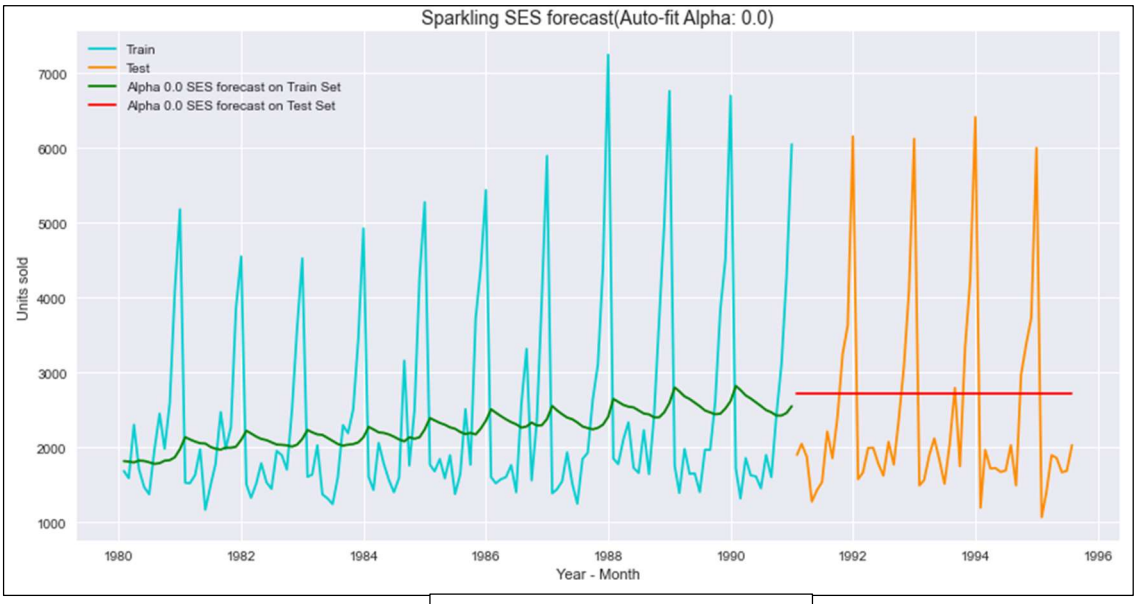


Fig 4. o Sparkling wines -best alpha

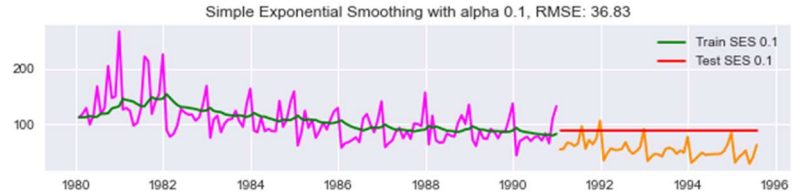
INFERENCE:

- The model auto picked alpha as zero being the best one. Hence, we see the forecasting is not having any specific trend but for test it is completely flat all throughout.
- ```
  For SES forecast on the Sparkling Training Data:  RMSE is 1315.232 and MAPE is 39.92
  ```

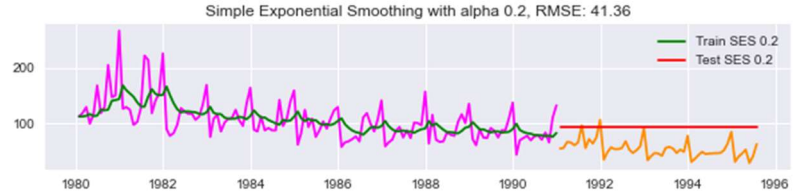
  ```
  For SES forecast on the Sparkling Testing Data:  RMSE is 1316.035 and MAPE is 45.47
  ```

  The RMSE scores are 1315 for training and 1316 for testing which is not a good score. The error percentage or MAPE score is also near to 50% which can be considered reasonable data but not fully accurate or reliable.
  ```

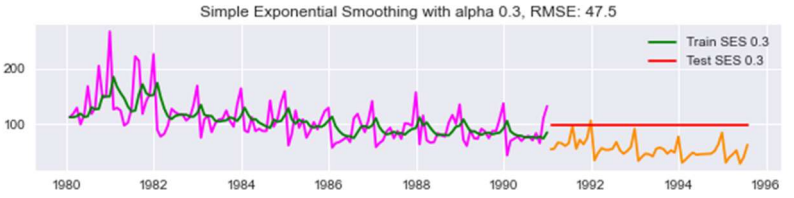
```
Test: For alpha = 0.10,  RMSE is 36.8278 MAPE is 63.94
For smoothing level = 0.10,  Initial level 112.00
```



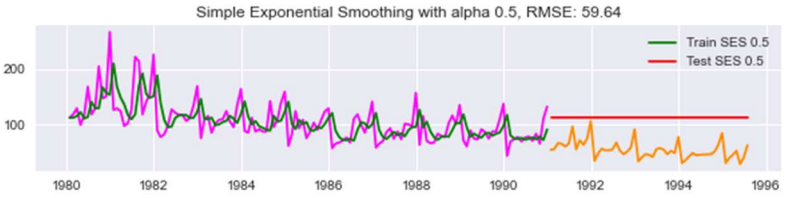

```
Test: For alpha = 0.20,  RMSE is 41.3617 MAPE is 72.21
For smoothing level = 0.20,  Initial level 112.00
```



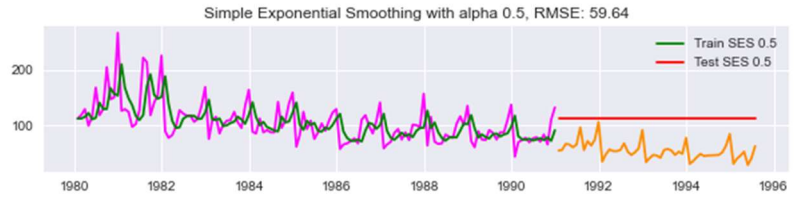

```
Test: For alpha = 0.30,  RMSE is 47.5046 MAPE is 83.71
For smoothing level = 0.30,  Initial level 112.00
```




```
Test: For alpha = 0.50,  RMSE is 59.6416 MAPE is 106.81
For smoothing level = 0.50,  Initial level 112.00
```




```
Test: For alpha = 0.50,  RMSE is 59.6416 MAPE is 106.81
For smoothing level = 0.50,  Initial level 112.00
```




```
Test: For alpha = 0.99,  RMSE is 79.4985 MAPE is 144.69
For smoothing level = 0.99,  Initial level 112.00
```

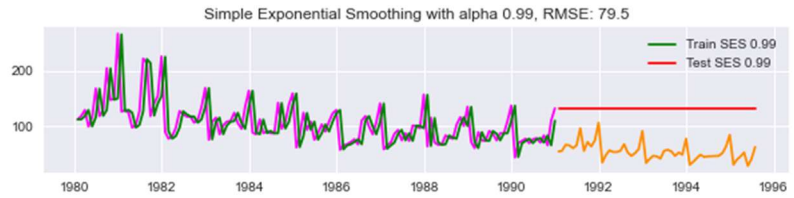


Fig 4. p Rose wines -best alpha

INFERENCE:

- The values are neither having a trend nor a seasonality as per the alpha values.
- The alpha values which are closer to one are better fit as a model and closer to zero should be inconsiderate or rejected.
- The values are as below:

Test: For alpha = 0.10, RMSE is 36.8278 MAPE is 63.94
For smoothing level = 0.10, Initial level 112.00
Test: For alpha = 0.20, RMSE is 41.3617 MAPE is 72.21
For smoothing level = 0.20, Initial level 112.00
Test: For alpha = 0.30, RMSE is 47.5046 MAPE is 83.71
For smoothing level = 0.30, Initial level 112.00
Test: For alpha = 0.50, RMSE is 59.6416 MAPE is 106.81
For smoothing level = 0.50, Initial level 112.00
Test: For alpha = 0.99, RMSE is 79.4985 MAPE is 144.69
For smoothing level = 0.99, Initial level 112.00

The alpha value that is the best is 0.10 is the best with low RMSE and error % being 45% which is the lowest of all alphas closer to one.

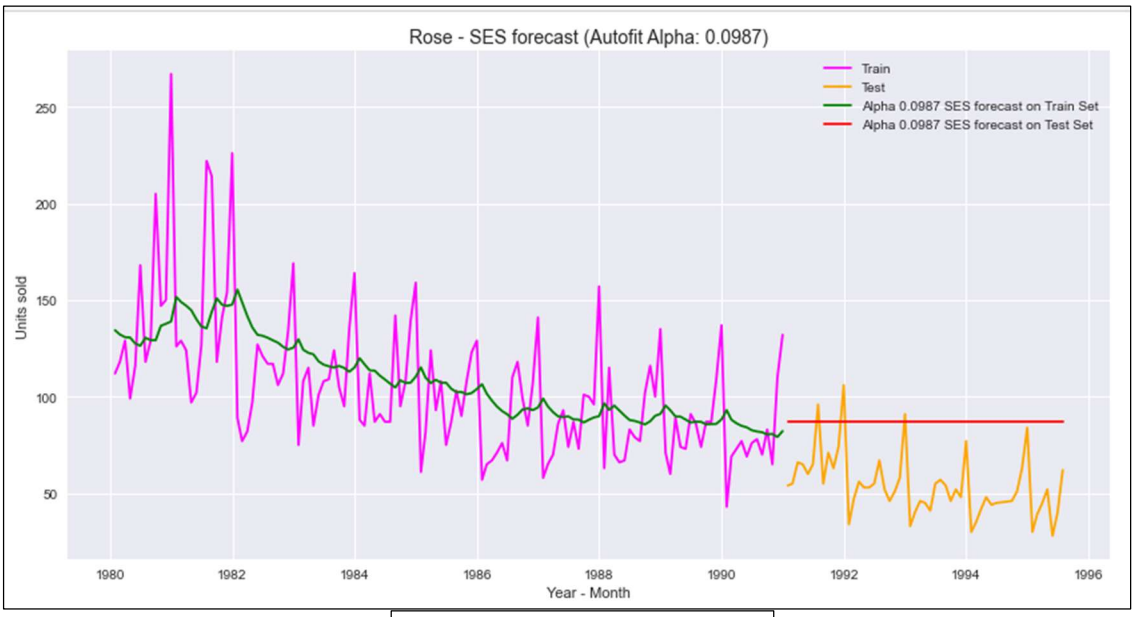


Fig 4. q Rose wines -best alpha

INFERENCE:

- The model auto picked alpha as 0.0987 being the best one. Hence, we see the forecasting is not having any specific trend but for test it is completely flat all throughout.
- ```
  For SES forecast on the Rose Training Data:  RMSE is 31.501 and MAPE is 22.73
  ```

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

  The RMSE scores are 31 for training and 37 for testing which is not a good score though but better than Sparkling wine. The error percentage or MAPE score is also near to 22% which can be considered in training and the testing data is having errors which is 64%.
  ```

Plot for Double Exponential Smoothing (Holt's Model)

Double exponential smoothing employs a level component and a trend component at each period. Double exponential smoothing uses two weights, (also called smoothing parameters), to update the components at each period.

For Sparkling Wine, we get the below:

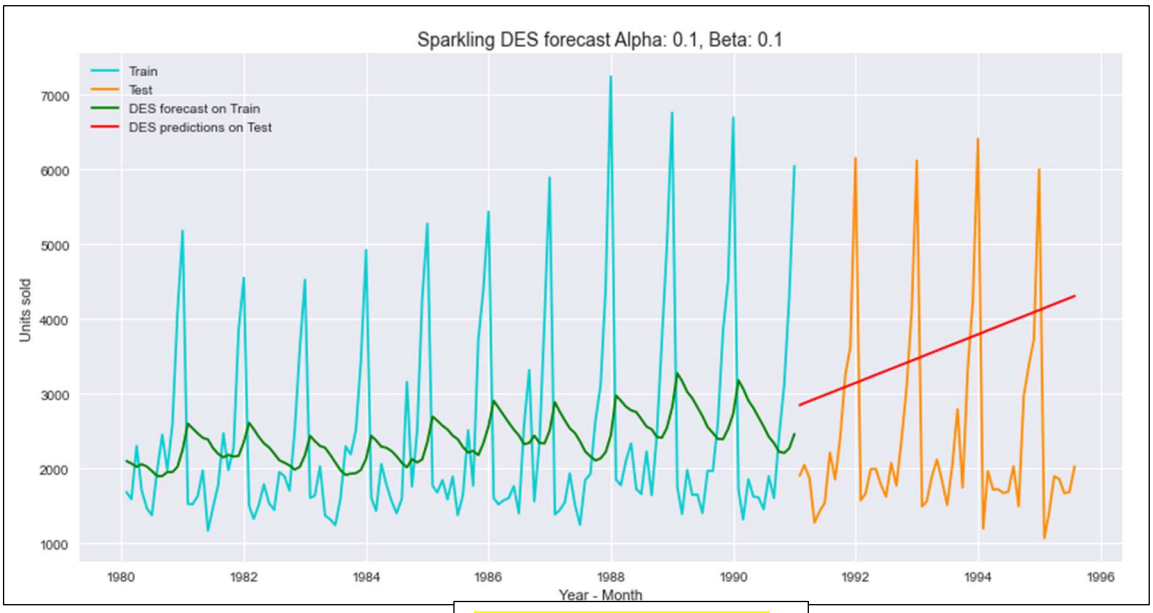


Fig 4.r. DES forecast- Sparkling

INFERENCE:

- Double exponential smoothing is applicable when data has a trend but no seasonality.
- The values in both train and test on this particular forecast is insignificant and cannot be considered however the train value looks a bit better than the test values which is just a straight line. The straight line means it is ignoring any kind of seasonality.
- We chose alpha and beta below 1 to know the best fit values and can see the best amongst all the below values is alpha = 0.1 and beta=0.1 because the RMSE is lesser as well as the MAPE which is the error percentage is also the least.

