TIME SERIES FORECASTING

Project Report

Part - 2

Submitted by,

Sindhu R Udupa.

BATCH: PGPDSBA.O. NOV22.B

Contents

List of Figures

List of Tables

Part 1: Report on the sales of the Rose wine.

The data contains the information on the sales of a wine called ‘Rose’ from January 1980 to July 1995.

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

The data contains two columns namely ‘Year Month’ and ‘Rose’. The data has been read as an appropriate time series by using the ‘parse date’ parameter in the read function of the pandas, which will help us identify the column ‘Year Month’ as date. We have set it as the index of our data frame, which makes it easier to analyze the data. The column ‘Rose’ contains information about the sales of that wine in that particular month.

The plot the time series is as given below.

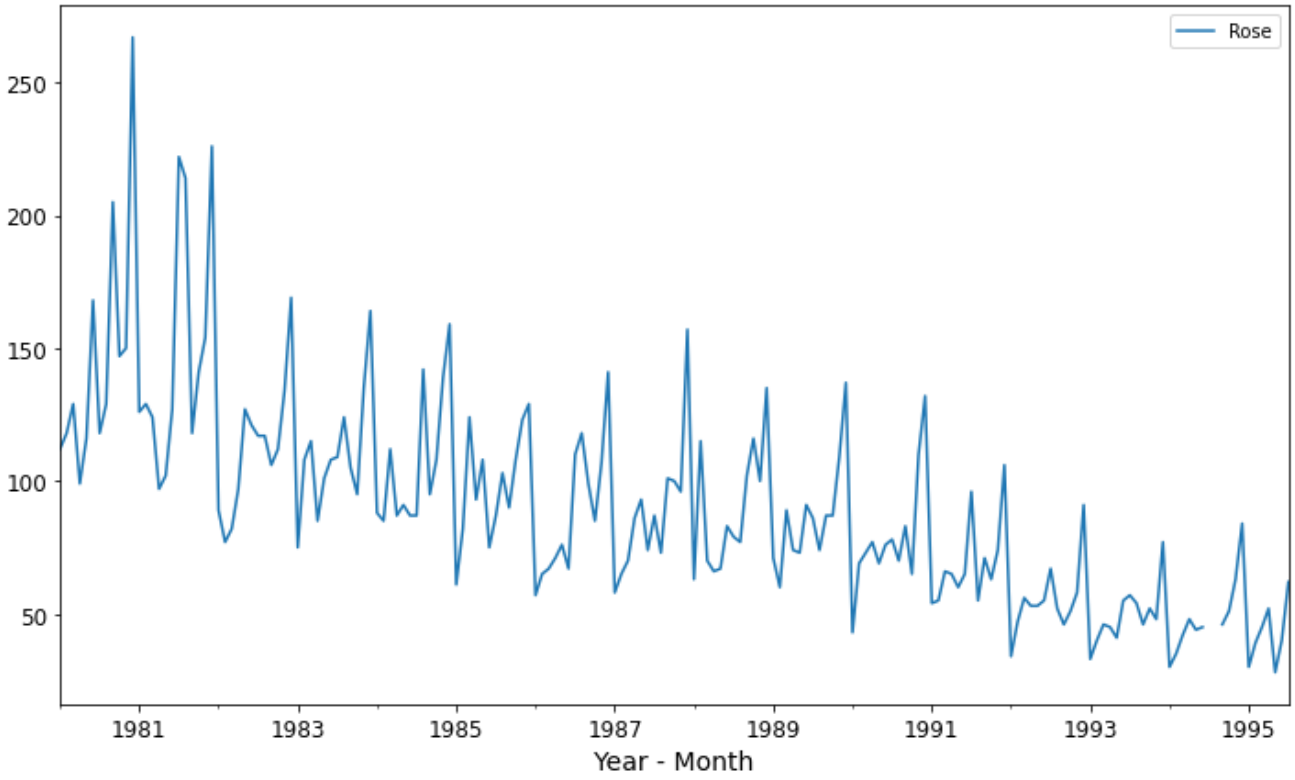


Fig. 1a: Plot of the sales of the wine Rose before missing value imputation.

We can see a hole in the time series between 1994 and 1995, which tells us that there are some missing values in the data.

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

Exploratory Data Analysis:

a. After making the Year Month column as index, there is only one column named ‘Rose’ of integer data type, which depicts the number of sales of the wine in that particular month.
b. There are no duplicate values present in the data.
c. The data starts from 1980 Jan and ends in 1995 July.
d. There are two missing values in the data. The sale of wine is not recorded in July and August 1994.
e. We use the interpolate function from the pandas library to impute the missing values.

The time series after missing value imputation is shown below.

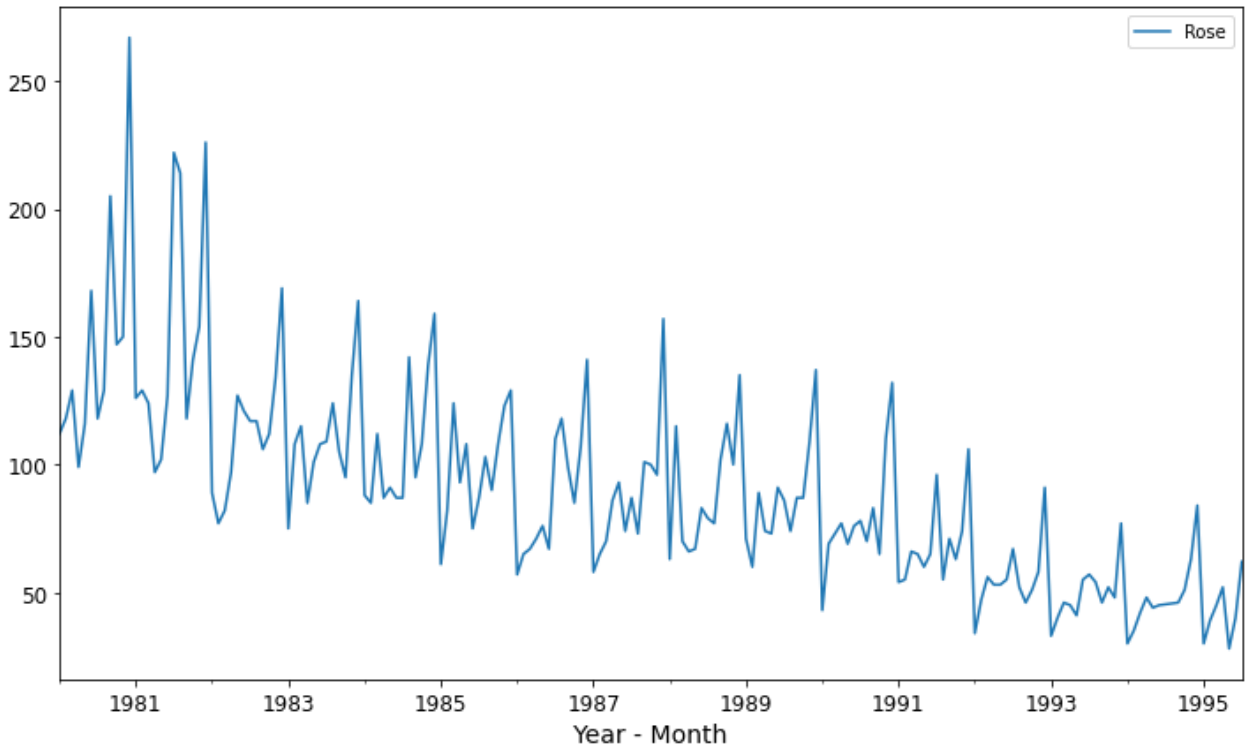


Fig. 1b: Plot of the sales of the wine Rose after missing value imputation.

f. The monthly sales, yearly sales of the wine are as follows:

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Month																
Jan	112.0	126.0	89.0	75.0	88.0	61.0	57.0	58.0	63.0	71.0	43.0	54.0	34.0	33.0	30.0	30.0
Feb	118.0	129.0	77.0	108.0	85.0	82.0	65.0	65.0	115.0	60.0	69.0	55.0	47.0	40.0	35.0	39.0
Mar	129.0	124.0	82.0	115.0	112.0	124.0	67.0	70.0	70.0	89.0	73.0	66.0	56.0	46.0	42.0	45.0
Apr	99.0	97.0	97.0	85.0	87.0	93.0	71.0	86.0	66.0	74.0	77.0	65.0	53.0	45.0	48.0	52.0
May	116.0	102.0	127.0	101.0	91.0	108.0	76.0	93.0	67.0	73.0	69.0	60.0	53.0	41.0	44.0	28.0
Jun	168.0	127.0	121.0	108.0	87.0	75.0	67.0	74.0	83.0	91.0	76.0	65.0	55.0	55.0	45.0	40.0
Jul	118.0	222.0	117.0	109.0	87.0	87.0	110.0	87.0	79.0	86.0	78.0	96.0	67.0	57.0	45.3	62.0
Aug	129.0	214.0	117.0	124.0	142.0	103.0	118.0	73.0	77.0	74.0	70.0	55.0	52.0	54.0	45.7	NaN
Sep	205.0	118.0	106.0	105.0	95.0	90.0	99.0	101.0	102.0	87.0	83.0	71.0	46.0	46.0	46.0	NaN
Oct	147.0	141.0	112.0	95.0	108.0	108.0	85.0	100.0	116.0	87.0	65.0	63.0	51.0	52.0	51.0	NaN
Nov	150.0	154.0	134.0	135.0	139.0	123.0	107.0	96.0	100.0	109.0	110.0	74.0	58.0	48.0	63.0	NaN
Dec	267.0	226.0	169.0	164.0	159.0	129.0	141.0	157.0	135.0	137.0	132.0	106.0	91.0	77.0	84.0	NaN

Table 1: Monthly Sales.

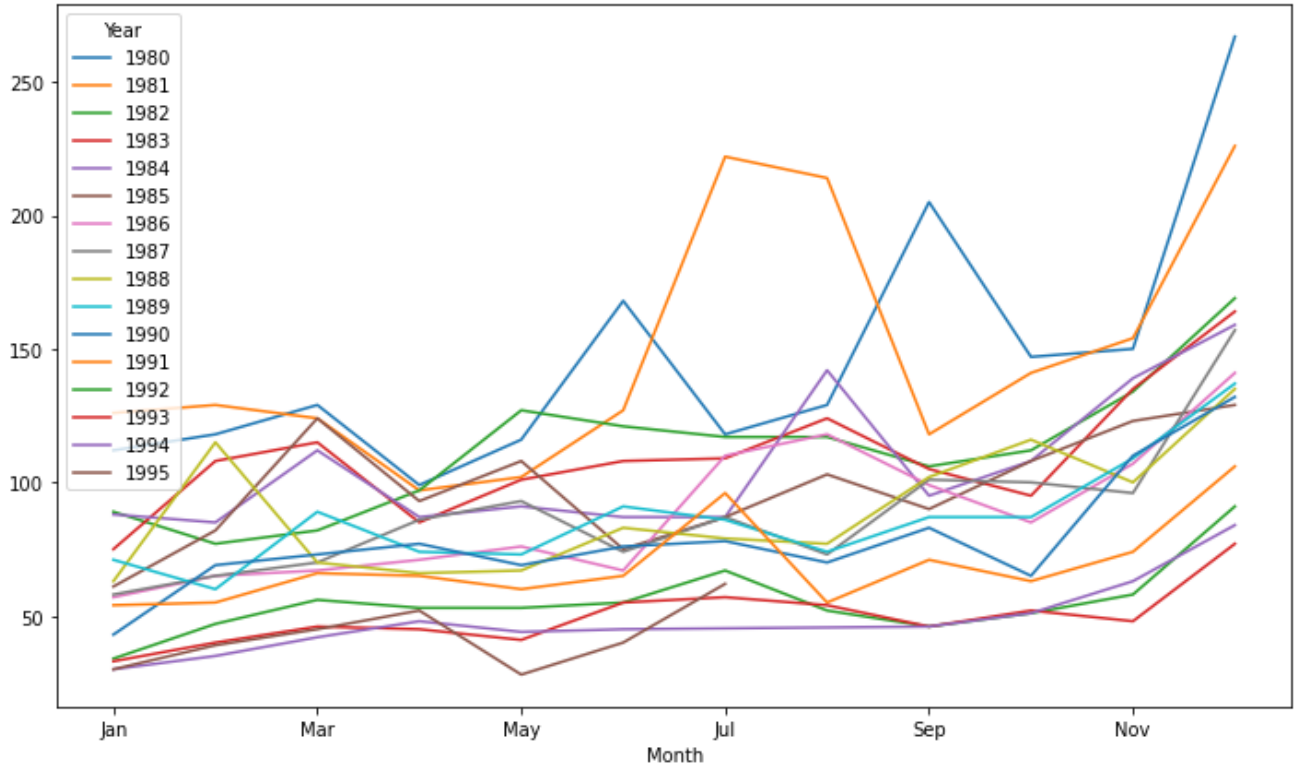


Fig. 2: Plot of the monthly sales.

