

INDEX

Problem.....	3
1. Read the data as an appropriate Time Series data and plot the data.....	4
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.....	6
3. Split the data into training and test. The test data should start in 1991....	10
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.....	11
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 $\alpha = 0.05$	22
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.....	25
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.....	29
8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.....	33
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.....	35
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.....	36

List of Figures:

Fig 1: Sample Data "Sparkling"	Fig 2: Time Stamp "Sparkling"
Fig 3: Time Series Plot "Sparkling"	Fig 4: Time Series Stamping "Sparkling"
Fig 5: Time Series Plot of "Sparkling" after Time Series Stamping	Fig 6: Descriptive Data of "Sparkling"
Fig 7: Yearly Box-Plot of "Sparkling"	Fig 8: Monthly Box-Plot of "Sparkling"
Fig 9: Month Plot of Time Series	Fig 10: Numerical Presentation of Year-Month Plot
Fig 11: Year-Month Plot	Fig 12: Data Decomposition Using Additive Model
Fig 13: Data Decomposition Using Multiplicative Model	Fig 14: Sample of Train Data
Fig 15: Sample of Test Data	Fig 16: Train & Test Data Plot
Fig 17: Training-Testing Data Time Instances	Fig 18: Training Sample Data
Fig 19: Testing Sample Data	Fig 20: Plot of Train & Test Data after adding Time Instances
Fig 21: RMSE of Test Data for Model Evaluation	Fig 22: Naïve Bayes Based Time Series Plot (Test Data)
Fig 23: RMSE of Test Data for Model Evaluation)	Fig 24: Sample of Predicted Value
Fig 25: Simple Average Based Time Series Plot (Test Data)	Fig 26: RMSE of Test Data for Model Evaluation
Fig 27: Moving Average Sample Data	Fig 28: Moving Average Plot (Train Data)
Fig 29: Moving Average Plot (Train & Test Data)	Fig 30: RMSE for Model Evaluation (Test Data)
Fig 31: LR, NB, SA, MA Comparison Plot	Fig 32: Parameters Used for Simple Exponential Smoothing Model
Fig 33: Predicted Value on the basis of Simple Exponential Smoothing	Fig 34: Predicted Value on the basis of Simple Exponential Smoothing
Fig 35: RMSE Value for Model Evaluation	Fig 36: Train & Test Time Series Plot
Fig 37: RMSE Values for Model Evaluation	Fig 38: Triple Exponential Smoothing Parameters
Fig 39: Triple Exponential Smoothing Predicted Sample	Fig 40: Train & Test Data at Chosen Parametric Values using Triple Exponential Smoothing (Prediction)
Fig 41: RMSE Value for Model Evaluation	Fig 42: Train-Test Data at Chose Parametric Value Using Triple Exponential Smoothing (Prediction)
Fig 43: Model Evaluation Comparison (RMSE Value Comparison) of all the Models Built	Fig 44: Augmented Dickey-Fuller Test Check
Fig 45: Augmented Dickey-Fuller Test Check After Dropping NA Values	Fig 46: Difference Plot
Fig 47: Auto-Correlation Plot	Fig 48: Partial Auto-Correlation Plot
Fig 49: Train Data Plot	Fig 50: Train Data Plot After Dropping NA Values (Making the Data Stationary)
Fig 51: Iterative values of P, D, Q	Fig 52: AIC Values of different P, D, Q
Fig 53: Arima Model Summary (Automated)	Fig 54: Diagnostic Plots
Fig 55: RMSE & MAPE for Model Evaluation	Fig 56: AIC Values of different P, D, Q
Fig 57: Sarima Model Summary (Automated)	Fig 58: Diagnostic Plot
Fig 59: RMSE/MAPE For Model Evaluation	Fig 60: ACF of ARIMA
Fig 61: PACF of ARIMA	Fig 62: Manual Arima Summary
Fig 63: Diagnostic Plot ARIMA	Fig 64: ACF of SARIMA
Fig 65: PACF of SARIMA	Fig 66: SARIMA Summary
Fig 67: Diagnostic Plot	Fig 68: RMSE/MAPE for Model Evaluation
Fig 69: RMSE Values of all the Models Built	Fig 70: Full Model Time Series Plot
Fig 71: RMSE for Model Evaluation	Fig 72: Full Model Prediction Plot
Fig 73: RMSE Values of all the Models Built	

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

Fig 1: Sample Data “Sparkling”

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 “Sparkling”

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:

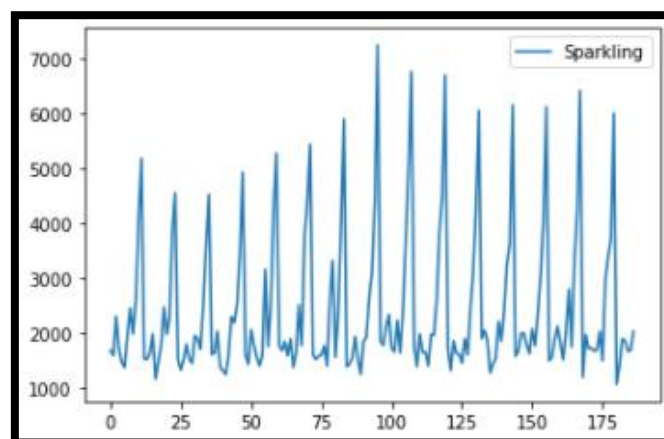


Fig 3: Time Series Plot “Sparkling”

Time Series Stamping with the Data Given:

Sparkling	
Time_Stamp	
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471
...	...
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031
187 rows × 1 columns	

Fig 4: Time Series Stamping “Sparkling”

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

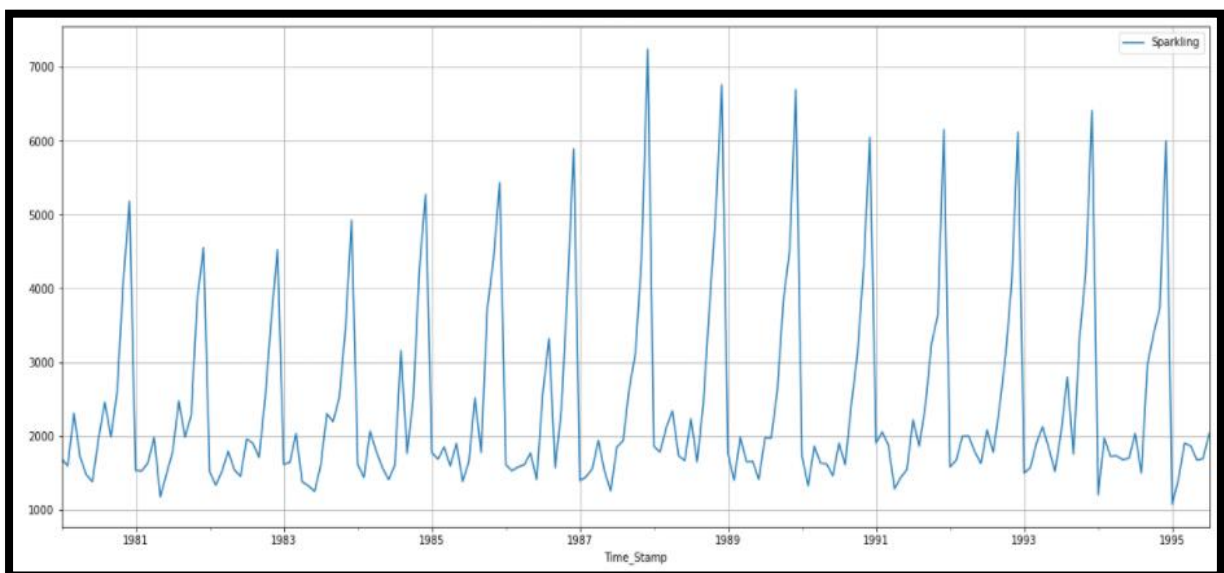


Fig 5: Time Series Plot of “Sparkling” after Time Series Stamping

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 7242 units and minimum sales is of 1070 units. average sales through series are 2402 units. and in total we have 187 months for which this data is speaking.

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

Fig 6: Descriptive Data of “Sparkling”

Missing Values: there are no missing values in data as clearly visible from the time series graph that trend lines are continuous over 187 months without any break of line in between.

Box plot:

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

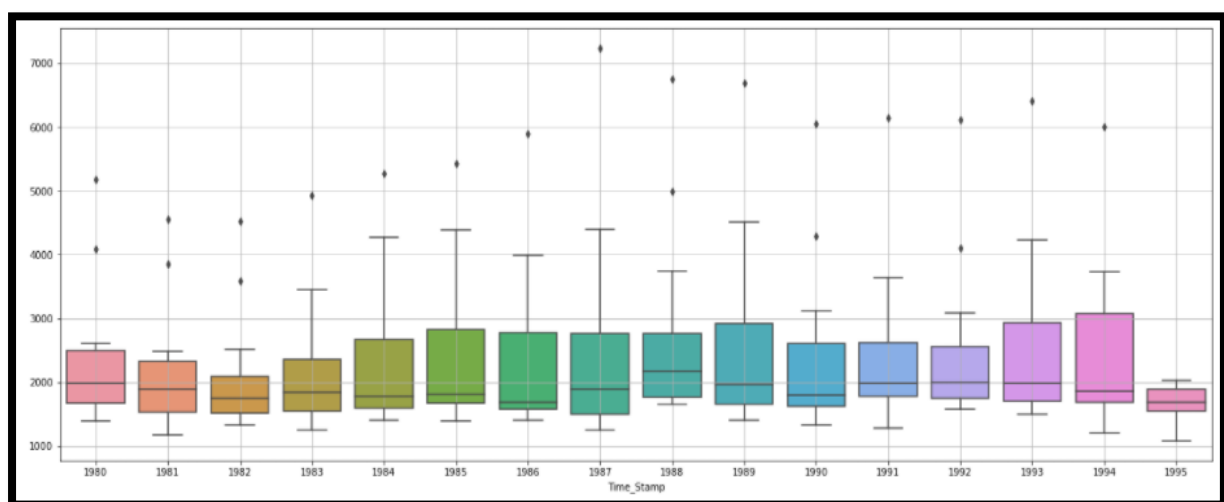


Fig 7: Yearly Box-Plot of “Sparkling”

Monthly Box-Plot: clearly indicates the trends is present in months as outliers are present 1st ,2nd, 3rd months.

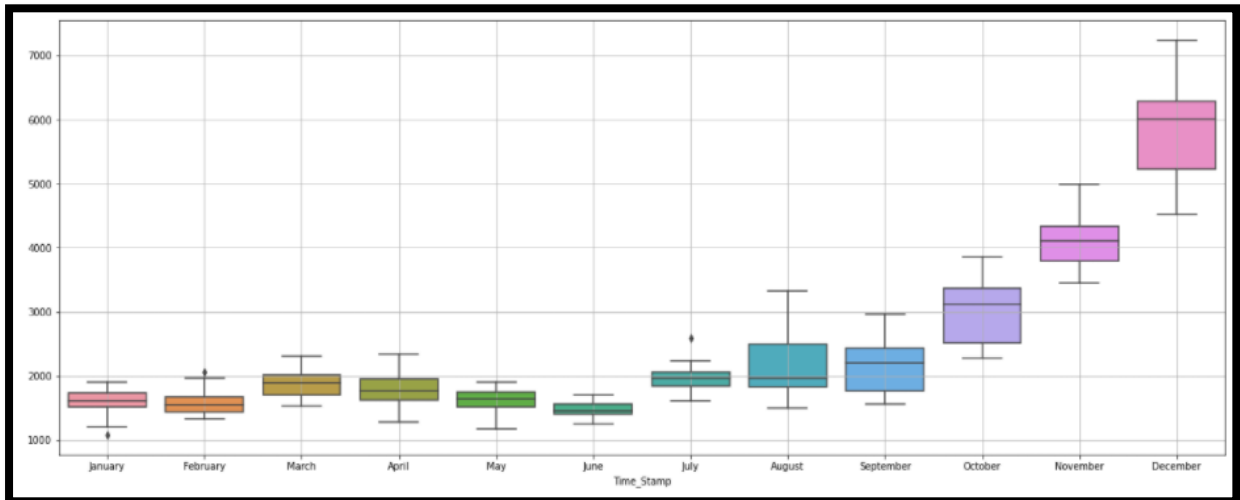


Fig 8: Monthly Box-Plot of “Sparkling”

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.

