Business Report

Time Series Forecasting

(Rose Wine Analysis)

Arnab Ghosal

19 September 2021

Executive Summary

For this assignment, the data of Rose wine sales in the 20th century is to be analyzed. As an analyst in the ABC Estate Wines, you are tasked to analyze and forecast Wine Sales in the 20th century. The dataset consists of the Rose wine sale. In this problem statement we will explore the different sale of the wine based on different months in different year. Depending on this data analysis we need to forecast the future sales.

Introduction

Forecast is a statistical method to predict an attribute using historical patterns in the data. Every business and organization apply different methods of forecasting in different situations. Therefore, it is imperative to identify what to forecast and which method of forecasting should be utilized in different scenarios so that the risk of forecasting is minimized.

The purpose of this whole exercise is to explore the dataset and various forecasting methods on this data, to understand which forecasting method is more suitable in this scenario. This assignment should help in exploring the time series, its components i.e., trend, seasonality & its effects on forecasting the sales in future.

Data Description:

1. Year Month: It represent the month and year of sale.
2. Rose: It represent the Rose wine sale in the corresponding year & it's month. The data is from year 1980 Jan to 1995 July.

Sample of the dataset:

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

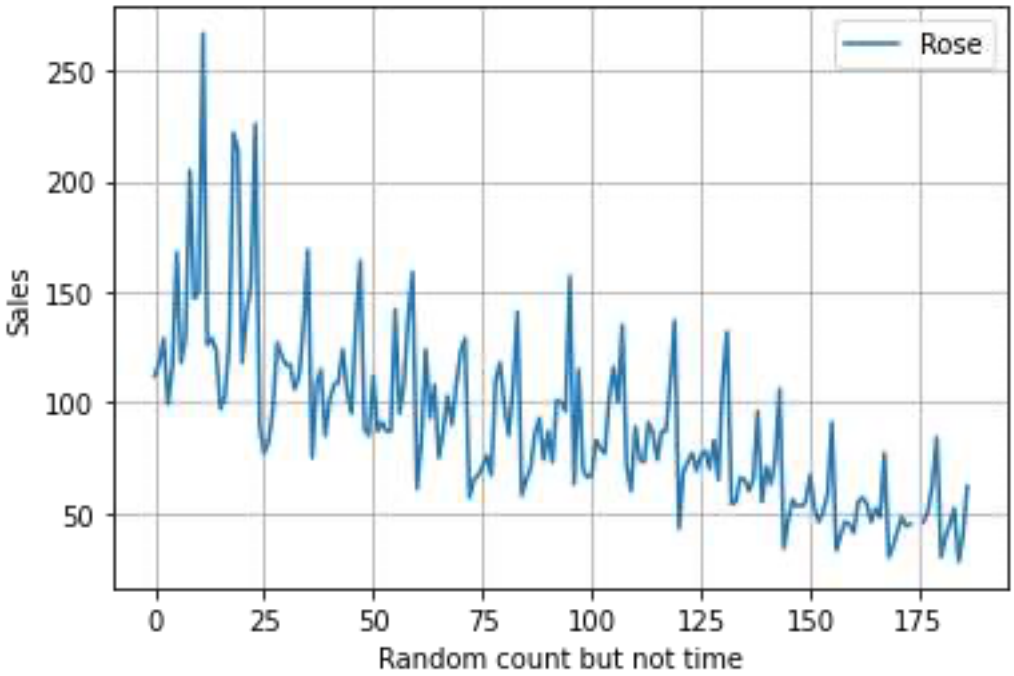
Q1. Read the data as an appropriate Time Series data and plot the data.

The Y axis represent the Sale of the Rose Wine

Though the above plot looks like a Time Series plot, notice that the X-Axis is not time.

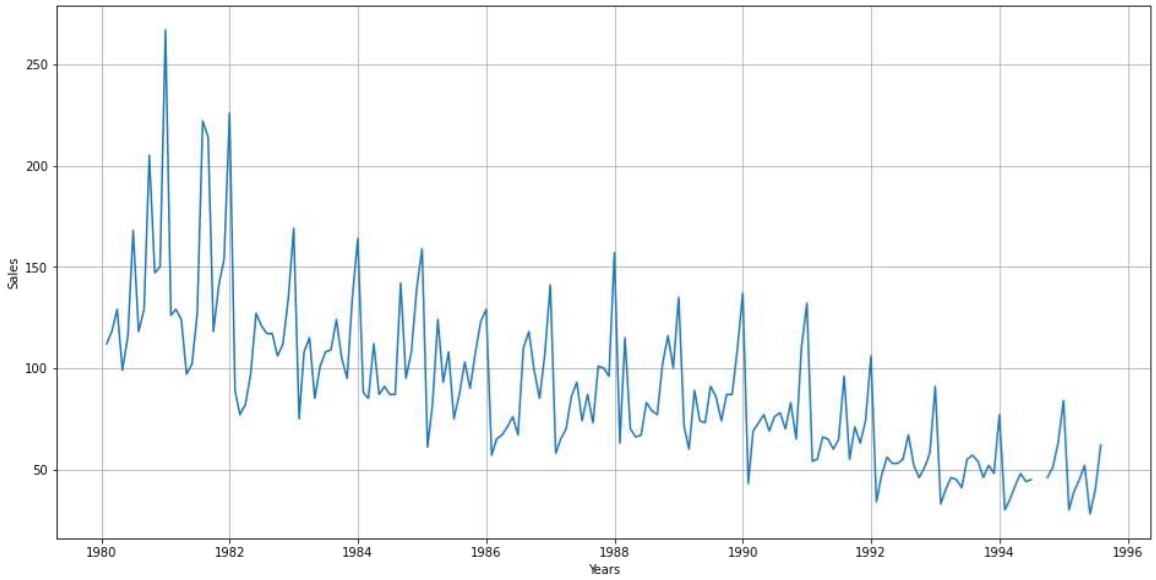
To make the X-Axis as a Time Series we need to pass the date range manually and add a column name Time stamp in the dataframe.

The plot of the data from the given csv file appears in Figure 1.
But we need to convert this data into Time series object to get a proper plot.



	YearMonth	Rose	Time_Stamp
0	1980-01	112.0	1980-01-31
1	1980-02	118.0	1980-02-29
2	1980-03	129.0	1980-03-31
3	1980-04	99.0	1980-04-30
4	1980-05	116.0	1980-05-31

Now we can plot the proper Time series Plot which will have Year's value on X axis & Sales value on Y axis.



From the above figure we can observe that there is decreasing/ negative trend in the dataset, seasonality can also be observed from the above graph.

Q2. Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition.

Exploratory Data Analysis:

1. Shape of the dataset:

The dataset has 187 rows and 2 columns.

2. Dataset Information:

```
df.shape
```

```
(187, 2)
```

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

The dataset has 2 column name YearMonth and Rose respectively.

The YearMonth has data type as object. The Rose has datatype as int64. The Rose column as only 185 non-null count that means 2 counts are null. We need to treat the missing values.

3. Null Value Check:

There are 2 null values in the dataset in the column Rose.

```
df.isnull().sum()
```

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

4. Removing Null Values from the dataset:

For this we use interpolate linear method.

```
df['Rose'].interpolate(method='linear', inplace=True)
```

```
df.isnull().sum()
```

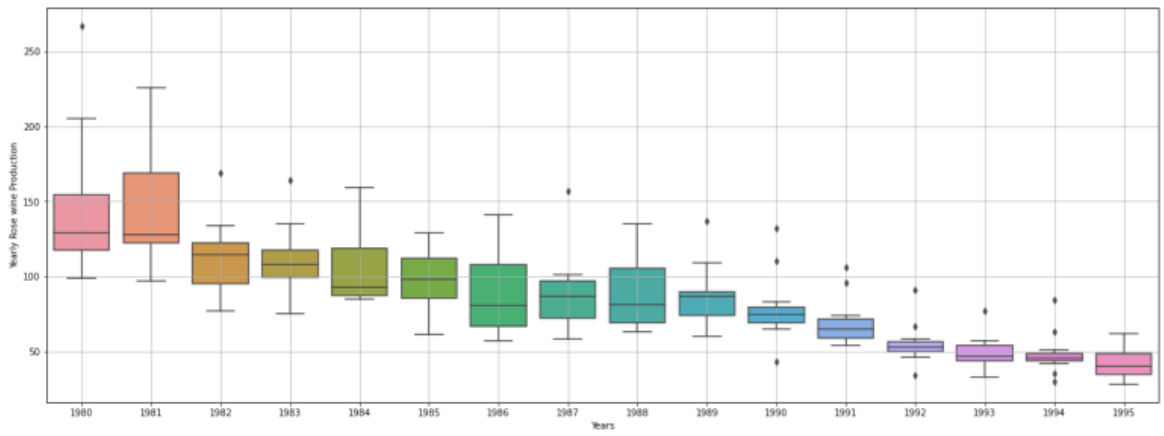
```
YearMonth    0
Rose         0
dtype: int64
```

5. Descriptive Statistics:

From year 1980 Jan to 1995 July
On an average the sale of Rose wine is 89.914439.
The 25% sale of Rose wine is 62.5.
The 50% sale of Rose wine is 85.0.
The 75% sale of Rose wine is 111.0.
The min sale of Rose is 28.0.
The max sale of Rose is 267.0.
As the max value is too large than the 75% so outlier at higher side is present in the data. Similarly, the min value is too small than the 25% so outlier at lower side is present in the data.

	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.914439	39.238325	28.0	62.5	85.0	111.0	267.0

Yearly Plot



We can see in Figure 3 as the year pass on the Sale of the Rose wine drops continuously. Among 1980 to 1995 year the highest sale of the Rose wine

was in the Year 1981. Among 1980 to 1995 year the lowest sale of the Rose wine was in the Year 1994.

Monthly Plot

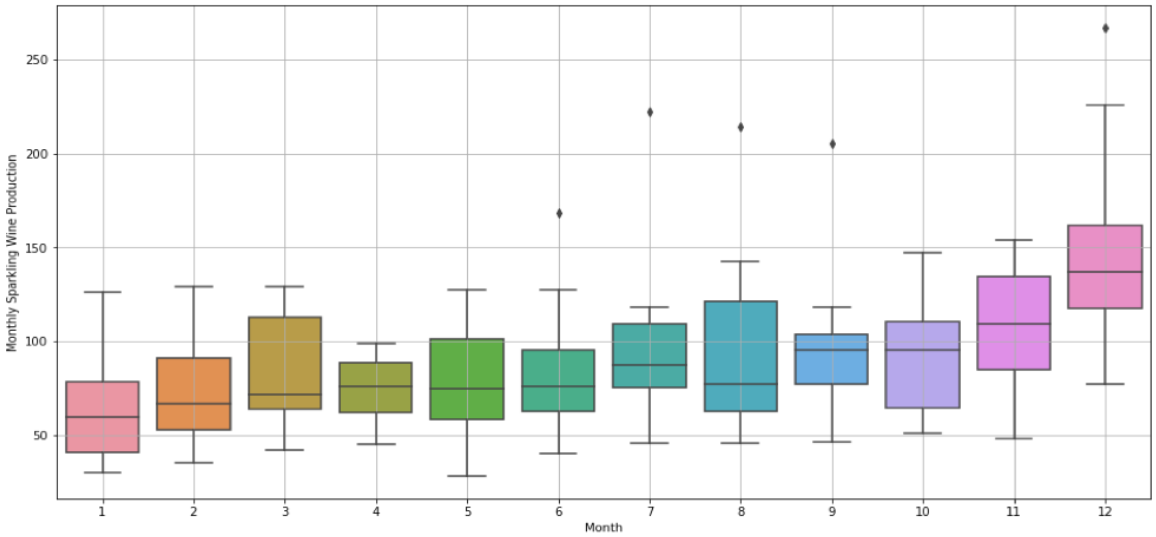


Figure 4 Month Plot (Box Plot)

For all the year from 1980 to 1995 the sale in December month is Highest as compared to rest of the months. This shows that there is seasonality in the data. As in every December the Sale gets increase.

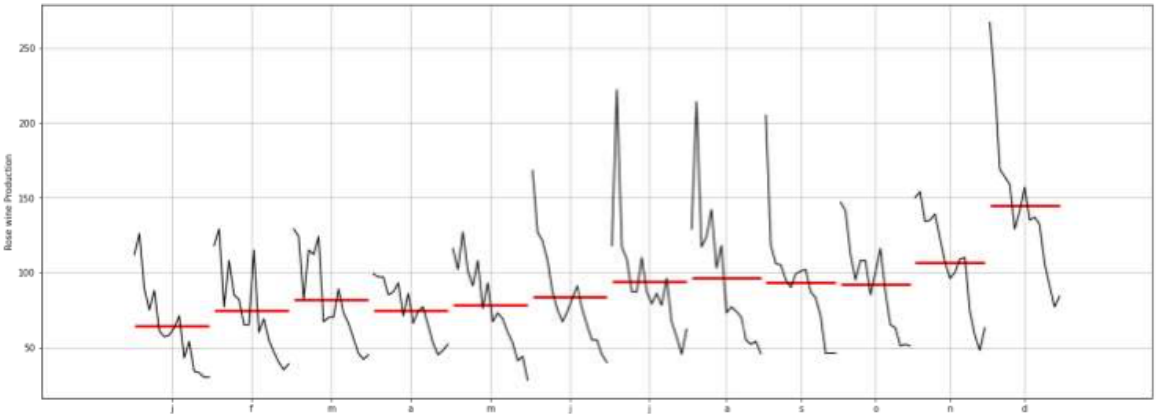


Figure 5 Month Plot with Average Sales

The vertical lines represent monthly sales, and the horizontal lines represent average sales of the given month.

- Here, it can be observed that average sales are higher in December as compared to other months.

- Average sales are lower in January as compared to other months.
- From July to august the average sales are near about same.