Month Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.3	45.7	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

Table 2: Yearly Sales.

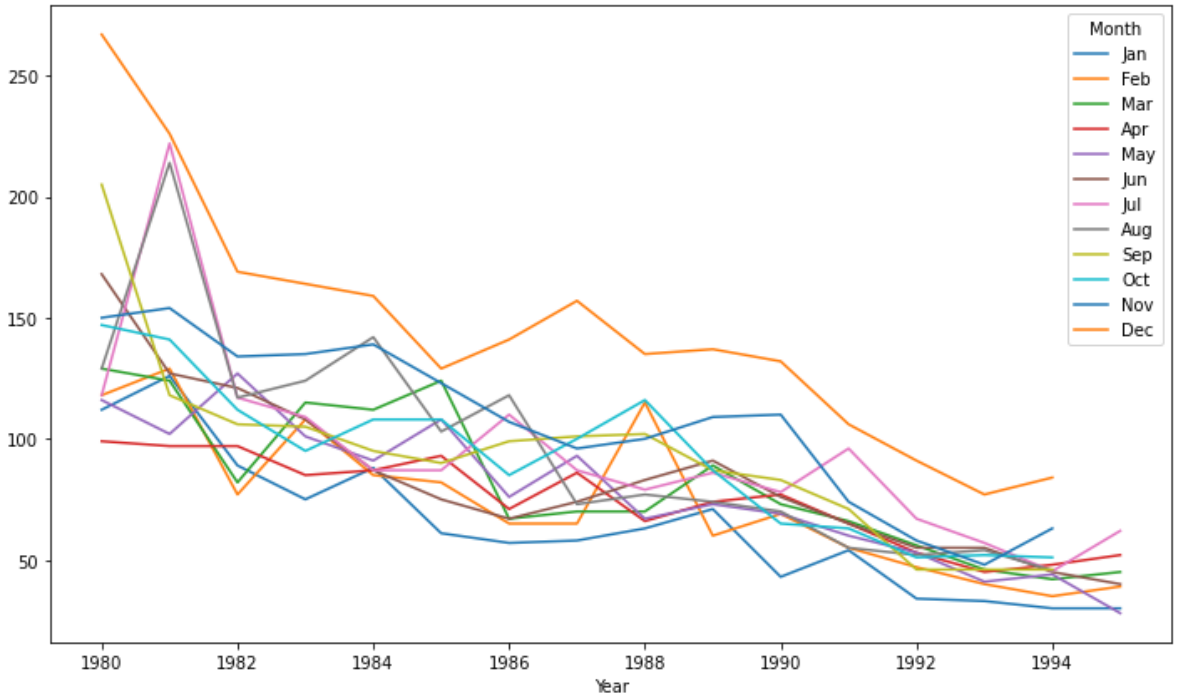
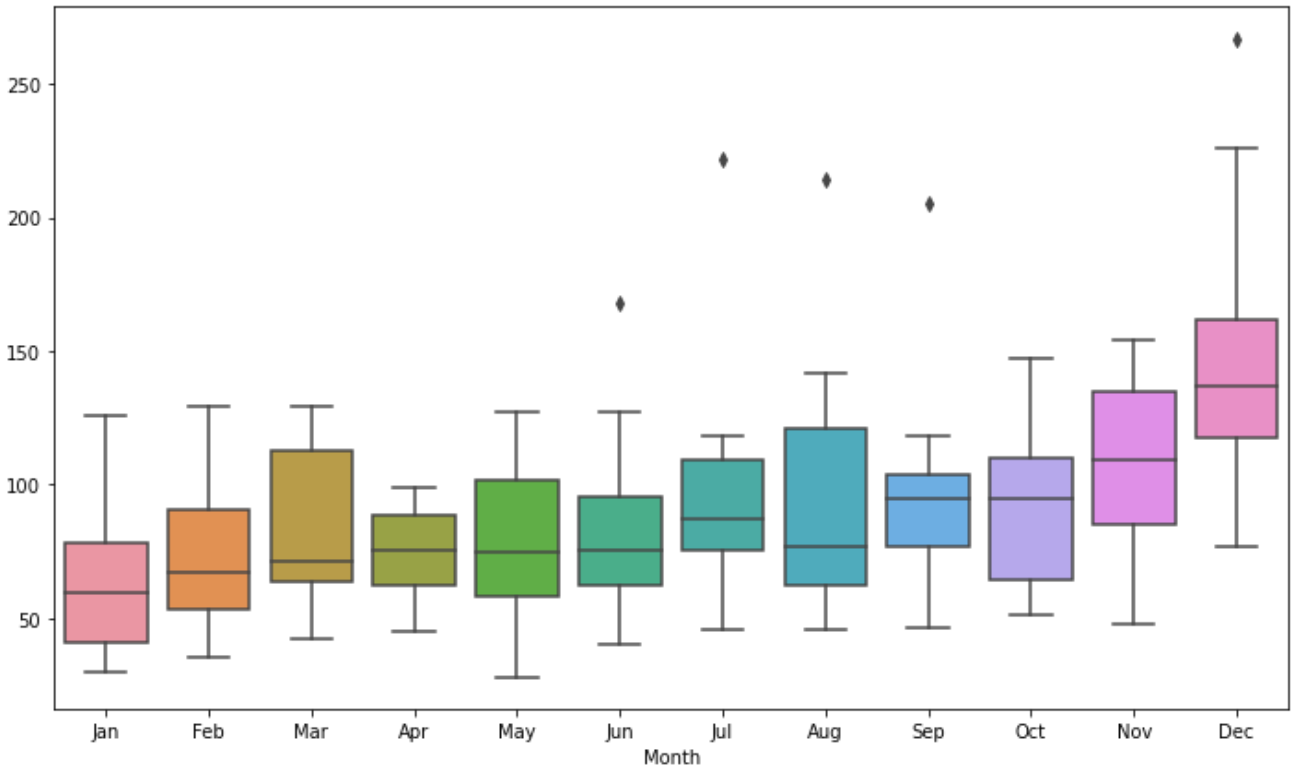


Fig. 3: Plot of the yearly sales.



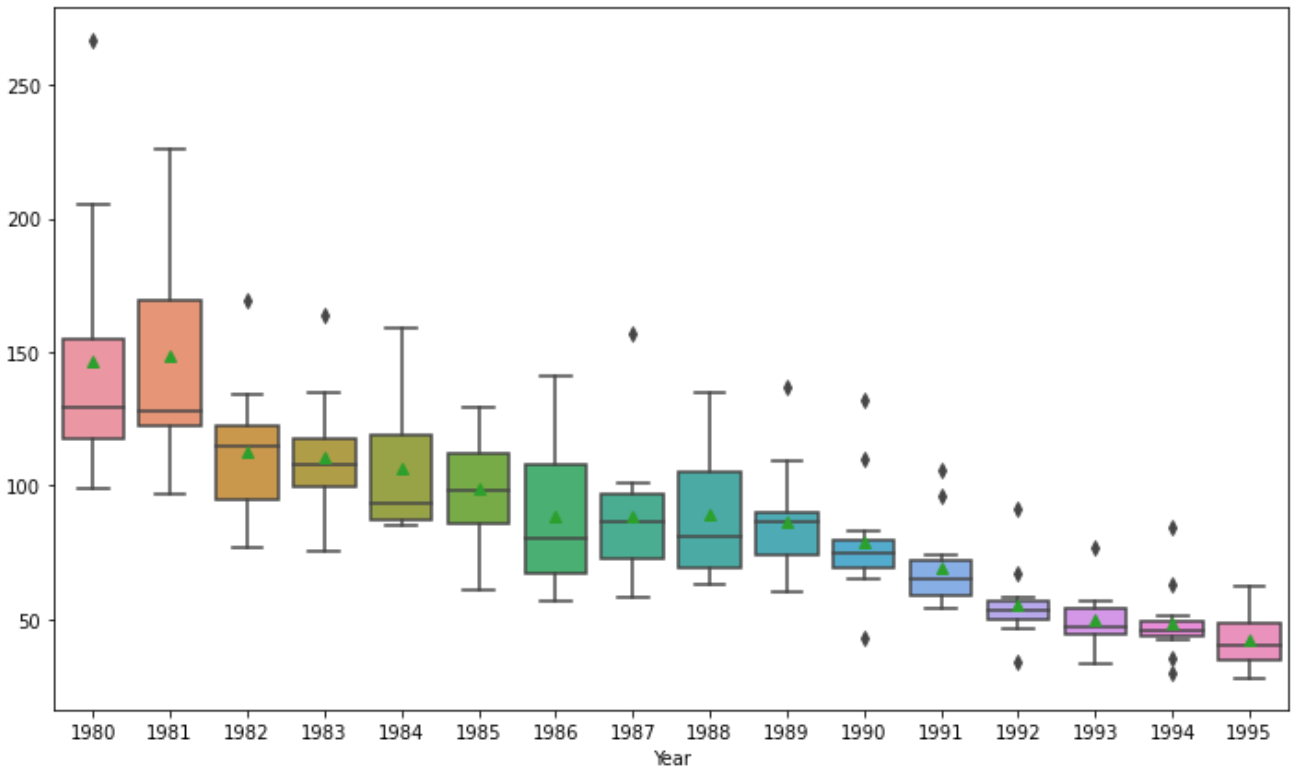


Fig. 4: Box plot of the monthly and yearly sales.

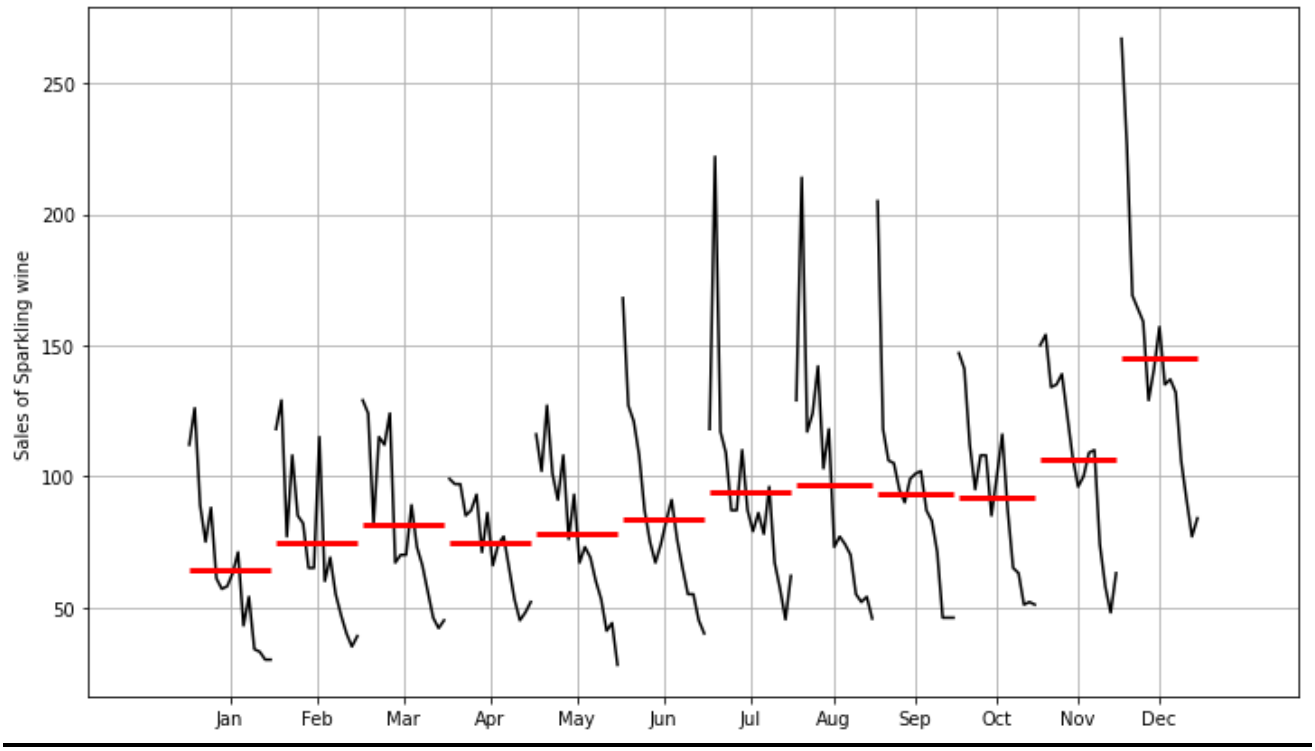


Fig. 5: Month plot of the sales.

