PROJECT

TIME SERIES FORECASTING

Table of contents

List of Figures

List of Tables

PROBLEM 1

Problem Statement

For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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 analyze and forecast Wine Sales in the 20th century.

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

Sample of the Time Series

	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
5	1980-06	168.0
6	1980-07	118.0
7	1980-08	129.0
8	1980-09	205.0
9	1980-10	147.0

Table 1. Sample of the Time Series.

Basic Information of the Dataset

```
<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
```

Data Types of Variables

	Data Type
YearMonth	object
Rose	float64

	Data Type
YearMonth	datetime64[ns]
Rose	float64

Table 2. Data Types of All Features in the Dataset.

In original data, Year-Month columns are of object datatype. It is converted into a time series stamp and the same is shown in the above table.

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0
1980-06-01	168.0
1980-07-01	118.0
1980-08-01	129.0
1980-09-01	205.0
1980-10-01	147.0

Table 3. Sample of the Time Series after Creating Timestamp.

Insights

1. There are 2 features (columns) with 187 observations (rows) in the dataset.
2. The dataset has one numerical feature i.e., It indicates the number of Rose wine sales. This is the forecast variable.
3. Another feature is the year-month timestamp and it is made as an index for this time series.

Checking for Null Values

- There are two null values in the given time series.
- As there is seasonality in the dataset, null values are filled with the average of the last two years respective months.

Ex: Sales of July-1994 is average sales of July-1993 and July-1992

Sales of August-1994 is average sales of August-1993 and August-1992

YearMonth	Rose
1994-07-01	NaN
1994-08-01	NaN

YearMonth	Rose
1994-06-01	45.0
1994-07-01	62.0
1994-08-01	53.0
1994-09-01	46.0

Description of the Dataset

Five number summaries

	count	mean	std	min	25%	50%	75%	max
Rose	187.0	90.042781	39.114366	28.0	62.5	85.0	111.0	267.0

Table 4. Five Number Summary.

Insights:

1. The Minimum number of sales in a month is 28.
2. The Maximum number of sales in a month is 267.
3. Mean sales per month is 90.04.

Plot the Time Series to understand the behavior of the data

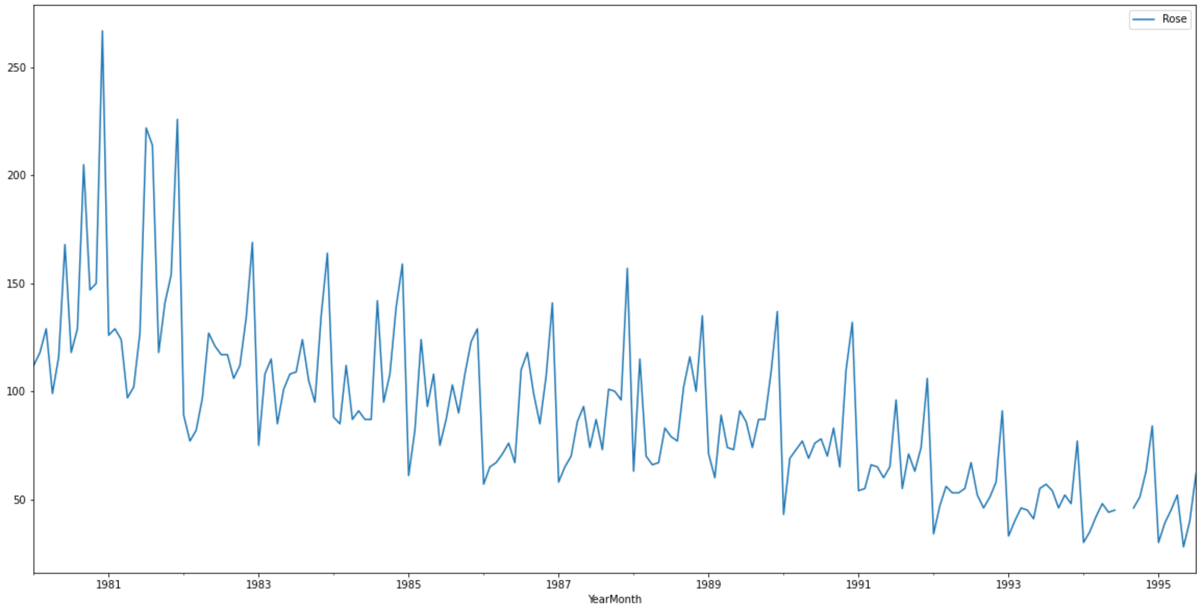


Figure 1. Time Series Plot.

- From the above plot, it can be noticed that wine sales **data has both trend and seasonality**.

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

EXPLORATORY DATA ANALYSIS

Yearly Sales Analysis - Box Plot

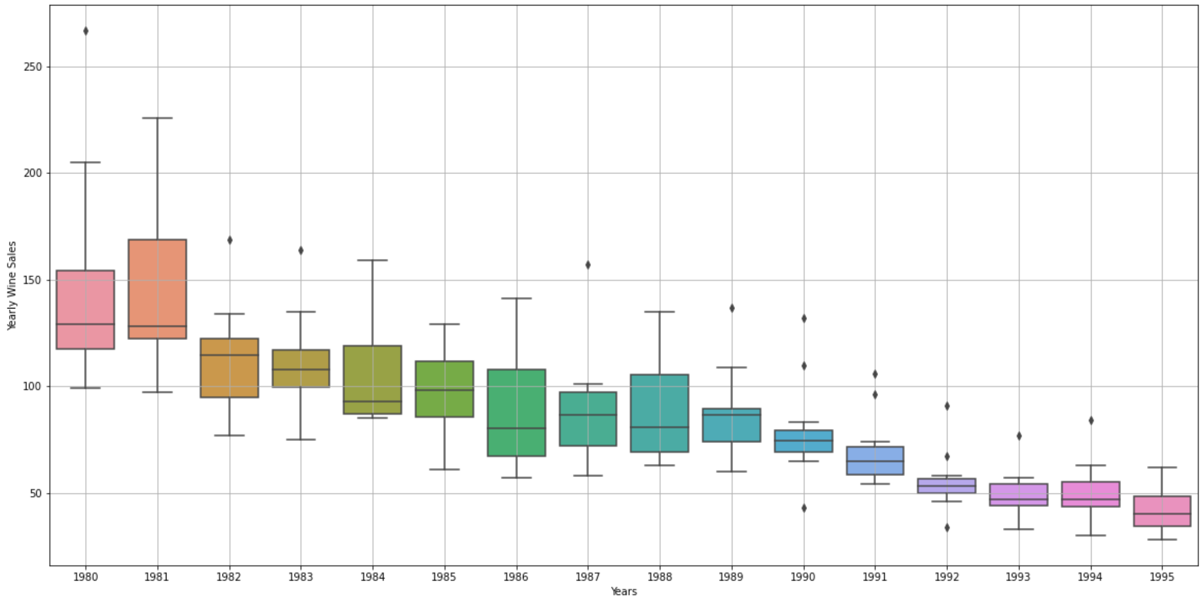


Figure 2. Box Plot of Yearly Sales.

- As we got to know from the Time Series plot, the boxplots over here also indicate a measure of the trend is present. Also, we see that the sales of wine have some outliers for certain years.

Yearly Mean Sales

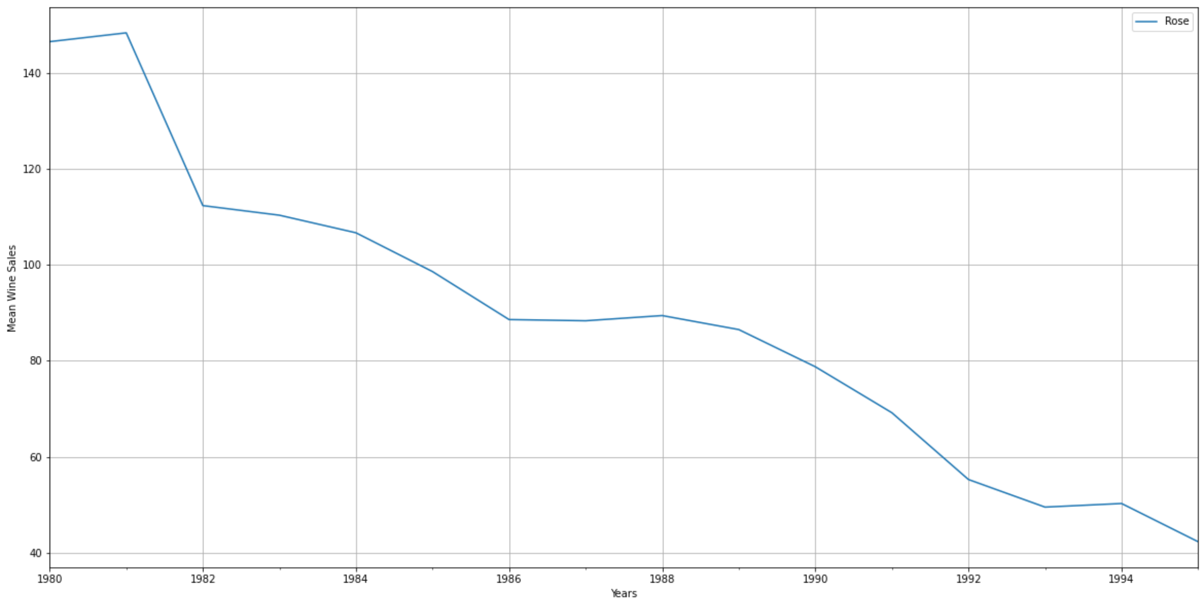


Figure 3. Yearly Mean Sales.

Insights

- Mean wine sales in a year are decreasing continuously.

Yearly Total Sales

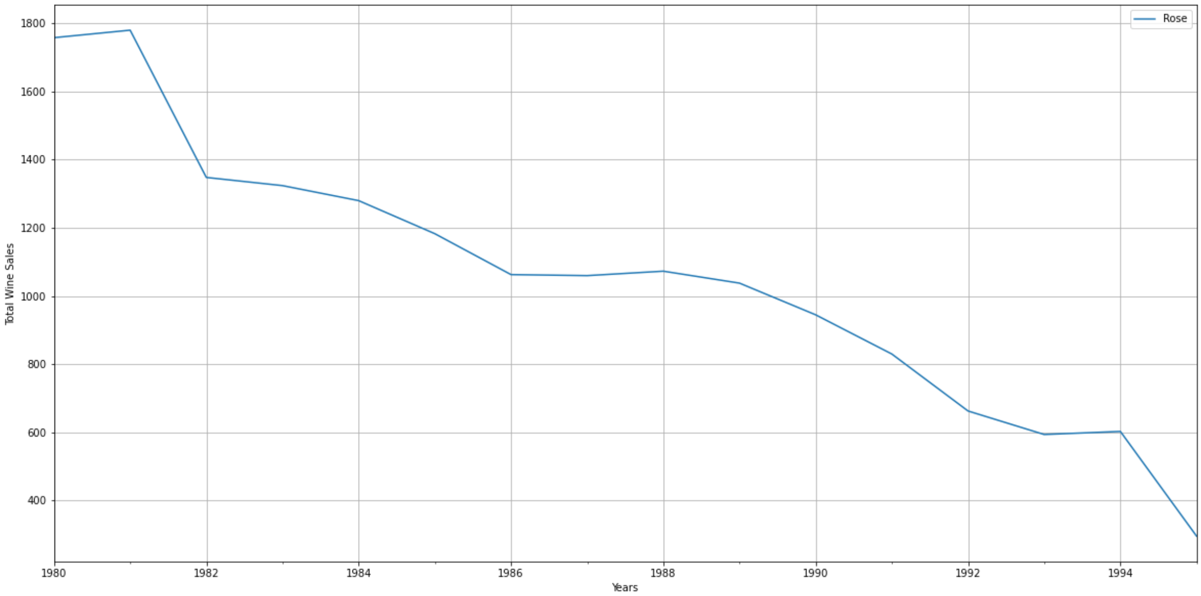


Figure 4. Yearly Total Wine Sales.