Empirical Cumulative Distribution

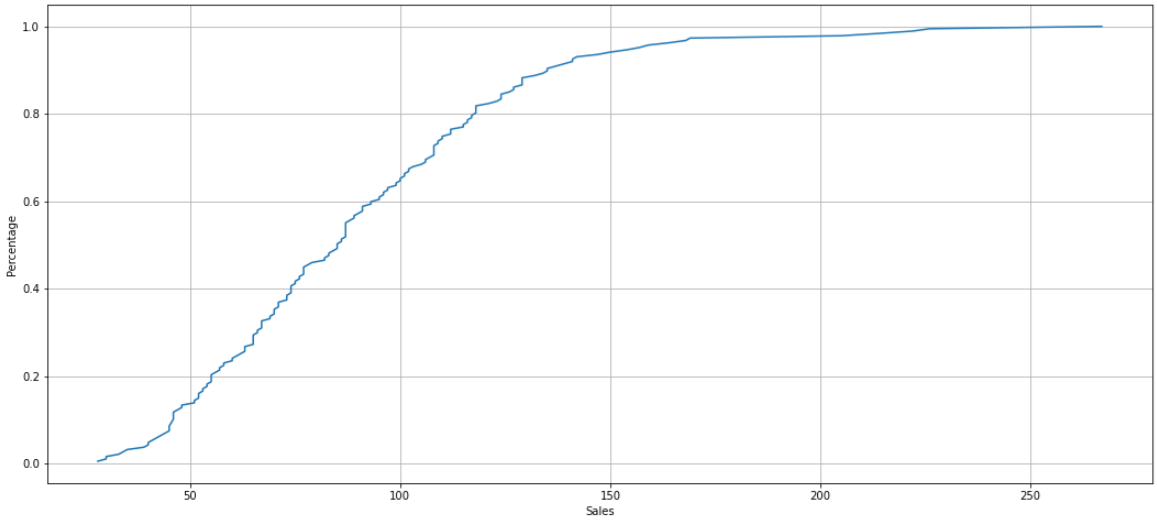


Figure 6 Empirical Cumulative Distribution Plot

- This plot y axis shows the percentage of sale.
- Among the total sale from year 1980 to 1985, 90% of total sale are under 150.
- Sale of 200 to 250 is of only 10% of total sale.
- From the Figure 7 average Sale we came to known that the trend of the time series is decreasing.

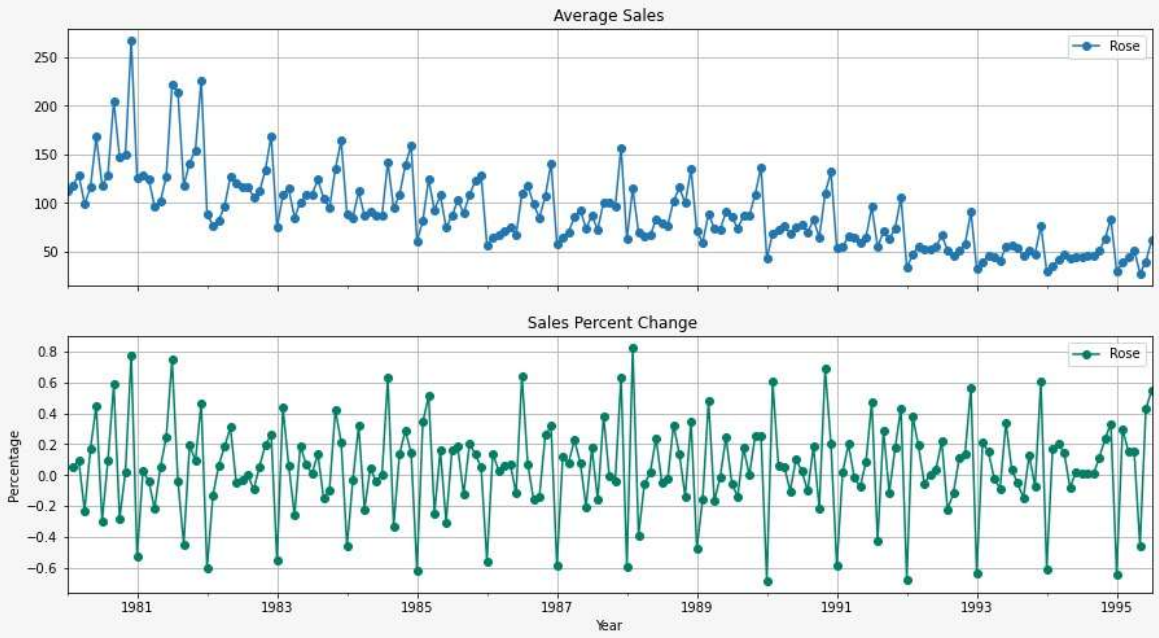


Figure 7 Percent Change in Sales

- The Figure 7 percent change graph is flat over the whole series, this shows that the amount of the change is constant over whole series.
- This graph shows that the Dec month has highest sale among all the month.

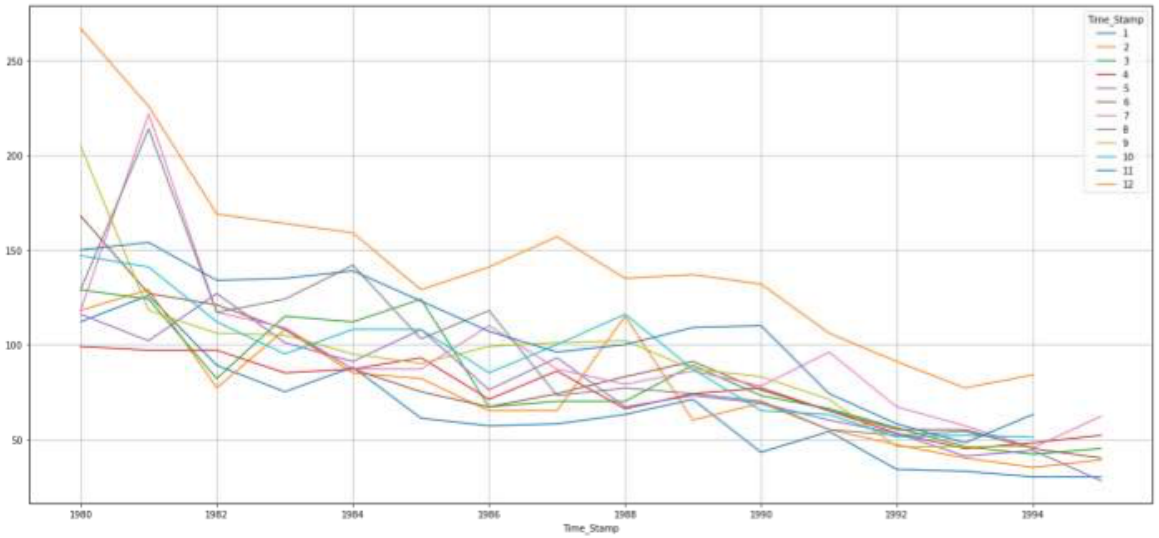


Figure 8 Yearly sales across months

- The July month sale suddenly increase after 1980 and decreases also suddenly.
- Similarly in all the months after 1980 the sale decreases, it increases for small duration but again
  drop out, and continuously in this pattern the sales are decrease till

1995.

Decomposition of Time Series

The time series have three components.

1. Trend (Long term movement)
2. Seasonal component: Intra-year stable fluctuations repeatable over the entire length of the series
3. Irregular component (Random movements)/Residuals/Noise.
   In decomposition of the time series, we came to known clearly on a broader platform regarding these three components.

- From the Decomposition plot we can clearly see that the Rose wine sales data do not have trend component in it. As we can see that there is no continuous increasing, decreasing behaviour in the data.

- There is Multiplicative Seasonality in the data.

Additive Model

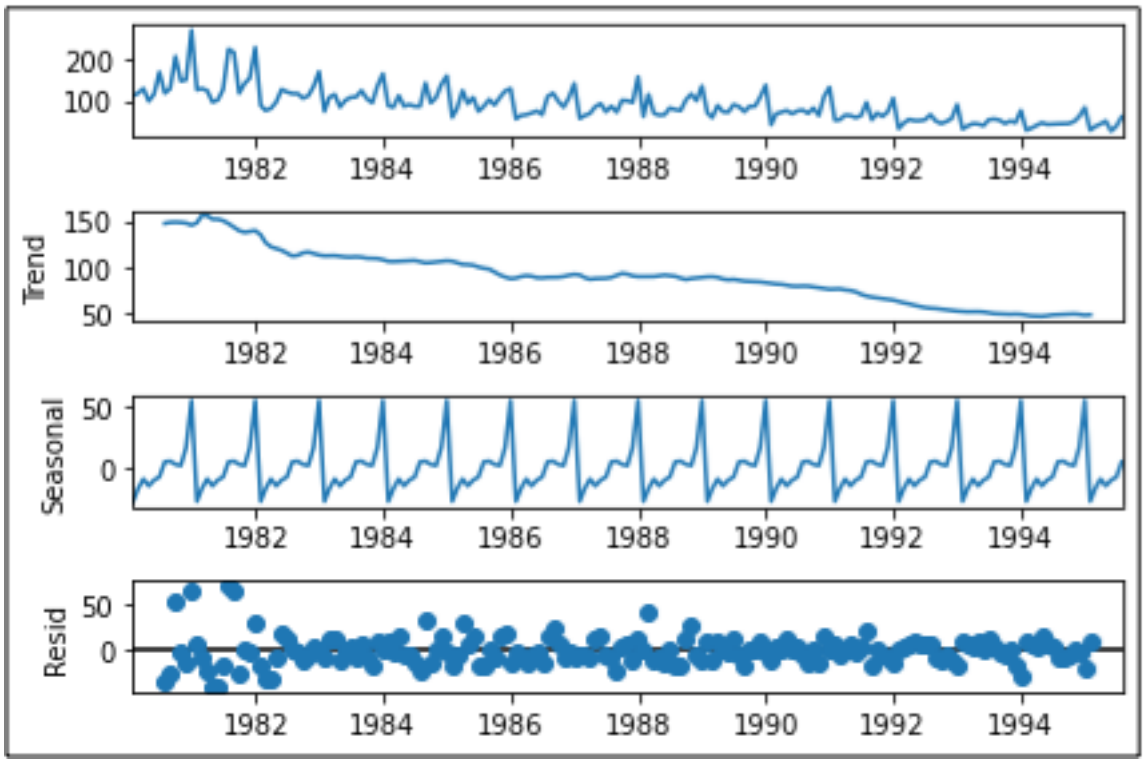


Figure 9 Additive Decomposition Modell

For Additive Model (Figure 9) we see that the residuals are located around 0 from the plot of the
residuals in the decomposition. But they are very much disperse not totally concentrated on 0

Multiplicative Model

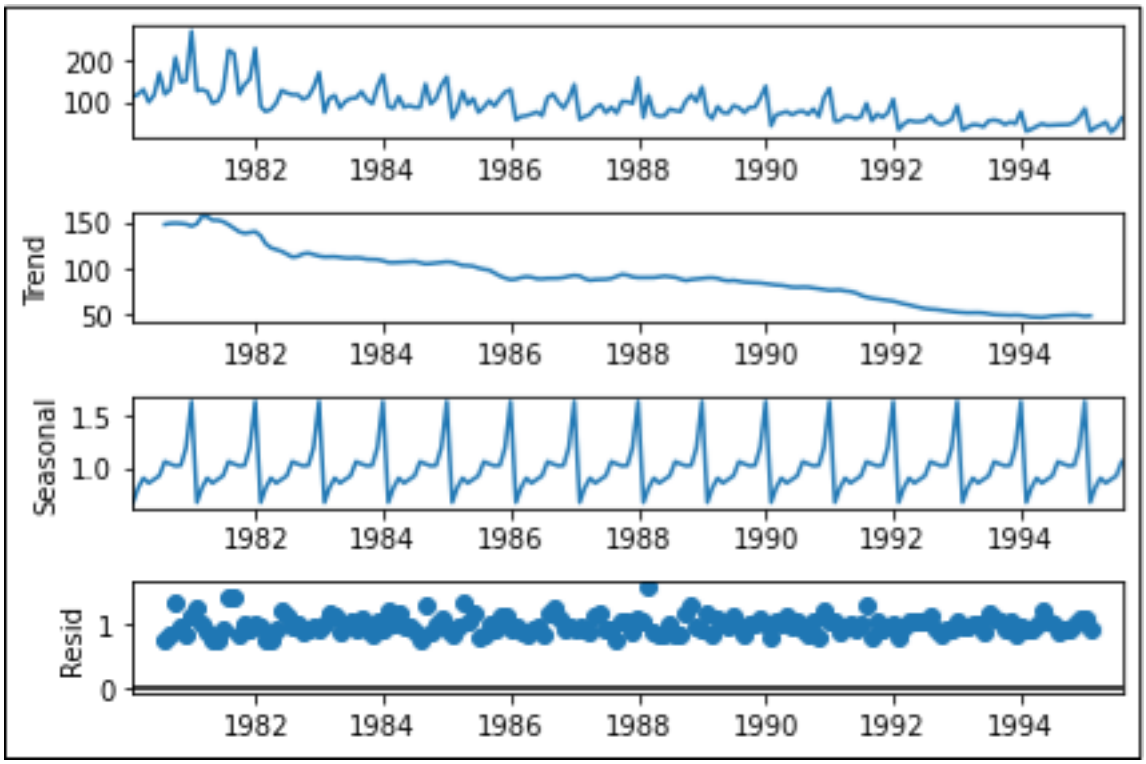


Figure 10 Multiplicative Decomposition Model

(Figure 10) For the multiplicative series, we can see that lot of residuals are located around 1. So here we can consider the seasonality in the dataset is Multiplicative seasonality.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0.67	0.81	0.9	0.85	0.89	0.92	1.06	1.04	1.02	1.02	1.19	1.63

Table 3 Seasonal Indices

Since this is monthly data, there are 12 seasonal indices.

In December Rose wine sales is the highest among all months in the same year, as the highest value of the seasonal component whereas in January (lowest value of the seasonality) sales is the lowest.

Q3. Split the data into training and test. The test data should start in 1991.

Before a forecast method is proposed, the method needs to be validated. For that purpose, data must be split into two sets i.e., training and testing. Training data helps in identifying and fitting right model(s) and test data is used to validate the same.

Testing Set starting few points and end few.

In case of time series data, the test data is the most recent part of the series so that the ordering in the data is preserved.

- As asked in the question for Rose wine sale series, the first 10 years of data from 1980 to 1990 is used for training purpose and last 5 years of data from 1991 to 1995 is used for testing purpose.

Training Dataset (1980 to 1990)

Training Set starting few points and end few points.

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

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

Testing Dataset (1991 to 1995)

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

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

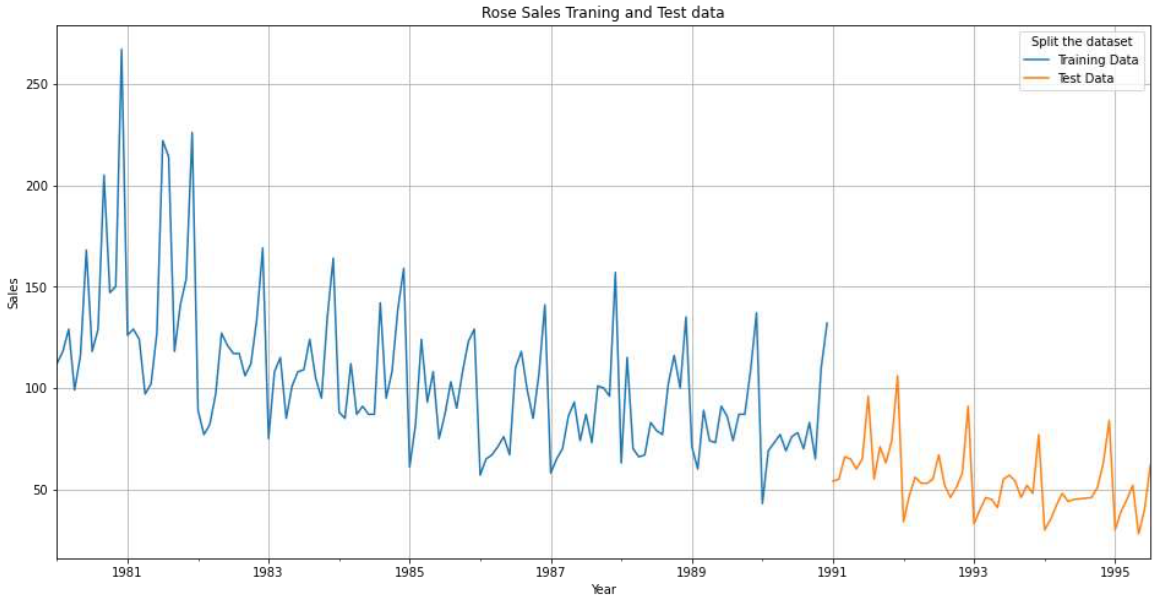


Figure 11 Plot of Training and test data division

So, from the graph Figure 11 we can see that the training and testing data are successfully splitted as per the requirement.