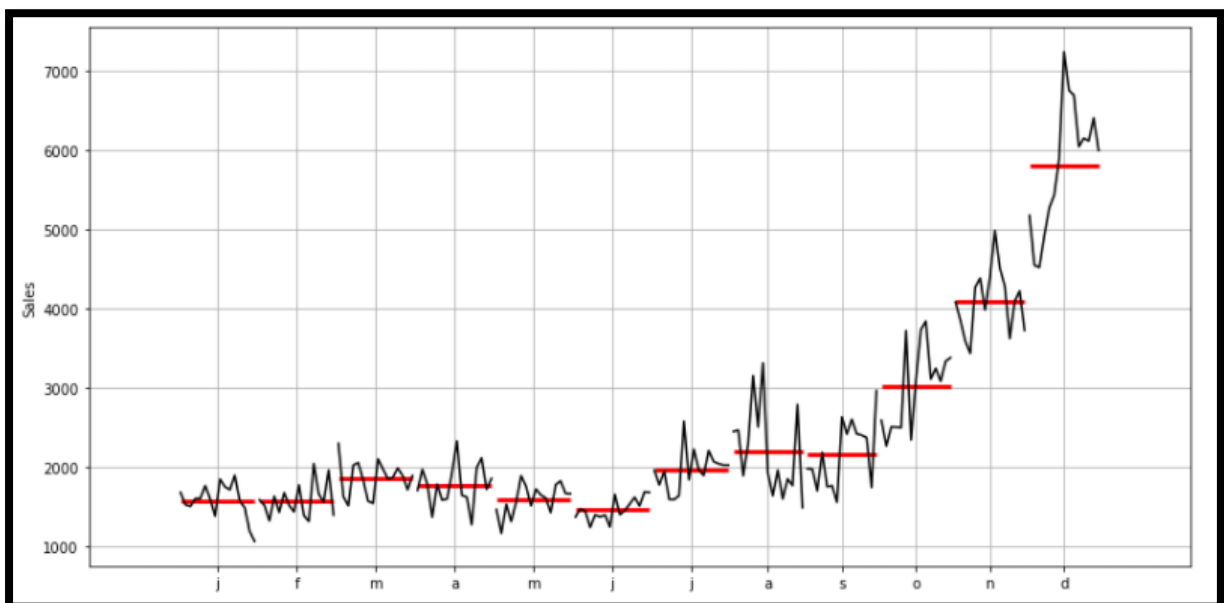


Fig 9: Month Plot of Time Series

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

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.

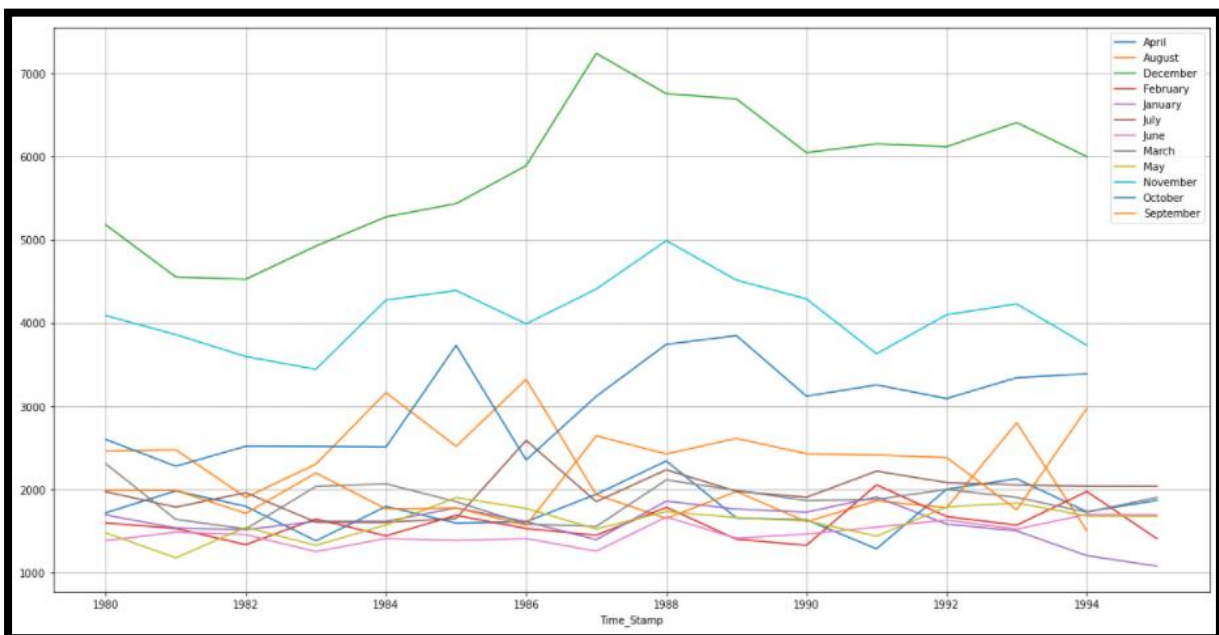


Fig 11: Year-Month Plot

December records the highest sale followed by November and October 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.

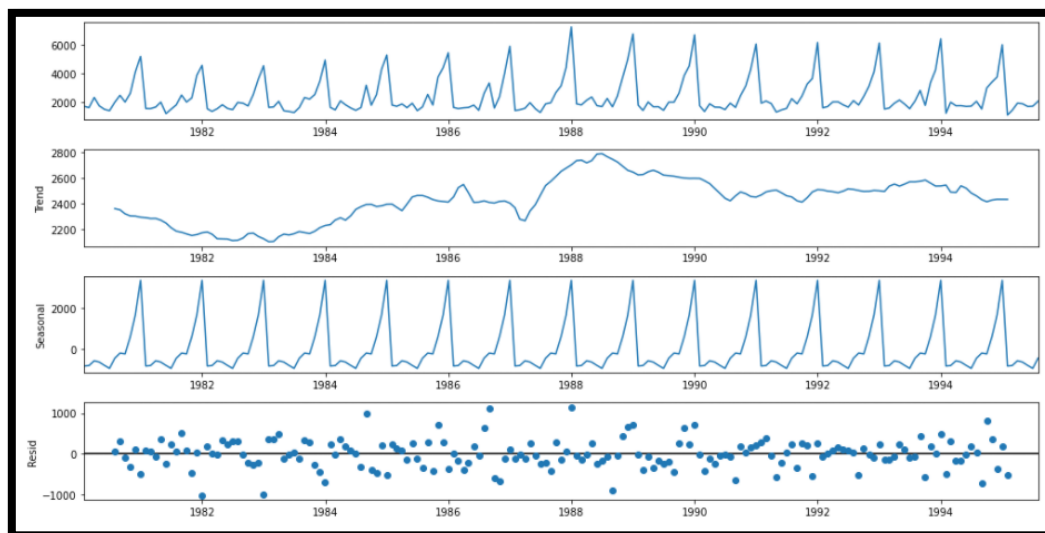


Fig 12: Data Decomposition Using Additive Model

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.

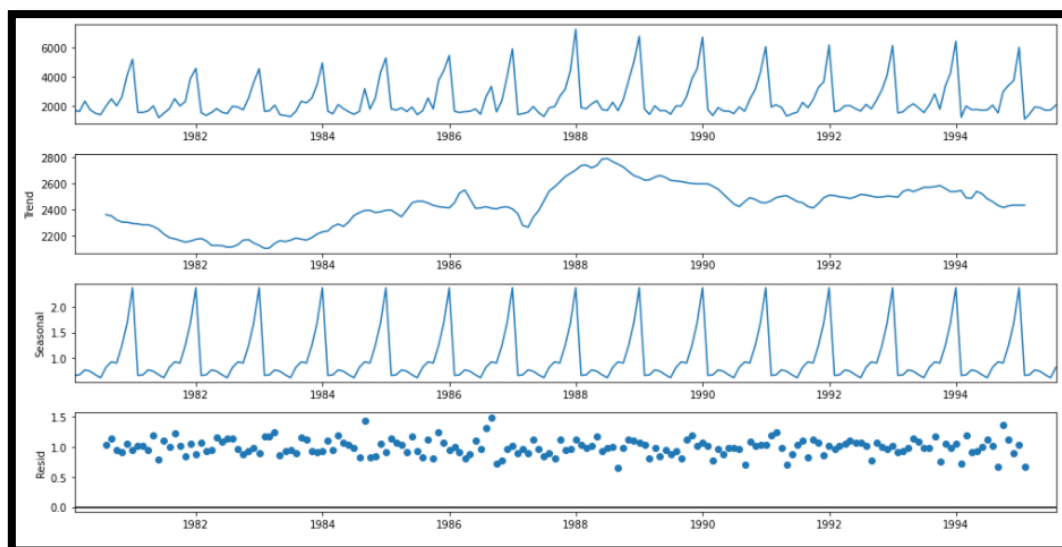


Fig 13: Data Decomposition Using Multiplicative Model

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		Last few rows of Training Data	
Sparkling		Sparkling	
Time_Stamp		Time_Stamp	
1980-01-31	1686	1990-08-31	1605
1980-02-29	1591	1990-09-30	2424
1980-03-31	2304	1990-10-31	3116
1980-04-30	1712	1990-11-30	4286
1980-05-31	1471	1990-12-31	6047

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		Last few rows of Test Data	
Sparkling		Sparkling	
Time_Stamp		Time_Stamp	
1991-01-31	1902	1995-03-31	1897
1991-02-28	2049	1995-04-30	1862
1991-03-31	1874	1995-05-31	1670
1991-04-30	1279	1995-06-30	1688
1991-05-31	1432	1995-07-31	2031

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 Train & Test Data:

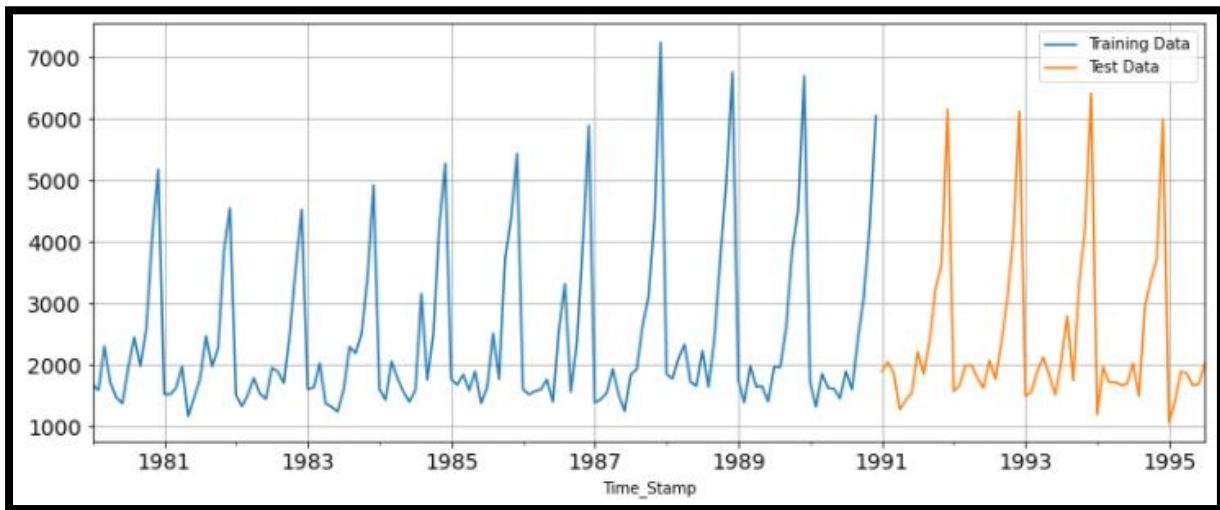


Fig 16: Train & Test Data Plot

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 Sparkling 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			Last few rows of Training Data		
Sparkling time			Sparkling time		
Time_Stamp			Time_Stamp		
1980-01-31	1686	1	1990-08-31	1605	128
1980-02-29	1591	2	1990-09-30	2424	129
1980-03-31	2304	3	1990-10-31	3116	130
1980-04-30	1712	4	1990-11-30	4286	131
1980-05-31	1471	5	1990-12-31	6047	132

Fig 18: Training Sample Data

Testing Sample Data:

First few rows of Test Data			Last few rows of Test Data		
Sparkling time			Sparkling time		
Time_Stamp			Time_Stamp		
1991-01-31	1902	133	1995-03-31	1897	183
1991-02-28	2049	134	1995-04-30	1862	184
1991-03-31	1874	135	1995-05-31	1670	185
1991-04-30	1279	136	1995-06-30	1688	186
1991-05-31	1432	137	1995-07-31	2031	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 sparkling wine for year 1991 units sold can be 2900 which can rise to 3200 units in 1996.

