Project:TIME SERIES FORECASTING

DSBA

Submitted by:

Name: Hima Jose Paliakkara

Batch: 2021-2022

Table of Contents

Contents

Problem – Sales data of Rose Wine

List of figures

List of Tables

Problem – Sales data of Rose 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: **Rose.csv**

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

Answer:

Summary of dataset

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

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 2 null values.
- No duplicates present

Plot the Time Series to understand the behaviour of the data

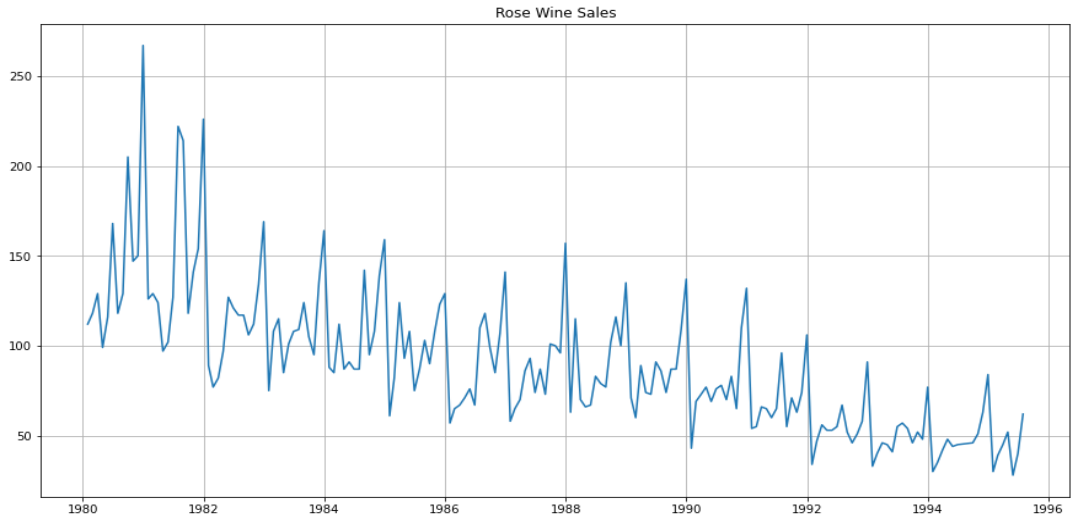


Figure 1.1 Time Series Plot

The sales for Rose Wine are showing a declining 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:

	Rose
count	187.000000
mean	89.914439
std	39.238325
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

```
<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   Rose    187 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Table 2.1 Summary& info of data

- The dataset contains 187 rows and 1 column
- The dataset has 2 null values, I am imputing the null values using linear interpolation method to fill in the missing values.
- No duplicates present

Yearly Box-plot

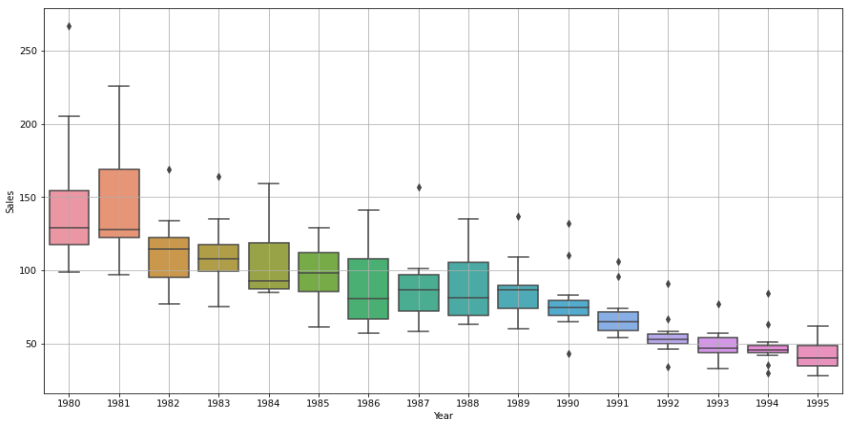


Figure 2.1 Yearly-box Plot

From the above plot, Rose wine has mostly a downward sales trend. The highest sales for Rose wine can be observed in 1981 and the lowest sales in 1994.The highest variation in monthly sales for Rose wine is in the year 1981 and on the year 1994 there seems to be the lowest variation in monthly sales. There are outliers in the yearly sales data, however as it is a Time Series, we can ignore the outlier data.

Monthly Box-plot

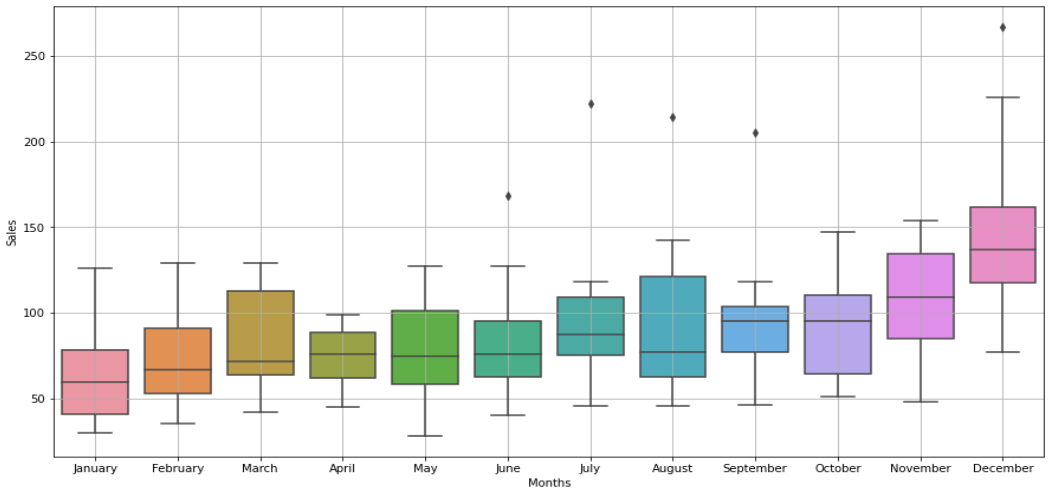


Figure 2.2 Monthly-box Plot

Above plot gives the monthly sales of rose wine. We can find that highest wine sales were in the month of December, followed by November, and the lowest in January. Hence, we can say there is seasonality present in our Rose data set. Few outliers are present.

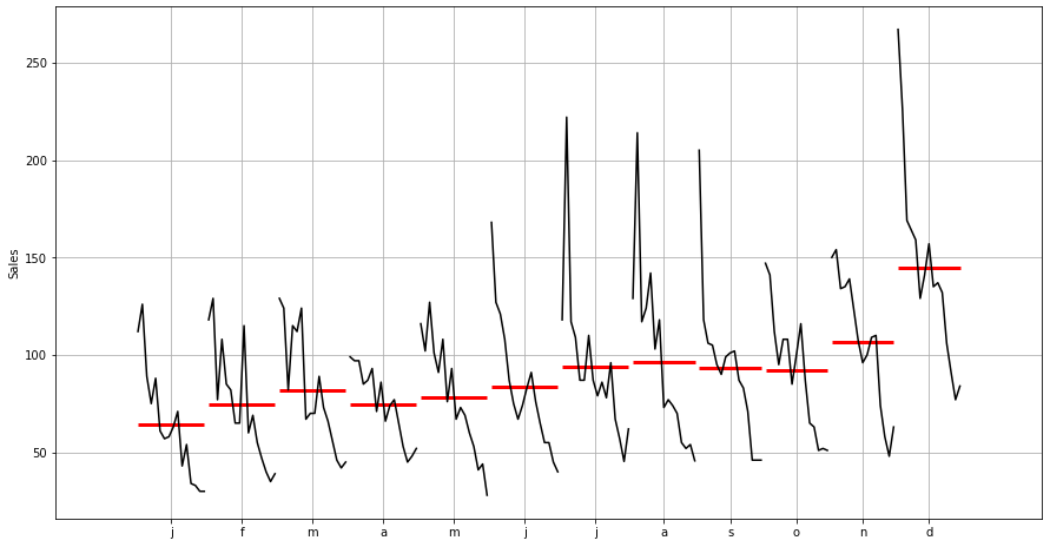


Figure 2.3 Monthly Plot

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	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.000000	124.000000	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.000000	142.000000	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.000000	103.000000	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.000000	118.000000	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.000000	73.000000	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.000000	77.000000	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.000000	74.000000	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.000000	70.000000	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.000000	55.000000	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.000000	52.000000	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.000000	54.000000	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.333333	45.666667	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.000000	NaN	NaN	NaN	NaN	NaN

Table 2.1 Pivot table Monthly sales across years

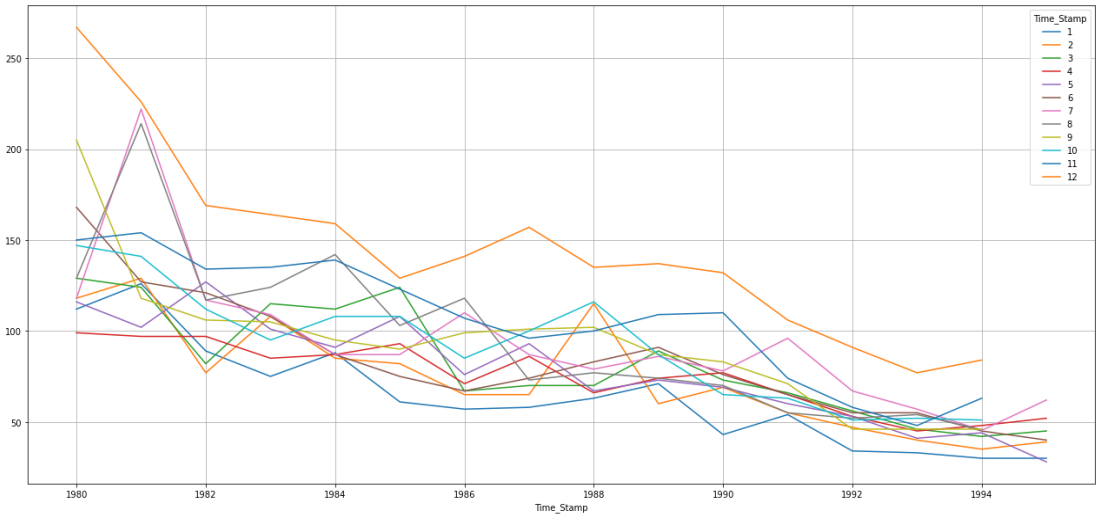


Figure 2.4 Plot of Monthly sales across years

- December records have the high number of rose wine sales
- May, January has 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	Rose
1980-12-31	1758.0
1981-12-31	1780.0
1982-12-31	1348.0
1983-12-31	1324.0
1984-12-31	1280.0

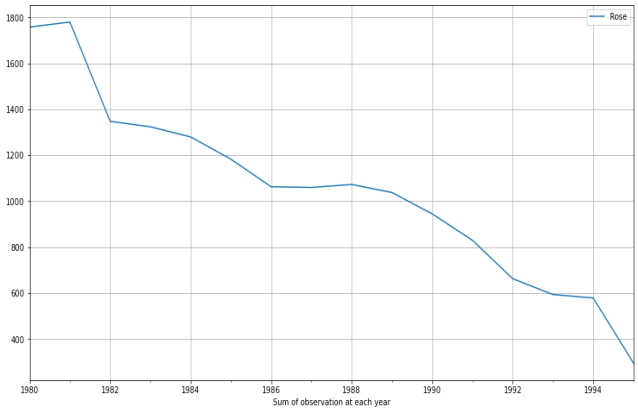


Figure 2.5 Sum of Yearly observations

We can observe that Rose wine annual sales year on year observe a downward sales trend. The steep drop post 1994 for Rose 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	Rose
1980-03-31	359.0
1980-06-30	383.0
1980-09-30	452.0
1980-12-31	564.0
1981-03-31	379.0

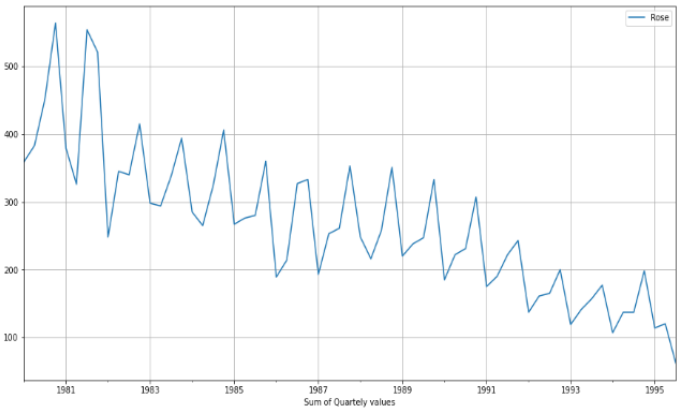


Figure 2.6 Sum of Quarterly observations

From the above tables and graphs we can find that the Quarterly sales show a downward trend for Rose wine and there is a slight element of seasonality in the time series datasets.

Empirical Cumulative Distribution

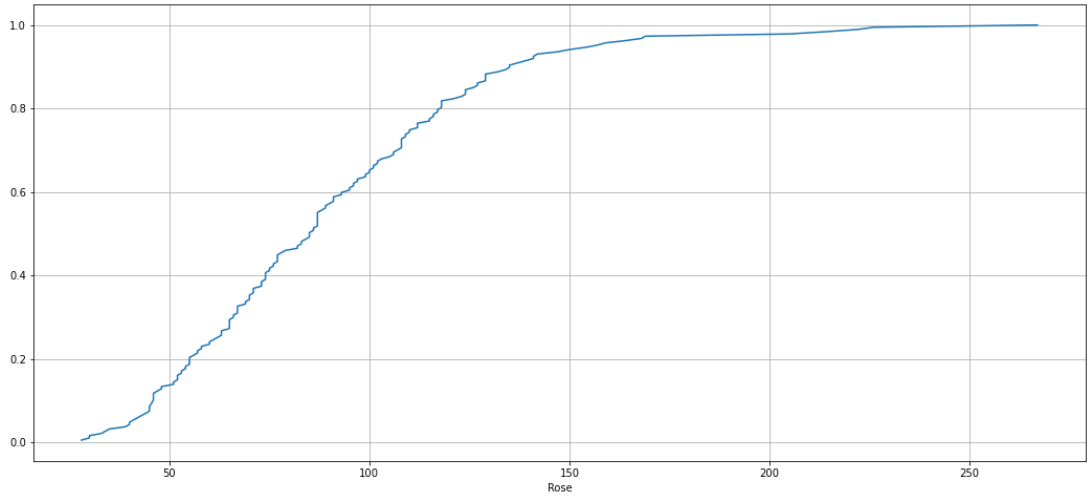


Figure 2.7 Empirical cumulative Distribution

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

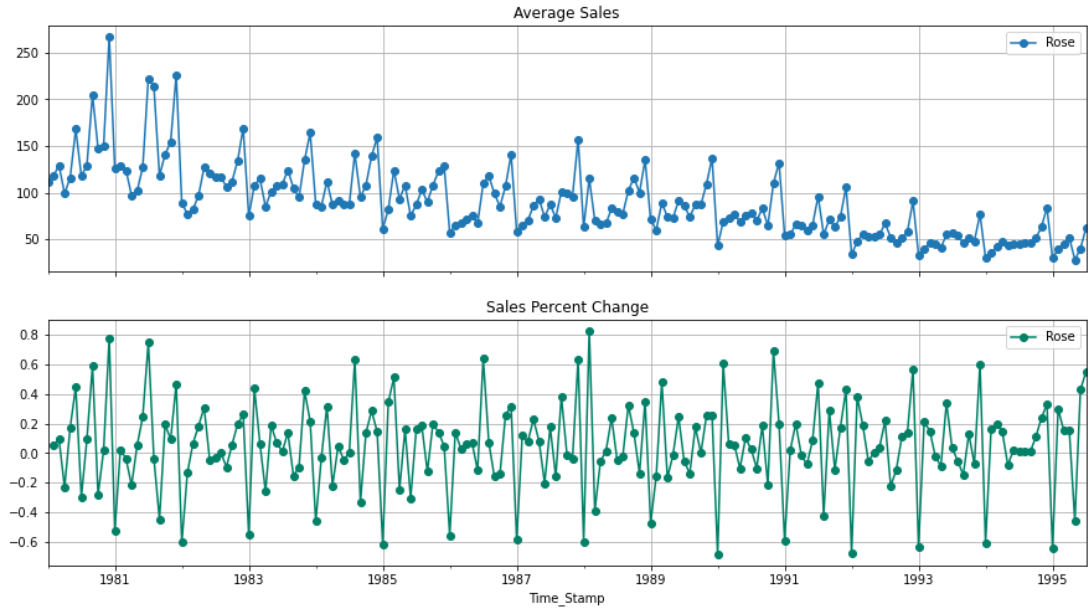


Figure 2.8 Average Sales per month & percentage change of sales

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 Rose

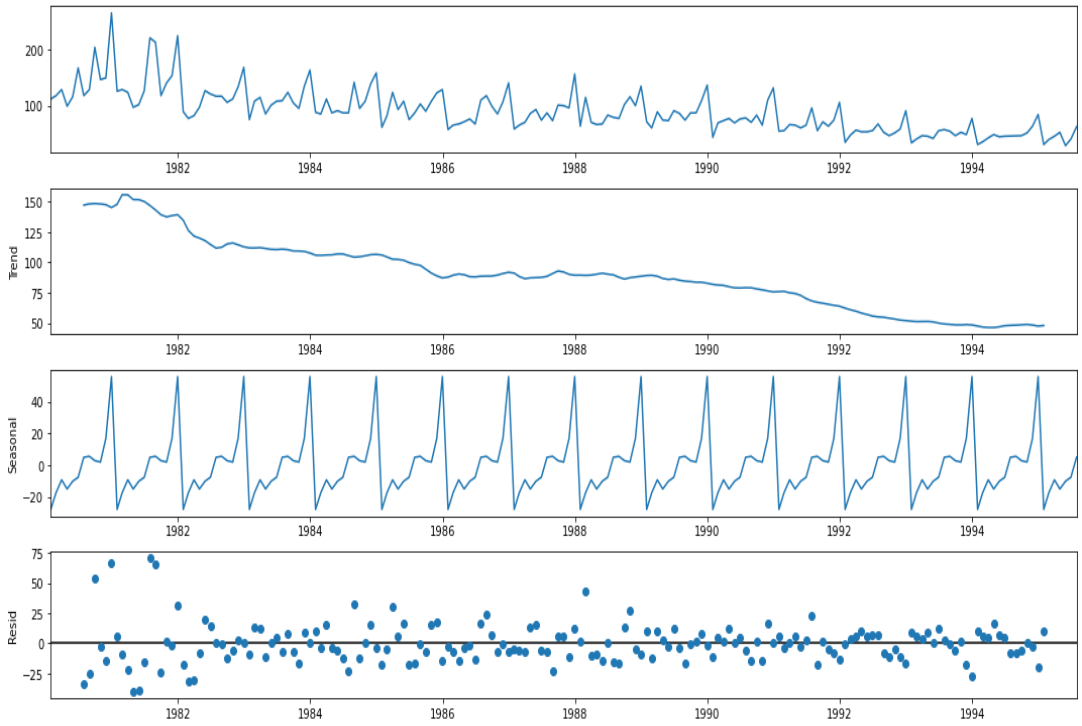


Figure 2.9 Additive decomposition

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	147.083
1980-08-31	148.125
1980-09-30	148.375
1980-10-31	148.083

Time_Stamp	Seasonality
1980-01-31	-27.909
1980-02-29	-17.436
1980-03-31	-9.286
1980-04-30	-15.098
1980-05-31	-10.197
1980-06-30	-7.679
1980-07-31	4.897
1980-08-31	5.500
1980-09-30	2.775
1980-10-31	1.872

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	-33.980
1980-08-31	-24.625
1980-09-30	53.850
1980-10-31	-2.955

Table 2.2 Trend, Seasonality, Residual (Additive)

Multiplicative Decomposition of Rose

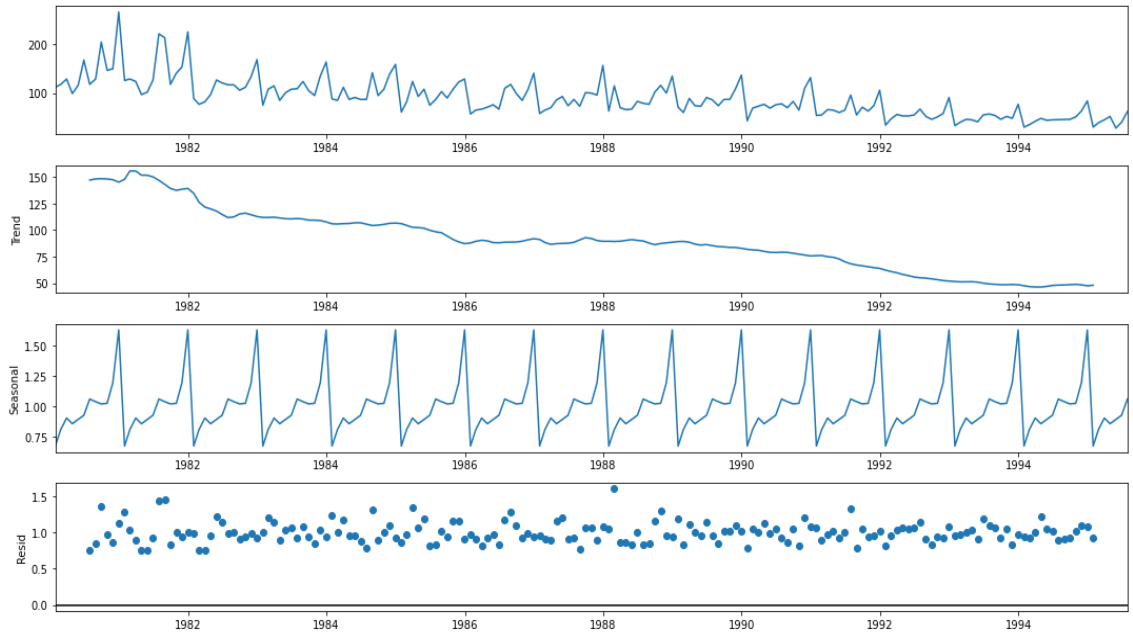


Figure 2.10 Multiplicative decomposition

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	147.083
1980-08-31	148.125
1980-09-30	148.375
1980-10-31	148.083

Time_Stamp	Seasonality
1980-01-31	0.670
1980-02-29	0.806
1980-03-31	0.901
1980-04-30	0.854
1980-05-31	0.889
1980-06-30	0.924
1980-07-31	1.058
1980-08-31	1.036
1980-09-30	1.018
1980-10-31	1.023

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	0.758
1980-08-31	0.841
1980-09-30	1.358
1980-10-31	0.971

Table 2.3 Trend, Seasonality, Residual (Multiplicative)

We can see the decomposition of the Rose time series above. I have tried with both additive and multiplicative decomposition so that I can determine if the wine datasets are 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 Rose is multiplicative in nature and have a seasonal component also rose wine is showing a downward sales trend.

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_rose[df_rose.index<'1991']
test = df_rose[df_rose.index>='1991']
```