Shape of the Train & Test dataset

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

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

Out of 181 Rows after splitting data into train & test dataset. Now train dataset has 132 rows and test dataset have 55 rows.

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

For Forecasting using Decomposition approach there are various models which can be used.

List of the models are:

1. **Linear Regression**
2. **Naive Forecast**
3. **Simple Average**
4. **Moving Average**
5. **Simple smoothing method**
6. **Holt's Method (Double Exponential Smoothing)**
7. **Holt-Winter's method (Triple Exponential Smoothing)**

Performance Evaluation of the Model:

Forecasting accuracy measures compare the predicted values against the observed values to quantify the predictive power of the proposed model. Mathematically, it can be defined as

Forecast error et for period t is given by: et = Ŷt - Yt Where

Ŷt = forecast value for time period t

Yt = actual value in time period t

n = No. of observations

Measure of Forecasting error

Root Mean Square Error (RMSE): $RMSE$=

Lower the RMSE value better is the model.

For Forecasting using Decomposition approach there are various models which can be used. List of the models are:

1. **Linear Regression**
2. **Naive Forecast**
3. **Simple Average**
4. **4. Moving Average**

$$\sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

1. Linear Regression

The linear regression model is built on the train dataset and forecasting is done on the test dataset.

The Forecast line does not follow the trend or seasonality in the test data. This forecast is not useful does not work well.

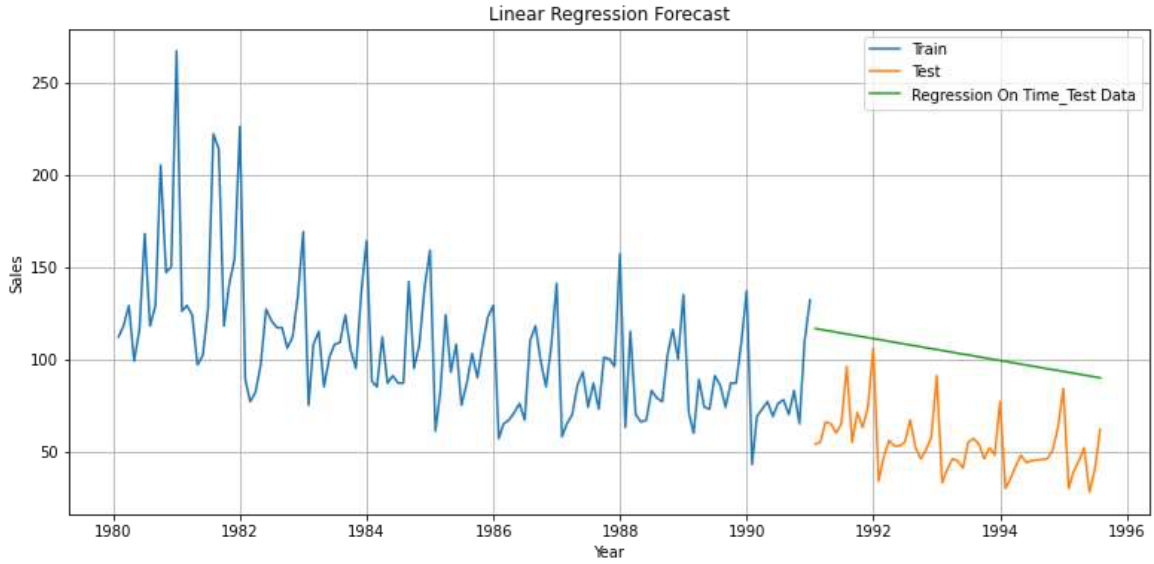


Figure 12 Linear Regression Forecasting Plot

Model Evaluation:

Model Evaluation by calculating the RMSE value.
- RMSE value for Linear Regression is **51.433312.**

2. Naive Forecast

Naïve Forecast uses the last observed value for forecasting.
The Naive model is built on the train dataset and forecasting is done on the test dataset. The Forecast line does not follow the trend or seasonality in the test data.
This forecast is not useful does not work well.

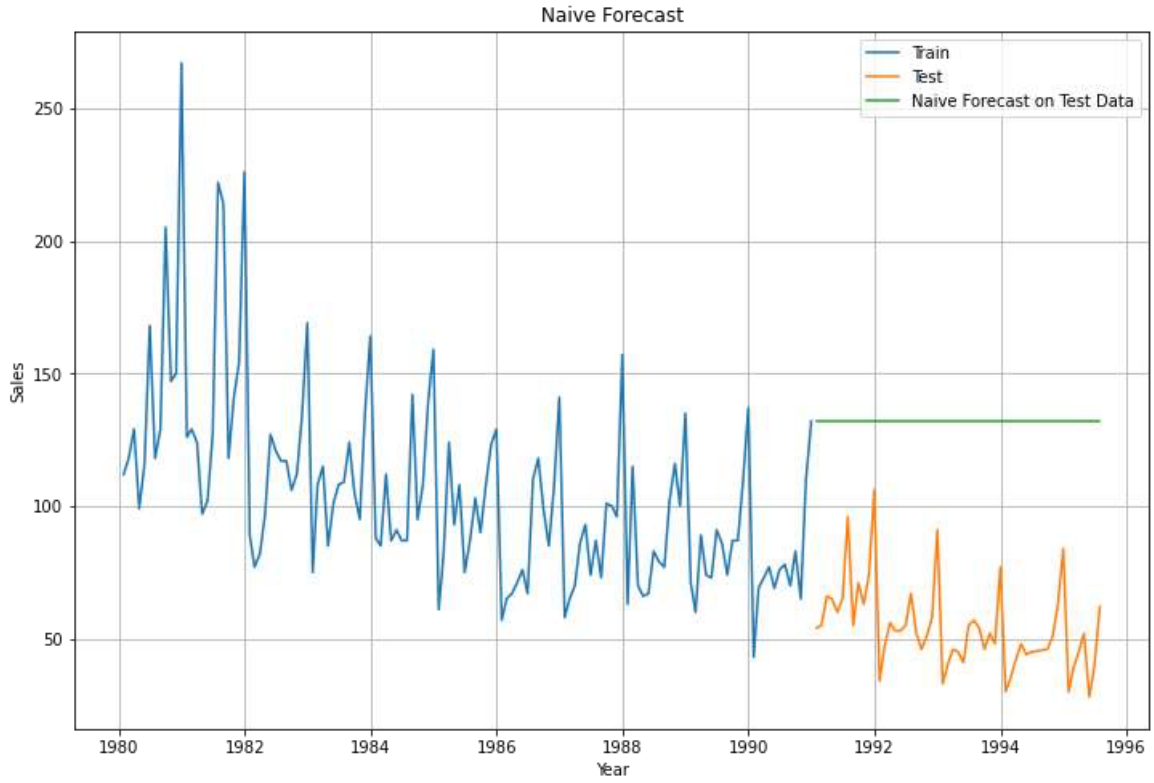


Figure 13 Naive Model Forecasting Plot

Model Evaluation:

Model Evaluation by calculating the RMSE value.
- RMSE value for Naïve Model is **79.718773.**

3. Simple Average

The Simple Average model is built on the train dataset and forecasting is done on the test dataset. The Forecast line does not follow the trend or seasonality in the test data.
This forecast is not useful.

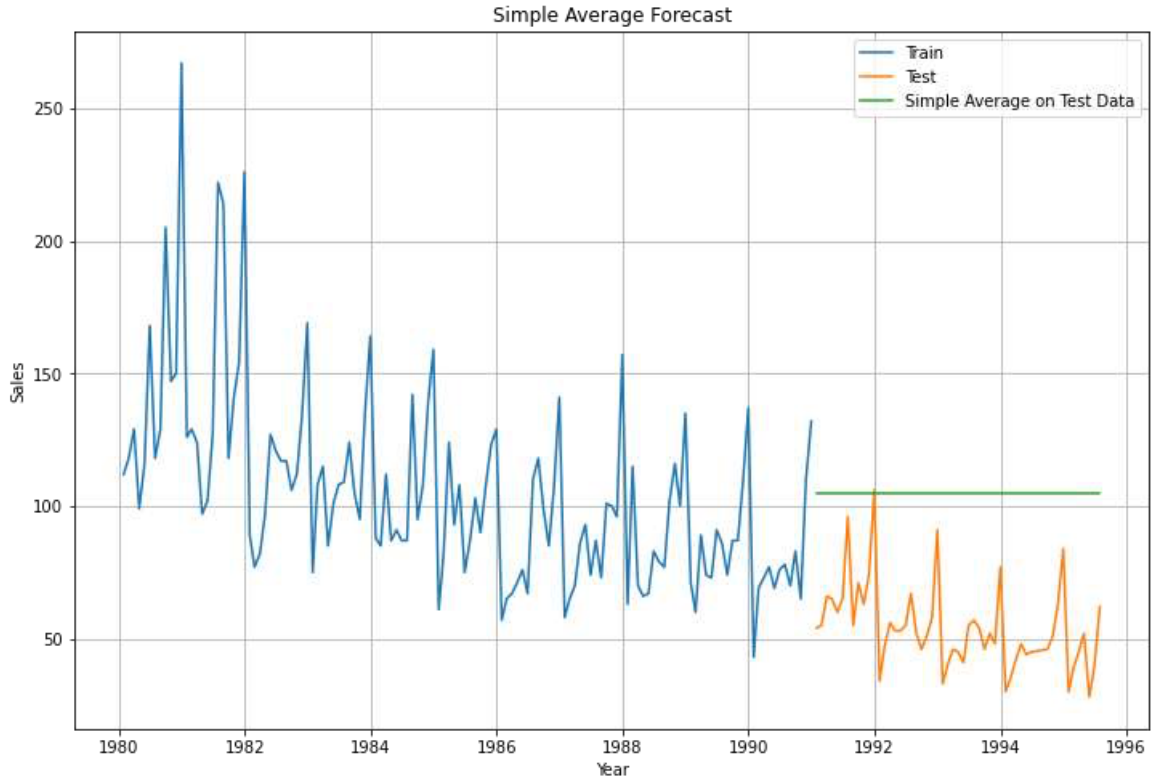


Figure 14 simple Average Forecasting Plot

Model Evaluation:

Model Evaluation by calculating the RMSE value.

- RMSE value for Simple Average Model is **53.460570.**

4. Moving Average

For forecasting take average over a window of certain width & move the window.

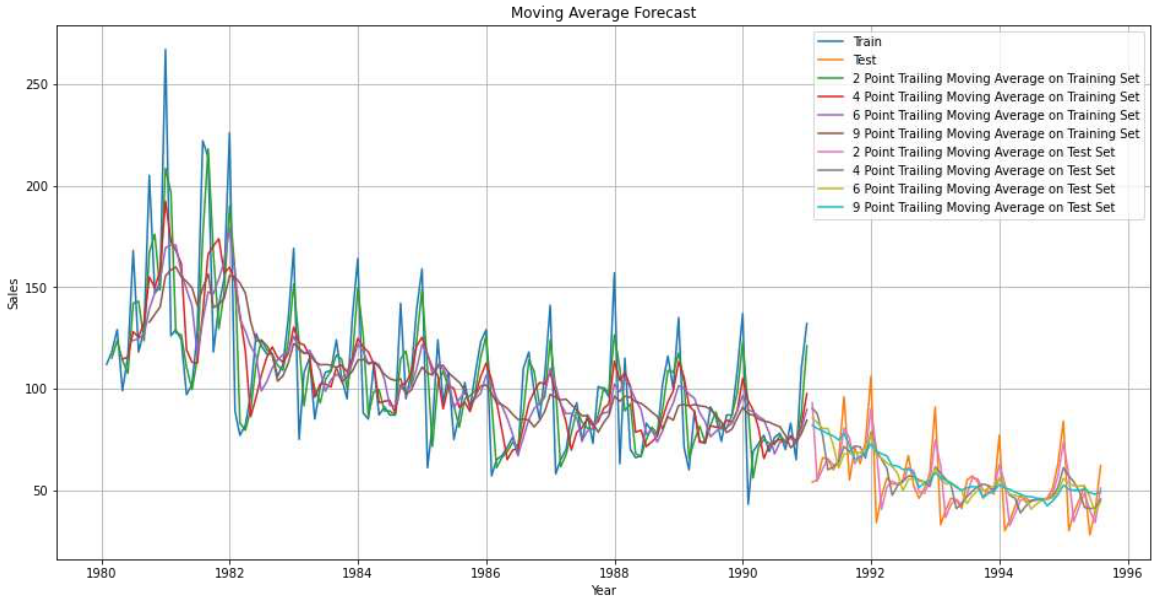


Figure 15 Moving average Forecasting Plot.

Model Evaluation:

Model Evaluation by calculating the RMSE value.

- RMSE value for 2point Training Moving Average is **11.529278.**
- RMSE value for 4point Training Moving Average is **14.451403.**
- RMSE value for 6point Training Moving Average is **14.566327.**
- RMSE value for 9point Training Moving Average is **14.727630.**
  2-point Training Moving average model is best from the all the moving average points model.

5. Simple Exponential Smoothing

SES or one-parameter exponential smoothing is applicable to time series which do not contain either of trend or seasonality. Forecast by SES is given by:

$\hat{Y}t+1 = \alpha Yt+\alpha(1-\alpha)\ Yt-1+\alpha(1-\alpha)2Yt-1+\cdots,\ 0<\alpha<1$

where, α is the smoothing parameter for the level. Such a series is hard to find. This is a one-step- ahead forecast where all the forecast values are identical.

We do the Simple Exponential modelling in two ways.