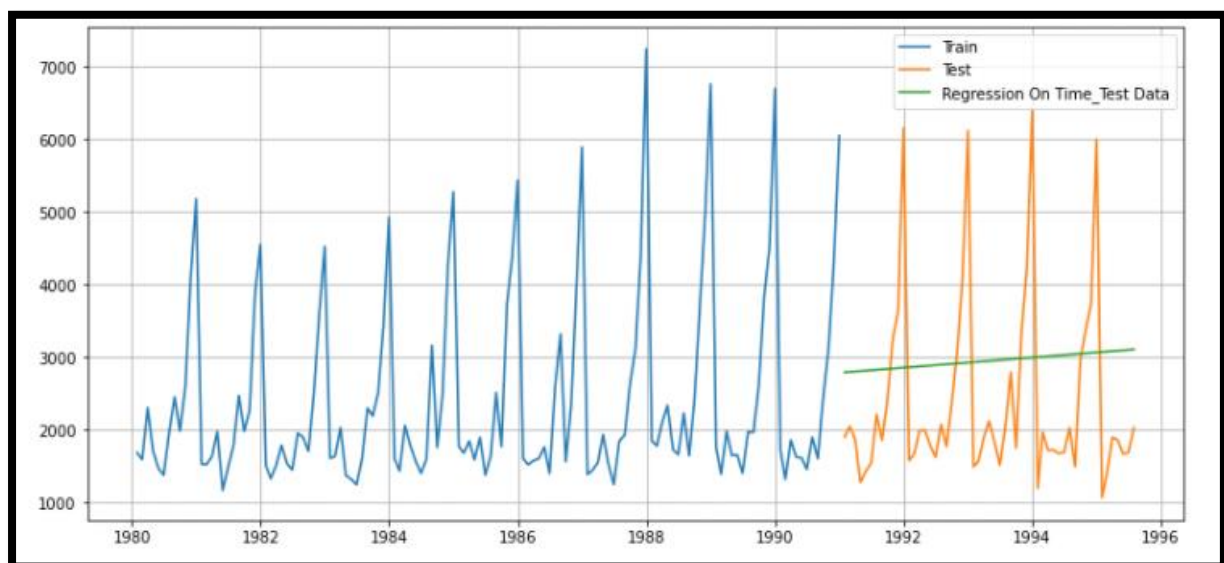


Fig 20: Plot of Train & Test Data after adding Time Instances

Model Evaluation:

For RegressionOnTime forecast on the Test Data, RMSE is 1389.135

Fig 21: RMSE of Test Data for Model Evaluation

The value of RMSE (1389.135) is quite high and since seasonality is also not taken care by model, therefore, this model is not suitable predictions on sparkling time series data.

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

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.

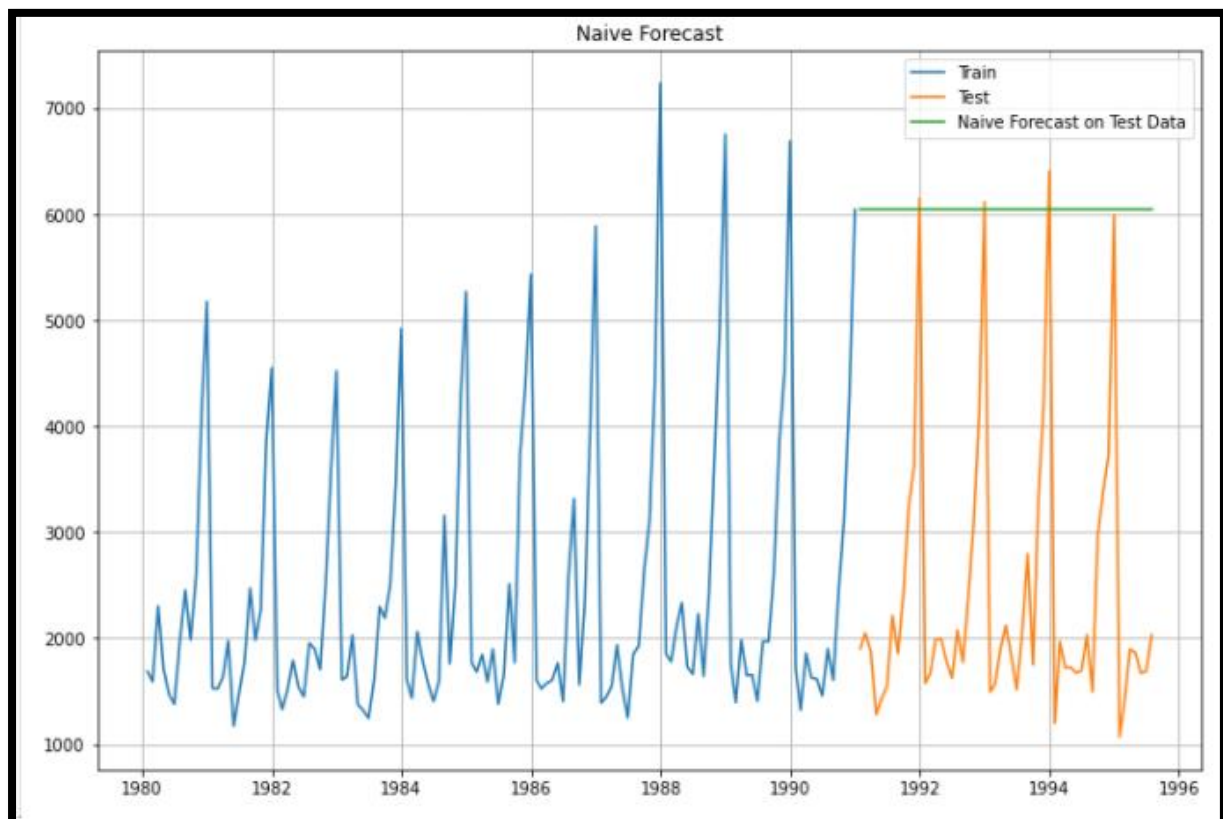


Fig 22: Naïve Bayes Based Time Series Plot (Test Data)

Model Evaluation:

For RegressionOnTime forecast on the Test Data, RMSE is 3864.279

Fig 23: RMSE of Test Data for Model Evaluation)

RSME is 3864.279, 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 2403.78 which is equal to the mean of the data as prevalent in EDA part showing data description.

	Sparkling	mean_forecast
Time_Stamp		
1991-01-31	1902	2403.780303
1991-02-28	2049	2403.780303
1991-03-31	1874	2403.780303
1991-04-30	1279	2403.780303
1991-05-31	1432	2403.780303

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 2403.

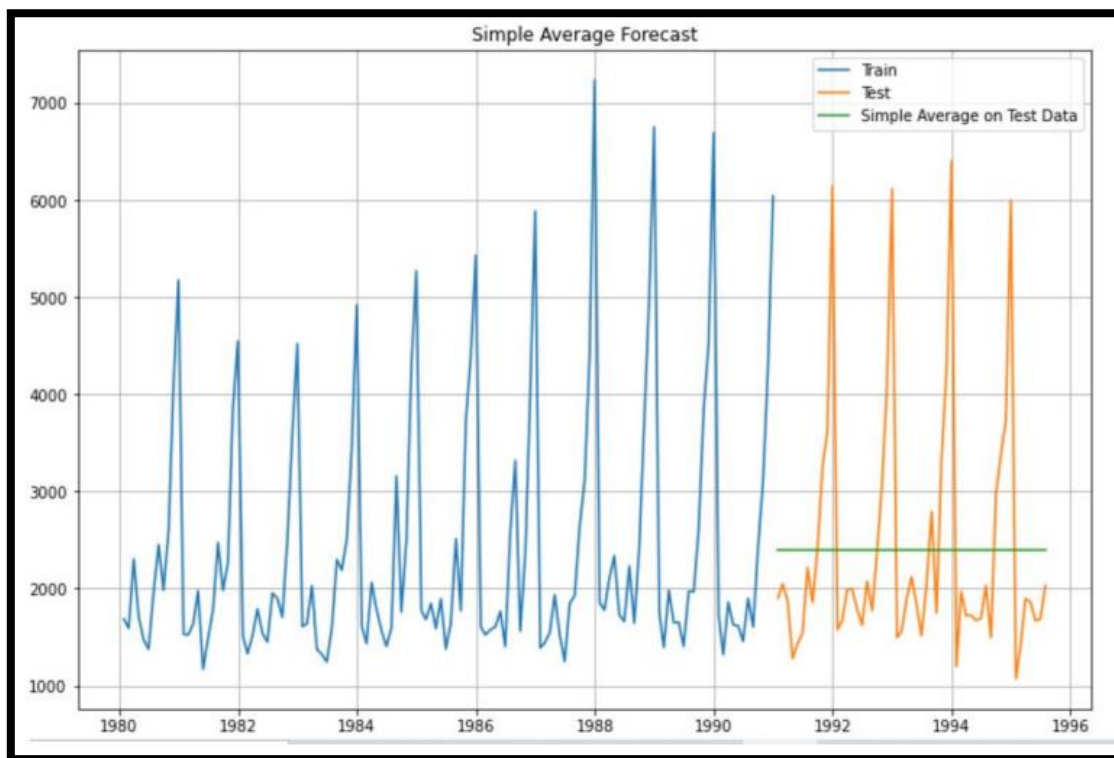


Fig 25: Simple Average Based Time Series Plot (Test Data)

Model Evaluation:

RMSE 1275.08 is quite high and without seasonality component in it we cannot consider this model for forecasting purposes.

For Simple Average forecast on the Test Data, RMSE is 1275.082

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.

	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
Time_Stamp					
1980-01-31	1686	NaN	NaN	NaN	NaN
1980-02-29	1591	1638.5	NaN	NaN	NaN
1980-03-31	2304	1947.5	NaN	NaN	NaN
1980-04-30	1712	2008.0	1823.25	NaN	NaN
1980-05-31	1471	1591.5	1769.50	NaN	NaN

Fig 27: Moving Average Sample Data

Moving Average Plot:

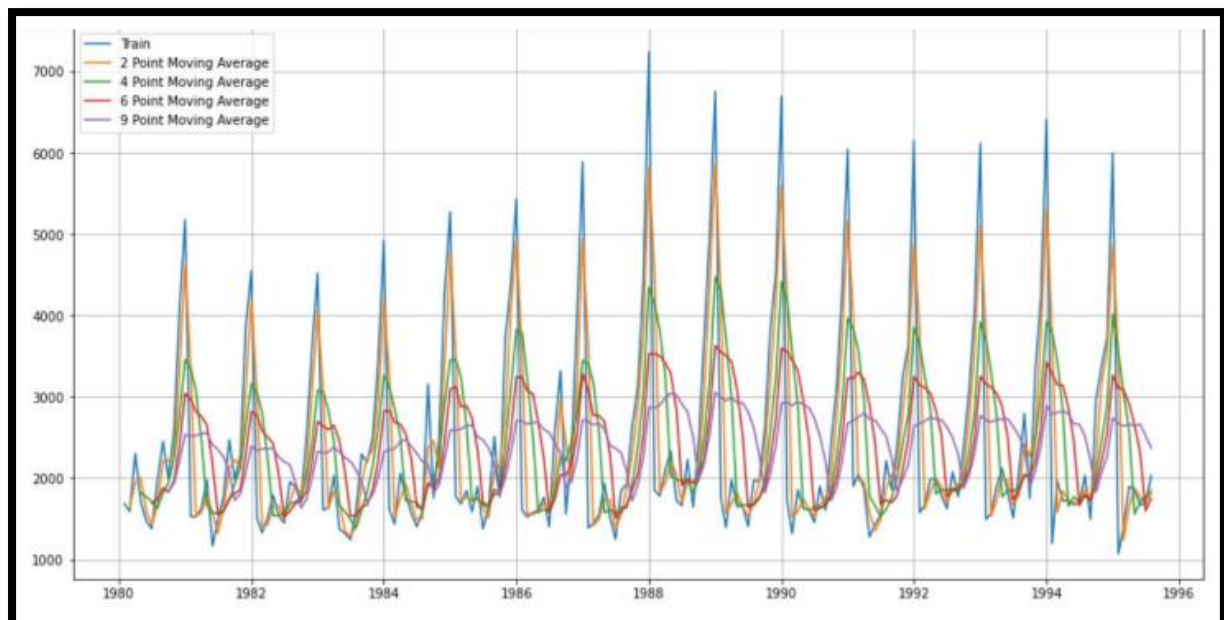


Fig 28: Moving Average Plot (Train Data)

