Project:TIME SERIES FORECASTING

DSBA

Submitted by:

Name: Hima Jose Paliakkara

Batch: 2021-2022

Table of Contents

Contents

Problem – Sales data of Sparkling Wine

List of figures

List of Tables

Problem – Sales data of Sparkling Wine

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

Data set for the Problem: **Sparkling.csv**

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

Answer:

Summary of dataset:

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Table 1.1 Read the data

- The dataset is the monthly sales of rose wine of a company from 1980 to 1995. It is a time series data with frequency of one month.
- The dataset contains 187 rows and 1 column
- The dataset has no-null values
- No duplicates present

Plot the Time Series to understand the behaviour of the data

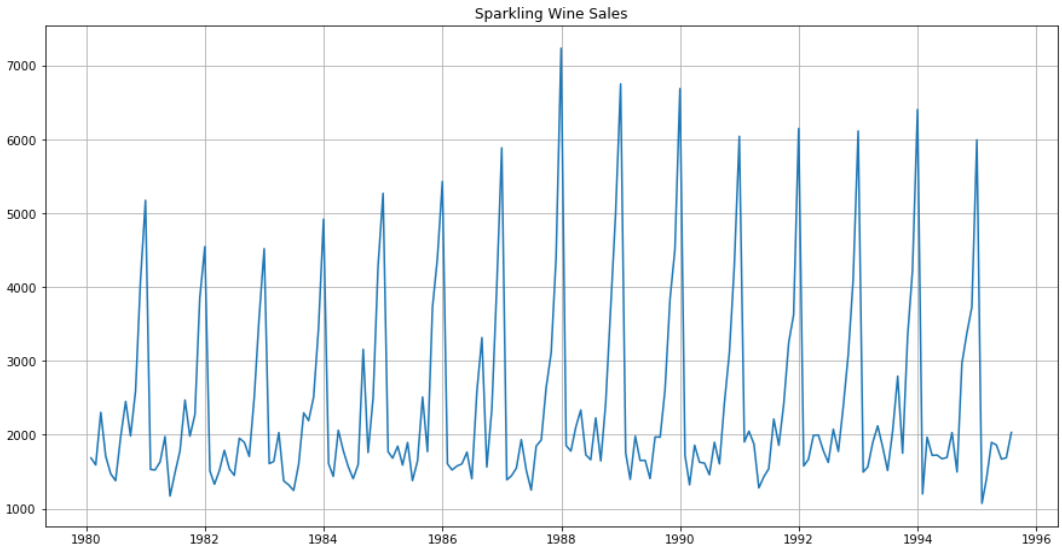


Figure 1.1 Time Series Plot

The sales for Sparkling wines is showing no much trend. There is a certain seasonality element that is visible in the graph. We will explore the trend and seasonality further during decomposition, where we will be able to view a much-detailed report on these two factors.

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

Answer:

	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

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Table 2.1 Summary & info of data

- The dataset contains 187 rows and 1 column
- The dataset has no-null values
- No duplicates present

Yearly Box-plot

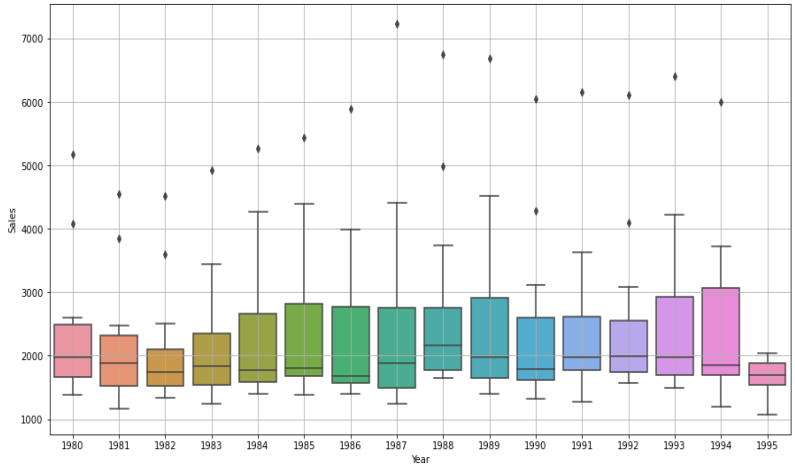


Figure 2.1 Yearly-box Plot

The highest sales for Sparkling Wine seem to happen in the year 1994 and the lowest in the year 1982. The Sparkling wine sales appear to be going down from the year 1980 and have started increasing from the year 1983. The variation in Sparkling Wine sales seem to be increasing for the period 1983-1986. There are outliers in the yearly sales data, however as it is a Time Series, we can ignore the outlier data.

Monthly Box-plot

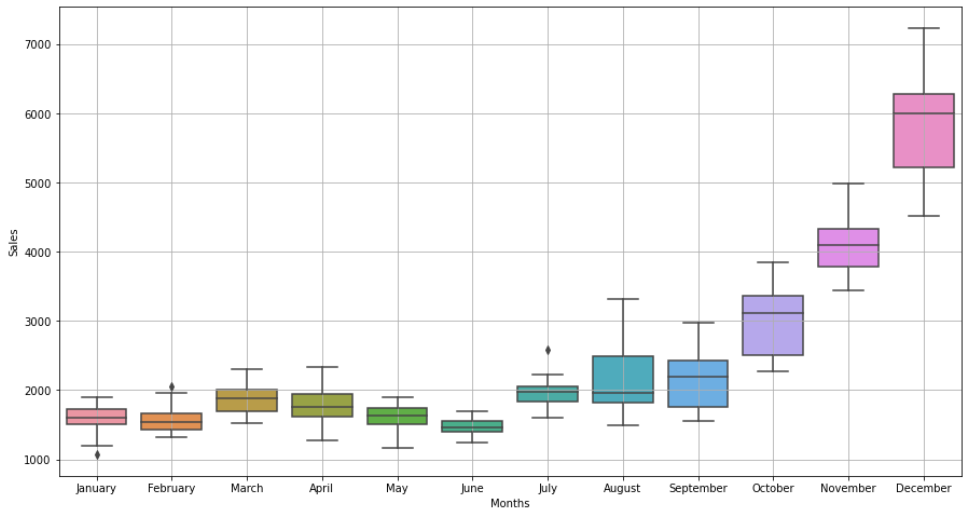


Figure 2.2 Monthly-box Plot

We can clearly see that there is a seasonality element visible in the dataset. There is an increasing sales trend in the last quarter of the year. The sales are relatively low in first two quarters, slowly picks up during the third quarter and goes on a rise till the end of the year. Few outliers are present.

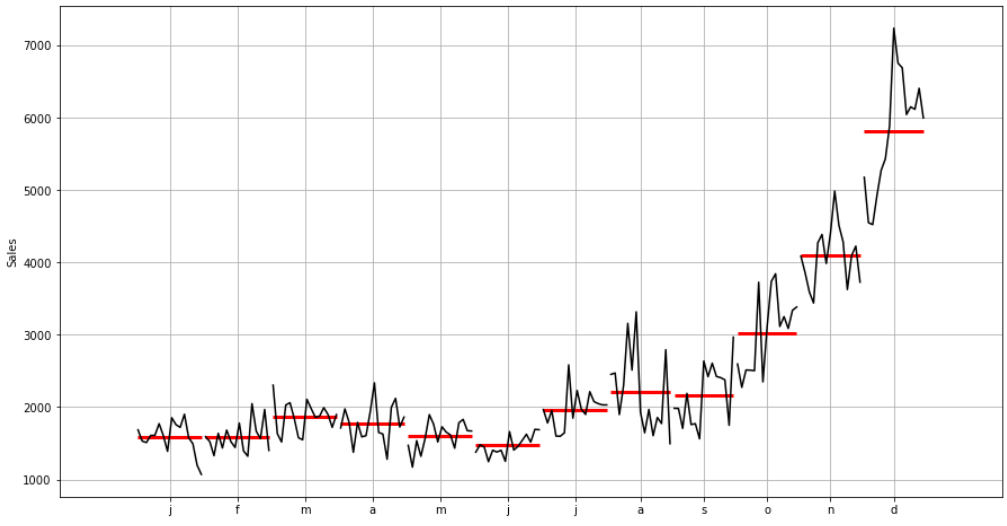


Figure 2.3 Monthly Plot

Above plot is another way of plotting the monthly median, highest and lowest values. December has the highest median also the highest peak.

Monthly Sales across Years

The monthly sales across years can be seen in the following Pivot Tables and the associated graphs.

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

Table 2.1 Pivot table Monthly sales across years

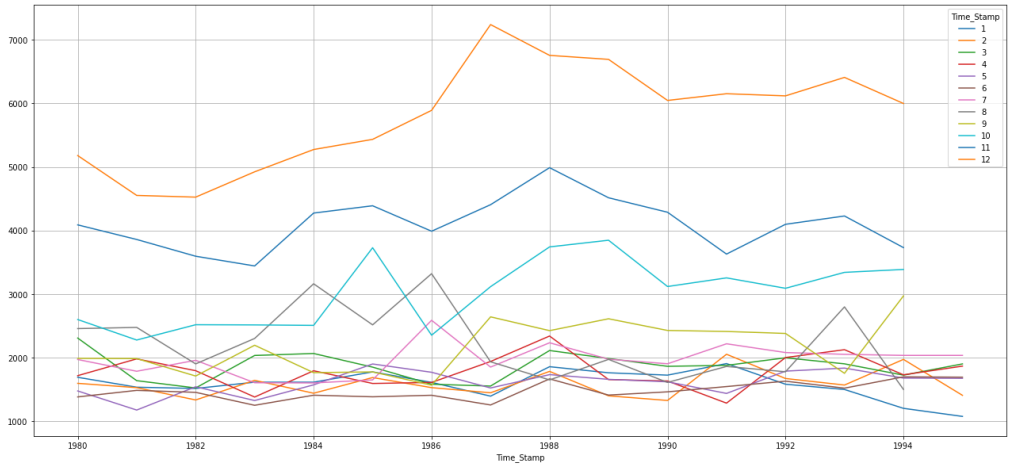


Figure 2.4 Plot of Monthly sales across years

- December records have the high number of wine sales followed by November.
- May, June and July have low number of wine sales.
- We can observe a seasonality element in the graph

Yearly Sum of observations

The yearly sum of sales numbers can be observed in the following tables and graphs:

Time_Stamp	Sparkling
1980-12-31	28406
1981-12-31	26227
1982-12-31	25321
1983-12-31	26180
1984-12-31	28431

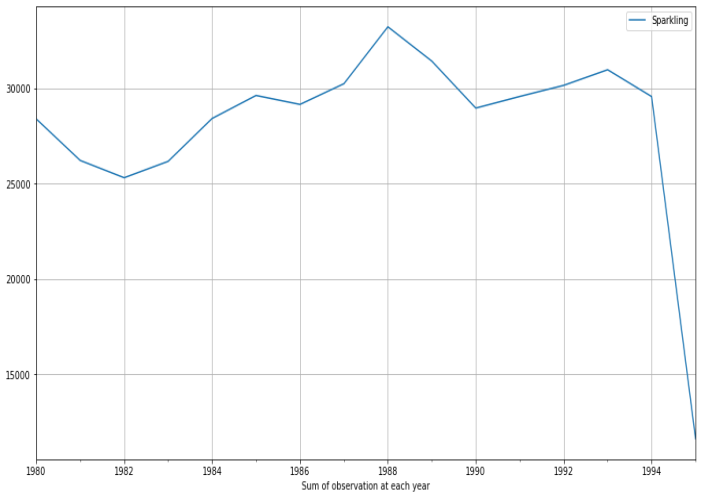


Figure 2.5 Sum of yearly observations

The sales figures for Sparkling wine show a dip initially with sales picking up from the year 1982 up to the year 1988 and then showing another dip in the sales. The steep drop post 1994 for Sparkling wine is because of the relatively less (till July only) data available for the year 1995.

Sum of Observations of each Quarter

The quarterly sum of sales numbers can be observed in the following tables and graphs:

Time_Stamp	Sparkling
1980-03-31	5581
1980-06-30	4560
1980-09-30	6403
1980-12-31	11862
1981-03-31	4686

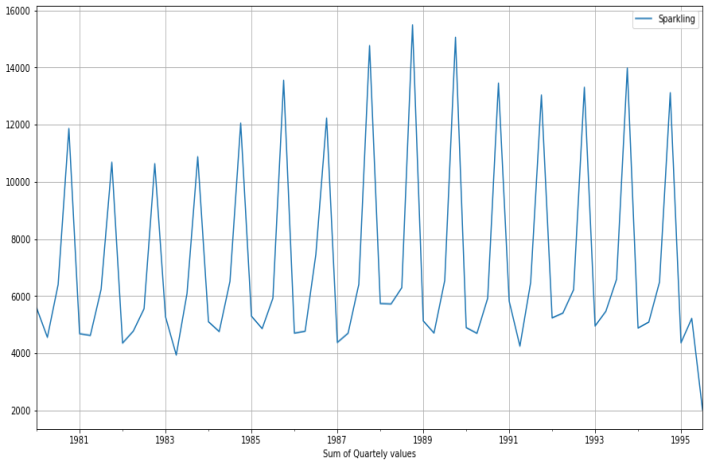


Figure 2.6 Sum of Quarterly observations

From the above tables and graphs we can find that the Quarterly sales shows trend for Sparkling wine and there is a slight element of seasonality in the time series dataset.