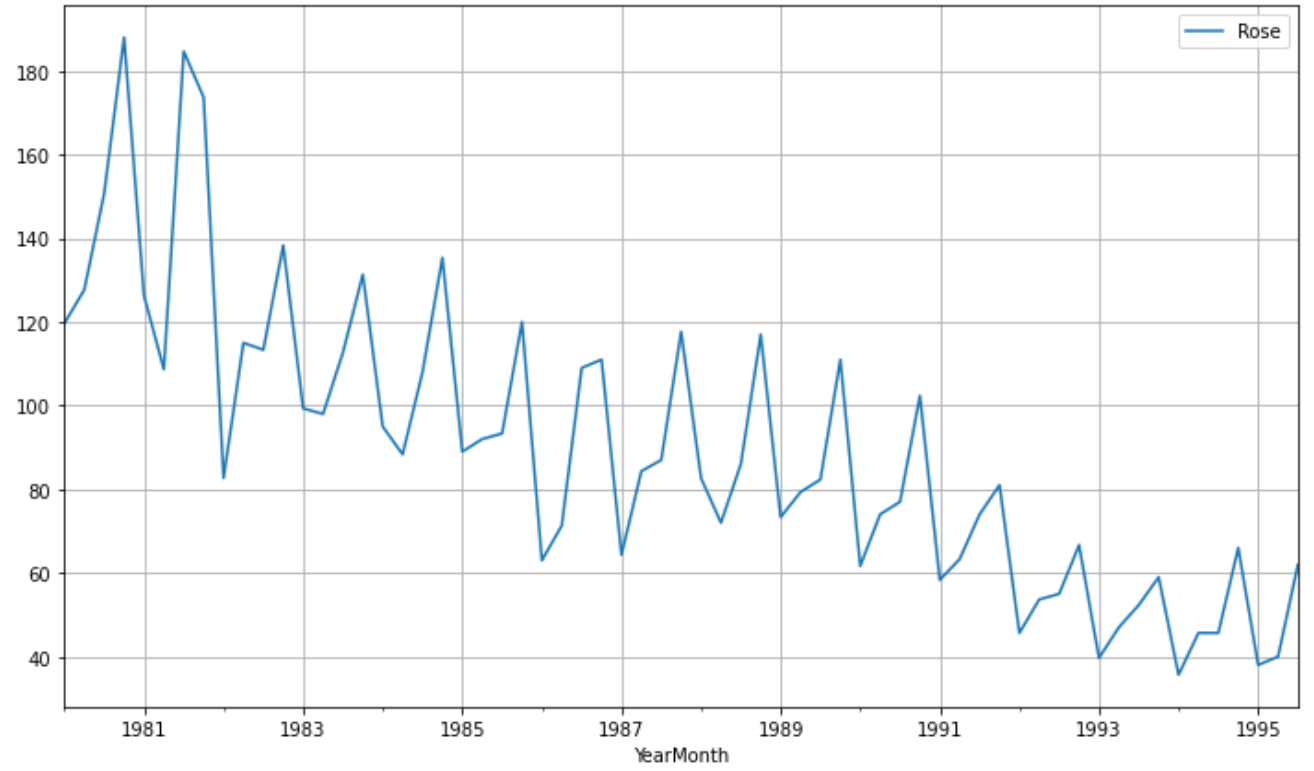


Fig. 6: Quarterly plot of the sales.

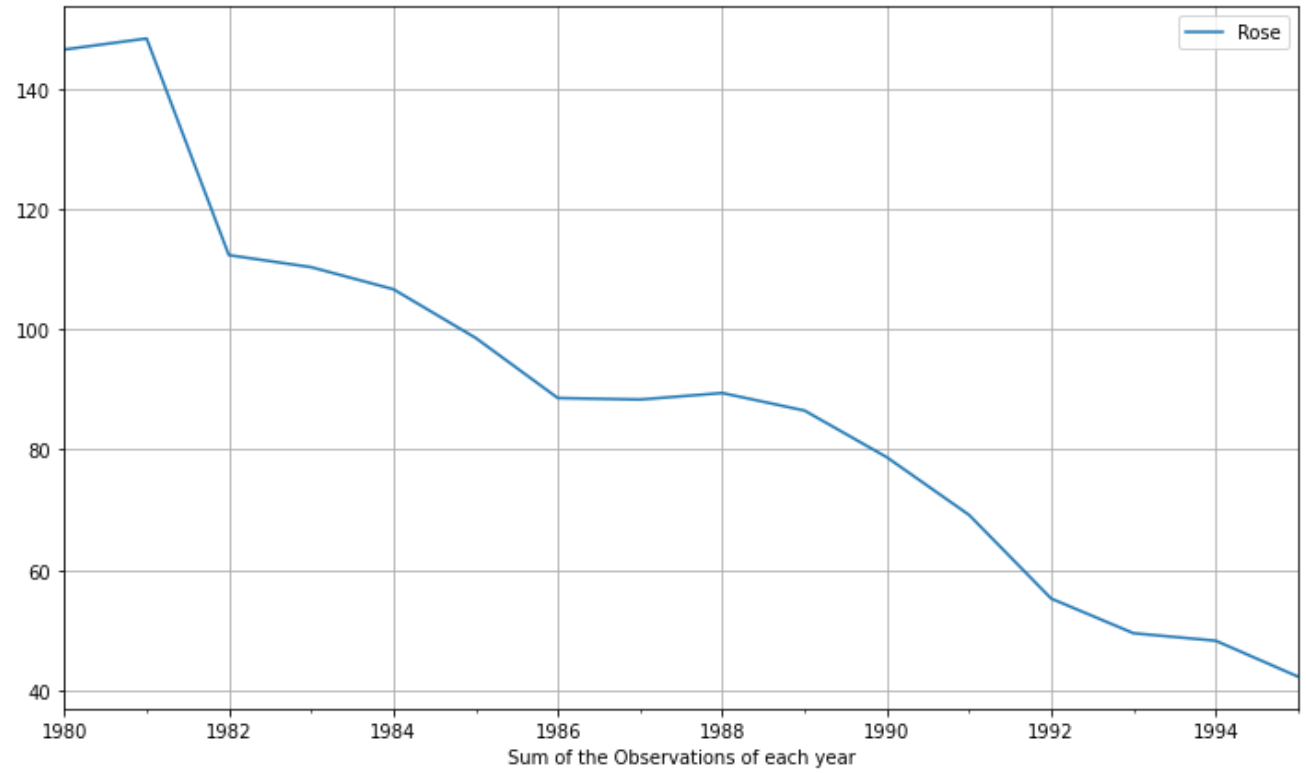


Fig. 7: Yearly plot of the sales.

. From the above tables and graphs we can see that

a. The sales of the wine are a bit more in the month of December.
b. The average sales of this wine are reducing every year.
c. From the quarterly plot, we can see that sales are more in the last quarter of each year.
d. From the yearly plot, the highest sale has happened in the year 1981, after that the sale has been reducing every year.

Additive Decomposition:

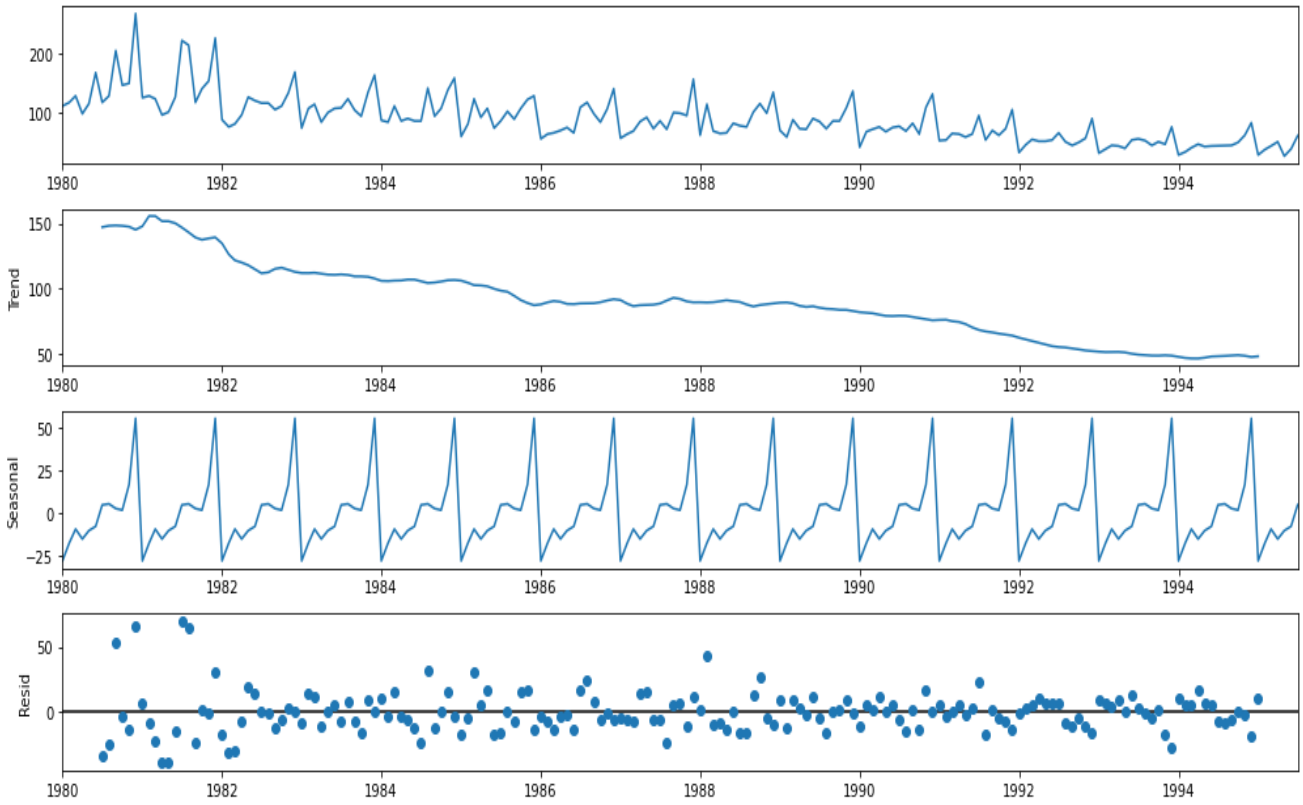


Fig. 8: Additive Decomposition.

The time series is decomposed into three components.

1. Trend

2. Seasonality and

3. Residual Component.

In an additive decomposition, the three components are added to get the time series.

We clearly see a decreasing trend in this decomposition.

Multiplicative Decomposition:

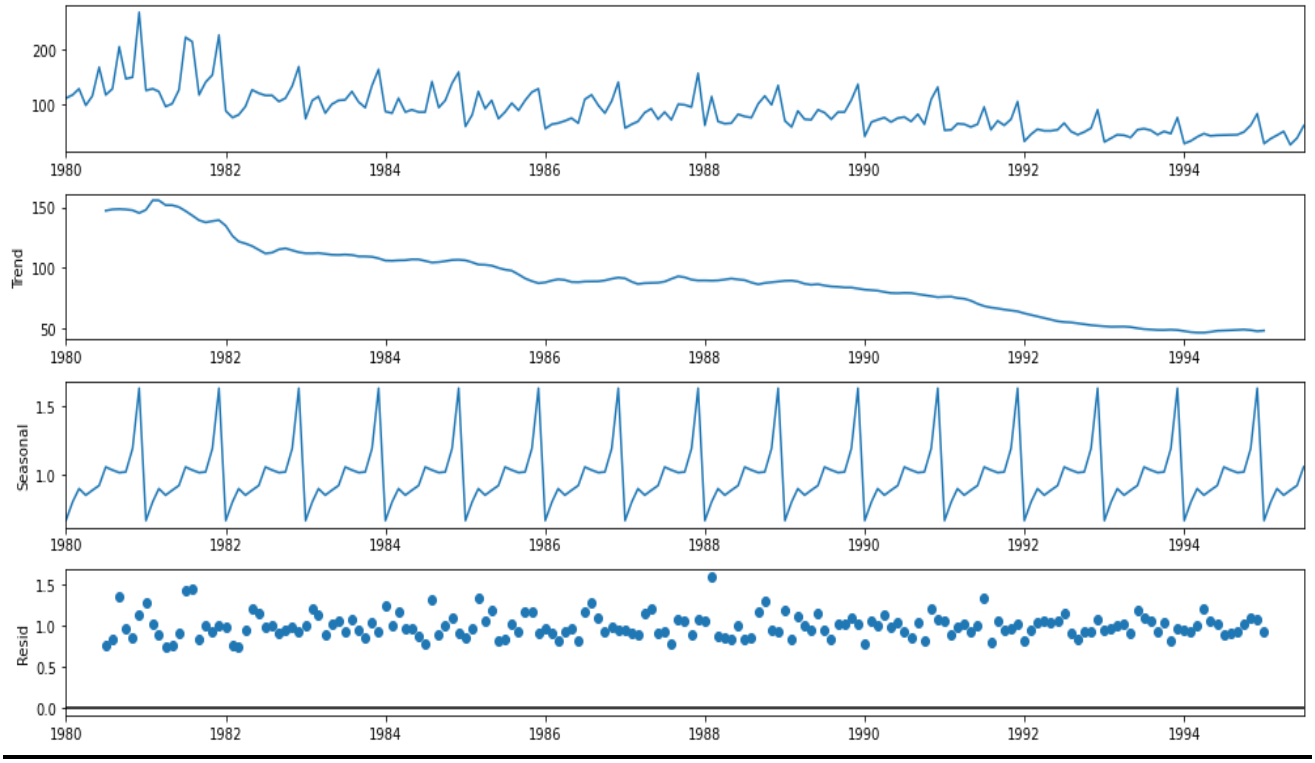


Fig. 9: Multiplicative Decomposition.