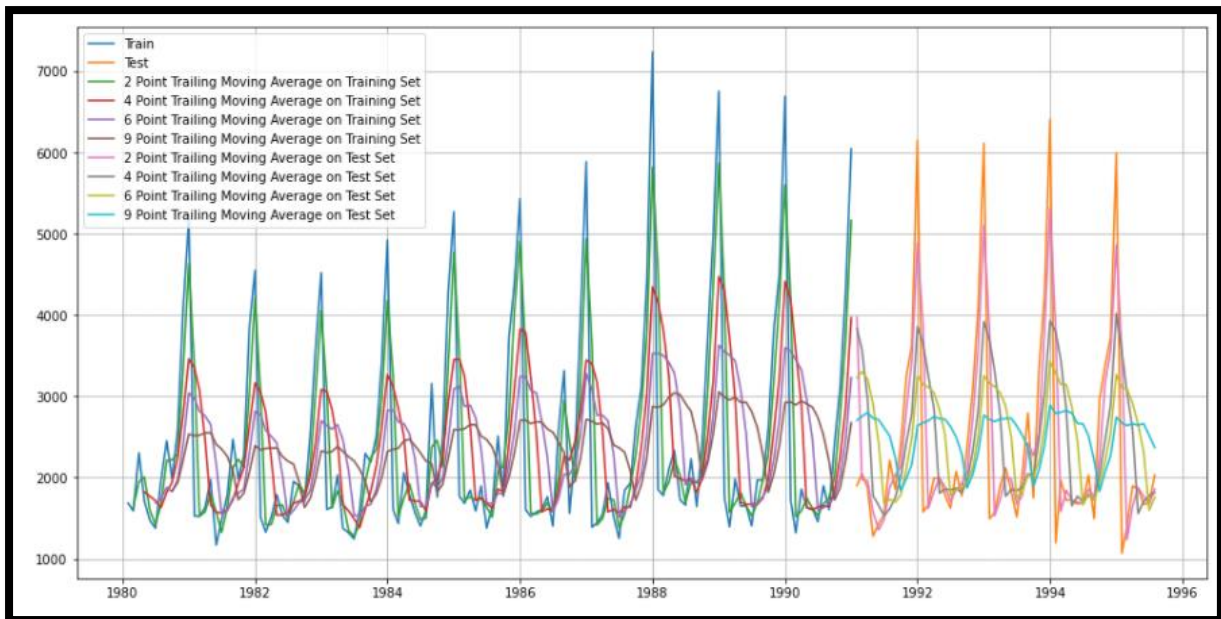


Fig 29: Moving Average Plot (Train & Test Data)

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 813.401
For 4 point Moving Average Model forecast on the Training Data,	RMSE is 1156.590
For 6 point Moving Average Model forecast on the Training Data,	RMSE is 1283.927
For 9 point Moving Average Model forecast on the Training Data,	RMSE is 1346.278

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.

<u>LR</u>	→	Linear Regression
<u>NB</u>	→	Naïve Bayes
<u>SA</u>	→	Simple Average
<u>MA</u>	→	Moving Average

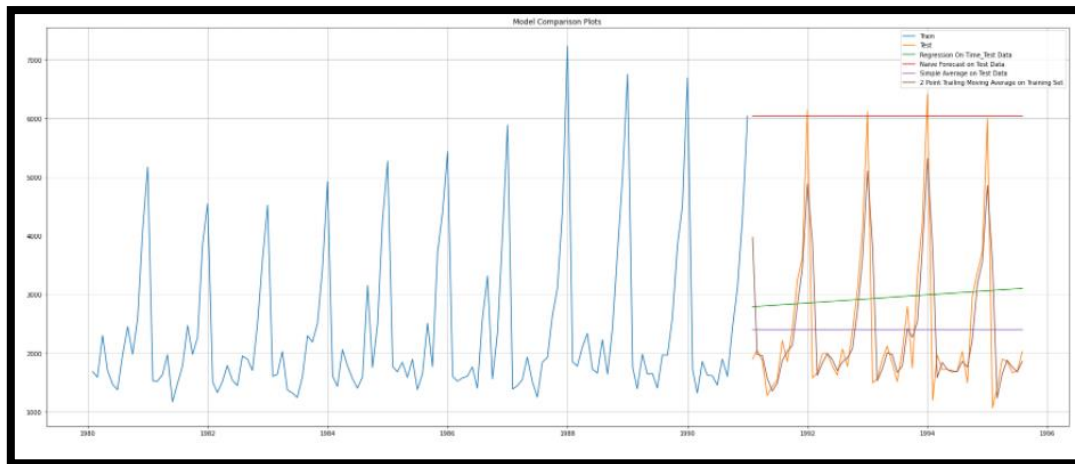


Fig 31: LR, NB, SA, MA Comparison Plot

Model 5: Simple Exponential Smoothing:

This model is build using the parmas as mentioned below.

```
{'smoothing_level': 0.04960736049406556,
'smoothing_trend': nan,
'smoothing_seasonal': nan,
'damping_trend': nan,
'initial_level': 2151.614314422547,
'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:

	Sparkling	predict
Time_Stamp		
1991-01-31	1902	2725.336037
1991-02-28	2049	2725.336037
1991-03-31	1874	2725.336037
1991-04-30	1279	2725.336037
1991-05-31	1432	2725.336037

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

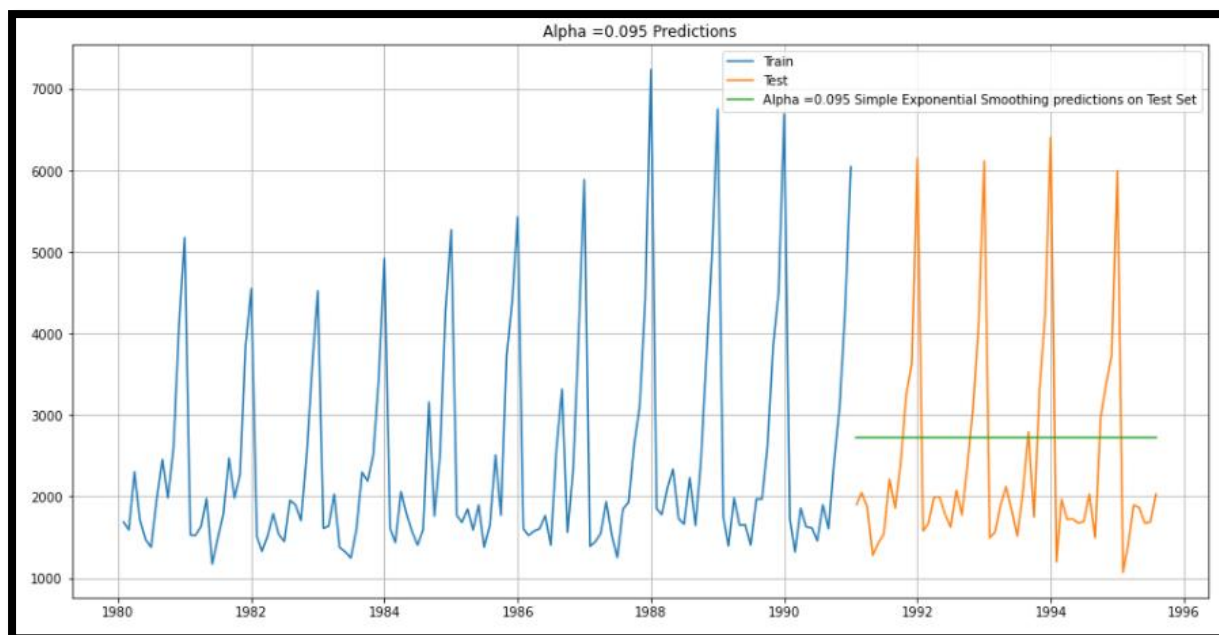


Fig 34: Predicted Value on the basis of Simple Exponential Smoothing

The graph below shows the simple trend line which is constant in predicting sales number as 2725.3306 units throughout the testing series data. This model ignores the seasonality.

Model Evaluation:

For Alpha = 0.095 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1316.135

Fig 35: RMSE Value for Model Evaluation

RMSE on test data is 1316.135 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.

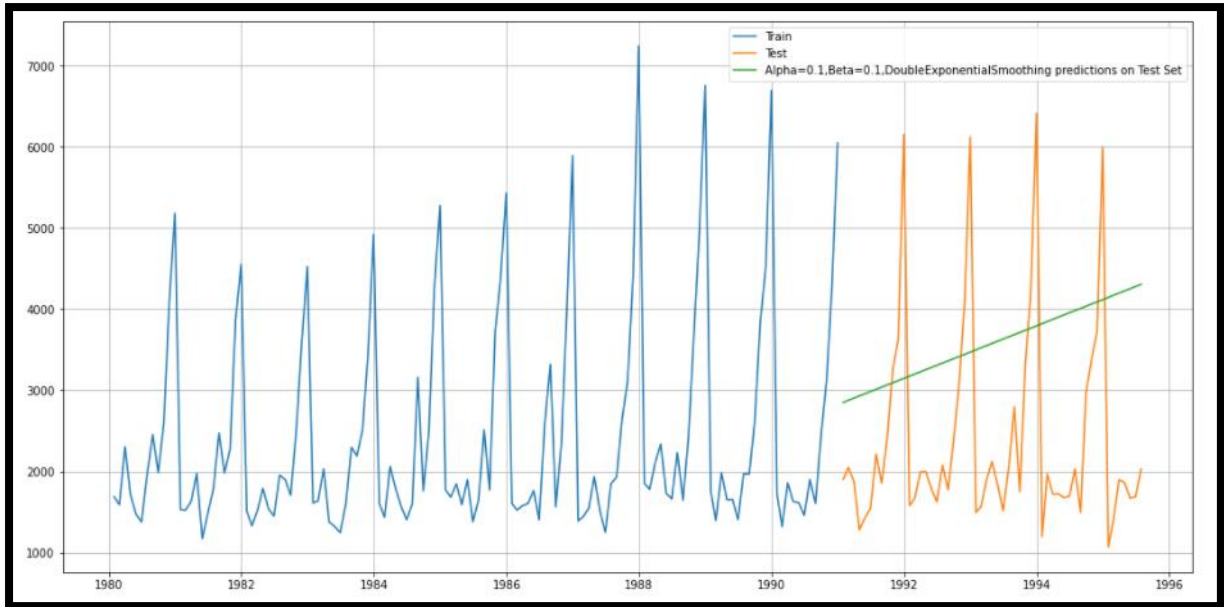


Fig 36: Train & Test Time Series Plot

Model Evaluation:

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	1382.520870	1778.564670
1	0.1	0.2	1413.598835	2599.439986
9	0.2	0.1	1418.041591	3611.763322
2	0.1	0.3	1445.762015	4293.084674
18	0.3	0.1	1431.169601	5908.185554

Fig 37: RMSE Values for Model Evaluation

The above RMSE values of Test data are shown in sorted order. 1778.564 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 γ .

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 γ).

Model Parameters:

```
{'smoothing_level': 0.11108840858679117,
'smoothing_trend': 0.061712060020663685,
'smoothing_seasonal': 0.3950814802151603,
'damping_trend': nan,
'initial_level': 1639.9088356475902,
'initial_trend': -11.928143593549056,
'initial_seasons': array([1.05065032, 1.02086214, 1.41078482, 1.20263518, 0.97315225,
0.96689379, 1.31724304, 1.70471609, 1.37289733, 1.81035002,
2.83962708, 3.60997333]),
'use_boxcox': False,
'lamda': None,
'remove_bias': False}
```

Fig 38: Triple Exponential Smoothing Parameters

Prediction Sample:

	Sparkling	auto_predict
Time_Stamp		
1991-01-31	1902	1577.208163
1991-02-28	2049	1333.663154
1991-03-31	1874	1745.977341
1991-04-30	1279	1630.435405
1991-05-31	1432	1523.306429

Fig 39: Triple Exponential Smoothing Predicted Sample

Above is the predicted data using Triple Exponential Smoothing model.

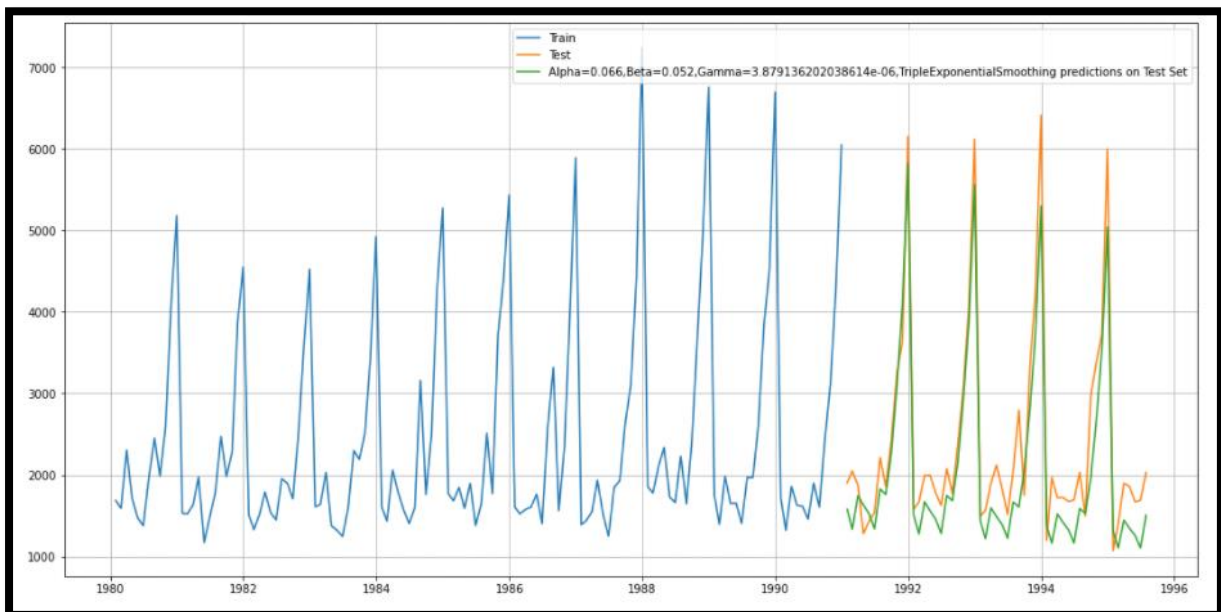


Fig 40: Train & Test Data at Chosen Parametric Values using Triple Exponential Smoothing (Prediction)

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 469.659 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 469.659

Fig 41: RMSE Value for Model Evaluation

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

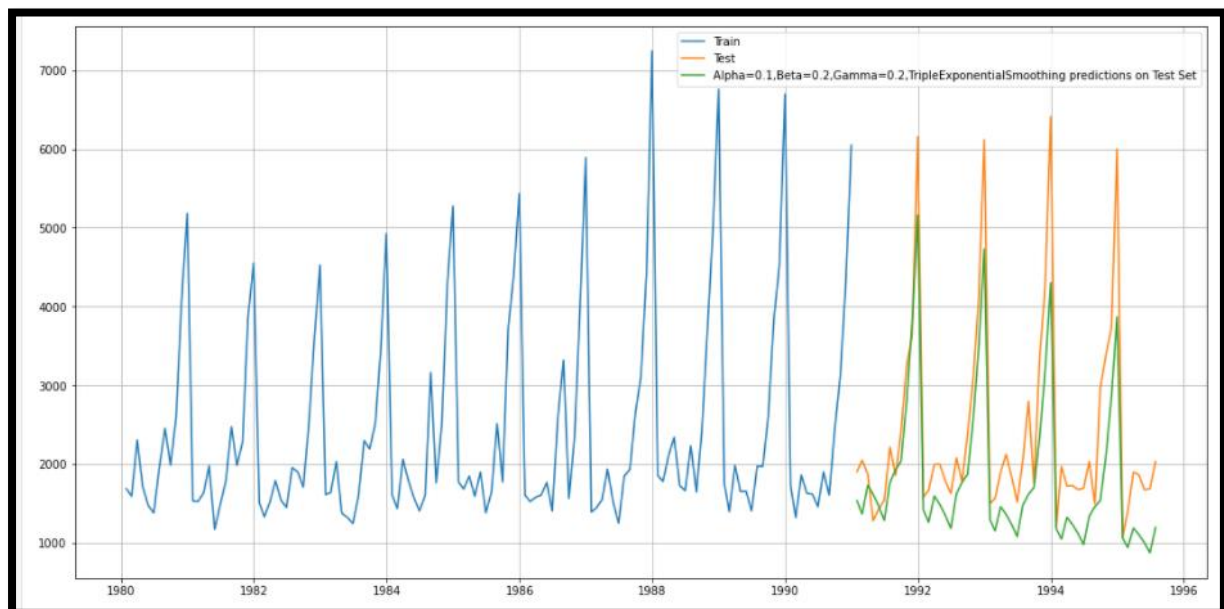


Fig 42: Train-Test Data at Chose Parametric Value Using Triple Exponential Smoothing (Prediction)

Model Evaluation:

The RMSE value with Alpha = 0.1, Beta = 0.2 and Gamma = 0.2 is 336.715.

Till now this is the lowest RMSE achieved amongst all the models.

Model Evaluation Comparison:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing	336.715250
Alpha=0.066,Beta=0.052,Gamma=3.879136202038614e-06,TripleExponentialSmoothing	469.659106
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
Alpha=0.099,SimpleExponentialSmoothing	1316.135411
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670
NaiveModel	3864.279352

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:

H_0 : The Time Series has a unit root and is thus non-stationary.

H_1 : 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 -1.798
DF test p-value is 0.7055958459932528
Number of lags used 12
```