Empirical Cumulative Distribution

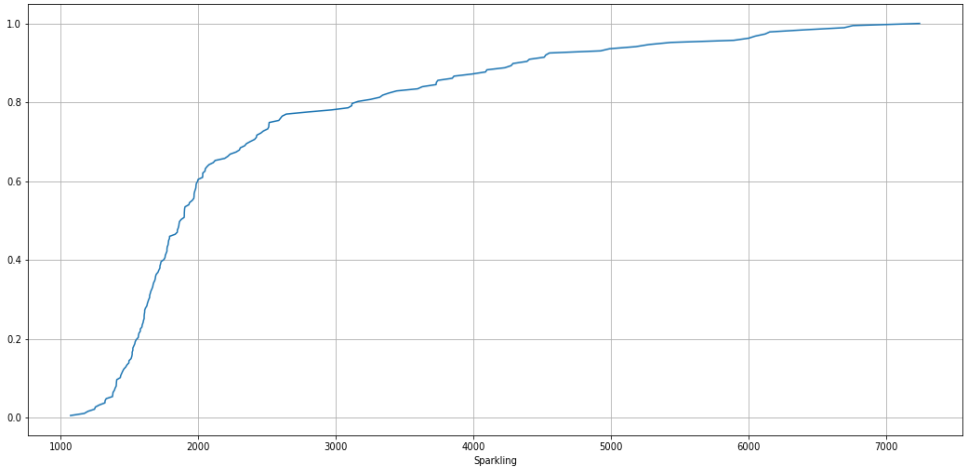


Figure 2.7 Empirical cumulative Distribution

Average Sales per month and the month-on-month percentage change of Sales

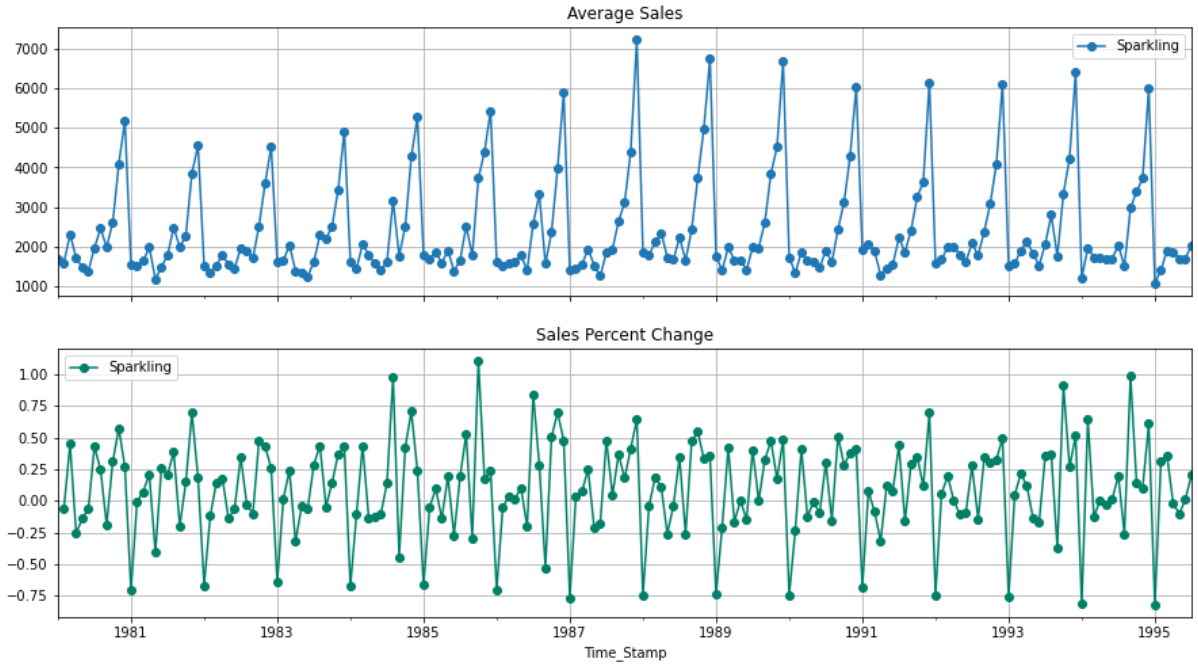


Figure 2.8 Average Sales per month & percentage change of sales

Decomposition of Time Series

This method is based on extraction of individual components of time series. There are various forces that may affect the observations in a time series. The three important components are:

- Trend (Long term movement)
- Seasonal component: Intra-year stable fluctuations repeatable over the entire length of the series
- Irregular component (Random movements)

Trend and seasonal components are part of systematic components of time series.

Additive Model: Yt= Tt+ St+ It is considered when the resultant series is the sum ofthe components.

Multiplicative Model: Yt= Tt* St* It is considered when the resultant time series is the product of the components

Additive Decomposition of Sparkling

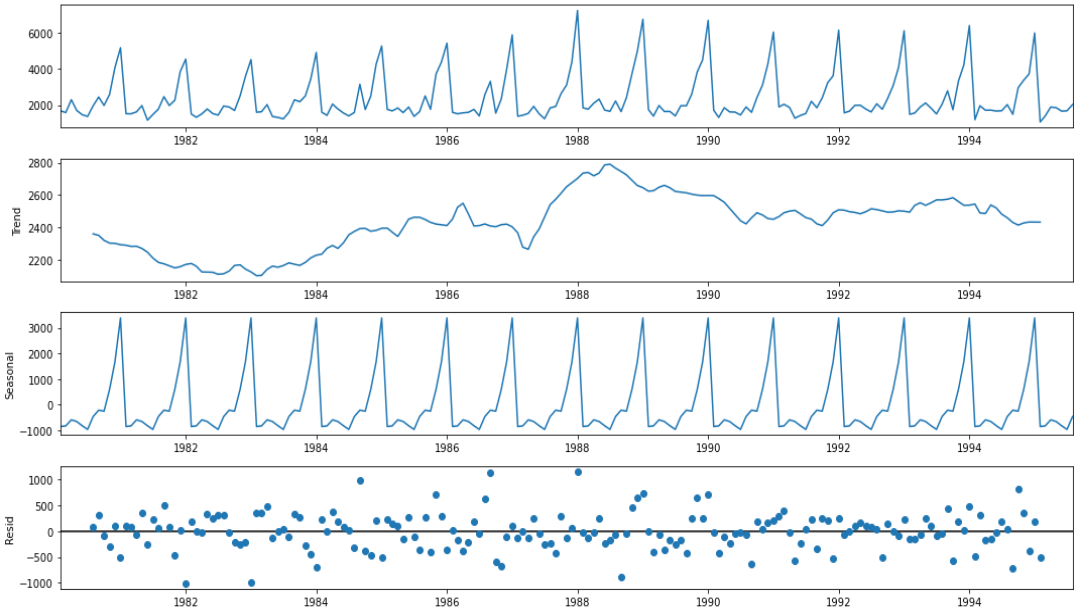


Figure 2.9 Additive decomposition

Time_Stamp	Trend
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	2360.667
1980-08-31	2351.333
1980-09-30	2320.542
1980-10-31	2303.583

Time_Stamp	Seasonality
1980-01-31	-854.261
1980-02-29	-830.351
1980-03-31	-592.357
1980-04-30	-658.491
1980-05-31	-824.416
1980-06-30	-967.434
1980-07-31	-465.502
1980-08-31	-214.333
1980-09-30	-254.677
1980-10-31	599.770

Time_Stamp	Residual
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	70.836
1980-08-31	315.999
1980-09-30	-81.864
1980-10-31	-307.353

Table 2.2 Trend, Seasonality, Residual (Additive)

Multiplicative Decomposition of Sparkling

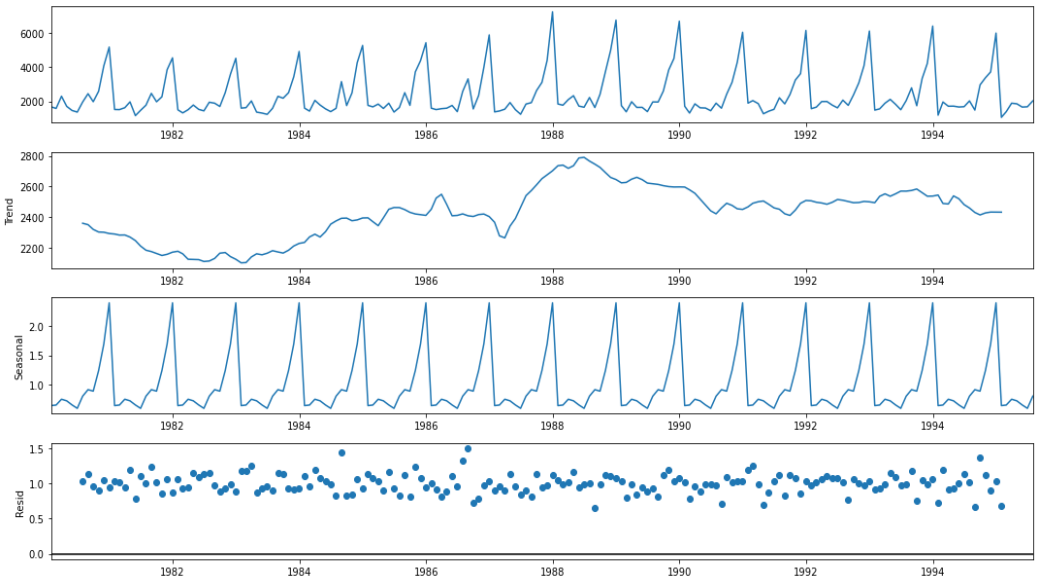


Figure 2.10 Multiplicative decomposition

Time_Stamp	Trend
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	2360.667
1980-08-31	2351.333
1980-09-30	2320.542
1980-10-31	2303.583

Time_Stamp	Seasonality
1980-01-31	0.650
1980-02-29	0.659
1980-03-31	0.757
1980-04-30	0.730
1980-05-31	0.661
1980-06-30	0.603
1980-07-31	0.809
1980-08-31	0.919
1980-09-30	0.894
1980-10-31	1.242

Time_Stamp	Residual
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	1.029
1980-08-31	1.135
1980-09-30	0.956
1980-10-31	0.908

Table 2.3 Trend, Seasonality, Residual (Multiplicative)

We can see the decomposition of the Sparkling time series above. I have tried with both additive and multiplicative decomposition so that I can determine if the dataset is a multiplicative or additive series.

- We see that the residuals are located around 0 from the plot of the residuals in the additive decomposition and showing some pattern.
- For the multiplicative series, we see that a lot of residuals are located around 1 and no pattern present.
- We can conclude that the time series of Sparkling is multiplicative.

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

Answer:

Next, we have divided the Rose datasets into train and test data. Test data starts in 1991, following is the shape of the test and train data.

```
train = df_spark[df_spark.index<'1991']
test = df_spark[df_spark.index>='1991']
```

```
print(train.shape)
print(test.shape)
```

```
(132, 1)
(55, 1)
```

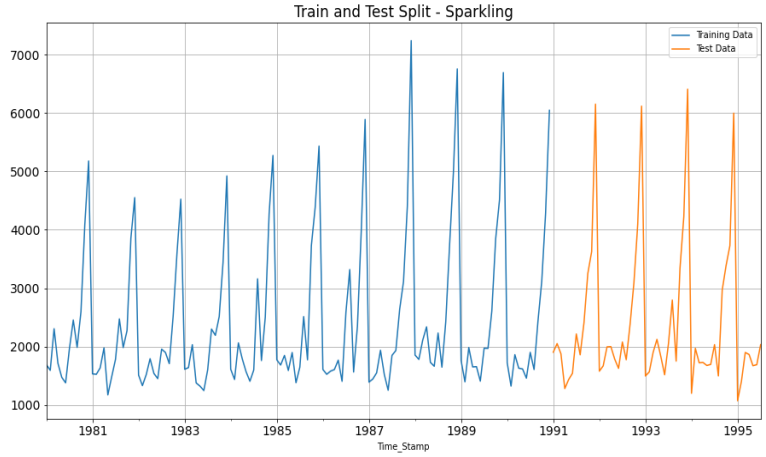


Figure 3.1 Plot of Train & Test Split

First few rows of Training Data

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

First few rows of Test Data

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

Last few rows of Training Data

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

Last few rows of Test Data

Time_Stamp	Sparkling
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

Table 3.1 Train & Test datasets

We can observe the training and test data in the above plot, the Blue part of the plots depicts the Train datasets (January ’80 – December ‘90), and the Orange part of the plots depict the testdatasets(January’91–July‘95).

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

Answer:

Model 1: Linear Regression

- Regress the "Sparkling" variable against the order of occurrence.
- Modifying the training set
- Generate the numerical instance order for both training and test set
- Printing the head and tail of test and train data

First few rows of Training Data

Time_Stamp	Sparkling	time
1980-01-31	1686	1
1980-02-29	1591	2
1980-03-31	2304	3
1980-04-30	1712	4
1980-05-31	1471	5

First few rows of Test Data

Time_Stamp	Sparkling	time
1991-01-31	1902	133
1991-02-28	2049	134
1991-03-31	1874	135
1991-04-30	1279	136
1991-05-31	1432	137

Last few rows of Training Data

Time_Stamp	Sparkling	time
1990-08-31	1605	128
1990-09-30	2424	129
1990-10-31	3116	130
1990-11-30	4286	131
1990-12-31	6047	132

Last few rows of Test Data

Time_Stamp	Sparkling	time
1995-03-31	1897	183
1995-04-30	1862	184
1995-05-31	1670	185
1995-06-30	1688	186
1995-07-31	2031	187

Table 4.1 Train & Test datasets for Linear Regression

Following are the results from a Linear Regression model

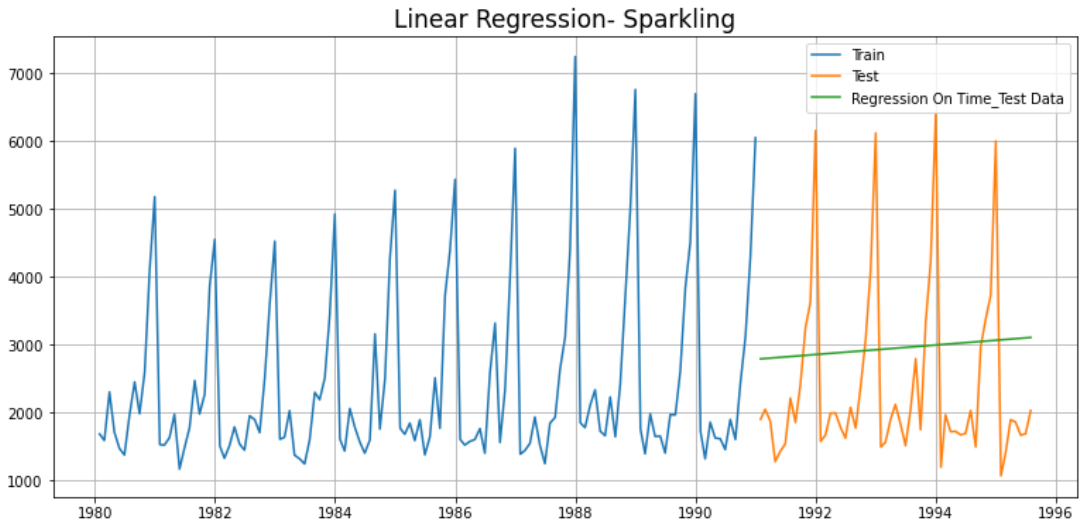


Figure 4.1 Plot for Linear Regression prediction

In the above plot we can see, blue line indicates train data, Orange as test and RegressionOnTime prediction on test set as the green line.