Year	Mean Sales	Total Sales
1980	146.5	1758.0
1981	148.3	1780.0
1982	112.3	1348.0
1983	110.3	1324.0
1984	106.7	1280.0
1985	98.6	1183.0
1986	88.6	1063.0
1987	88.3	1060.0
1988	89.4	1073.0
1989	86.5	1038.0
1990	78.8	945.0
1991	69.2	830.0
1992	55.2	663.0
1993	49.5	594.0
1994	50.2	603.0
1995	42.3	296.0

Table 5. Yearly Mean and Total Wine Sales.

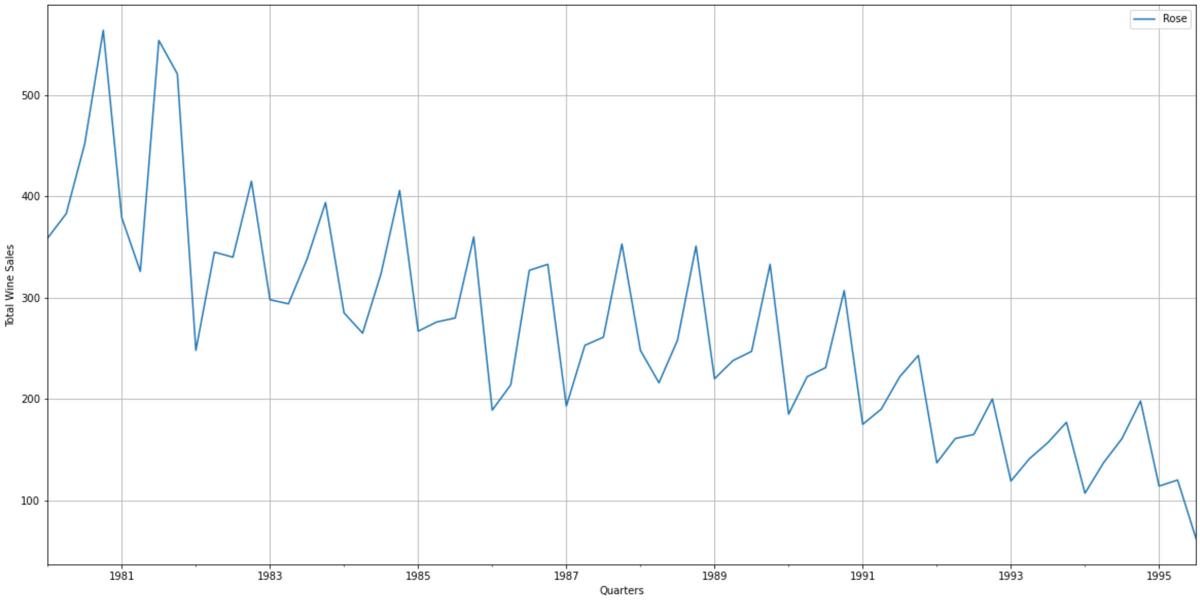
Insights

- Total wine sales in a year are following almost the same pattern as that of the mean wine sales in a year.
- Total wine sales in a year are decreasing continuously.

Quarterly Sales Analysis

	Rose
YearMonth	
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

Table 6. Quarterly Resampled Time Series.



. Figure 5. Quarterly Resampled Time Series Plot.

Box Plot of Quarterly Sales

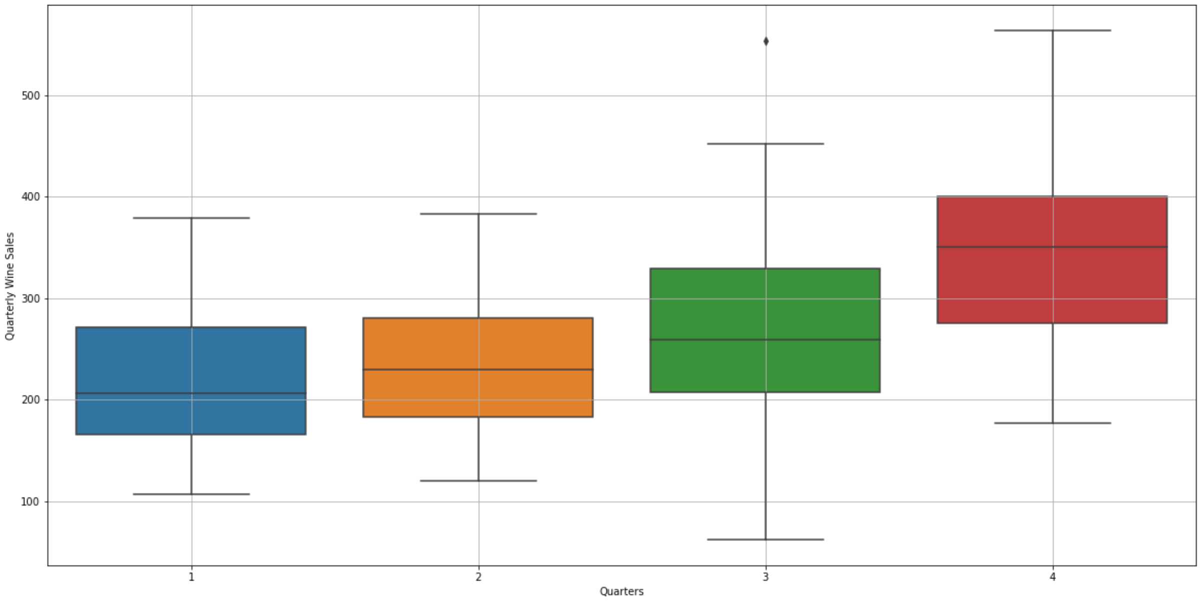


Figure 6. Box Plot of Quarterly Sales.

Insights

- Sales are increasing gradually from quarter1 to quarter 4.

Quarterly Mean Sales

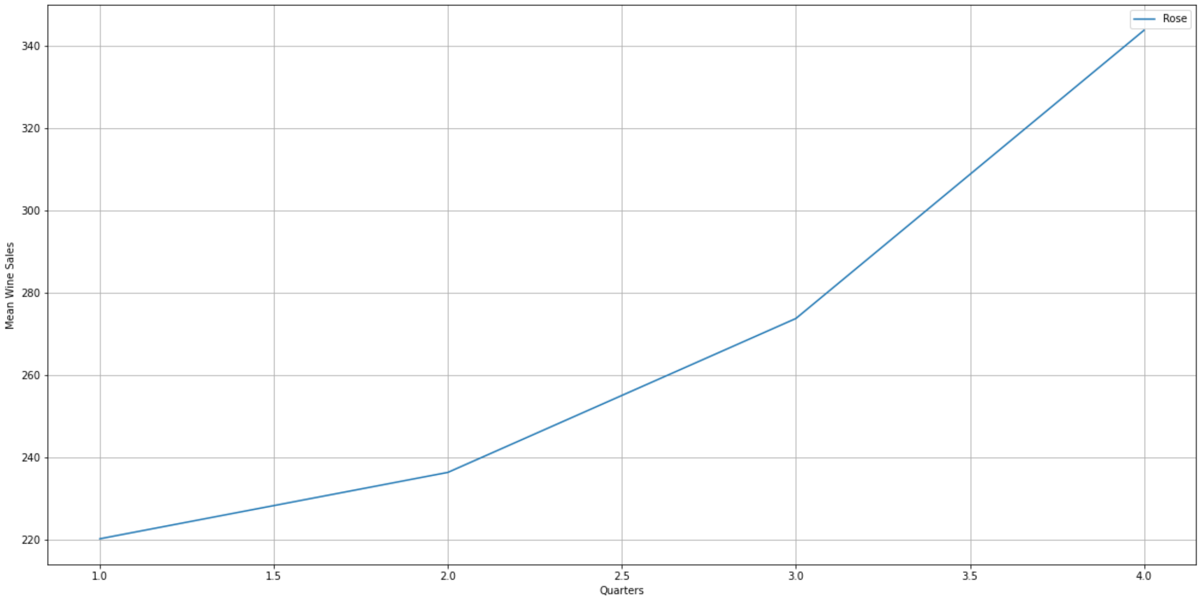


Figure 7. Quarterly Mean Sales.

Quarterly Total Sales

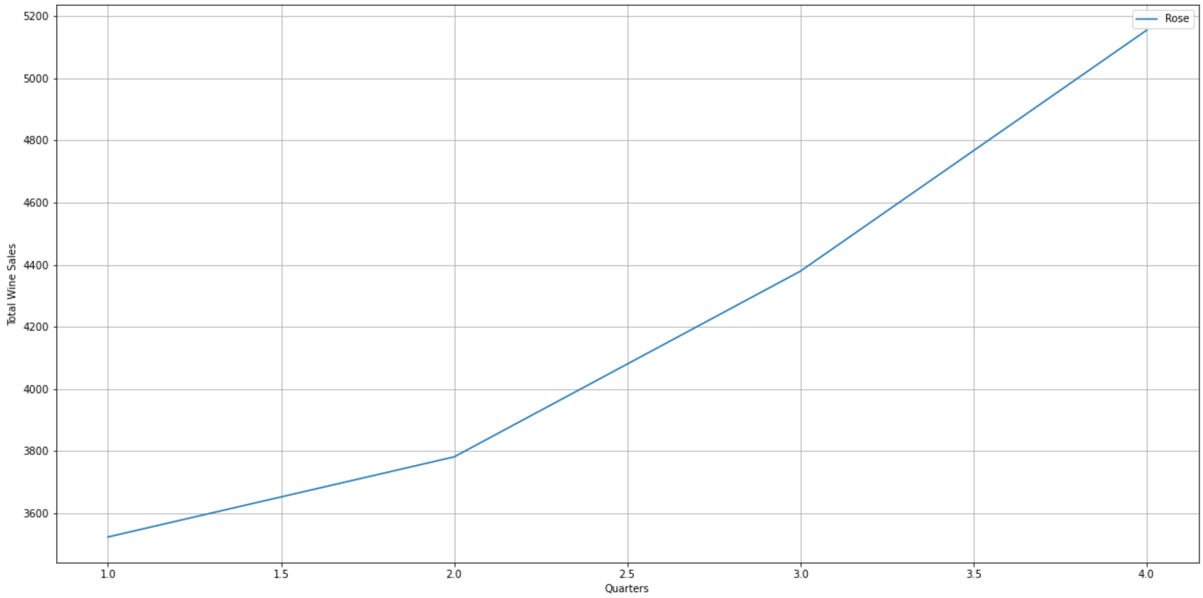


Figure 8. Quarterly Total Wine Sales.

	Mean Sales	Total Sales
Quarter		
1	220.2	3523.0
2	236.3	3781.0
3	273.7	4379.0
4	343.7	5155.0

Table 7. Quarterly Mean and Total Sales.