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

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

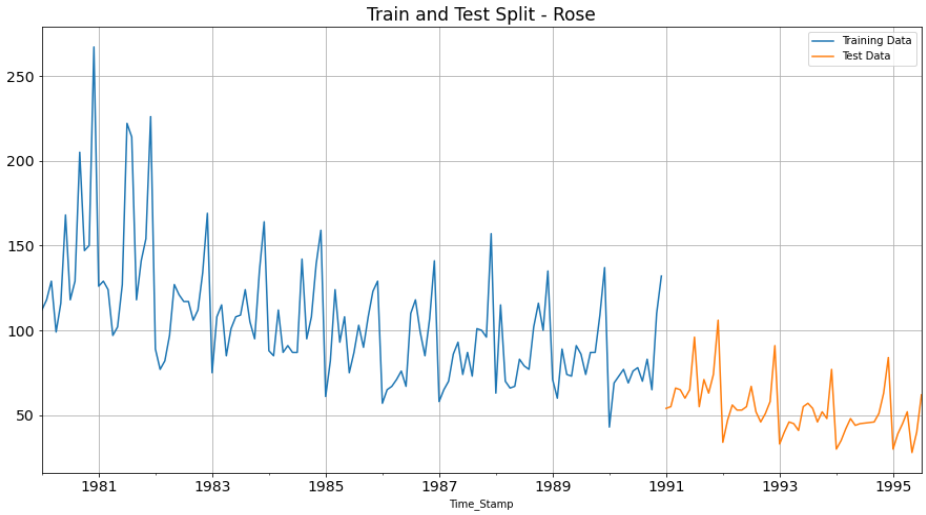


Figure 3.1 Plot of Train & Test Split

First few rows of Training Data

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

First few rows of Test Data

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

Last few rows of Training Data

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

Last few rows of Test Data

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

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

Last few rows of Training Data

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

First few rows of Test Data

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

Last few rows of Test Data

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

Table 4.1 Train & Test datasets for Linear Regression

Following are the results from a Linear Regression model

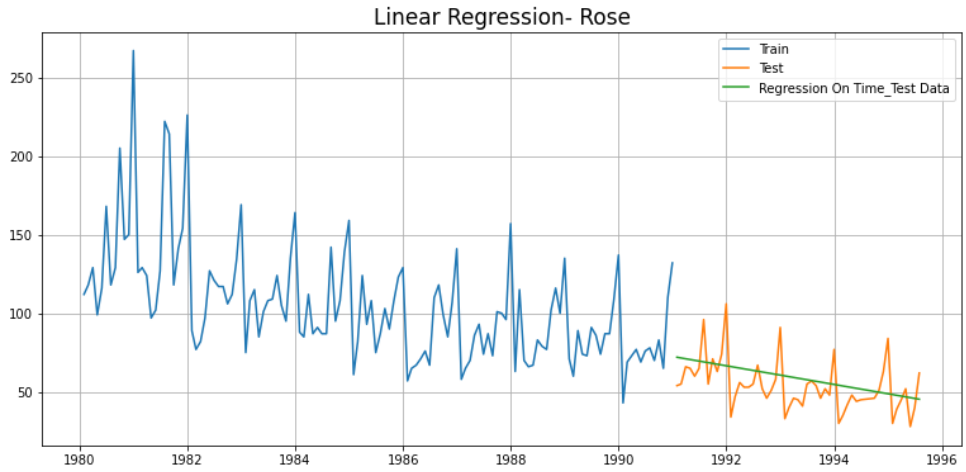


Figure 4.1 Plot for Linear Regression prediction

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

For RegressionOnTime forecast on the Test Data: RMSE is 15.269

	Test RMSE
RegressionOnTime	15.269

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

Time_Stamp	Rose	naive
1991-01-31	54.0	132.0
1991-02-28	55.0	132.0
1991-03-31	66.0	132.0
1991-04-30	65.0	132.0
1991-05-31	60.0	132.0

Table 4.2 Train & Test datasets for Naïve Bayes

Following are the results from a Naïve Bayes model

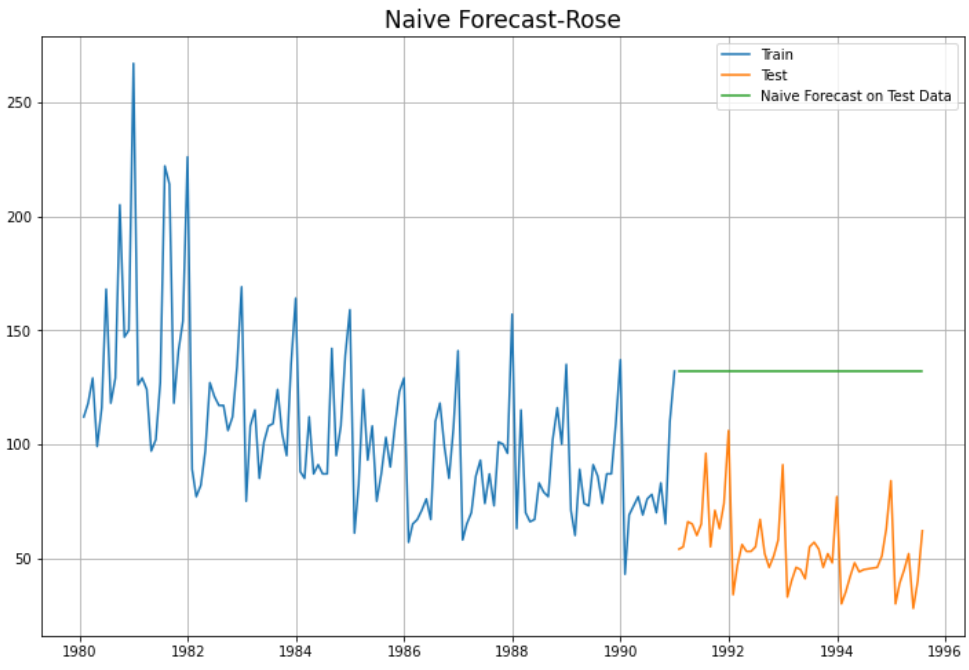


Figure 4.2 Plot for Naïve Bayes prediction

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

For Naive Model forecast on the Test Data: RMSE is 79.719

	Test RMSE
NaiveModel	79.719

Model 3- Simple Average

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

Time_Stamp	Rose	mean_forecast
1991-01-31	54.0	104.939
1991-02-28	55.0	104.939
1991-03-31	66.0	104.939
1991-04-30	65.0	104.939
1991-05-31	60.0	104.939

Table 4.3 Forecast of test data for Simple Average

Following are the results from a Simple Average model

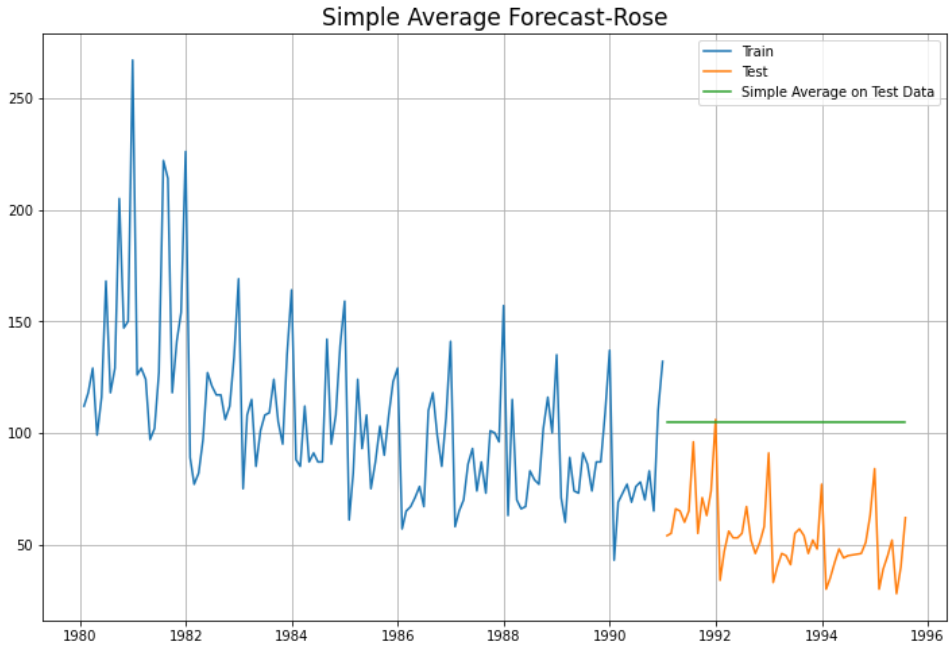


Figure 4.3 Plot for Simple Average prediction

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

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

	Test RMSE
SimpleAverageModel	53.461

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	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.50	NaN	NaN
1980-05-31	116.0	107.5	115.50	NaN	NaN
1980-06-30	168.0	142.0	128.00	123.666667	NaN
1980-07-31	118.0	143.0	125.25	124.666667	NaN
1980-08-31	129.0	123.5	132.75	126.500000	NaN
1980-09-30	205.0	167.0	155.00	139.166667	132.666667
1980-10-31	147.0	176.0	149.75	147.166667	136.555556

Table 4.4 Moving Average data

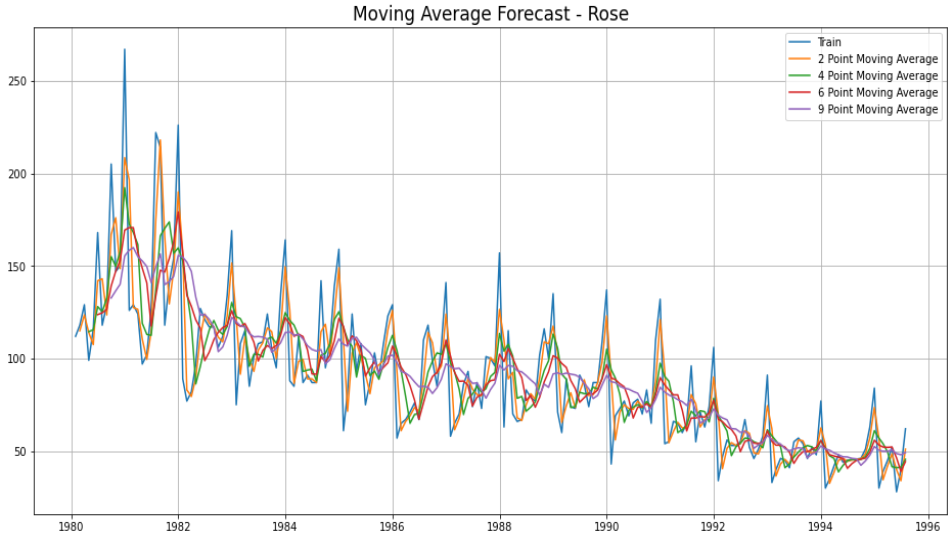


Figure 4.4 Plot for Moving Average data

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.

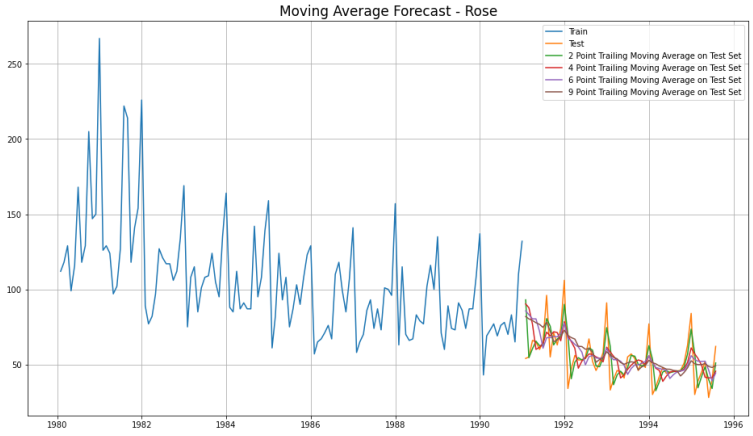


Figure 4.5 Plot for Moving Average prediction

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

	Test RMSE
2pointTrailingMovingAverage	11.529
4pointTrailingMovingAverage	14.451
6pointTrailingMovingAverage	14.566
9pointTrailingMovingAverage	14.728

I have applied 2, 4, 6- and 9-point trailing averages on the Sparkling wine data sets. As we can observe from the above plots, the 9- point trailing average plot shows the lowest prediction of all the plots. 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

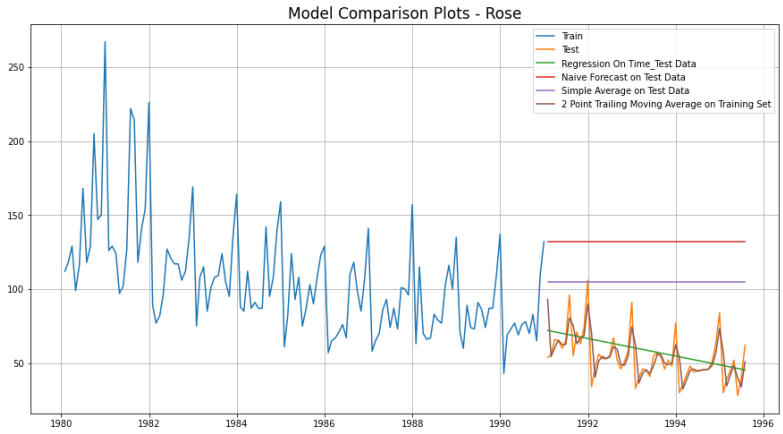


Figure 4.6 Plot for modes comparison till now

Model 5- Simple Exponential Smoothing (with best param 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 Rose wine datasets after fitting the model are:

	name	param	optimized
smoothing_level	alpha	0.098750	True
initial_level	l.0	134.387023	True