- By setting the default values of the parameters.
- By taking the best value of the parameter from range 0.3 to 1

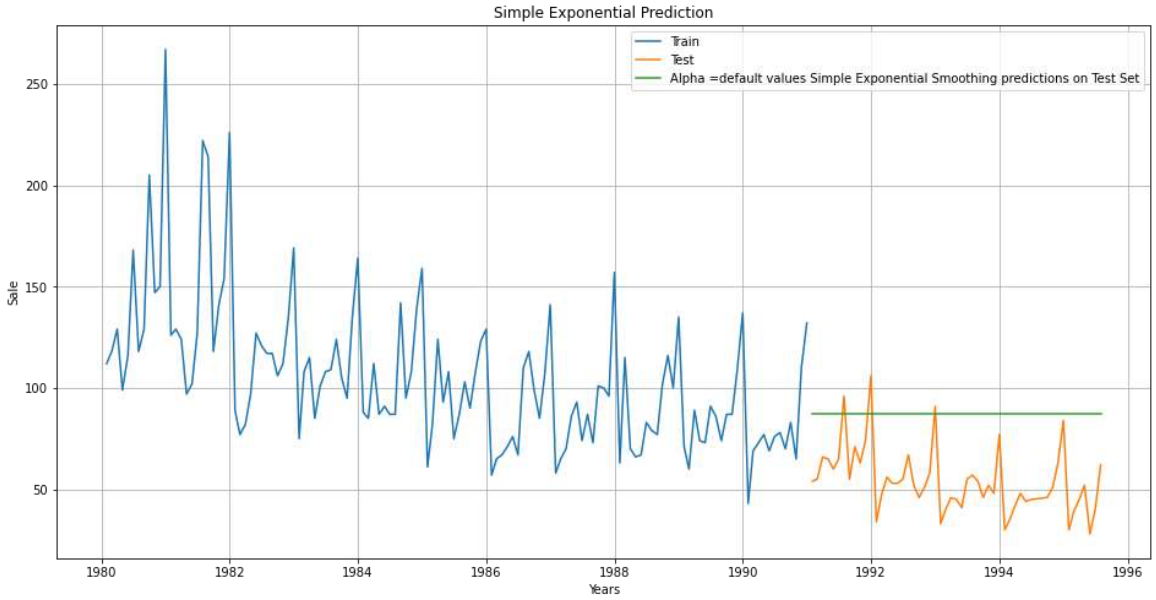


Figure 16 Simple Exponential Smoothening Forecasting Plot

We apply the Simple Exponential Smoothing at the default alpha value and at alpha = 0.3.

Model Evaluation:

Model Evaluation by calculating the RMSE value.

- RMSE value for SES at default alpha value is **36.796241.**

- RMSE value for SES at alpha value = 0.3 is **47.504821.**
  From the above two the result for default alpha value is better than the alpha value 0.3.
  In Simple Exponential Smoothing determines the level parameter only which is indicated in the graph above i.e., it forecast only the level parameter. But no trend and seasonality are observed in forecasting, so for our case this simple Exponential Smoothing is not useful.

6. Double Holt's Method (Double Exponential Smoothing)

This method is an extension of SES method, proposed by Holt in 1957. 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.
But in our data trend is not present, so this modelling type will not appropriate for our case. We do the Double Exponential modelling in two ways.

- By setting the default values of the parameters.

- By taking the best value of the parameter from range 0.3 to 1.

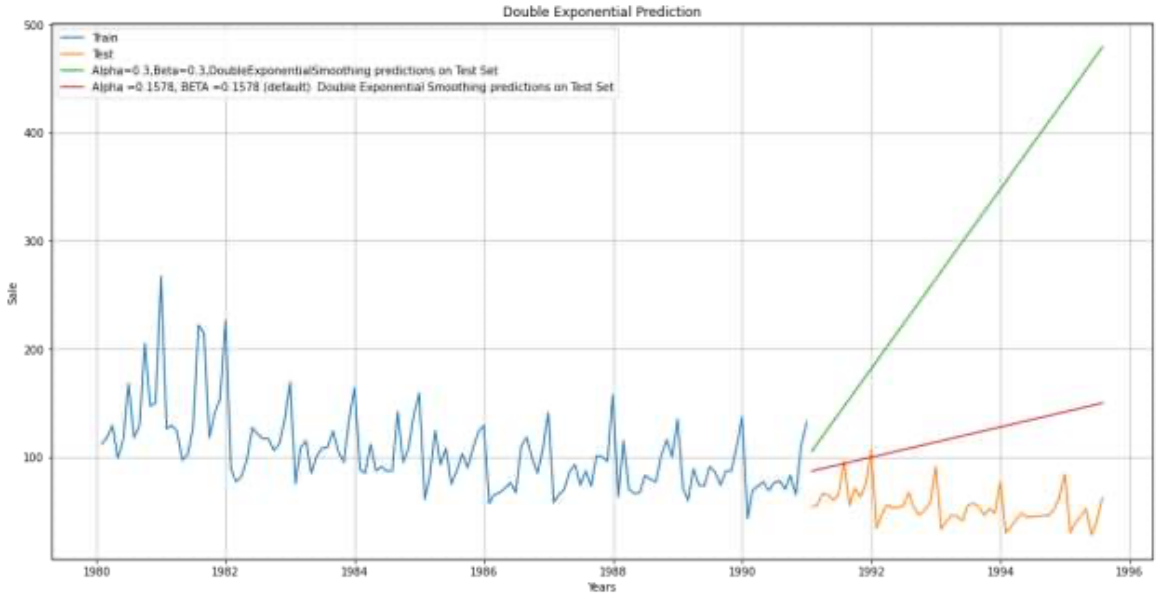


Figure 17 Double Exponential Smoothening Forecasting Plot

We apply the DES for the default alpha & beta value, for alpha 0.3, Beta=0.3

Model Evaluation:

Model Evaluation by calculating the RMSE value.

- RMSE value for SES at default alpha value is **70.572452.**

- RMSE value for SES at alpha & beta value = 0.3 is **265.567594.**
  From the above two the result for alpha & beta default value is better than the default values.
  In Double Exponential Smoothing determines the level & trend parameter only which is indicated in the graph above i.e., it forecast only the level & trend parameter. But no seasonality is observed in forecasting, so for our case this Double Exponential Smoothing is not useful.

7. Holt-Winter's method (Triple Exponential Smoothing)

This is an extension of Holt's method where along with trend seasonality is also 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. But in our data trend is not present, but seasonality is there so let's implement and check the TES

model performance in this case.
We do the Triple Exponential modelling in two ways.

- By setting the default values of the parameters.
- By taking the best value of the parameter from range 0.3 to 1.

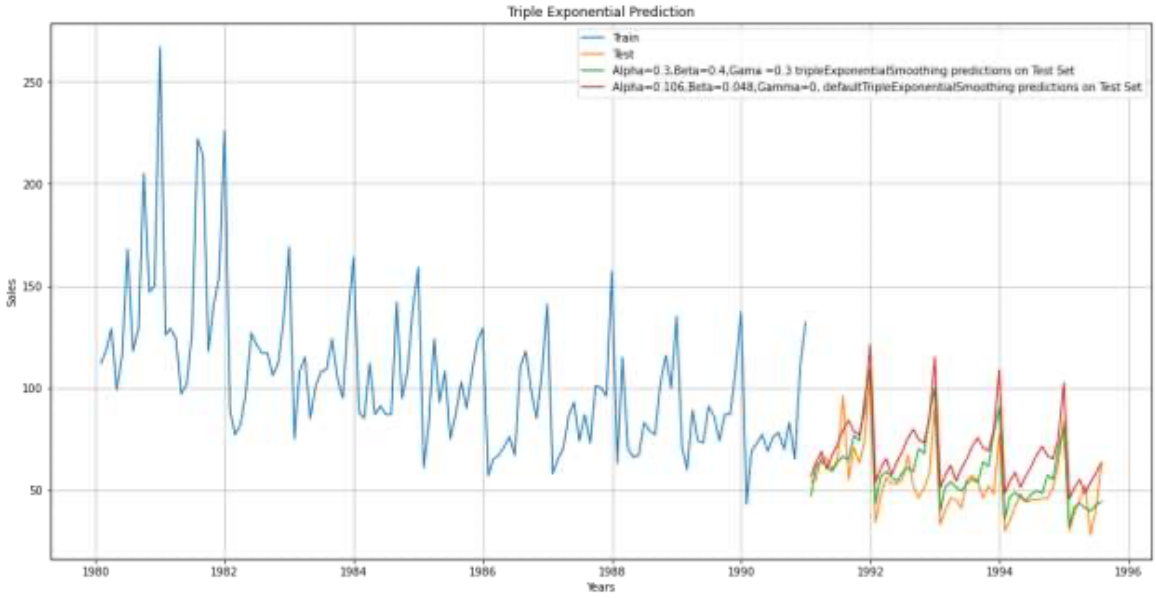


Figure 18 Triple Exponential Smoothening Forecasting Plot

We apply the TES at the default alpha & beta value and at alpha 0.3, Beta=0.3, Gamma = 0.3 Model Evaluation by calculating the RMSE value.

- RMSE value for TES at default alpha value is **17.369489.**
- RMSE value for TES at alpha & gamma value = 0.3 & beta= 0.4 is **10.945435.**
  From the above two the result RMSE value for TES at alpha & gamma value = 0.3 & beta= 0.4 for values is better than default.
  As compared to another exponential model Triple Exponential model has forecasted quite well, as we can see this from graph above.

Sorting of the RMSE values of all the models

	Test RMSE
Alpha=0.3,Beta=0.4,Gamma=0.3,TripleExponentialSmoothing	10.945435
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
Alpha=0.106,Beta=0.048,Gamma=0, default TripleExponentialSmoothing	17.369489
Alpha=(default),SimpleExponentialSmoothing	36.796241
Alpha=0.3,SimpleExponentialSmoothing	47.504821
RegressionOnTime	51.433312
SimpleAverageModel	53.460570
Alpha=0.1578,Beta=0.1578,default DoubleExponentialSmoothing	70.572452
NaiveModel	79.718773
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.567594

Table 4 Sorting of RMSE values for different models

Among all the values the TES (default values) model RMSE is less, so this model is best till now.

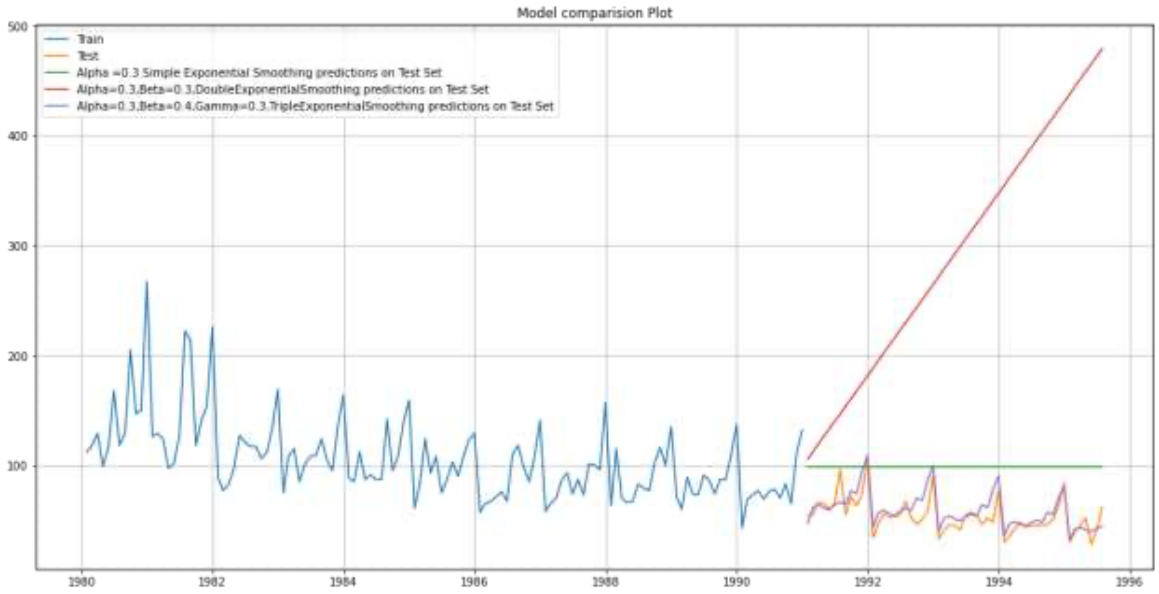


Figure 19 All model Comparison Plot

Among all the Model we can see that the Triple Exponential model forecast is closer to the test data while other models are directly showing the tangent line which is not at a relevant forecast.

Q.5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and 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.

- To build the ARIMA/SARIMA model the time series must be Stationary.
- So first we need to check whether the series is stationary or not.
- If not, then make the series stationary to apply the ARIMA/SARIMA model on time series
  for forecasting.
- Since ARIMA model requires a stationary series, a formal stationarity test needs to be applied to the time series under consideration.
  • **Augmented Dickey-Fuller Test:** A formal test to check whether time series data follows stationary process.

HO: Time series is non-stationary.
H1: Time series is stationary.
If p>0.05 we fail to reject the Null Hypothesis i.e., our Time Series will be non-Stationary. If p<0.05 we fail to reject the Alternate Hypothesis i.e., our Time Series will be Stationary.

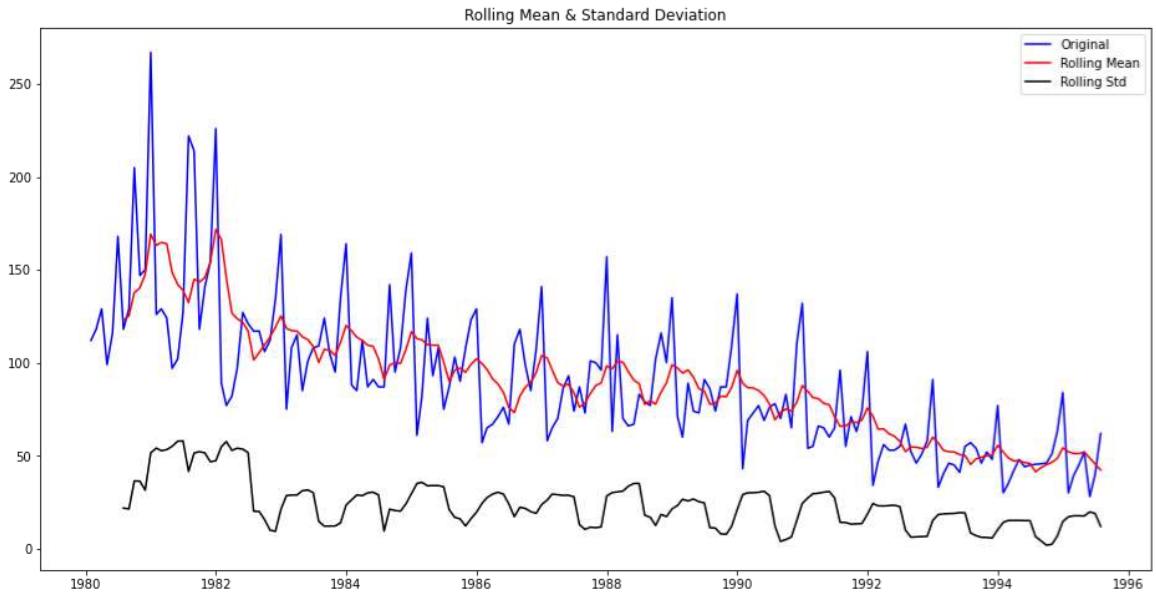


Figure 20 Time Series Stationarity Check on whole data

From the graph we can see that the mean and standard deviation are not constant, so we can predict that the series is non-stationary.
But to be more accurate we will execute Augmented Dickey-Fuller Test, this test gives us the p value. This p value tells us whether the series is stationary or not.

