INDEX

List of Figures:

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

Brief List of Actions:

1. Read the data as an appropriate Time Series data and plot the data.
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.
   Note: Stationarity should be checked at alpha = 0.05.
6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.
8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

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

We have been assigned the task of analysing and forecasting the sales of 20th century wine for the company named ABC Estate Wines.

Loading the Data & Plotting the Appropriate Time Series:

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

Fig 1: Sample Data "Rose"

Dataset is loaded and read successfully in jupyter file.

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

Fig 2: Time Stamp "Rose"

Formulised time stamp for the given data successfully. This time stamp is created as per the time data available in the data given.

Plotting the Time Series:

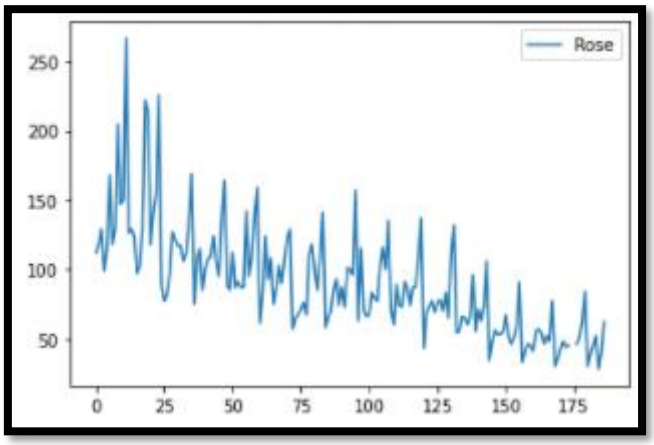

Fig 3: Time Series Plot "Rose"

Time Series Stamping with the Data Given:

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0
...	...
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

187 rows × 1 columns

Fig 4: Time Series Stamping "Rose"

The graph plotted after converting months year to time stamping is showing years on the X axis now.

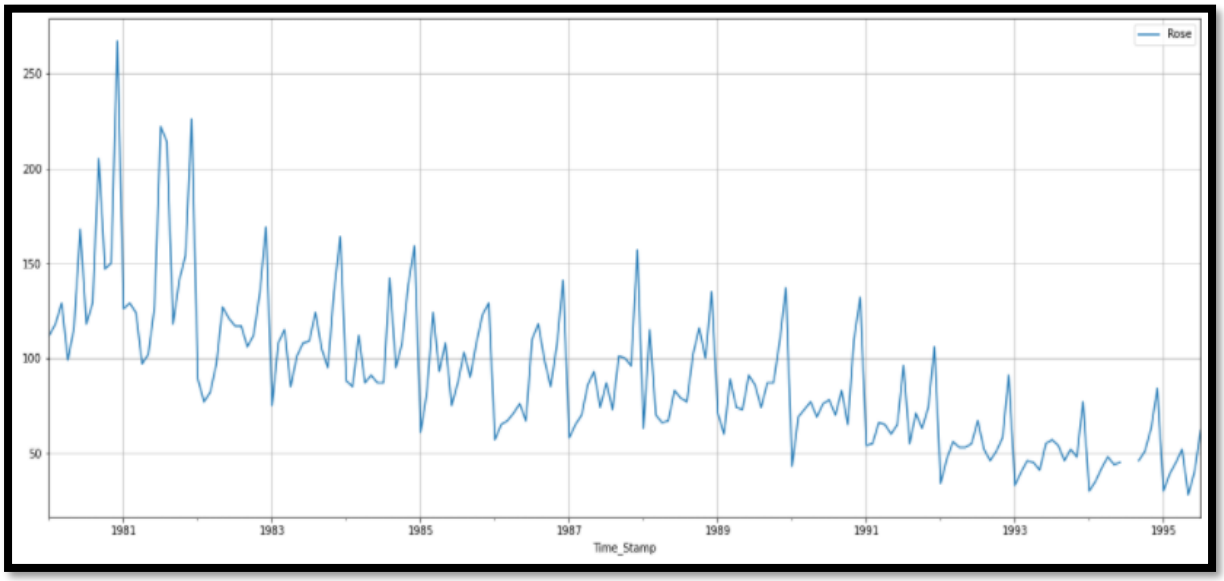

Fig 5: Time Series Plot of "Rose" after Time Series Stamping

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

Trend and seasonality: We notice that there is clear trend in data. From 1980 to 1988 there is upward rise in sales, after that sales falls from 1990 to 1996. Seasonality is also clearly visible from the intra year stable fluctuations repeating over the entire length of series.

Description of entire data: shows that maximum sales is of 267 units and minimum sales is of 28 units. average sales through series are 90 units. and in total we have 185 months for which this data is speaking.

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

Fig 6: Descriptive Data of "Rose"

Missing Values: there are 2 missing values in data as clearly visible from the time series graph that trend lines are continuous over 185 months without any break of line in between. We have used interpolation method to impute the data. (Please refer python file)

Box plot:
Yearly Box plot: clearly indicates the trends as outliers are present in data.

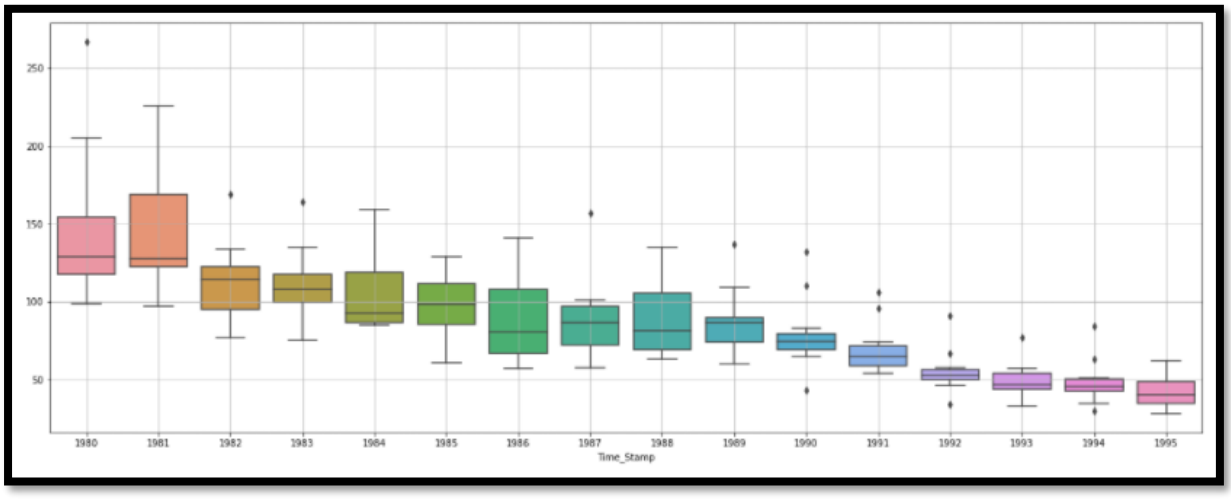

Fig 7: Yearly Box-Plot of "Rose"

Monthly Box-Plot: clearly indicates the trends is present in months as outliers are present 6th ,7th, 8th & 12th months. Max Sales seems to be in December Month.

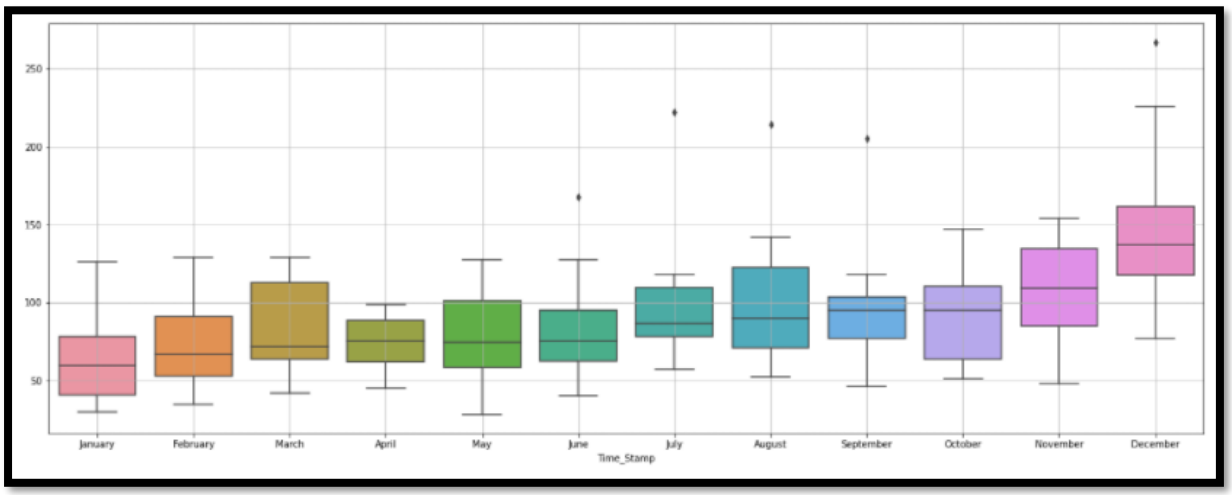

Fig 8: Monthly Box-Plot of "Rose"

Month plot of time series: The month plot shown below represents in terms of red lines the average sale of unit for particular month and black lines are actual number of units sold.

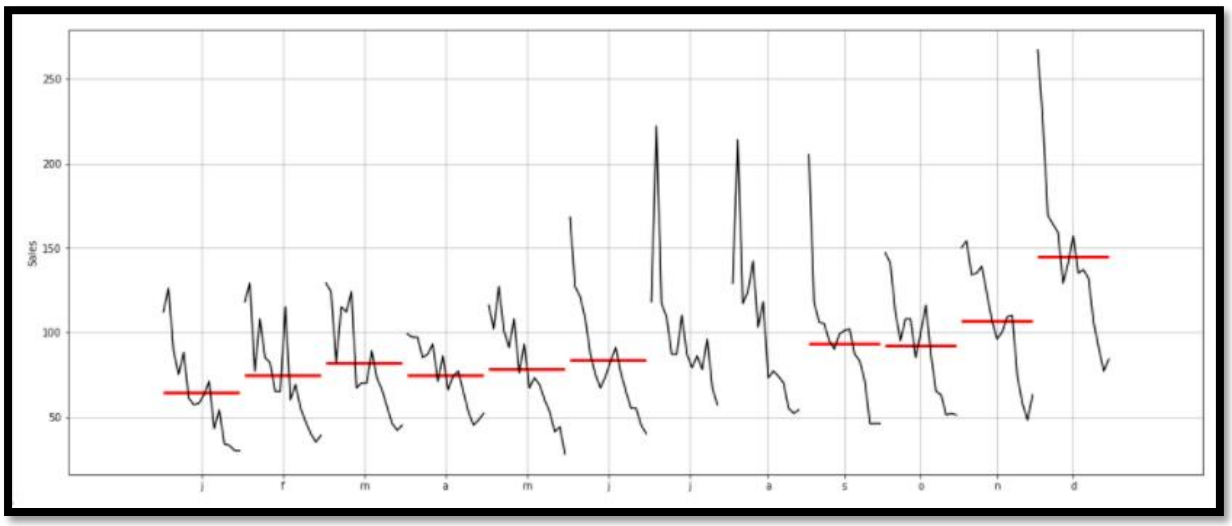

Fig 9: Month Plot of Time Series

Summarised table month wise for each year showing sales of Rose wine. Average sales are quite low from January to June months from July onwards it increases and reaches maximum in December.

Year and month number chart is made to understand how sales is behaving for each month of a particular year. Its graphical presentation is easy to interpret what is happening in data.

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

Fig 10: Numerical Presentation of Year-Month Plot

To understand this data, we have plotted this on time series graph for various months for all years.

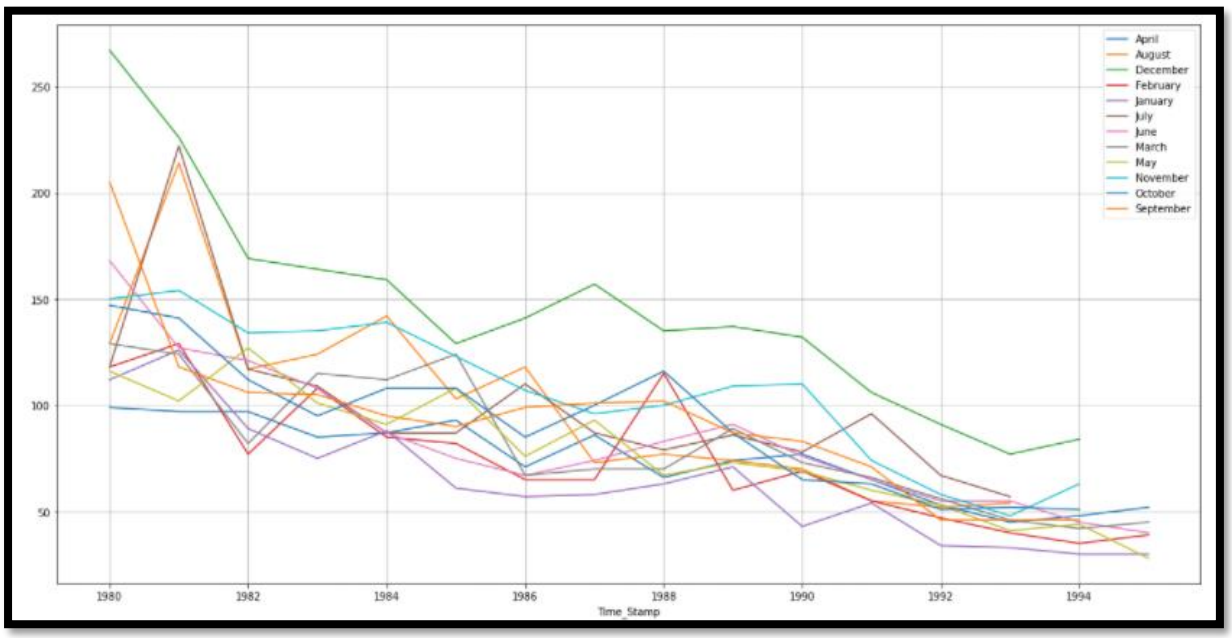

Fig 11: Year-Month Plot

December records the highest sale followed by August and November respectively. May, June and July months record the lowest sales across all the years.