Fig 44: Augmented Dickey-Fuller Test Check

The test gave the p-value to be 0.705 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 -44.912
DF test p-value is 0.0
Number of lags used 10
```

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

Difference Plot:

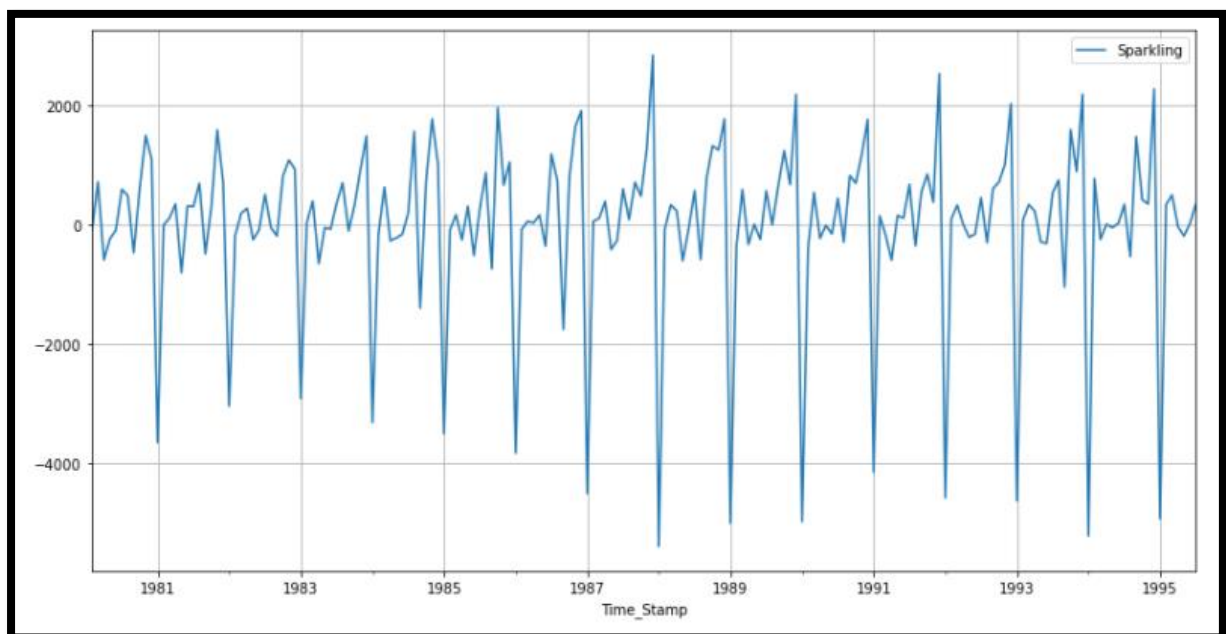


Fig 46: Difference Plot

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:

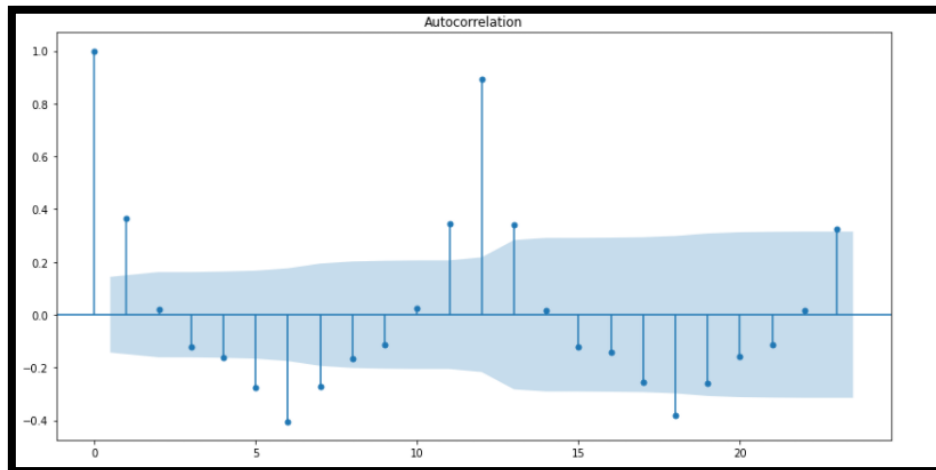


Fig 47: Auto-Correlation Plot

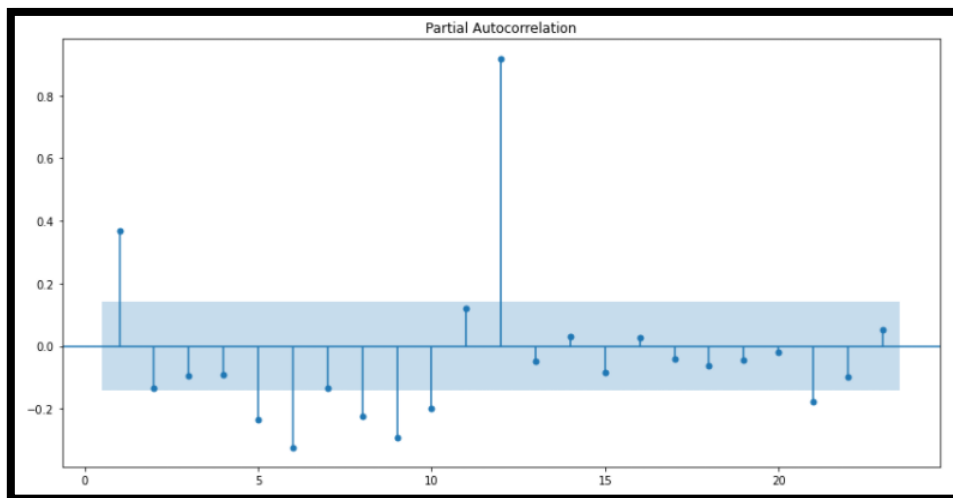


Fig 48: Partial Auto-Correlation Plot

For Train Data:

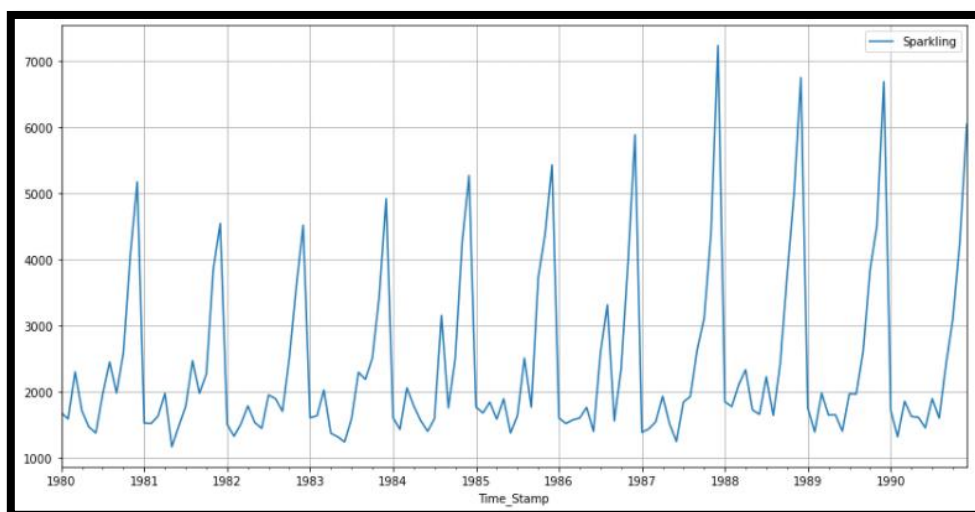


Fig 49: Train Data Plot

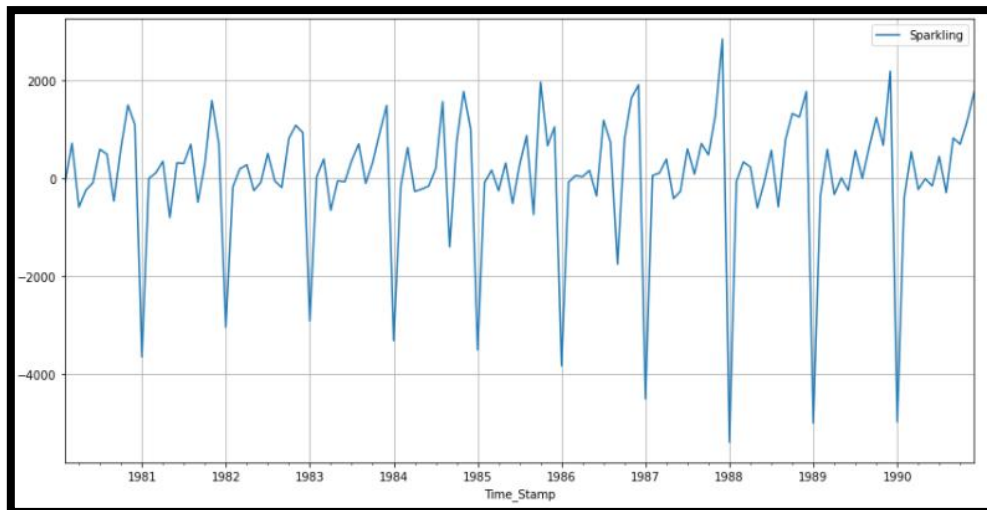


Fig 50: Train Data Plot After Dropping NA Values (Making the Data Stationary)

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, 2 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
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.458917
14	(3, 1, 2)	2230.781965
11	(2, 1, 3)	2232.893329
9	(2, 1, 1)	2233.777626

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

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

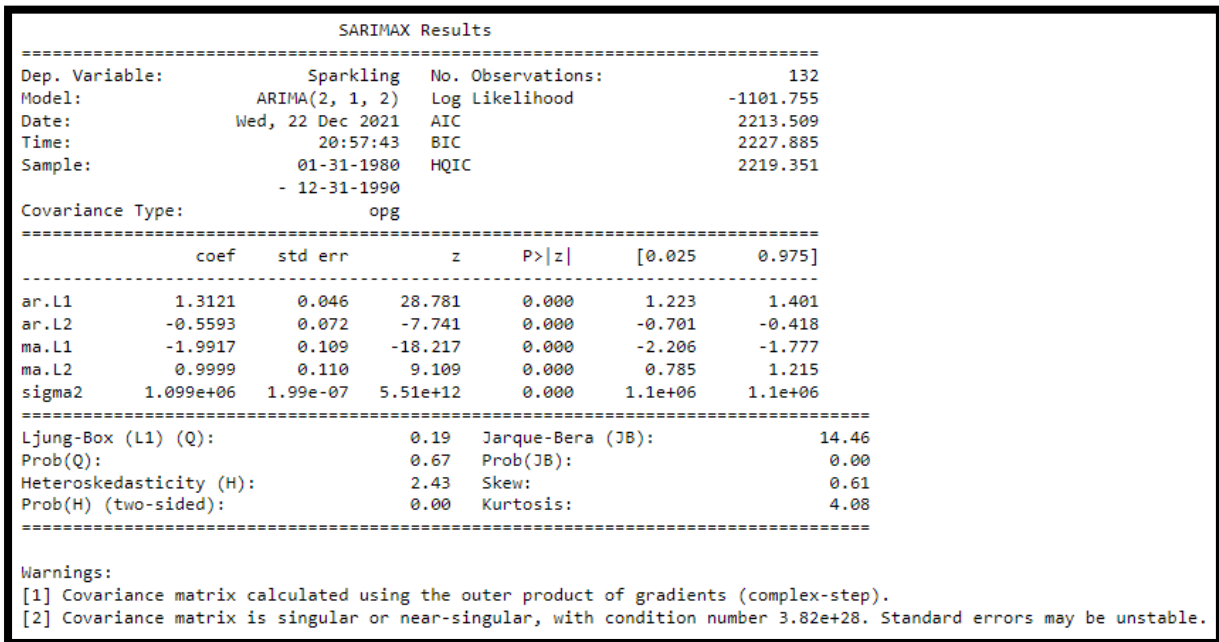


Fig 53: Arima Model Summary (Automated)

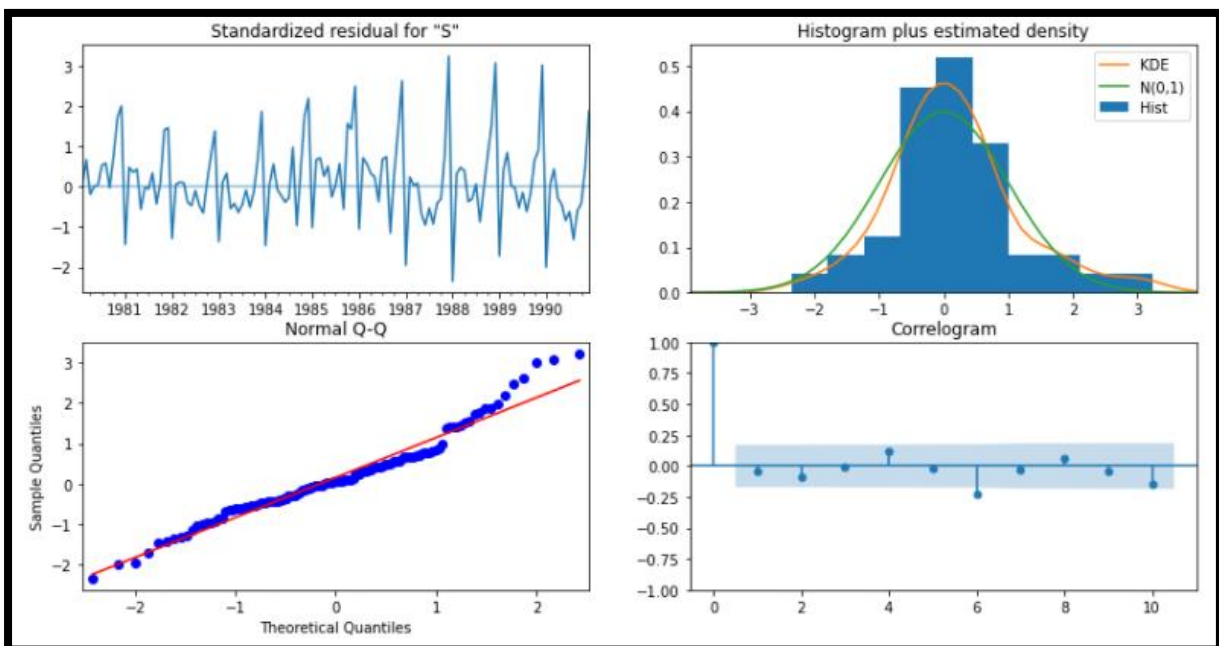


Fig 54: Diagnostic Plots

In ARIMA, AR = p = 2; MA = q = 2, as per our model. Since there are total 5 parameters (2 of AR, 2 of MA and 1 sigma), AIC value is 2213, we see 5 parameters listed with their values. The p-values for all are 0 which is less than 0.05 (claiming the necessary condition for being stationary).

Model Evaluation:

RMSE of ARIMA is 1299.97 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.

RMSE: 1299.979155446021
MAPE: 47.09989943547311

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)	1629.052487
59	(0, 1, 3)	(2, 0, 3, 6)	1633.327862
123	(1, 1, 3)	(2, 0, 3, 6)	1633.964217
191	(2, 1, 3)	(3, 0, 3, 6)	1634.547841
251	(3, 1, 3)	(2, 0, 3, 6)	1634.617364

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

the AIC values which we get are as follows. The lowest AIC returned is 1629.05 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.

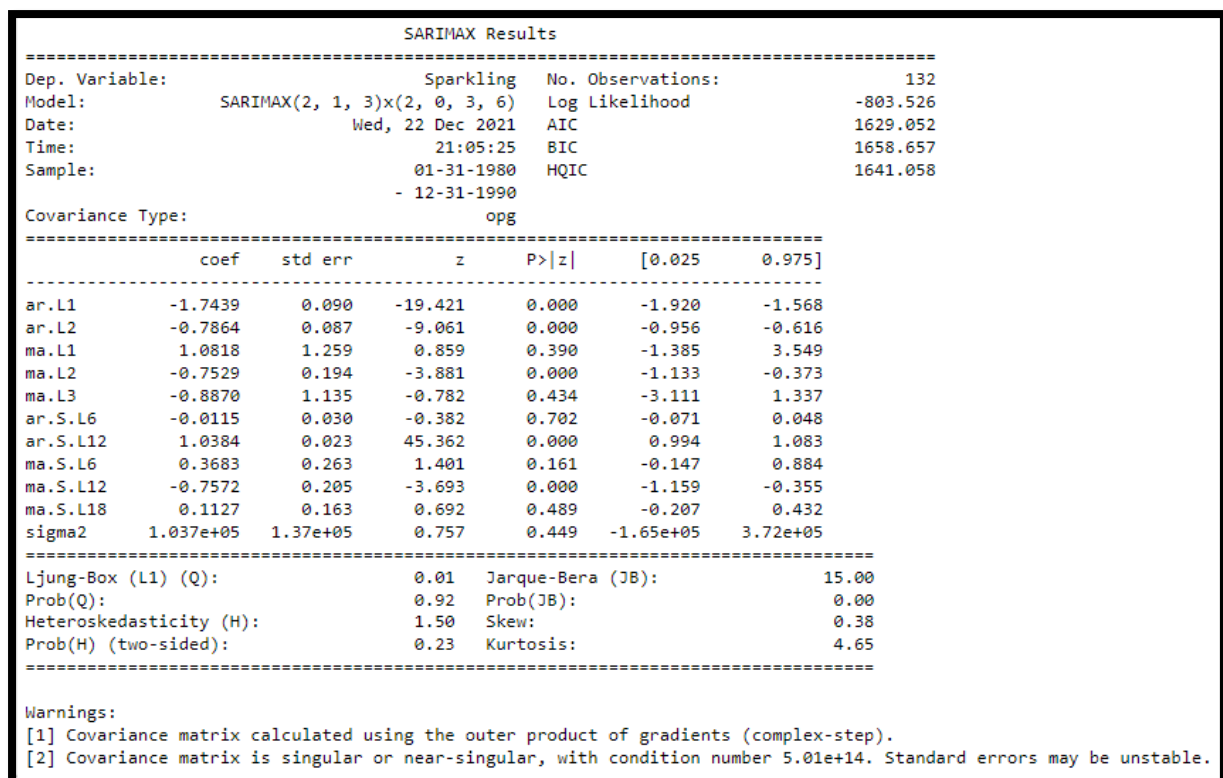


Fig 57: Sarima Model Summary (Automated)

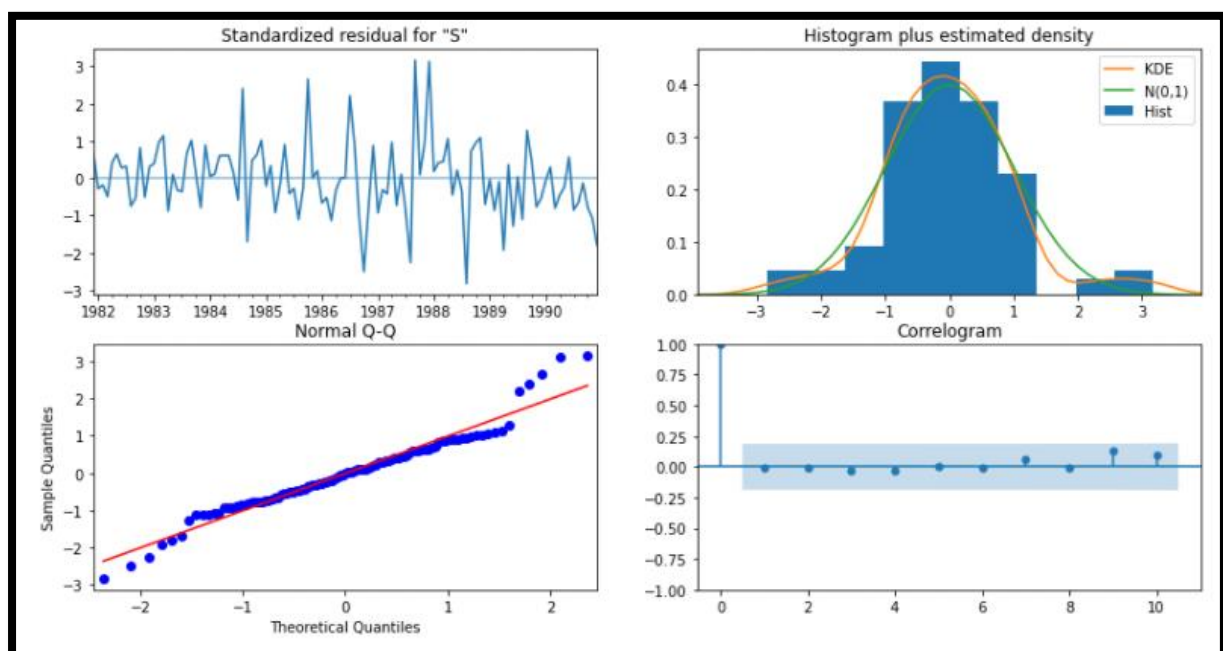


Fig 58: Diagnostic Plot

The model returns 11 parameters as seen above out of which 5. The AIC value returned in 1629.05 which has reduced with the seasonality factor as compared to the auto SARIMA model AIC.

Model Evaluation:

RMSE: 826.0533841576093
MAPE: 36.358173110127886

Fig 59: RMSE/MAPE For Model Evaluation

RMSE is 826.05 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:

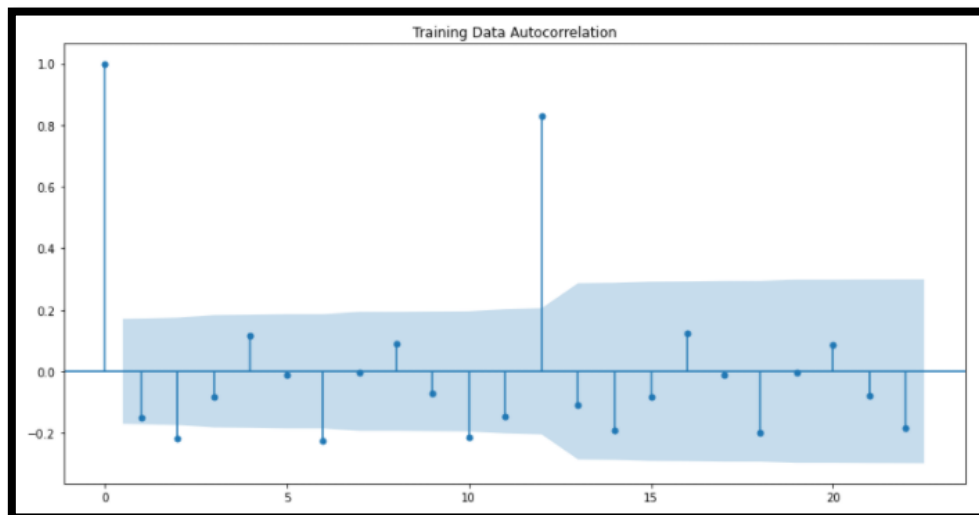


Fig 60: ACF of ARIMA

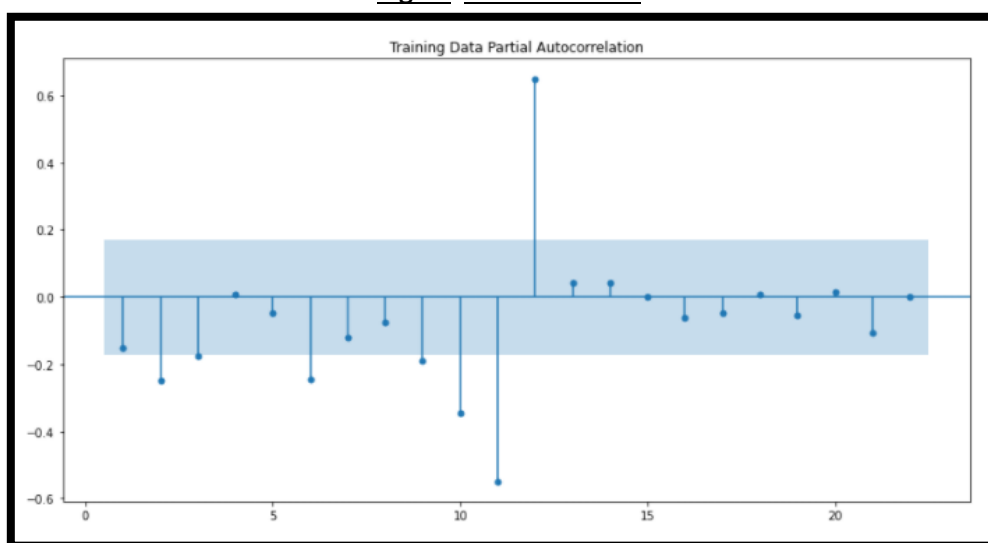


Fig 61: PACF of ARIMA

- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off is clearly 1 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 2,1,1 gives the following results for ARIMA