In the multiplicative decomposition, the three components – trend, seasonality and the residuals are multiplied to get the time series. The random component or the residual component should not follow any pattern, we see they do not follow any pattern in both the additive and the multiplicative decomposition. In this case also we see the trend following a decreasing pattern.

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

- The data has been split into train and test.
- The train data contains data from 1980 Jan to 1989 Dec.
- There are total of 132 data points in the train set.
- The test data contains data from 1991 Jan to 1995 July.
- There are total of 55 data points in the test set.

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471
...	...
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

Table 3: Train Set

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432
...	...
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Table 4: Test Set

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.

Regression Model

We are building a linear regression model using the sklearn library. First, we assigned integers to each data point in order to build the model. The sample is as follows.

First few rows of Training Data

YearMonth	Rose	time
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

First few rows of Test Data

YearMonth	Rose	time
1991-01-01	54.0	133
1991-02-01	55.0	134
1991-03-01	66.0	135
1991-04-01	65.0	136
1991-05-01	60.0	137

Last few rows of Training Data

YearMonth	Rose	time
1990-08-01	70.0	128
1990-09-01	83.0	129
1990-10-01	65.0	130
1990-11-01	110.0	131
1990-12-01	132.0	132

Last few rows of Test Data

YearMonth	Rose	time
1995-03-01	45.0	183
1995-04-01	52.0	184
1995-05-01	28.0	185
1995-06-01	40.0	186
1995-07-01	62.0	187

Table 5: Time points in train and test data.

Predictions on the test set:

The below table gives the predicted values of the Regression model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	72.063266
1991-02-01	55.0	71.568888
1991-03-01	66.0	71.074511
1991-04-01	65.0	70.580133
1991-05-01	60.0	70.085755
1991-06-01	65.0	69.591377
1991-07-01	96.0	69.096999
1991-08-01	55.0	68.602621
1991-09-01	71.0	68.108243
1991-10-01	63.0	67.613866

Table 6: Predictions of the Linear Regression model.

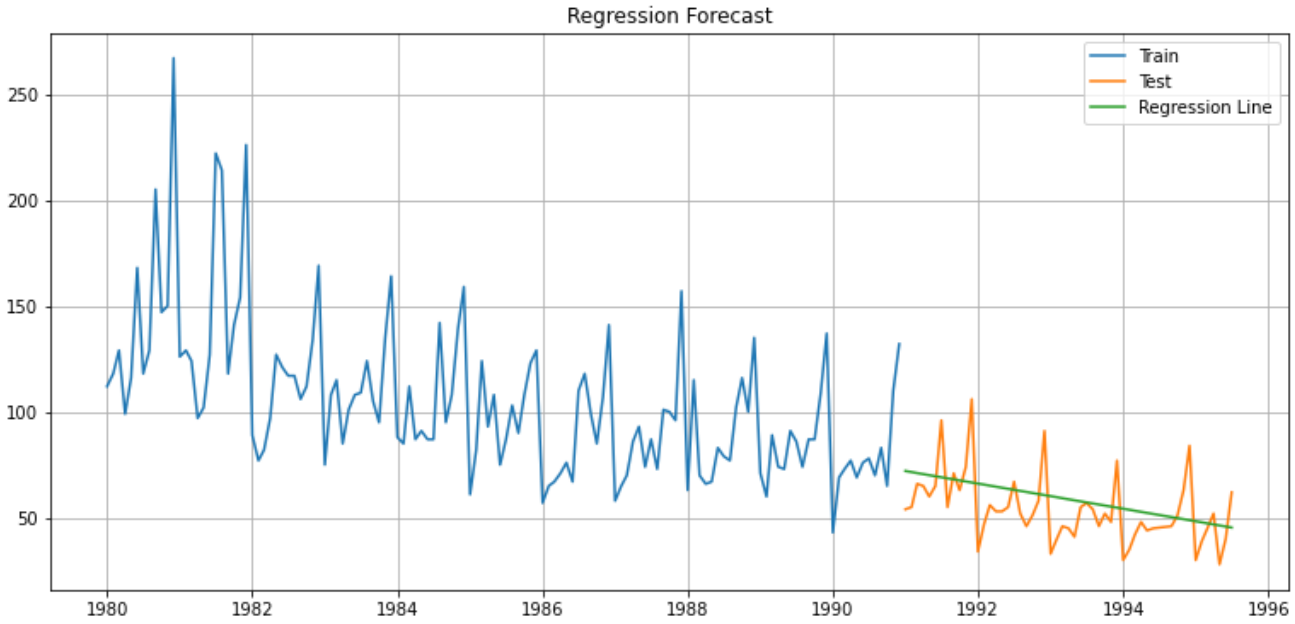


Fig. 10: Predictions of the Linear Regression model.

Model Evaluation:

The RMSE on the test set comes out to be 15.269.

Naïve Forecast Model

This model naively forecasts the last recorded data to the future. So the predictions of this model on the test data is the number of sales happened in Dec. 1990.

Predictions on the test set:

The below table gives the predicted values of the Naïve Forecast model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	132.0
1991-02-01	55.0	132.0
1991-03-01	66.0	132.0
1991-04-01	65.0	132.0
1991-05-01	60.0	132.0
1991-06-01	65.0	132.0
1991-07-01	96.0	132.0
1991-08-01	55.0	132.0
1991-09-01	71.0	132.0
1991-10-01	63.0	132.0

Table 7: Predictions of the Naïve Forecast model.

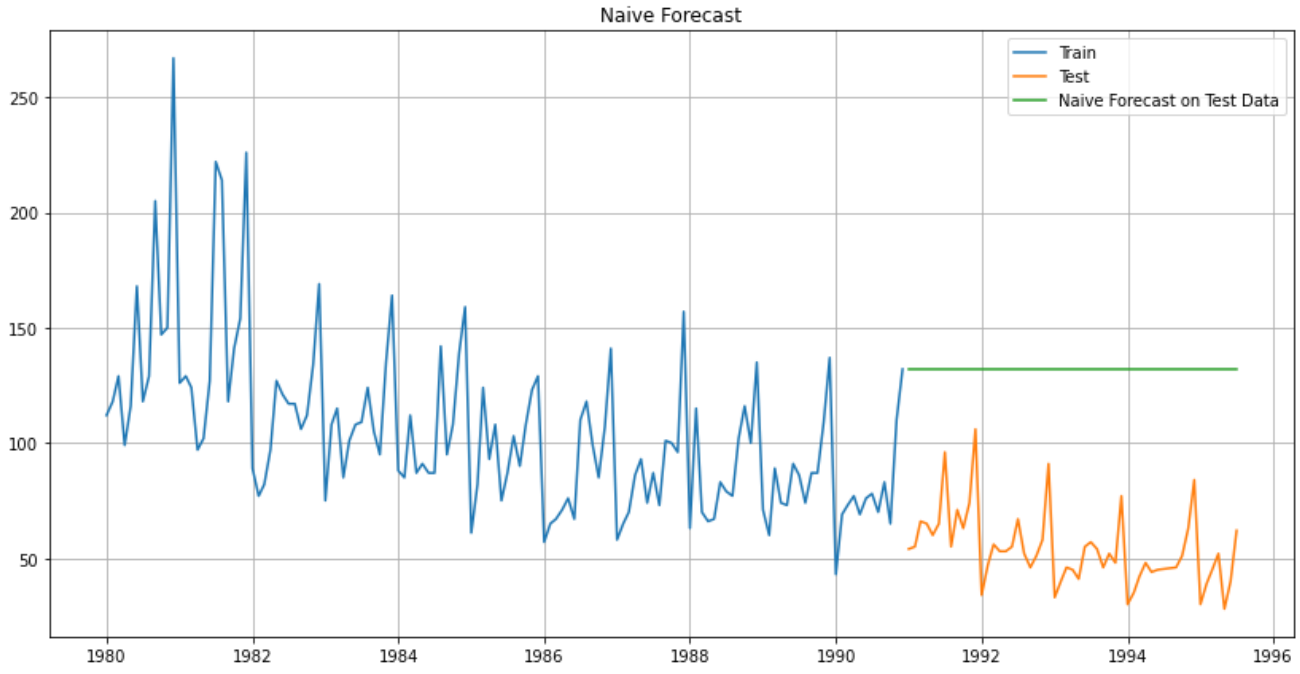


Fig. 11: Predictions of the Naïve Forecast model.

Model Evaluation:

The RMSE on the test set comes out to be 79.719

Simple Average Model

This model forecasts the average of the previous values to the future.

Predictions on the test set:

The below table gives the predicted values of the Simple Average model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	104.939394
1991-02-01	55.0	104.939394
1991-03-01	66.0	104.939394
1991-04-01	65.0	104.939394
1991-05-01	60.0	104.939394
1991-06-01	65.0	104.939394
1991-07-01	96.0	104.939394
1991-08-01	55.0	104.939394
1991-09-01	71.0	104.939394
1991-10-01	63.0	104.939394

Table 8: Predictions of the Simple Average model.

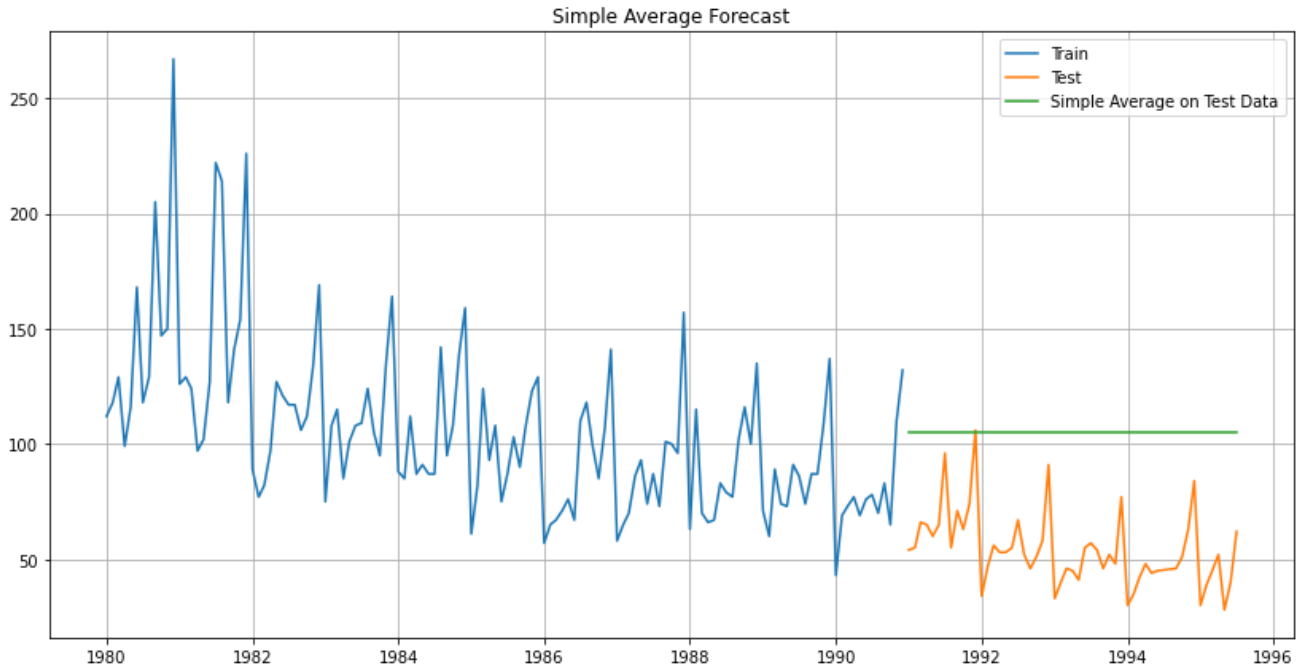


Fig. 12: Predictions of the Simple Average model.