Insights

- Mean sales are increasing gradually from quarter1 to quarter 4.
- Total wine sales in a quarter are following almost the same pattern as that of the mean wine sales in a quarter.
- Total sales are increasing gradually from quarter1 to quarter 4

Monthly Sales Analysis

Box Plot of Monthly Sales

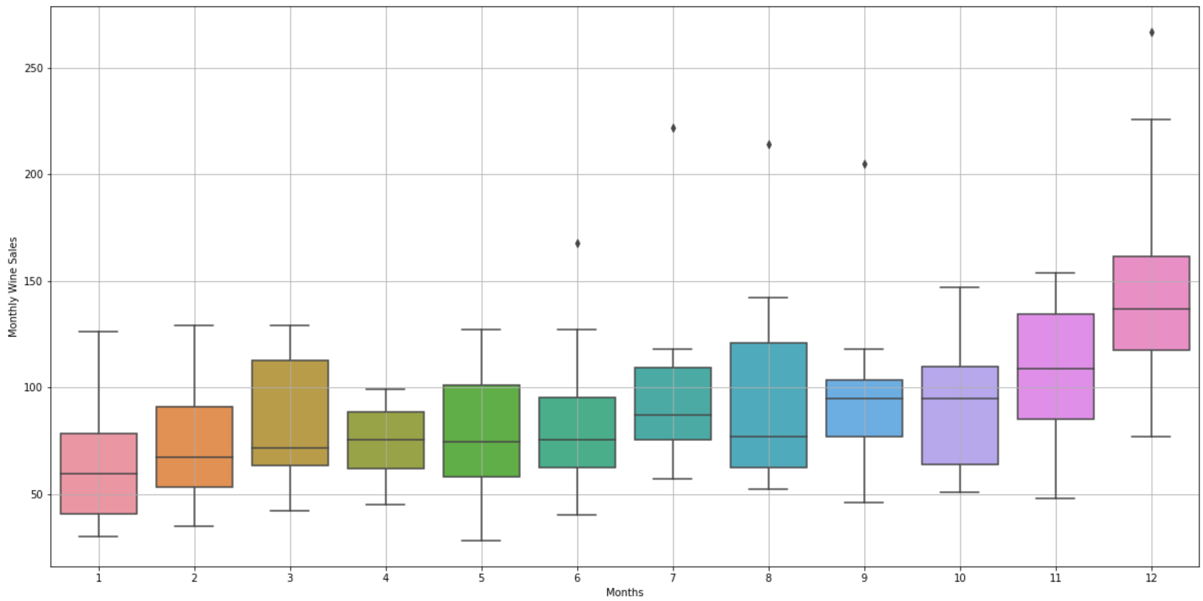


Figure 9. Box Plot of Monthly Sales.

Insights

- Sales are increasing gradually from January to December.

Monthly Mean Sales

- Mean sales in a month are increasing gradually from January to December.

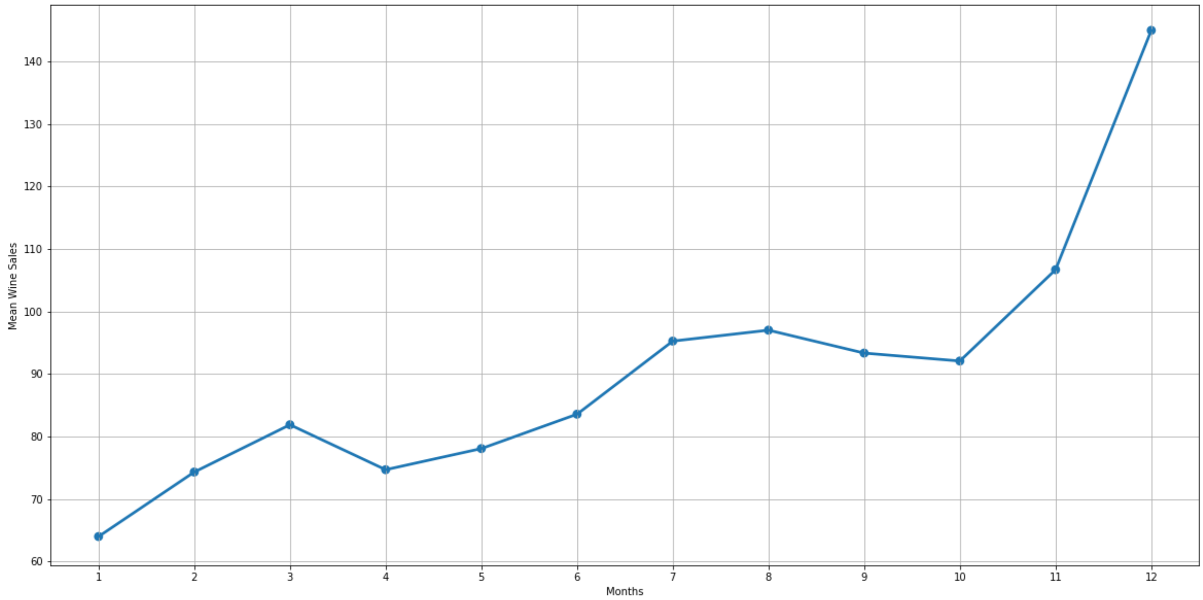


Figure 10. Monthly Mean Sales.

Monthly Total Sales

- Total sales in a month are following almost the same pattern as that of the mean sales in a month.
- Total sales in a month are increasing gradually from January to December.

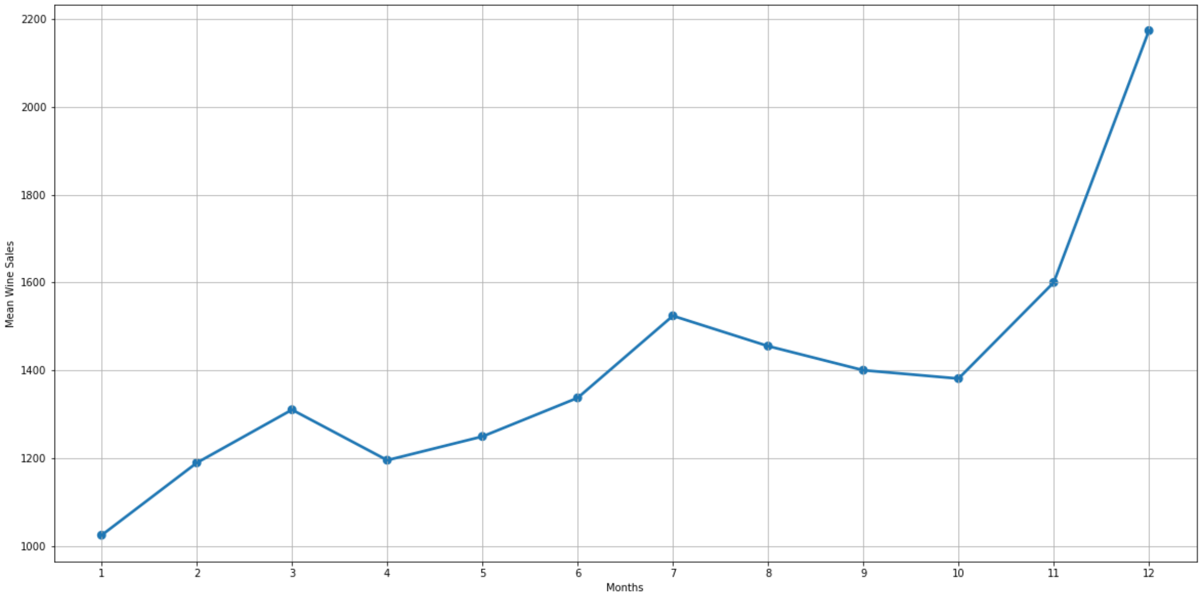


Figure 11. Monthly Total Sales.

Month	Mean Sales	Total Sales
1	64.0	1024.0
2	74.3	1189.0
3	81.9	1310.0
4	74.7	1195.0
5	78.1	1249.0
6	83.6	1337.0
7	95.2	1524.0
8	97.0	1455.0
9	93.3	1400.0
10	92.1	1381.0
11	106.7	1600.0
12	144.9	2174.0

Table 8. Monthly Mean and Total Sales.

Monthly Time Series

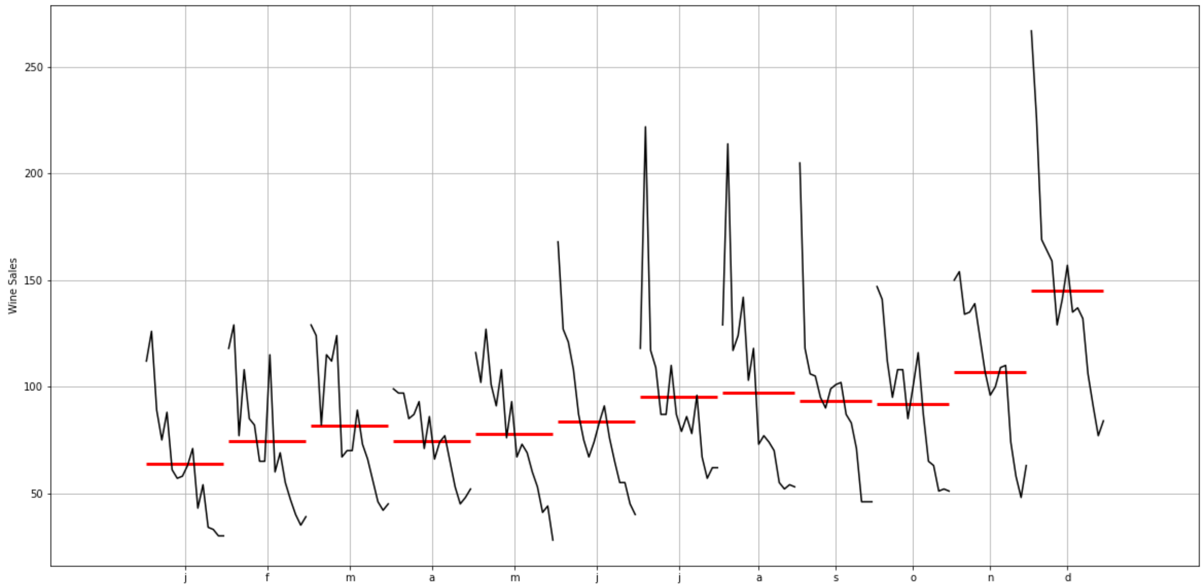


Figure 12. Monthly Time Series.

- Sales are increasing gradually from January to December.

Yearly Sales across Months

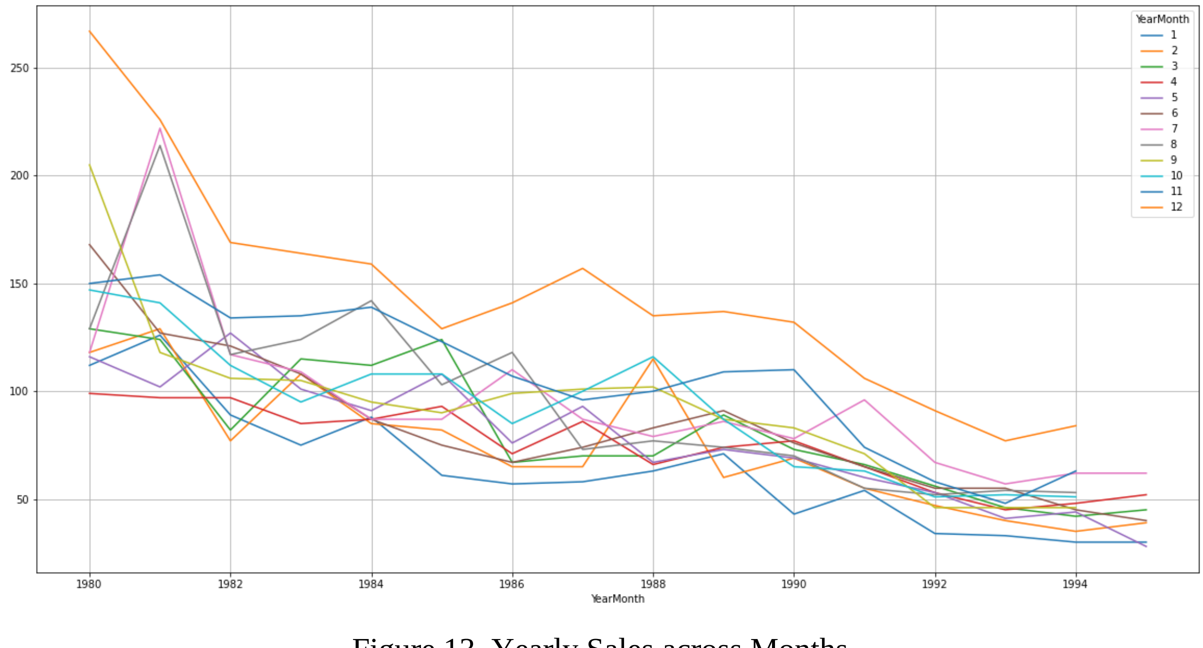


Figure 13. Yearly Sales across Months.