To have a better idea and understanding, we must decompose the data into **seasonality**, **trend** and **residuals**.

Data Decomposition:

Decomposition using Additive Model:

As per the 'additive' decomposition, we see that there is a decreasing trend in the earlier years of the data. After 1984 sales raised till 1986 after which it further dropped from 1990 onwards.

Highest sales were achieved in year 1988. Seasonality is also clearly visible from the seasonal graph where trend lines are forming the peaks with different height every year. Residuals seems to be scattered from the 0 level, indicating that the series is not additive.

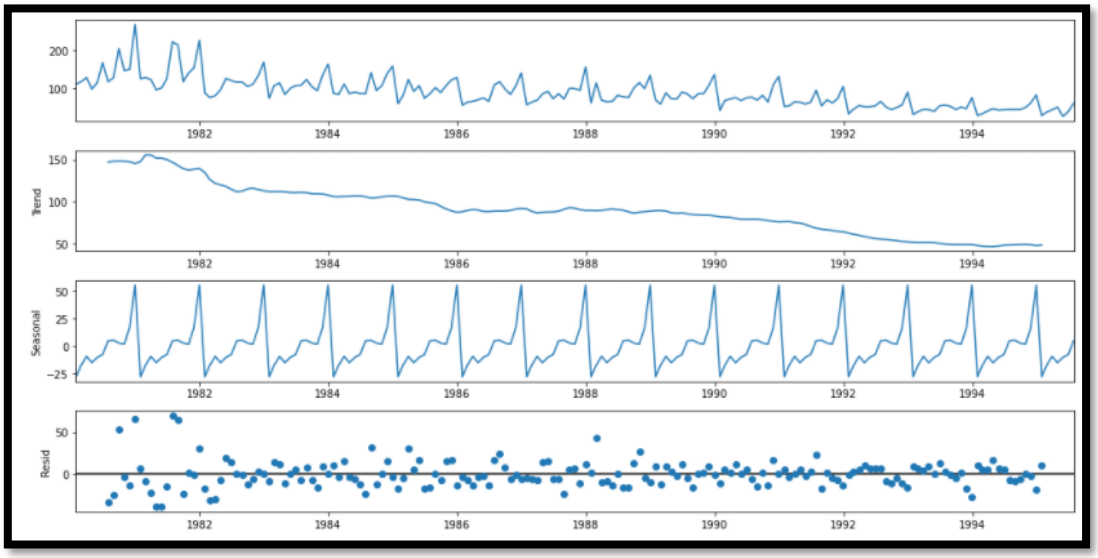

Fig 12: Data Decomposition Using Additive Model

Decomposition Using Multiplicative Model:

The trend and seasonality are same as in the case of decomposition using additive model. But residuals plot is clearly showing the concentration of data towards point 1. Hence it can be concluded that series is multiplicative.

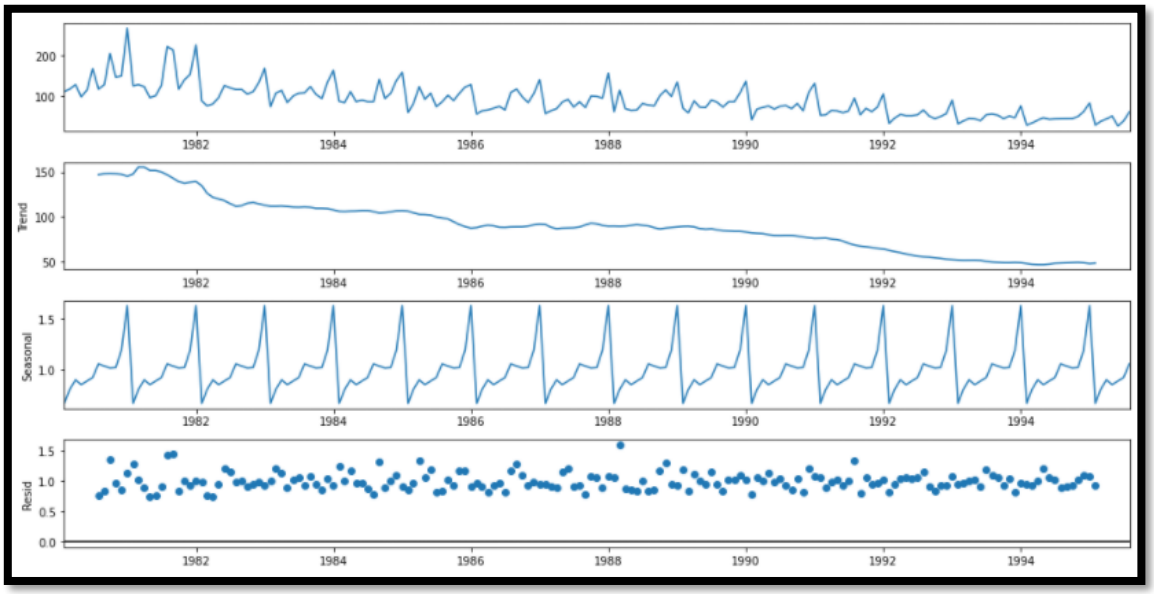

Fig 13: Data Decomposition Using Multiplicative Model

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

The data which is given is from the year 1980 to 1995. The required data for bifurcation is from 1991 into train and test data, which approximately is 71% or 0.71 of the whole data division.

Below are the images of some first and last few rows of the train data.

First few rows of Training Data

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

Last few rows of Training Data

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

Fig 14: Sample of Train Data

Below are images of some first and last few rows of the test data.

First few rows of Test Data

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

Last few rows of Test Data

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

Fig 15: Sample of Test Data

From the images above, it is clearly observable that the train data is starting from 1980 and contains the record till the end of 1990 whereas the test data contains the data from 1991 till the end of 1995.

Hereby, if we want to conclude the splitting of the data in train and test in numerical terms then train contains the 71% (~70%) of the data while the 29% (~30%) is given under test data.

Note: The above division of the data is done on the basis of the data asked in the question.

Plotting the Split of Train & Test Data:

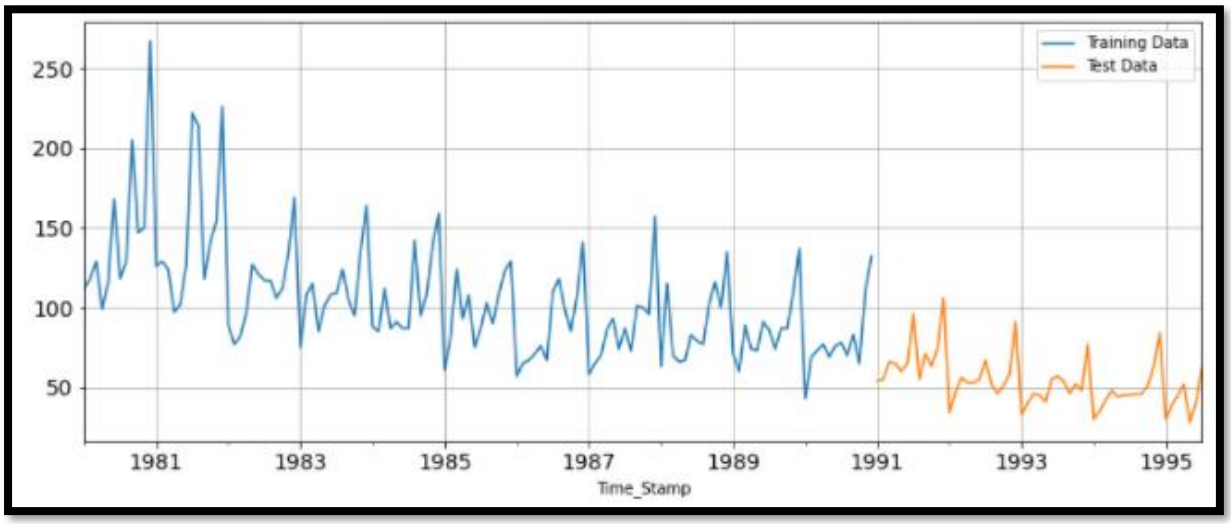

Fig 16: Train & Test Data Plot

The graph shows the training data is well separated from the test data. Blue line is showing test data from 1980 till 1990 and from 1991, test data is shown in orange.

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

Model 1: Linear Regression:

Model For this particular linear regression, we are going to regress the 'monthly sales of Rose wine' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression. Training and Testing data time instances are generated as below and are now ready for adding in the training and testing data set.

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

Fig 17: Training-Testing Data Time Instances

Below is the sample of the first and last few rows of the training and testing data.

Training Sample Data:

First few rows of Training Data

Time_Stamp	Rose	time
1980-01-31	112.0	1
1980-02-29	118.0	2
1980-03-31	129.0	3
1980-04-30	99.0	4
1980-05-31	116.0	5

Last few rows of Training Data

Time_Stamp	Rose	time
1990-08-31	70.0	128
1990-09-30	83.0	129
1990-10-31	65.0	130
1990-11-30	110.0	131
1990-12-31	132.0	132

Fig 18: Training Sample Data

Testing Sample Data:

First few rows of Test Data

Time_Stamp	Rose	time
1991-01-31	54.0	133
1991-02-28	55.0	134
1991-03-31	66.0	135
1991-04-30	65.0	136
1991-05-31	60.0	137

Last few rows of Test Data

Time_Stamp	Rose	time
1995-03-31	45.0	183
1995-04-30	52.0	184
1995-05-31	28.0	185
1995-06-30	40.0	186
1995-07-31	62.0	187

Fig 19: Testing Sample Data

Now we will build the model using sklearn liner model Linear Regression and fit it on training data to have predictions on test data.
The graph below shows the Linear Regression Line represented in green on the test data showing only the trend as Linear Regression cannot take care of seasonality component at all.
Trend line shows a little upwards trend transition in sales of Rose wine for year 1991 units sold is around 65 which can fell to 50 units by the end of 1995.

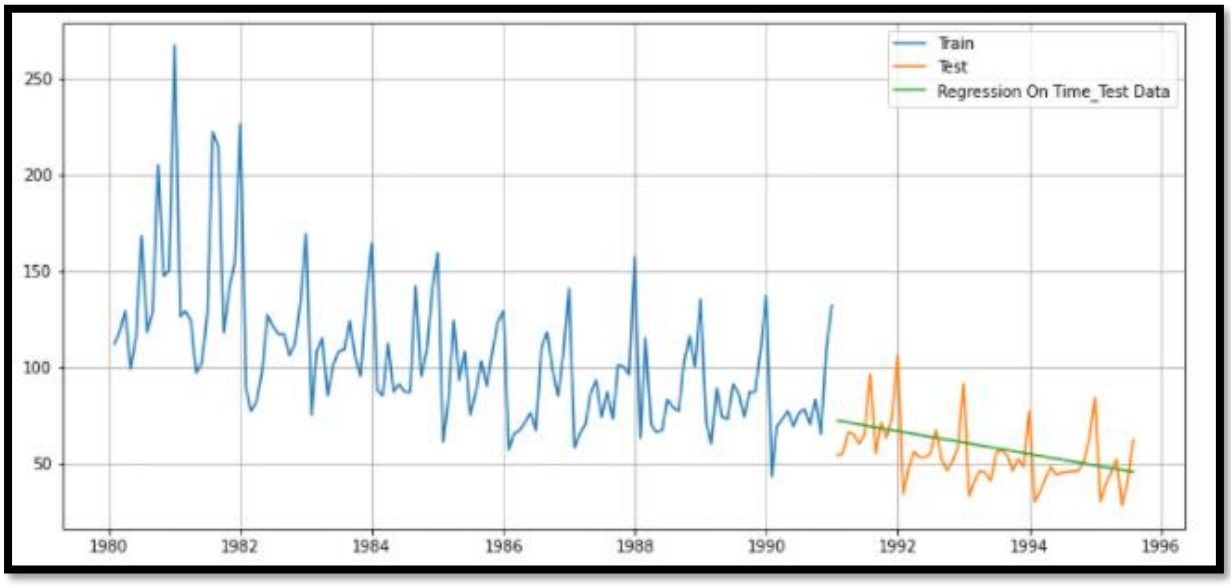

Fig 20: Plot of Train & Test Data after adding Time Instances

Model Evaluation:

```
For RegressionOnTime forecast on the Test Data,  RMSE is 15.269
```

Fig 21: RMSE of Test Data for Model Evaluation

The value of RMSE (15.269) is quite high and since seasonality is also not taken care by model, therefore, this model is not suitable predictions on Rose time series data.

Model 2: Naïve Approach $\hat{y}t+1=yt$:

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.
The graph below shows in green predictions done on test data. A straight green line indicated that model has forecasted the 1990 sales of 6000 units for years 1991 till 1996.Seasonality is not taken care in this model.