Model Evaluation:

The RMSE on the test set comes out to be 53.461.

Moving Average Model

For the moving average model, we are going to calculate rolling means for different intervals. The best interval can be determined by the minimum error.

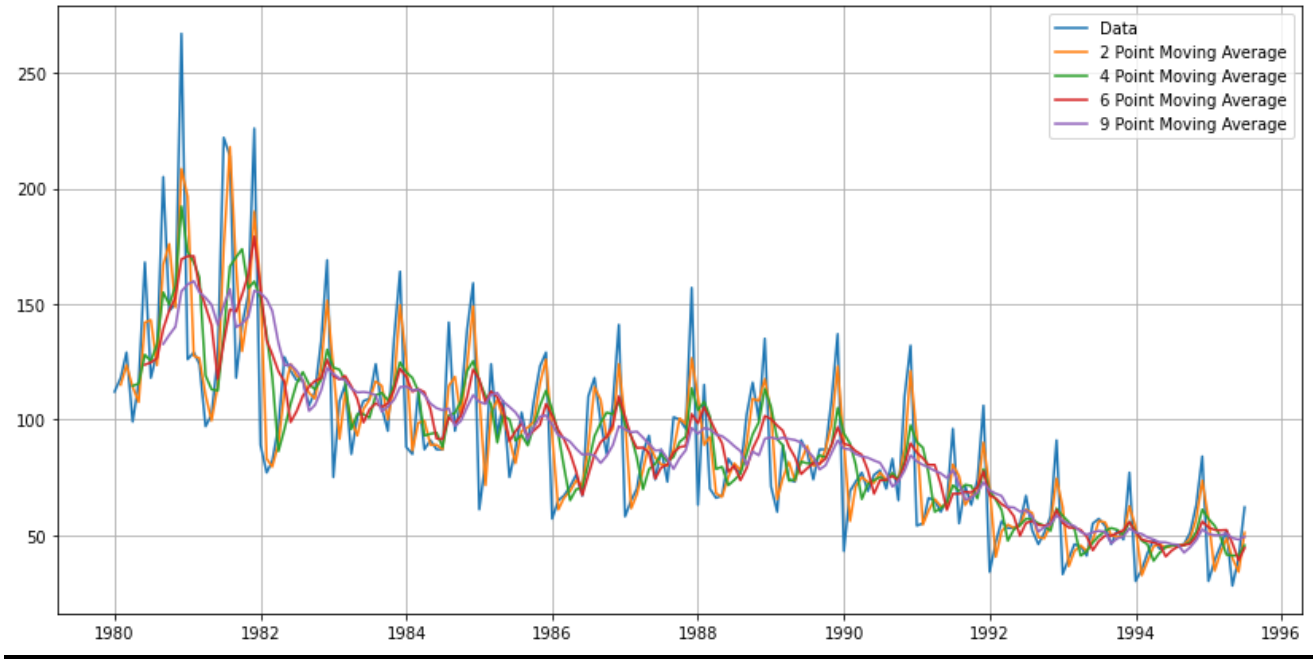


Fig. 13: Moving Average on the whole data:

Predictions on the test set:

YearMonth	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1991-01-01	54.0	93.0	90.25	85.666667	81.888889
1991-02-01	55.0	54.5	87.75	83.166667	80.333333
1991-03-01	66.0	60.5	76.75	80.333333	79.222222
1991-04-01	65.0	65.5	60.00	80.333333	77.777778
1991-05-01	60.0	62.5	61.50	72.000000	76.666667
1991-06-01	65.0	62.5	64.00	60.833333	74.666667
1991-07-01	96.0	80.5	71.50	67.833333	78.111111
1991-08-01	55.0	75.5	69.00	67.833333	72.000000
1991-09-01	71.0	63.0	71.75	68.666667	65.222222
1991-10-01	63.0	67.0	71.25	68.333333	66.222222

Table 9: Predictions of the Moving Average model.

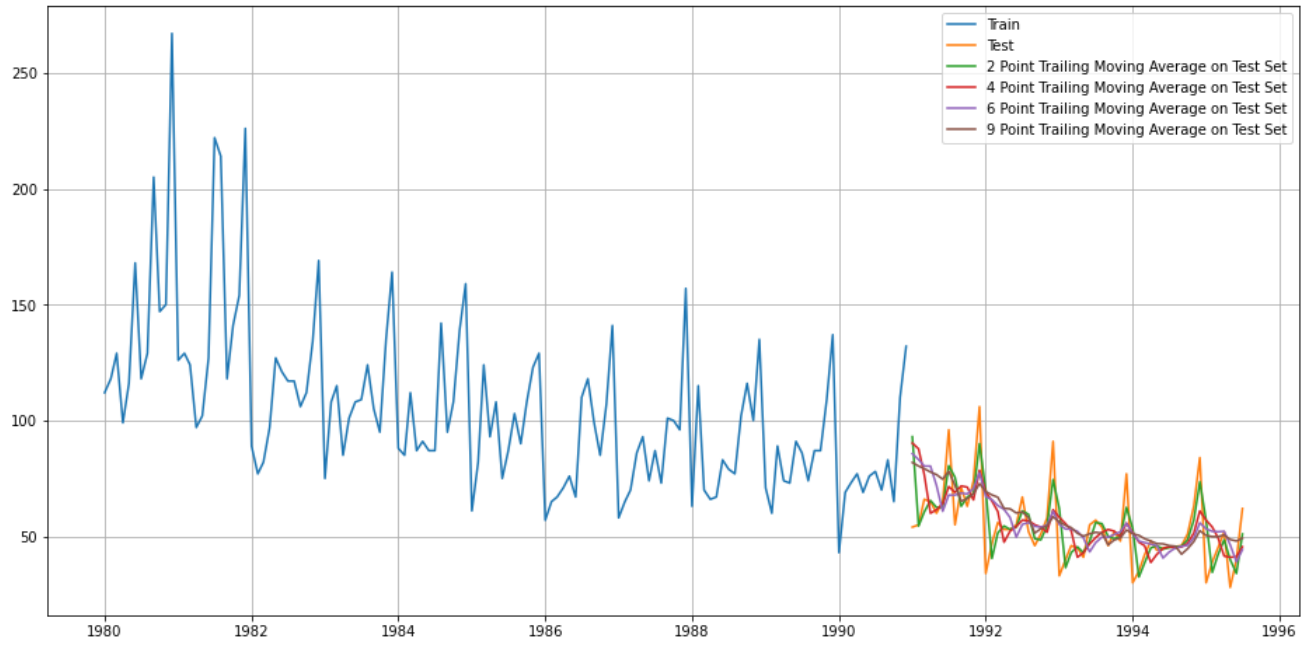


Fig. 14: Predictions of the Moving Average model.

Model Evaluation:

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

Simple Exponential Smoothing Model

The level is taken into account while building the simple exponential smoothing model. The parameter considered here is alpha.

Predictions on the test set:

The below table gives the predicted values of the Simple Average model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	87.104997
1991-02-01	55.0	87.104997
1991-03-01	66.0	87.104997
1991-04-01	65.0	87.104997
1991-05-01	60.0	87.104997
1991-06-01	65.0	87.104997
1991-07-01	96.0	87.104997
1991-08-01	55.0	87.104997
1991-09-01	71.0	87.104997
1991-10-01	63.0	87.104997

Table 10: Predictions of the Simple Exponential Smoothing model.

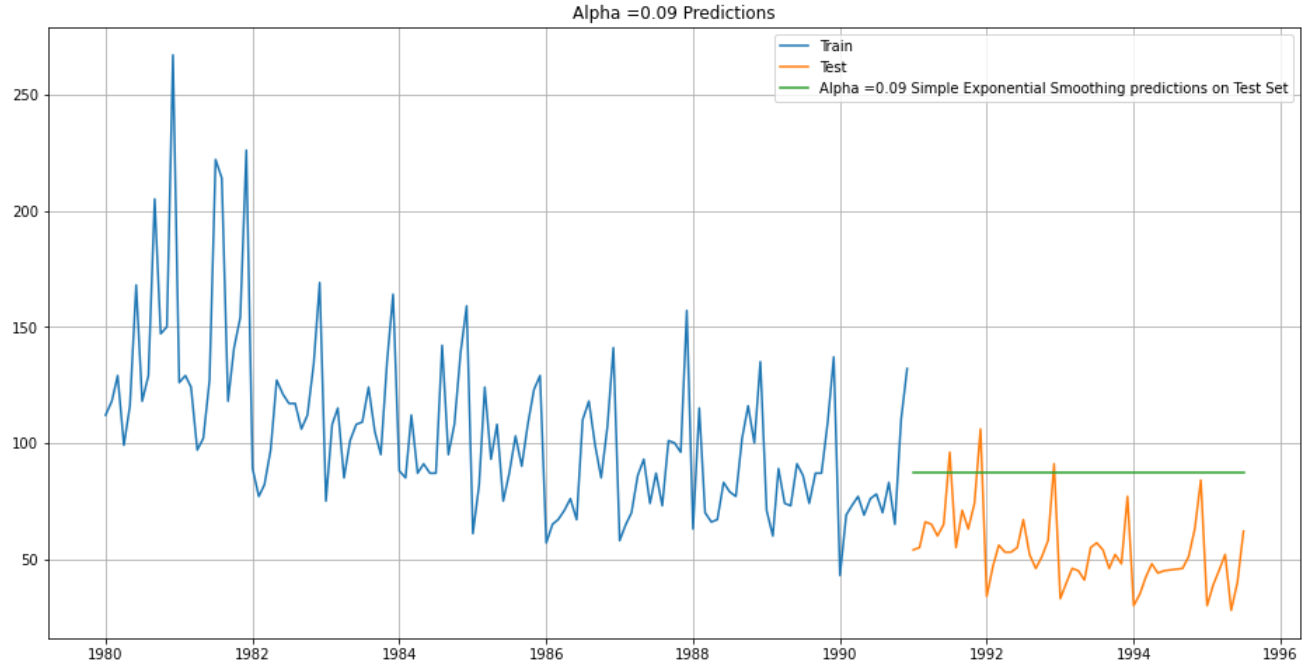


Fig. 15: Predictions of the Simple Exponential Smoothing model.

Model Evaluation:

The RMSE on the test set comes out to be 36.796.

Double Exponential Smoothing – Holt Model

The level and trend are taken into account while building the double exponential smoothing model. The parameter considered here is alpha and beta.

Predictions on the test set:

The below table gives the predicted values of the Holt model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	72.063238
1991-02-01	55.0	71.568859
1991-03-01	66.0	71.074481
1991-04-01	65.0	70.580103
1991-05-01	60.0	70.085725
1991-06-01	65.0	69.591347
1991-07-01	96.0	69.096969
1991-08-01	55.0	68.602590
1991-09-01	71.0	68.108212
1991-10-01	63.0	67.613834

Table 11: Predictions of the Holt model.

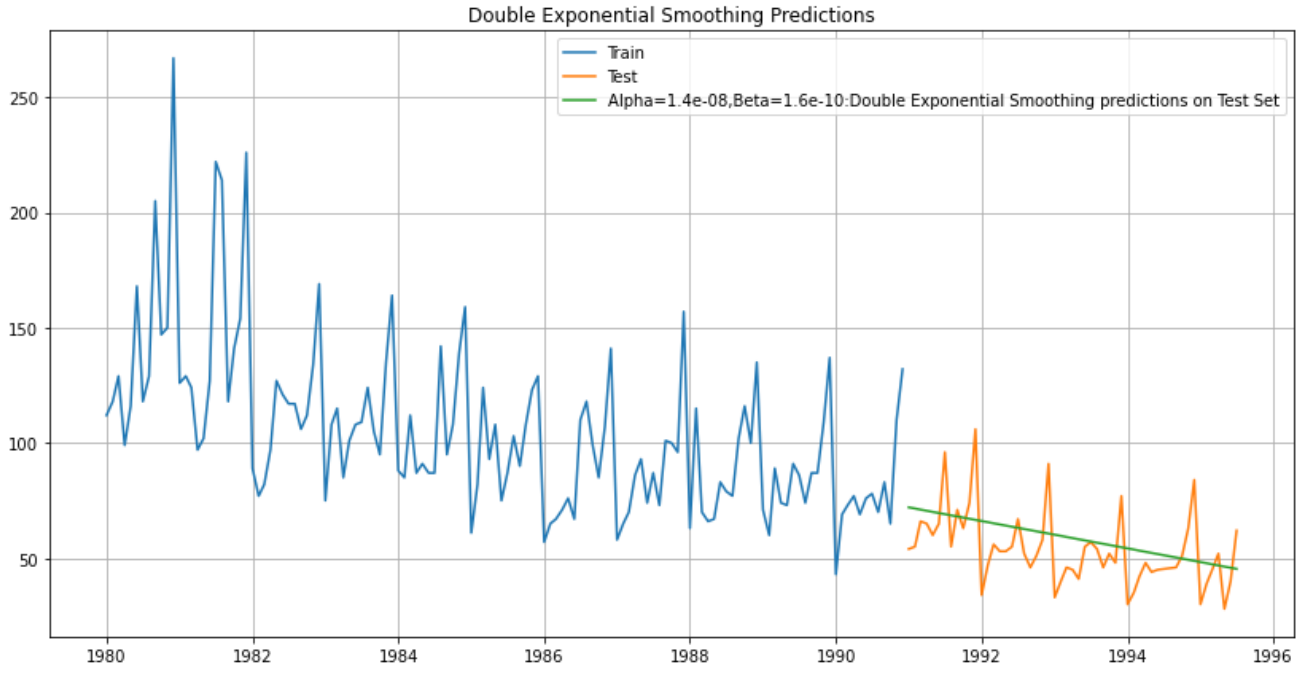


Fig. 16: Predictions of the Holt model.