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	1363.47	44.26	1779.42	67.23
1	0.1	0.2	1398.19	45.61	2601.54	95.50
10	0.2	0.1	1412.03	46.62	3611.77	135.41
2	0.1	0.3	1431.37	46.90	4288.43	155.25
3	0.1	0.4	1466.77	48.27	6042.38	219.09

Fig 4.s. RMSE- Sparkling

- We will not be able to take into consideration any of the forecasting done under this model since the error percentage is too high which suggest forecasting is inaccurate and filled with errors.

For ROSE Wine, we get the below:

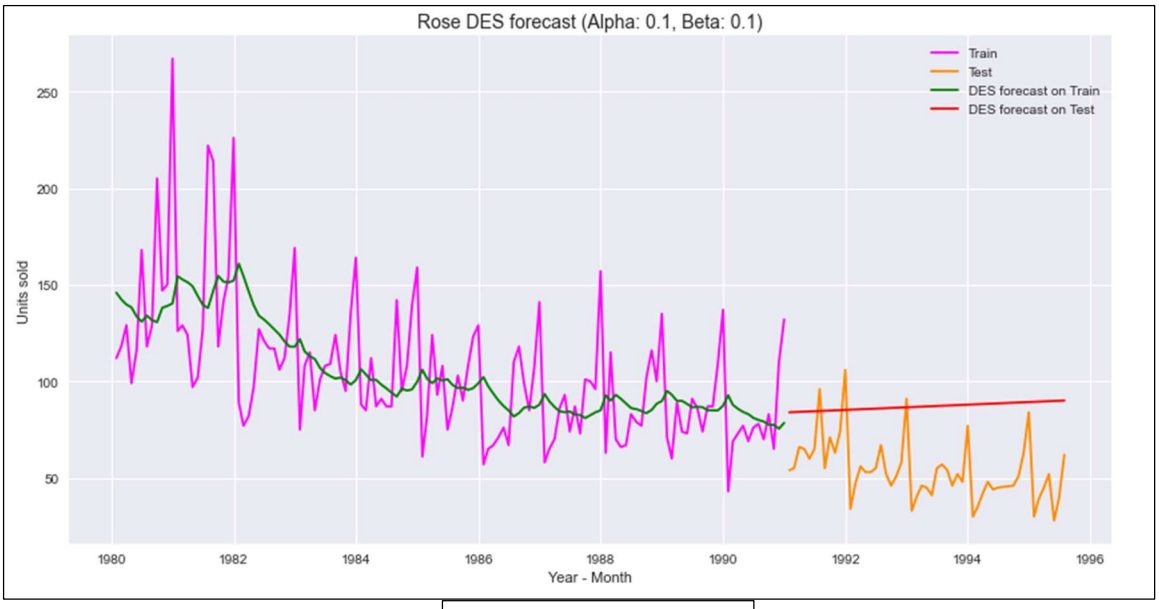


Fig 4.t. DES forecast- ROSE

INFERENCE:

- Double exponential smoothing is applicable when data has a trend but no seasonality.
- The values in both train and test on this particular forecast is insignificant and cannot be considered however the train value looks a bit better than the test values which is just a straight line. The straight line means it is ignoring any kind of seasonality.
- We chose alpha and beta below 1 to know the best fit values and can see the best amongst all the below values is alpha = 0.1 and beta=0.1 because the RMSE is lesser as well as the MAPE which is the error percentage is also the least at 64%.

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	32.026565	22.78	37.056911	64.02
1	0.1	0.2	32.685228	23.63	48.806921	83.29
10	0.2	0.1	32.796403	23.06	65.731352	113.20
2	0.1	0.3	32.925494	24.23	78.209401	131.33
3	0.1	0.4	33.179749	24.75	99.554566	165.49

Fig 4.u. RMSE/MAPE- ROSE

- We will not be able to take into consideration any of the forecasting done under this model since the error percentage is too high which suggest forecasting is inaccurate and filled with errors.

Plot for Triple Exponential Smoothing (Holt – Winter's Model)

Triple Exponential Smoothing is an extension of Exponential Smoothing that explicitly adds support for seasonality to the univariate time series. This method is sometimes called Holt-Winters Exponential Smoothing, named for two contributors to the method: Charles Holt and Peter Winters.

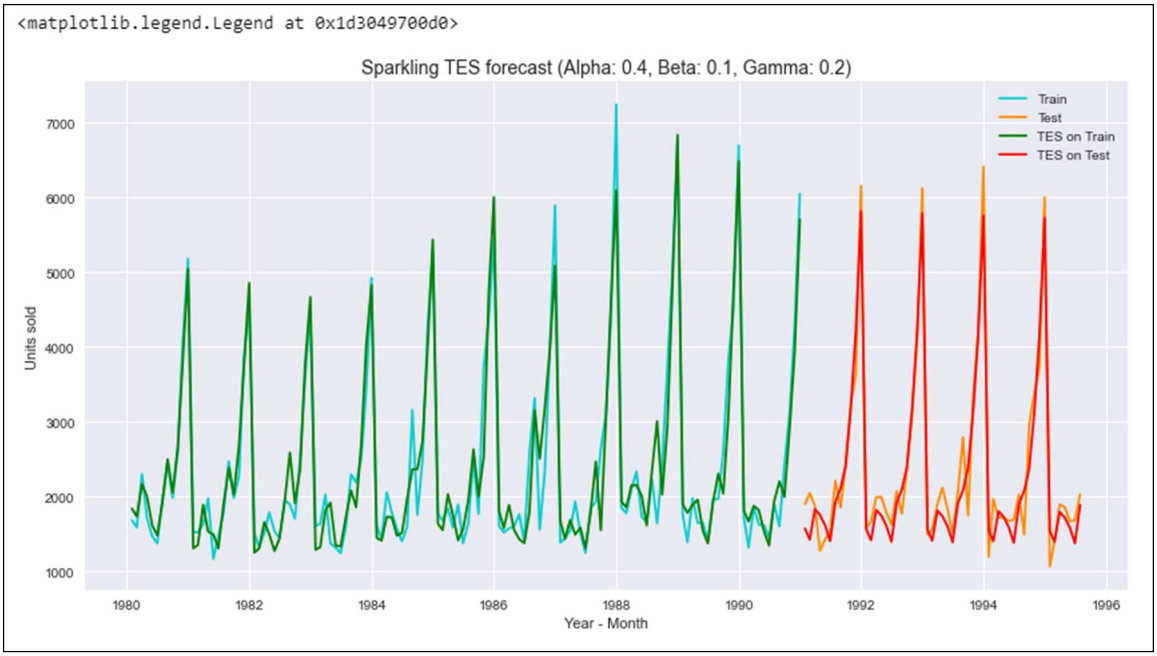


Fig 4.v. TES Forecasting-Sparkling

INFERENCE:

- Triple exponential smoothing is applicable when data has both trend and seasonality.
- The graph shows train and test is same as the prediction with little diversion.
- Smoothing level =alpha, trend = beta and seasonality = gamma is fitted to the model iteratively for values of 0.01-1 and the best amongst them all is the one in the plot which is alpha =0.4, beta =0.1 and gamma =0.2.
- We chose alpha and beta below 1 to know the best fit values and can see the best amongst all the below values is alpha = 0.5, gamma 0.3 and beta=0.1 because the error percentage is the least.

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
402	0.5	0.1	0.3	390.175608	11.54	325.545203	9.99
211	0.3	0.2	0.2	378.189776	11.28	314.882349	10.10
300	0.4	0.1	0.1	371.341930	11.14	318.045761	10.24
301	0.4	0.1	0.2	373.815525	11.13	315.533374	10.41
30	0.1	0.4	0.1	403.937167	11.72	330.772119	10.56

Fig 4.w. RMSE/MAPE- Sparkling

- Since error percentage is less, we can consider this data.

ROSE:

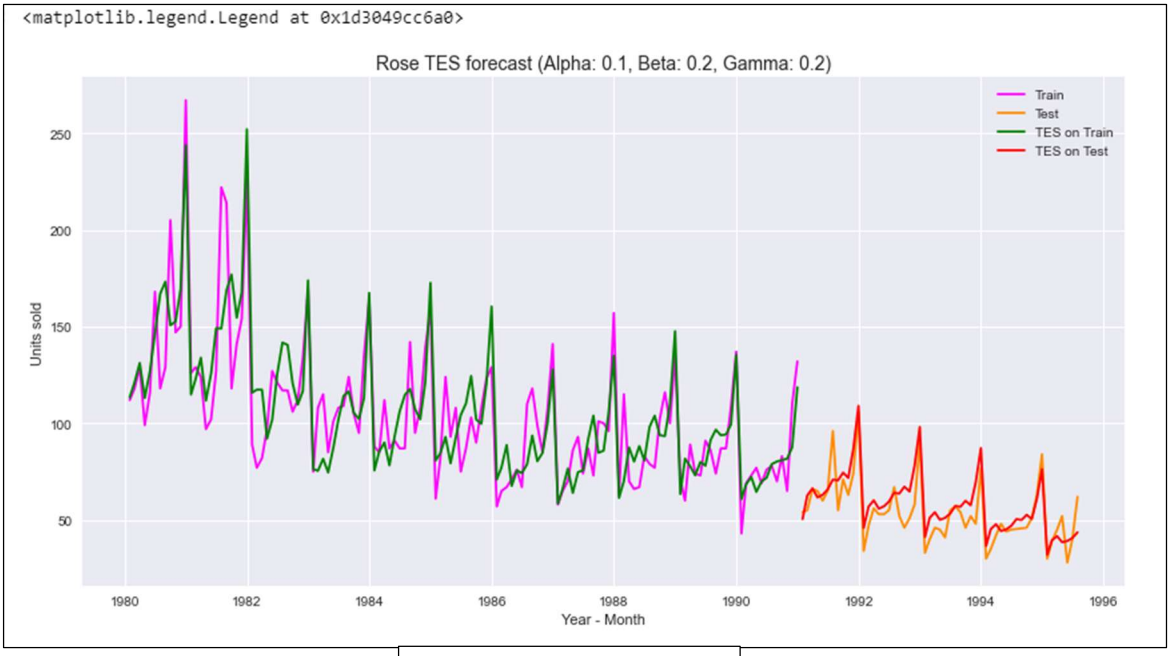


Fig 4.x. Forecasting plot- ROSE

INFERENCE:

- Triple exponential smoothing is applicable when data has both trend and seasonality.
- The graph shows train and test is same as the prediction with little diversion.
- Smoothing level =alpha, trend = beta and seasonality = gamma is fitted to the model iteratively for values of 0.01-1 and the best amongst them all is the one in the plot which is alpha =0.1, beta =0.2 and gamma =0.2.
- We chose alpha and beta below 1 to know the best fit values and can see the best amongst all the below values is alpha = 0.1, gamma 0.1 and beta=0.2 because the error percentage is the least at 13.19%.

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
10	0.1	0.2	0.1	19.651464	14.31	9.171621	13.19
11	0.1	0.2	0.2	20.140683	14.66	9.493832	13.68
151	0.2	0.6	0.2	22.793871	17.02	9.682585	13.71
142	0.2	0.5	0.3	23.300524	17.35	9.885717	14.21
12	0.1	0.2	0.3	20.725703	14.88	9.896169	14.16

Fig 4.y. RMSE/MAPE- ROSE

- Since error percentage is less, we can consider this data.

LET'S COMPARE ALL MODELS TO SEE BEST ONE:

For Sparkling we see the below RMSE (Root Mean Square Error) and MAPE(mean absolute percentage error)

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.533374	10.41
TES Alpha 0.15, Beta 0.00, Gamma 0.37	482.892737	14.95
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
6 point TMA	1283.927428	43.86
SES Alpha 0.00	1316.035487	45.47
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87

Fig 4.z. RMSE/MAPE- Sparkling

INFERENCE:

- The best model with the Least error in the RMSE and the low error percentage is Triple Exponential Smoothing. Hence, we can use the same for predicting our forecasting sales.
- The model which is 2 point Moving Average is also below 20% hence we can refer the data for the same as accurate as well.
- Concluding to say the model which should not be used in Naïve Model and which should be used is Triple Exponential Smoothing.

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.493832	13.68
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
SES Alpha 0.01	36.796004	63.88
TES Alpha 0.11, Beta 0.05, Gamma 0.00	45.036273	76.86
SimpleAverage	53.460350	94.93
NaiveModel	79.718559	145.10

Fig 4.1a. RMSE/MAPE- Rose

INFERENCE:

- The best model with the Least error in the RMSE and the low error percentage is Triple Exponential Smoothing. Hence, we can use the same for predicting our forecasting sales.
- The model which is 2 point Moving Average is also below 20% hence we can refer the data for the same as accurate as well.
- Concluding to say the model which should not be used in Naïve Model and which should be used is Triple Exponential Smoothing.

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.