For RegressionOnTime forecast on the Test Data: RMSE is 1389.135

	Test RMSE
RegressionOnTime	1389.135

Model 2 - Naive Bayes

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

The extracts of Training and Test data for the Naïve Model can be seen below:

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

Time_Stamp	Sparkling	naive
1991-01-31	1902	6047
1991-02-28	2049	6047
1991-03-31	1874	6047
1991-04-30	1279	6047
1991-05-31	1432	6047

Table 4.2 Train & Test datasets for Naïve Bayes

Following are the results from a Naïve Bayes model

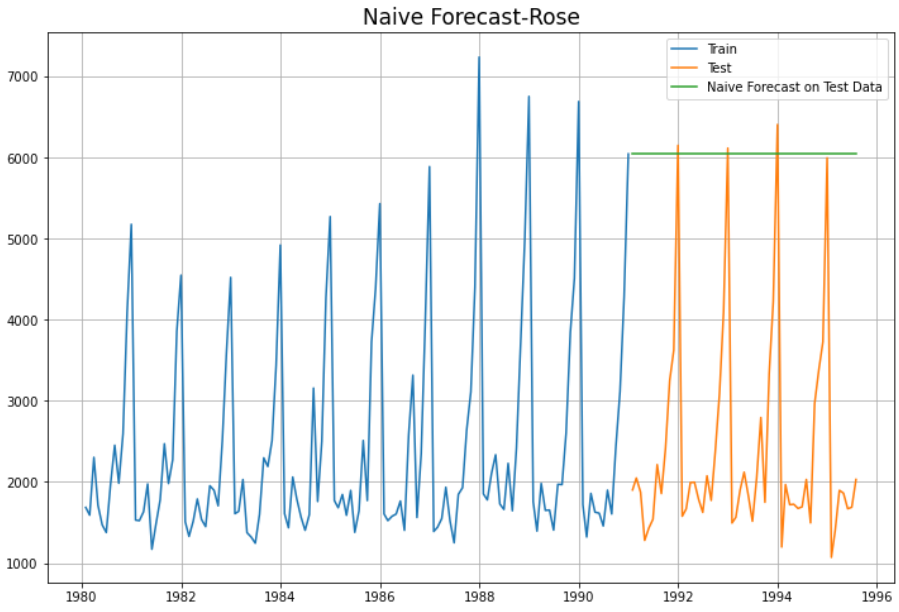


Figure 4.2 Plot for Naïve Bayes prediction

In the above plot we can see, blue line indicates train data, Orange as test and Naive bayes prediction on test set as the green line.

For Naive Model forecast on the Test Data: RMSE is 3864.279

	Test RMSE
NaiveModel	3864.279

Model 3- Simple Average

For this particular simple average method, we will forecast by using the average of the training values.

Time_Stamp	Sparkling	mean_forecast
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

Table 4.3 Forecast of test data for Simple Average

Following are the results from a Simple Average model

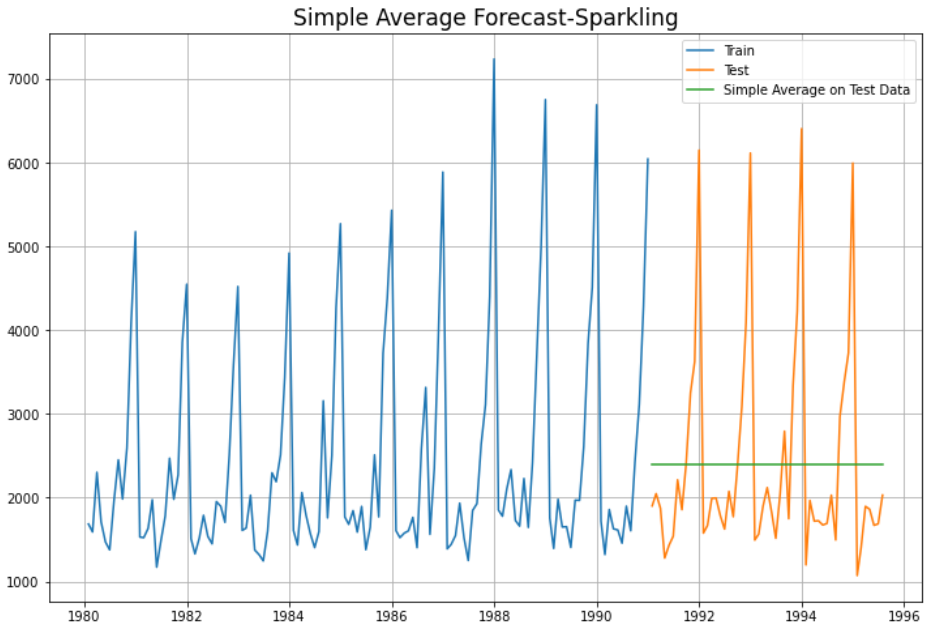


Figure 4.3 Plot for Simple Average prediction

In the above plot we can see, blue line indicates train data, Orange as test and Simple average prdiction on test set as the green line.

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

	Test RMSE
SimpleAverageModel	1275.082

Model 4- Moving Average (MA)

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here. For Moving Average, we are going to average over the entire data.

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
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
1980-06-30	1377	1424.0	1716.00	1690.166667	NaN
1980-07-31	1966	1671.5	1631.50	1736.833333	NaN
1980-08-31	2453	2209.5	1816.75	1880.500000	NaN
1980-09-30	1984	2218.5	1945.00	1827.166667	1838.222222
1980-10-31	2596	2290.0	2249.75	1974.500000	1939.333333

Table 4.4 Moving Average data

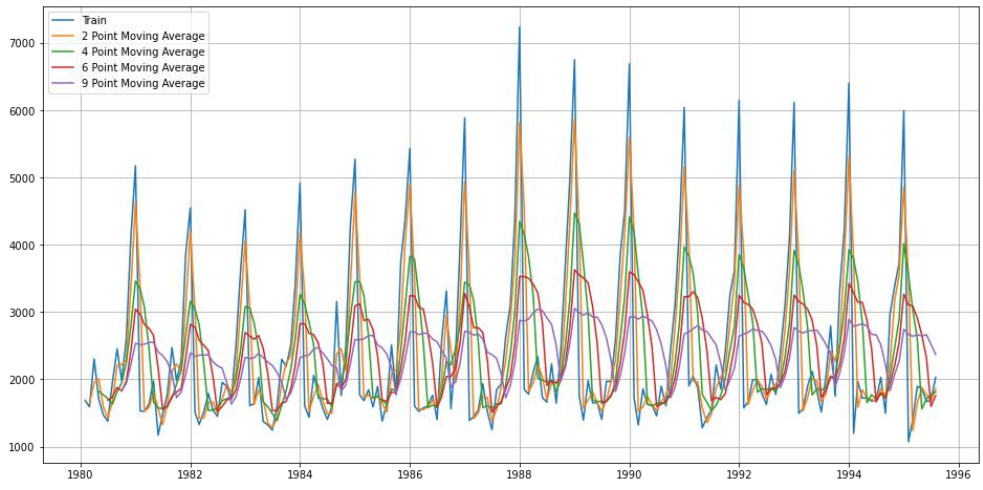


Figure 4.4 Plot for Moving Average data

Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.

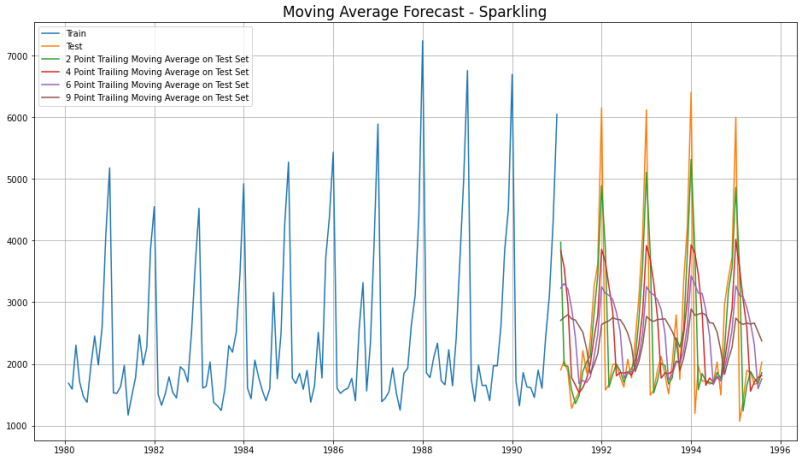


Figure 4.5 Plot for Moving Average prediction

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

	Test RMSE
2pointTrailingMovingAverage	813.401
4pointTrailingMovingAverage	1156.590
6pointTrailingMovingAverage	1283.927
9pointTrailingMovingAverage	1346.278

I have applied 2, 4, 6- and 9-point trailing averages on the Sparkling wine data sets. The closest prediction to actual data is shown by the 2- point trailing moving average model. This observation is confirmed by the RMSE scores for each of these moving average models. As can be seen from the summarized performance of all the models, the 2- point moving average has shown the best performance of all the 4 models.

Before we go on to build the various Exponential Smoothing models, let us plot all the models and compare the Time Series plots.

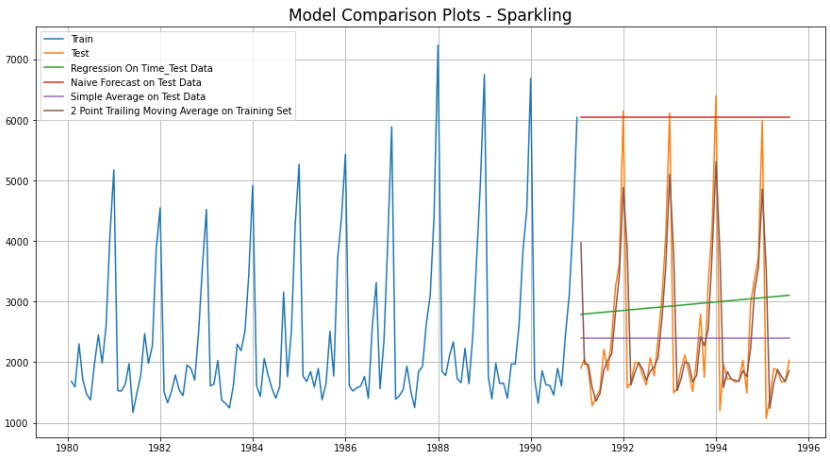


Figure 4.6 Plot for modes comparison till now

Model 5- Simple Exponential Smoothing (with α= 0.0496)

SES or one-parameter exponential smoothing is applicable to time series which do not contain either of trend or seasonality where, α is the smoothing parameter for the level.

The SES Parameters for Sparkling wine datasets after fitting the model are:

	name	param	optimized
smoothing_level	alpha	0.049607	True
initial_level	l.0	2151.614314	True

Table 4.5 Table of SES best params

From the above result we can find that the best param of alpha = 0.0496

Time_Stamp	Sparkling	predict
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

Table 4.6 Table of SES prediction data for α= 0.0496

Following are the results from a Simple Exponential model with α= 0.0496

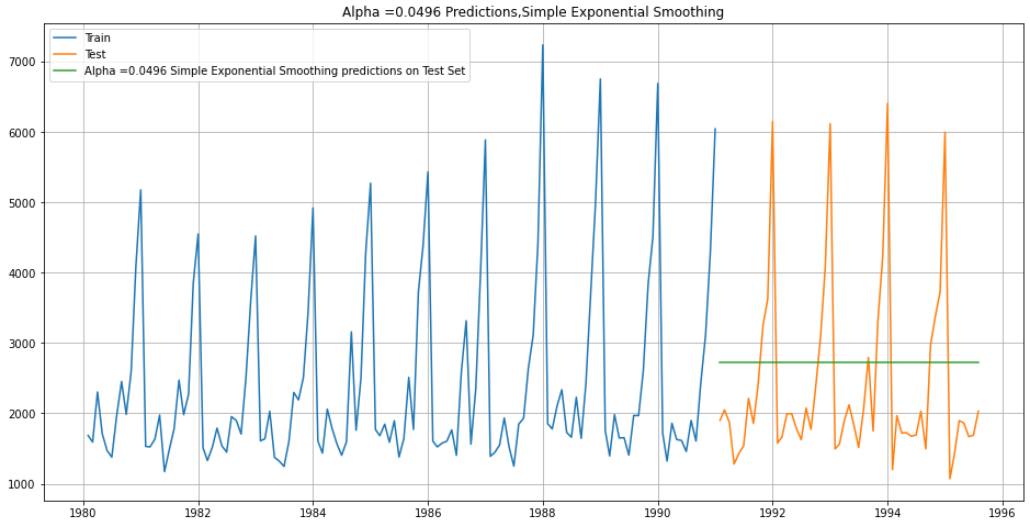


Figure 4.7 Plot of SES prediction for α= 0.0496

For Alpha =0.0496 Simple Exponential Smoothing Model forecast on the Test Data:
RMSE is 1316.135

	Test RMSE
Alpha=0.0496:SimpleExponentialSmoothing	1316.135

Model 6: Simple Exponential Smoothing (α = 0.1)

Setting different alpha values. We will run a loop with different alpha values to understand which particular value works best for alpha on the test set. After running the loop with alpha(smoothing_level) range of (0.1,1,0.1) below is the table consisting of alpha values with lowest Test RMSE.