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
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	62.0	53.0	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 9. Yearly Sales across Months.

Empirical Cumulative Distribution of Sales

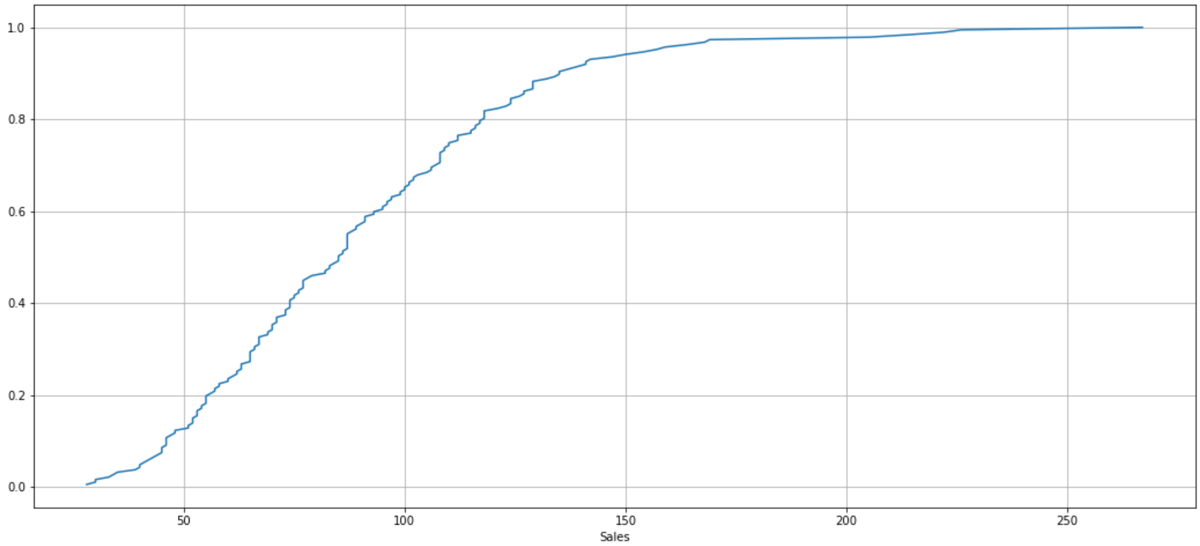


Figure 14. Empirical Cumulative Distribution of Sales.

- From the above plot, the probability to have sales more/less than a certain value can be found.

DECOMPOSITION

Additive Decomposition

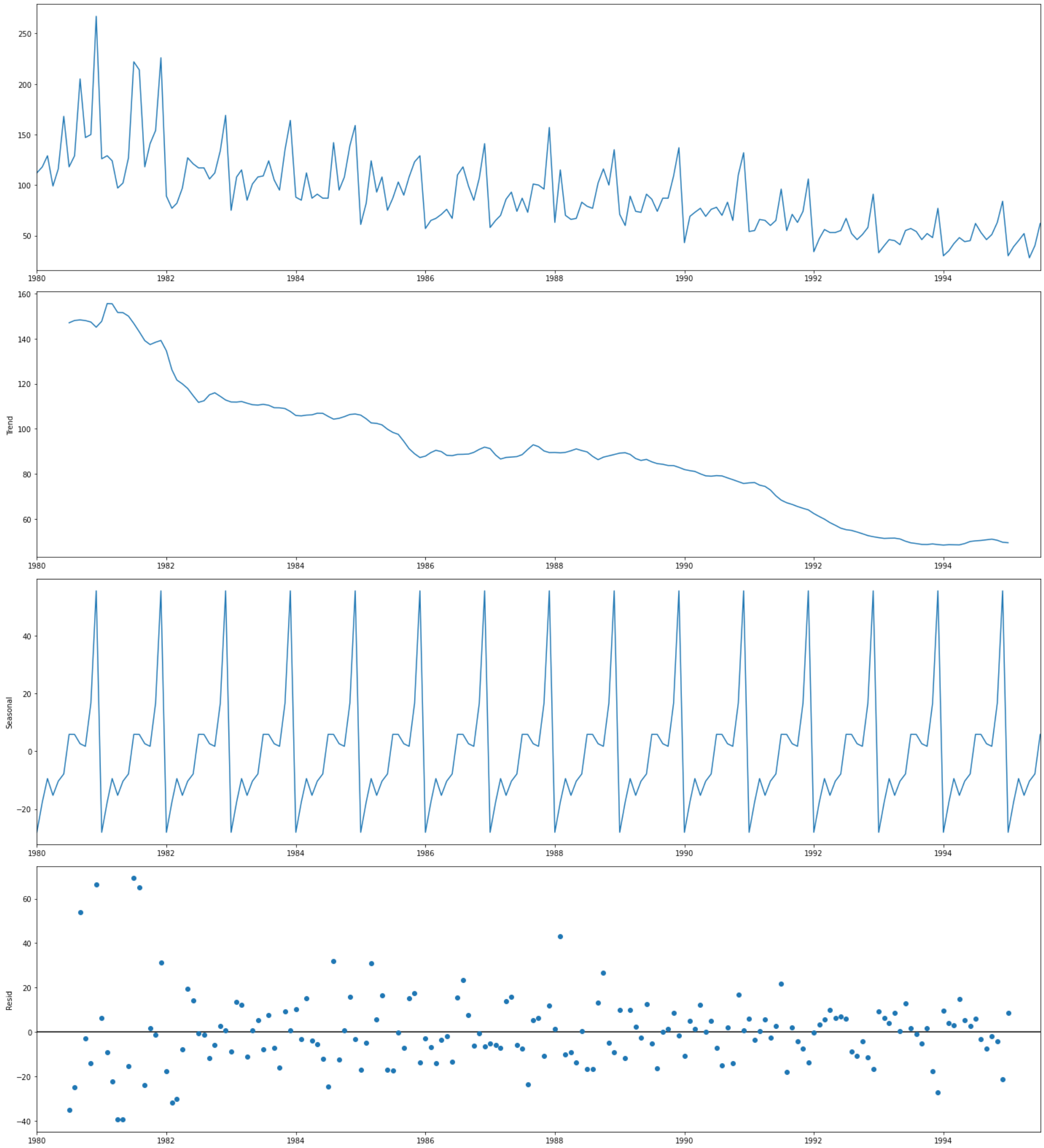


Figure 15. Trend, Seasonality, and Residual Plots after Additive Decomposition.

Insights

- Decreasing trend is observed.
- Yearly seasonality is present in the time series. Sales reach to maximum in December every year.
- Residuals are not following any pattern.

Multiplicative Decomposition

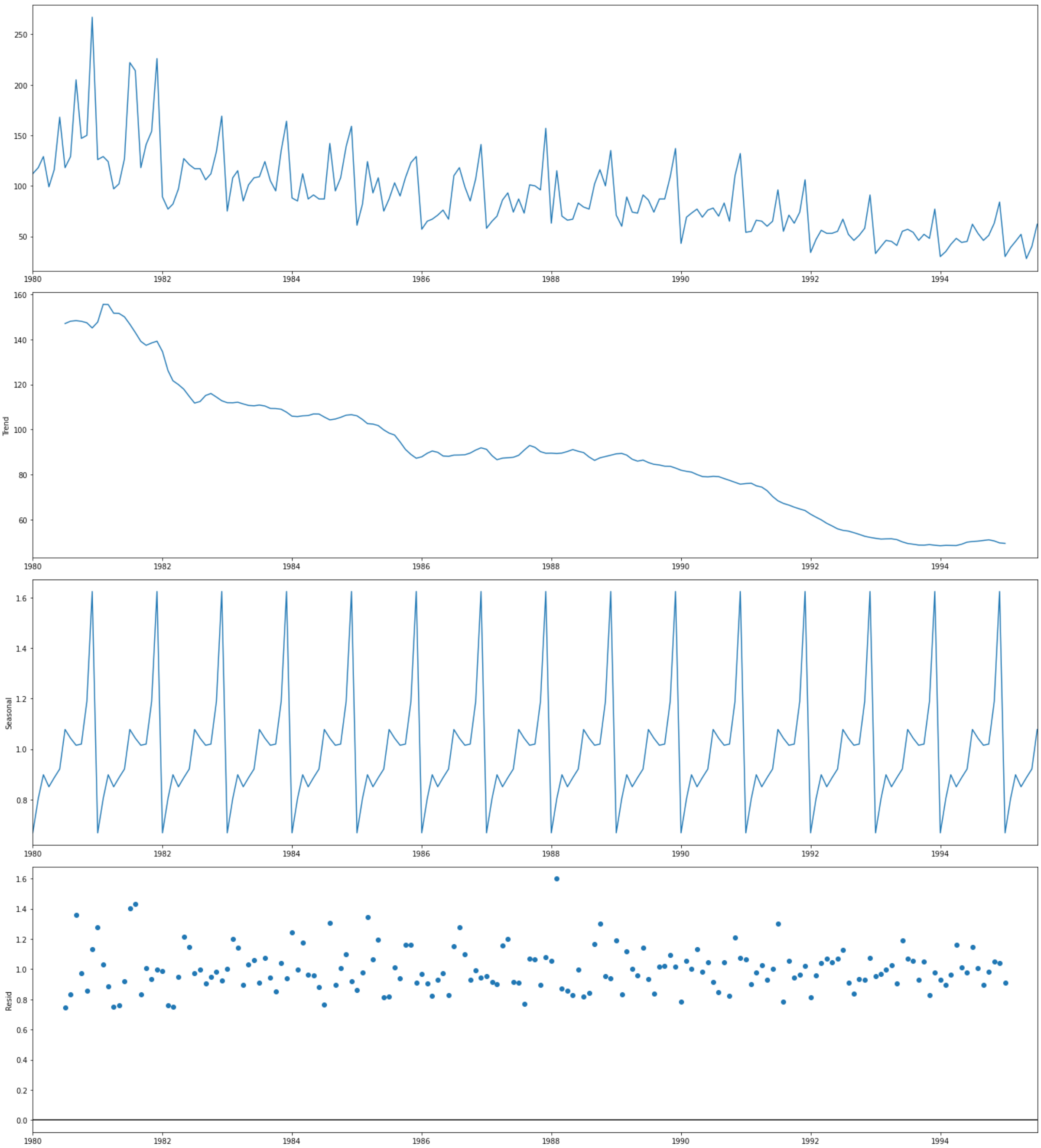


Figure 16. Trend, Seasonality, and Residual Plots after Multiplicative Decomposition.