Auto Regression equation is $X_t = \Phi X_{t-1} + \varepsilon_t$

$X_t = X_{t-1} + \varepsilon_t$ (if $\Phi=1$)
$X_t - X_{t-1} = \varepsilon_t$ (Stationary Series)

So, for Time Series to be Stationary the Φ must be 1, the ADF test finds what is the probability

```
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 21 ADF Test Report for whole data

that the Φ is 1. This probability is nothing but the p-value.

From the above Result we see that the p value is 0.343101 which is greater than 0.05. So, the Time Series is not stationary.
Often differencing a non-stationary time series leads to a stationary series.
So just differencing the series by 1 the plot is as below:

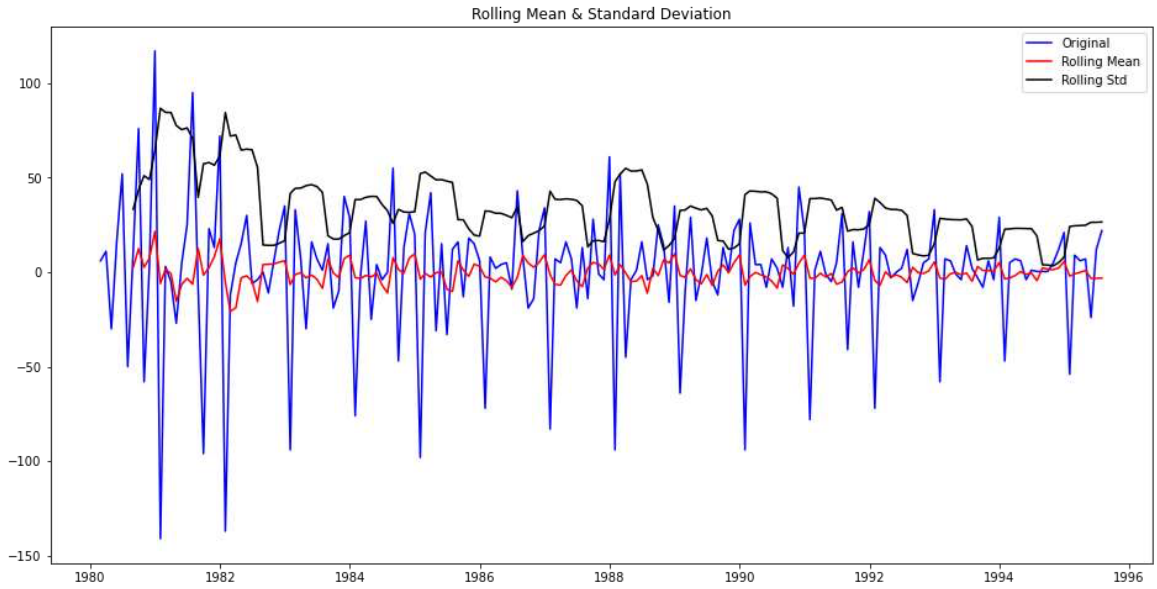


Figure 22 Time Series Stationarity Check after differentiation by 1

The mean and std dev of the series is constant all over the series, this shows that the series is now stationary.

The ADF test Result is

Now the p-value is less than 0.05, so the series is now become Stationary.

Now we must split the data into the train and test for the further modelling process, so we need to check the stationary of the train time series.

Train time series Stationary Check

```
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 23 ADF Test report after differentiation by 1

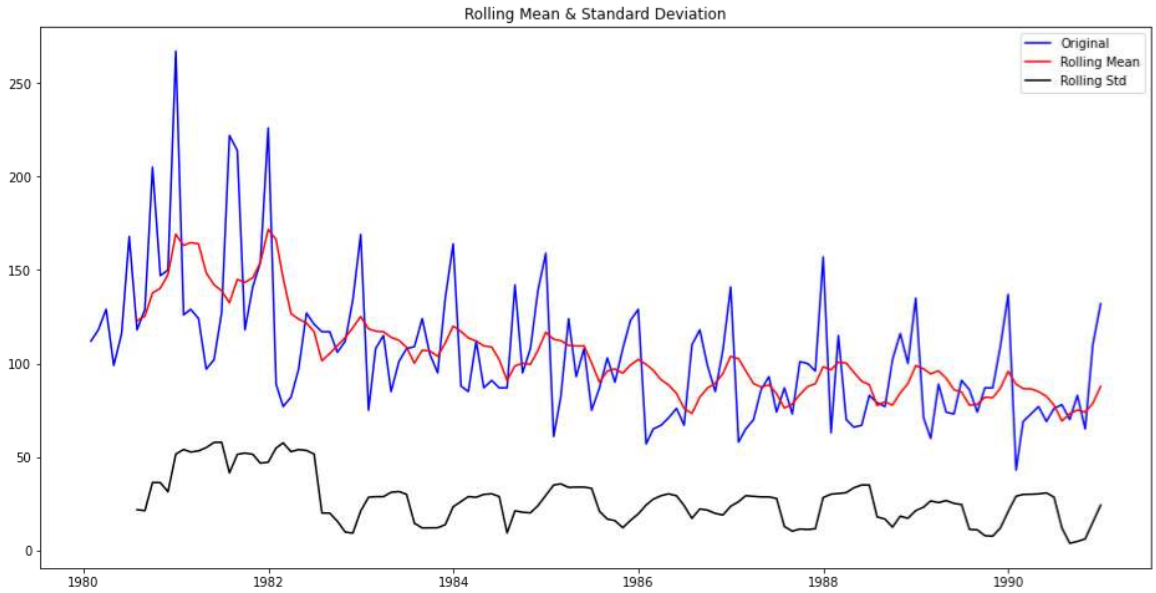


Figure 24 Time Series Stationary Check on Train dataset

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 25 ADF test report on Train dataset

From the above Result we see that the p value is 0.219476 which is greater than 0.05. So, the Time Series is not stationary.
Often differencing a non-stationary time series leads to a stationary series.
So just differencing the series by 1 the plot is as below:

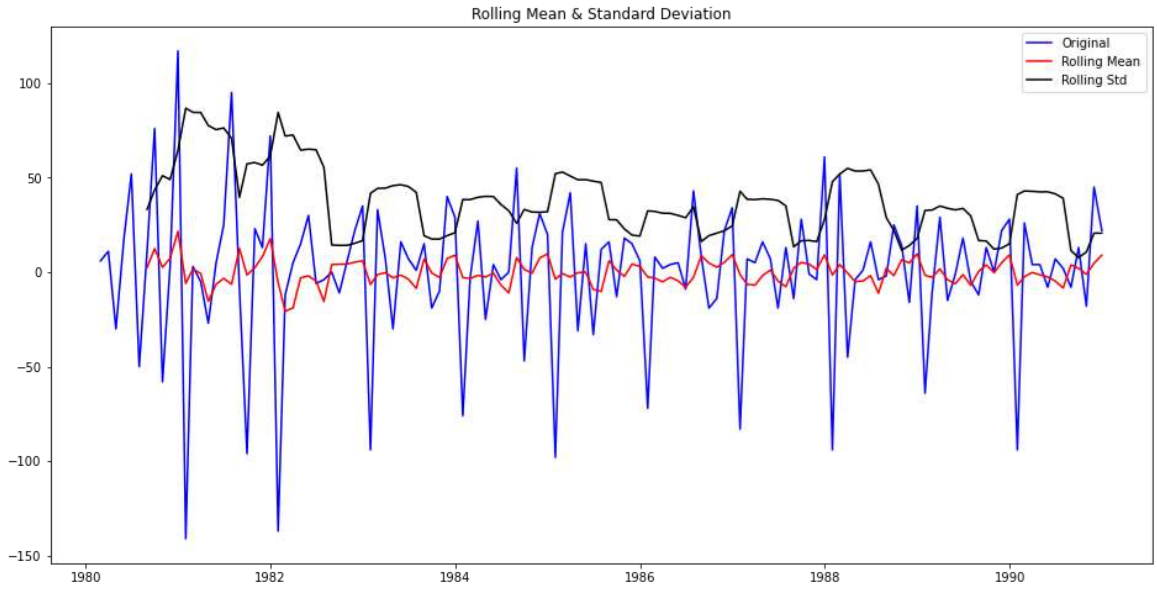


Figure 26 Time Series Stationary Check on Train dataset differentiation by 1

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 27 ADF test report on Train dataset after differentiation by 1

Now the p-value is less than 0.05, so the Train Time series is now become Stationary.

Q.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.

$ARIMA$ **(p, d, q) Model**: ARIMA is defined by 3 parameters. p: No of autoregressive terms

d: No of differencing to stationaries the series

q: No of moving average terms.

For building model, we will consider values as p = q = range (0, 3), d= range (1,2)

So, the Model for all combination of p, d, q will be taken under consideration & the one with least AIC score will be choose as best fit. Some parameter combinations for the Model are as follow:

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)

For our case study Sorted AIC for respective p, d, q values are as follows:

	param	AIC
2	(0, 1, 2)	1276.835372
5	(1, 1, 2)	1277.359224
4	(1, 1, 1)	1277.775753
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

Table 5 Sorted AIC values for ARIMA.

So, the best fit is (0,1,2) with lowest AIC value of 1276.835372.
Now we will use (0,1,2) values as p, d, q value and build the ARIMA model. ARIMA model summary is as follow:

Figure 28 Summary test report of AIC ARIMA model

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value.

RMSE value for ARIMA (0,1,2) is **15.617828821296156.**

S$ARIMA$ (p, d, q) (P, D, Q, F) Model

As we can see while doing the EDA that there is Seasonality in the data. So SARIMA model is specially use for such kind of cases.

- Seasonal ARIMA models are more complex models with seasonal adjustments.
- These models are used when time series data has significant seasonality.
- The most general form of seasonal ARIMA is $ARIMA\ (p, d, q) * ARIMA\ (P, D, Q)$ [m],
- where P, D, Q are defined as seasonal AR component, seasonal difference and seasonal MA component respectively. And 'm' represents the frequency (time interval) at which the data is observed.
- We will build SARIMA model for seasonality 6 & seasonality 12 and check which is best model.
- We will build the SARIMA model by using the least AIC terms as we build for ARIMA.

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(0, 1, 2)	Log Likelihood	-634.418
Method:	css-mle	S.D. of innovations	30.167
Date:	Sun, 23 May 2021	AIC	1276.835
Time:	11:34:23	BIC	1288.336
Sample:	02-29-1980 - 12-31-1990	HQIC	1281.509

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4885	0.085	-5.742	0.000	-0.655	-0.322
ma.L1.D.Rose	-0.7601	0.101	-7.499	0.000	-0.959	-0.561
ma.L2.D.Rose	-0.2398	0.095	-2.518	0.012	-0.427	-0.053

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-4.1695	+0.0000j	4.1695	0.5000

For building model, we will consider values as p = q = range (0, 3), d= range (1,2),
D = range (0,1)
So, the Model for all combination given below will be taken under consideration & the one with least AIC score will be choose as best fit.

SARIMA AIC model for Seasonality 6:

There will be total 81 combination few are listed below:

Model: (0, 1, 1) (0, 0, 1, 6)

Model: (0, 1, 2) (0, 0, 2, 6)

Model: (1, 1, 0) (1, 0, 0, 6)

Model: (1, 1, 1) (1, 0, 1, 6)

Model: (1, 1, 2) (1, 0, 2, 6)

Model: (2, 1, 0) (2, 0, 0, 6)

Model: (2, 1, 1) (2, 0, 1, 6)

Model: (2, 1, 2) (2, 0, 2, 6)

Sorted AIC for respective (p, d, q) (P, D, Q) values are as follows:

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655818
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.288741
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778469

Table 6 Sorted AIC model for Seasonality 6

So, the best fit is (1,1,2) (2,0,2,6) with lowest AIC value of 1041.655818.

Now we will use (1,1,2) values as p, d, q value and (2,0,2,6) values as P, D, Q, F and build the SARIMA model.

SARIMA model summary is as follow:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-512.828
Date:	Sun, 23 May 2021	AIC	1041.656
Time:	11:40:52	BIC	1063.685
Sample:	0	HQIC	1050.598
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5939	0.152	-3.899	0.000	-0.892	-0.295
ma.L1	-0.1954	158.955	-0.001	0.999	-311.742	311.351
ma.L2	-0.8046	127.938	-0.006	0.995	-251.559	249.950
ar.S.L6	-0.0625	0.035	-1.763	0.078	-0.132	0.007
ar.S.L12	0.8451	0.039	21.884	0.000	0.769	0.921
ma.S.L6	0.2225	405.708	0.001	1.000	-794.950	795.395
ma.S.L12	-0.7774	315.452	-0.002	0.998	-619.051	617.496
sigma2	335.2115	1.5e+05	0.002	0.998	-2.95e+05	2.95e+05

Ljung-Box (Q):	15.89	Jarque-Bera (JB):	56.68
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.47	Skew:	0.52
Prob(H) (two-sided):	0.02	Kurtosis:	6.26

Figure 30 Summary test report of AIC SARIMA seasonality 6 model

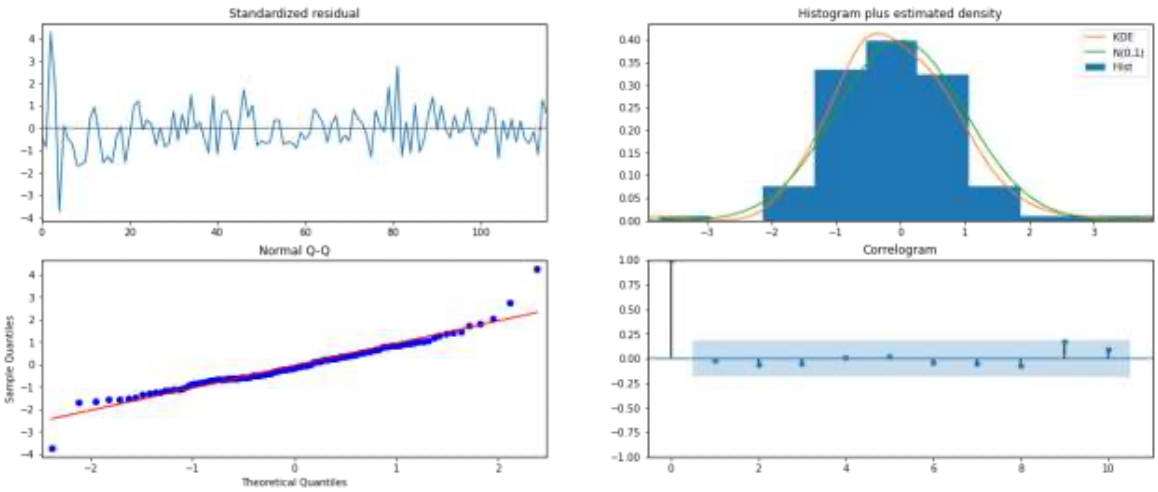


Figure 31 Model Performance check dist plot (SARIMA Seasonality 6)

From the above plot we can see that the Residual are around the zero line.
The histogram is normalized.
The near about all points in the Q-Q plot lies on the line.
We can see there are no significant lags in the series.