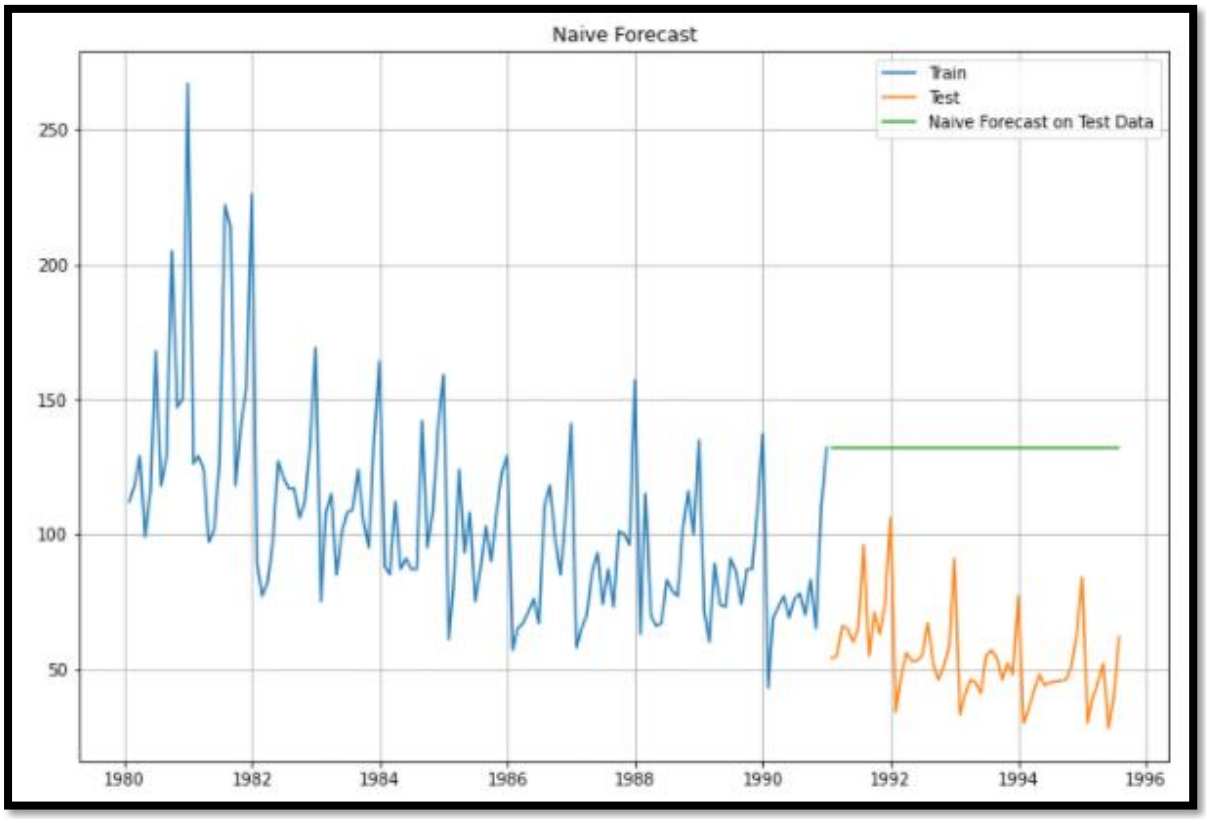

Fig 22: Naïve Bayes Based Time Series Plot (Test Data)

Model Evaluation:

```
For RegressionOnTime forecast on the Test Data,  RMSE is 79.719
```

Fig 23: RMSE of Test Data for Model Evaluation)

RSME is 79.719, which is higher than the linear regression model. Thus, this model being too simple. Model with very high RSME is not good model at all. Values predicted using this model will be significantly away from the actual data points in test data set.

Model 3: Simple Average:

Using the simple average of data to predict the values on test data will definitely not take care of seasonality and trends at all and will produce a flat line predicting same number of units sold of wine for each month for every year.
So is happening in our case. Predicted values are 105 which is equal to the mean of the data as prevalent in EDA part showing data description.

Time_Stamp	Rose	mean_forecast
1991-01-31	54.0	104.939394
1991-02-28	55.0	104.939394
1991-03-31	66.0	104.939394
1991-04-30	65.0	104.939394
1991-05-31	60.0	104.939394

Fig 24: Sample of Predicted Value

The Graph shows the green line shows the predicted values based on simple average which are constant at number of units sold as 104.

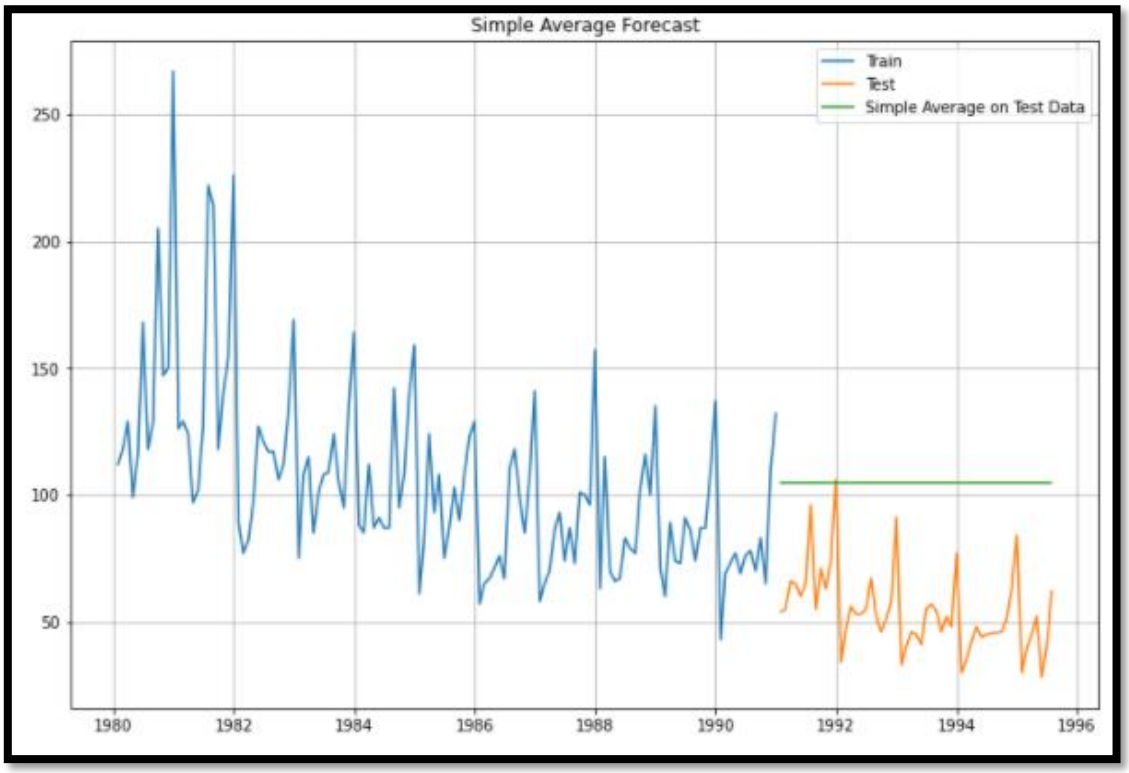

Fig 25: Simple Average Based Time Series Plot (Test Data)

Model Evaluation:

RMSE 53.461 is less than Naïve model but higher than linear regression and without seasonality component in it we cannot consider this model for forecasting purposes.

```
For Simple Average forecast on the Test Data,  RMSE is 53.461
```

Fig 26: RMSE of Test Data for Model Evaluation

Model 4: Moving Averages:

Considering moving averages as 2,4,6 and 9 data is plotted to see which moving average is nearest to the actual data set. Moving average 2 shown in orange colour trend line is nearest fit to actual data. Hence it will predictions should be made using 2 as moving average.

Time_Stamp	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.5	NaN	NaN
1980-05-31	116.0	107.5	115.5	NaN	NaN

Fig 27: Moving Average Sample Data

Moving Average Plot:

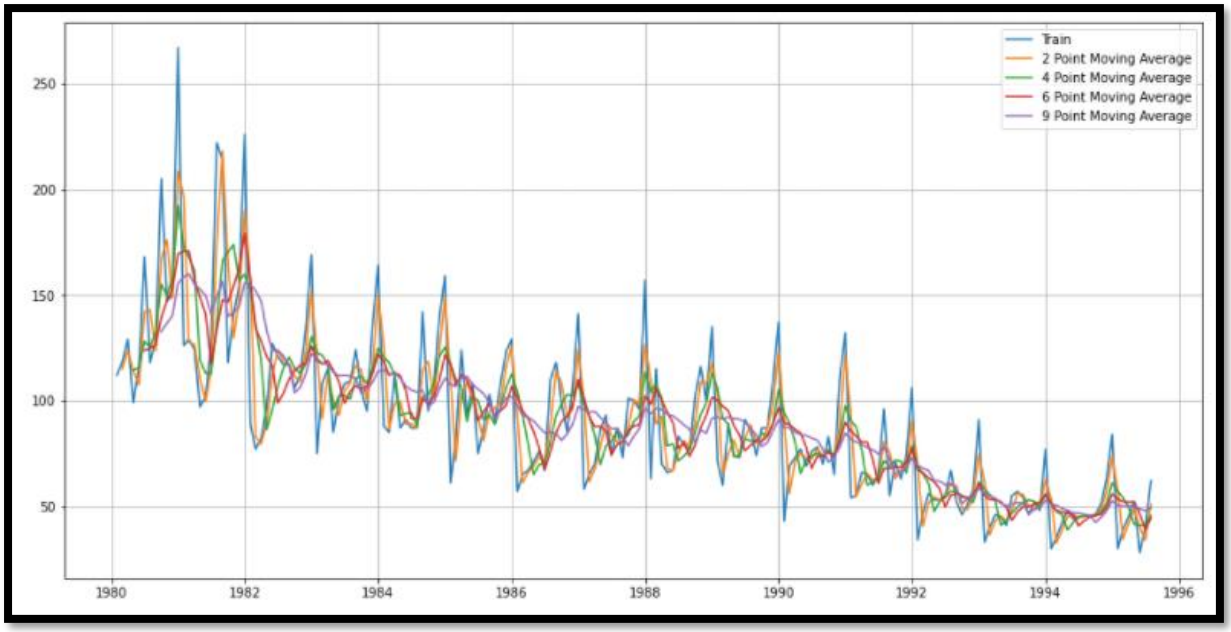

Fig 28: Moving Average Plot (Train Data)

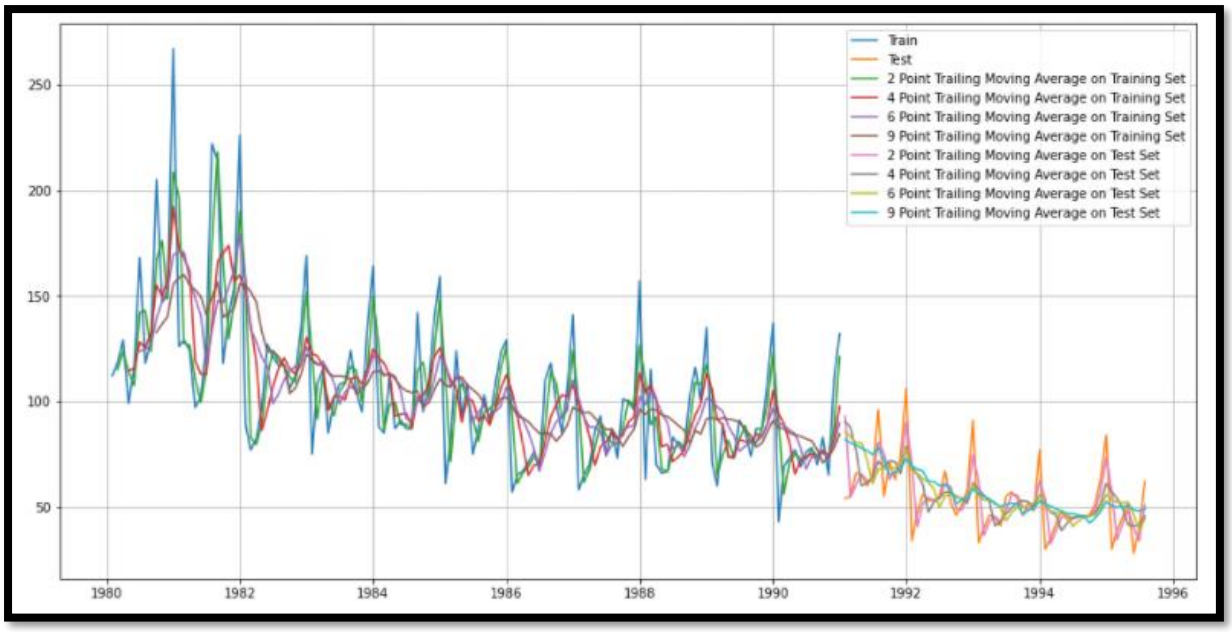

Fig 29: Moving Average Plot (Train & Test Data)

On the testing data set having the orange colour trend line the predicted values in red colour trend line fits the most which is for the 2-point trailing moving average.

Model Evaluation:

RMSE for all these moving averages is as below, lowest is for 2-point moving average.