Insights

- Decreasing trend is observed.
- Yearly seasonality is present in the time series. Sales reach to maximum in December every year.
- Residuals are not following any pattern.
- As residuals are randomly distributed in both additive and multiplicative decomposition. It is recommended to try both of them in forecasting to find out the best model.

	trend	seasonal	resid
YearMonth			
1980-01-01	NaN	-28.0	NaN
1980-02-01	NaN	-17.6	NaN
1980-03-01	NaN	-9.4	NaN
1980-04-01	NaN	-15.2	NaN
1980-05-01	NaN	-10.3	NaN
1980-06-01	NaN	-7.8	NaN
1980-07-01	147.1	5.9	-35.0
1980-08-01	148.1	5.9	-25.0
1980-09-01	148.4	2.6	54.0
1980-10-01	148.1	1.7	-2.8
1980-11-01	147.4	16.7	-14.1
1980-12-01	145.1	55.6	66.3
1981-01-01	147.8	-28.0	6.3
1981-02-01	155.6	-17.6	-9.1
1981-03-01	155.5	-9.4	-22.1
1981-04-01	151.7	-15.2	-39.4
1981-05-01	151.6	-10.3	-39.2
1981-06-01	150.0	-7.8	-15.2
1981-07-01	146.8	5.9	69.3
1981-08-01	143.1	5.9	65.1
1981-09-01	139.2	2.6	-23.8
1981-10-01	137.4	1.7	1.8
1981-11-01	138.5	16.7	-1.2

	trend	seasonal	resid
YearMonth			
1980-01-01	NaN	0.7	NaN
1980-02-01	NaN	0.8	NaN
1980-03-01	NaN	0.9	NaN
1980-04-01	NaN	0.9	NaN
1980-05-01	NaN	0.9	NaN
1980-06-01	NaN	0.9	NaN
1980-07-01	147.1	1.1	0.7
1980-08-01	148.1	1.0	0.8
1980-09-01	148.4	1.0	1.4
1980-10-01	148.1	1.0	1.0
1980-11-01	147.4	1.2	0.9
1980-12-01	145.1	1.6	1.1
1981-01-01	147.8	0.7	1.3
1981-02-01	155.6	0.8	1.0
1981-03-01	155.5	0.9	0.9
1981-04-01	151.7	0.9	0.8
1981-05-01	151.6	0.9	0.8
1981-06-01	150.0	0.9	0.9
1981-07-01	146.8	1.1	1.4
1981-08-01	143.1	1.0	1.4
1981-09-01	139.2	1.0	0.8
1981-10-01	137.4	1.0	1.0
1981-11-01	138.5	1.2	0.9

Table 10. Components of both Additive and Multiplicative Decomposition.

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

- Data is divided into train and test sets. The sales before 1991 are included in the train set. The sales from 1991 are included in the test set.
- Train set has 132 records and test set has 55 records.

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

First Few Rows of Train Data

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

First Few Rows of Test Data

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

Last Few Rows of Train Data

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Last Few Rows of Test Data

Table 11. Sample Train and Test Datasets.

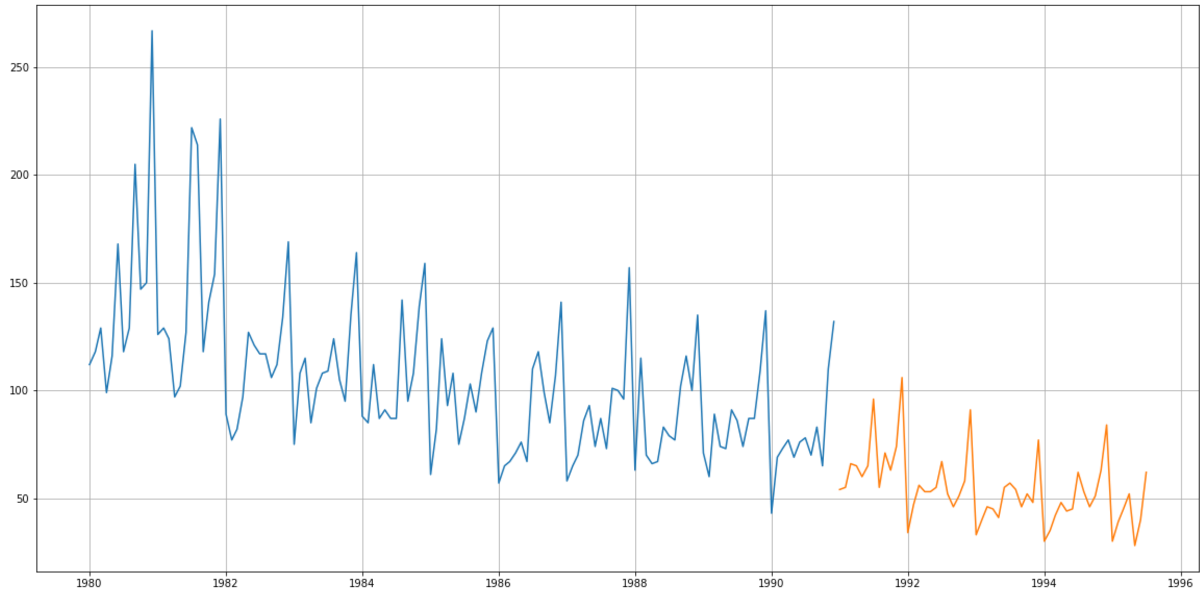


Figure 17. Train and Test Datasets Plot.

Q4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Model 1: Linear Regression