Table 4.5 Table of SES best params

From the above result we can find that the best param of alpha = 0.099

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

Table 4.6 Table of SES prediction data for α= 0.099

Following are the results from a Simple Exponential model with α= 0.099

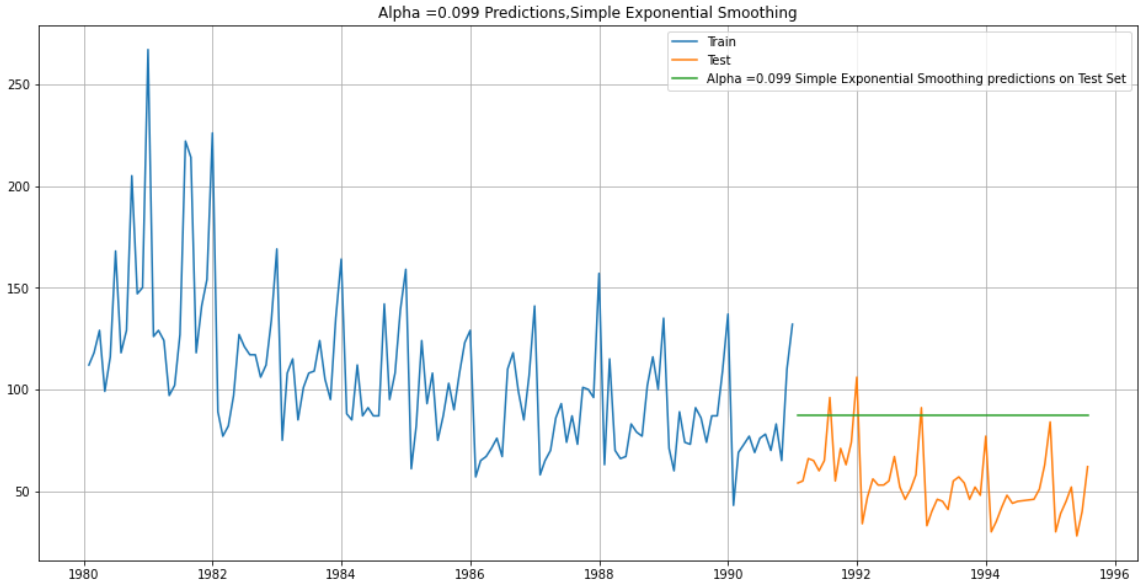


Figure 4.7 Plot of SES prediction for α= 0.099

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

	Test RMSE
Alpha=0.099:SimpleExponentialSmoothing	36.796

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	Test RMSE	Train RMSE
0	0.1	36.828033	31.815610
1	0.2	41.361876	31.979391
2	0.3	47.504821	32.470164
3	0.4	53.767406	33.035130
4	0.5	59.641786	33.682839
5	0.6	64.971288	34.441171
6	0.7	69.698162	35.323261
7	0.8	73.773992	36.334596
8	0.9	77.139276	37.482782

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

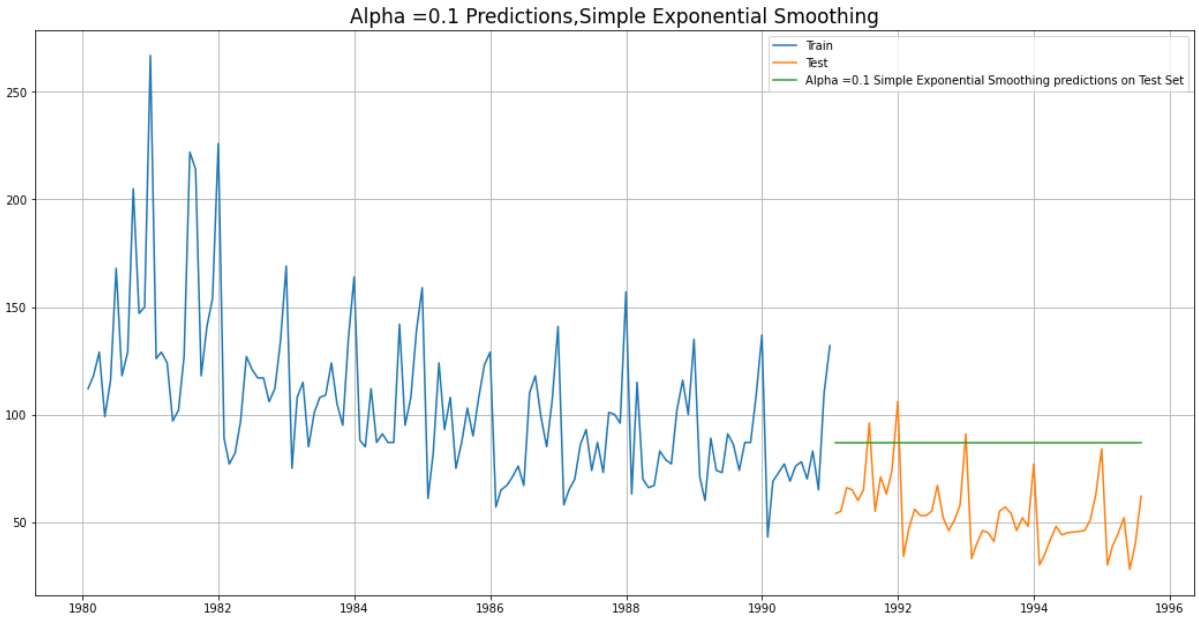


Figure 4.7 Plot of SES prediction for α= 0.1

For Alpha =0.1 Simple Exponential Smoothing Model forecast on the Test Data:
RMSE is 36.828

	Test RMSE
Alpha=0.1,SimpleExponentialSmoothing	36.828

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 =0.00013, β =0)

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 Rose wine datasets after fitting the model are:

	name	param	optimized
smoothing_level	alpha	1.321340e-04	True
smoothing_trend	beta	1.051388e-16	True
initial_level	l.0	1.362244e+02	True
initial_trend	b.0	-4.786758e-01	True

Table 4.8 Table of DES best params

From the above result we can find that the best param of α = 0.00013, β =0.

Time_Stamp	Rose	auto_predict
1991-01-31	54.0	72.569944
1991-02-28	55.0	72.091268
1991-03-31	66.0	71.612593
1991-04-30	65.0	71.133917
1991-05-31	60.0	70.655241

Table 4.9 DES prediction data with α = 0.00013, β =0

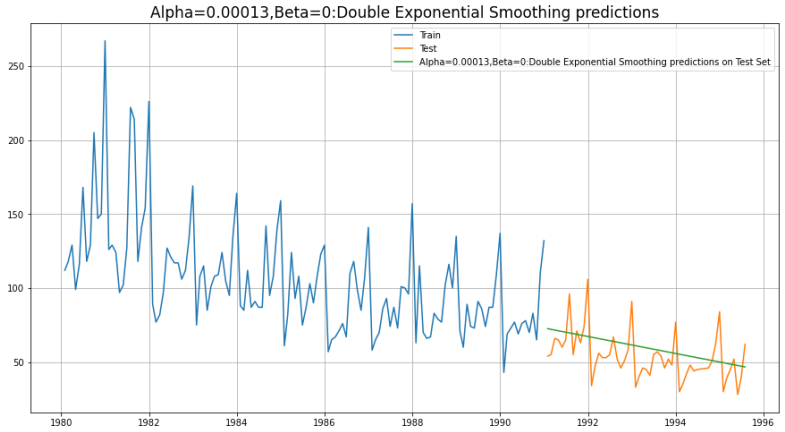


Figure 4.8 Plot of DES prediction for α = 0.00013, β =0

Following are the results from a Double Exponential model with α = 0.00013, β =0. 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=0.00013, Beta=0: Double Exponential Smoothing predictions forecast on the Test Data: RMSE is 15.569

	Test RMSE
Alpha=0.00013,Beta=0 :DoubleExponentialSmoothing	15.569

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	34.439111	36.923416
1	0.1	0.2	33.450729	48.688648
9	0.2	0.1	33.097427	65.731702
2	0.1	0.3	33.145789	78.156641
18	0.3	0.1	33.611269	98.653317

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:

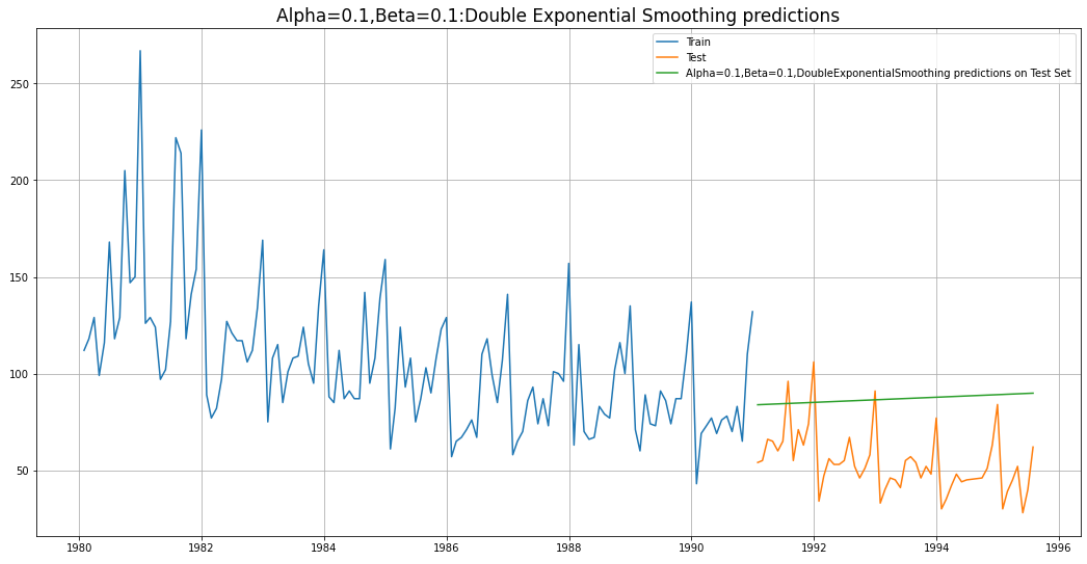


Figure 4.9 Plot of DES prediction for α = 0.1, β =0.1

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 36.923. We can see the DES model is not performing well.

	Test RMSE
Alpha=0.1&Beta=0.1,DoubleExponentialSmoothing	36.923

Model 9: Triple Exponential Smoothing (Holt - Winter's Model)(α =0.064 , β =0.053 & γ=0

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 Rose wine datasets after fitting the model are:

	name	param	optimized
smoothing_level	alpha	0.064672	True
smoothing_trend	beta	0.053159	True
smoothing_seasonal	gamma	0.000000	True
initial_level	l.0	50.880913	True
initial_trend	b.0	-0.316568	True
initial_seasons.0	s.0	2.215837	True
initial_seasons.1	s.1	2.514395	True
initial_seasons.2	s.2	2.746930	True
initial_seasons.3	s.3	2.401184	True
initial_seasons.4	s.4	2.699363	True
initial_seasons.5	s.5	2.943381	True
initial_seasons.6	s.6	3.235389	True
initial_seasons.7	s.7	3.440529	True
initial_seasons.8	s.8	3.264207	True
initial_seasons.9	s.9	3.193652	True
initial_seasons.10	s.10	3.722694	True
initial_seasons.11	s.11	5.134358	True

Table 4.11 TES best params

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

Table 4.12 TES prediction data for α =0.064, β =0.053 & γ=0

Following are the results from a Triple Exponential model with α =0.064, β =0.053 & γ=0. We see that the Triple Exponential Smoothing is picking up the seasonal component as well.