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.529
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.451
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.566
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.728
```

Fig 30: RMSE for Model Evaluation (Test Data)

Model Comparison:

Till now for Linear regression, Naïve Bayes, Simple average, moving average best performing model with lowest RMSE is 2 point moving average.
Comparison plot shows the best fit model in brown colour line for 2nd moving average appropriately fitting on the actual test values.

LR → Linear Regression
NB → Naïve Bayes
SA → Simple Average
MA → Moving Average

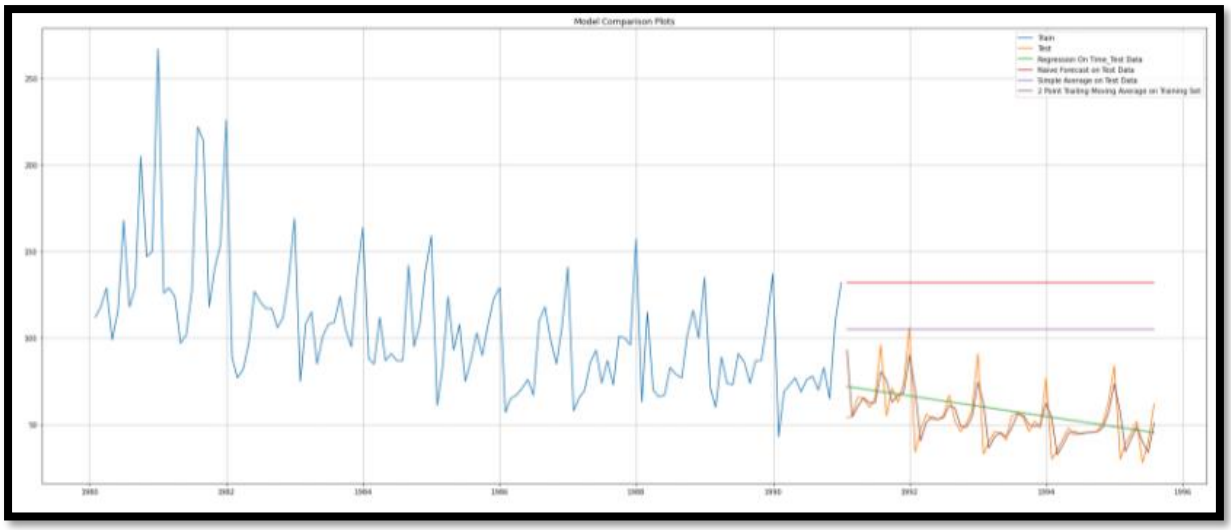

Fig 31: LR, NB, SA, MA Comparison Plot

Model 5: Simple Exponential Smoothing:

This model is build using the parmas as mentioned below.

```
{'smoothing_level': 0.09874989825614361,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38702255613862,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig 32: Parameters Used for Simple Exponential Smoothing Model

The number chart and graph of values predicted on basis of this model is as shown below:

Time_Stamp	Rose	predict
1991-01-31	54.0	87.104999
1991-02-28	55.0	87.104999
1991-03-31	66.0	87.104999
1991-04-30	65.0	87.104999
1991-05-31	60.0	87.104999

Fig 33: Predicted Value on the basis of Simple Exponential Smoothing

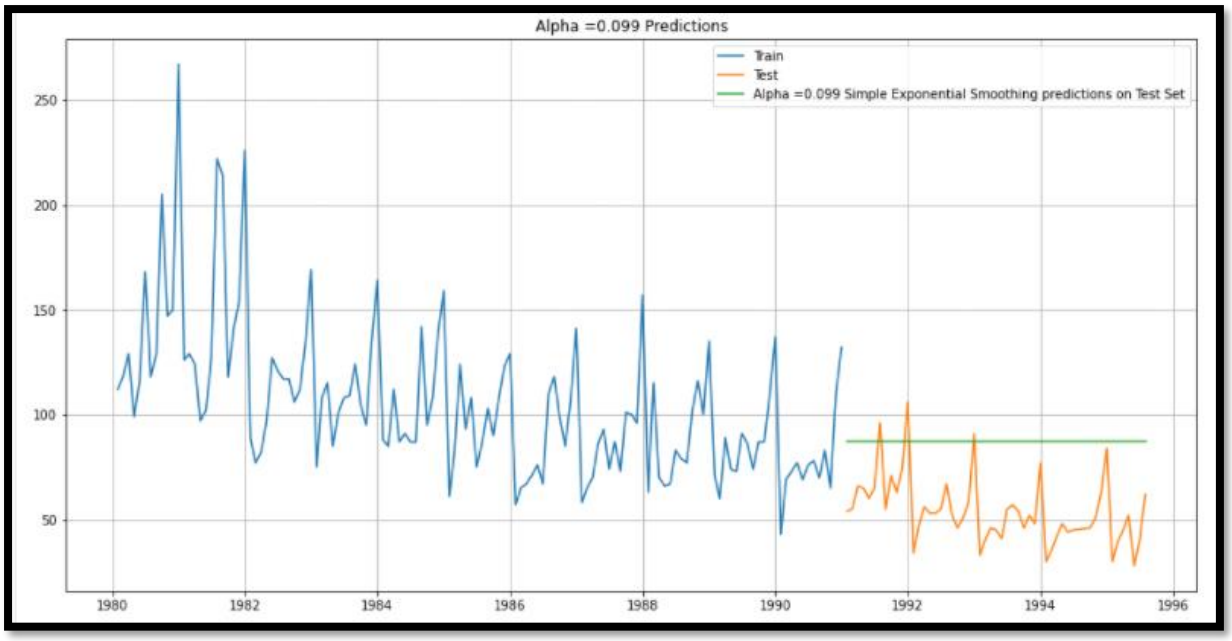

Fig 34: Predicted Value on the basis of Simple Exponential Smoothing

The graph below shows the simple trend line which is constant in predicting sales number as 87 units throughout the testing series data. This model ignores the seasonality.

Model Evaluation:

```
For Alpha =0.099 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.796
```

Fig 35: RMSE Value for Model Evaluation

RMSE on test data is 36.769 which is quite high.
There is another evaluation which is done using different values of alpha (please refer to the jupyter python file).

Model 6: Double Exponential Smoothing (Holt's Model):

This method is useful where trend is not present in data. However, in our data trend is present therefore this will not be suitable model for us.

Test-Train Time Series Plot:

The graph shows an upward trend over the test data section. This plot is build having alpha = 0.1 and beta = 0.1.

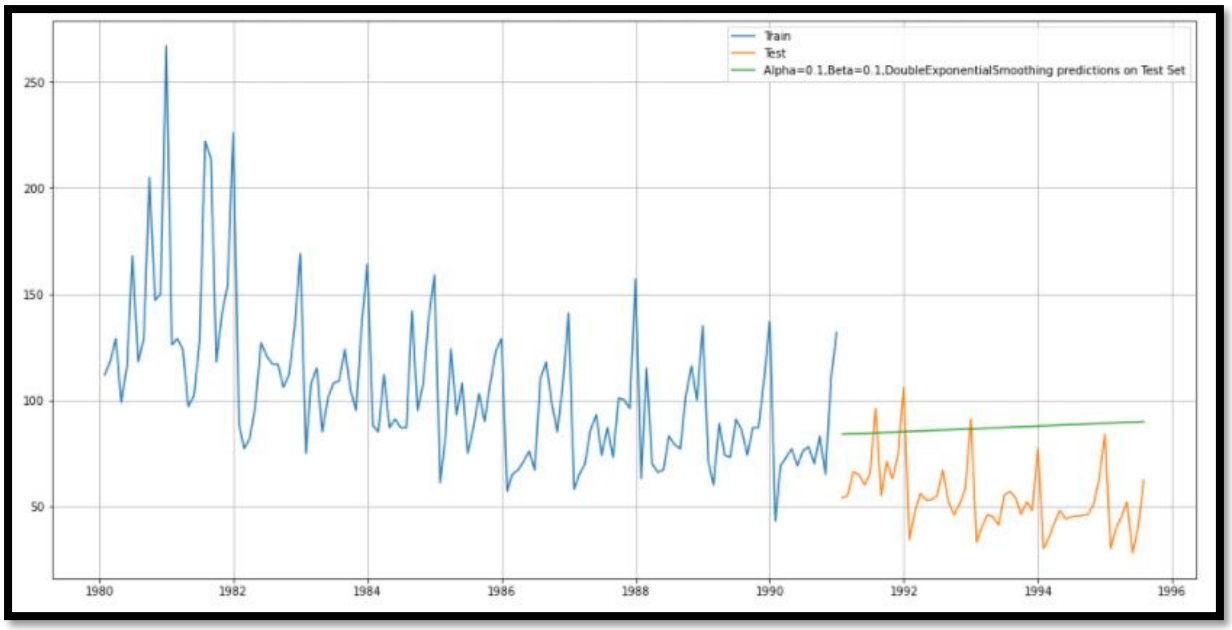

Fig 36: Train & Test Time Series Plot

Model Evaluation:

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	34.439111	36.923416
1	0.1	0.2	33.450729	48.688648
9	0.2	0.1	33.097427	65.731702
2	0.1	0.3	33.145789	78.156641
18	0.3	0.1	33.611269	98.653317

Fig 37: RMSE Values for Model Evaluation

The above RMSE values of Test data are shown in sorted order. 34.43 comes to be the lowest test RMSE value.

Model 7: Triple Exponential Smoothing:

This method takes care of trend and seasonality of data and is basically extension of Holt's method.
We will be having three smoothing parameters as α, β and y.
First model is based on taking trend as additive and seasonality as multiplicative.
Second model is based on taking trend as additive and seasonality as multiplicative (same as above but with different α, β and y).

Model Parameters:

```
{'smoothing_level': 0.06467234615091698,
 'smoothing_trend': 0.05315920636255018,
 'smoothing_seasonal': 0.0,
 'damping_trend': nan,
 'initial_level': 50.880912909225756,
 'initial_trend': -0.31656840824205823,
 'initial_seasons': array([2.21583703, 2.51439498, 2.74693025, 2.40118428, 2.69936273,
       2.94338111, 3.2353888 , 3.44052906, 3.26420741, 3.19365239,
       3.72269442, 5.13435788]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig 38: Triple Exponential Smoothing Parameters

Prediction Sample:

Time_Stamp	Rose	auto_predict
1991-01-31	54.0	56.755640
1991-02-28	55.0	64.211013
1991-03-31	66.0	69.939833
1991-04-30	65.0	60.953618
1991-05-31	60.0	68.316934

Fig 39: Triple Exponential Smoothing Predicted Sample

Above is the predicted data using Triple Exponential Smoothing model.

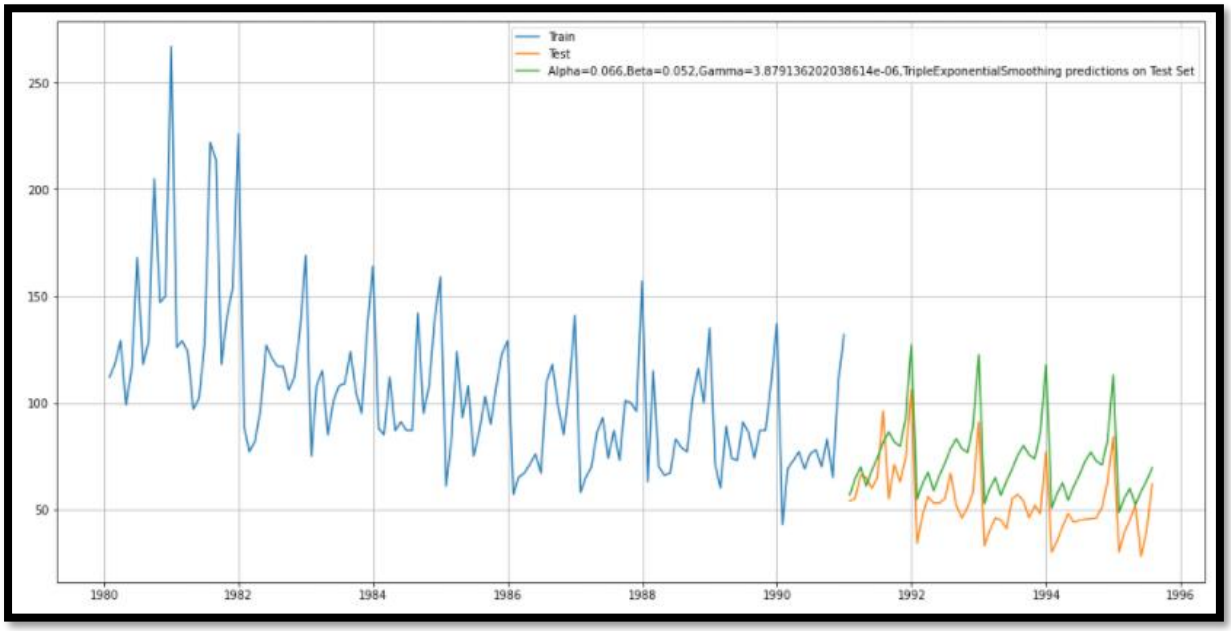

Fig 40: Train & Test Data at Chosen Parametric Values using Triple Exponential Smoothing (Prediction)

Here, green line represents the predicted values given by the model which are more or less fitting aptly on to the actual values of test data.

Model evaluation:

RMSE for this is 21.155 which is significantly less from Exponential smoothing and Naïve model, however it is significantly high as compared to 2-point moving average data.