- In the linear regression model, we are going **to regress the 'Rose' variable against the order of the occurrence**. For this, we need to add time instance to our training data before fitting it into a linear regression.
- We have successfully generated the numerical time instance order for both the training and test set.

```
Training Time
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

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 Train 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

First Few Rows of Test 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 Train 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

Last Few Rows of Test Data

Table 12. Sample Train and Test Datasets for Linear Regression Model.

YearMonth	Rose	forecast_lr
1991-01-01	54.0	72.1
1991-02-01	55.0	71.6
1991-03-01	66.0	71.1
1991-04-01	65.0	70.6
1991-05-01	60.0	70.1
1991-06-01	65.0	69.6
1991-07-01	96.0	69.1
1991-08-01	55.0	68.6
1991-09-01	71.0	68.1
1991-10-01	63.0	67.6

Table 13. Sample of Forecasted Sales in Linear Regression Model.

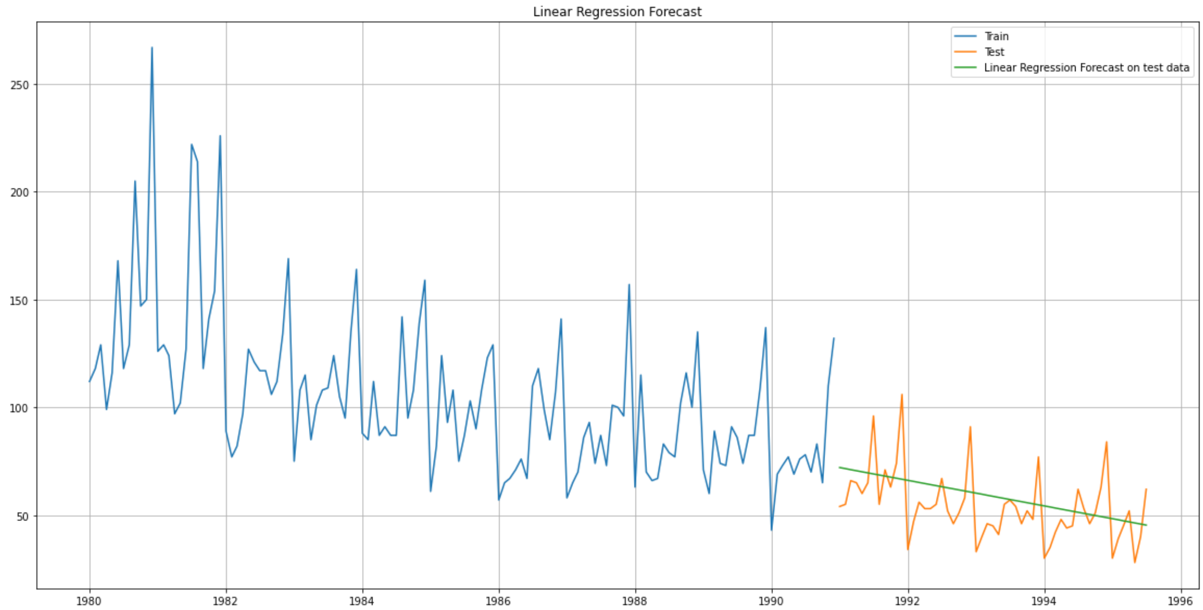


Figure 18. Plot of Forecasted Sales in Linear Regression Model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Linear Regression may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in Linear Regression Model is 15.3.

Model 2: Naïve Approach

- In this model, the forecast for tomorrow is the same as today and the forecast for the day after tomorrow is tomorrow therefore the forecast for the day after tomorrow is also today.

YearMonth	Rose	forecast_naive
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 14. Sample of Forecasted Sales in Naive Model.

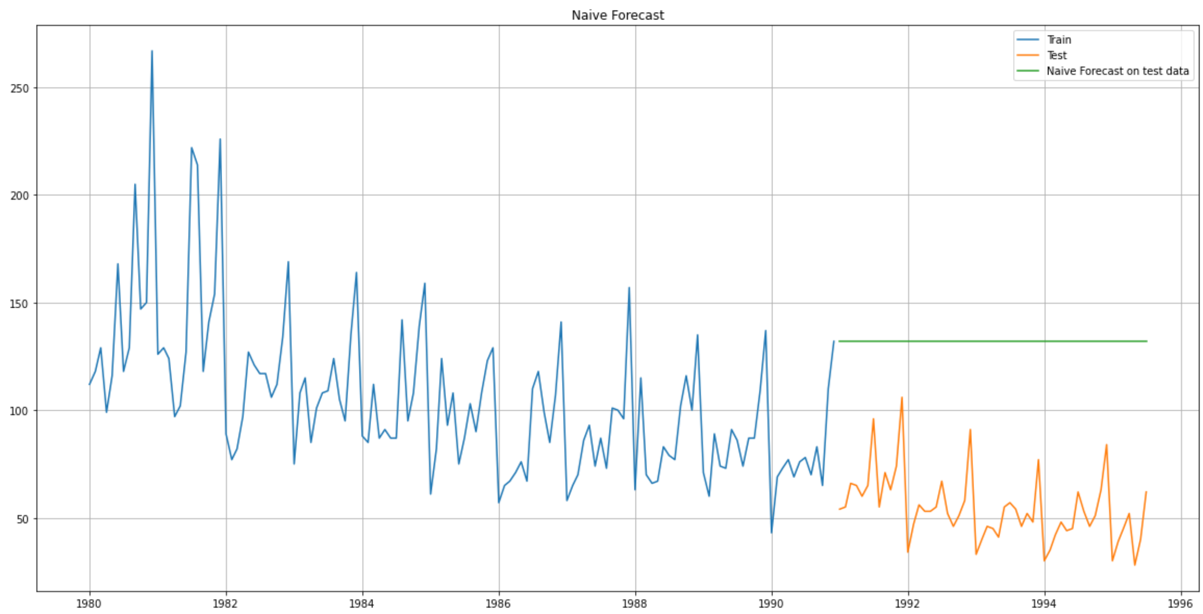


Figure 19. Plot of Forecasted Sales in Naive Model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the Naïve model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.

- RMSE in Naive Model is 79.3

Model 3: Simple Average Approach

- In this model, the forecast for the future is the **simple average of training values**.

	Rose	forecast_sa
YearMonth		
1991-01-01	54.0	104.9
1991-02-01	55.0	104.9
1991-03-01	66.0	104.9
1991-04-01	65.0	104.9
1991-05-01	60.0	104.9
1991-06-01	65.0	104.9
1991-07-01	96.0	104.9
1991-08-01	55.0	104.9
1991-09-01	71.0	104.9
1991-10-01	63.0	104.9

Table 15. Sample of Forecasted Sales in Simple Average Model.

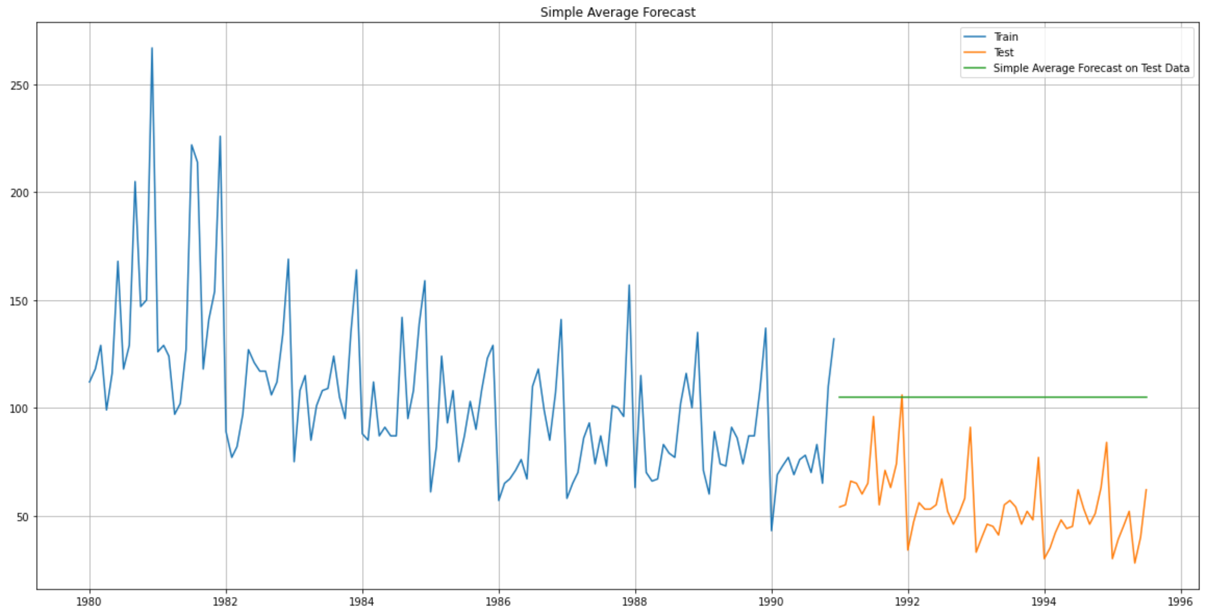


Figure 20. Plot of Forecasted Sales in Simple Average Model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the Simple Average model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in Simple Average Model is 53.02

Exponential Smoothing

- Exponential smoothing methods consist of flattening time series data.
- Exponential smoothing averages or exponentially weighted moving averages consist of forecasts based on previous periods' data with exponentially declining influence on the older observations.
- Exponential smoothing methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality) where each can be none(N), additive (N), additive damped (Ad), Multiplicative (M), or multiplicative damped (Md).
- One or more parameters control how fast the weights decay.
- These parameters have values between 0 and 1

Model 4: Simple Exponential Smoothing

- The simplest of the exponentially smoothing methods is called simple exponential smoothing (SES).
- This method is suitable for forecasting data when a **time series does not have trend and seasonality**.
- In simple exponential smoothing, the forecast at the time (t + 1) is given by,

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t$$

- Parameter **α** is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.

I. Optimized Model

A simple exponential smoothing model is built and fitted with optimum parameters.

Parameters

	name	param	optimized
smoothing_level	alpha	0.098750	True
initial_level	l.0	134.387025	True

Table 16. Smoothing Parameters in Simple Exponential Smoothing Optimized Model.

YearMonth	Rose	forecast_ses_optimized
1991-01-01	54.0	87.1
1991-02-01	55.0	87.1
1991-03-01	66.0	87.1
1991-04-01	65.0	87.1
1991-05-01	60.0	87.1
1991-06-01	65.0	87.1
1991-07-01	96.0	87.1
1991-08-01	55.0	87.1
1991-09-01	71.0	87.1
1991-10-01	63.0	87.1

Table 17. Sample of Forecasted Sales in Simple Exponential Smoothing Optimized Model.

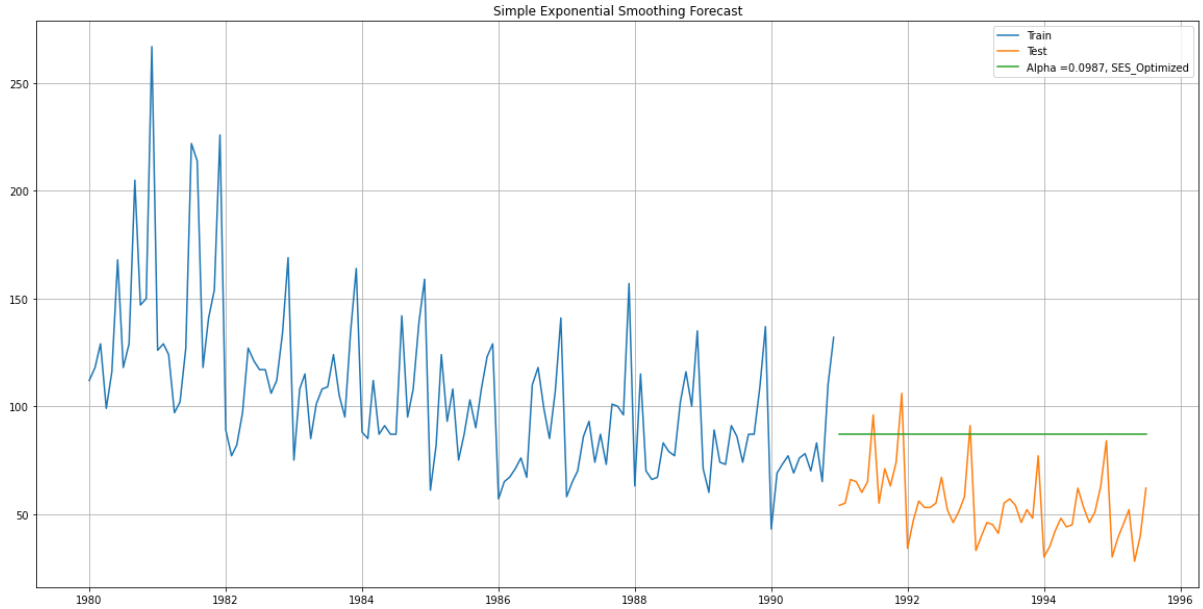


Figure 21. Plot of Forecasted Sales in Simple Exponential Smoothing Optimized Model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the Simple Exponential Smoothing Optimized Model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in Simple Exponential Smoothing Optimized Model is 36.38.

II. Iteration Model – Finding best **α** to minimize RMSE on the test dataset.

- The higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.
- Different SES models are built and fitted with different α values (0.1 to 1) and RMSE values are calculated on the test dataset. Below are the top five models with low RMSE values on the test dataset.

	Alpha_Values	RMSE_Train	RMSE_Test
0	0.1	32.3	36.4
1	0.2	32.2	40.9