SARIMAX Results						
=====						
Dep. Variable:	Sparkling	No. Observations:	132			
Model:	ARIMA(2, 1, 1)	Log Likelihood	-1112.889			
Date:	Fri, 24 Dec 2021	AIC	2233.778			
Time:	22:22:46	BIC	2245.278			
Sample:	01-31-1980	HQIC	2238.451			
	- 12-31-1990					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

ar.L1	0.5194	0.119	4.359	0.000	0.286	0.753
ar.L2	-0.1782	0.196	-0.908	0.364	-0.563	0.206
ma.L1	-0.9993	0.667	-1.499	0.134	-2.306	0.307
sigma2	1.354e+06	8.31e+05	1.630	0.103	-2.74e+05	2.98e+06
=====						
Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	12.66			
Prob(Q):	0.76	Prob(JB):	0.00			
Heteroskedasticity (H):	2.77	Skew:	0.50			
Prob(H) (two-sided):	0.00	Kurtosis:	4.14			
=====						
Warnings:						
[1] Covariance matrix calculated using the outer product of gradients (complex-step).						

Fig 62: Manual Arima Summary

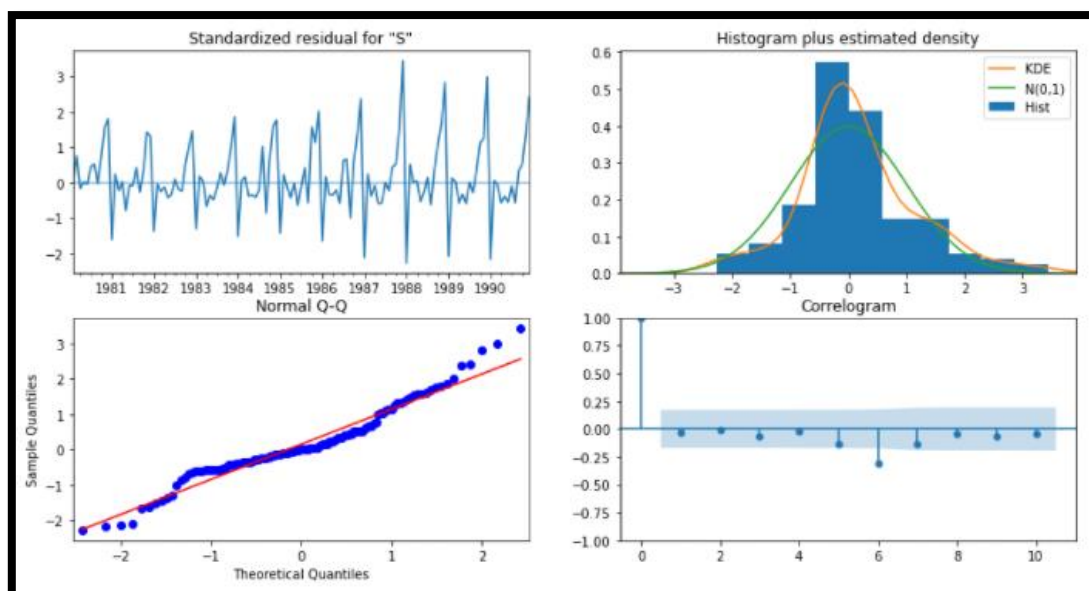


Fig 63: Diagnostic Plot Arima

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 2233.778 v/s an AIC value of 2213.509 for the auto ARIMA model. An increase of around 20-point hence the manual model is not a good model.

Model Evaluation:

RMSE: 1300.7213817353916
MAPE: 40.22566908011195

Fig 64: Diagnostic Plot ARIMA

RMSE is 1300. Still not preferable as a model to be used here in this project.

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.

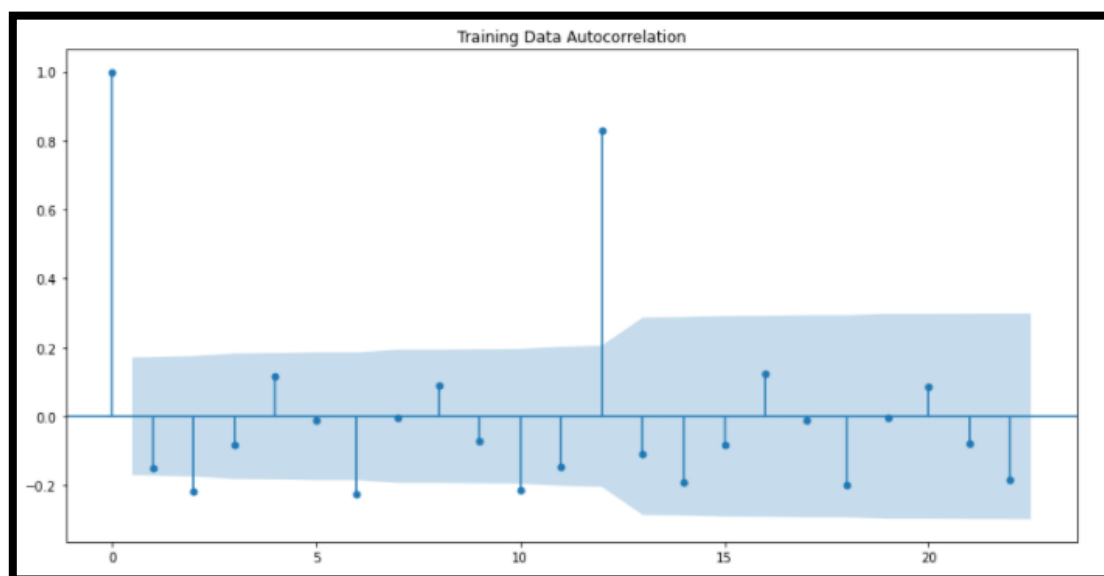


Fig 65: ACF of SARIMA

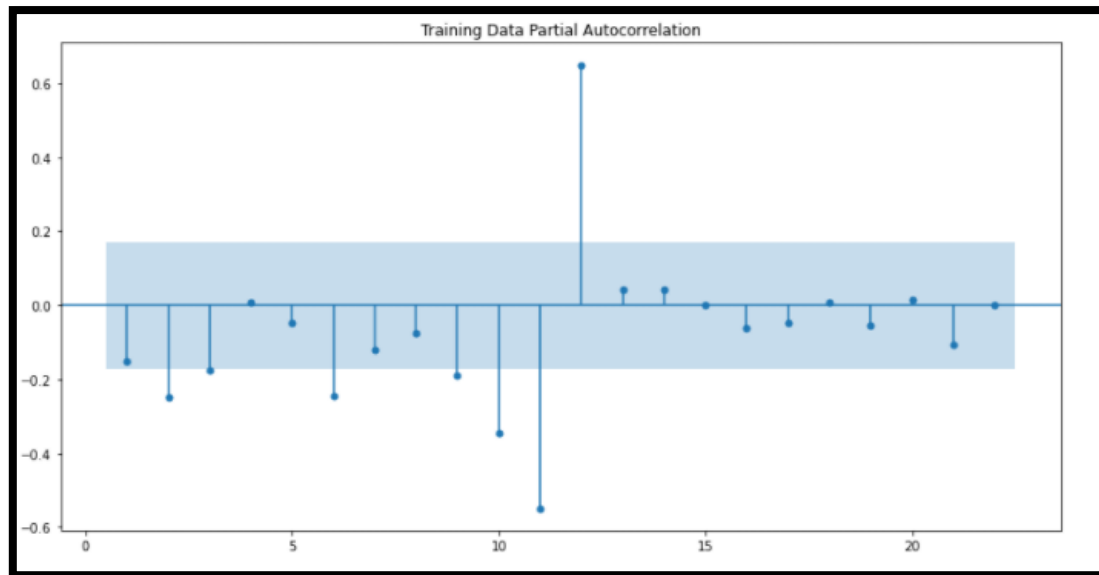


Fig 66: PACF of SARIMA

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:	Sparkling		No. Observations:	132		
Model:	SARIMAX(2, 1, 2)x(2, 0, 2, 6)		Log Likelihood	-855.682		
Date:	Fri, 24 Dec 2021		AIC	1729.364		
Time:	22:31:46		BIC	1754.146		
Sample:	01-31-1980		HQIC	1739.424		
	- 12-31-1990					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

ar.L1	-0.6771	0.289	-2.343	0.019	-1.243	-0.111
ar.L2	-0.0633	0.143	-0.444	0.657	-0.343	0.216
ma.L1	-0.1068	0.271	-0.394	0.694	-0.638	0.424
ma.L2	-0.6721	0.250	-2.689	0.007	-1.162	-0.182
ar.S.L6	-0.0033	0.025	-0.129	0.897	-0.053	0.047
ar.S.L12	1.0387	0.017	60.043	0.000	1.005	1.073
ma.S.L6	0.0602	0.151	0.399	0.690	-0.236	0.356
ma.S.L12	-0.6345	0.093	-6.794	0.000	-0.818	-0.451
sigma2	1.443e+05	1.73e+04	8.345	0.000	1.1e+05	1.78e+05
=====						
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	21.83			
Prob(Q):	0.97	Prob(JB):	0.00			
Heteroskedasticity (H):	2.69	Skew:	0.40			
Prob(H) (two-sided):	0.00	Kurtosis:	4.97			
=====						
Warnings:						
[1] Covariance matrix calculated using the outer product of gradients (complex-step).						

Fig 67: SARIMA Summary

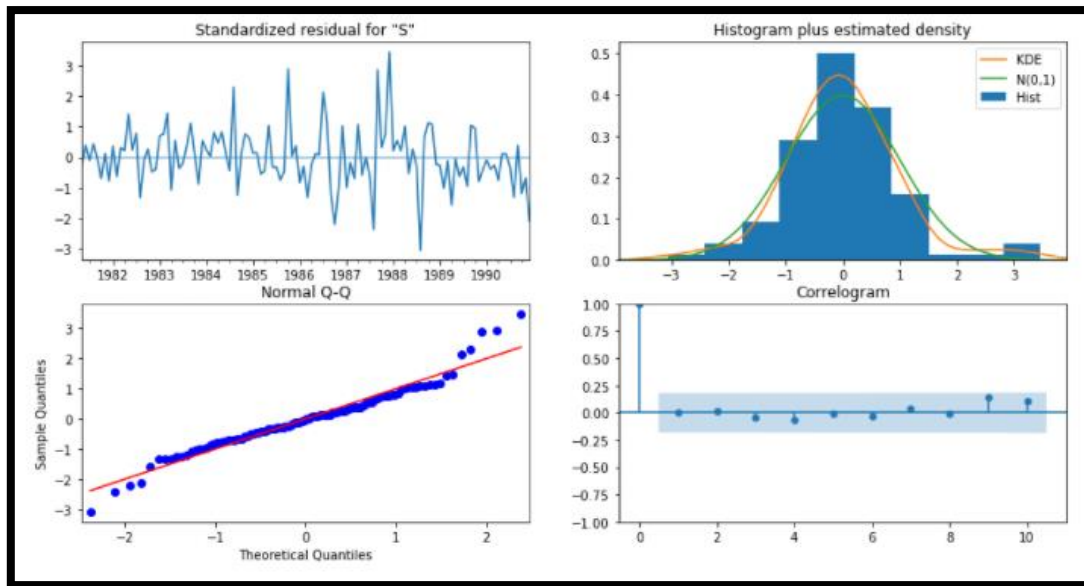


Fig 68: Diagnostic Plot

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 1729.364 and the automated SARIMA has 1629.052. A difference in 100-point. The manual SARIMA model here is not fit to be considered.