```
For Alpha=0.066,Beta=0.052,Gamma=3.879136202038614e-06, Triple Exponential Smoothing Model forecast on the Test Data,  RMSE is 21.155
```

Fig 41: RMSE Value for Model Evaluation

For another values of alpha, beta and gamma, we did another evaluation.

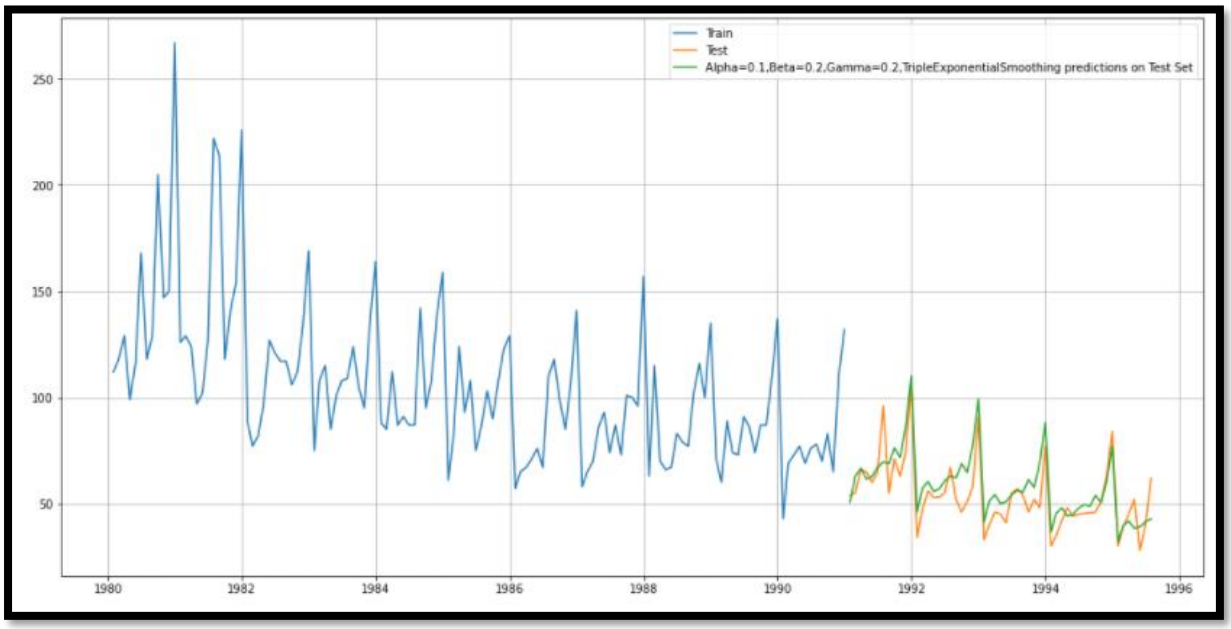

Fig 42: Train-Test Data at Chose Parametric Value Using Triple Exponential Smoothing (Prediction)

Model Evaluation:

The RMSE value with Alpha = 0.1, Beta = 0.2 and Gamma = 0.2 is 9.64.
Till now this is the lowest RMSE achieved amongst all the models.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
10	0.1	0.2	0.2	24.365597	9.640687
11	0.1	0.2	0.3	23.969166	9.935740
9	0.1	0.2	0.1	25.529854	9.943539
119	0.2	0.5	0.3	27.631767	10.026210
127	0.2	0.6	0.2	28.289836	10.031639

Fig. 43: Different RMSE Values for Different Parametric Values

Model Evaluation Comparison:

Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing	9.640687
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
RegressionOnTime	15.268955
Alpha=0.066,Beta=0.052,Gamma=3.879136202038614e-06,TripleExponentialSmoothing	21.154772
Alpha=0.099,SimpleExponentialSmoothing	36.796242
Alpha=0.1,SimpleExponentialSmoothing	36.828033
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416
SimpleAverageModel	53.460570
NaiveModel	79.718773

Fig 43: **Model Evaluation Comparison (RMSE Value Comparison) of all the Models Built**

Note: Please refer to the jupyter file for the actions taken on training data for all the models (since we included the model evaluation only for the test).

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.

There is a test called **Augmented Dickey-Fuller Test**, which is used to check whether the data is stationary or non-stationary.

The **Augmented Dickey-Fuller test** is a unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.
The hypothesis in a simple form for the ADF test is:

$H0$: The Time Series has a unit root and is thus non-stationary.
$H1$: The Time Series does not have a unit root and is thus stationary.
We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

```
DF test statistic is -2.240
DF test p-value is 0.46713716277931344
Number of lags used 13
```

Fig 44: **Augmented Dickey-Fuller Test Check**

The test gave the p-value to be 0.467 which is greater than alpha which is 0.05.
Thus, the null hypothesis can't be rejected that means the data is non-stationary at the moment.
We will have to make the data stationary in order to proceed further.

Augmented Dickey-Fuller Test After Removing NA Values:

```
DF test statistic is -8.162
DF test p-value is 3.015976115828492e-11
Number of lags used 12
```

Fig 45: Augmented Dickey-Fuller Test Check After Dropping NA Values

Difference Plot:

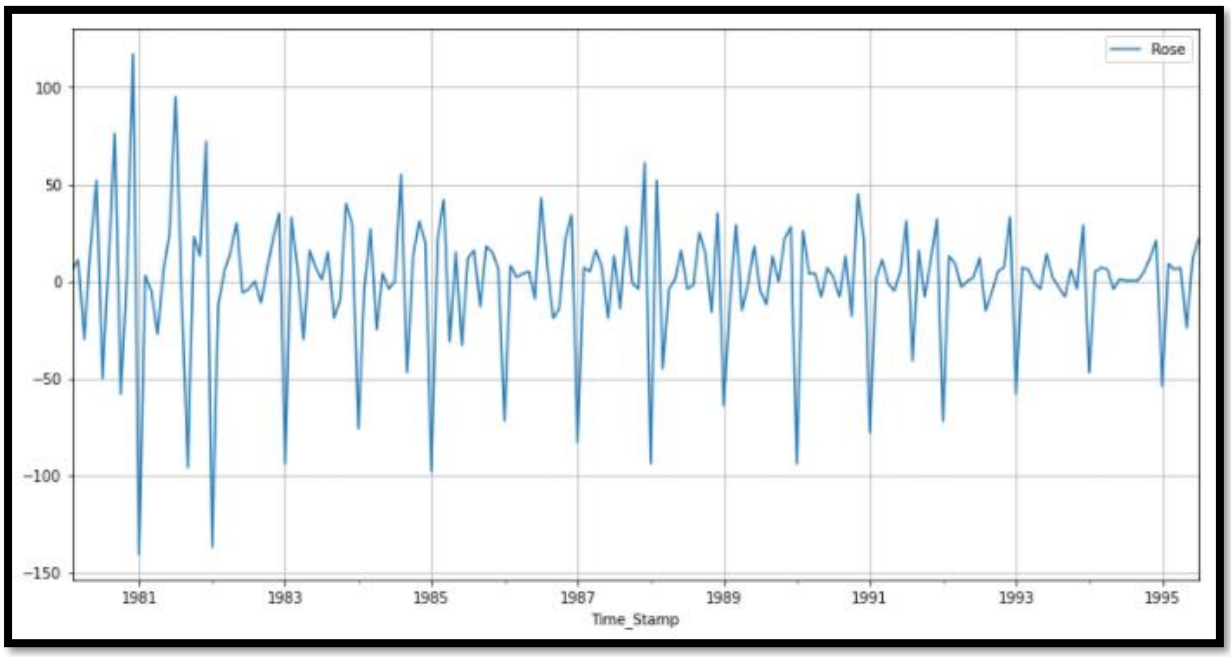

Fig 46: Difference Plot

Note: The above test results are that of the test data, for train data please refer to the jupyter (python) file.

Auto-Correlation and Partial Auto-Correlation Function Plots:

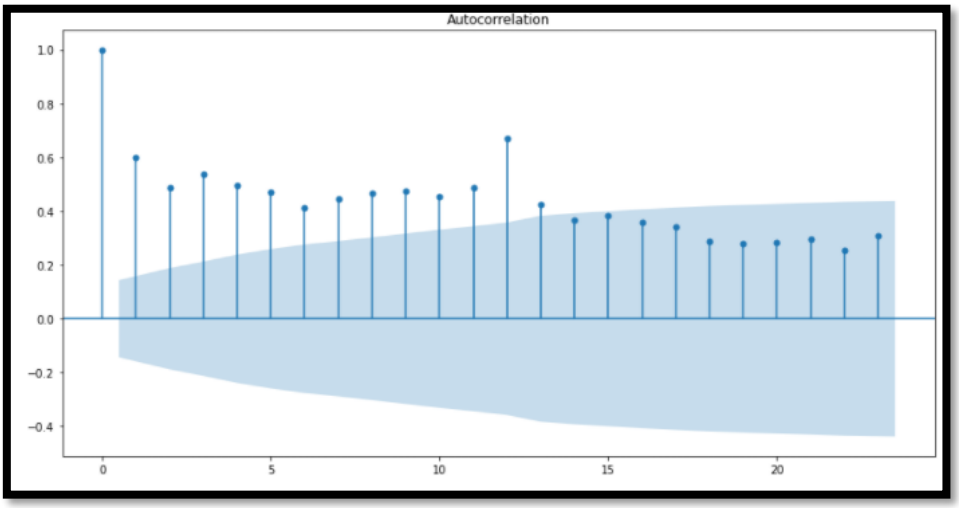

Fig 47: Auto-Correlation Plot

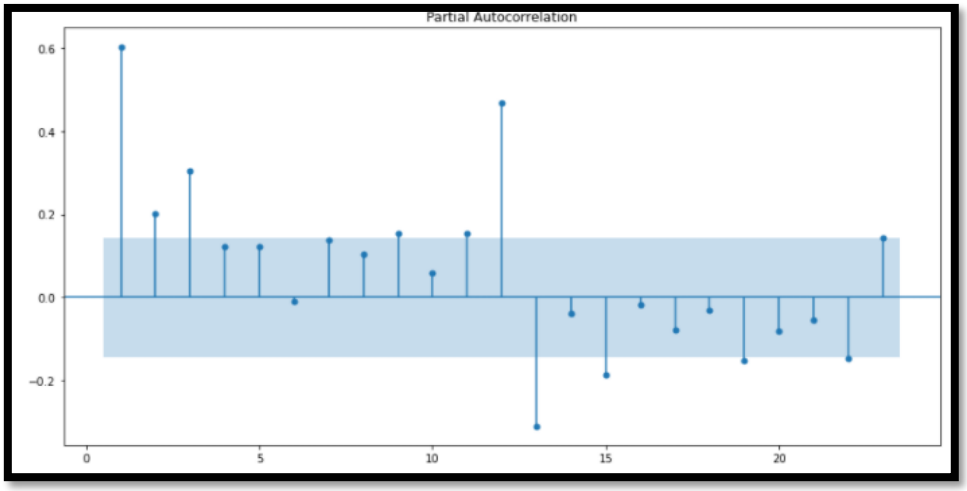

Fig 48: Partial Auto-Correlation Plot

For Train Data:

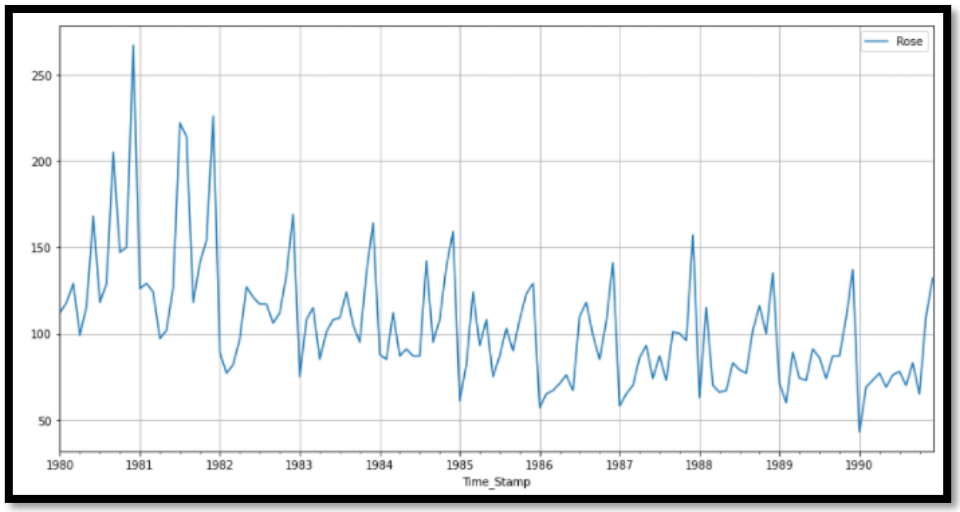

Fig 49: Train Data Plot

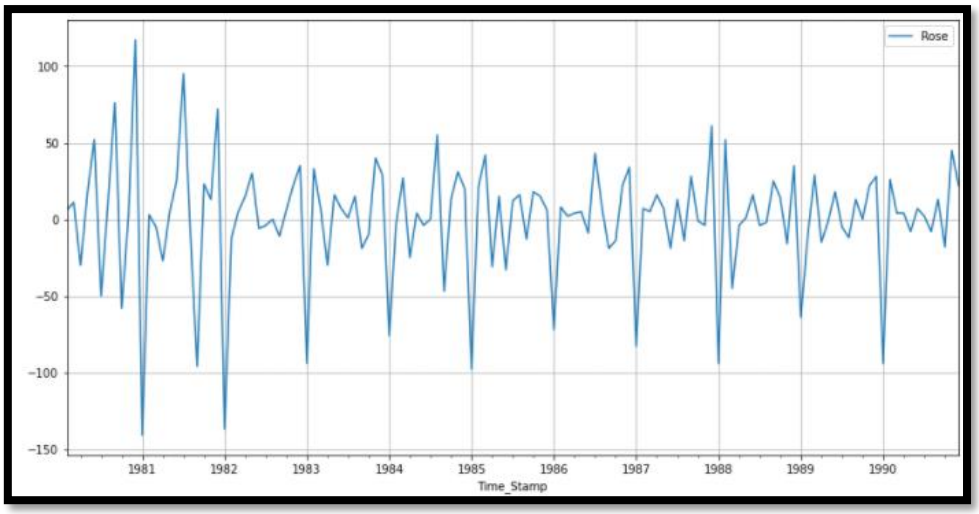

Fig 50: Train Data Plot After Dropping NA Values (Making the Data Stationary)

Note: Please refer to python file for the train data augmented dicky fuller test.

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.

Note: The data has some seasonality so ideally, we should build a SARIMA model. But for demonstration purposes we are building an ARIMA model both by looking at the minimum AIC criterion and by looking at the ACF and the PACF plots.

ARIMA:

We have chosen the value of d as 1 since we need to take a difference of the series to make it stationary. The combinations for AIC values are as under, best combination is p, d, q of 2, 1, 3 having lowest AIC:

```
Examples of the parameter combinations for the Model
Model: (0, 1, 0)
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

Fig 51: Iterative values of P, D, Q

	param	AIC
11	(2, 1, 3)	1274.695578
15	(3, 1, 3)	1278.671368
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

Fig 52: AIC Values of different P, D, Q

Best Parameter as per the AIC values is 2, 1, 3 for P, D, Q.