So, this model is quite good enough.

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for SARIMA (1,1,2) (2,0,2,6) is **26.13366867588045**

This is less than the ARIMA model. So, we can say performance of AIC SARIMA model with seasonality 6 is better than the AIC ARIMA model.

SARIMA AIC model for Seasonality 12:

There will be total 81 combination few are listed below:

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)

Sorted AIC for respective (p, d, q) (P, D, Q) values are as follows:

So, the best fit is **(1,0,2) (2,0,2,12)** with lowest AIC value of **900.256361**.

Now we will use (1,0,2) values as p, d, q value and (2,0,2,12) values as P, D, Q, F and build the SARIMA model.

	param	seasonal	AIC
53	(1, 0, 2)	(2, 0, 2, 12)	900.256361
80	(2, 0, 2)	(2, 0, 2, 12)	902.352661
78	(2, 0, 2)	(2, 0, 0, 12)	908.895138
70	(2, 0, 1)	(2, 0, 1, 12)	909.003971
26	(0, 0, 2)	(2, 0, 2, 12)	910.624932

Table 7 Sorted AIC model for Seasonality 12

SARIMA model summary is as follow:

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Sun, 23 May 2021   AIC                            887.938
Time:                                    11:46:07   BIC                            906.448
Sample:                                         0   HQIC                           895.437
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8428    174.383     -0.005      0.996    -342.626     340.941
ma.L2         -0.1572     27.446     -0.006      0.995     -53.951      53.637
ar.S.L12       0.3467      0.079      4.374      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       251.3159   4.38e+04      0.006      0.995   -8.56e+04    8.61e+04
===================================================================================
Ljung-Box (Q):                       24.56   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.97   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================
```

Figure 32 Summary test report of AIC SARIMA seasonality 12 model

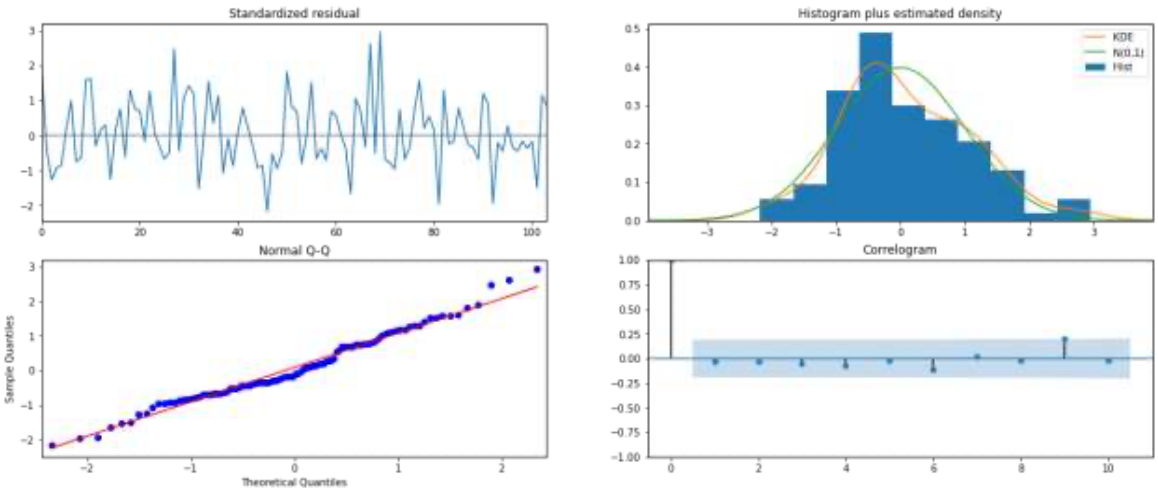


Figure 33 Model Performance check dist-plot (SARIMA Seasonality 12)

From the above plot we can see that the Residual are around the zero line.
The histogram is normalized.
The near about all points in the Q-Q plot lies on the line.
We can see there are no significant lags in the series.

So, this model is quite good enough.

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for SARIMA (1,0,2) (2,0,2,12) is **29.749419253563993.**

This is less than the ARIMA & SARIMA seasonality 6 model. So, we can say performance of AIC SARIMA model with seasonality 12 is better than the AIC ARIMA & SARIMA seasonality 6 model.

Sr No	Model	RMSE
1	AIC ARIMA (0,1,2)	15.617828
2	AIC SARMA (Seasonality 6) (1,1,2) (2,0,2,6)	26.133668
3	AIC SARMA (Seasonality 12) (1,0,2) (2,0,2,12)	29.749419

Table 8 Comparison of AIC models RMSE values

So, among the models build using the lowest Akaike Information Criteria (AIC) ARIMA model is performing best for our case study.

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

ARIMA Model based on the cut-off points of ACF and PACF.

Another way to build the model is by taking the (p, d, q) (P, D, Q) values manually by observing the ACF and PACF plot.
As we know that to make train data stationary, we have taken the difference of 1.
So now plotting the ACF and PACF plot of the differenced train data time series.

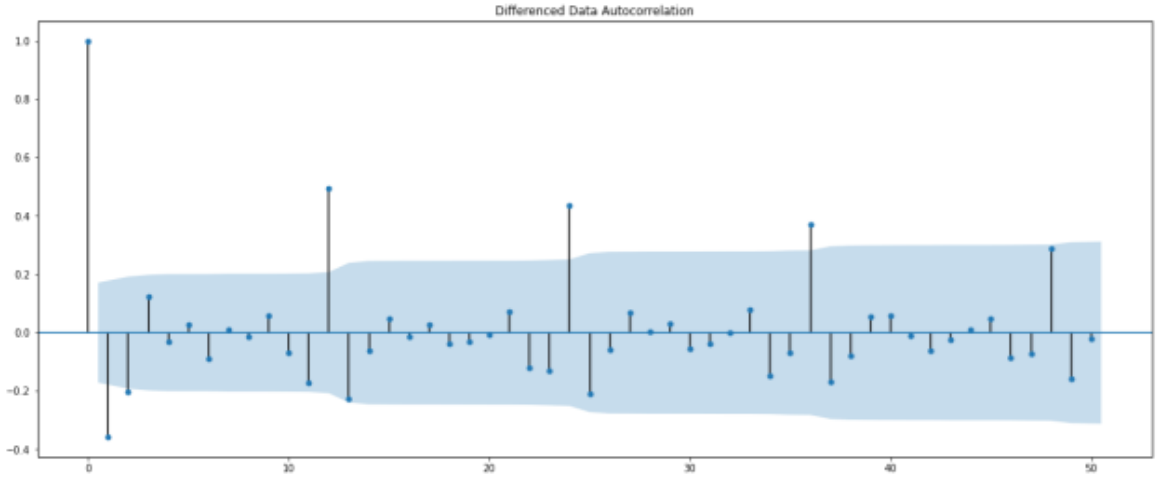


Figure 33 Differential ACF Plot of Training data

From the ACF plot we can see the significant lag count is 2, so we can take **q value as 2**.

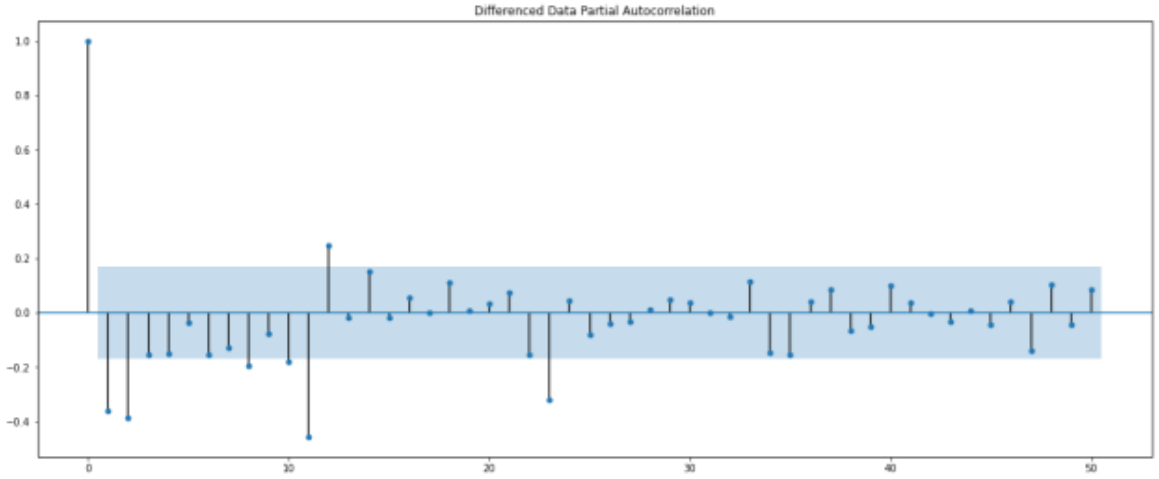


Figure 34 Differential PACF Plot of Training data

From the PACF plot we can see the significant lag count is 0, so we can take **p value as 2.** So, the order will be (2,1,2)

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-633.649
Method:	css-mle	S.D. of innovations	29.975
Date:	Sun, 23 May 2021	AIC	1279.299
Time:	11:56:04	BIC	1296.550
Sample:	02-29-1980 - 12-31-1990	HQIC	1286.309

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4911	0.081	-6.076	0.000	-0.649	-0.333
ar.L1.D.Rose	-0.4383	0.218	-2.015	0.044	-0.865	-0.012
ar.L2.D.Rose	0.0269	0.109	0.246	0.806	-0.188	0.241
ma.L1.D.Rose	-0.3316	0.203	-1.633	0.102	-0.729	0.066
ma.L2.D.Rose	-0.6684	0.201	-3.332	0.001	-1.062	-0.275

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-2.0289	+0.0000j	2.0289	0.5000
AR.2	18.3387	+0.0000j	18.3387	0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.4961	+0.0000j	1.4961	0.5000

Figure 35 Summary test report of ARIMA model based on the cutoff points of ACF and PCF plot

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for ARIMA (2,1,2) is **15.354877.**

SARIMA Model with Seasonality 6 based on the cut-off points of ACF and PACF.

We have taken differentiation of 6 for Seasonality of 6 so diff (6) is D=1. Time Series plot after taking diff (6)

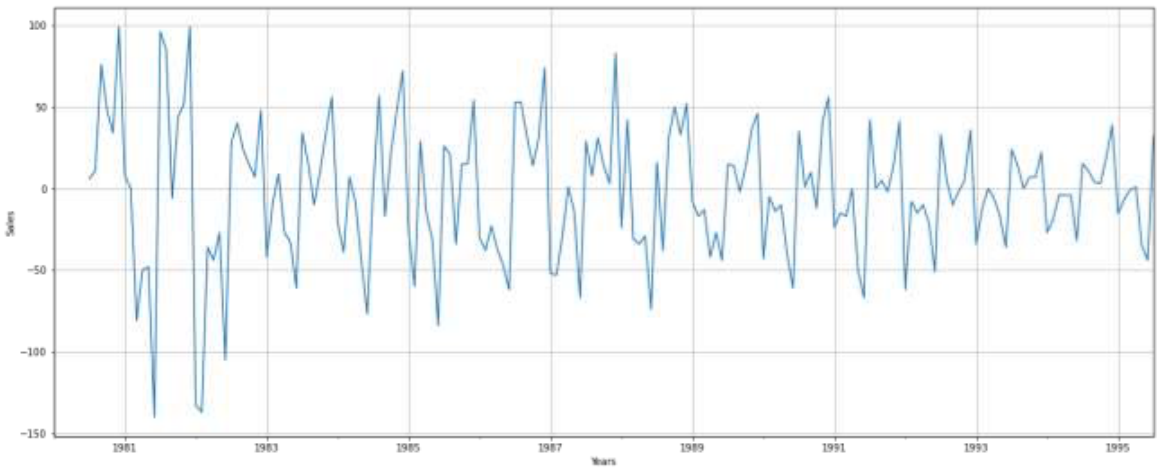


Figure 36 Time Series Plot after Differentiation of 6 on Whole data

Applying the diff (6) on Train series and checking stationary property

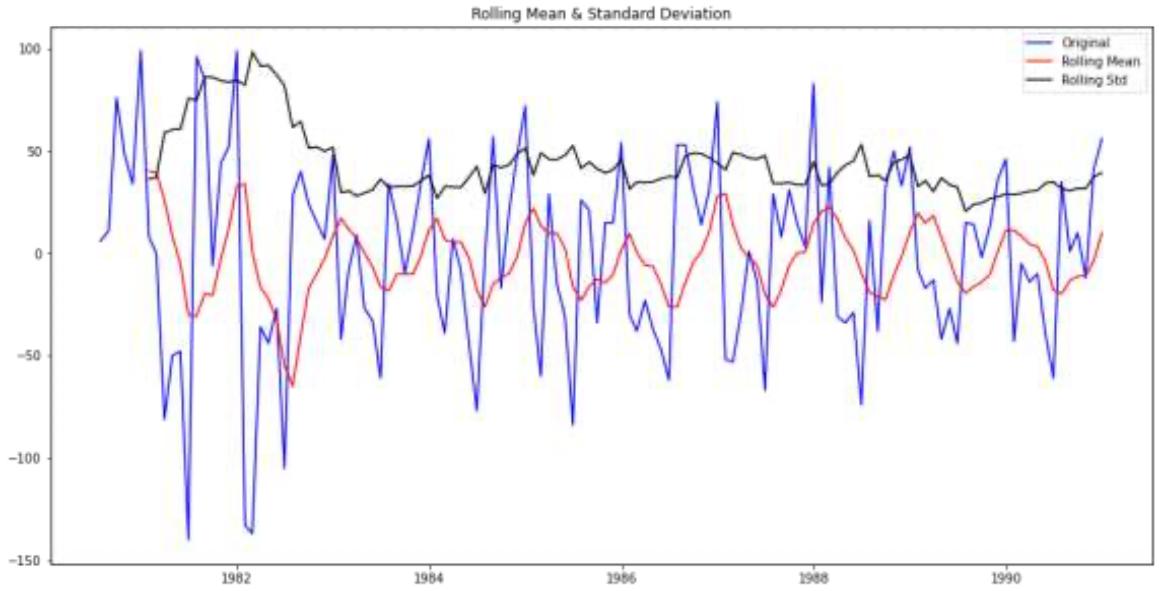


Figure 37 Time Series Stationary check after Differentiation of 6 on Train data.