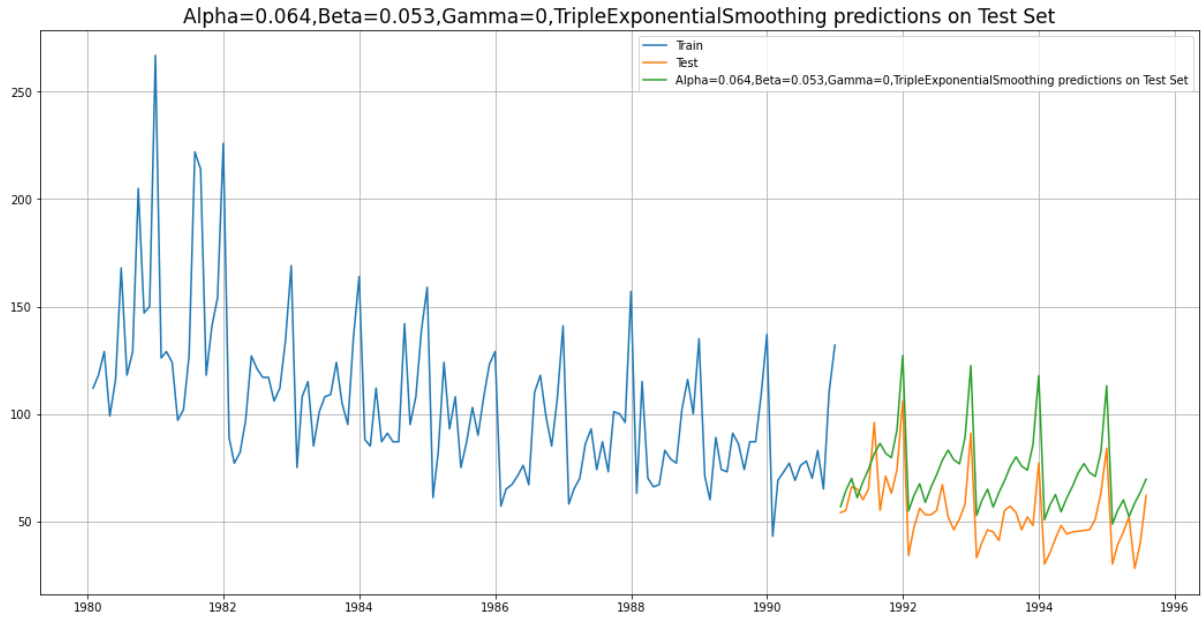


Figure 4.10 Plot of TES prediction data for α =0.064, β =0.053 & γ=0

For Alpha=0.064 ,Beta=0.053,Gamma=0,TripleExponentialSmoothing predictions forecast o n the Test Data RMSE is 21.155

	Test RMSE
Alpha=0.064,Beta=0.053,Gamma=0,TripleExponentialSmoothing	21.155

Model 10: Triple Exponential Smoothing (Holt - Winter's Model) (α =0.1 , β =0.2 & 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
10	0.1	0.2	0.2	24.365597	9.640687
11	0.1	0.2	0.3	23.969166	9.935740
9	0.1	0.2	0.1	25.529854	9.943539
119	0.2	0.5	0.3	27.631767	10.026210
127	0.2	0.6	0.2	28.289836	10.031639

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

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

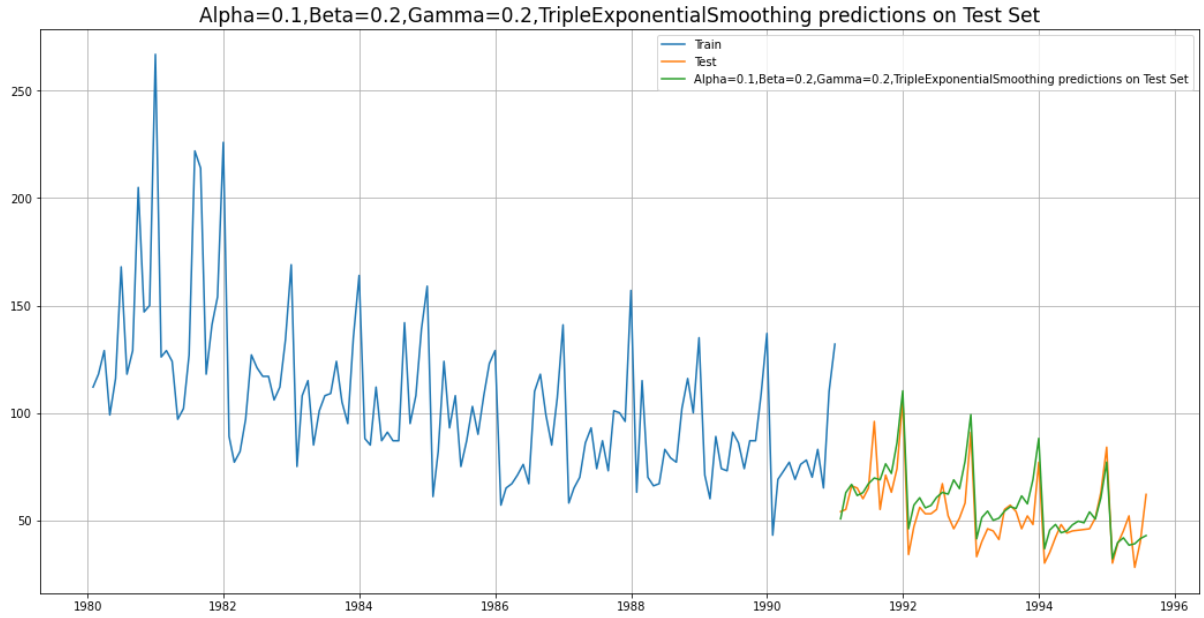


Figure 4.11 Plot of TES prediction data for α =0.1, β =0.2 & γ=0.2

For Alpha=0.4, Beta=0.1, Gamma=0.2, TripleExponentialSmoothing predictions forecast on the Test Data RMSE is 9.641

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2:TripleExponentialSmoothing	9.641

Triple Exponential Smoothing with Alpha=0.1, Beta=0.2, 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

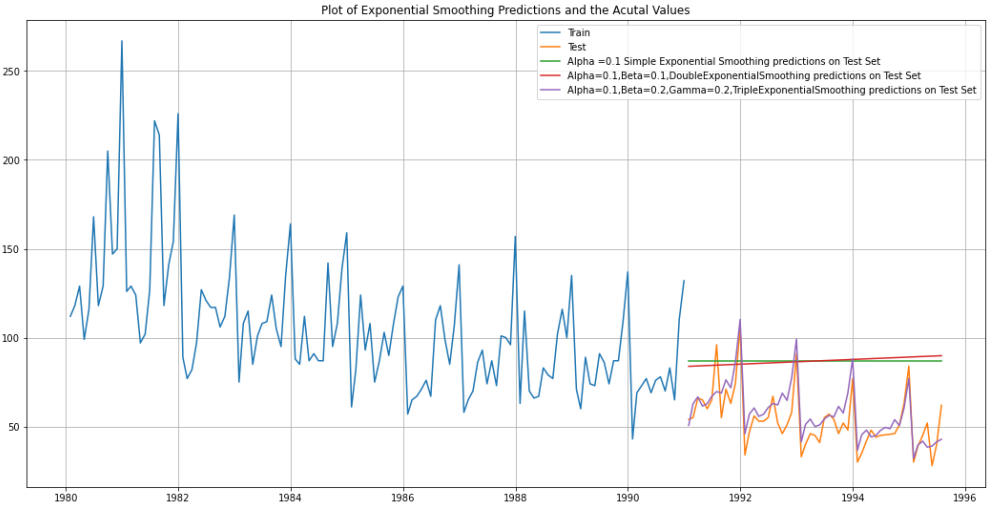


Figure 4.12 Plot of Exponential Smoothing predictions

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2:TripleExponentialSmoothing	9.641
2pointTrailingMovingAverage	11.529
4pointTrailingMovingAverage	14.451
6pointTrailingMovingAverage	14.566
9pointTrailingMovingAverage	14.728
RegressionOnTime	15.269
Alpha=0,Beta=0 :DoubleExponentialSmoothing	15.569
Alpha=0.064,Beta=0.053,Gamma=0,TripleExponentialSmoothing	21.155
Alpha=0.099:SimpleExponentialSmoothing	36.796
Alpha=0.1,SimpleExponentialSmoothing	36.828
Alpha=0.1&Beta=0.1,DoubleExponentialSmoothing	36.923
SimpleAverageModel	53.461
NaiveModel	79.719

Table 4.14 Dataframe with all models Test RMSE values

Triple Exponential Smoothing with Alpha=0.1, Beta=0.2, Gamma=0.2 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:

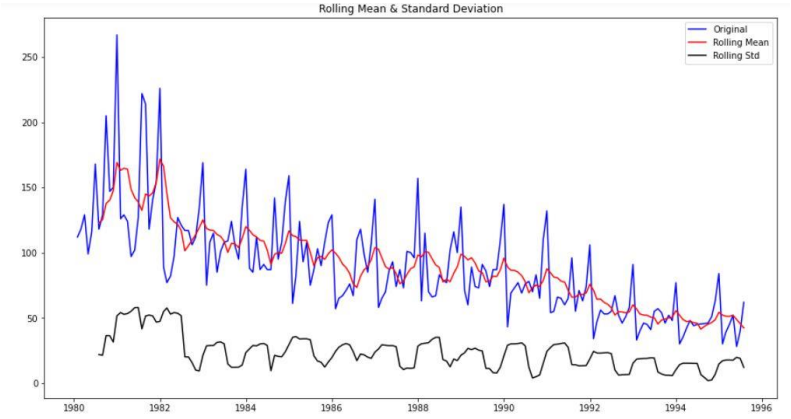
- $H0$: The Time Series has a unit root and is thus non-stationary.
- $H1$: The Time Series does not have a unit root and is thus stationary.