2	0.3	32.6	47.1
3	0.4	33.1	53.3
4	0.5	33.7	59.2

Table 18. SES models with low test RMSE values.

From the above table, it can be noticed that the SES model with smoothing constant 0.1 can be considered as the best and final simple exponential smoothing model.

	Rose	forecast_ses_optimized	forecast_ses
YearMonth			
1991-01-01	54.0	87.1	87.1
1991-02-01	55.0	87.1	87.1
1991-03-01	66.0	87.1	87.1
1991-04-01	65.0	87.1	87.1
1991-05-01	60.0	87.1	87.1
1991-06-01	65.0	87.1	87.1
1991-07-01	96.0	87.1	87.1
1991-08-01	55.0	87.1	87.1
1991-09-01	71.0	87.1	87.1
1991-10-01	63.0	87.1	87.1

Table 19. Sample of Forecasted Sales in Simple Exponential Smoothing Iteration Model.

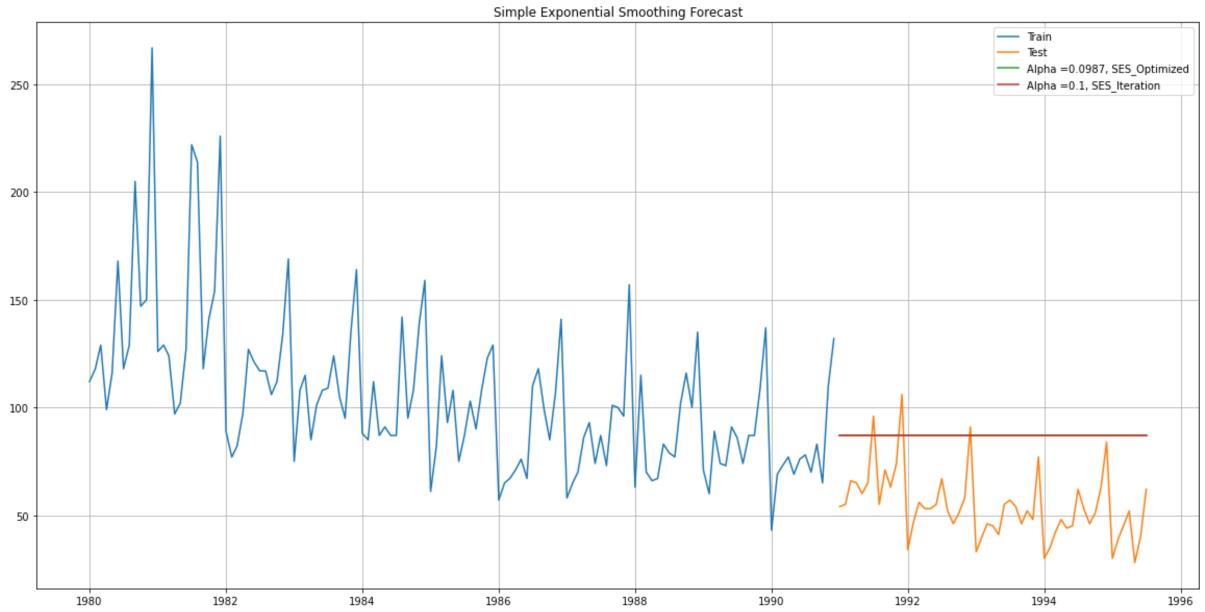


Figure 22. Plot of Forecasted Sales in Simple Exponential Smoothing Iteration Model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the Simple Exponential Smoothing Iteration Model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in Simple Exponential Smoothing Iteration Model is 36.41.

Model 5: Double Exponential Smoothing

- This model is applicable when the dataset has a trend but no seasonality.
- Two separate components are considered: Level and Trend.
- Level is the local mean.
- One smoothing parameter α corresponds to the level series which has values between 0 and 1.
- A second smoothing parameter β corresponds to the trend series which has values between 0 and 1.

I. Optimized Model

A double exponential smoothing model is built and fitted with optimum parameters.

	name	param	optimized
smoothing_level	alpha	1.490116e-08	True
smoothing_trend	beta	1.661039e-10	True

Table 20. Smoothing Parameters in Double Exponential Smoothing Optimized Model.

YearMonth	Rose	forecast_des_optimized
1991-01-01	54.0	72.1
1991-02-01	55.0	71.6
1991-03-01	66.0	71.1
1991-04-01	65.0	70.6
1991-05-01	60.0	70.1
1991-06-01	65.0	69.6
1991-07-01	96.0	69.1
1991-08-01	55.0	68.6
1991-09-01	71.0	68.1
1991-10-01	63.0	67.6

Table 21. Sample of Forecasted Sales in Double Exponential Smoothing Optimized Model.

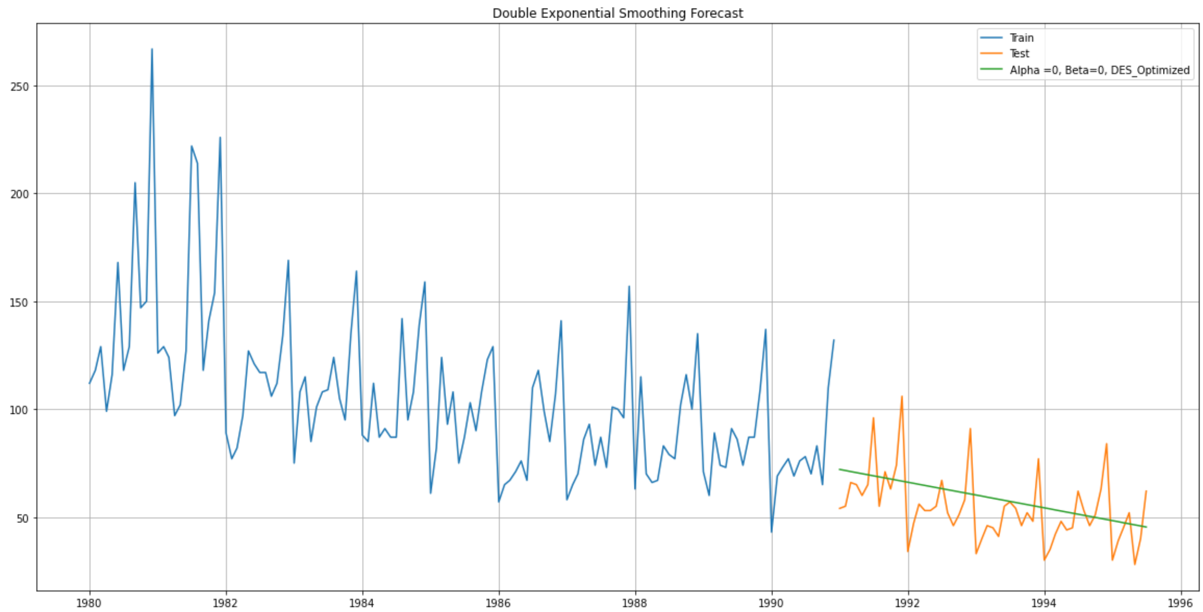


Figure 23. Plot of Forecasted Sales in Double Exponential Smoothing Optimized Model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the Double Exponential Smoothing Optimized Model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in Double Exponential Smoothing Optimized Model is 15.3

II. Iteration Model – Finding best **α** and **β** to minimize RMSE on the test dataset.

- We will run a loop with different **α and β** values to understand which particular values work best for **α and β** on the test set.
- Different DES models are built and fitted with different **α and β** values (0.1 to 1) and RMSE values are calculated on the test dataset. Below are the top five models with low RMSE values on the test dataset.

	Alpha_Values	Beta Values	RMSE_Train	RMSE_Test
0	0.1	0.1	35.1	36.4
1	0.1	0.2	34.0	48.1
10	0.2	0.1	33.3	65.2
2	0.1	0.3	33.5	77.6
20	0.3	0.1	33.7	98.1

Table 22. DES models with low test RMSE values.

From the above table, it can be noticed that the DES model with smoothing constants α=0.1 & β=0.1 can be considered as the best and final double exponential smoothing model.

	Rose	forecast_des_optimized	forecast_des
YearMonth			
1991-01-01	54.0	72.1	83.9
1991-02-01	55.0	71.6	84.0
1991-03-01	66.0	71.1	84.1
1991-04-01	65.0	70.6	84.2
1991-05-01	60.0	70.1	84.3
1991-06-01	65.0	69.6	84.4
1991-07-01	96.0	69.1	84.5
1991-08-01	55.0	68.6	84.6
1991-09-01	71.0	68.1	84.7
1991-10-01	63.0	67.6	84.8

Table 23. Sample of Forecasted Sales in Double Exponential Smoothing Iteration Model.

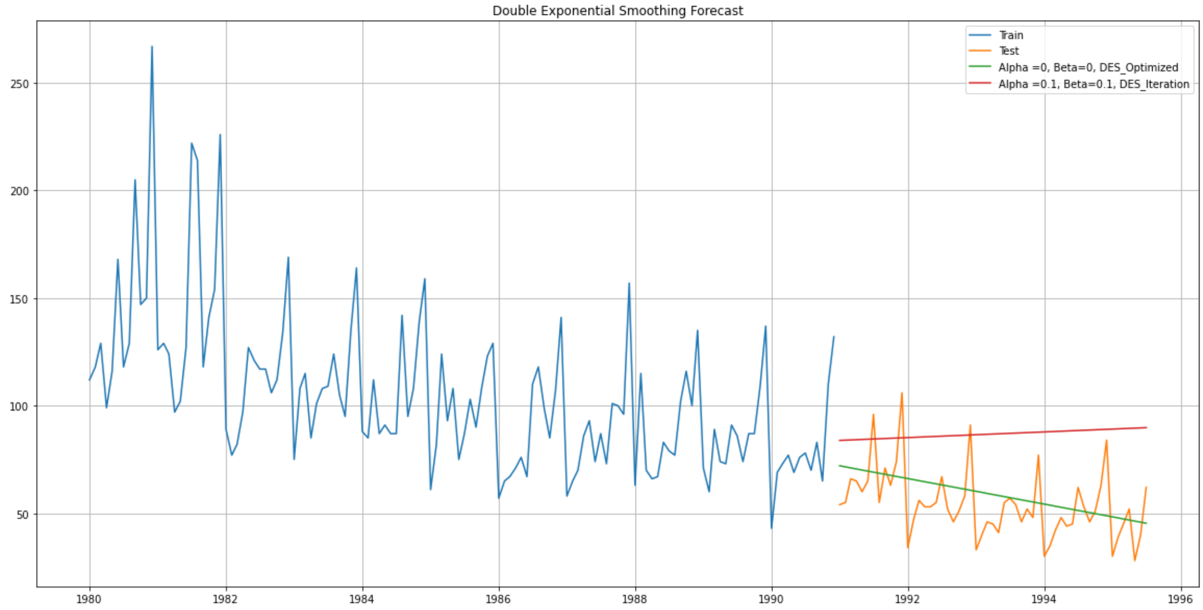


Figure 24. Plot of Forecasted Sales in Double Exponential Smoothing Iteration Model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the Double Exponential Smoothing Iteration Model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in Double Exponential Smoothing Iteration Model is 1777.7

Model 6: Triple Exponential Smoothing with Additive trend & Additive seasonality

I. Optimized Model

A TES with the additive trend and additive seasonality model is built and fitted with optimum parameters.

Parameters

	name	param	optimized
smoothing_level	alpha	0.089541	True
smoothing_trend	beta	0.000240	True
smoothing_seasonal	gamma	0.003467	True

Table 24. Smoothing Parameters in TES with the additive trend and additive seasonality optimized model.

	Rose	forecast_tes_add_add_optimized
YearMonth		
1991-01-01	54.0	42.7
1991-02-01	55.0	54.6
1991-03-01	66.0	62.0
1991-04-01	65.0	50.9
1991-05-01	60.0	59.0
1991-06-01	65.0	63.9
1991-07-01	96.0	73.2
1991-08-01	55.0	78.7
1991-09-01	71.0	74.3
1991-10-01	63.0	71.9

Table 25. Sample of Forecasted Sales in TES with the additive trend and additive seasonality optimized model.

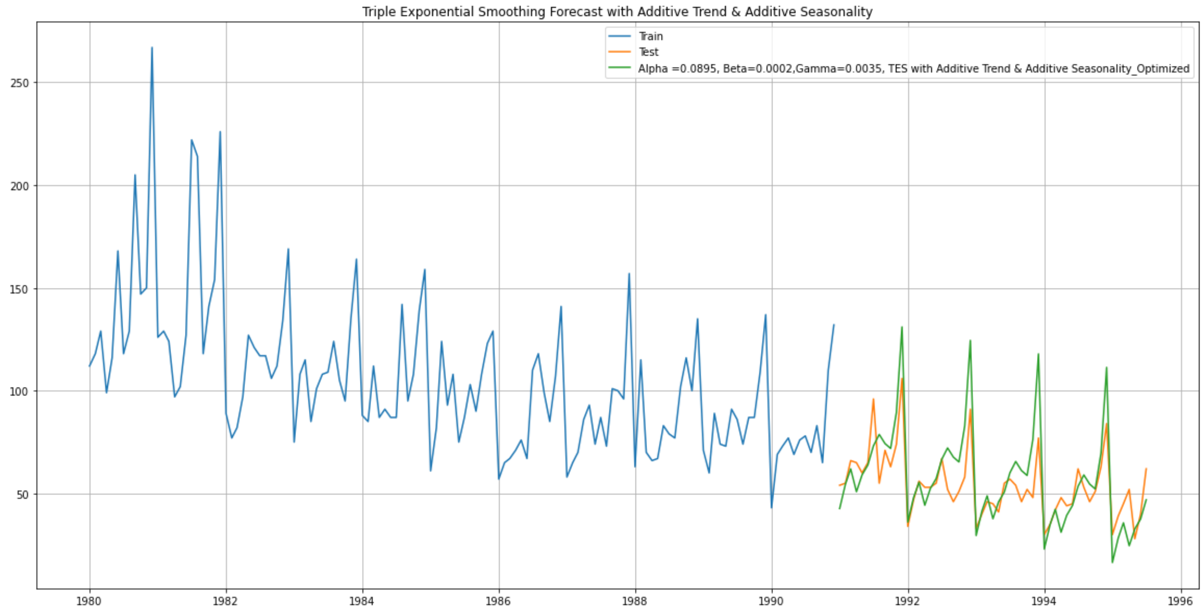


Figure 25. Plot of Forecasted Sales in TES with the additive trend and additive seasonality optimized model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is approximately matching with the actual plot of the test set**. Hence, the TES with the additive trend and additive seasonality optimized model may be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.