	Alpha Values	Train RMSE	Test RMSE
0	0.1	1333.873836	1375.393398
1	0.2	1356.042987	1595.206839
2	0.3	1359.511747	1935.507132
3	0.4	1352.588879	2311.919615
4	0.5	1344.004369	2666.351413
5	0.6	1338.805381	2979.204388
6	0.7	1338.844308	3249.944092
7	0.8	1344.462091	3483.801006
8	0.9	1355.723518	3686.794285

Table 4.7 SES prediction data with α = 0.1

From the above table we can find that alpha =0.1 gives the lower RMSE value. Following are the results from a Simple Exponential model with α= 0.1

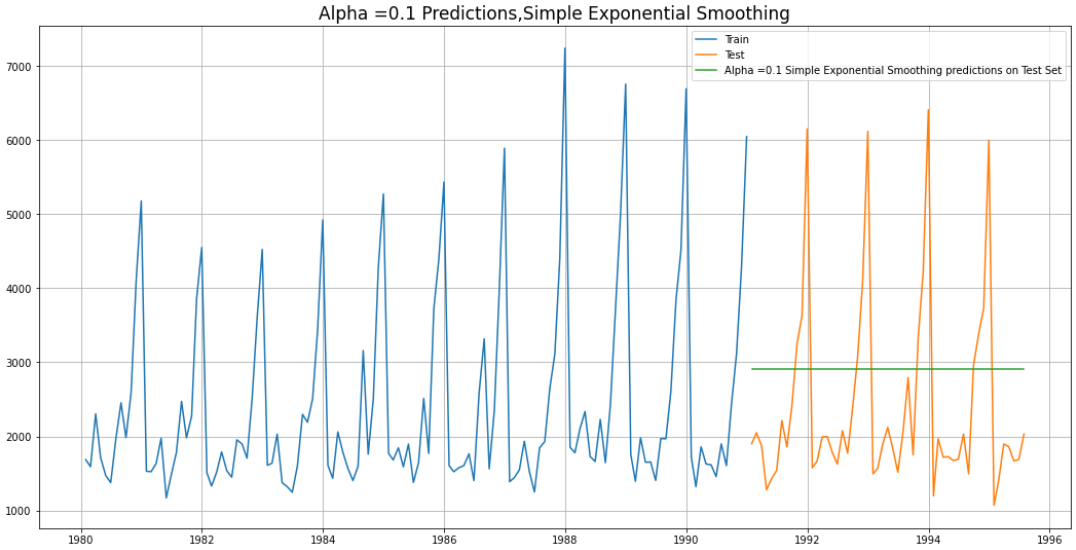


Figure 4.7 Plot of SES prediction for α= 0.1

For Alpha =0.1 Simple Exponential Smoothing Model forecast on the Test Data:
RMSE is 1375.393

	Test RMSE
Alpha=0.1,SimpleExponentialSmoothing	1375.393

I used Alpha =0. 1 for the SES model and as expected, it did not perform well as compared to previously run models.

Model 7: Double Exponential Smoothing (Holt's Model) (with α =0.688, β =0.0001)

This method is an extension of SES method. This method is applicable where trend is present in the data but no seasonality. α is the smoothing parameter for the level and β is the smoothing parameter for trend.

The DES Parameters for Sparkling wine datasets after fitting the model are:

	name	param	optimized
smoothing_level	alpha	0.688571	True
smoothing_trend	beta	0.000100	True
initial_level	l.0	1686.000000	True
initial_trend	b.0	-95.000000	True

Table 4.8 Table of DES best params

From the above result we can find that the best param of α = 0.688, β =0.0001.

Time_Stamp	Sparkling	predict_test
1991-01-31	1902	5221.278699
1991-02-28	2049	5127.886554
1991-03-31	1874	5034.494409
1991-04-30	1279	4941.102264
1991-05-31	1432	4847.710119

Table 4.9 DES prediction data with α = 0.688, β =0.0001

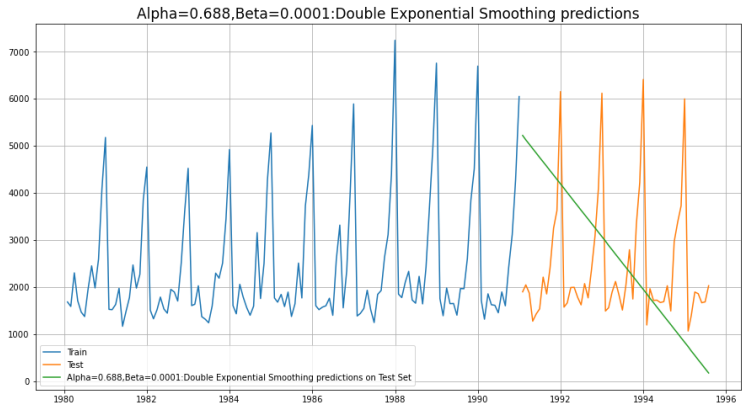


Figure 4.8 Plot of DES prediction for α = 0.688, β =0.0001

Following are the results from a Double Exponential model with α = 0.688, β =0.0001. Here, we see that the Double Exponential Smoothing has actually done well when compared to the Simple Exponential Smoothing. This is because of the fact that the Double Exponential Smoothing model has picked up the trend component as well.

For Alpha=0.6885, Beta=0.0001: Double Exponential Smoothing predictions forecast on the Test Data: RMSE is 2007.239

	Test RMSE
Alpha=0.688,Beta=0.0001 :DoubleExponentialSmoothing	2007.239

Model 8 : Double Exponential Smoothing (Holt's Model) (with α =0.1, β =0.1)

Setting different alpha & beta values. We will run a loop with different alpha & beta values to understand which particular value works best on the test set. After running the loop with range of alpha (smoothing_level) & & beta (smoothing_trend)0.1,1,0.1 below is the table consisting of alpha beta values with lowest Test RMSE.

	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

Table 4.10 DES prediction data with α = 0.1, β =0.1

From the above table we can find that α =0.1 & β =0.1 gives the lower RMSE value. Following are the results from a Double Exponential model with α =0.1 & β =0.1:

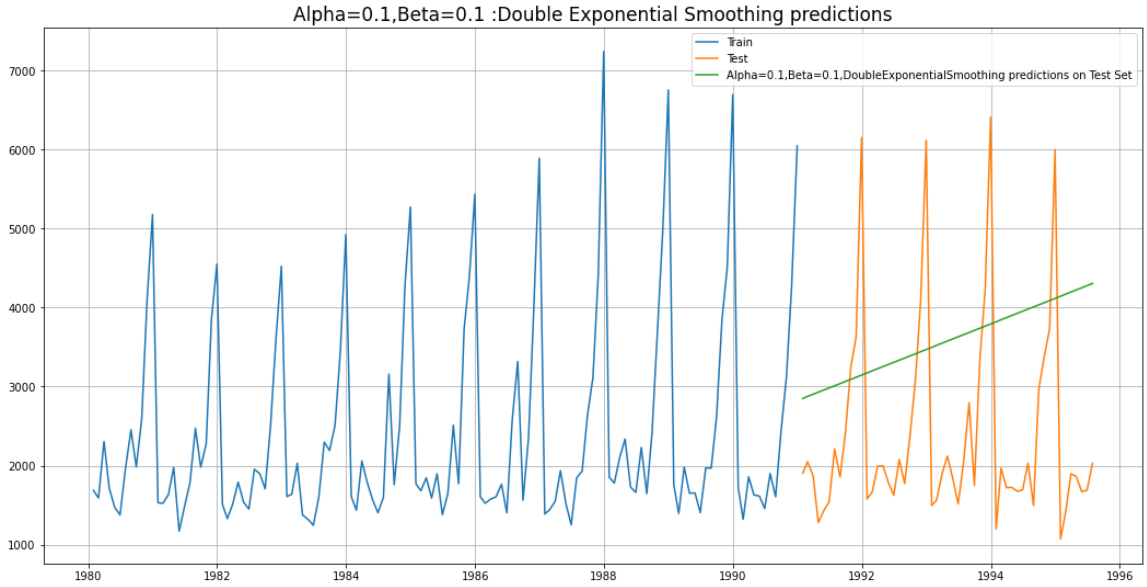


Figure 4.9 Plot of DES prediction for α = 0.1, β =0.1

From the plot we can find that only trend is present. For Alpha=0.1, Beta=0.1: Double Exponential Smoothing predictions forecast on the Test Data: RMSE is 1778.565. We can see the DES model is not performing well.

	Test RMSE
Alpha=0.1&Beta=0.1,DoubleExponentialSmoothing	1778.565

Model 9: Triple Exponential Smoothing (Holt - Winter's Model)(α =0.676 , β =0.088 & γ=0.323

This is an extension of Holt's method when seasonality is found in the data. This is also known as three parameters exponential or triple exponential because of the three smoothing parameters α, β and γ. This is a general method and a true multi-step ahead forecast. Here we are giving the parameters trend=additive & seasonality = multiplicative.

The TES Parameters for Sparkling wine datasets after fitting the model are:

	name	param	optimized
smoothing_level	alpha	0.111088	True
smoothing_trend	beta	0.061712	True
smoothing_seasonal	gamma	0.395081	True
initial_level	l.0	1639.908836	True
initial_trend	b.0	-11.928144	True
initial_seasons.0	s.0	1.050650	True
initial_seasons.1	s.1	1.020862	True
initial_seasons.2	s.2	1.410785	True
initial_seasons.3	s.3	1.202635	True
initial_seasons.4	s.4	0.973152	True
initial_seasons.5	s.5	0.966894	True
initial_seasons.6	s.6	1.317243	True
initial_seasons.7	s.7	1.704716	True
initial_seasons.8	s.8	1.372897	True
initial_seasons.9	s.9	1.810350	True
initial_seasons.10	s.10	2.839627	True
initial_seasons.11	s.11	3.609973	True

Table 4.11 TES best params

Time_Stamp	Sparkling	auto_predict
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

Table 4.12 TES prediction data for α =0.111, β =0.0617 & γ=0.395

Following are the results from a Triple Exponential model with α =0.676, β =0.088 & γ=0.323. We see that the Triple Exponential Smoothing is picking up the seasonal component as well.

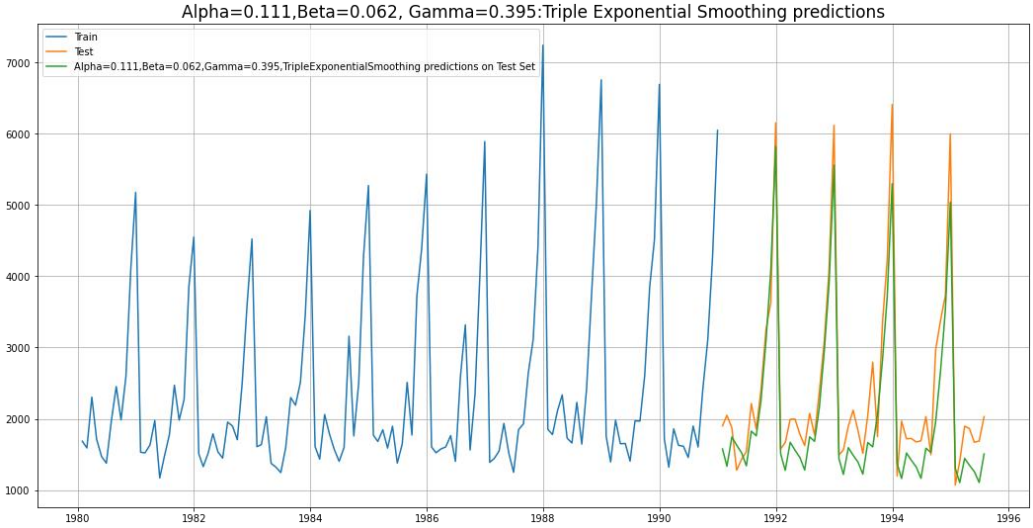


Figure 4.10 Plot of TES prediction data for α =0.111, β =0.0617 & γ=0.395

For Alpha=0.111,Beta=0.062,Gamma=0.395,TripleExponentialSmoothing predictions foreca st on the Test Data RMSE is 469.659

	Test RMSE
Alpha=0.111,Beta=0.062,Gamma=0.395,TripleExponentialSmoothing	469.659

Model 10: Triple Exponential Smoothing (Holt - Winter's Model) (α =0.4 , β =0.1 & γ=0.2)

Setting different alpha & beta values. We will run a loop with different alpha, beta & gamma values to understand which particular value works best on the test set. After running the loop with range of alpha(smoothing_level),beta(smoothing_trend) & gamma(smoothing_seasonal) as 0.1,1,0.1 below is the table consisting of alpha, beta & gamma values with lowest Test RMSE.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
244	0.4	0.1	0.2	389.772245	336.715250
172	0.3	0.2	0.2	395.529174	350.145204
90	0.2	0.2	0.1	405.333164	352.571689
162	0.3	0.1	0.1	394.630053	352.607849
18	0.1	0.3	0.1	414.423963	354.534561