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

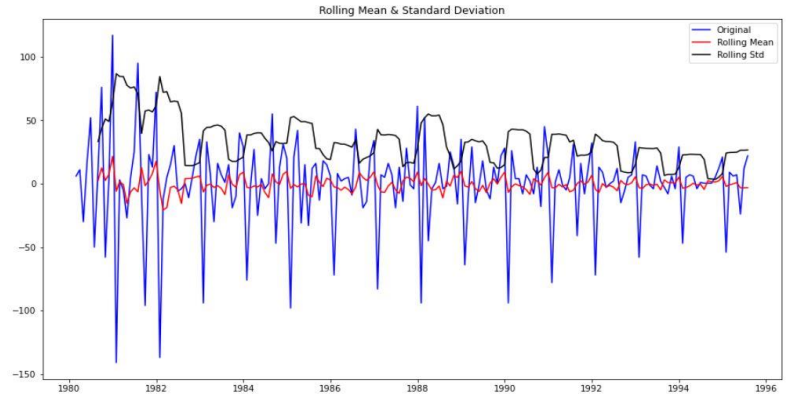



```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

Figure 5.1 adfuller test for whole data

Augmented Dicky-Fuller Test was applied to the whole Rose dataset.

We found, p-value = 0.3431. Here, p-value > alpha=0.05. We fail to reject the Null Hypothesis and hence conclude that Rose 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                -8.044392e+00
p-value                        1.810895e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

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

Now, p-value = 1.810895e-12. As, p-value < alpha=0.05 we reject the Null Hypothesis and c onclude that Rose Time Series is Stationary with a lag of 1.

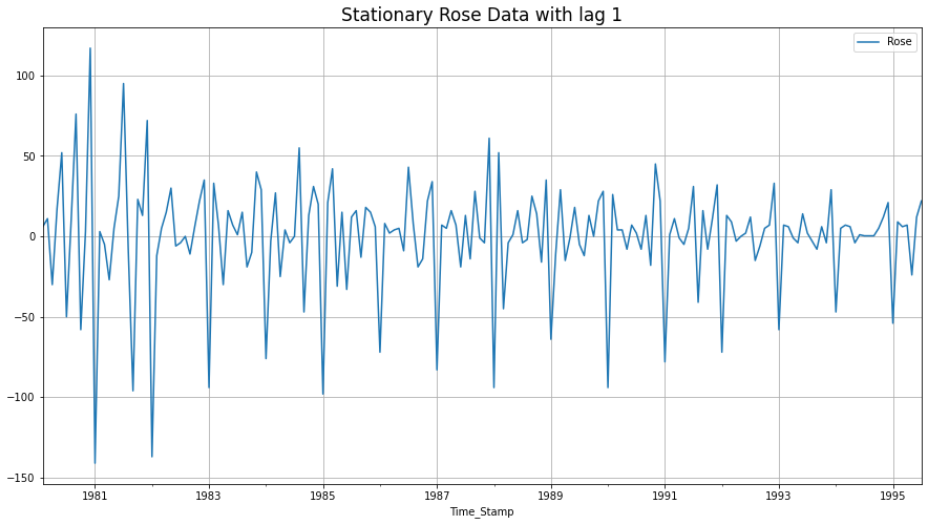


Figure 5.3 is Stationary od data with a lag of 1

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

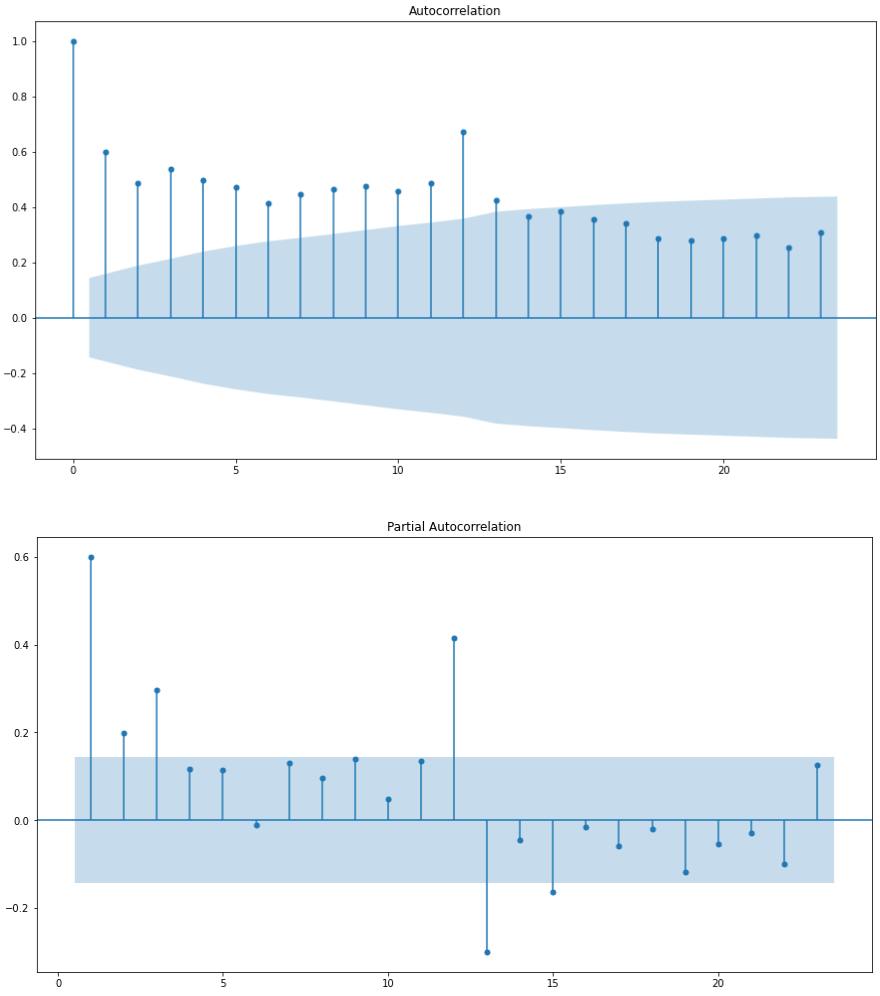
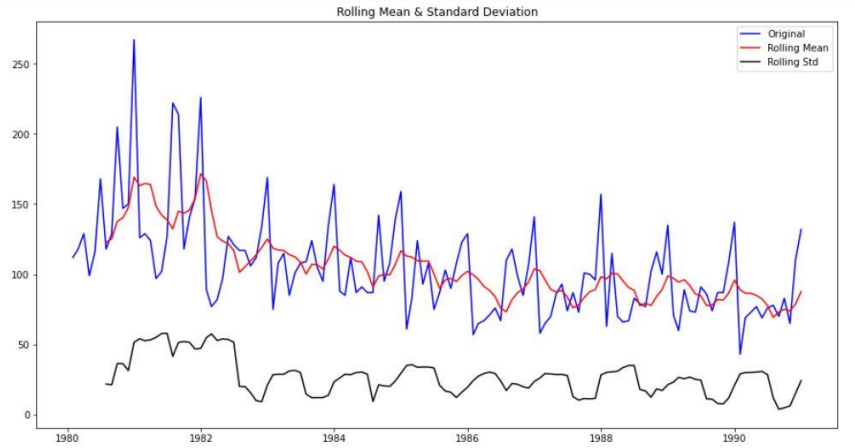


Figure 5.4 ACF & PACF plots for whole data with lag 1

Check for stationarity of the Training Data Time Series

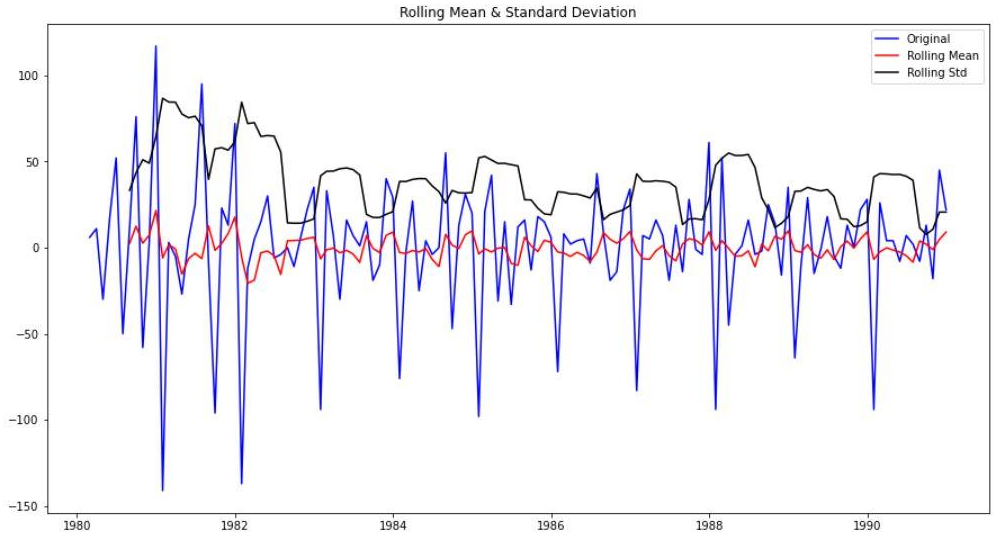
We check for stationarity of Train Rose data by using Augmented Dicky Fuller Test. We take a difference of 1 and make the dataset Stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

Figure 5.5 adfuller test for train data

We found, p-value = 0.21947. Here, p-value > alpha=0.05. We fail to reject the Null Hypothesis and hence conclude that Rose 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                -6.592372e+00
p-value                        7.061944e-09
#Lags Used                     1.200000e+01
Number of Observations Used    1.180000e+02
Critical Value (1%)           -3.487022e+00
Critical Value (5%)           -2.886363e+00
Critical Value (10%)          -2.580009e+00
dtype: float64
```

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

Now, p-value = 7.061944e-09. 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
2	(0, 1, 2)	1279.671529
5	(1, 1, 2)	1279.870723
4	(1, 1, 1)	1280.574230
7	(2, 1, 1)	1281.507862
8	(2, 1, 2)	1281.870722

Table 6.1 Auto ARIMA with lowest AIC values

- The best combination with Least AIC is - (p, d, q) is (0,1,2)

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(0, 1, 2)	Log Likelihood	-636.836
Date:	Thu, 17 Mar 2022	AIC	1279.672
Time:	23:47:19	BIC	1288.297
Sample:	01-31-1980 - 12-31-1990	HQIC	1283.176
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.6970	0.072	-9.689	0.000	-0.838	-0.556
ma.L2	-0.2042	0.073	-2.794	0.005	-0.347	-0.061
sigma2	965.8407	88.305	10.938	0.000	792.766	1138.915

Ljung-Box (L1) (Q):	0.14	Jarque-Bera (JB):	39.24
Prob(Q):	0.71	Prob(JB):	0.00
Heteroskedasticity (H):	0.36	Skew:	0.82
Prob(H) (two-sided):	0.00	Kurtosis:	5.13

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

Figure 6.1 Auto ARIMA (0,1,2) Summary

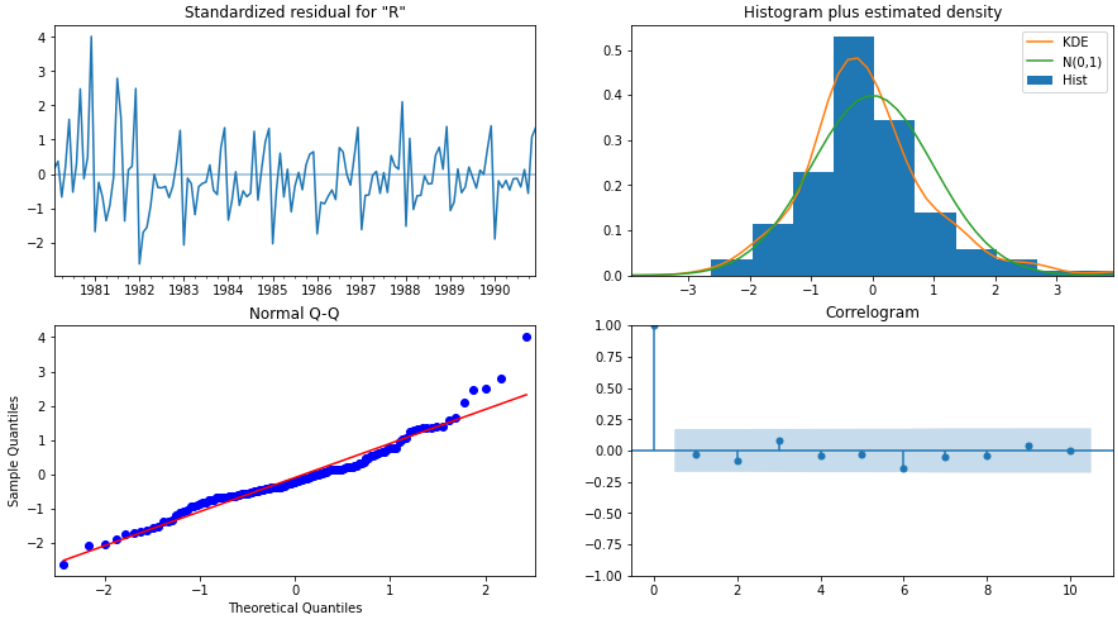


Figure 6.2 Residual Diagnostics of Auto ARIMA (0,1,2)

RMSE for ARIMA (2,1,2) based on AIC value is: 37.306

	Test RMSE
ARIMA(0,1,2)based on AIC values	37.306

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

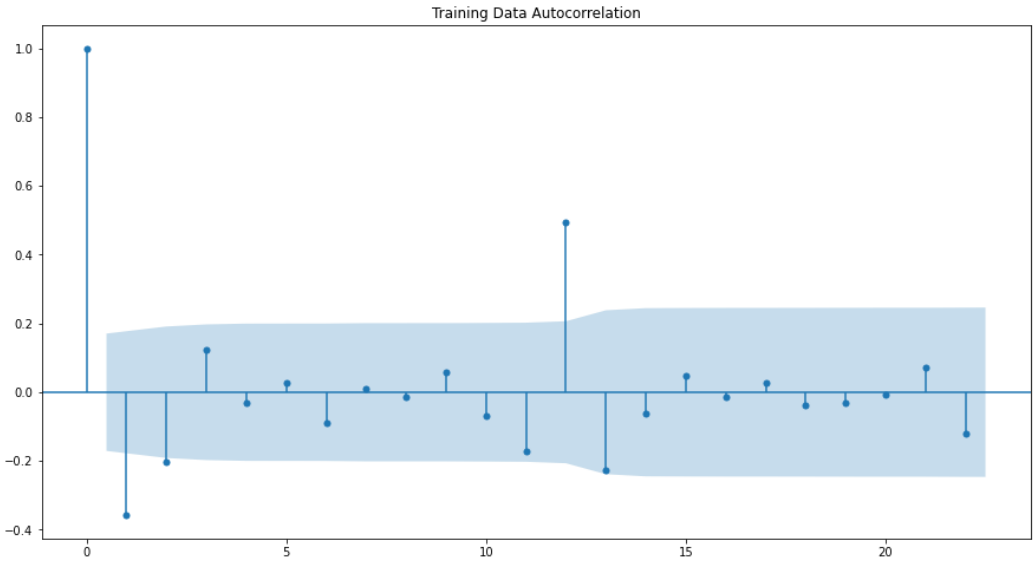


Figure 6.3 ACF plot to check seasonality value

- 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
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
53	(1, 1, 2)	(2, 0, 2, 12)	896.686896
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

Table 6.2 Auto SARIMA with lowest AIC values

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

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-436.969
Date:	Thu, 17 Mar 2022	AIC	887.938
Time:	23:58:01	BIC	906.448
Sample:	0	HQIC	895.437
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8427	189.757	-0.004	0.996	-372.760	371.074
ma.L2	-0.1573	29.812	-0.005	0.996	-58.587	58.273
ar.S.L12	0.3467	0.079	4.375	0.000	0.191	0.502
ar.S.L24	0.3023	0.076	3.996	0.000	0.154	0.451
ma.S.L12	0.0767	0.133	0.577	0.564	-0.184	0.337
ma.S.L24	-0.0726	0.146	-0.498	0.618	-0.358	0.213
sigma2	251.3136	4.77e+04	0.005	0.996	-9.32e+04	9.37e+04

Ljung-Box (L1) (Q):	0.10	Jarque-Bera (JB):	2.33
Prob(Q):	0.75	Prob(JB):	0.31
Heteroskedasticity (H):	0.88	Skew:	0.37
Prob(H) (two-sided):	0.70	Kurtosis:	3.03