When it comes to identifying if the data is stationary, it means identifying the fine-grained notions of stationarity in the data. The types of stationarity observed in time series data include

- Trend Stationary – A time series that does not show a trend.
- Seasonal Stationary – A time series that does not show seasonal changes.
- Strictly Stationary – The joint distribution of observations is invariant to time shift.

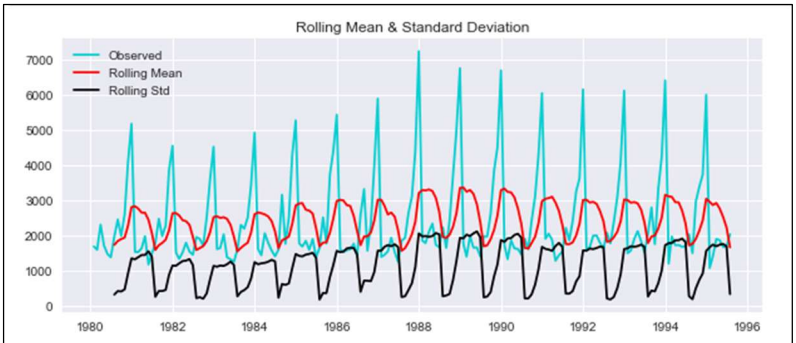
Let's define the null and alternate hypotheses for both Sparkling and Rose,

- Ho (Null Hypothesis): The time series data is non-stationary
- H1 (alternate Hypothesis): The time series data is stationary

Assume alpha = 0.05, meaning (95% confidence). The test results are interpreted with a p-value if p > 0.05 fails to reject the null hypothesis, else if p <= 0.05 reject the null hypothesis.

ADF is the test that we are going to perform. The Augmented Dickey-Fuller Test is a well-known statistical test that can help determine if your time series is stationary. In this article, I will show you how to perform the Augmented Dickey-Fuller Test (ADF) test in python.

SPARKLING:



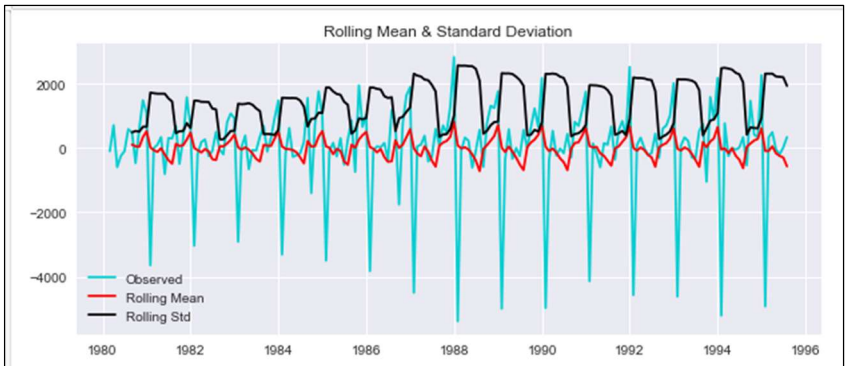

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value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

Fig 5-a Stationarity for TEST-Sparkling

INFERENCE:

- The observed values can be seen in blue which is more than the Rolling mean and Standard deviation.
- The Dickey Filler Test has given us a p-value of 0.60 which is greater than 0.05 so we fail to reject the null hypothesis.

- The critical value for 0.05 is -2.87 which is lesser than -1.36. Therefore, we again fail to reject the null hypothesis.
- Hence, we conclude series is non-stationary in Sparkling wines.
- We tried checking ADF with logarithmic transformation of train data and differencing of seasonal order (12) to understand if removing multiplicity of seasonal component will hamper data.

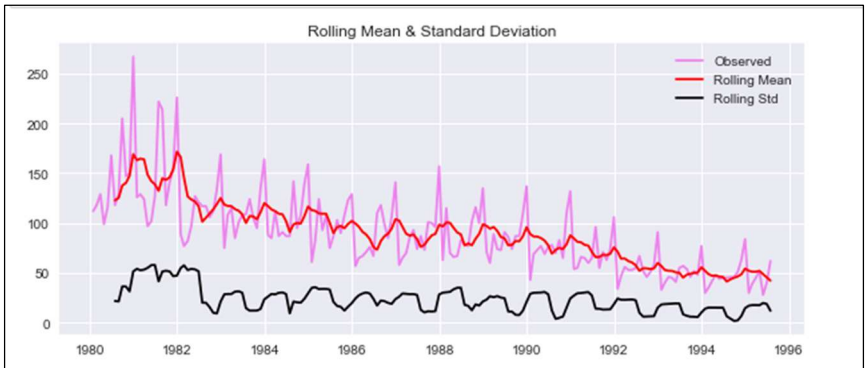



```
Results of Dickey-Fuller Test:
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical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

Fig 5-b Stationarity for TEST-Sparkling

We see above the differencing of order one we have stationary in the series.The altitude of rolling mean and standard deviation change with change in slope which indicates multiplicity.

ROSE:



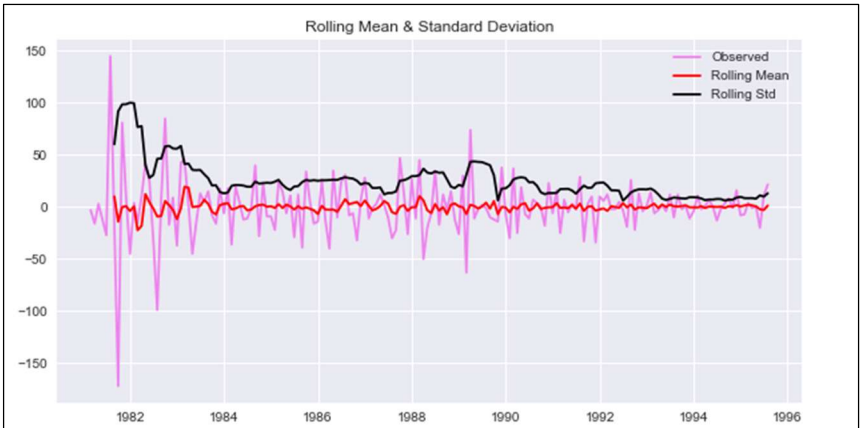

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876719
p-value                          0.343091
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

Fig 5-c Stationarity for TEST-ROSE

INFERENCE:

- The observed values can be seen in blue which is more than the Rolling mean and Standard deviation. Standard deviation is way below mean and observed values.
- The Dickey Filler Test has given us a p-value of 0.34 which is greater than 0.05 so we fail to reject the null hypothesis.

- The critical value for 0.05 is -1.87 which is lesser than -2.87. Therefore, we again fail to reject the null hypothesis.
- Hence, we conclude series is non-stationary in Sparkling wines.
- We tried checking ADF with logarithmic transformation of train data and differencing of seasonal order (12) to understand if removing multiplicity of seasonal component will hamper data.




```
Results of Dickey-Fuller Test:
Test Statistic                  -4.605732
p-value                          0.000126
#Lags Used                      11.000000
Number of Observations Used    162.000000
Critical Value (1%)             -3.471374
Critical Value (5%)             -2.879552
Critical Value (10%)            -2.576373
dtype: float64
```

Fig 5-d Stationarity for TEST-ROSE

We see above the differencing of order one we have stationary in the series. The altitude of rolling mean and standard deviation change with change in slope which indicates multiplicity. The p value is lesser than 0.00126 hence we can reject the hypothesis since p value is lesser than 0.05 with 95% confidence level.

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.

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of **ARIMA** that explicitly supports univariate time series data with a seasonal component. If your data is seasonal, like it happens after a certain period of time. then we will use **SARIMA**. p, q, d values will remain the same. period value will be the value after what period of time seasonality occurs.

An ARIMA model stands for Autoregressive Integrated Moving Average Model, and the key difference is that the model is designed to work with non-stationary data.

ARIMA MODEL

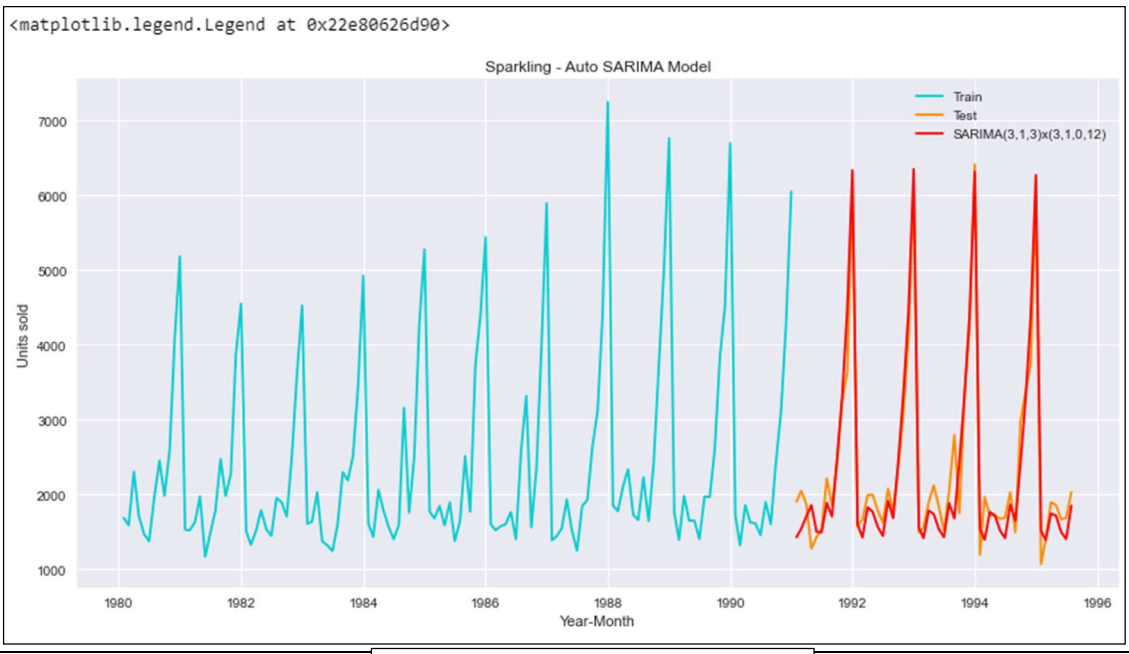


Fig 6 a .Sparkling AUTO SARIMA MODEL

INFERENCE:

- Sparkling has seasonality so we take auto SARIMA.
- The below is the result for RMSE and MAPE :