Table 4.13 TES prediction data for α =0.4, β =0.1 & γ=0.2

From the above table we can find that α =0.4, β =0.1 & γ=0.2 gives the lower RMSE value. Following are the results from a Triple Exponential model with α =0.4, β =0.1 & γ=0.2:

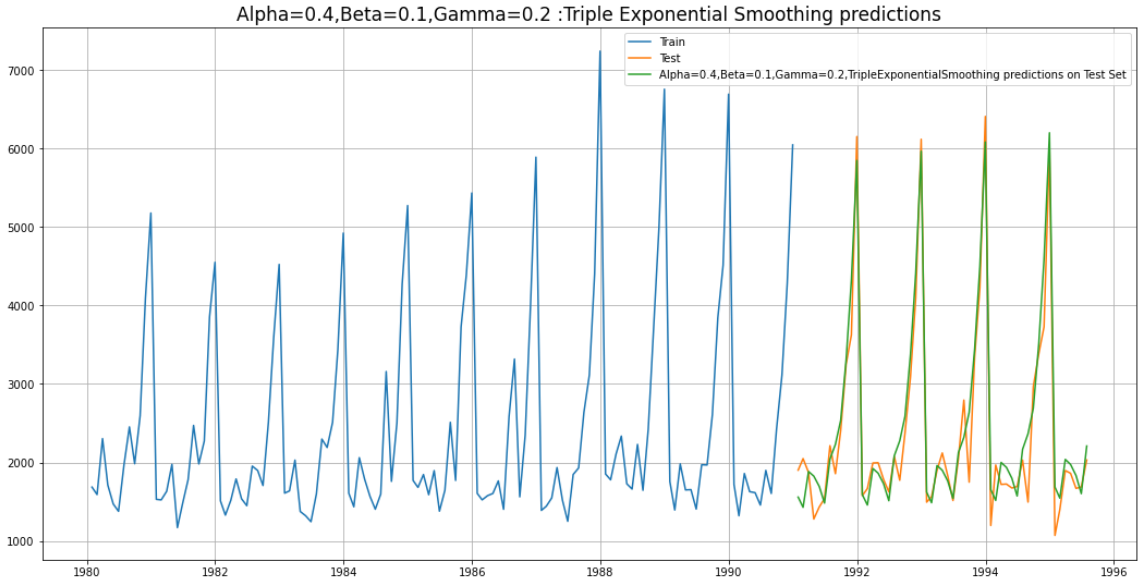


Figure 4.11 Plot of TES prediction data for α =0.4, β =0.1 & γ=0.2

For Alpha=0.4, Beta=0.1, Gamma=0.2, TripleExponentialSmoothing predictions forecast on the Test Data RMSE is 336.715.

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2:TripleExponentialSmoothing	336.715

Triple Exponential Smoothing with Alpha=0.4, Beta=0.1, Gamma=0.2 has performed the well on the test as expected since the data had both trend and seasonality with lowest RMSE

Plot of Exponential Smoothing Predictions and the Actual Values

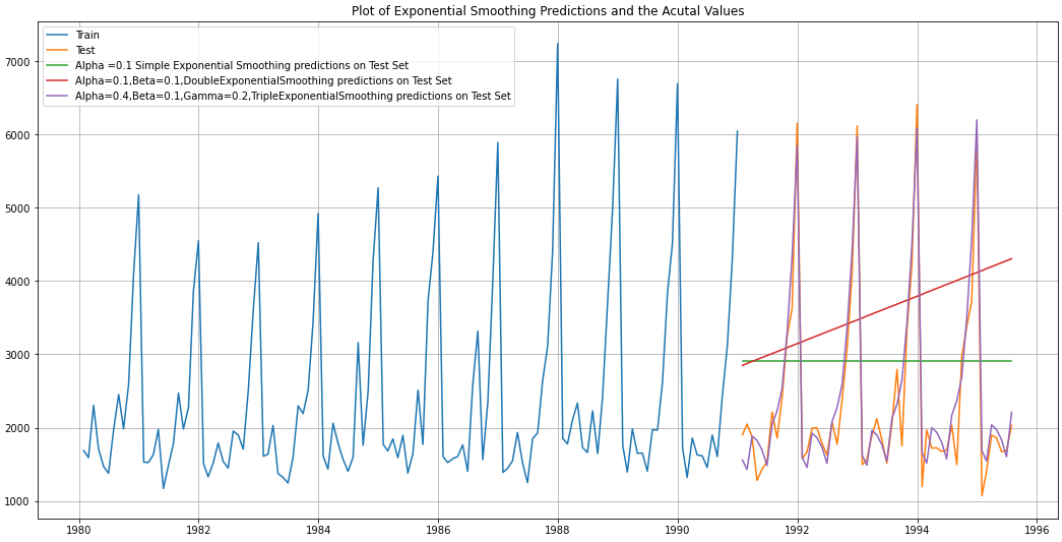


Figure 4.12 Plot of Exponential Smoothing predictions

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2:TripleExponentialSmoothing	336.715
Alpha=0.111,Beta=0.062,Gamma=0.395,TripleExponentialSmoothing	469.659
2pointTrailingMovingAverage	813.401
4pointTrailingMovingAverage	1156.590
SimpleAverageModel	1275.082
6pointTrailingMovingAverage	1283.927
Alpha=0.0496:SimpleExponentialSmoothing	1316.135
9pointTrailingMovingAverage	1346.278
Alpha=0.1,SimpleExponentialSmoothing	1375.393
RegressionOnTime	1389.135
Alpha=0.1&Beta=0.1,DoubleExponentialSmoothing	1778.565
Alpha=0.688,Beta=0.0001 :DoubleExponentialSmoothing	2007.239
NaiveModel	3864.279

Table 4.14 Dataframe with all models Test RMSE values

Triple Exponential Smoothing with Alpha=0.4, Beta=0.1, Gamma=0.2 2 has performed the best among all the models performed till now with lowest RMSE. We will further do another few more models like ARIMA/SARIMA in the below questions and finalize later which model performs well overall.

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.

Answer:

Check for stationarity of the whole Time Series data

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:

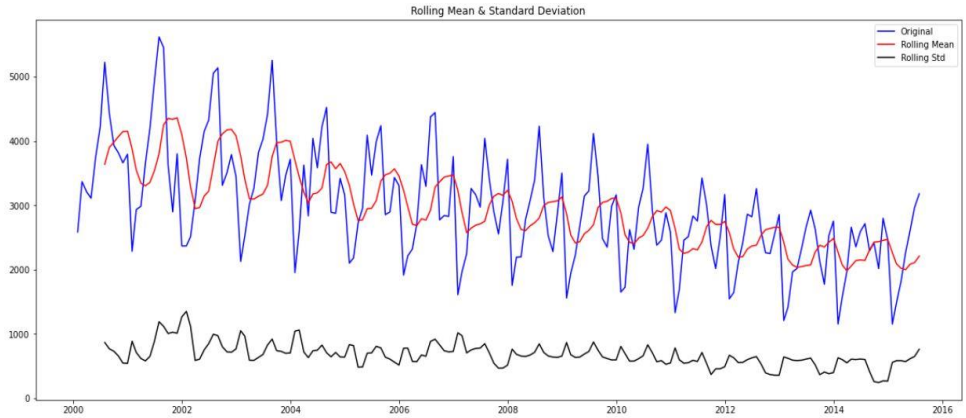
- $H0$: The Time Series has a unit root and is thus non-stationary.
- $H1$: The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA/SARIMA models and thus we would want the p-value of this test to be less than the α value.
$\alpha = 0.05$

So in ADF Test,

- if p-value < alpha ===> We reject the Null Hypothesis and hence conclude that given Time Series is Stationary.
- if p-value > alpha ===> We fail to reject the Null Hypothesis and hence conclude that given Time Series is Not Stationary
- If Time Series is not Stationary then we apply one level of differencing and check for Stationarity again.

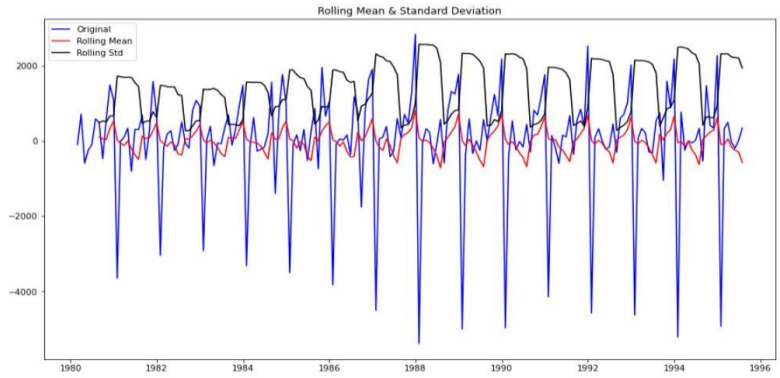
After running the adfuller test for the whole data below are is the result




```
Results of Dickey-Fuller Test:
Test Statistic                  -0.907380
p-value                          0.785532
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

Figure 5.1 adfuller test for whole data

Augmented Dicky-Fuller Test was applied to the whole Sparkling dataset. We found, p-value = 0.7855. Here, p-value > alpha=0.05. We fail to reject the Null Hypothesis and hence conclude that Sparkling Wine Time Series is Not Stationary. Let us take a difference of order 1 and check whether the Time Series is stationary or not. Below is the output for that




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

Figure 5.2 adfuller test for whole data (difference of order 1)

Now, p-value = 0. As, p-value < alpha=0.05 we reject the Null Hypothesis and conclude that Sparkling Time Series is Stationary with a lag of 1.

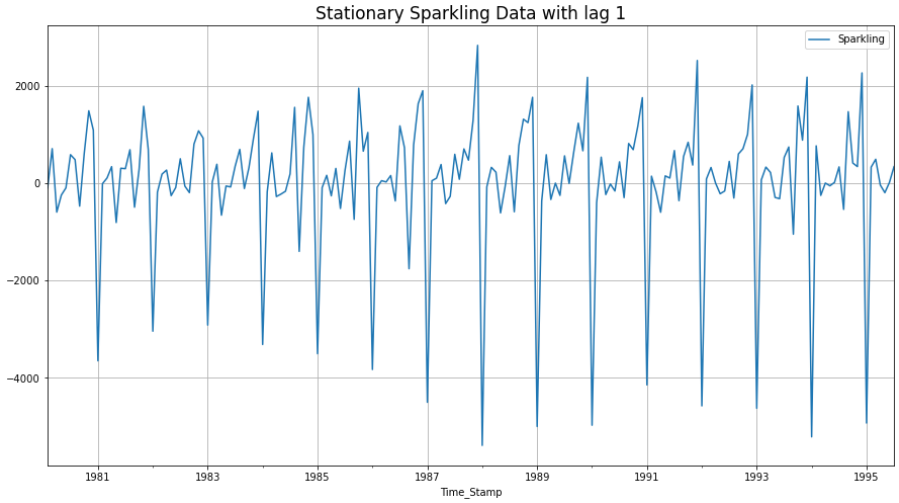


Figure 5.3 is Stationary od data with a lag of 1

Autocorrelation and the Partial Autocorrelation function plots on the whole data.

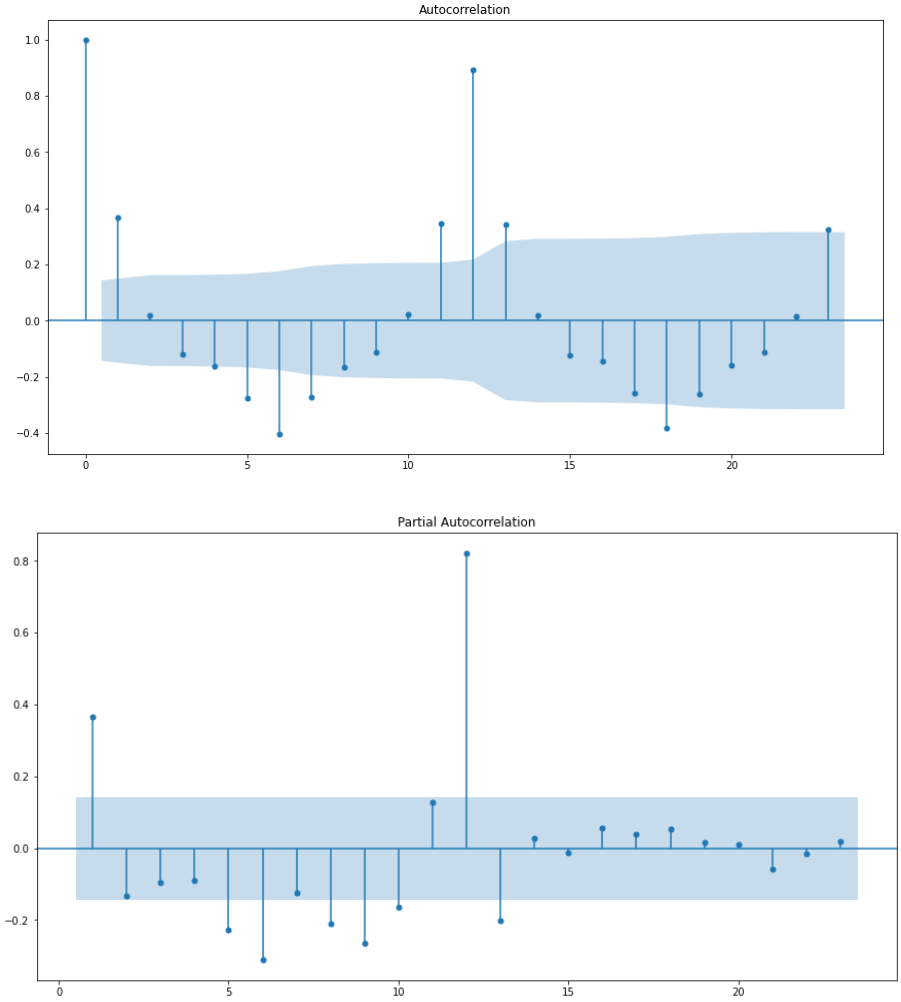
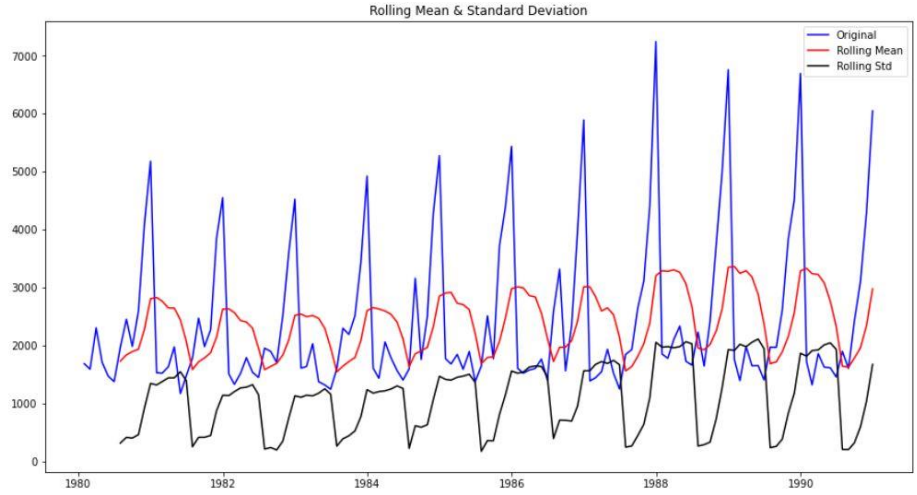


Figure 5.4 ACF & PACF plots for whole data with lag 1

Check for stationarity of the Training Data Time Series

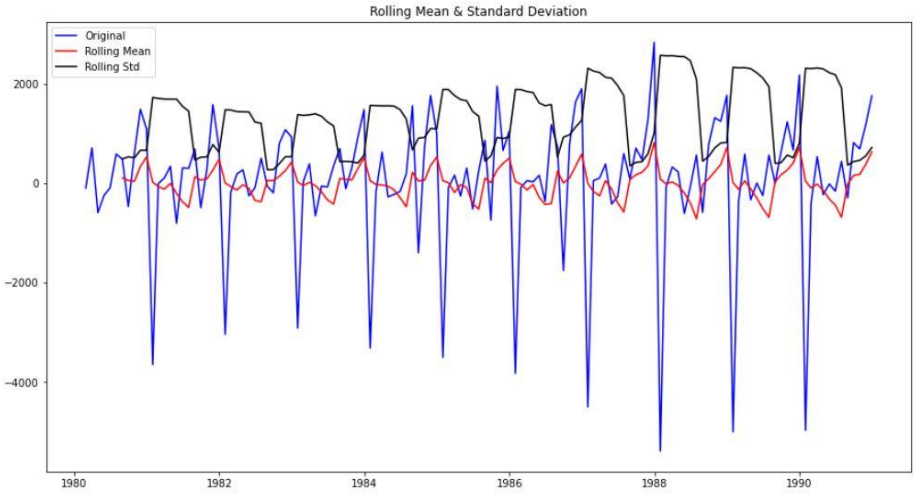
We check for stationarity of Train Sparkling data by using Augmented Dicky Fuller Test. We take a difference of 1 and make the dataset Stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.208926
p-value                          0.669744
#Lags Used                      12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
dtype: float64
```