Model Evaluation:

The RMSE on the test set comes out to be 15.269.

Triple Exponential Smoothing – Holt Winter Model

Multiplicative seasonality

The level, trend and seasonality are taken into account while building the triple exponential smoothing model. The parameter considered here is alpha, beta and gamma.

Predictions on the test set:

The below table gives the predicted values of the Holt Winter model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	56.321655
1991-02-01	55.0	63.664690
1991-03-01	66.0	69.374024
1991-04-01	65.0	60.435528
1991-05-01	60.0	67.758341
1991-06-01	65.0	73.546478
1991-07-01	96.0	80.630117
1991-08-01	55.0	85.541323
1991-09-01	71.0	80.707713
1991-10-01	63.0	78.764555

Table 12: Predictions of the Holt Winter model with multiplicative seasonality.

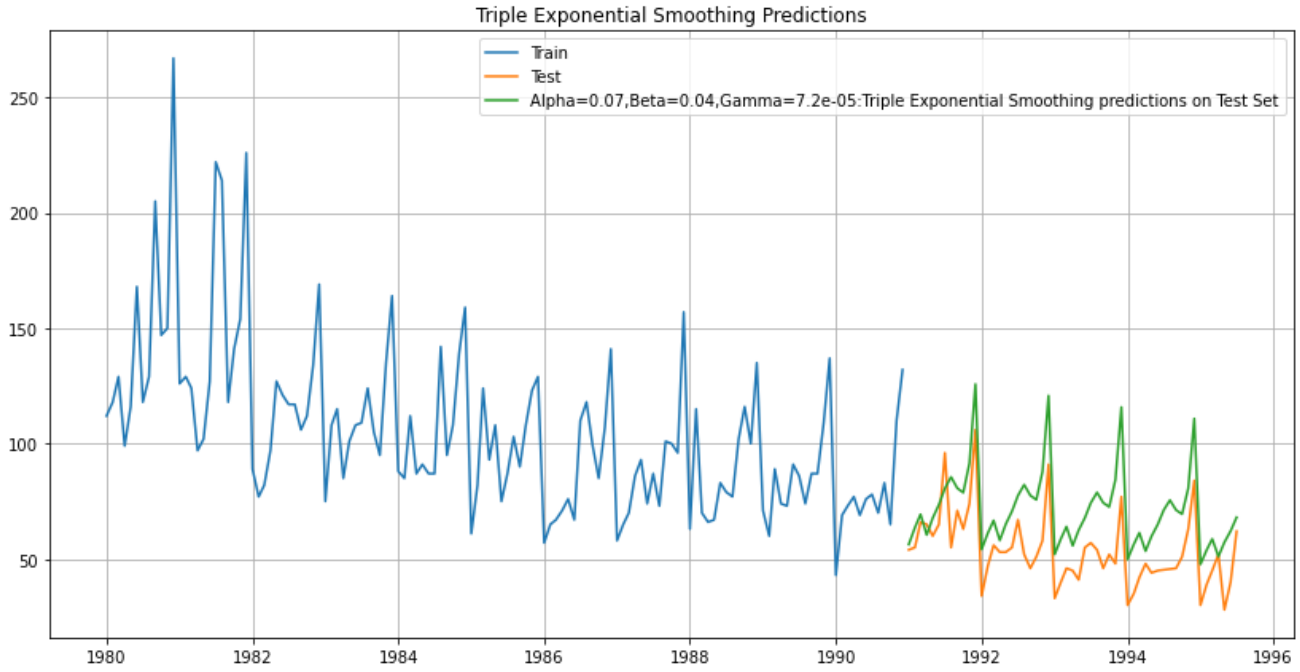


Fig. 17: Predictions of the Holt Winter model with multiplicative seasonality

Model Evaluation:

The RMSE on the test set comes out to be 20.157.

Triple Exponential Smoothing – Holt Winter Model

Additive seasonality

The level, trend and seasonality are taken into account while building the triple exponential smoothing model. The parameter considered here is alpha, beta and gamma.

Predictions on the test set:

The below table gives the predicted values of the Holt Winter model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	42.684928
1991-02-01	55.0	54.564005
1991-03-01	66.0	61.995209
1991-04-01	65.0	50.852018
1991-05-01	60.0	59.034271
1991-06-01	65.0	63.850901
1991-07-01	96.0	73.190805
1991-08-01	55.0	78.724624
1991-09-01	71.0	74.276280
1991-10-01	63.0	71.895000

Table 13: Predictions of the Holt Winter model with additive seasonality.

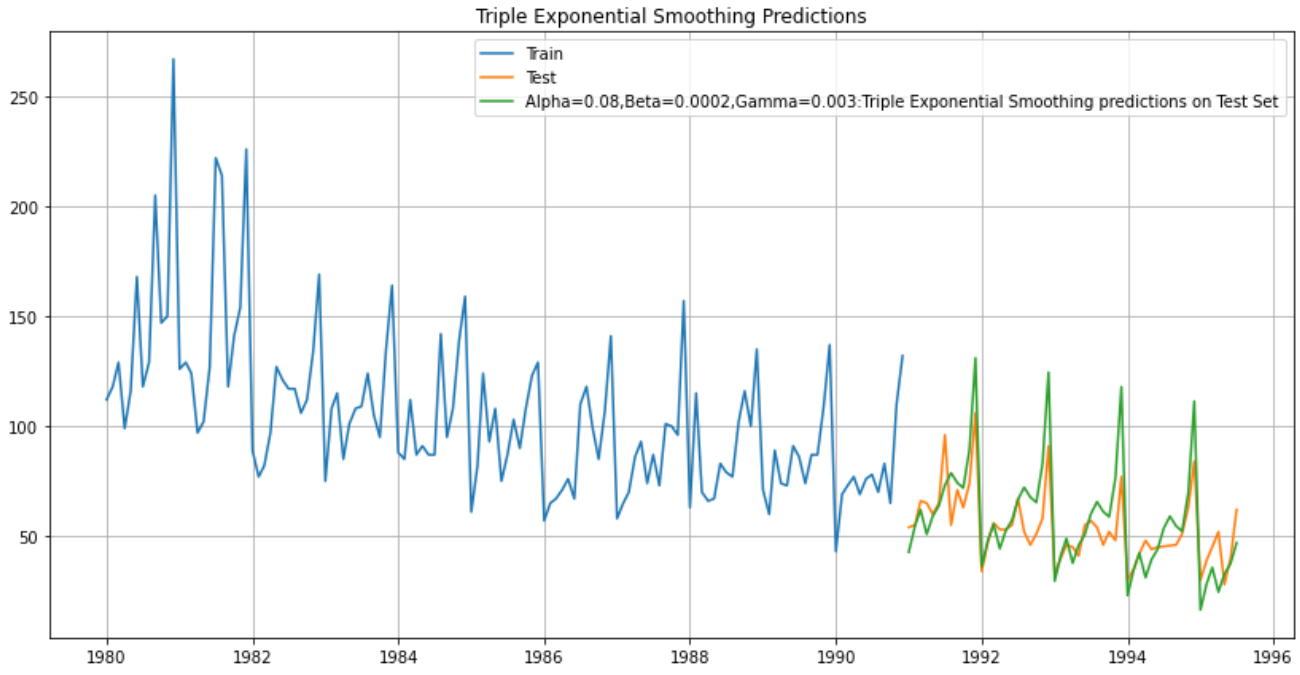


Fig. 18: Predictions of the Holt Winter model with additive seasonality

Model Evaluation:

The RMSE on the test set comes out to be 14.249.

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.

Stationarity Check:

Dickey-Fuller Test – is a statistical test on the timeseries to check for stationarity of data.

- Null Hypothesis - H_o: Time Series is non-stationary.
- Alternate Hypothesis – H_a: Time Series is stationary.

If p-value < alpha = 0.05 then null hypothesis is rejected else we fail to reject the null hypothesis.

When we run the Dickey-Fuller Test on our time series, we find the p-value to be 0.3431, which is greater than 0.05. Hence, we fail to reject the null hypothesis. That is the time series is not stationary.

To make the series stationary, we shall difference the series once, i.e., d=1. Then the time series looks as below:

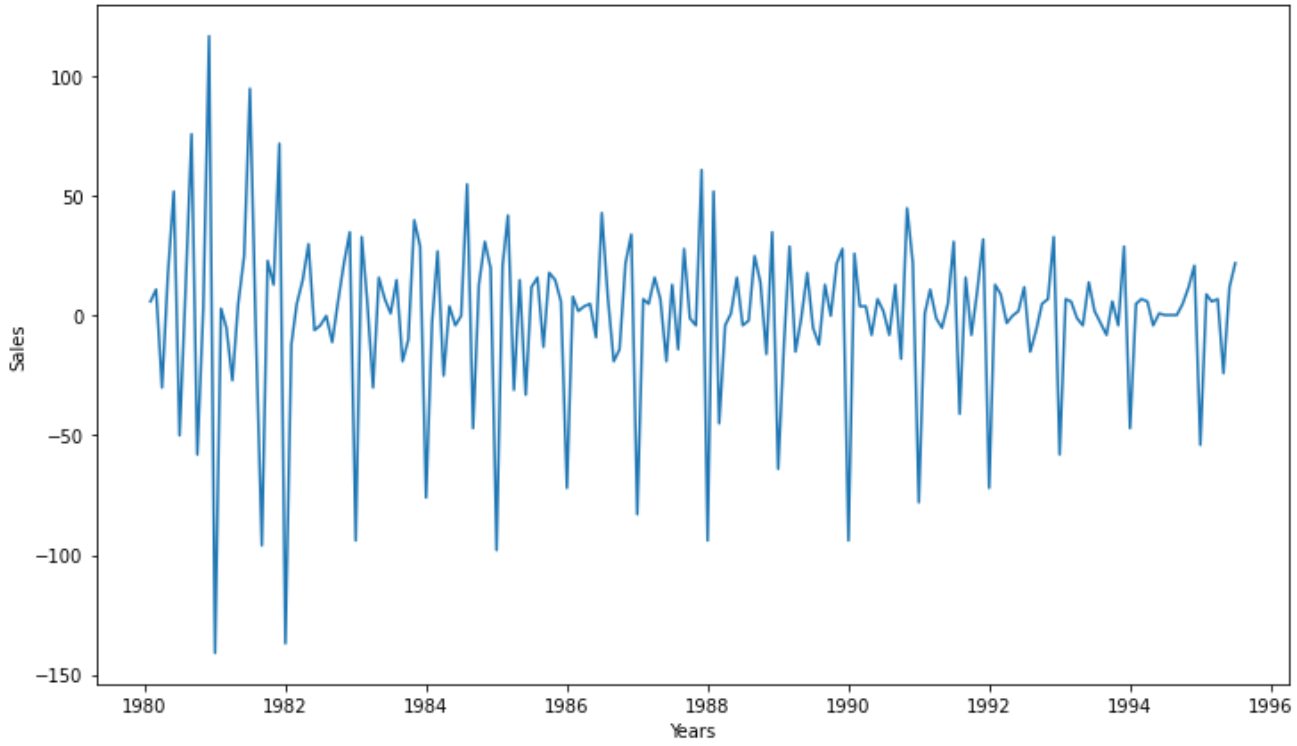


Fig. 19: First order differentiated time series.