```
For SARIMA forecast on the Sparkling Testing Data:  RMSE is 336.796 and MAPE is 11.18
```

The MAPE suggests that the error is low and accurate forecasting is done. The RMSE is 336.79 which is too high and the data can be significantly retried to get better results with other models.

- We have done one model with log and one with original data.
- For the original model the data looks like below:

	param	seasonal	AIC
107	(1, 1, 2)	(2, 1, 3, 12)	18.000000
155	(2, 1, 1)	(2, 1, 3, 12)	18.000000
27	(0, 1, 1)	(2, 1, 3, 12)	88.058295
179	(2, 1, 3)	(0, 1, 3, 12)	173.539206
183	(2, 1, 3)	(1, 1, 3, 12)	236.059945

We did try the different AIC scores of which the below one comes as the best combinations:

- Parameters (p,d,q)x(P,D,Q) here is (1,1,2)x(2,1,3,12) which is best for the model.
  The original data selected the optimum values for final SARIMA MODEL as (1,1,2)x(2,1,3,12).The AIC score being 1213.283.Since this values is giving us the lowest AIC score hence this model is the best model.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 1, [1, 2, 3], 12)	Log Likelihood	0.000
Date:	Sun, 20 Feb 2022	AIC	18.000
Time:	12:36:25	BIC	39.438
Sample:	0	HQIC	26.595
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.8356	-0	-inf	0.000	1.836	1.836
ma.L1	6.1968	-0	-inf	0.000	6.197	6.197
ma.L2	-1.0846	-0	inf	0.000	-1.085	-1.085
ar.S.L12	-0.4026	-0	inf	0.000	-0.403	-0.403
ar.S.L24	-0.1507	-0	inf	0.000	-0.151	-0.151
ma.S.L12	1.472e+13	-0	-inf	0.000	1.47e+13	1.47e+13
ma.S.L24	4.92e+12	-0	-inf	0.000	4.92e+12	4.92e+12
ma.S.L36	-8.08e+13	-0	inf	0.000	-8.08e+13	-8.08e+13
sigma2	6.487e+05	-0	-inf	0.000	6.49e+05	6.49e+05

Ljung-Box (L1) (Q):	nan	Jarque-Bera (JB):	30.00
Prob(Q):	nan	Prob(JB):	0.00
Heteroskedasticity (H):	nan	Skew:	0.00
Prob(H) (two-sided):	nan	Kurtosis:	0.00

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

Fig 6 a. Sparkling SARIMAX results

INFERENCE:

- We can see all values is p>|z| is 0 which means statistically its significant.
- The AIC for this model stands at 18.
- The Ljung Box ,Jarque-Beram ,Prob and others are the different test results for the parameters selected and given in the modal.

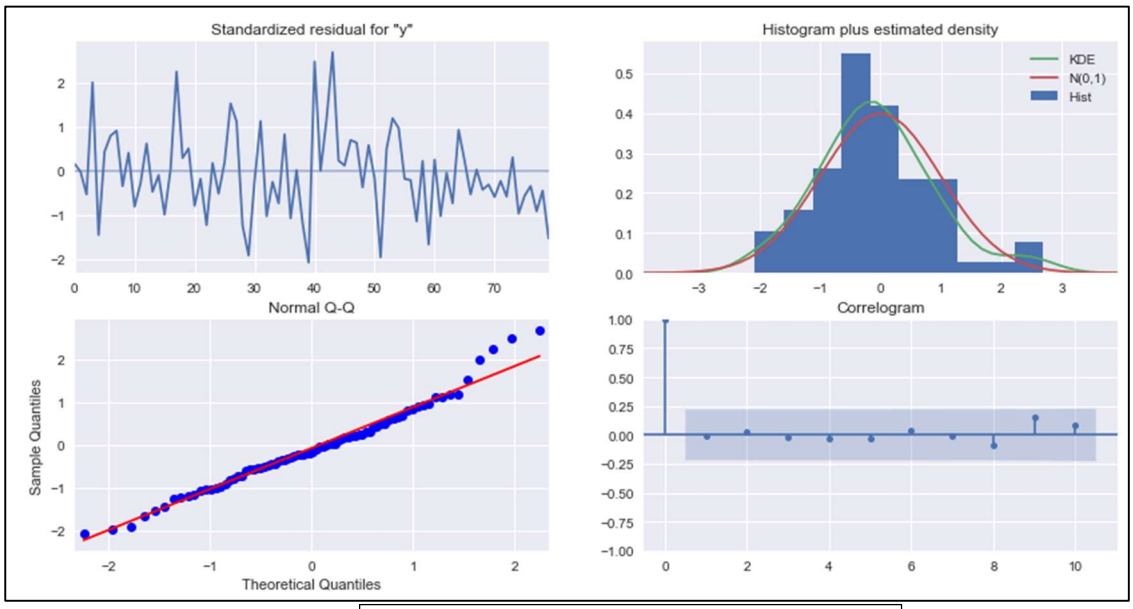


Fig 6 b. Sparkling diagnostic results-PLOT RESIDUAL

INFERENCE:

- The histogram QQ plot represents our residual. The Residuals if not in line it means our model is not working well.
- The normal QQ plot shows that the quartiles come from a normal distribution as the points form a roughly straight line.
- The correlogram shows auto correlation of the residuals and there are no significant lags above the confidence level.
- The RMSE and MAPE values of the automated SARIMA models built are given here.

ROSE:

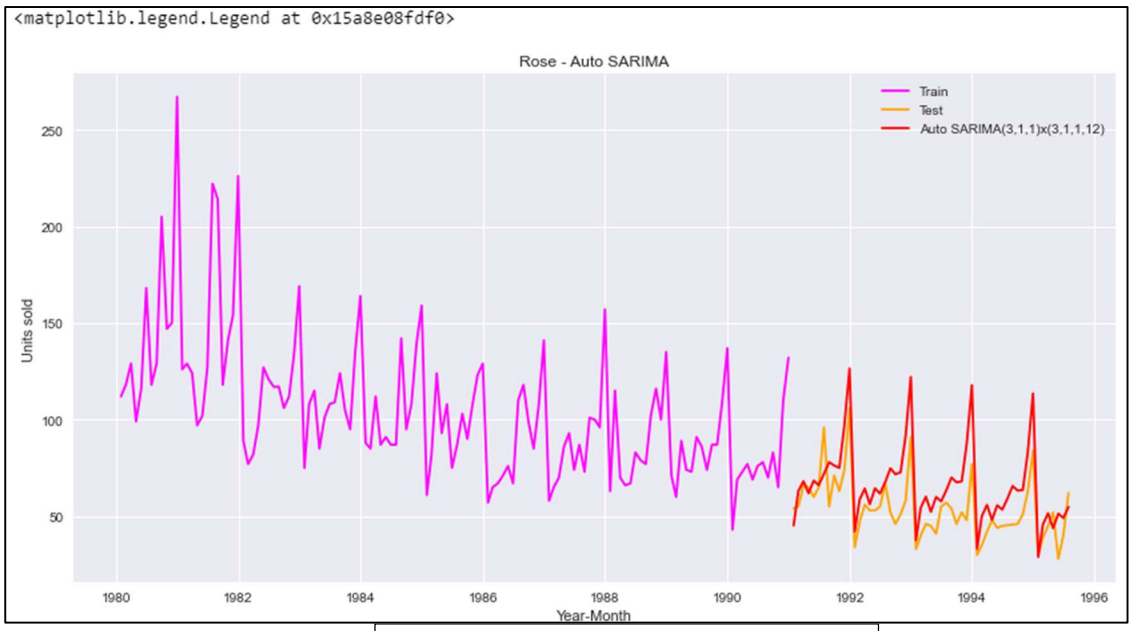


Fig 6 c. Auto SARIMA train and Test data

INFERENCE:

- We see maximum data is in train is perfectly fine and the red one in the plot is the SARIMA model which we ran on the test data to get the below RMSE and MAPE scores.
- `For SARIMA forecast on the SRose Testing Data:  RMSE is 16.823 and MAPE is 25.48`

  The model can be said to have good accuracy since MAPE is low and has a good RMSE score as well.

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(1, 0, 0)x(1, 0, [1], 12)   Log Likelihood                 132.810
Date:                              Sun, 20 Feb 2022   AIC                           -257.621
Time:                                      09:44:38   BIC                           -246.504
Sample:                                  01-31-1980   HQIC                          -253.107
                                       - 12-31-1990
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1689      0.078      2.179      0.029       0.017       0.321
ar.S.L12       0.9872      0.001    751.584      0.000       0.985       0.990
ma.S.L12      -0.9407      0.349     -2.698      0.007      -1.624      -0.257
sigma2         0.0052      0.002      2.900      0.004       0.002       0.009
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 4.00
Prob(Q):                              0.89   Prob(JB):                         0.14
Heteroskedasticity (H):               0.86   Skew:                             0.40
Prob(H) (two-sided):                  0.64   Kurtosis:                         3.40
===================================================================================

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

Fig 6 d. ROSE SARIMAX results

INFERENCE:

- From the below model we can infer Seasonal AR(2) has the highest weight and the next is seasonal MA(2)
- From p-values it is knows that AR and MA all are below 0.05 or alpha.
- The Ljung Box, Jarque-Beram, Prob and others are the different test results for the parameters selected and given in the modal.

- The AIC stands at -257.621.
- The lowest Akaike Information Criteria or AIC is as below which we have considered to get our SARIMA results above:

	param	seasonal	AIC
115	(1, 0, 0)	(1, 0, 1, 12)	-257.620750
7	(0, 0, 0)	(1, 0, 1, 12)	-256.170281
133	(1, 0, 1)	(1, 0, 1, 12)	-255.482061
25	(0, 0, 1)	(1, 0, 1, 12)	-254.978844
223	(2, 0, 0)	(1, 0, 1, 12)	-253.620649

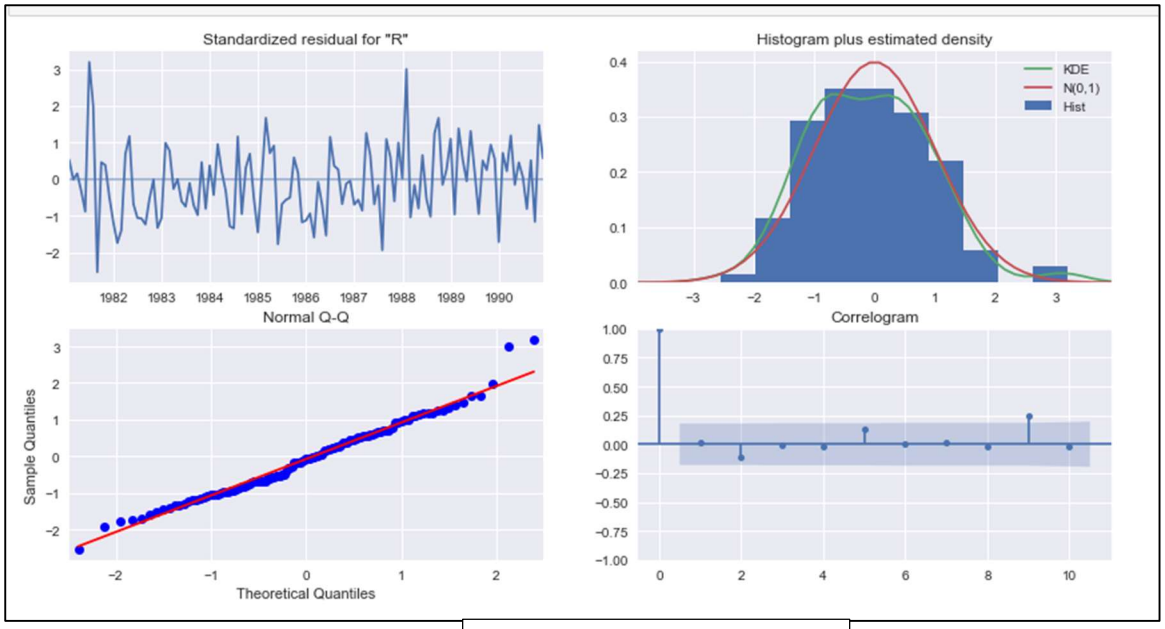


Fig 6 e. ROSE diagnostic results

INFERENCE:

- The histogram QQ plot represents our residual. The Residuals if not in line it means our model is not working well.