Figure 5.5 adfuller test for train data

We found, p-value = 0.6697. Here, p-value > alpha=0.05. We fail to reject the Null Hypothesis and hence conclude that Sparkling Wine Train data is Not Stationary. Let us take a difference of order 1 and check whether the train data is stationary or not. Below is the output for that




```
Results of Dickey-Fuller Test:
Test Statistic                -8.005007e+00
p-value                        2.280104e-12
#Lags Used                     1.100000e+01
Number of Observations Used    1.190000e+02
Critical Value (1%)           -3.486535e+00
Critical Value (5%)           -2.886151e+00
Critical Value (10%)          -2.579896e+00
dtype: float64
```

Figure 5.6 adfuller test for train data (difference of order 1)

Now, p-value = 2.280104e-12. As, p-value < alpha=0.05 we reject the Null Hypothesis and c onclude that Sparkling train data is Stationary with a lag of 1.

Now we can use this particular differenced series to train the ARIMA/SARIMA models. We do not need to worry about stationarity for the Test Data because we are not building any models on the Test Data, we are evaluating our models over there.

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.

Answer:

We have already performed the stationary check in question 5 and made the train data stationary.

ARIMA Automated(p,q,d):

An ARIMA model consists of the Auto-Regressive (AR) part and the Moving Average (MA) part after we have made the Time Series stationary by taking the correct degree/order of differencing

- We create a grid of all possible combinations of (p, d, q)
- Range of p = Range of q = 0 to 3, Constant d = Range of 1 to 2
- Few Examples of the grid:

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
```

- We fit ARIMA models to each of these combinations of dataset
- We choose the combination with the least Akaike Information Criteria (AIC)
- We fit ARIMA to this combination of (p, d, q) to the Train set and forecast on the Test set
- Finally, we check the accuracy of this model by checking RMSE of Test set

	param	AIC
8	(2, 1, 2)	2213.509212
7	(2, 1, 1)	2233.777626
2	(0, 1, 2)	2234.408323
5	(1, 1, 2)	2234.527200
4	(1, 1, 1)	2235.755095

Table 6.1 Auto ARIMA with lowest AIC values

- The best combination with Least AIC is - (p, d, q) is (2,1,2)

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Wed, 16 Mar 2022	AIC	2213.509
Time:	23:27:09	BIC	2227.885
Sample:	01-31-1980 - 12-31-1990	HQIC	2219.351
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.782	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.741	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.217	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.109	0.000	0.785	1.215
sigma2	1.099e+06	1.99e-07	5.51e+12	0.000	1.1e+06	1.1e+06

Ljung-Box (L1) (Q):	0.19	Jarque-Bera (JB):	14.46
Prob(Q):	0.67	Prob(JB):	0.00
Heteroskedasticity (H):	2.43	Skew:	0.61
Prob(H) (two-sided):	0.00	Kurtosis:	4.08

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.02e+28. Standard errors may be unstable.

Figure 6.1 Auto ARIMA (2,1,2) Summary

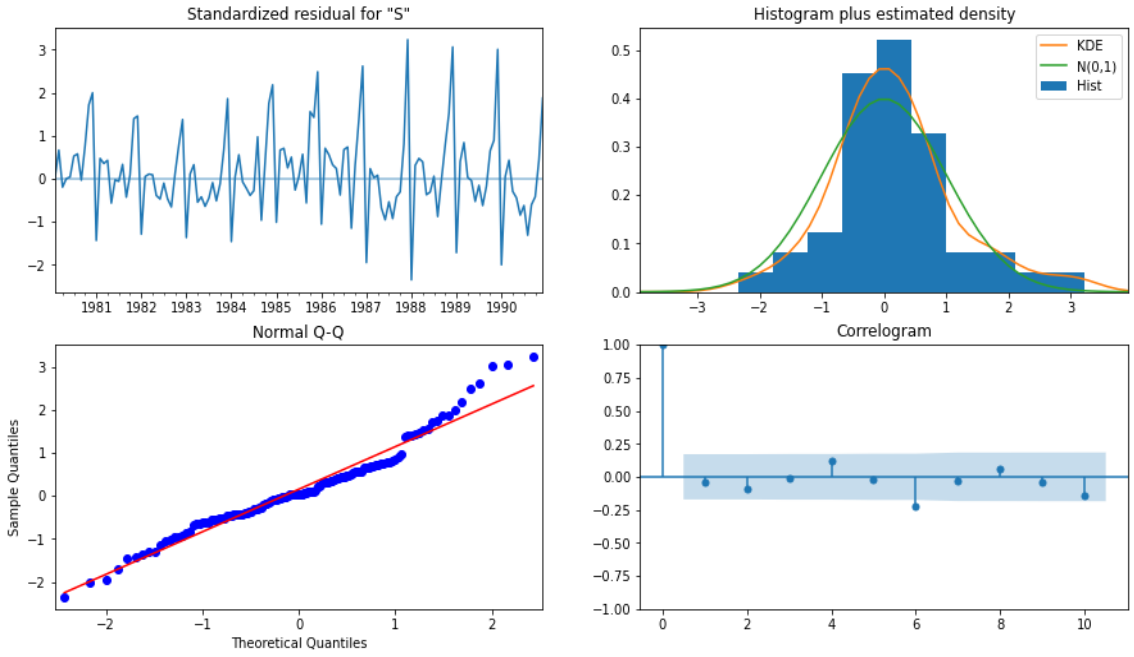


Figure 6.2 Residual Diagnostics of Auto ARIMA (2,1,2)

RMSE for ARIMA (2,1,2) based on AIC value is: 1299.98

	Test RMSE
ARIMA(2,1,2) based on AIC value	1299.98

SARIMA Automated(p,q,d)(P,D,Q)F:

For a Seasonal Auto-Regressive Integrated Moving Average we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d) and seasonal differencing (D). Here, the 'F' parameter indicates the seasonality/seasonal effects over a particular period

Let us look at the ACF plot once more to understand the seasonal parameter for the SARIMA model

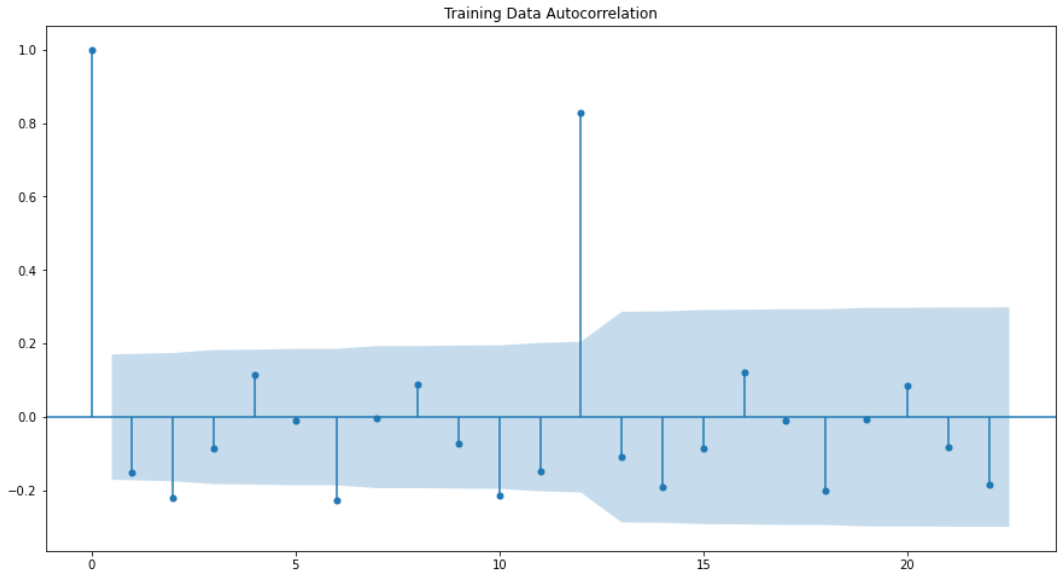


Figure 6.3 ACF plot to check seasonality value

- We can find from the above ACF plot that the seasonality is 12.
- We create a grid of all possible combinations of (p, d, q) along with Seasonal (P, D, Q) & Seasonality of 12
- Range of p = Range of q = 0 to 3, Constant d = Range of 1 to 2
- Range of Seasonal P = Range of Seasonal Q = 0 to 3, Constant D = Range of 0 to 1, Seasonality m = 12
- Few Examples of the grid (p, d, q) (P, D, Q, m) - Model:

```
Examples of some parameter combinations for Model.
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

- We fit SARIMA models to each of these combinations and select with least AIC
- We fit SARIMA to this best combination of (p, d, q) (P, D, Q, m) to the Train set and forecast on the Test set. Then, we check accuracy using RMSE on Test set.