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                      Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sat, 25 Dec 2021   AIC                           1274.696
Time:                        20:33:24   BIC                           1291.947
Sample:                    01-31-1980   HQIC                          1281.706
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6774      0.084    -20.019      0.000      -1.842      -1.513
ar.L2         -0.7283      0.084     -8.691      0.000      -0.893      -0.564
ma.L1          1.0446      0.593      1.761      0.078      -0.118       2.207
ma.L2         -0.7716      0.130     -5.928      0.000      -1.027      -0.517
ma.L3         -0.9043      0.537     -1.685      0.092      -1.956       0.148
sigma2       859.7658    498.268      1.726      0.084    -116.821    1836.353
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.37
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================

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

Fig 53: Arima Model Summary (Automated)

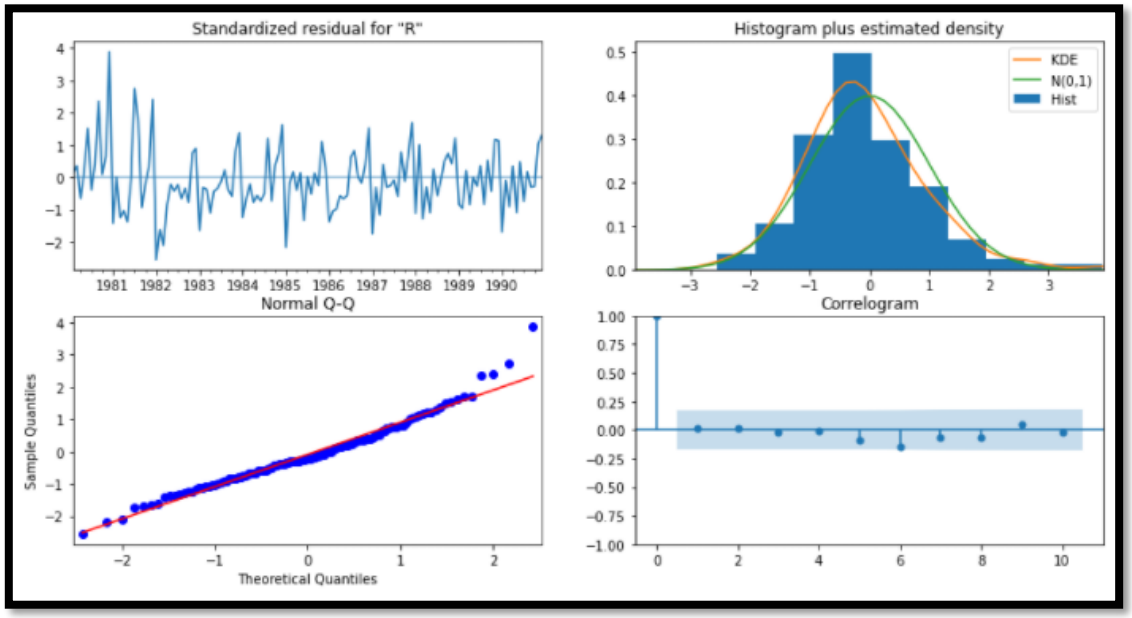

Fig 54: Diagnostic Plots

In ARIMA, AR = p = 2; MA = q = 3, as per our model. Since there are total 5 parameters (2 of AR, 2 of MA and 1 sigma), AIC value is 1274, we see 6 parameters listed with their values. The p-values for almost all are 0 or less than 0.05 (claiming the necessary condition for being stationary).

Model Evaluation:

RMSE of ARIMA is 36.81 which is quite high. This shows that SARIMA model should work out as best for our data set as seasonality is present in our data.

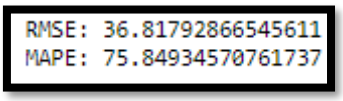

Fig 55: RMSE & MAPE for Model Evaluation

SARIMA:

Auto SARIMA model is built using seasonality as 6 as seasonality is visible at 6 months or its multiple intervals.

Below is the sample of the parameters generated automatically.

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	951.744300
59	(0, 1, 3)	(2, 0, 3, 6)	952.073632
251	(3, 1, 3)	(2, 0, 3, 6)	952.582103
191	(2, 1, 3)	(3, 0, 3, 6)	953.205620
123	(1, 1, 3)	(2, 0, 3, 6)	953.684951

Fig 56: AIC Values of different P, D, Q

the AIC values which we get are as follows. The lowest AIC returned is 951.744 for combination 2,1,3 and 2,0,3,6.

Best Parameter as per the AIC values is 2, 1, 3 for P, D, Q.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 0, 3, 6)	Log Likelihood	-464.872
Date:	Sat, 25 Dec 2021	AIC	951.744
Time:	20:37:11	BIC	981.349
Sample:	01-31-1980	HQIC	963.750
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5028	0.083	-6.083	0.000	-0.665	-0.341
ar.L2	-0.6628	0.084	-7.919	0.000	-0.827	-0.499
ma.L1	-0.3713	187.414	-0.002	0.998	-367.696	366.953
ma.L2	0.2033	117.808	0.002	0.999	-230.695	231.102
ma.L3	-0.8320	155.887	-0.005	0.996	-306.366	304.702
ar.S.L6	-0.0838	0.049	-1.720	0.085	-0.179	0.012
ar.S.L12	0.8099	0.052	15.462	0.000	0.707	0.913
ma.S.L6	0.1702	0.248	0.687	0.492	-0.315	0.655
ma.S.L12	-0.5645	0.199	-2.838	0.005	-0.954	-0.175
ma.S.L18	0.1709	0.143	1.198	0.231	-0.109	0.451
sigma2	260.8222	4.89e+04	0.005	0.996	-9.56e+04	9.61e+04

Ljung-Box (L1) (Q):	0.72	Jarque-Bera (JB):	4.77
Prob(Q):	0.40	Prob(JB):	0.09
Heteroskedasticity (H):	0.54	Skew:	-0.36
Prob(H) (two-sided):	0.06	Kurtosis:	3.73

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

Fig 57: Sarima Model Summary (Automated)

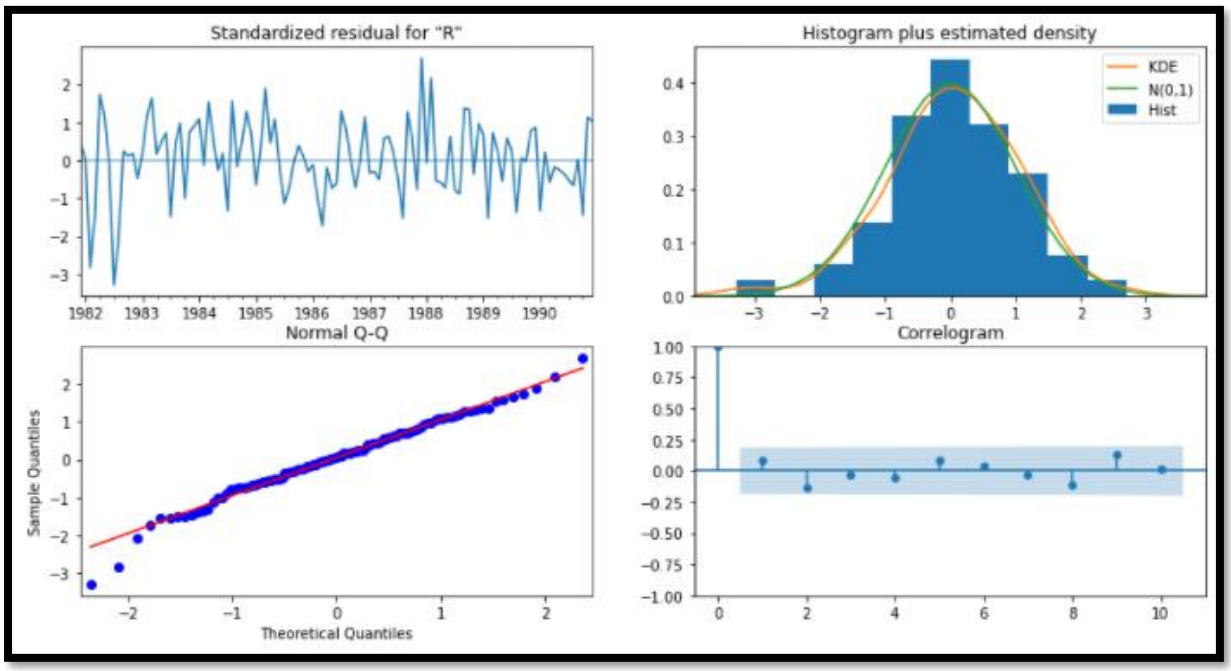

Fig 58: Diagnostic Plot

The model returns 11 parameters as seen above out of which 5. The AIC value returned in 951 which has reduced with the seasonality factor as compared to the auto SARIMA model AIC.

Model Evaluation:

```
RMSE: 27.125517502693572
MAPE: 55.241787855610404
```

Fig 59: RMSE/MAPE For Model Evaluation

RMSE is 27.12 which is significantly less than ARIMA model or even from all the models built here in this project so far.

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.

ARIMA:

ACF & PACF of ARIMA:

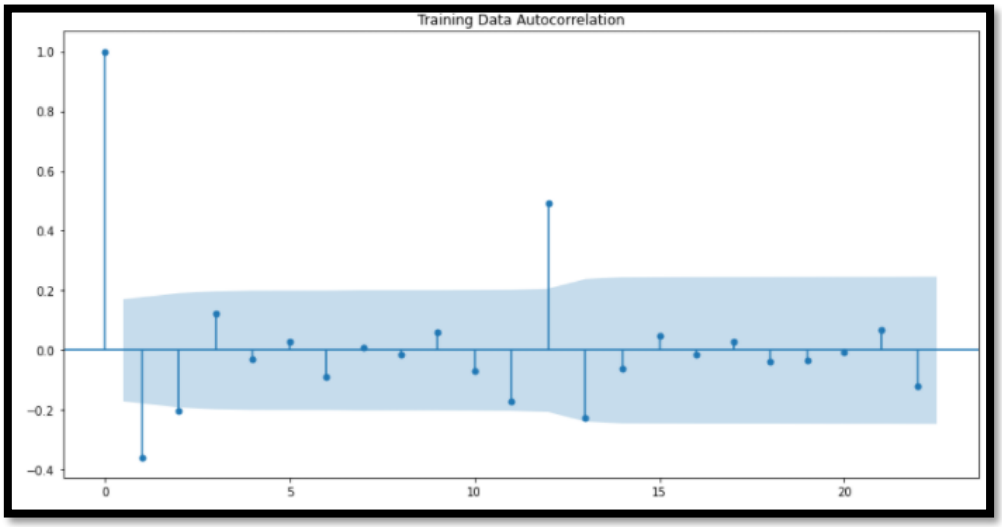

Fig 60: ACF of ARIMA

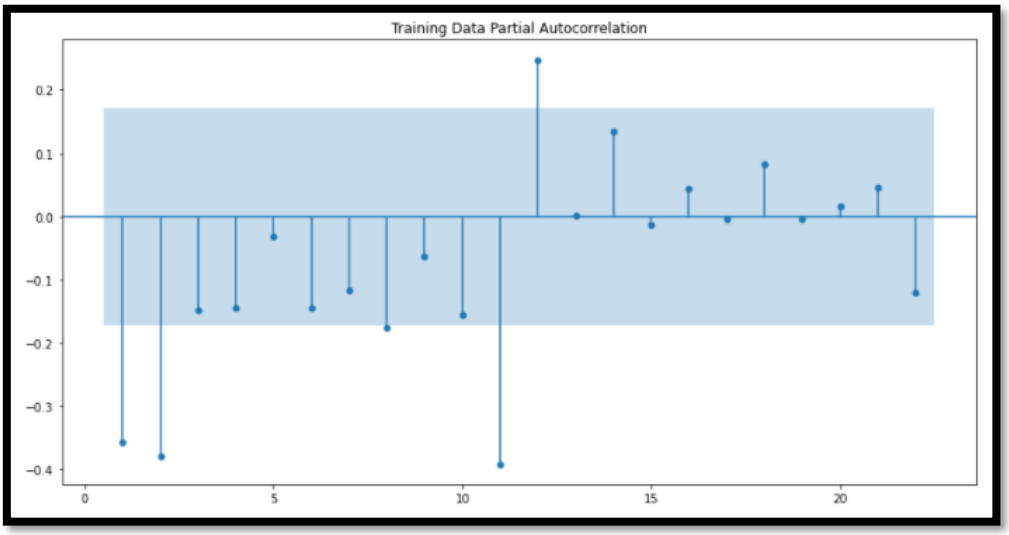

Fig 61: PACF of ARIMA

- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off is clearly 2 here.
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off which we should take as 2 in our case.