```
Results of Dickey-Fuller Test:
Test Statistic                  -7.442449e+00
p-value                          5.956534e-11
#Lags Used                       7.000000e+00
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 38 ADF test report SARIMA model seasonality 6 based on the cutoff points of ACF and PCF plot

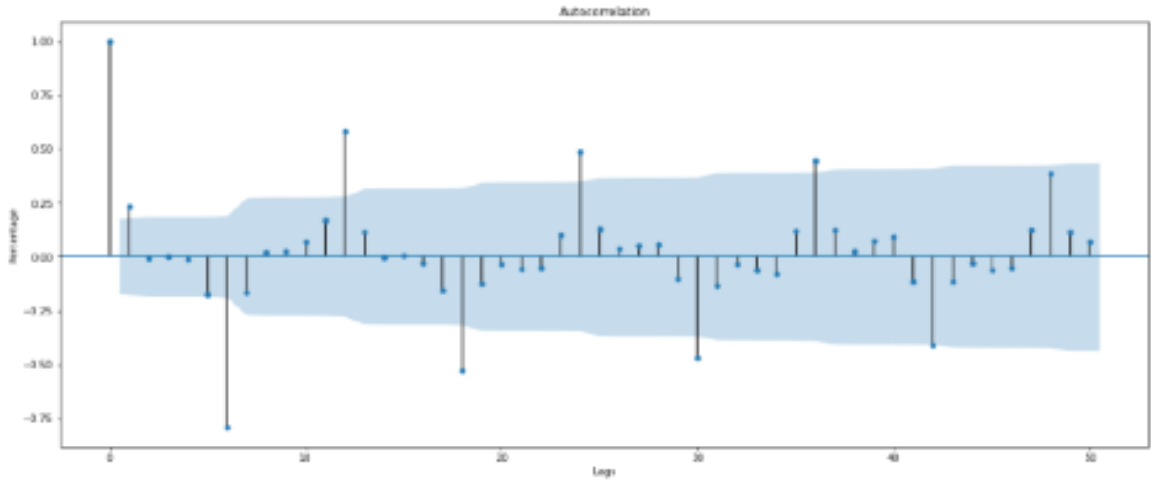


Figure 39 ACF Plot after differentiation of 6

So we can say from p value that train series is stationary after taking seasonal diff (6)

From the ACF plot we can see the significant lag count is 1, so we can take **Q value as 1.**

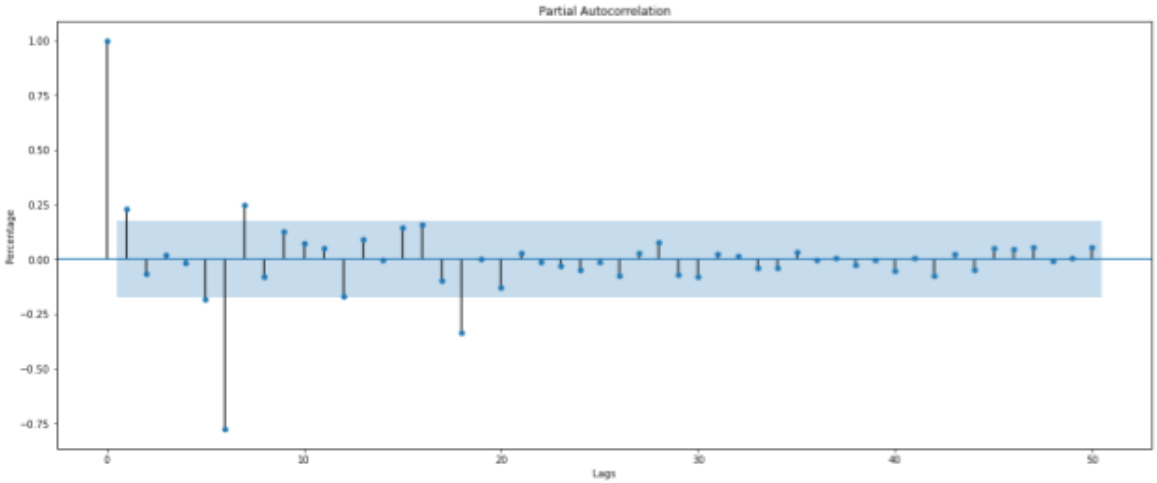


Figure 40 PACF Plot after differentiation of 6

From the PACF plot we can see the significant lag count is 1, so we can take **P value as 1.** So, the order will be (2,1,2) (1,1,1,6)

SARIMA model seasonality 6 summary is as follow:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(1, 1, [1], 6)	Log Likelihood	-536.814
Date:	Sun, 23 May 2021	AIC	1087.627
Time:	14:24:51	BIC	1106.902
Sample:	0	HQIC	1095.452
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6933	0.066	-10.464	0.000	-0.823	-0.563
ar.L2	0.0624	0.069	0.907	0.365	-0.073	0.197
ma.L1	-0.0087	642.724	-1.36e-05	1.000	-1259.724	1259.706
ma.L2	-1.0087	648.287	-0.002	0.999	-1271.628	1269.610
ar.S.L6	-0.8176	0.056	-14.578	0.000	-0.928	-0.708
ma.S.L6	0.0625	0.102	0.612	0.540	-0.138	0.263
sigma2	572.9511	3.68e+05	0.002	0.999	-7.21e+05	7.22e+05

Ljung-Box (Q):	32.85	Jarque-Bera (JB):	50.02
Prob(Q):	0.78	Prob(JB):	0.00
Heteroskedasticity (H):	0.29	Skew:	0.21
Prob(H) (two-sided):	0.00	Kurtosis:	6.19

Figure 41 Summary test report of SARIMA seasonality 6 model based on the cutoff points of ACF and PCF plot

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for SARIMA (2,1,2) (1,1,1,6) is **15.757185397935 75.**

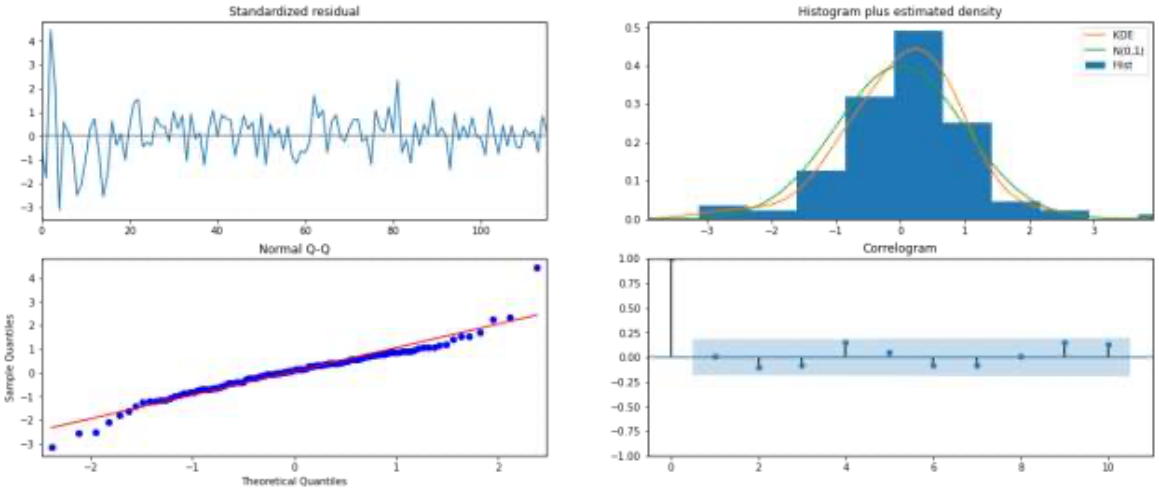


Figure 42 Model Performance check dist plot (SARIMA Seasonality 6) based on cutoff point of ACF and PACF plot

From the above plot we can see that the Residual are around the zero line. The histogram is normalized.

The near about all points in the Q-Q plot lies on the line.
We can see there is one significant lag in the series.

So, this model is not good enough.

SARIMA Model with Seasonality 12 based on the cut-off points of ACF and PACF.

We have taken differentiation of 12 for Seasonality of 12 so diff (12) is D=2
Time Series plot after taking diff (12)