To check if the differentiated new series is stationary or not, we run the Dickey Fuller test on the new series again and find that p-value is 0.000, which is less than 0.05. Hence, we can say that the new series is stationary.

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

The ARIMA model takes three parameters into account. The parameters are p, d and q. The parameter d depicts the order of the differencing that makes the series stationary. The parameter p is the order of the auto regression model and the parameter q is the order of the moving average model. Different combinations of p, d and q are considered and the one with the lowest AIC value is considered to build the model.

Different parameter combinations and AIC:

	param	AIC
11	(2, 1, 3)	1274.695412
15	(3, 1, 3)	1278.667917
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.57423
9	(2, 1, 1)	1281.507862
10	(2, 1, 2)	1281.870722
7	(1, 1, 3)	1281.870722
1	(0, 1, 1)	1282.309832
13	(3, 1, 1)	1282.419278
14	(3, 1, 2)	1283.720741
12	(3, 1, 0)	1297.481092
8	(2, 1, 0)	1298.611034
4	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

Table 14: Parameter combinations and AIC values.

We shall build an ARIMA model with p=2, d=1, q=3, since this gives the least AIC.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	-631.348
Date:	Thu, 03 Aug 2023	AIC	1274.695
Time:	16:45:00	BIC	1291.947
Sample:	01-01-1980 - 12-01-1990	HQIC	1281.705
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6783	0.084	-19.999	0.000	-1.843	-1.514
ar.L2	-0.7291	0.084	-8.687	0.000	-0.894	-0.565
ma.L1	1.0446	0.618	1.691	0.091	-0.166	2.255
ma.L2	-0.7720	0.132	-5.858	0.000	-1.030	-0.514
ma.L3	-0.9045	0.560	-1.616	0.106	-2.002	0.192
sigma2	860.3101	519.823	1.655	0.098	-158.525	1879.145

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.51
Prob(Q):	0.87	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	4.57

Table 15: Summary of the ARIMA model.

Predictions on the test set:

The below table gives the predicted values of the ARIMA model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	85.595789
1991-02-01	55.0	90.535998
1991-03-01	66.0	81.967217
1991-04-01	65.0	92.746555
1991-05-01	60.0	80.902653
1991-06-01	65.0	92.921603
1991-07-01	96.0	81.384994
1991-08-01	55.0	91.984470
1991-09-01	71.0	82.606135
1991-10-01	63.0	90.618245

Table 16: Predictions of the ARIMA model.

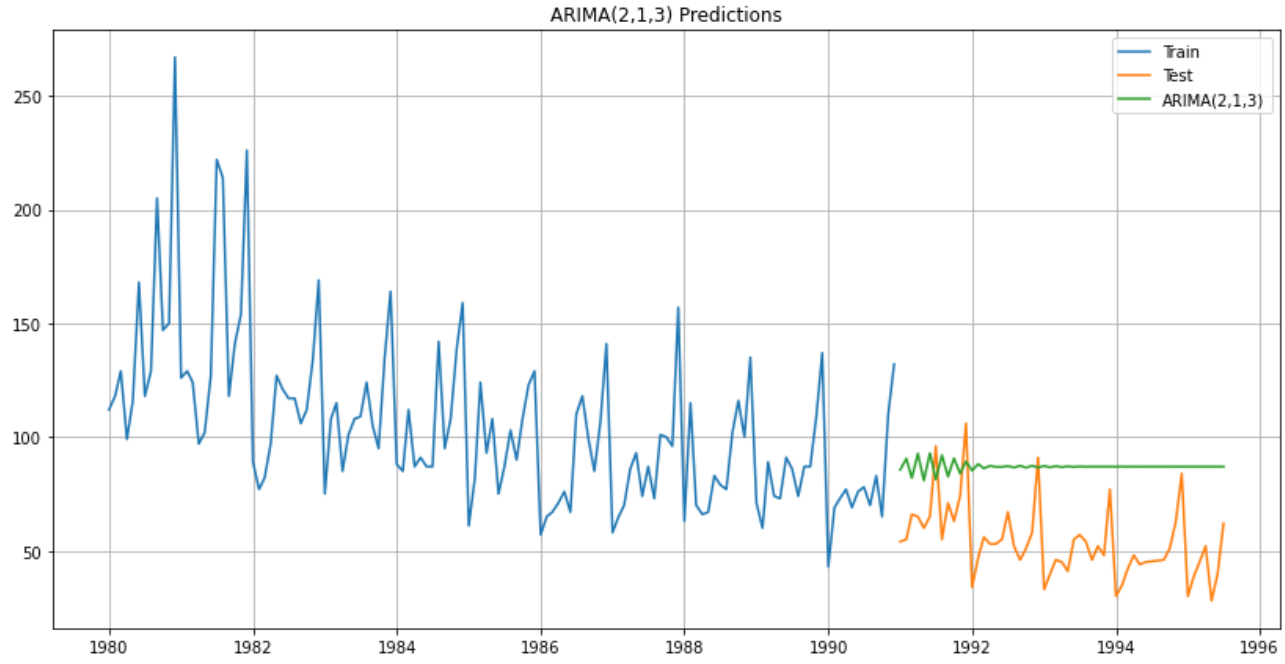


Fig. 20: Predictions of the ARIMA model.

Plot Diagnostics:

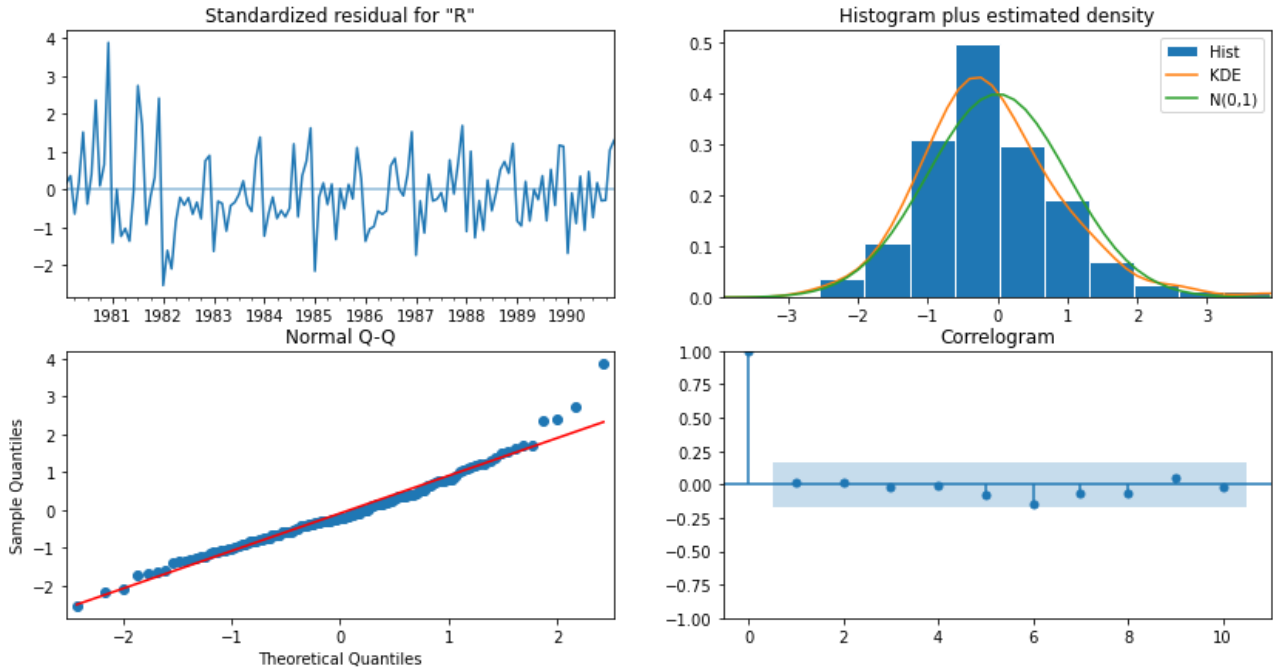


Fig. 21: Plot diagnostics of the ARIMA model.

Model Evaluation:

The RMSE on the test set comes out to be 36.812.

SARIMA

The SARIMA model, in addition to the p, d, q parameters, it takes into account the seasonal parameters – P, D, Q and S. The parameter S is determined by looking at the auto correlation plot.

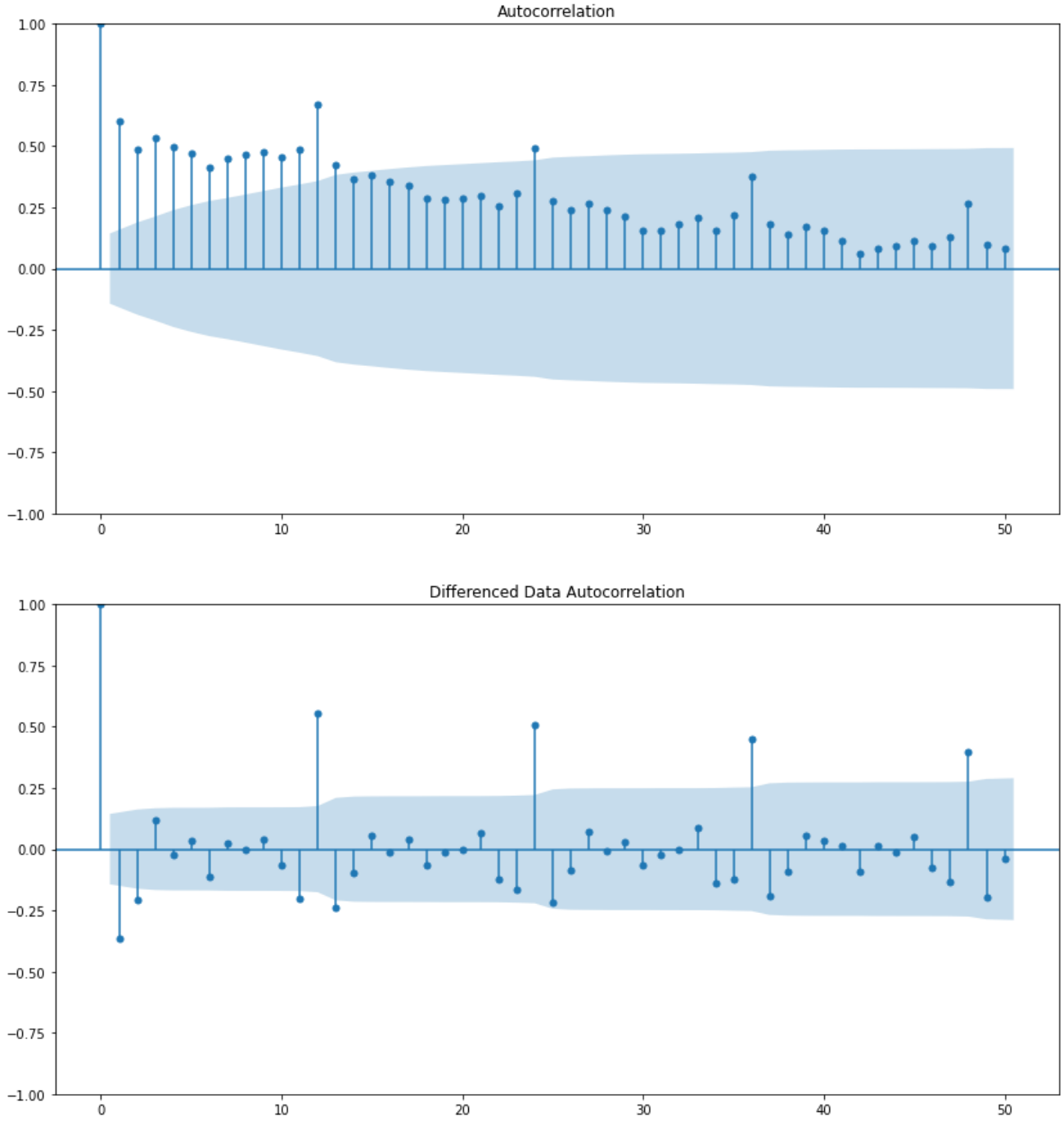


Fig. 22: Auto correlation of the original and differenced series.

It is clear that the seasonal component is 12 as we see a spike in every 12th data point.

Different parameter combinations and AIC:

	param	seasonal	AIC
222	(3, 1, 1)	(3, 0, 2, 12)	774.400285
238	(3, 1, 2)	(3, 0, 2, 12)	774.880934
220	(3, 1, 1)	(3, 0, 0, 12)	775.426699
221	(3, 1, 1)	(3, 0, 1, 12)	775.49533
252	(3, 1, 3)	(3, 0, 0, 12)	775.561018

Table 17: Parameter combinations and AIC values.