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

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

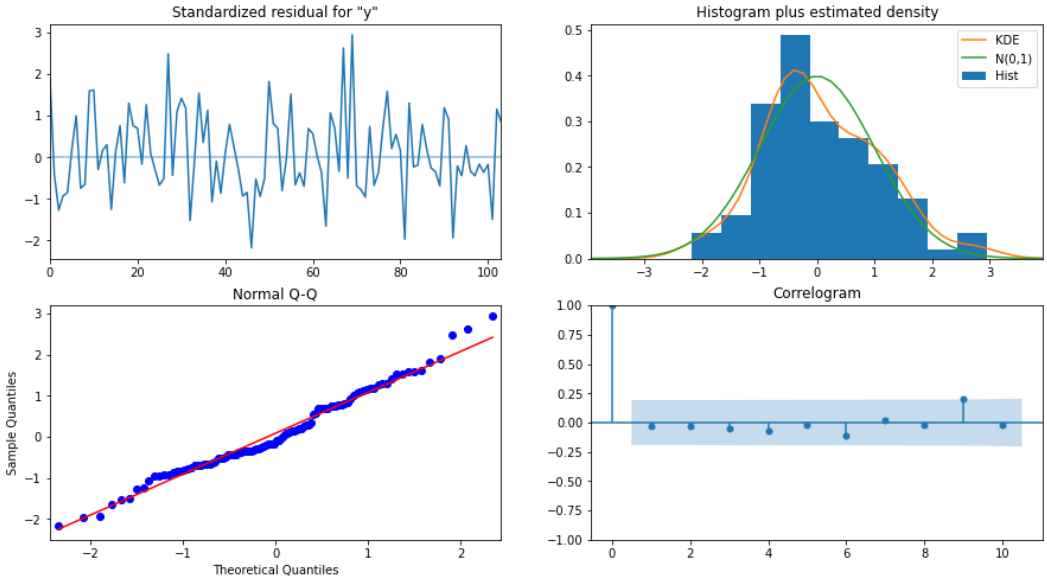


Figure 6.5 Residual Diagnostics of Auto SARIMA (0,1,2)(2,0,2,12)

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.867263	15.928501	31.647976	94.086550
1	70.541190	16.147658	38.892361	102.190019
2	77.356410	16.147656	45.707586	109.005235
3	76.208814	16.147656	44.559989	107.857638
4	72.747398	16.147656	41.098573	104.396222

Table 6.3 Auto SARIMA model

RMSE for SARIMA (0,1,2) (2,0,2,12) based on AIC value 26.928

	Test RMSE
SARIMA(0,1,2)(2,0,2,12) based on AIC value	26.928

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(0,1,2)based on AIC value	37.306
SARIMA(0,1,2)(2,0,2,12)based on ACF & PACF values	26.928

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 (2,1,2)

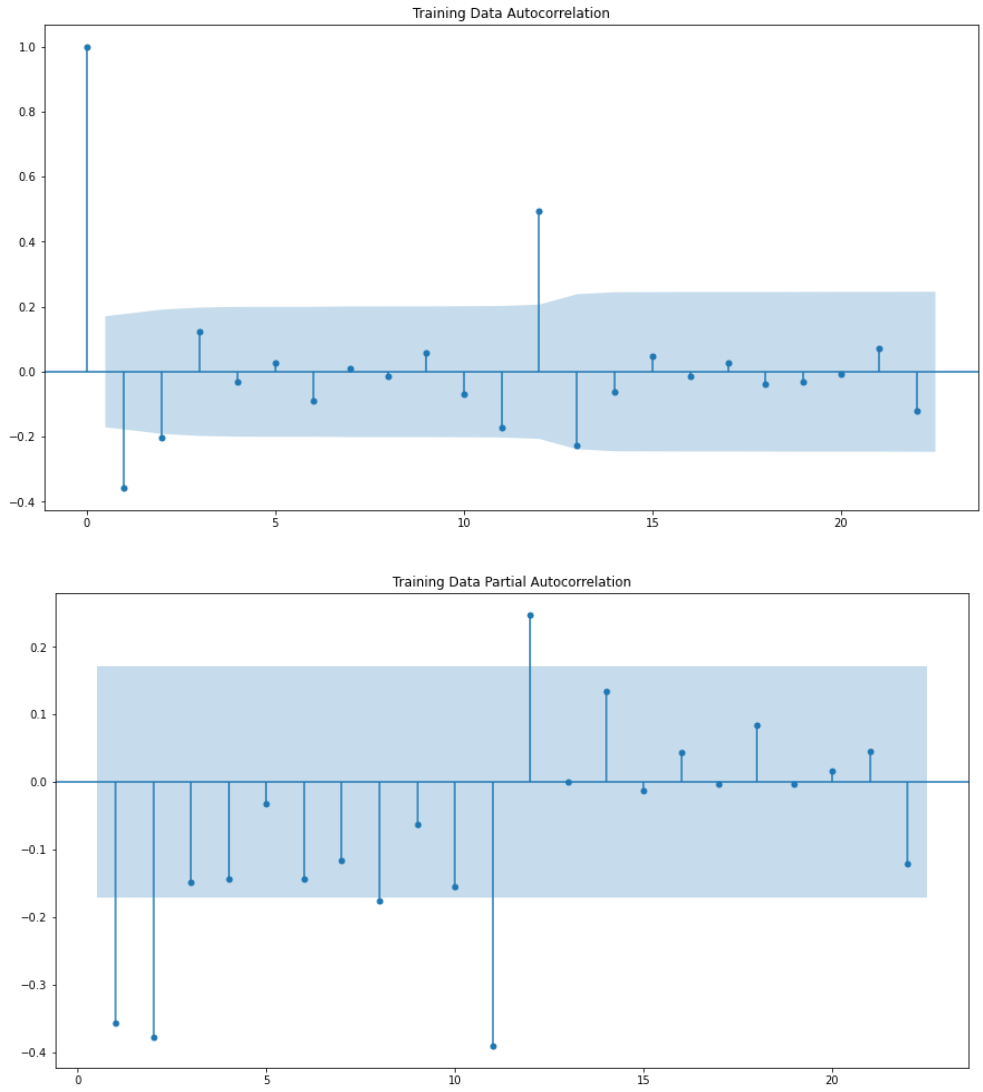


Figure 7.1 ACF & PACF plot for first difference

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 2.

The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.

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 = 2, q = 2 and difference d = 1 i.e (2,1,2)

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Thu, 17 Mar 2022	AIC	1281.871
Time:	23:58:23	BIC	1296.247
Sample:	01-31-1980	HQIC	1287.712
	- 12-31-1990		
Covariance Type:	opg		

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

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

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

Figure 7.2 Manual ARIMA (2,1,2) Summary

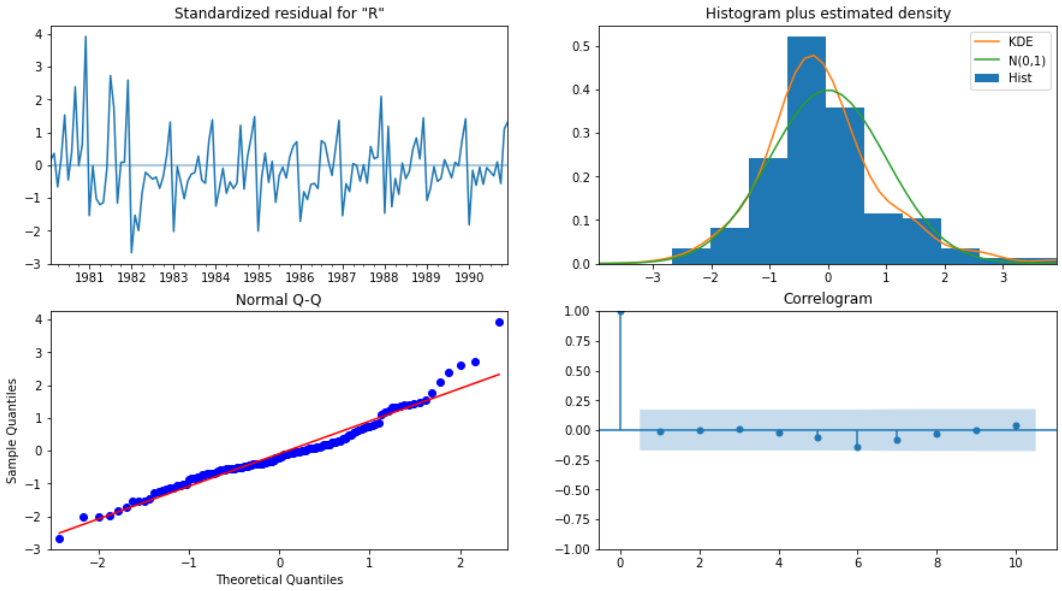


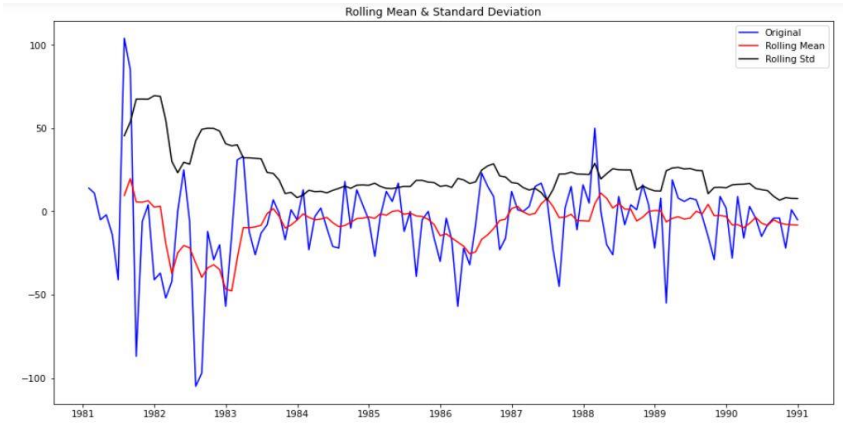
Figure 7.3 Residual Diagnostics of manual ARIMA (2,1,2)

RMSE for manual ARIMA(2,1,2) based on ACF & PACF values is: 36.871

	Test RMSE
ARIMA(2,1,2)based on ACF & PACF values	36.871

SARIMA Manual with seasonality 12 (2,1,2) (2,0,1,12)