- The normal QQ plot shows that the quartiles come from a normal distribution as the points form a roughly straight line which means the model is significant.
- The correlogram shows auto correlation of the residuals and there are no significant lags above the confidence level.
- The RMSE and MAPE values of the automated SARIMA models built are given here.

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

SPARKLING:

From the ACF plot of the observed train data we can infer that seasonal interval of 12, the plot is not quickly vailing away. So, difference has been taken to be 12.

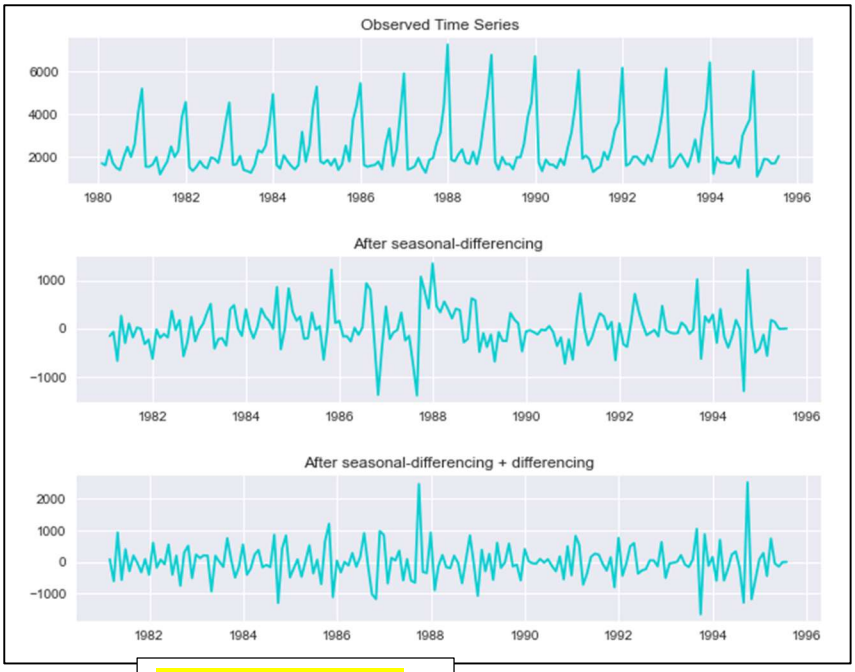


Fig7 a. ACF plot-Sparkling

From the plots below an apparently slight trend is still existing after differencing of the seasonal order at 12. If we move on post 12, we will see no trend.

- An ADF test need to be done to check the stationarity after the above differencing is done.
- P>0.05 and test statistic below critical values it can be concluded data is stationary.

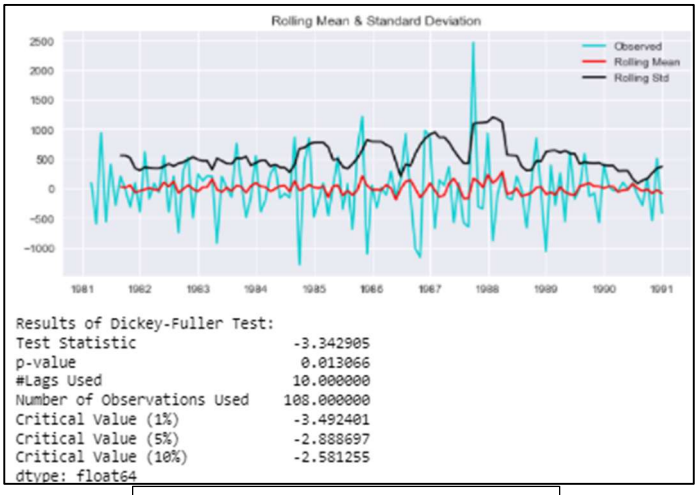


Fig7 b. Stationarity check-Sparkling

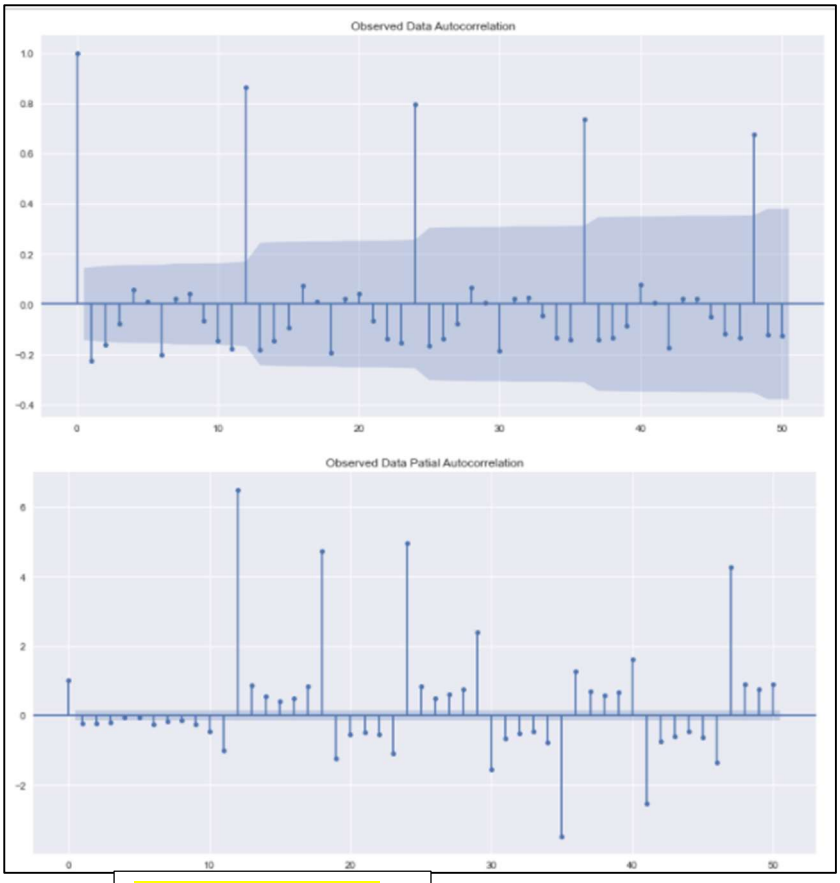


Fig7 c. ACF plot-Sparkling

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

Fig7 d Stationary check

We do see alpha less than 0.05 and the p value is also less than 0.05 proving data stationarity.

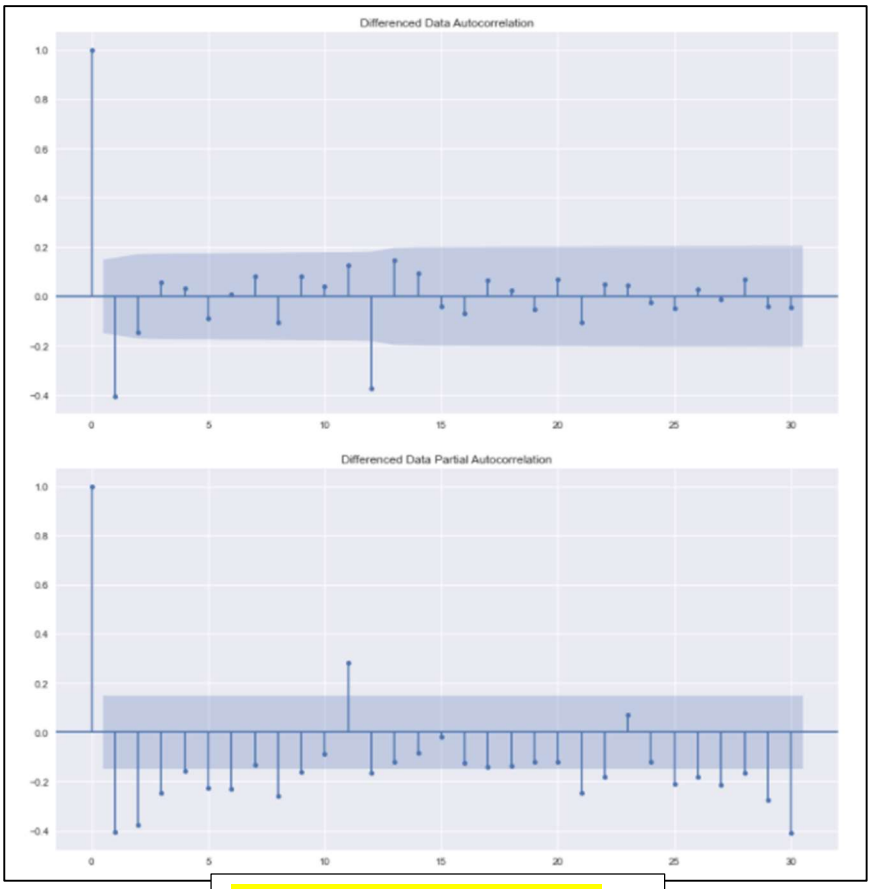


Fig7 e. ACF/PACF lag 30 plot-Sparkling

1. Alpha taken as 0.05 and seasonal period as 12.
2. From PACF we see after every 3 lags there is a cut-off, so AR term ‘p=3’ is chosen. At seasonal lag of 12 it almost cuts off so seasonal AR ‘p=1’/
3. From ACF plot we see lag 1 is significant before it cuts off, so MA term ‘q=1’ is selected and at seasonal lag of 12 a significant lag is apparent, so kept seasonal MA term ‘Q=1’ initially.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(1, 1, [1, 2], 12)	Log Likelihood	-693.697
Date:	Sun, 20 Feb 2022	AIC	1403.394
Time:	09:16:48	BIC	1423.654
Sample:	0	HQIC	1411.574
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2229	0.130	1.713	0.087	-0.032	0.478
ar.L2	-0.0798	0.131	-0.607	0.544	-0.337	0.178
ar.L3	0.0921	0.122	0.756	0.450	-0.147	0.331
ma.L1	-1.0241	0.094	-10.925	0.000	-1.208	-0.840
ar.S.L12	-0.1992	0.866	-0.230	0.818	-1.897	1.499
ma.S.L12	-0.2109	0.881	-0.239	0.811	-1.938	1.516
ma.S.L24	-0.1299	0.381	-0.341	0.733	-0.877	0.617
sigma2	1.654e+05	2.62e+04	6.302	0.000	1.14e+05	2.17e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	19.66
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	0.81	Skew:	0.69
Prob(H) (two-sided):	0.56	Kurtosis:	4.78

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

Fig7 f. SARIMA plot-Sparkling

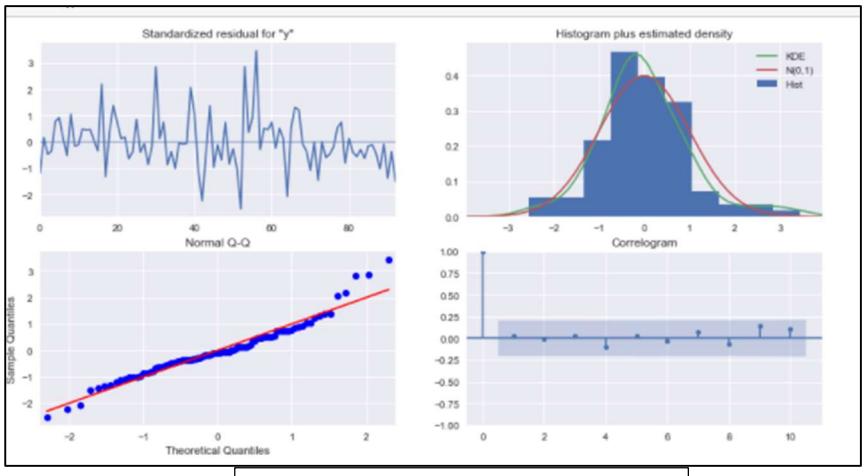


Fig7 g. Residual plot-Sparkling

Inference fig 7 (d&e):

- The seasonal MA term 'Q' was later reduced to 2, by validating model performance, as the data might be under-differenced.
- The final selected terms for SARIMA model are (3, 1, 1).
- The diagnostic plot for the model is as below, which clearly shows a normal distribution of the residuals, where more values are around zero.
- The normal QQ plot shows that quantiles come from a normal distribution as the points forms roughly a straight line.
- The correlogram shows auto correlation of the residuals and there are no points significant above the confidence index.
- From the multiple iterations of SARIMA model or scores for accuracy and RMSE , MAPE :

	Test RMSE	Test MAPE
Auto SARIMA(3,1,3)x(3,1,0,12)	331.586044	10.33
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.795755	11.18
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106737	9.48

The best model out of the above three is Manual SARIMA for sparkling wines. This is because accuracy is high as per the results of RMSE and MAPE results.

ROSE

Log transformation of the data is done to handle multiplicity of seasonality.

- From the ACF plot of the log transformed data, it can be seen that at seasonal interval of 12, the plot is not quickly tapering off. So, we need to take a seasonal differencing of 12.
- From the plots below it can be seen that a slight trend is still existing after the differencing of the seasonal order of 12. With a further differencing of order one, no trend is present.

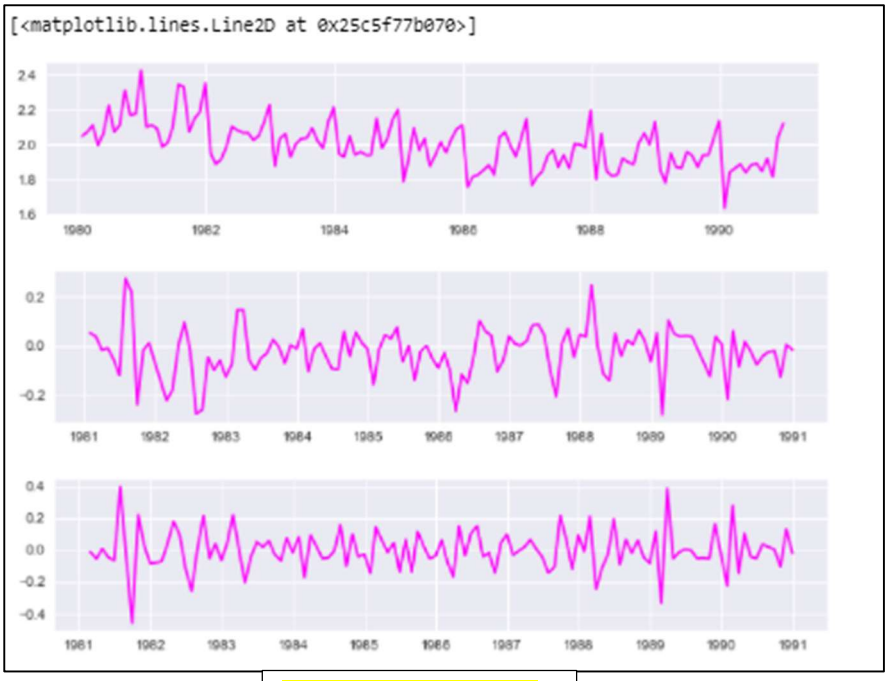
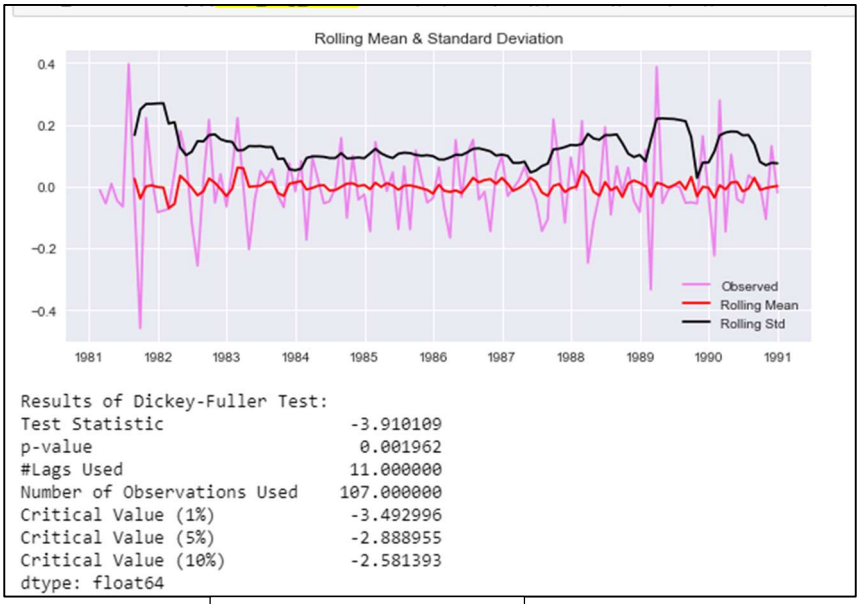


Fig7 h ACF PACF plot-Rose

- Have done an ADF test to check stationarity of the data after differencing. With a p>value below alpha 0.05 and the test statistics below critical values, it can be confirmed that the data stationary is true.
- ACF and PACF plots of the seasonal difference + one order difference data is created to find the value for the (pxdxq)x(P,D,Q).



-

Fig7 I ACF PACF plot-Rose

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

ACF and PACF plots of the seasonal difference + one order difference data is created to find the value for the (pxdxq)x (P, D, Q).

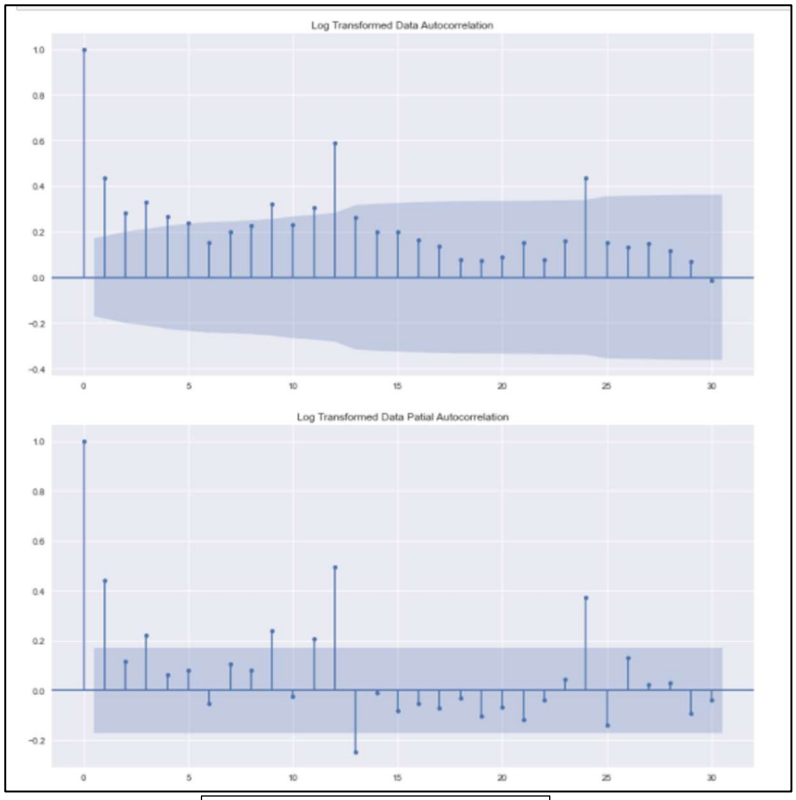


Fig7 j ACF PACF plot log -Rose

1. Alpha taken as 0.05 and seasonal period as 12.
2. From PACF we see after every 4 lags there is a cut-off, so AR term 'p=4' is chosen. At seasonal lag of 12 it almost cuts off so seasonal AR 'p=1'/
3. From ACF plot we see lag 1 and 2 are significant before it cuts off, so MA term 'q=1' is selected and at seasonal lag of 12 a significant lag is apparent, so kept seasonal MA term 'Q=1'.