Manual filling the pdq values as 2,1,2 gives the following results for ARIMA

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Sat, 25 Dec 2021	AIC	1281.871
Time:	20:33:25	BIC	1296.247
Sample:	01-31-1980 - 12-31-1990	HQIC	1287.712
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4540	0.469	-0.969	0.333	-1.372	0.464
ar.L2	0.0001	0.170	0.001	0.999	-0.334	0.334
ma.L1	-0.2541	0.459	-0.554	0.580	-1.154	0.646
ma.L2	-0.5984	0.430	-1.390	0.164	-1.442	0.245
sigma2	952.1601	91.424	10.415	0.000	772.973	1131.347

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	34.16
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.79
Prob(H) (two-sided):	0.00	Kurtosis:	4.94

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

Fig 62: Manual ARIMA Summary

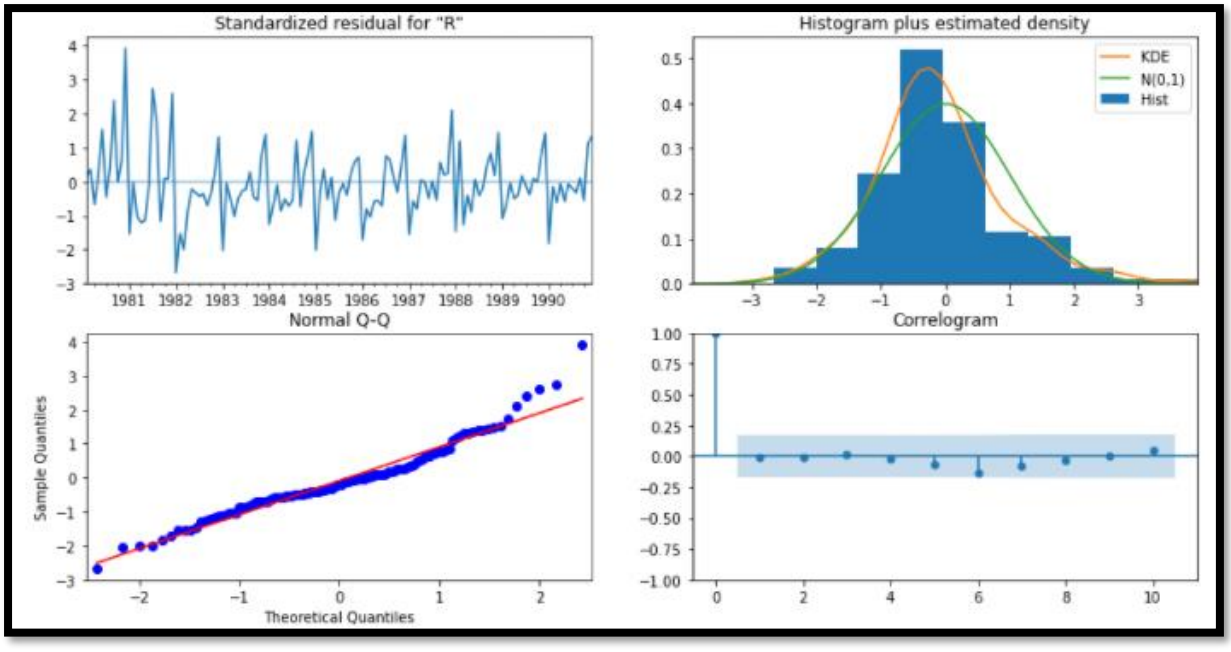

Fig 63: Diagnostic Plot Arima

We get a comparatively simpler model by looking at the ACF and the PACF plots. The AIC value obtained in the manual ARIMA model is 1281.871 v/s an AIC value of 1274 for the auto ARIMA model. An increase of around 7-point hence the manual model is not a good model.

Model Evaluation:

```
RMSE: 36.871196613369605
MAPE: 76.05621270446787
```

Fig 64: Diagnostic Plot ARIMA

RMSE is 36.87. Still not preferable as a model to be used here in this project, since the triple exponential smoothing has better RMSE value.

SARIMA:

ACF & PACF of ARIMA:

We have built the differenced ACF and PACF graphs below to find out the relevant seasonal interval. Looking at ACF plot the seasonal interval (6) holds well. So, we go ahead and take a seasonal differencing of the original series.

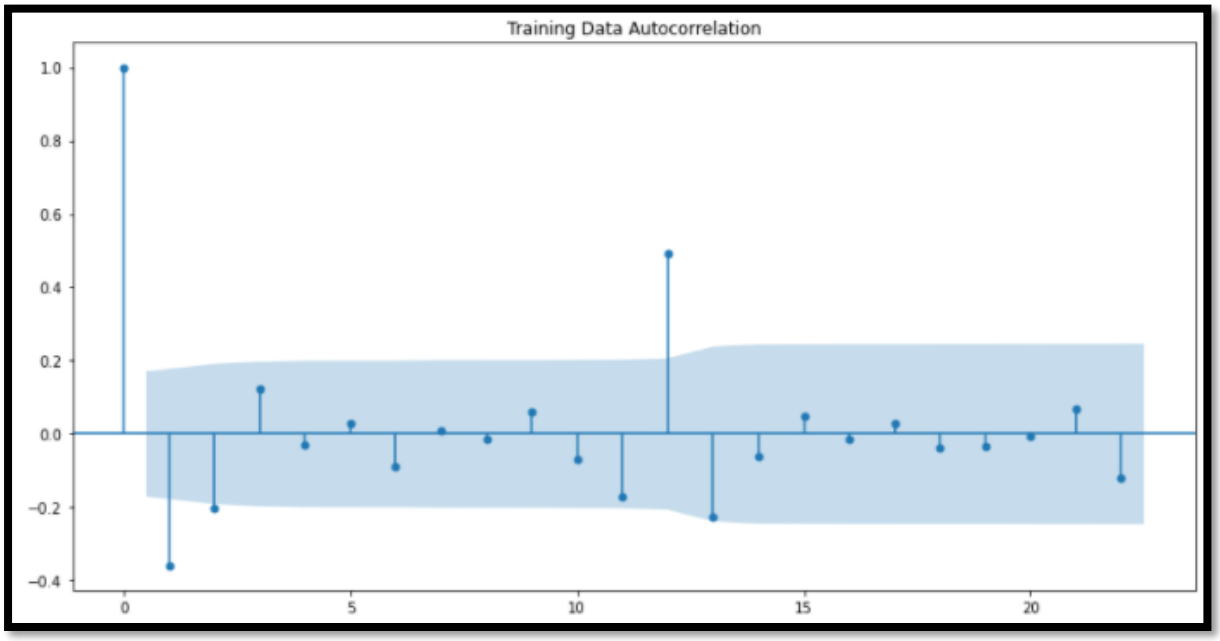

Fig 65: ACF of SARIMA

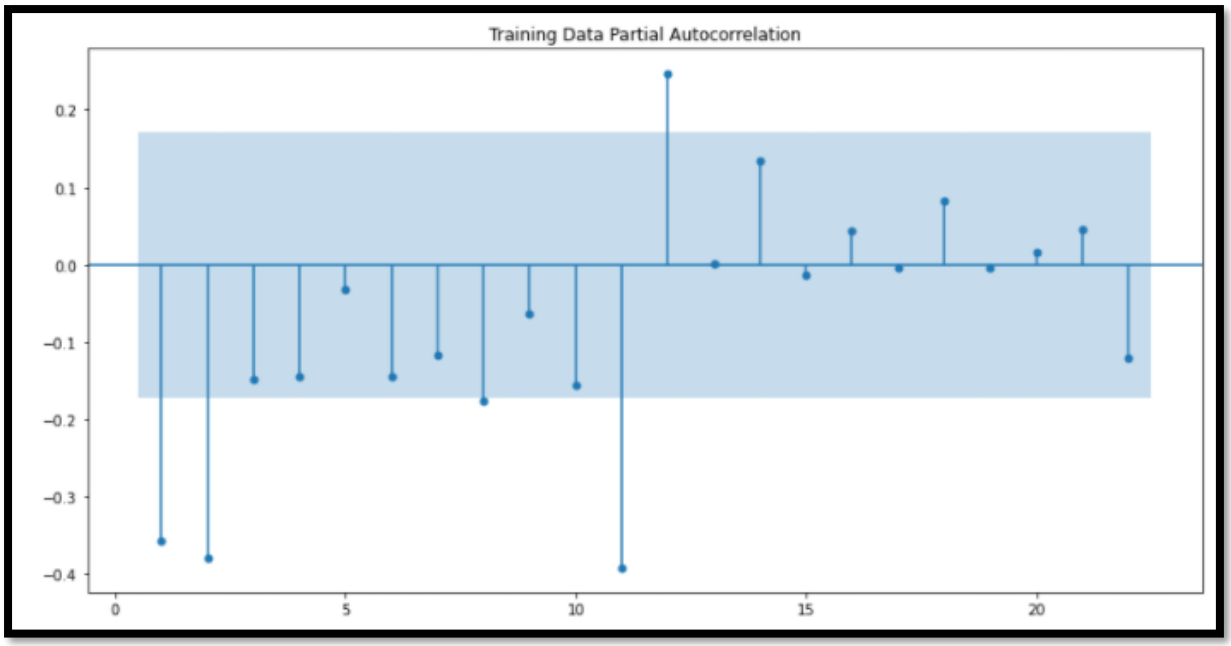

Fig 66: PACF of SARIMA

Here, we have taken alpha=0.05.
We are going to take the seasonal period as 2 or its multiple e.g., 6. We are taking the p value to be 2 and the q value also to be 2 as the parameters same as the ARIMA model.
The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 2. The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 2.

SARIMAX Results

Dep. Variable:		Rose	No. Observations:		132	
Model:	SARIMAX(2, 1, 2)x(3, 0, 2, 6)		Log Likelihood		-471.498	
Date:	Sat, 25 Dec 2021		AIC		962.996	
Time:	20:37:14		BIC		990.092	
Sample:	01-31-1980 - 12-31-1990		HQIC		973.988	
Covariance Type:	opg					
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4011	0.183	-2.190	0.029	-0.760	-0.042
ar.L2	0.1363	0.093	1.463	0.144	-0.046	0.319
ma.L1	-0.3344	279.689	-0.001	0.999	-548.514	547.846
ma.L2	-0.6656	186.214	-0.004	0.997	-365.638	364.307
ar.S.L6	-0.1943	0.066	-2.938	0.003	-0.324	-0.065
ar.S.L12	0.8209	0.043	19.134	0.000	0.737	0.905
ar.S.L18	0.1025	0.056	1.815	0.070	-0.008	0.213
ma.S.L6	0.2860	0.193	1.479	0.139	-0.093	0.665
ma.S.L12	-0.6011	0.137	-4.393	0.000	-0.869	-0.333
sigma2	253.8713	7.1e+04	0.004	0.997	-1.39e+05	1.39e+05

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	0.92
Prob(Q):	0.81	Prob(JB):	0.63
Heteroskedasticity (H):	0.81	Skew:	0.19
Prob(H) (two-sided):	0.54	Kurtosis:	3.22

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

Fig 67: SARIMA Summary

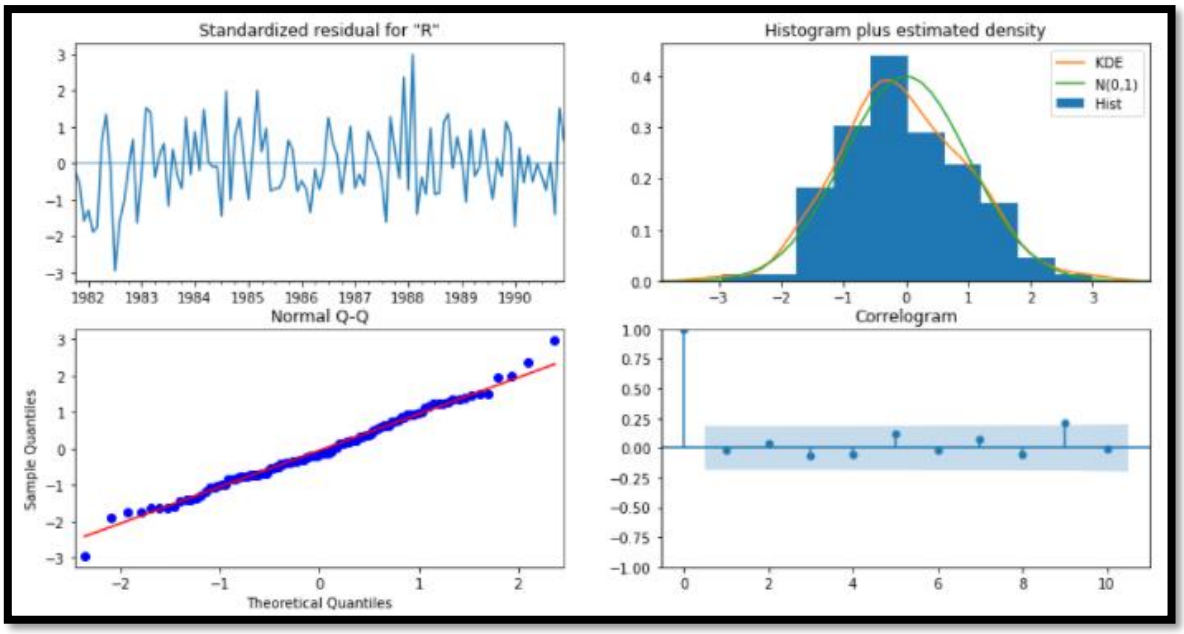

Fig 68: Diagnostic Plot

The p-value is greater than 0.05 here therefore we can't reject null hypothesis for some factors. The moving average is having p-value greater than 0.05. The manual model has 962 and the automated SARIMA has 951. A difference in 100-point. The manual SARIMA model here is not fit to be considered.

Model Evaluation:

```
RMSE: 26.622319253529184
MAPE: 54.277602216222796
```

Fig 69: RMSE/MAPE for Model Evaluation

The RMSE for manual SARIMA model is 26.62. One of the lowest but since the model has its own inconsistencies. Suggestion is not to consider this model.

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.

Shown below is the table for comparison of all the models built by far using RMSE values as the basis of comparison. The lowest the RMSE value is, the better the model will be (one of the necessary conditions).
Also, the below table contains the parameters along which they were built for the time forecast series.

Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing	9.640687
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
RegressionOnTime	15.268955
Alpha=0.066,Beta=0.052,Gamma=3.879136202038614e-06,TripleExponentialSmoothing	21.154772
SARIMA(2,1,2)(3,0,2,6)	26.622319
SARIMA(2,1,3)(2,0,3,6)	27.125518
Alpha=0.099,SimpleExponentialSmoothing	36.796242
ARIMA(2,1,3)	36.817929
Alpha=0.1,SimpleExponentialSmoothing	36.828033
ARIMA(2,1,2)	36.871197
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416
SimpleAverageModel	53.460570
NaiveModel	79.718773

Fig 70: RMSE Values of all the Models Built

Here as per the list the Triple Exponential Smoothing (Holt – Winter's Model) is having the lowest RMSE value for Test Model (9.64) whereas the Naïve Model is having the highest RMSE value for the test model (79.71).

Therefore, the best model to consider here will be the Triple Exponential Smoothing (Holt-Winter's Model) with Alpha = 0.1, Beta = 0.2 and Gamma = 0.2 for building the final model.

Therefore, the big idea of choosing the Triple Exponential Smoothing (Holt – Winter's Model) is that it will have better and closer predictions to the actual scenario, having less of errors and can prove to be a better model for the organization to use for its evaluations in order to better the sales with better decisions.

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.

Full Model:

Since the best model is coming to be with Triple Exponential Smoothing, therefore we will build the full model using Triple Exponential Smoothing.
Smoothing level = 0.1, smoothing trend = 0.2, smoothing seasonal = 0.2
Time series plot for the complete data for this model is shown below.

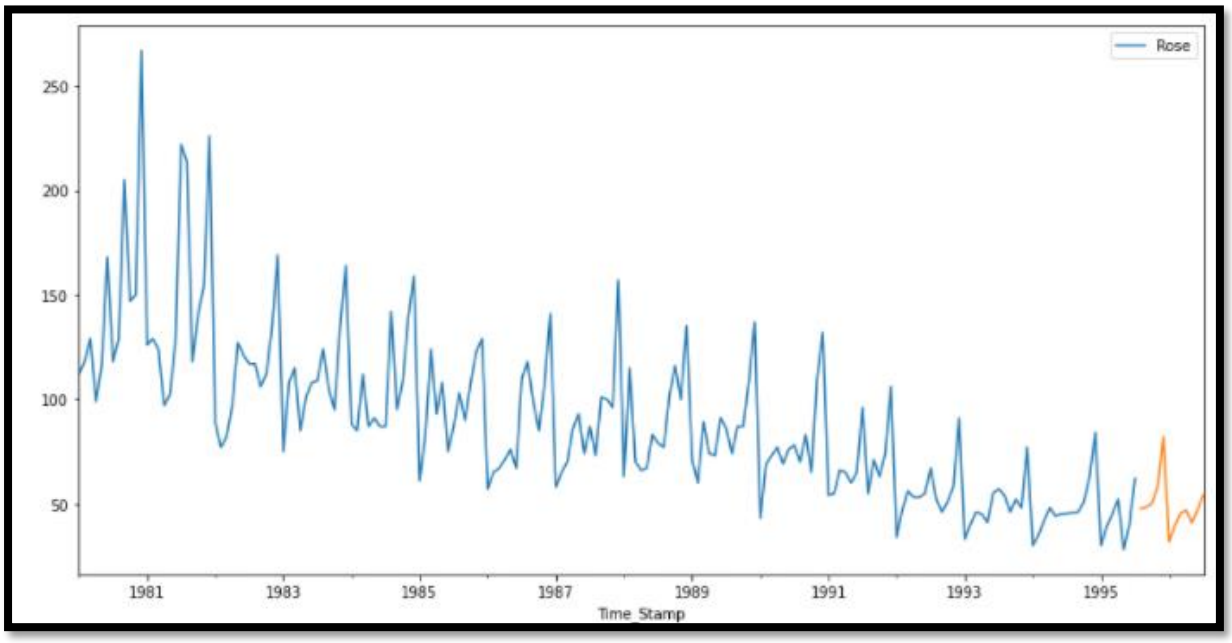

Fig 71: Full Model Time Series Plot

Prediction (Confidence Level=95% i.e., alpha (for p-value comparison = 0.05):

	lower_CI	prediction	upper_ci
1995-08-31	13.422360	47.607992	81.793624
1995-09-30	14.098851	48.284483	82.470114
1995-10-31	16.094027	50.279659	84.465291
1995-11-30	24.275597	58.461229	92.646861
1995-12-31	47.930965	82.116597	116.302229

Fig 72: Predictions with Lower & Upper Confidence Bands (Confidence Level = 95%)

Above is the sample of the values predicted for 12 months. The lower ci and the upper ci for the correction factor.
We can see as per the predictions; the sales are increasing by the end of year 1995 but decreases a bit more in 1996. (Please refer to python file for complete 12 months data)

Model Evaluation:

This model was made using **smoothing level = 0.1**, **smoothing trend = 0.2**, **smoothing seasonal = 0.2**.
The RMSE for this model is 17.404

```
RMSE: 17.404362273430372
```

Fig 73: RMSE for Model Evaluation

Prediction Plot:

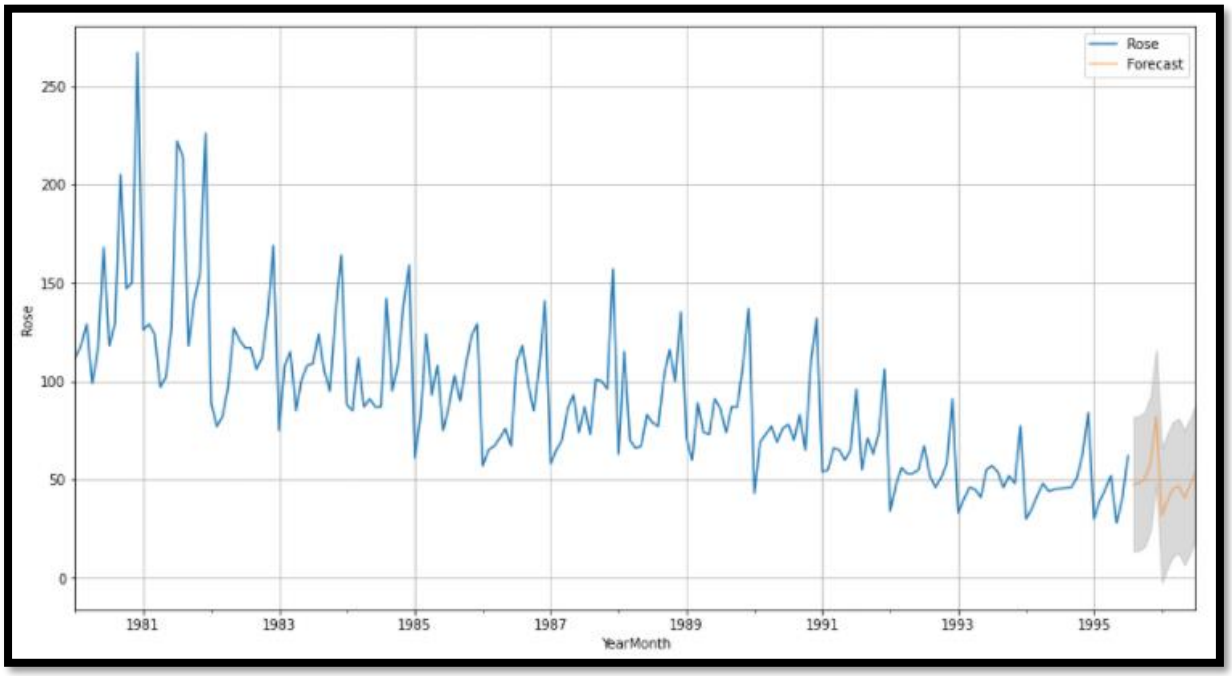

Fig 74: Full Model Prediction Plot

The above shown is plot as per the predictions made. Which confirms on the same fact that the sales go a bit up by the end of the year 1995 but decreases and then increases later on.

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

Brief:

We have been given the data of a company which sells wine and we have analysed the same in order to have a time series forecast and to take the necessary actions in order to better the same.

We have built many models (list below) and have analysed the data and applied operations on the data to come up with the predictions, forecasting to be precise and to take necessary actions to make the outcomes better than they can be.

Also, we have split the data into test and train using 1991 as the dividing year as was instructed.

Models:

1. **Linear Regression:** We applied operations on the training data and then tested it over the test data. Plot the time series forecast plot and did the model evaluation in which we found that the model is not much of a fit for our purpose where reason being the RMSE value that was observed, is yet to be compared. So, we didn't use it to build the final model for then but later was rejected.
2. **Naïve Bayes:** Naïve Bayes can be extremely fast relative to other classification algorithms. It majorly works over the Bayes theorem of probability to predict the class of unknown dataset. Here it acts similarly from the class of causal and non-causal systems where it uses the previous data to bring out the calculation for the future data. The model used the previous data from the train data and used it on the test data to predict. But this model here comes out to have a large RMSE value making it unfit for the time series forecast of the data available here.
3. **Simple Average:** The simple average of a set of observations is computed as the sum of the individual observations divided by the number of observations in the set. The RMSE value for this model as well came out to be large for us to consider it to build the full model for the data available to us for the wine sales.
4. **Moving Average:** Moving average is a calculation used to analyse data points by creating a series of averages of different subsets of the full data set. By calculating the moving average, the impacts of random, short-term fluctuations on the price of any stock over a specified time frame are mitigated. Here it had a variety of the observations but had a good 2-point moving average RMSE value. But still we put that on hold for the situation in case we find a better RMSE value for the other models to be built further.
5. **Exponential Smoothing:** Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. It is a powerful forecasting method that may be used as an alternative to the popular Box-Jenkins ARIMA family of methods. Here the RMSE value was also unfit to get this model considered for building the full data model for time series forecasting.
6. **Double Exponential Smoothing:** Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. It is a powerful forecasting method that may be used as an alternative to the popular Box-Jenkins ARIMA family of methods. We did the training on the train data and then further tested on the test data. Also, we did test with two different sets of the parameter's alpha and beta with different values in both the testing. The RMSE value is still high for us to consider this model fit for making the model building with full data and further predicting the same.
7. **Triple Exponential Smoothing (Holt-Winter's Model):** Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. It is a powerful forecasting method that may be used as an alternative to the popular Box-Jenkins ARIMA family of methods. We did the same training on the train data, tested on test data, with two parameters having two different values each to

further test the model. The model was fitter than others. The RMSE value is lowest till now. But there are ARIMA and SARIMA yet to come.

8. **Automated & Manual ARIMA:** The automated ARIMA was the built and the parameter values were selected on the basis of the AIC value (lowest AIC is preferred) and the manual ARIMA was built using the ACF and PACF plots for getting the values of "q" and "p" respectively for us to get the desired order. The RMSE value is still high than the triple exponential smoothing.
9. **Automated & Manual SARIMA:** The automated SARIMA was the built and the parameter values were selected on the basis of the AIC value (lowest AIC is preferred) and the manual SARIMA was built using the ACF and PACF plots for getting the values of "q" and "p" respectively for us to get the desired order and the seasonality. The RMSE value is still high than the triple exponential smoothing.

At final, we have selected triple exponential smoothing for building the full model as the RMSE value is the lowest for the triple exponential model (all the RMSEs are listed below).

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing	9.640687
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
RegressionOnTime	15.268955
Alpha=0.066,Beta=0.052,Gamma=3.879136202038614e-06,TripleExponentialSmoothing	21.154772
SARIMA(2,1,2)(3,0,2,6)	26.622319
SARIMA(2,1,3)(2,0,3,6)	27.125518
Alpha=0.099,SimpleExponentialSmoothing	36.796242
ARIMA(2,1,3)	36.817929
Alpha=0.1,SimpleExponentialSmoothing	36.828033
ARIMA(2,1,2)	36.871197
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416
SimpleAverageModel	53.460570
NaiveModel	79.718773

Fig 75: RMSE Values of all the Models Built

Measures/Suggestions for the Future Sales:

1. **Cross-Market with Other Local Businesses (Including Wineries):** Consumers are savvy with both their money and their time. If you want people to visit your tasting room or winery, create content promoting other local businesses nearby (and encourage them to promote your business too). People who are searching for wine-related activities in your area will be intrigued when they see that by visiting your establishment, they can also visit several other local attractions.
2. **Offer Tasting Packages:** People make more purchases after sampling a product, just ask Costco. As a winery, you likely offer tastings at your primary location.
3. **Know when to serve what:** Knowing when to serve red, and when to serve white or rose can also help customers know what they should prefer at what occasion— since many consumers can't distinguish a cabernet from another dry red wine.
4. **Online Blogs/Using Social Media:** Using social media to advertise the wine will also prove to be beneficial since almost every person carrying cell phones are using social media accounts and see a dozen of content every day.
5. **Use Influencers:** A social media influencer is simply someone who has a sphere of influence. I tend to think of the most significant influencers as those having 100,000 followers or more, but even someone with half that following can help you sell wine.
6. **Create A Sub-Brand:** Sub-brands are brands that fall under your larger brand umbrella but are directed at a different demographic than your typical product. Think Diet Coke versus Coca-Cola. Millennials make up a huge segment of the wine-consuming audience, drinking 42% of all wine sold in the U.S. If your brand isn't appealing to this demographic, consider a sub-brand that is focused on earning their loyalty.
7. **Unique Offers:** Coming on to the data part, we see that there are off seasons like may, June, July when the wine sales were very less. Seems like there were less or no occasions during that period. There should be offers in the off seasons for the Rose wine so that people can consider it more to buy even in off seasons as well.
8. We also can predict that the wine sales go high during some occasion. Now considering the average sales of 90 unit during all seasons and minimum of 28 unit. The company should organize events to maintain the feeling of occasion even during the off seasons so that people buy the wine more that it is sold in off seasons.