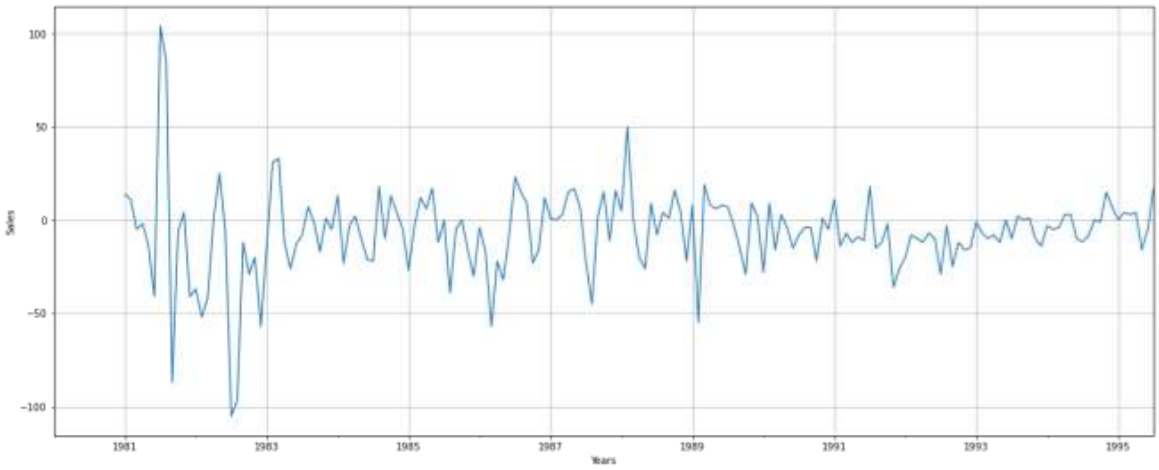


Figure 43 Time Series Plot after Differentiation of 12 on whole data

Applying the diff (12) on Train series and checking stationary property

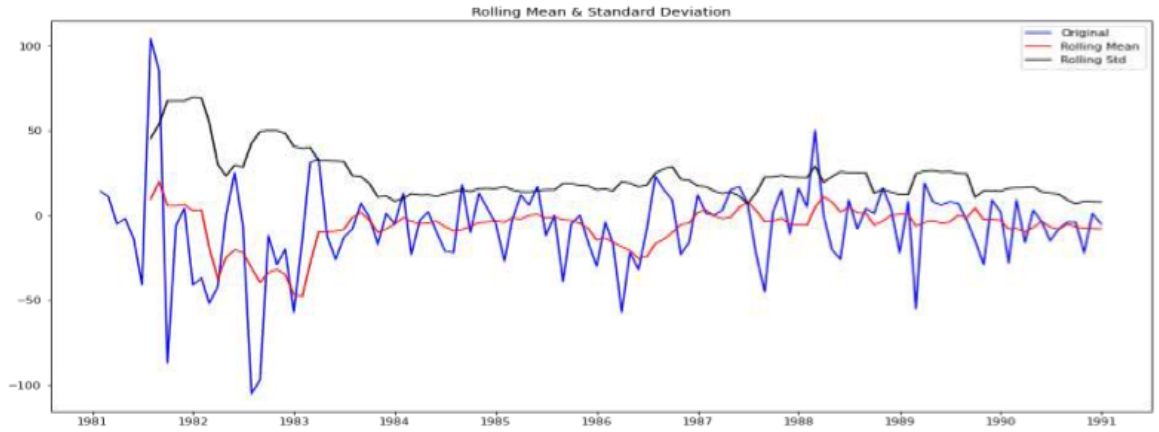


Figure 44 Time Series Stationary check after Differentiation of 12 on Train data

```
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 45 ADF test report SARIMA model seasonality 12 based on the cutoff points of ACF and PCF plot

So we can say from p value that train series is stationary after taking seasonal diff (12)

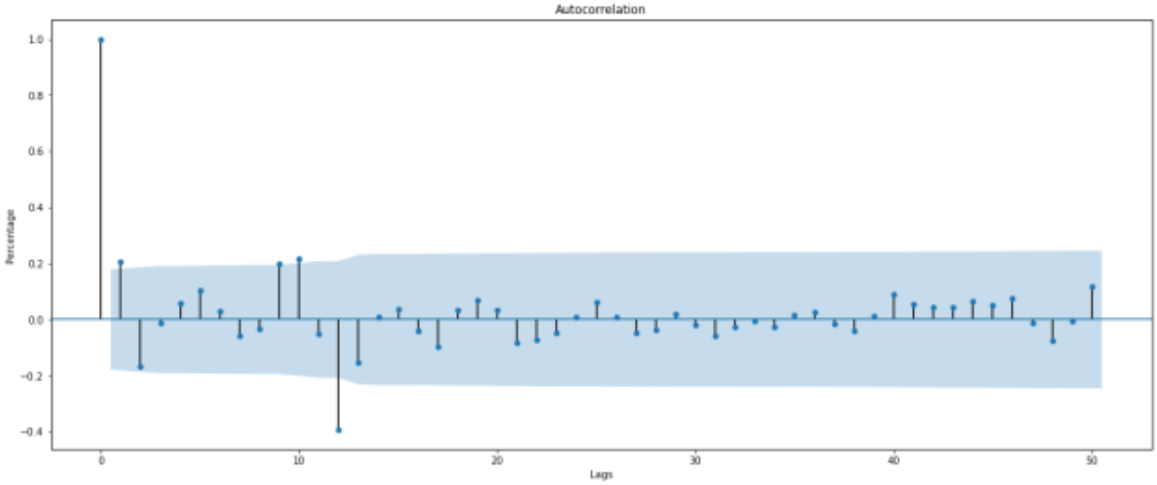


Figure 46 ACF Plot after differentiation of 12
From the ACF plot we can see the significant lag count is 1, so we can take **Q value as 1.**

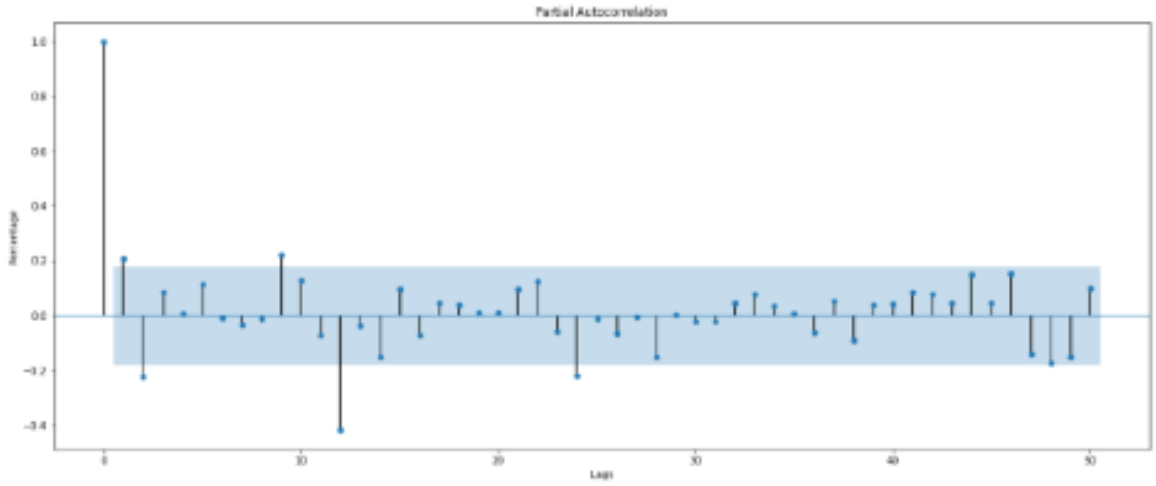


Figure 47 PACF Plot after differentiation of 12

SARIMA model seasonality 12 summary is as follow:

From the PACF plot we can see the significant lag count is 1, so we can take **P value as 2.** So, the order will be (2,1,2) (2, 2, 1, 12)

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(2, 2, [1], 12)	Log Likelihood	-355.585
Date:	Sun, 23 May 2021	AIC	727.170
Time:	14:26:40	BIC	746.325
Sample:	0	HQIC	734.855
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3784	3.296	-0.115	0.909	-6.838	6.081
ar.L2	0.0596	0.289	0.206	0.837	-0.507	0.626
ma.L1	-0.5368	55.923	-0.010	0.992	-110.143	109.069
ma.L2	-0.4631	26.745	-0.017	0.986	-52.883	51.956
ar.S.L12	-0.4172	0.150	-2.780	0.005	-0.711	-0.123
ar.S.L24	-0.1949	0.106	-1.840	0.066	-0.403	0.013
ma.S.L12	-1.0001	814.955	-0.001	0.999	-1598.282	1596.282
sigma2	270.0564	2.18e+05	0.001	0.999	-4.28e+05	4.28e+05

Ljung-Box (Q):	27.42	Jarque-Bera (JB):	5.58
Prob(Q):	0.93	Prob(JB):	0.06
Heteroskedasticity (H):	0.89	Skew:	-0.23
Prob(H) (two-sided):	0.75	Kurtosis:	4.20

Figure 48 Summary test report of SARIMA sesonality 12 model based on the cutoff points of ACF and PCF plot

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for SARIMA (2,1,2) (2,2,1,12) is **33.9846357881655.**

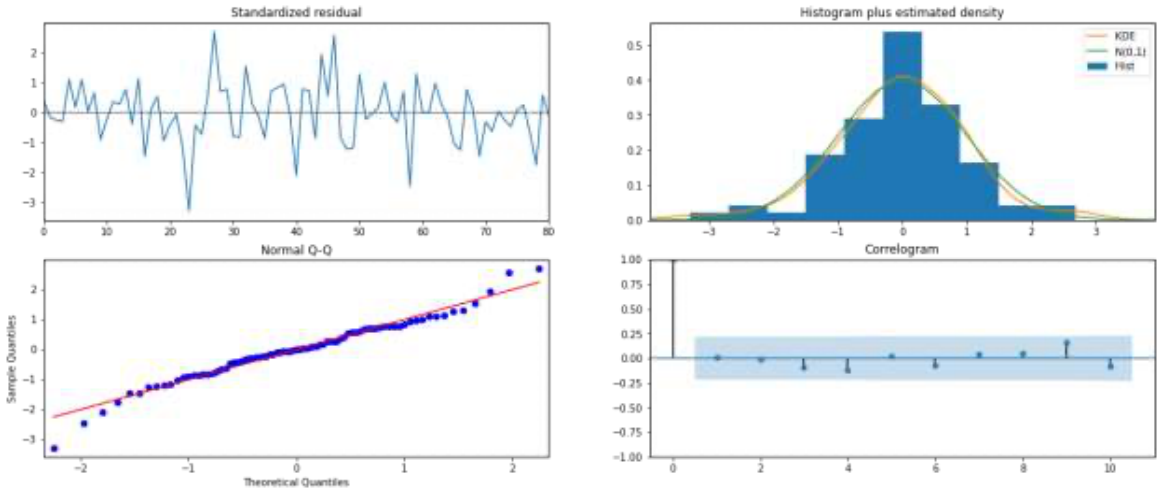


Figure 49 Model Performance check dist plot (SARIMA Seasonality 12) based on cutoff point of ACF and PACF plot

From the above plot we can see that the Residual are around the zero line.
The histogram is normalized.
The near about all points in the Q-Q plot lies on the line.
But we can see there is no significant lag in the series.

So, this model is good enough.

Sr No	Model	RMSE
1	Manual ARIMA (2,1,2)	15.354877
2	Manual SARIMA at seasonality 6 (2,1,2) (1,1,1,6)	15.757185
3	Manual SARIMA at seasonality 12 (2,1,2) (2,2,1,12)	33.984636

Table 9 RMSE values comparison based on cutoff points of ACF and PACF.

So, among ARIMA/SARIMA models based on the cut-off points of ACF and PACF Manual ARIMA is better among all.

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

Sr. No	Model	RMSE
1	Alpha=0.3, Beta=0.4, Gamma=0.3, Triple Exponential Smoothing	10.94544
2	2pointTrailingMovingAverage	11.52928
3	4pointTrailingMovingAverage	14.4514
4	6pointTrailingMovingAverage	14.56633
5	9pointTrailingMovingAverage	14.72763
6	manual ARIMA (2,1,2)	15.35488
7	auto ARIMA (0,1,2)	15.61783
8	manual SARIMA (2,1,2) (1,1,1,6)	15.75719
9	Alpha=0.106, Beta=0.048, Gamma=0, Triple Exponential Smoothing	17.36949
10	auto SARIMA (1,1,2) (2,0,2,6)	26.13367
11	auto SARIMA (1,0,2) (2,0,2,12)	29.74942
12	manual SARIMA (2,1,2) (2,2,1,12)	33.98464
13	Alpha=(default) Simple Exponential Smoothing	36.79624
14	Alpha=0.3, Simple Exponential Smoothing	47.50482
15	Linear Regression	51.43331
16	Simple Average Model	53.46057
17	Alpha=0.1578, Beta=0.1578, Double Exponential Smoothing	70.57245
18	Naïve Model	79.71877
19	Alpha=0.3, Beta=0.3, Double Exponential Smoothing	265.5676

Table 10 All models RMSE results

Among all the model the Triple Exponential model RMSE value is least. So, for this case study the Triple Exponential model is the best fit.

Q.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.

- The most optimum model among all is Triple Exponential model.
- The Triple Exponential model is built on whole data for future forecast.
- Best fit alpha, beta, gamma value (Alpha=0.3, Beta=0.4, Gamma=0.3) are places in the model and
  built.
- While forecasting gives the step of 12 as we have to forecast for next 12 months.

After building the model the forecast is as shown in the Figure 48

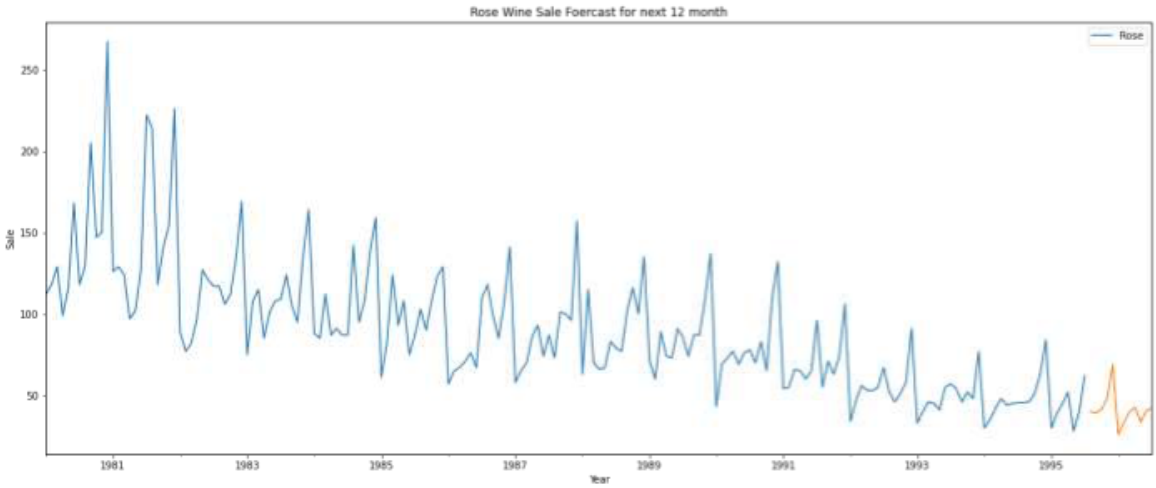


Figure 50 Rose Wine Sales Forecast for next 12 months

Model Evaluation:

- Calculating the RMSE value for the full model build
- RMSE of the Full Model 24.266535961104537

Confidence Band/Interval:

• When we predict certain time points into the future, we might need to have a concept of

Confidence Band for our predictions.
• This gives us range of values in which our predictions will be lying in the future or for the future

time stamps.

Hence the upper and lower confidence bands at 95% confidence level are calculated.

	lower_CI	prediction	upper_ci
1995-08-31	-7.559277	40.074648	87.708574
1995-09-30	-8.388196	39.245730	86.879656
1995-10-31	-6.321556	41.312369	88.946295
1995-11-30	0.433543	48.067469	95.701394
1995-12-31	21.101528	68.735454	116.369379
1996-01-31	-21.639634	25.994292	73.628217
1996-02-29	-14.296324	33.337602	80.971527
1996-03-31	-7.899125	39.734801	87.368726
1996-04-30	-4.886942	42.746984	90.380909
1996-05-31	-14.060390	33.573536	81.207461
1996-06-30	-7.128660	40.505265	88.139191
1996-07-31	-5.223399	42.410526	90.044452

Table 11 Upper and Lower confidence bands at 95% confidence level

Forecast Graph with the confidence band.

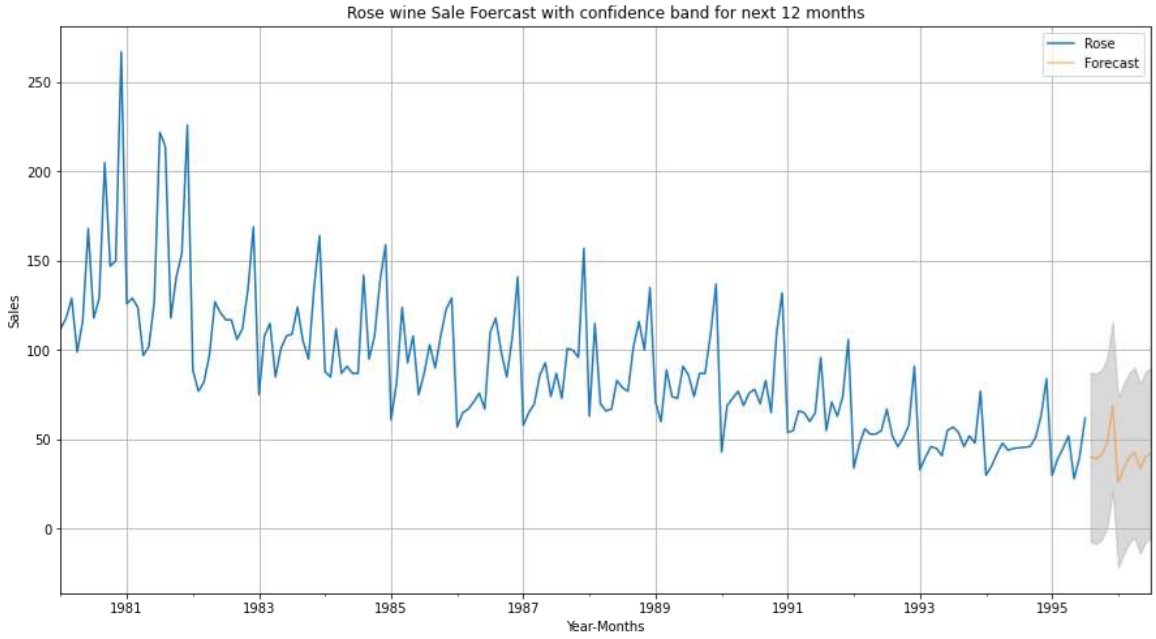


Figure 51 Forecast Graph with the confidence band

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

Summary the various steps performed in this project:

1. Do EDA, if any missing value present do the missing value treatment.
2. Plot the time series.
3. There are two approaches to proceed do Decomposition or ARIMA.

Decomposition Approach

1. See if trend & seasonality is present in the data.
2. If trend & seasonality is not in the data then apply linear regression, naive, simple average,
   moving averages, simple exponential smoothing.
3. Check Accuracy of the model by calculating RMSE.
4. If only trend is available in the data, then apply Double Exponential Smoothing model & check accuracy.
5. If both trend and seasonality is present, then apply Triple Exponential Smoothing model & check accuracy.

ARIMA Approach

1. If data is not stationary, then do differencing and make series stationary.

2. If data is stationary, then plot the ACF &PACF plot

3. If the data is seasonal then apply ARIMA (p, d, q) (P, D, Q) F model & check accuracy.

4. If the data is not seasonal then apply ARIMA (p, d, q) model & check accuracy.

Comment on the model

- Among all the model the Triple Exponential model RMSE value is least.

- So, for this case study the Triple Exponential model is the best fit.

- The forecasting the wine sale using this model is good enough, the confidence band is not much
away from the actual forecast.

Findings

- We can see there is highest sale of Rose Wine is in month of December for every year, means in December the demand of Rose wine is generated i.e., there is seasonality in the sales.

- We can see the month January; April Sale is low among all the months.

Measures that the company should be taking for future sales.

• So, as we come to known that the December is peak point for sale so accordingly production and stock of the Rose wine must be maintained in the month of the December.

- Company must be able to grab this December season and increase the sale and earn profits.

- Again, we can see the month January, April Sale is low among all the months, so some strategy must be planned to increase the sale, such as some discounts/ offers, so that the sale will get

increase in those respective months.

- As we can see year by year the sale is dropping, so the study of the Market Place where the wine
is sale out must be done. Reason why the demand is decreased need to be find out and corrective
actions must be taken over it to increase the sale.

- Demand may be decrease due to many reasons such as Taste of the wine, cost, Market area,
crowd in the locality there can be various reason. So, finding the appropriate reason and working on it could increase the future sale of the ABC Estate.

THE END