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




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.619482
p-value                          0.005399
#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
dtype: float64
```

Figure 7.4 adfuller test for train data with seasonality =12

From the above check we can find that p value = 0.0053. 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.

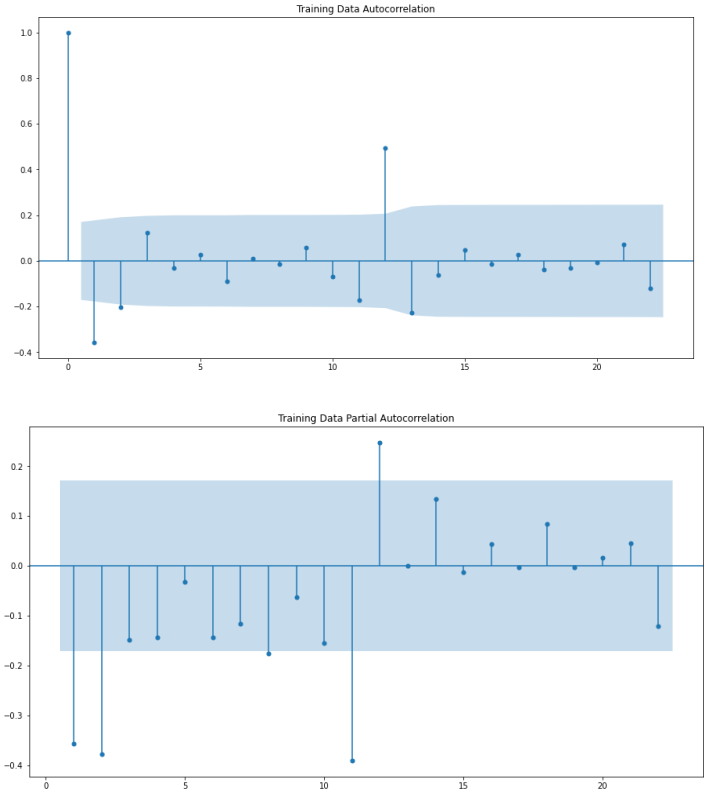


Figure 7.5 ACF & PACF plot of train with seasonality =12

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 2 and the q value also to be 2 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 2.(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 = 2, q = 2, d = 1 and P = 2, D = 0, Q = 1, Seasonality=12 (2,1,2) (2,0,1,12)

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(2, 0, [1], 12)	Log Likelihood	-441.189
Date:	Thu, 17 Mar 2022	AIC	898.378
Time:	23:58:51	BIC	919.610
Sample:	0	HQIC	906.982
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4772	0.305	1.565	0.118	-0.121	1.075
ar.L2	-0.1667	0.104	-1.608	0.108	-0.370	0.037
ma.L1	-1.3270	287.087	-0.005	0.996	-564.008	561.354
ma.L2	0.3270	93.924	0.003	0.997	-183.760	184.414
ar.S.L12	0.3280	0.082	3.983	0.000	0.167	0.489
ar.S.L24	0.2831	0.070	4.047	0.000	0.146	0.420
ma.S.L12	0.1309	0.131	0.998	0.318	-0.126	0.388
sigma2	248.8224	7.14e+04	0.003	0.997	-1.4e+05	1.4e+05

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	2.96
Prob(Q):	0.90	Prob(JB):	0.23
Heteroskedasticity (H):	1.01	Skew:	0.37
Prob(H) (two-sided):	0.99	Kurtosis:	3.34

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

Figure 7.6 Manual SARIMA (2,1,2) (2,0,1,12) Summary

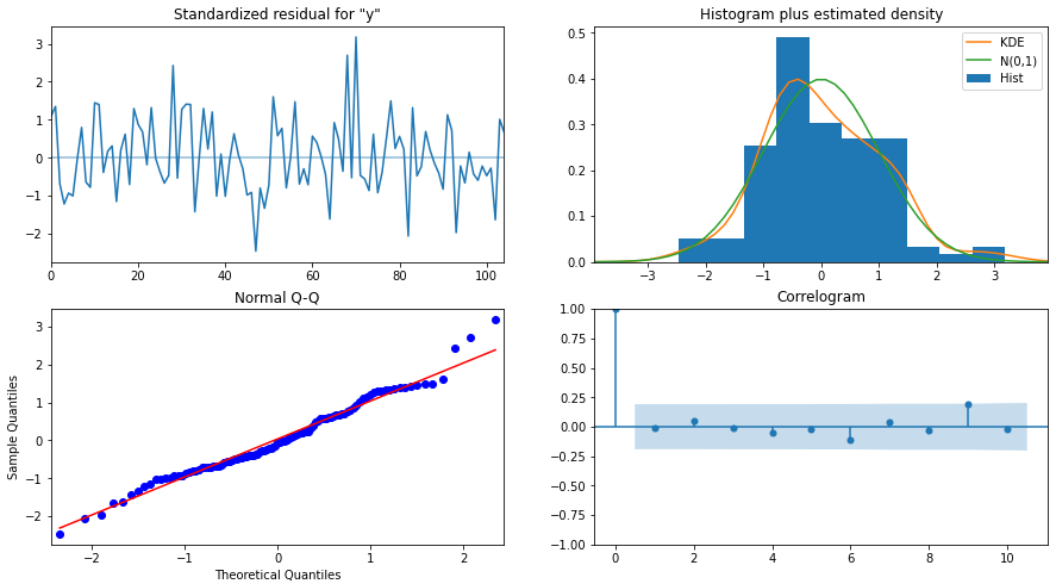


Figure 7.7 Residual Diagnostics of manual SARIMA (2,1,2) (2,0,1,12)

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	63.184121	15.846990	32.124593	94.243650
1	67.518949	16.046319	36.068742	98.969157
2	77.724938	16.101126	46.167311	109.282565
3	77.269078	16.129080	45.656662	108.881494
4	73.696201	16.129036	42.083872	105.308530

Table 6.5 Manual SARIMA model

RMSE for manual SARIMA(0,1,0)(1,0,1,12) based on ACF & PACF values is: 28.22

	Test RMSE
SARIMA(2,1,2)(2,0,1,12)based on ACF & PACF values	28.222

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(2,1,2)based on ACF & PACF values	36.871
SARIMA(2,1,2)(2,0,1,12)based on ACF & PACF values	28.222

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=0.1,Beta=0.2,Gamma=0.2:TripleExponentialSmoothing	9.641
2pointTrailingMovingAverage	11.529
4pointTrailingMovingAverage	14.451
6pointTrailingMovingAverage	14.566
9pointTrailingMovingAverage	14.728
RegressionOnTime	15.269
Alpha=0.00013,Beta=0 :DoubleExponentialSmoothing	15.569
Alpha=0.064,Beta=0.053,Gamma=0,TripleExponentialSmoothing	21.155
SARIMA(0,1,2)(2,0,2,12) based on AIC value	26.928
SARIMA(2,1,2)(2,0,1,12)based on ACF & PACF values	28.222
Alpha=0.099:SimpleExponentialSmoothing	36.796
Alpha=0.1,SimpleExponentialSmoothing	36.828
ARIMA(2,1,2)based on ACF & PACF values	36.871
Alpha=0.1&Beta=0.1,DoubleExponentialSmoothing	36.923
ARIMA(0,1,2)based on AIC values	37.306
SimpleAverageModel	53.461
NaiveModel	79.719

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=0.1, Beta=0.2, Gamma=0.2: TripleExponentialSmoothing = 9.641** 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 best model is the Triple Exponential Smoothing with additive trend & multiplicative seasonality and the parameters α = 0.1, β = 0.2 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=0.1, smoothing_trend=0.2, smoothing_seasonal=0.2

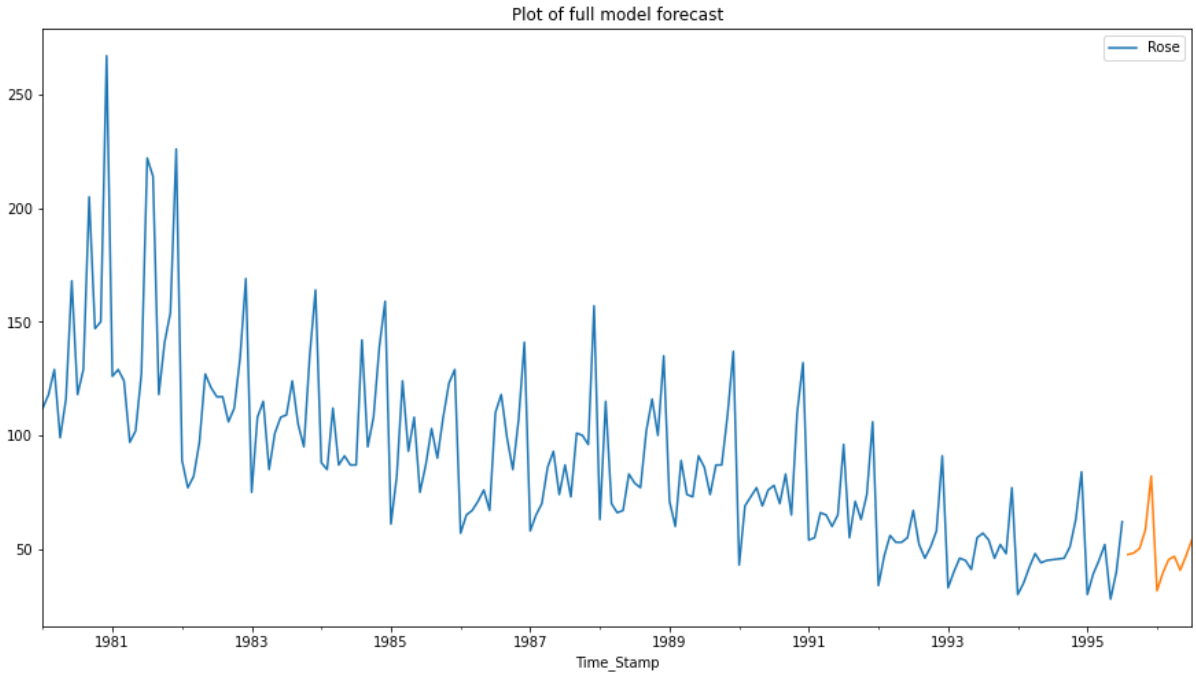


Figure 9.1 Plot of full model forecast

RMSE for full model with Triple Exponential Smoothing α = 0.1, β = 0.2 and γ = 0.2 is: 17.404

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	13.422360	47.607992	81.793624
1995-09-30	14.098851	48.284483	82.470114
1995-10-31	16.094027	50.279659	84.465291
1995-11-30	24.275597	58.461229	92.646861
1995-12-31	47.930965	82.116597	116.302229
1996-01-31	-2.489284	31.696347	65.881979
1996-02-29	5.246178	39.431810	73.617442
1996-03-31	11.174395	45.360027	79.545658
1996-04-30	12.617372	46.803004	80.988636
1996-05-31	6.536603	40.722235	74.907866
1996-06-30	12.808923	46.994555	81.180186
1996-07-31	19.855470	54.041102	88.226733

Table 9.1 12-month forecast along with confidence band

The forecasted value is from August 1995 – July 1996

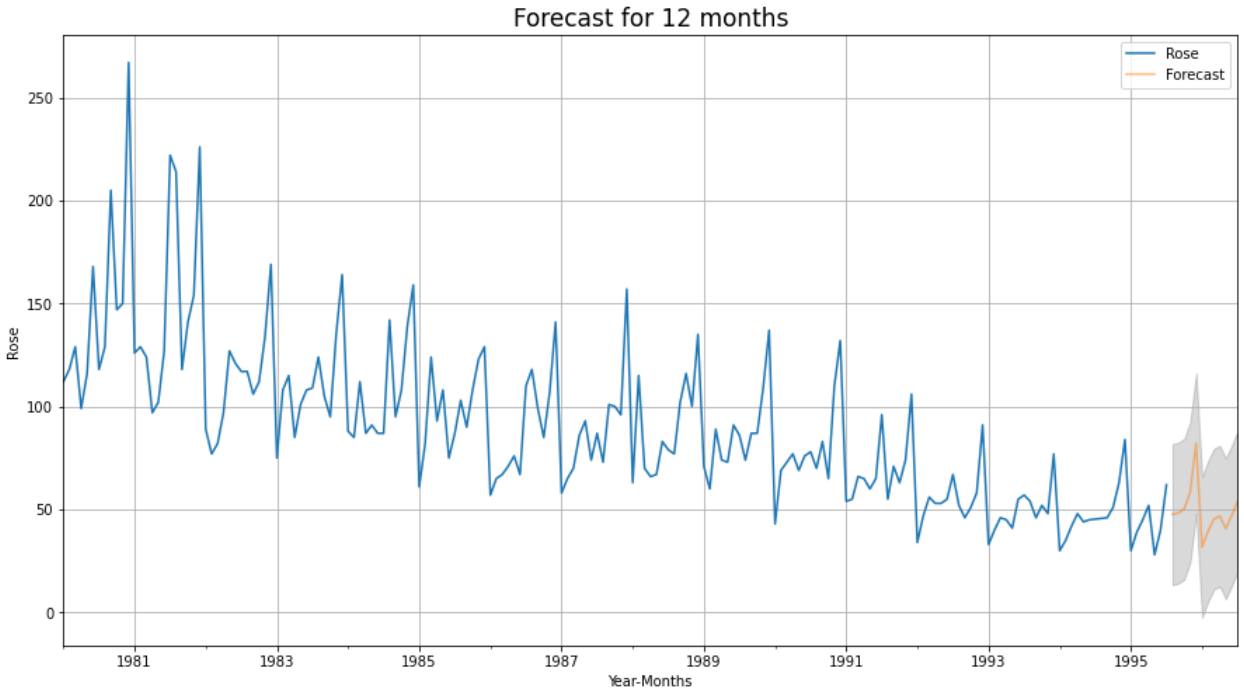


Figure 9.2 Plot of 12-month forecast along with confidence band

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:

- Rose wine shows a clear trend of declining sales from 1980 to 1995. This shows decline in popularity of this variant of wine.
- There is seasonality present in our dataset, we have peaks in December followed by November, and the least in January. This might be due to the Holiday season in this period.
- The lowest sale of rose wine is in every year from January. This could be due to some New Year resolutions and the end of vacation.
- From 1980 to 1982 the peaks of rose wine consumption in November/December was high, but the sales dropped significantly after that. Need to study further what happened in 1982.Need to study further what happened in 1982.
- 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 $\alpha = 0.1$, $\beta = 0.2$ and $\gamma = 0.2$ 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 Rose with cheaper price and special designed bottles to increase the sales.

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