- The final selected terms for the SARIMA model is (4x1x1)x(0,1,1,12) as inferred from ACF and PACF plots.
- The diagnostics plot for the model is as below, which clearly shows a normal distribution of the residuals since all points are near to zero.
- The Normal QQ plot also shows that the quartiles come from a normal distribution as the points forms roughly a straight line.
- The correlogram shows the auto correlation of the residuals and there are no points significant above the confidence index.

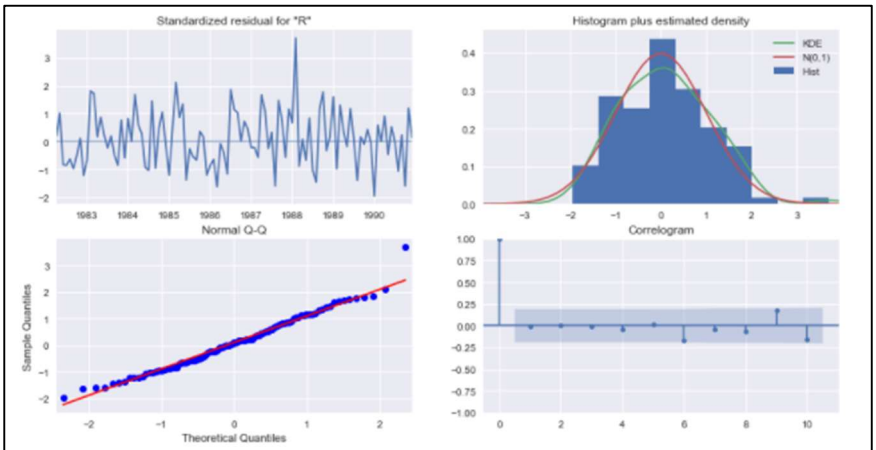


Fig7 k diagnostic plot -Rose

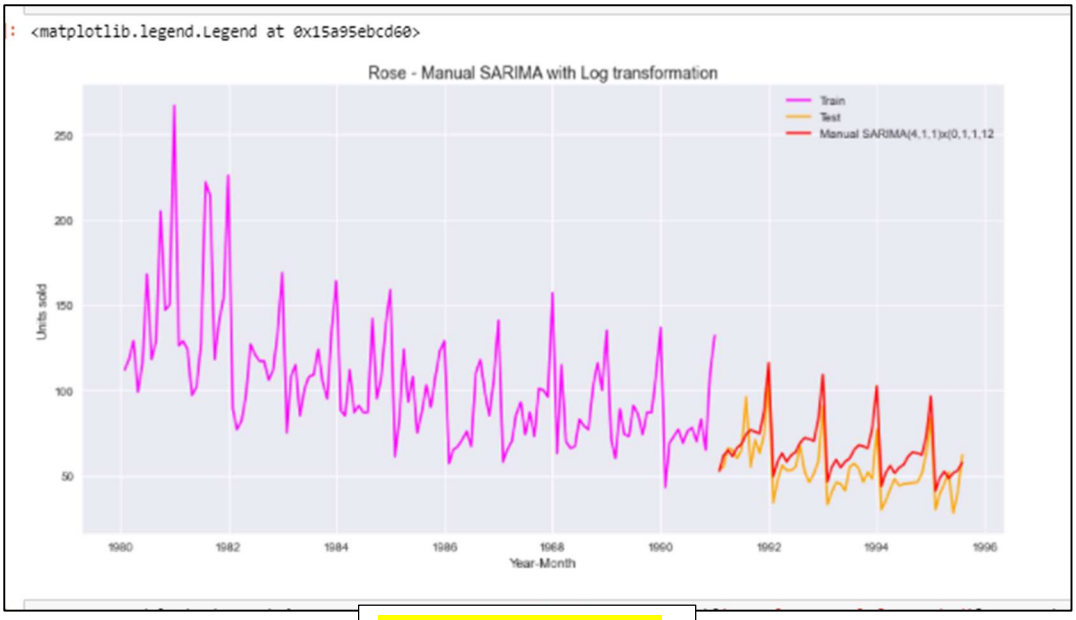
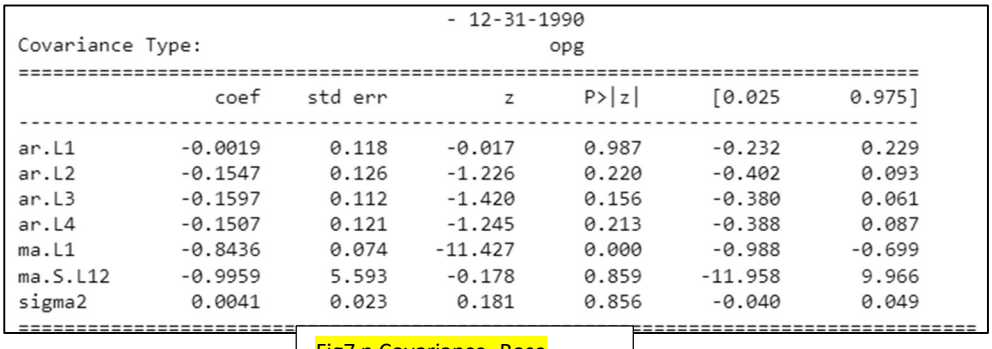


Fig7 l Model forecasted-Rose

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.533374	10.41
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106737	9.48
Auto SARIMA(3,1,3)x(3,1,0,12)	331.586044	10.33
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.795755	11.18
TES Alpha 0.15, Beta 0.00, Gamma 0.37	482.892737	14.95

Fig7 m RMSE and MAPE- Rose

The forecasting check as denoted by the fig 7g is being done on test data.From multiple iterations of SARIMA models below is the comparison of the models in terms of accuracy attributes of RMSE and MAPE.



- 12-31-1990

Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0019	0.118	-0.017	0.987	-0.232	0.229
ar.L2	-0.1547	0.126	-1.226	0.220	-0.402	0.093
ar.L3	-0.1597	0.112	-1.420	0.156	-0.380	0.061
ar.L4	-0.1507	0.121	-1.245	0.213	-0.388	0.087
ma.L1	-0.8436	0.074	-11.427	0.000	-0.988	-0.699
ma.S.L12	-0.9959	5.593	-0.178	0.859	-11.958	9.966
sigma2	0.0041	0.023	0.181	0.856	-0.040	0.049

Fig7 n Covariance- Rose

The model summary indicates that none of the terms used in the model are significant in terms of pvalues.

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

- 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.

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.533374	10.41
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106737	9.48
Auto SARIMA(3,1,3)x(3,1,0,12)	331.586044	10.33
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.795755	11.18
TES Alpha 0.15, Beta 0.00, Gamma 0.37	482.892737	14.95
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
6 point TMA	1283.927428	43.86
SES Alpha 0.00	1316.035487	45.47
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87

Fig 8 a : RMSE or MAPE - Sparkling

- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data.
- The SARIMA and Triple Exponential Smoothing are found to be comparable in terms of Performance and fitment with the test data.

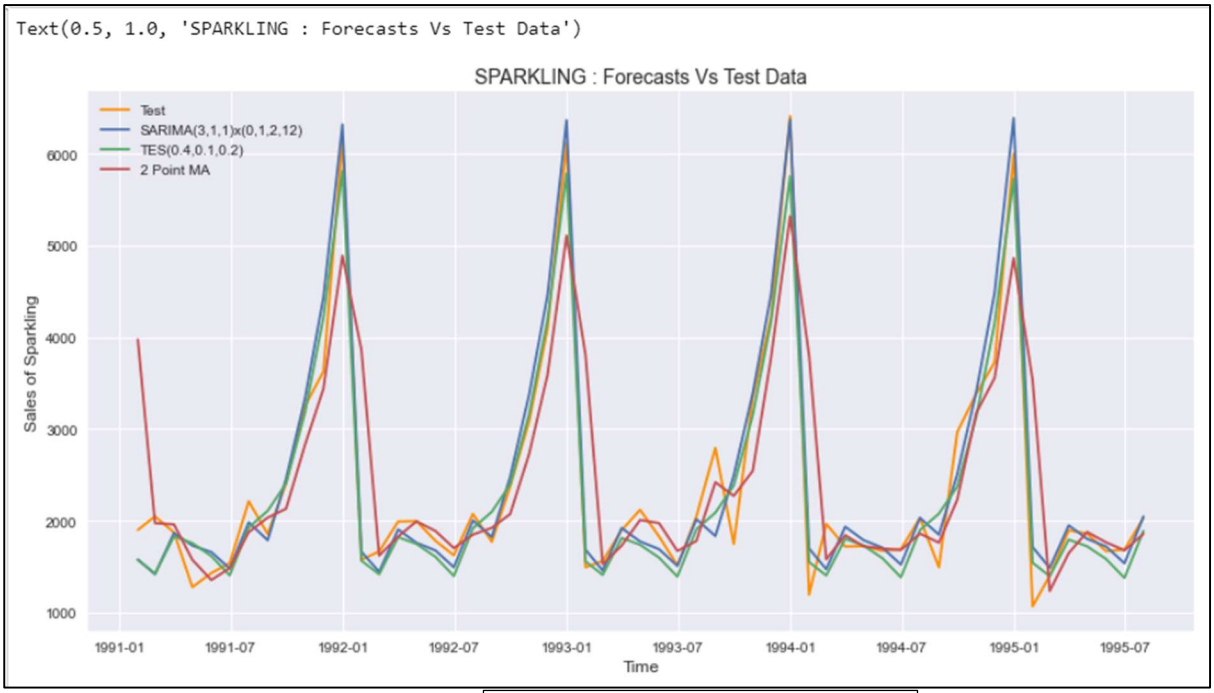


Fig 8 b- All predictions of all models-Plot

- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data.
- The SARIMA and Triple Exponential Smoothing are found to be comparable in terms of Performance and fitment with the test data.

ROSE

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.493832	13.68
2 point TMA	11.529278	13.54
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.590795	21.92
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.177101	23.10
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
Auto SARIMA(3,1,1)x(3,1,1,12)	16.823277	25.48
SES Alpha 0.01	36.796004	63.88
TES Alpha 0.11, Beta 0.05, Gamma 0.00	45.036273	76.86
SimpleAverage	53.460350	94.93
NaiveModel	79.718559	145.10

Fig 8c- All predictions of all models scores

- 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 point 2 Moving Average

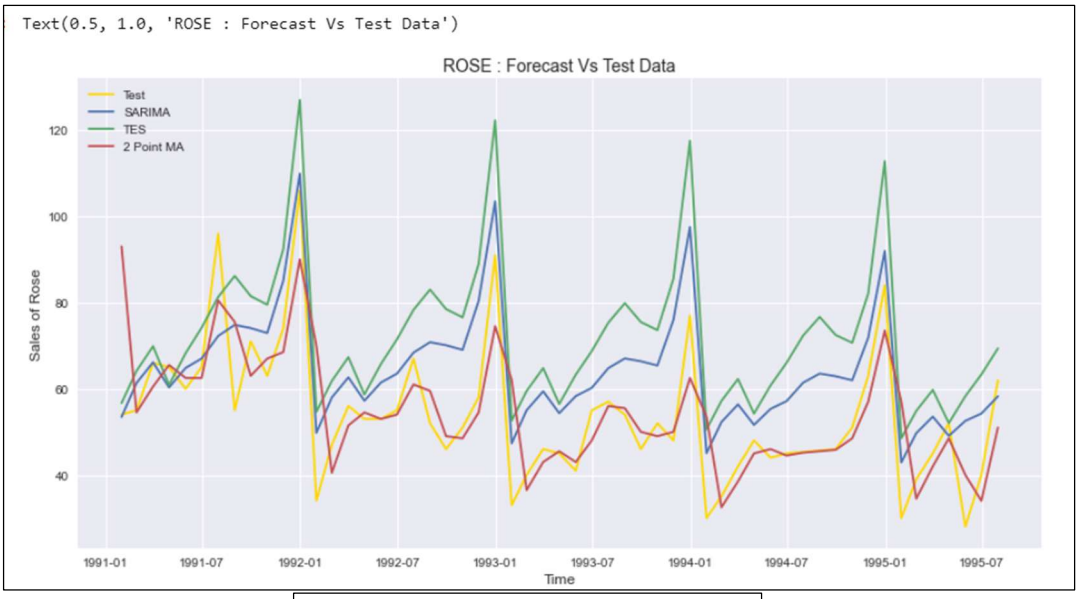


Fig 8d- All predictions of all models scores

- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data.
- 2-point trailing moving average is found to be having the best fitment against the test data, through with a lag 2 and falling short at times.
- Both SARIMA and TES forecasts are a bit higher than the actuals at any given point in time.

9. 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.

- Based on the overall evaluation of the data and comparison. Tripple Exponential Smoothing (Holt Winter's) and SARIMA are selected for final prediction of 12 months.
- TES model alpha:0.4, beta:0.1 and gamma :0.2 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data.
- The model predicts an upward trend and continuation of the seasonal surge in sales in the upcoming 12 months. According to the model the seasonal sale will be more than that of the previous years.
- The 12-month prediction of TES model is as below.

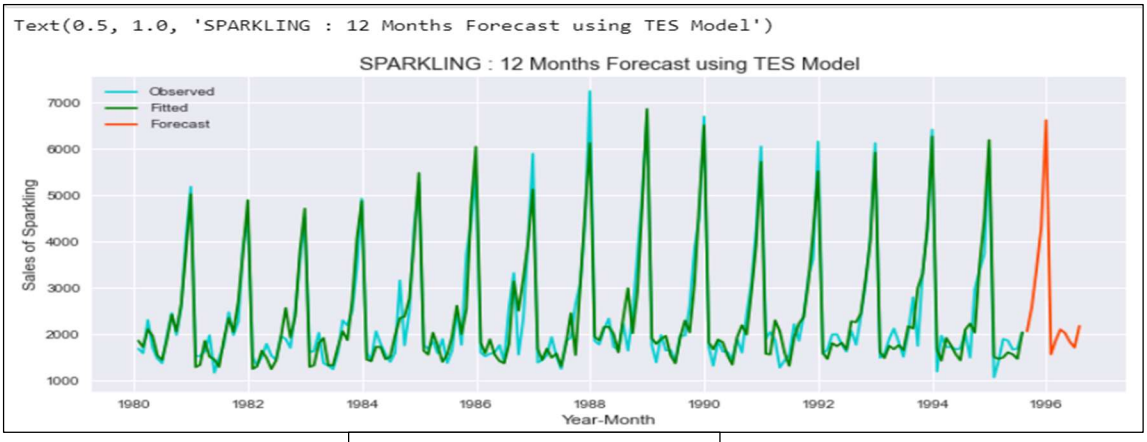


Fig 9a : TES model over the years

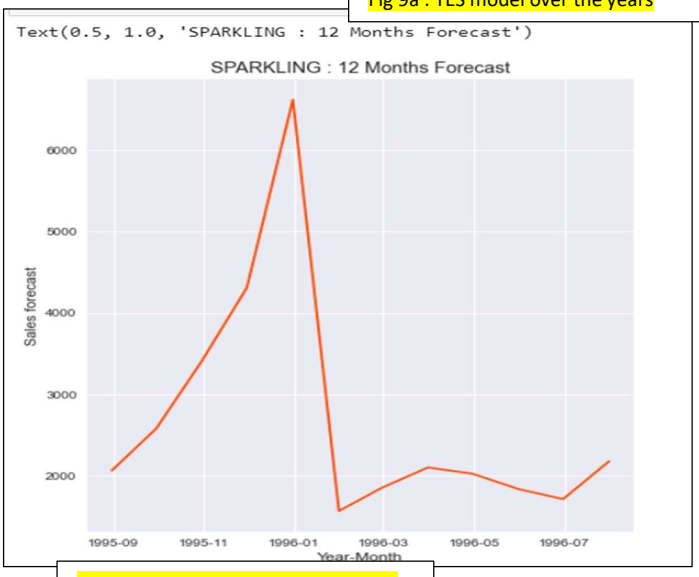


Fig 9a : TES model over the years highest forecasted sales

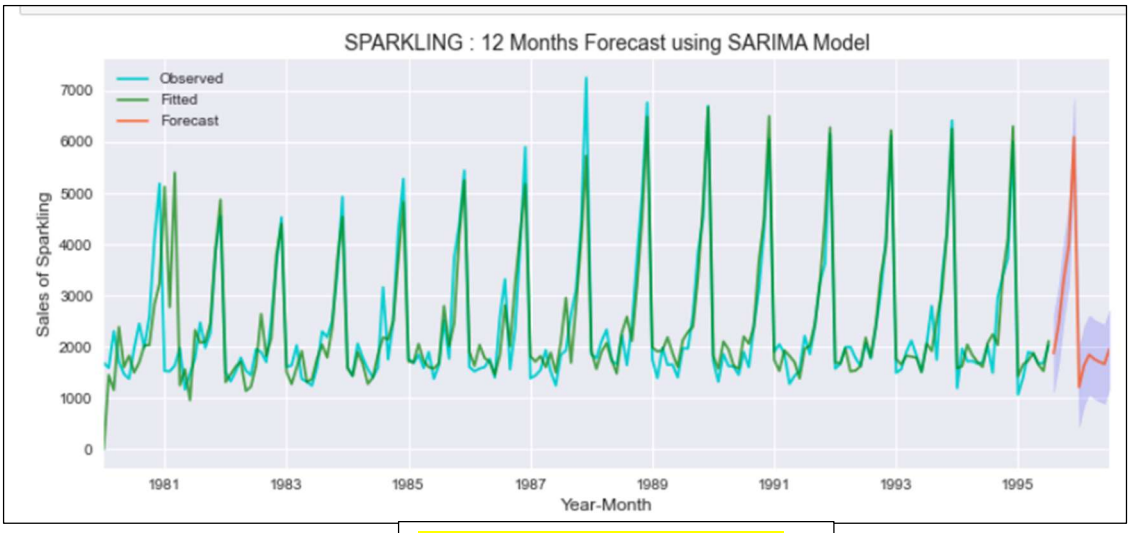


Fig 9b : SARIMA model over the years

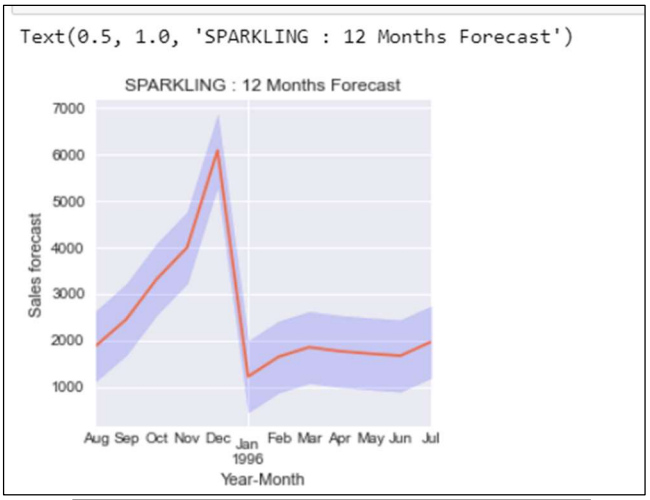


Fig 9b : SARIMA of the maximum sales forecasted

INFERENCE:

- The SARIMA model is built with parameters (3,1,3)x(1,1,2,12), is found to be the most optimal SARIMA model
- SARIMA model has reflected the trend and seasonality of the series continuously into the future as well. The seasonal altitude predicted us more conservative a nature than TES model.
- SARIMA model is seen to have better fitment with the most recent observations of the data and show higher variations in the farthest periods of observations, which explain the high RMSE and MAPE values:

```
TES forecast on the Sparkling Full Data:  RMSE is 377.290 and MAPE is 11.36
```

```
For SARIMA forecast on the Sparkling Full Data:  RMSE is 591.255 and MAPE is 14.86
```

ROSE:

- Based on the overall evaluation of the data and comparison. Triple Exponential Smoothing (Holt Winter's) and SARIMA are selected for final prediction of 12 months.
- TES model alpha:0.1, beta:0.1 and gamma :0.2 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data.
- The model predicts a combination 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.

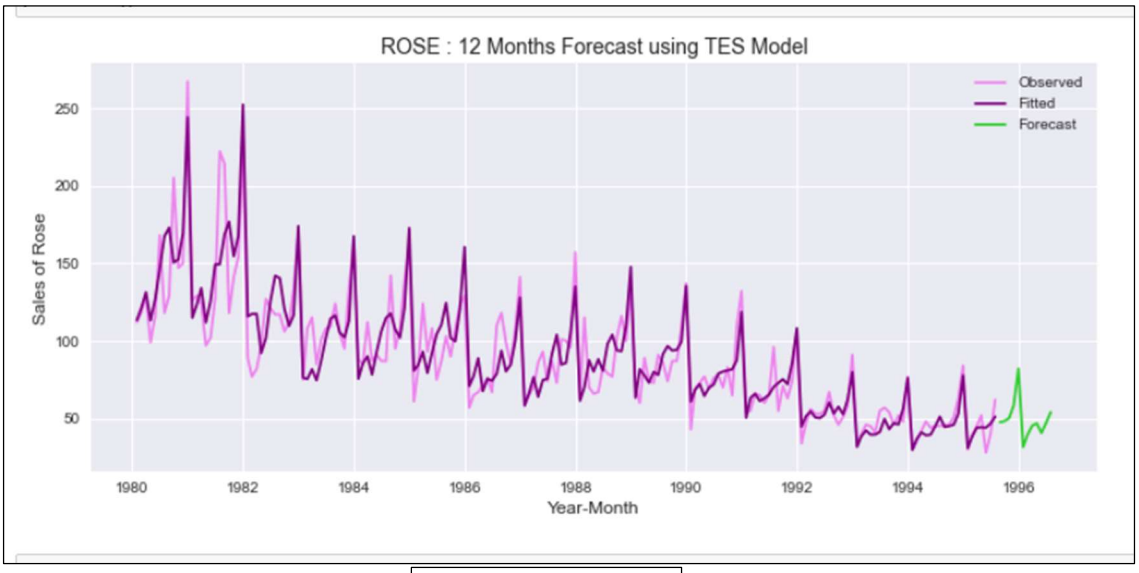


Fig 9c : TES forecasted

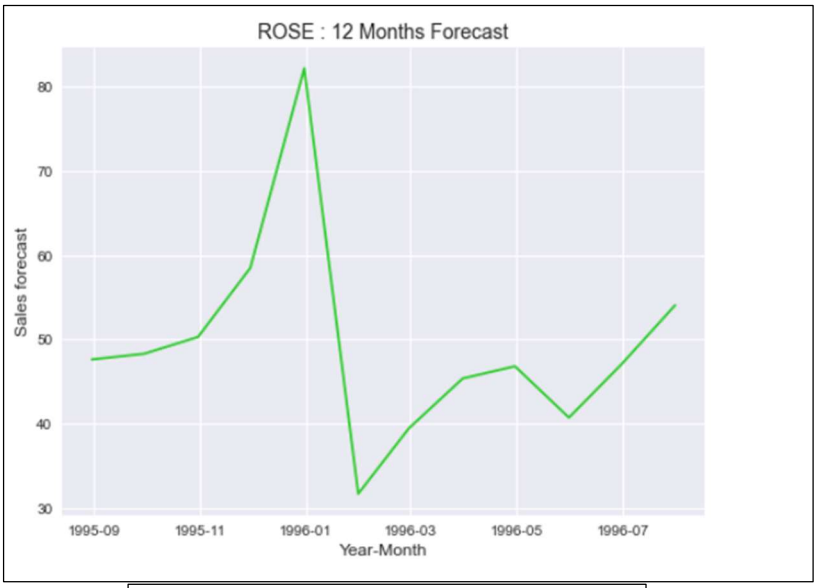


Fig 9c : TES of the maximum sales forecasted

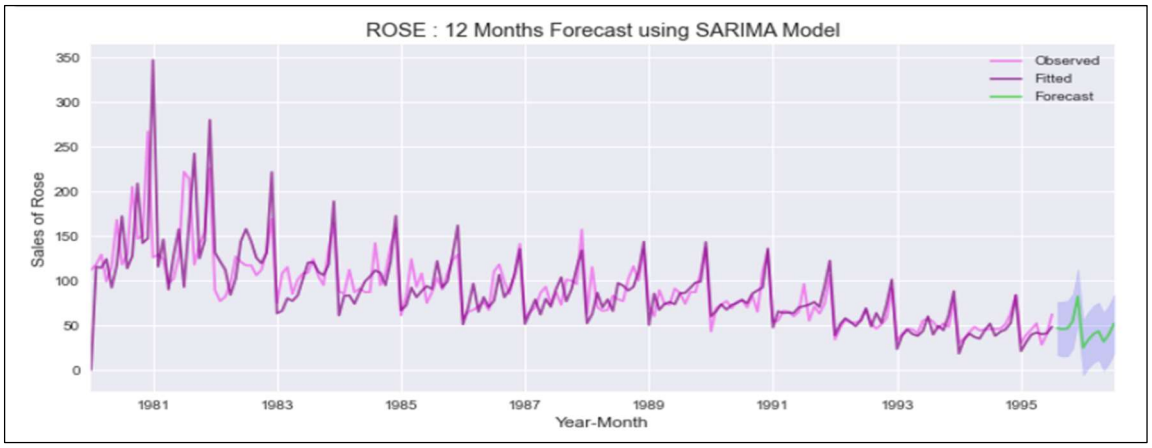


Fig 9 d : SARIMA of the maximum sales forecasted

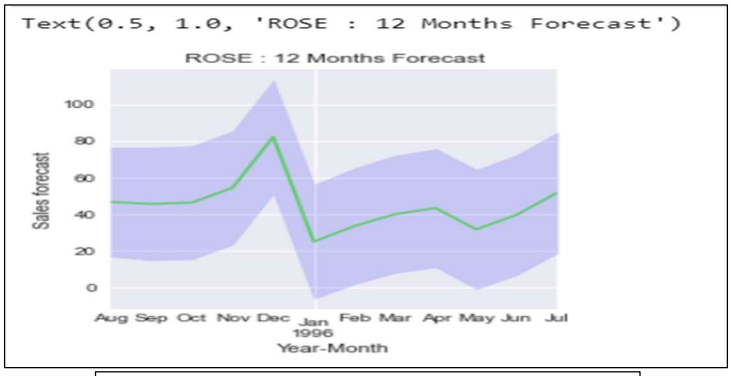


Fig 9 d: SARIMA of the maximum sales forecasted

INFERENCE:

- 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 entire time series.
- SARIMA model has reflected the trend and seasonality of the series continuing into the future as well.
- SARIMA model is seen to have better fitment with the most recent observed data and shows high variations in the farthest periods of observations, which explain the high RMSE and MAPE values:

```
TES forecast on the Rose Full Data:  RMSE is 17.404 and MAPE is 13.87
```

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

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

- The SARIMA model built on the complete Sparkling timeseries is chosen as prediction provide confidence interval which give better explainability and confidence to the forecasts.
- The diagnostics plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in normal QQ plot.
- The model summary also provides valuable insights in the model. From the snapshot of summary below it can be understood that AR (2), MA (3) terms has the highest absolute weightage. The p-value indicates that the terms AR (1), AR (2), MA (1), MA (2) and MA (3) are the most significant terms.
- The rest of the pvalues got values higher than alpha 0.05 which fails to reject the null hypothesis that these terms are not significant.

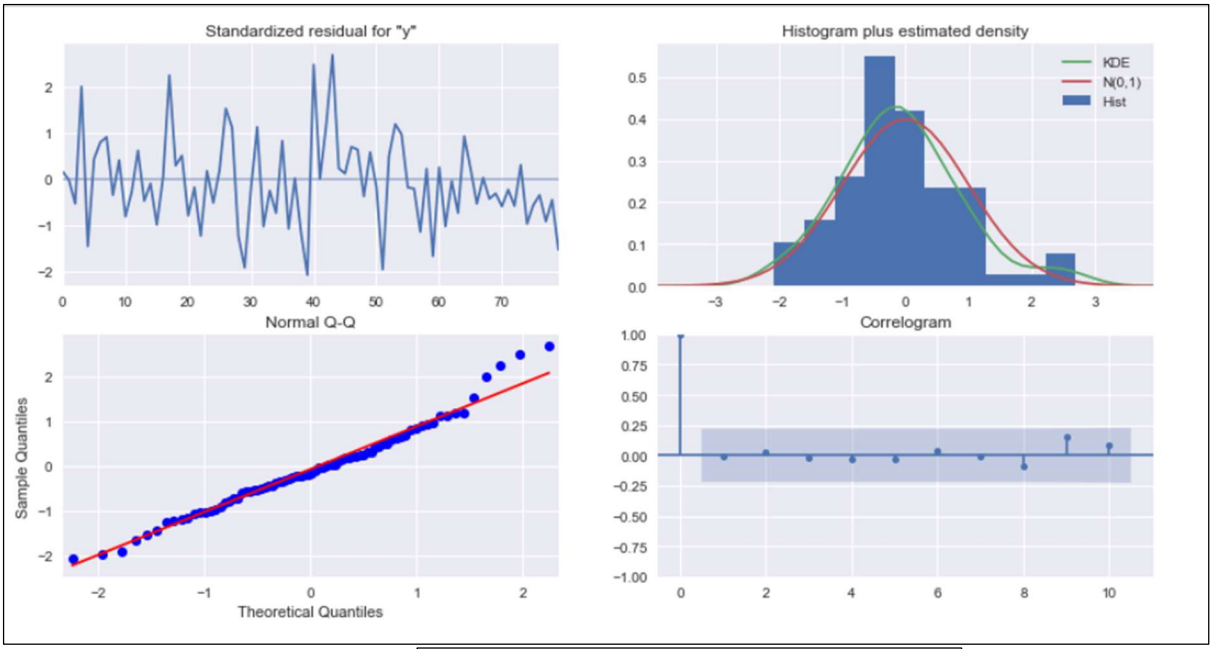


Fig 10a: Residual of the Forecasted best model- Sparkling

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                  187
Model:             SARIMAX(3, 1, 3)x(1, 1, [1, 2], 12)   Log Likelihood           -1078.437
Date:                                 Sun, 20 Feb 2022   AIC                       2176.875
Time:                                        16:26:20   BIC                       2206.711
Sample:                                  01-31-1980   HQIC                      2188.998
                                       - 07-31-1995
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4230      0.086     -4.914      0.000      -0.592      -0.254
ar.L2         -0.9094      0.053    -17.290      0.000      -1.013      -0.806
ar.L3          0.1424      0.087      1.637      0.102      -0.028       0.313
ma.L1         -0.4113      0.078     -5.270      0.000      -0.564      -0.258
ma.L2          0.4622      0.083      5.574      0.000       0.300       0.625
ma.L3         -0.9673      0.104     -9.307      0.000      -1.171      -0.764
ar.S.L12      -0.0698      0.710     -0.098      0.922      -1.461       1.322
ma.S.L12      -0.4551      0.722     -0.630      0.528      -1.870       0.960
ma.S.L24      -0.0808      0.397     -0.204      0.839      -0.859       0.697
sigma2     1.461e+05   1.06e-06   1.37e+11      0.000    1.46e+05    1.46e+05
==============================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                35.58
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               0.72   Skew:                             0.66
Prob(H) (two-sided):                  0.26   Kurtosis:                         5.03
==============================================================================

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

Fig 10 b: The best model for

ROSE

- The SARIMA model built on the complete ROSE timeseries is chosen as prediction provide confidence interval which give better explainability and confidence to the forecasts.
- The diagnostics plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in normal QQ plot.
- The model summary also provides valuable insights in the model. From the snapshot of summary below it can be understood that AR(2), MA(3) terms has the highest absolute weightage. The p-value indicates that the terms AR(1),AR(2) , MA(1), MA(2) and MA(3) are the most significant terms.
- The rest of the pvalues got values higher than alpha 0.05 which fails to reject the null hypothesis that these terms are not significant.
- Prediction on the ROSE time-series is on a wider confidence band than sparkling.

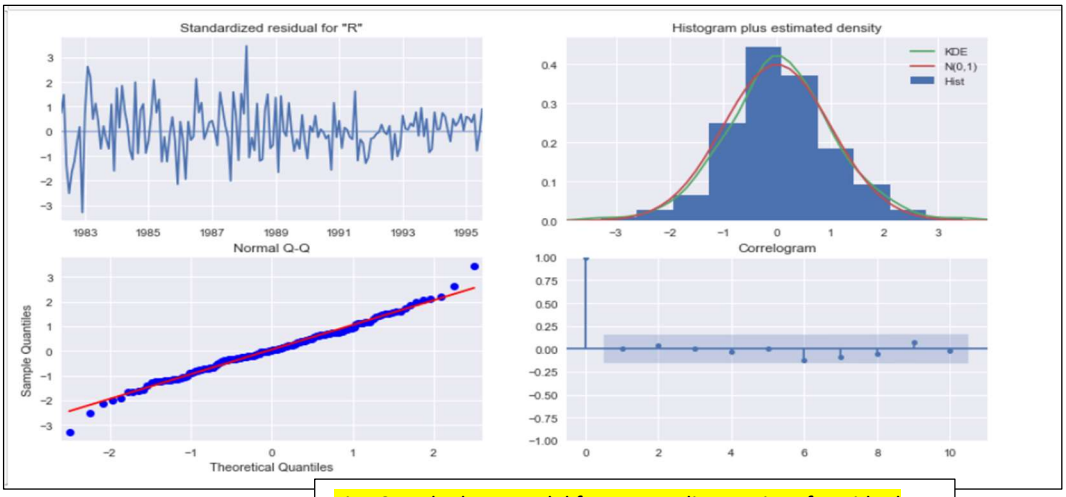
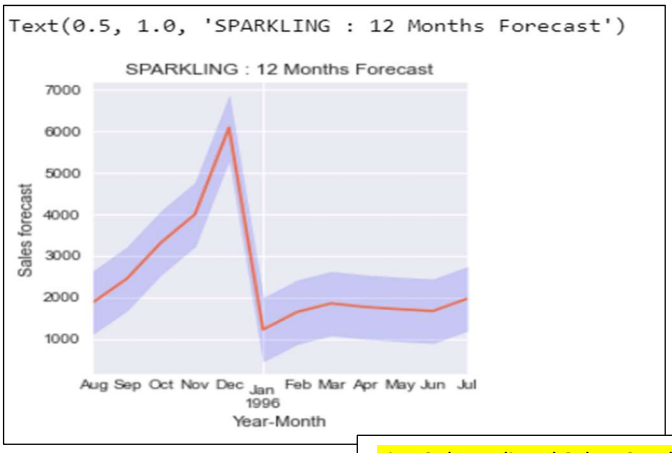


Fig 10 c: The best model for Rose – diagnostics of Residual

future sales:

SPARKLING:



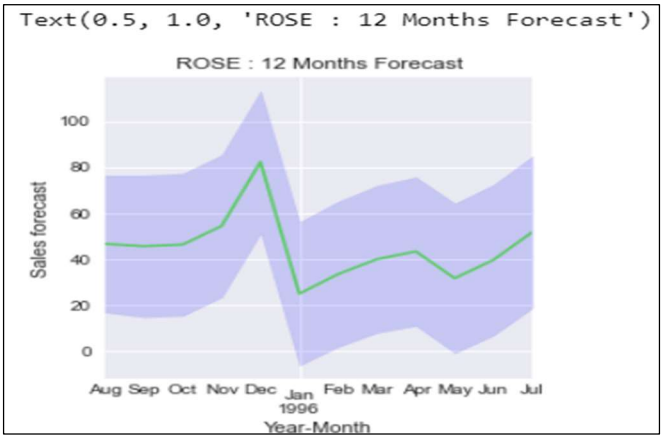
	Sparkling
31-08-1995	1873.39
30-09-1995	2444.94
31-10-1995	3312.6
30-11-1995	3994.61
31-12-1995	6084.08
31-01-1996	1215.91
29-02-1996	1640.57
31-03-1996	1847.28
30-04-1996	1761.99
31-05-1996	1708.3
30-06-1996	1664.02
31-07-1996	1961.28
Total Sales	29508.97

Fig 10 d Predicted Sales- Sparkling

INFERENCE:

- The model forecasts a sale of 29510 units of Sparkling wine in 12 months into future. Which is an average sale of 2459 units per month.
- The seasonal sale in December 1995 will hit a maximum of 6084 units before it drops to the lowest sale in January 1996 at 1216 units.
- The wine company is recommended to ramp up their procurement and production line in accordance with the above forecasts for the third quarter of 1995(October, November and December), which is a total of 13,392 units of sparkling wine is expected to be sold.
- The forecast also indicates that the year-on- year sales of sparkling wine is not showing an upward trend. The winery must adopt innovative marketing skills to improve the sale compared to previous years.
- Adding more exogenous variable into the time series data can improve forecasts

ROSE:



	ROSE
31-08-1995	46.54
30-09-1995	45.51
31-10-1995	46.23
30-11-1995	54.32
31-12-1995	82.21
31-01-1996	24.81
29-02-1996	33.35
31-03-1996	39.87
30-04-1996	43.23
31-05-1996	31.53
30-06-1996	39.56
31-07-1996	51.7
Total	538.86

Fig 10 e Predicted Sales- ROSE

INFERENCE:

- The model forecasts a sale of 593 units of Sparkling wine in 12 months into future. Which is an average sale of 45 units per month.
- The seasonal sale in December 1995 will hit a maximum of 82 units before it drops to the lowest sale in January 1996 at 25 units.
- Unlike Sparkling wine, Rose wine sells very low number of units and the Standard deviation is only 14.5, which means that higher demand does not impact production or procurement.
- Apart from higher sale in November and December months, Rosé sales will be above average in the summer months of July and August.
- The winery should investigate on the low demand for Rose wines in the market and take corrective actions in promotions and marketing.

In between Sparkling and Rose, Sparkling has more sales and production compared to Rose. Company should try to know more of their customer preferences and induce them in ROSE to increase sales.

Apart from marketing the wines quality should also be improved if that would help sales boost. So a research is needed for ROSE WINE thoroughly in terms of customer preferences, marketing and cost.

……THE END…….