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.934564
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121565
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

Table 6.2 Auto SARIMA with lowest AIC values

- The best combination with Least AIC is - (p, d, q) (P, D, Q, m) is (1,1,2) (1,0,2,12)

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Wed, 16 Mar 2022	AIC	1555.584
Time:	23:33:01	BIC	1574.095
Sample:	0	HQIC	1563.083
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6282	0.255	-2.463	0.014	-1.128	-0.128
ma.L1	-0.1041	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7276	0.154	-4.735	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.840	0.000	1.016	1.072
ma.S.L12	-0.5550	0.098	-5.663	0.000	-0.747	-0.363
ma.S.L24	-0.1355	0.120	-1.133	0.257	-0.370	0.099
sigma2	1.506e+05	2.03e+04	7.401	0.000	1.11e+05	1.9e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	11.72
Prob(Q):	0.84	Prob(JB):	0.00
Heteroskedasticity (H):	1.47	Skew:	0.36
Prob(H) (two-sided):	0.26	Kurtosis:	4.48

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

Figure 6.4 Auto SARIMA (1,1,2)(1,0,2,12) Summary

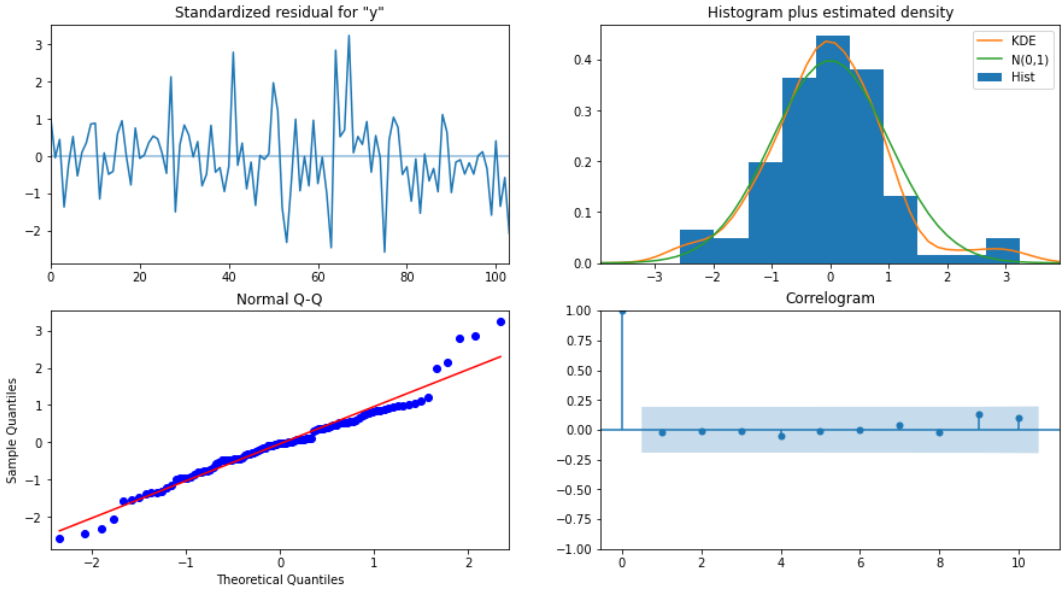


Figure 6.5 Residual Diagnostics of Auto SARIMA (1,1,2)(1,0,2,12)

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1327.385682	388.343707	566.246002	2088.525361
1	1315.127259	402.007354	527.207323	2103.047194
2	1621.591592	402.000958	833.684193	2409.498992
3	1598.879086	407.238889	800.705530	2397.052642
4	1392.693371	407.968895	593.089030	2192.297712

Table 6.3 Auto SARIMA model

RMSE for SARIMA (1,1,2) (1,0,2,12) based on AIC value 528.60

	Test RMSE
SARIMA(1,1,2)(1,0,2,12)based on AIC value	528.607

From the above 2 model of ARIMA and SARIMA we can see that SARIMA model is performing well with lowest RMSE as this model includes the seasonality as well.

	Test RMSE
ARIMA(2,1,2) based on AIC value	1299.98
SARIMA(1,1,2)(1,0,2,12)based on AIC value	528.607

Table 6.4 Auto ARIMA/SARIMA test RMSE values

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.

Answer:

Autocorrelation Function (ACF): Autocorrelation of order $p$is the correlation between$Yt$and $Yt+k$for all values of k=0, 1, $-1 \leq ACF \leq 1$ and ACF (0) =1.ACF measures strength of dependency of current observations on past observations.

Partial Autocorrelation Function (PACF): PACF of order $k$is the autocorrelation between$Yt$and $Yt+k$adjusting for all the intervening periods i.e., it provides the correlation value between current and k-lagged series by removing the influence of all other observations that exist in between.

- ACF and PACF used together to identify the order of the ARIMA
- Seasonal ACF and PACF examines correlations for seasonal data

ARIMA Manual (0,1,0)

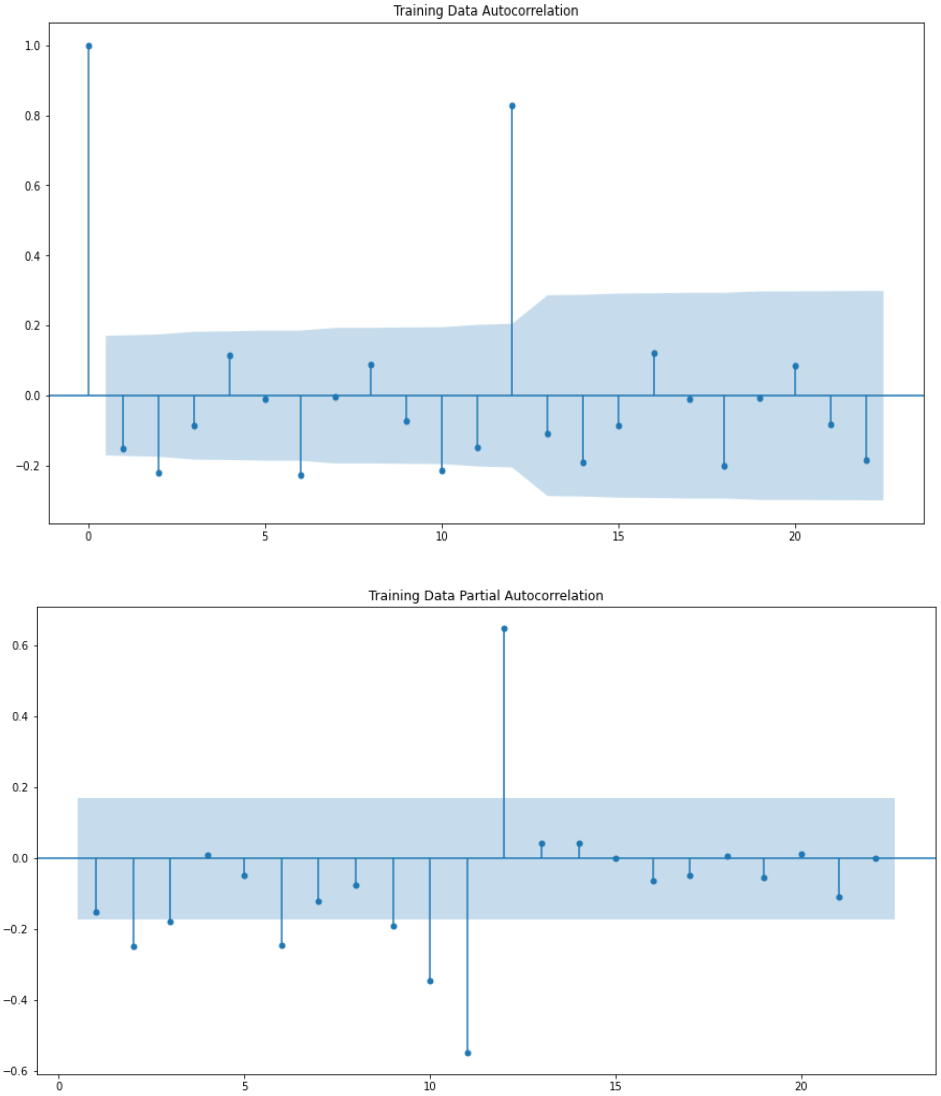


Figure 7.1 ACF & PACF plot for first difference

Here, we have taken alpha=0.05.

The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0.

The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.

The'd' value is 1 as we have taken the train data 1 lag differentiation for stationarity(train.diff(), question5)

Observing the cut-offs in ACF and PACF plots for dataset, we get:

ARIMA —-> p = 0, q = 0 and difference d = 1 i.e (0,1,0)

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1132.832
Date:	Wed, 16 Mar 2022	AIC	2267.663
Time:	23:34:09	BIC	2270.538
Sample:	01-31-1980	HQIC	2268.831
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1.885e+06	1.29e+05	14.658	0.000	1.63e+06	2.14e+06

Ljung-Box (L1) (Q):	3.07	Jarque-Bera (JB):	198.83
Prob(Q):	0.08	Prob(JB):	0.00
Heteroskedasticity (H):	2.46	Skew:	-1.92
Prob(H) (two-sided):	0.00	Kurtosis:	7.65

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

Figure 7.2 Manual ARIMA (0,1,0) Summary

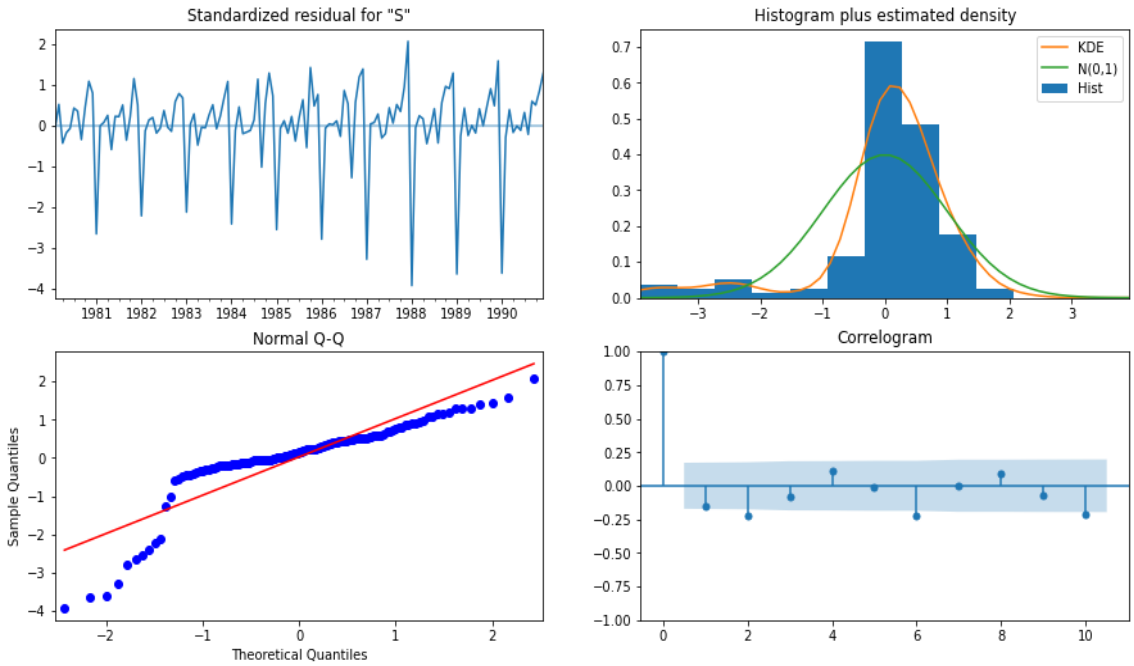


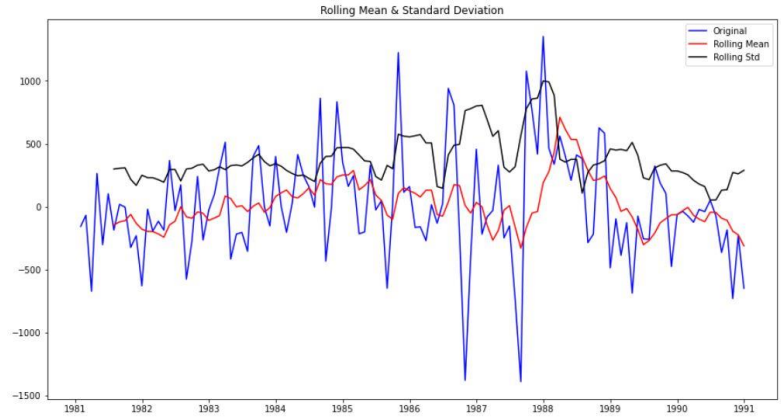
Figure 7.3 Residual Diagnostics of manual ARIMA (0,1,0)

RMSE for manual ARIMA(0,1,0) based on ACF & PACF values is: 3864.279

	Test RMSE
ARIMA(0,1,0)based on ACF & PACF values	3864.279

SARIMA Manual with seasonality 12 (0,1,0) (1,0,1,12)