Model Evaluation:

RMSE: 655.5192385809944
MAPE: 28.580570867273625

Fig 69: RMSE/MAPE for Model Evaluation

The RMSE for manual SARIMA model is 655.519. 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 RMSE values on the Test Data:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponential Smoothing	336.715250
Alpha=0.066,Beta=0.052,Gamma=3.879136202038614e-06,TripleExponential Smoothing	469.659106
SARIMA(2,1,2)(2,0,2,6)	655.519239
2pointTrailingMovingAverage	813.400684
SARIMA(2,1,3)(2,0,3,6)	826.053384
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,2)	1299.979155
ARIMA(2,1,1)	1300.721382
Alpha=0.099,SimpleExponential Smoothing	1316.135411
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponential Smoothing	1375.393398
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponential Smoothing	1778.564670
NaiveModel	3864.279352

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 (336) whereas the Naïve Model is having the highest RMSE value for the test model (3864).

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.

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.

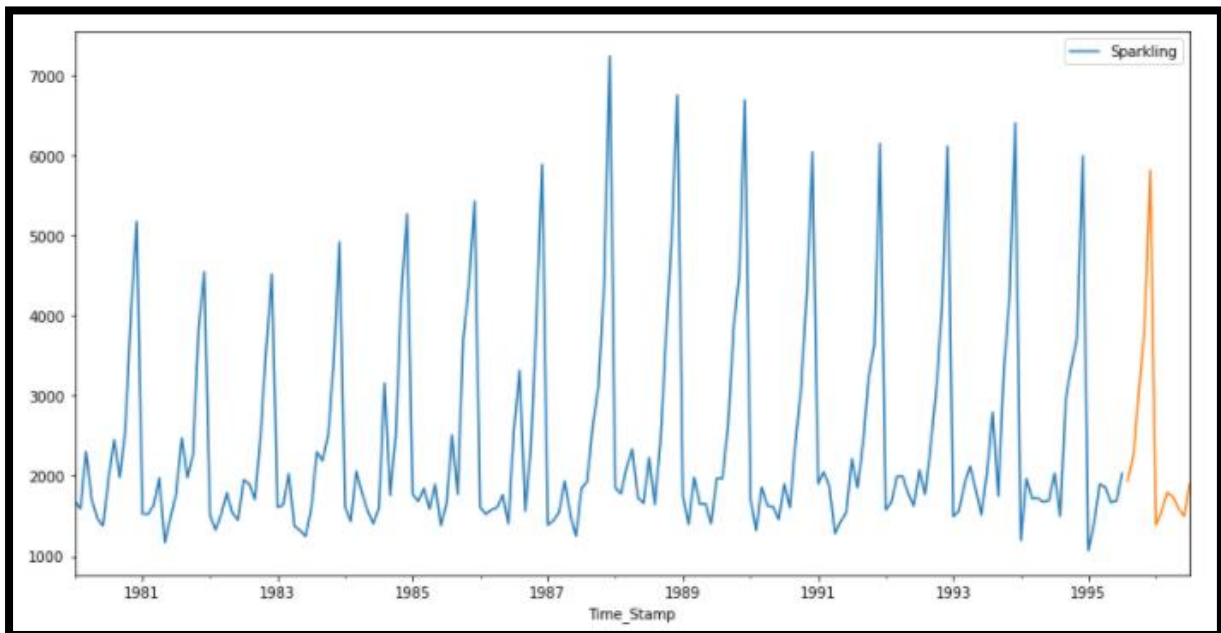


Fig 71: Full Model Time Series Plot

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

	lower_CI	prediction	upper_ci
1995-08-31	1225.011928	1939.502343	2653.992758
1995-09-30	1547.548022	2262.038436	2976.528851
1995-10-31	2358.711507	3073.201922	3787.692337
1995-11-30	3157.670950	3872.161365	4586.651779
1995-12-31	5105.490401	5819.980816	6534.471230

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 363.604

RMSE: 363.60434345947067

Fig 73: RMSE for Model Evaluation

Prediction Plot:

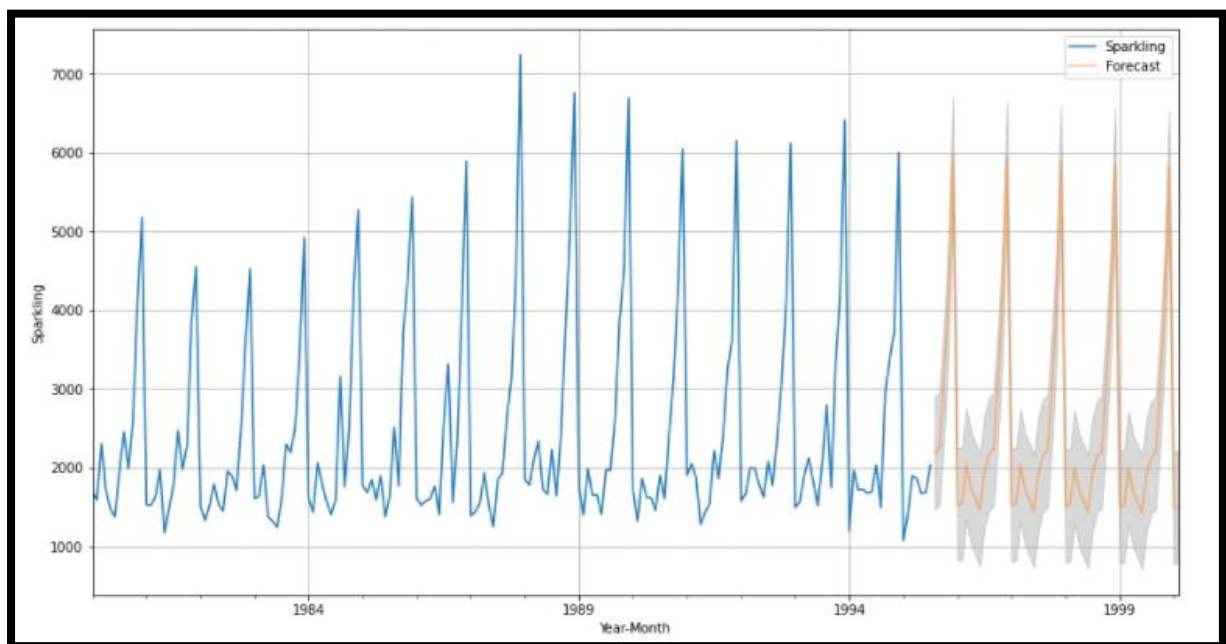


Fig 74: Full Model Prediction Plot

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 very high. So, we didn't use it to build the final model.
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).

Sorted RMSE values on the Test Data:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponential Smoothing	336.715250
Alpha=0.066,Beta=0.052,Gamma=3.879136202038614e-06,TripleExponential Smoothing	469.659106
SARIMA(2,1,2)(2,0,2,6)	655.519239
2pointTrailingMovingAverage	813.400684
SARIMA(2,1,3)(2,0,3,6)	826.053384
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,2)	1299.979155
ARIMA(2,1,1)	1300.721382
Alpha=0.099,SimpleExponential Smoothing	1316.135411
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponential Smoothing	1375.393398
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponential Smoothing	1778.564670
NaiveModel	3864.279352

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 sparkling 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 2402 unit during all seasons and minimum of 1070 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.