We shall build a SARIMA model with p=3, d=1, q=1, P=3, D=0, Q=2 and S=12 since this gives the least AIC.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)	Log Likelihood	-377.519
Date:	Wed, 09 Aug 2023	AIC	775.038
Time:	13:30:10	BIC	800.256
Sample:	01-01-1980 - 12-01-1990	HQIC	785.216
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0785	0.123	0.638	0.524	-0.163	0.320
ar.L2	0.0123	0.112	0.109	0.913	-0.208	0.233
ar.L3	-0.1546	0.104	-1.481	0.138	-0.359	0.050
ma.L1	-0.9967	0.529	-1.882	0.060	-2.034	0.041
ar.S.L12	0.7871	0.161	4.881	0.000	0.471	1.103
ar.S.L24	0.0815	0.162	0.502	0.616	-0.237	0.400
ar.S.L36	0.0477	0.099	0.484	0.628	-0.146	0.241
ma.S.L12	-0.5003	0.261	-1.919	0.055	-1.011	0.011
ma.S.L24	-0.2216	0.212	-1.043	0.297	-0.638	0.195
sigma2	187.6323	101.415	1.850	0.064	-11.137	386.402

Ljung-Box (L1) (Q):	0.26	Jarque-Bera (JB):	1.70
Prob(Q):	0.61	Prob(JB):	0.43
Heteroskedasticity (H):	1.14	Skew:	0.33
Prob(H) (two-sided):	0.72	Kurtosis:	3.04

Table 18: Summary of the SARIMA model.

Predictions on the test set:

The below table gives the predicted values of the SARIMA model on the first ten data points of the test set.

YearMonth	Actual Values	Prediction
1991-01-01	54.0	54.877546
1991-02-01	55.0	67.562759
1991-03-01	66.0	67.331698
1991-04-01	65.0	66.671740
1991-05-01	60.0	69.091041
1991-06-01	65.0	69.986101
1991-07-01	96.0	74.576856
1991-08-01	55.0	75.349420
1991-09-01	71.0	77.995322
1991-10-01	63.0	74.975432

Table 19: Predictions of the SARIMA model.

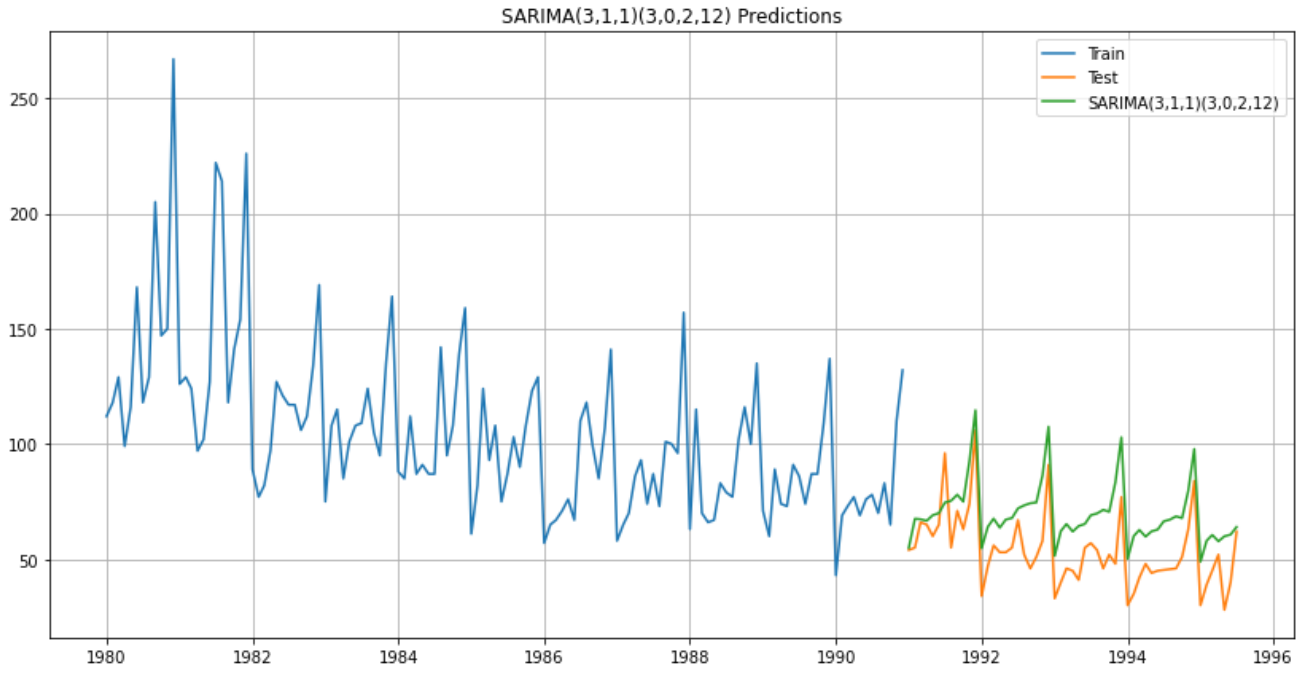


Fig. 23: Predictions of the SARIMA model.

Plot Diagnostics:

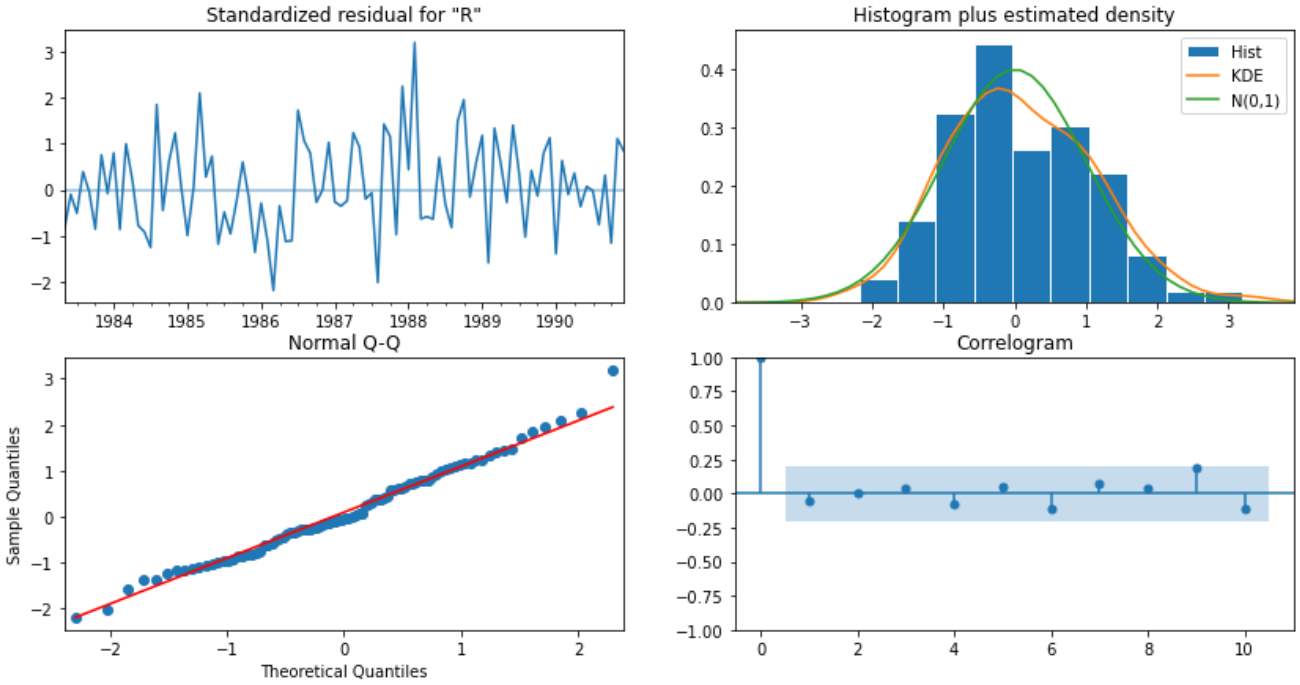


Fig. 24: Plot diagnostics of the SARIMA model.

Model Evaluation:

The RMSE on the test set comes out to be 18.259

7. 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.

	Parameters	Test RMSE
Moving Average -1	2 point	11.529278
TES(A,A,A)	Alpha=0.08, Beta=0.0002, Gamma=0.003	14.249661
Moving Average -2	4 point	14.451403
Moving Average-3	6 point	14.566327
Moving Average-3	9 point	14.727630
DES	Alpha=1.4e-08, Beta=1.6e-10	15.268944
Linear Regression		15.269000
SARIMA	(3,1,1)(3,0,2,12)	18.259388
TES(A,A,M)	Alpha=0.07,Beta=0.04,Gamma=7.2e-05	20.156763
SES	Alpha = 0.09	36.796241
ARIMA	(0,1,2)	36.812984
Simple Average Model		53.460570
Naive Model		79.718773

Table 20: Different models and their RMSE values on test data

8. 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.

From the above table, we see that the triple exponential models have the lowest RMSE scores. Hence, we shall build these models on the whole data to forecast 12 months into future with 95% confidence interval.

Predictions of Triple Exponential Smoothing – with Additive Seasonality:

	lower_CI	prediction	upper_CI
1995-08-01	14.790479	49.559654	84.328828
1995-09-01	11.635608	46.404783	81.173957
1995-10-01	10.374718	45.143892	79.913067
1995-11-01	24.977411	59.746585	94.515760
1995-12-01	63.269027	98.038201	132.807376
1996-01-01	-21.196473	13.572702	48.341876
1996-02-01	-10.898772	23.870402	58.639577
1996-03-01	-3.350807	31.418367	66.187542
1996-04-01	-10.560022	24.209153	58.978328
1996-05-01	-7.214159	27.555016	62.324190
1996-06-01	-1.718932	33.050243	67.819417
1996-07-01	8.880321	43.649495	78.418670

Table 21: Future Prediction with 95% confidence interval from TES (A,A,A) Model.

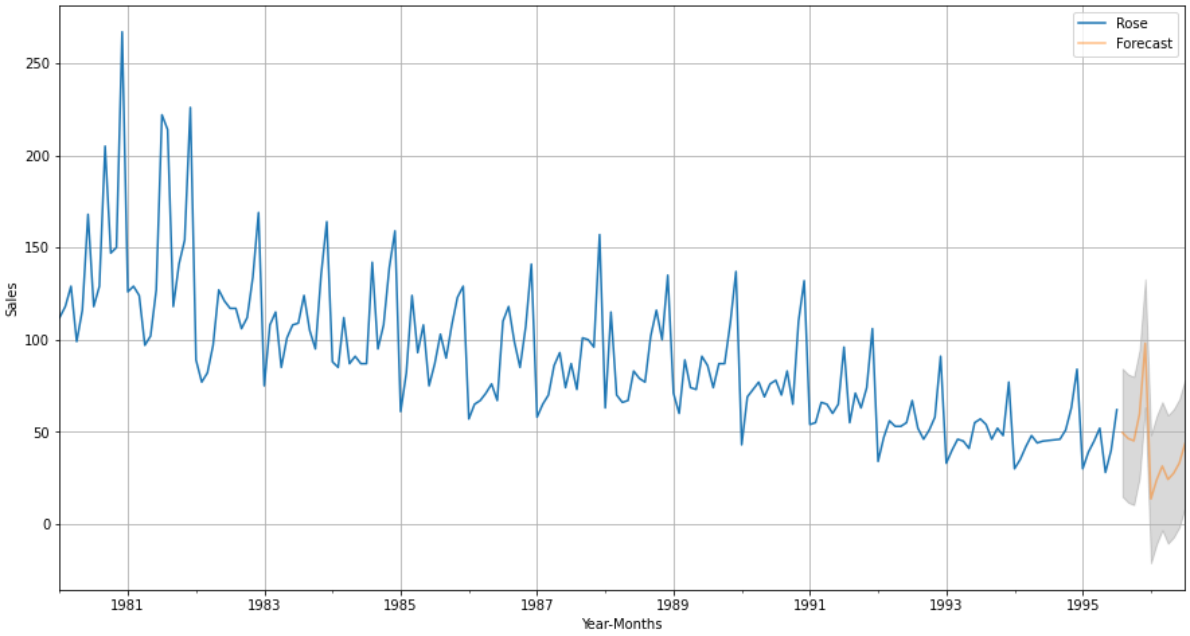


Fig. 25: Future Prediction with 95% confidence interval from TES (A,A,A) Model.

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

Inference:

- Triple Exponential Smoothing model is used to forecast into the future for next 12 months.
- The forecast tells that the sale will be more in the month of November and December 1995.
- The company should make sure the stocks are not emptied in these months.
- The forecast projects a downward trend for the sale of this wine.
- The company should come up with a plan to increase the sales.
- If the company does not intervene, there is a danger that this wine will disappear from the market.