Checking the stationarity of train with seasonality 12, below is the result




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.136812
p-value                          0.023946
#Lags Used                      11.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
```

Figure 7.4 adfuller test for train data with seasonality =12

From the above check we can find that p value = 0.0239. Since the p value < alpha=0.05, we reject the null hypothesis and confirm that the data is stationary. From this we can say that the D is 0. Checking the P,Q values using ACF,PACF plots.

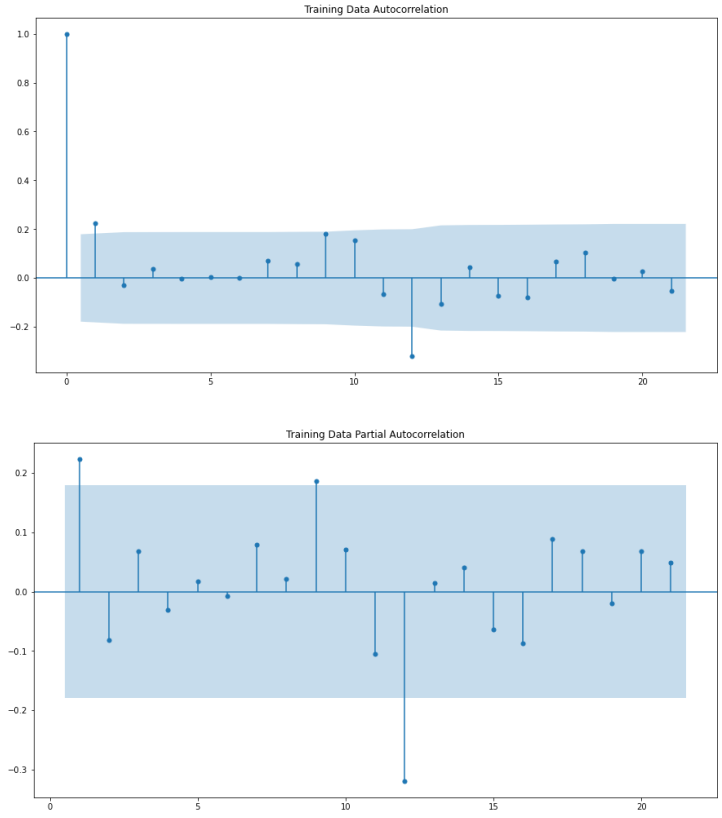


Figure 7.5 ACF & PACF plot of train with seasonality =12

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12. We are taking the p value to be 0 and the q value also to be 0 as the parameters same as the manual 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 1.(from the above ACF plot)
- The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 1. (from the above PACF plot)

SARIMA —-> p = 0, q = 0, d = 1 and P = 1, D = 0, Q = 1, Seasonality=12 (0,1,0) (1,0,1,12)

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 0)x(1, 0, [1], 12)	Log Likelihood	-900.495
Date:	Wed, 16 Mar 2022	AIC	1806.991
Time:	23:35:03	BIC	1815.303
Sample:	0	HQIC	1810.365
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L12	1.0325	0.019	52.957	0.000	0.994	1.071
ma.S.L12	-0.5384	0.078	-6.896	0.000	-0.691	-0.385
sigma2	2.463e+05	2.34e+04	10.520	0.000	2e+05	2.92e+05

Ljung-Box (L1) (Q):	19.69	Jarque-Bera (JB):	31.97
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.88	Skew:	0.66
Prob(H) (two-sided):	0.05	Kurtosis:	5.18

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

Figure 7.6 Manual SARIMA (0,1,0) (1,0,1,12) Summary

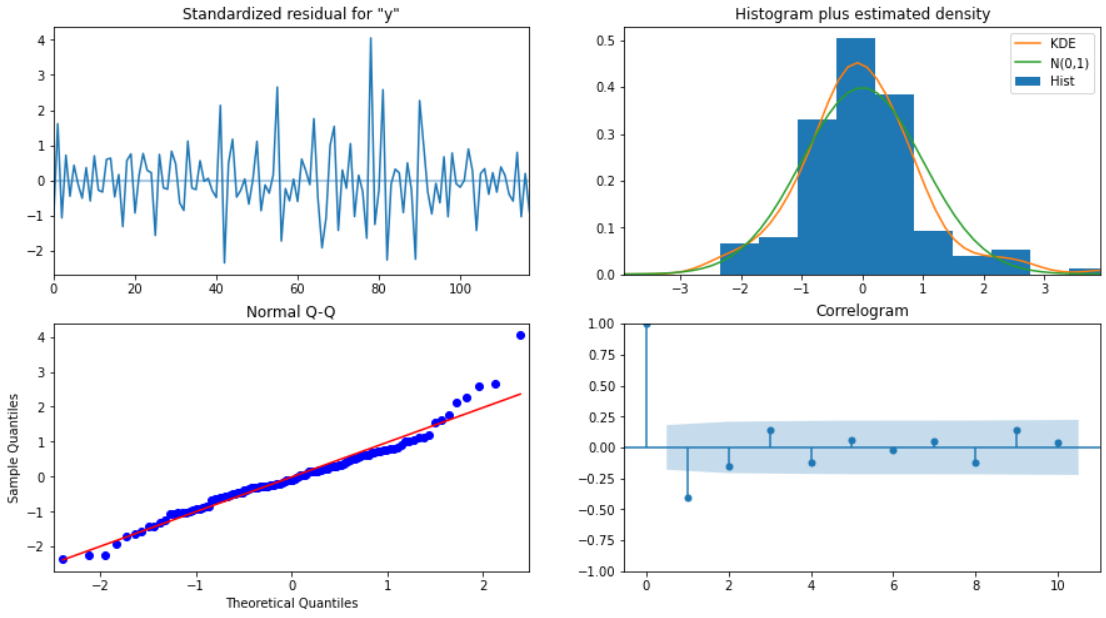


Figure 7.7 Residual Diagnostics of manual SARIMA (0,1,0) (1,0,1,12)

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	820.888032	496.283504	-151.809763	1793.585826
1	512.512844	701.850523	-863.088904	1888.114591
2	1008.023356	859.587690	-676.737558	2692.784269
3	856.770553	992.566288	-1088.623624	2802.164730
4	739.687026	1109.722792	-1435.329680	2914.703732

Table 6.5 Manual SARIMA model

RMSE for manual SARIMA(0,1,0)(1,0,1,12) based on ACF & PACF values is: 1787.707

	Test RMSE
SARIMA(0,1,0)(1,0,1,12)based on ACF & PACF values	1787.707

From the above 2 model of manual ARIMA and SARIMA we can see that manual SARIMA model is performing well with lowest RMSE as this model includes the seasonality as well.

	Test RMSE
ARIMA(0,1,0)based on ACF & PACF values	3864.279
SARIMA(0,1,0)(1,0,1,12)based on ACF & PACF values	1787.707

Table 6.6 Manual ARIMA/SARIMA test RMSE values

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.

Answer:

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2:TripleExponentialSmoothing	336.715
Alpha=0.111,Beta=0.062,Gamma=0.395,TripleExponentialSmoothing	469.659
SARIMA(1,1,2)(1,0,2,12)based on AIC value	528.607
2pointTrailingMovingAverage	813.401
4pointTrailingMovingAverage	1156.590
SimpleAverageModel	1275.082
6pointTrailingMovingAverage	1283.927
ARIMA(2,1,2) based on AIC value	1299.980
Alpha=0.0496:SimpleExponentialSmoothing	1316.135
9pointTrailingMovingAverage	1346.278
Alpha=0.1,SimpleExponentialSmoothing	1375.393
RegressionOnTime	1389.135
Alpha=0.1&Beta=0.1,DoubleExponentialSmoothing	1778.565
SARIMA(0,1,0)(1,0,1,12)based on ACF & PACF values	1787.707
Alpha=0.688,Beta=0.0001 :DoubleExponentialSmoothing	2007.239
ARIMA(0,1,0)based on ACF & PACF values	3864.279
NaiveModel	3864.279

Table 8.1 dataframe with all models in ascending order test RMSE values

As we can observe from above table that the among the models build we can find that **Alpha=0.4,Beta=0.1,Gamma=0.2:TripleExponentialSmoothing = 336.715** is performing well with lowest RMSE.

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.

Answer:

We see that the most optimum model is the Triple Exponential Smoothing with additive trend & multiplicative seasonality and the parameters α = 0.4, β = 0.1 and γ = 0.2 Let us now build the model using the same parameters on the full data and check the confidence bands when we forecast into the future for the length of the test set.

Parameters given:

trend='additive', seasonal='multiplicative'

(smoothing_level=0.4, smoothing_trend=0.1, smoothing_seasonal=0.2

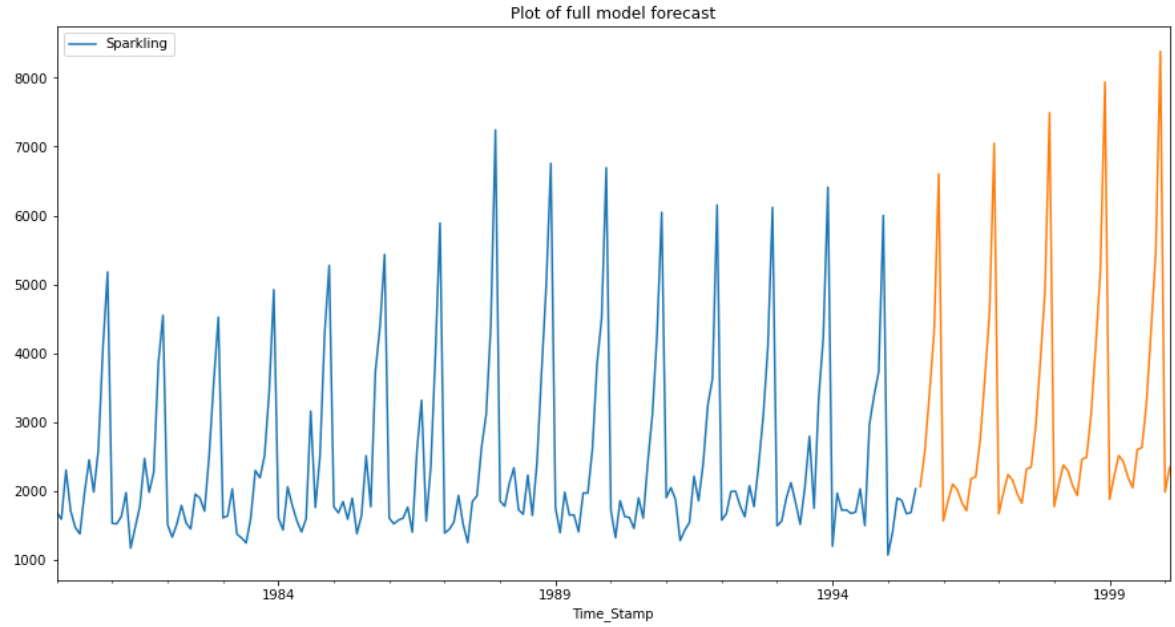


Figure 9.1 Plot of full model forecast

RMSE for full model with Triple Exponential Smoothing α = 0.4, β = 0.1 and γ = 0.2 is: 376.775

The model predicting 12 months into the future with appropriate confidence intervals to see h ow the predictions look, here we have calculated the upper and lower confidence bands at 95 % confidence level , below is the result:

	lower_CI	prediction	upper_ci
1995-08-31	1322.989337	2063.449003	2803.908669
1995-09-30	1838.947945	2579.407611	3319.867277
1995-10-31	2676.194903	3416.654569	4157.114235
1995-11-30	3564.017841	4304.477507	5044.937173
1995-12-31	5864.417326	6604.876992	7345.336658
1996-01-31	824.080268	1564.539934	2304.999600
1996-02-29	1109.300430	1849.760096	2590.219762
1996-03-31	1358.419272	2098.878937	2839.338603
1996-04-30	1281.969204	2022.428870	2762.888536
1996-05-31	1094.080979	1834.540644	2575.000310
1996-06-30	971.949271	1712.408937	2452.868603
1996-07-31	1435.965864	2176.425530	2916.885196

Table 9.1 12-month forecast along with confidence band

The forecasted value is from August 1995 – July 1996

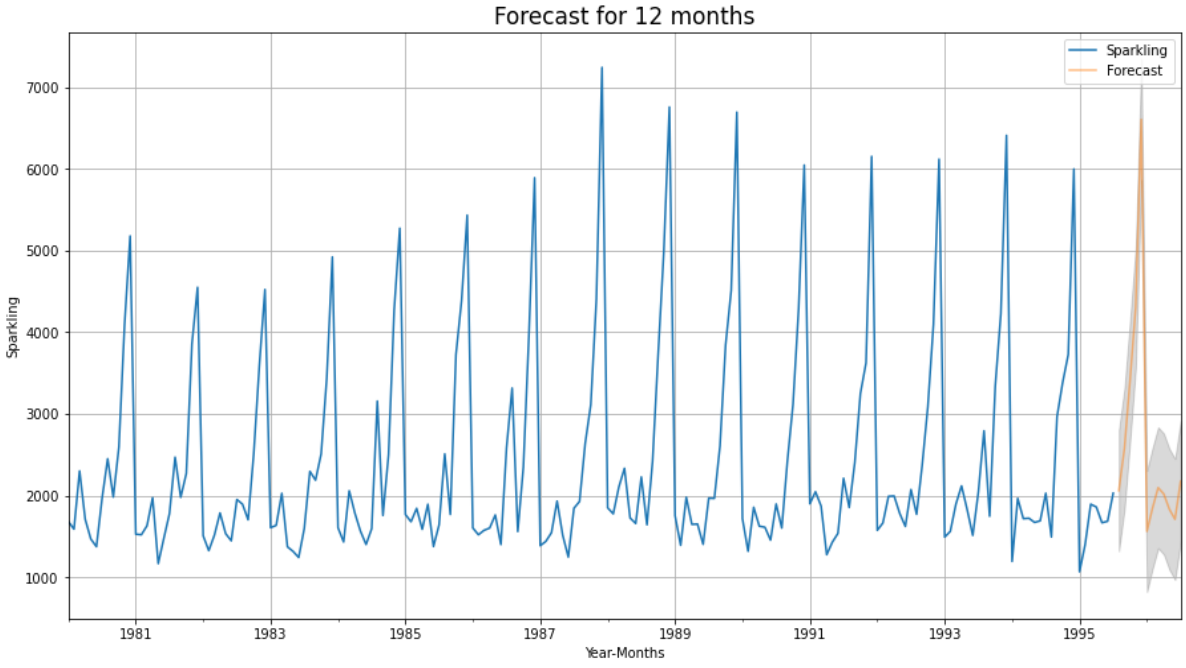


Figure 9.2 Plot of 12-month forecast along with confidence band

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

Answer:

- Sparkling wine sales don't show any upward or downward trend. This shows flat sales over long term range
- There is very high spike in sales seen in the last quarter of every year from October to December, this may be due to the holiday season. Highest peak is sales seen in every December.
- The lowest sale of sparkling wine is seen in first quarter of every year from January. This could be due to some New Year resolutions and the end of vacation.
- Sales slowly pick up only from month of July.
- Different models were run to find the optimum model using the lowest RMSE value. The best optimum model selected was Holt-Winters Triple Exponential Smoothing with α = 0.4, β = 0.1 and γ = 0.2 with trend as additive and seasonality as multiplicative
- We predicted for the next 12 months into the future with appropriate confidence intervals to see how the predictions look and plotted the forecast.

Suggestions:

- Company should stock up more during the holiday season as we can see the forecast showing increasing sales and peak during December
- Would recommend running marketing campaigns, discounts and offers to increase consumption.
- Offer tasting packages- this will make the customers to purchase after sampling a product.
- Offer home delivery services for events that can drive big time sales. Consumers love the convenience of at-home delivery, especially if they know you are delivering delicious product that will make their event.
- Can introduce new variants of Sparkling with cheaper price and special designed bottles to increase the sales.

----------END OF REPORT----------