- RMSE in TES with the additive trend and additive seasonality optimized model is 14.16

II. Iteration Model – Finding best **α**, **β,** and **γ** to minimize RMSE on the test dataset.

- We will run a loop with different **α, β**, and **γ** values to understand which particular values work best for **α, β**, and **γ** on the test set.
- Different TES models are built and fitted with different **α, β**, and **γ** values (0.1 to 1), and RMSE values are calculated on the test dataset. Below are the top five models with low RMSE values on the test dataset.

	Alpha_Values	Beta Values	Gamma Values	RMSE_Train	RMSE_Test
32	0.1	0.4	0.3	22.7	12.3
22	0.1	0.3	0.3	22.3	12.3
13	0.1	0.2	0.4	22.5	12.3
12	0.1	0.2	0.3	22.0	12.4
23	0.1	0.3	0.4	22.9	12.4

Table 26. TES with the additive trend and additive seasonality models with low test RMSE values.

From the above table, it can be noticed that the TES model with smoothing constants α=0.1, β=0.4, and γ=0.3 can be considered as the best and final triple exponential smoothing model with additive trend and additive seasonality.

	Rose	forecast_tes_add_add_optimized	forecast_tes_add_add
YearMonth			
1991-01-01	54.0	42.7	44.8
1991-02-01	55.0	54.6	61.1
1991-03-01	66.0	62.0	65.5
1991-04-01	65.0	50.9	60.7
1991-05-01	60.0	59.0	60.3
1991-06-01	65.0	63.9	64.6
1991-07-01	96.0	73.2	68.6
1991-08-01	55.0	78.7	65.5
1991-09-01	71.0	74.3	74.8
1991-10-01	63.0	71.9	71.0

Table 27. Sample of Forecasted Sales in TES with the additive trend and additive seasonality Iteration Model.

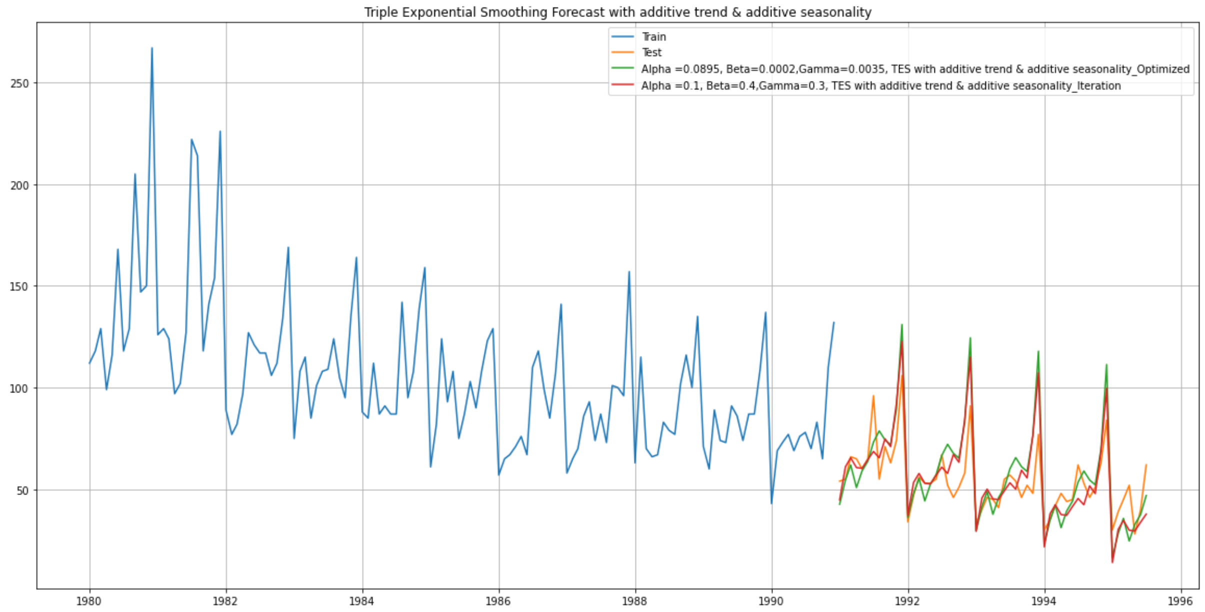


Figure 26. Plot of Forecasted Sales in TES with the additive trend and additive seasonality iteration model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is approximately matching with the actual plot of the test set**. Hence, the triple exponential smoothing with the additive trend and additive seasonality iteration model may be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in triple exponential smoothing with the additive trend and additive seasonality iteration model is 12.25

Model 7: Triple Exponential Smoothing with Additive Trend & Multiplicative Seasonality.

I. Optimized Model

A TES with the additive trend and multiplicative seasonality model is built and fitted with optimum parameters.

	name	param	optimized
smoothing_level	alpha	0.071511	True
smoothing_trend	beta	0.045292	True
smoothing_seasonal	gamma	0.000072	True

Table 28. Smoothing Parameters in TES with the additive trend and multiplicative seasonality optimized model.

YearMonth	Rose	forecast_tes_add_mult_optimized
1991-01-01	54.0	56.3
1991-02-01	55.0	63.7
1991-03-01	66.0	69.4
1991-04-01	65.0	60.4
1991-05-01	60.0	67.8
1991-06-01	65.0	73.5
1991-07-01	96.0	80.6
1991-08-01	55.0	85.5
1991-09-01	71.0	80.7
1991-10-01	63.0	78.8

Table 29. Sample of Forecasted Sales in TES with the additive trend and multiplicative seasonality optimized model.

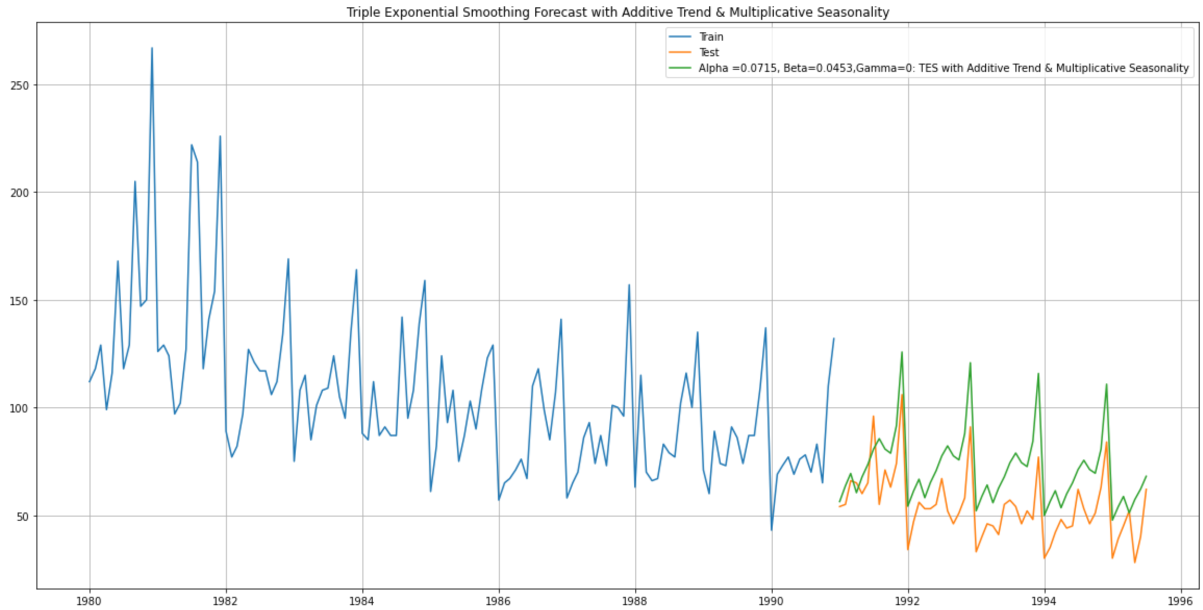


Figure 27. Plot of Forecasted Sales in TES with the additive trend and multiplicative seasonality optimized model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is approximately matching with the actual plot of the test set**. Hence, the TES with the additive trend and multiplicative seasonality optimized model may be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.

- RMSE in TES with the additive trend and multiplicative seasonality optimized model is 19.71

II. Iteration Model – Finding best **α**, **β**, and **γ** to minimize RMSE on the test dataset.

- We will run a loop with different **α, β**, and **γ** values to understand which particular values work best for **α, β**, and **γ** on the test set.
- Different TES models are built and fitted with different **α, β**, and **γ** values (0.1 to 1), and RMSE values are calculated on the test dataset. Below are the top five models with low RMSE values on the test dataset.

	Alpha_Values	Beta Values	Gamma Values	RMSE_Train	RMSE_Test
10	0.1	0.2	0.1	19.770392	9.266047
11	0.1	0.2	0.2	20.253487	9.585590
151	0.2	0.6	0.2	23.129850	9.982780
12	0.1	0.2	0.3	20.871304	9.989167
142	0.2	0.5	0.3	23.656276	10.336096

Table 30. TES with the additive trend and multiplicative seasonality models with low test RMSE values.

From the above table, it can be noticed that the TES model with smoothing constants α=0.1, β=0.2, and γ=0.1 can be considered as the best and final triple exponential smoothing model with additive trend and multiplicative seasonality.

	Rose	forecast_tes_add_mult_optimized	forecast_tes_add_mult
YearMonth			
1991-01-01	54.0	56.3	52.6
1991-02-01	55.0	63.7	61.4
1991-03-01	66.0	69.4	66.1
1991-04-01	65.0	60.4	59.0
1991-05-01	60.0	67.8	63.4
1991-06-01	65.0	73.5	65.3
1991-07-01	96.0	80.6	72.8
1991-08-01	55.0	85.5	74.7
1991-09-01	71.0	80.7	73.4
1991-10-01	63.0	78.8	70.3

Table 31. Sample of Forecasted Sales in TES with the additive trend and multiplicative seasonality Iteration Model.

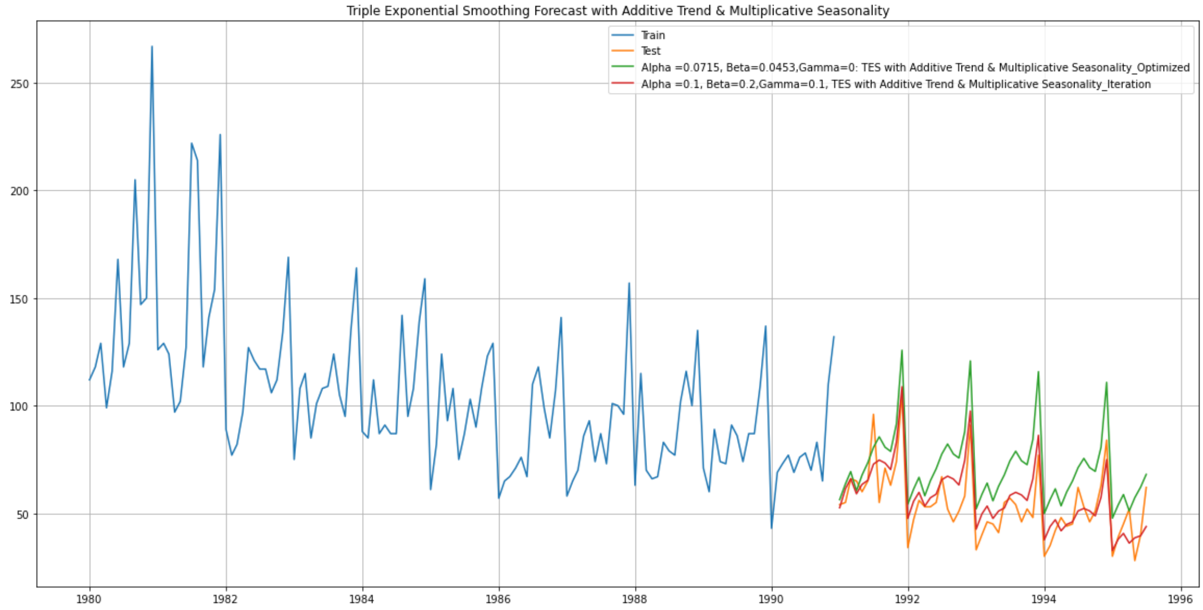


Figure 28. Plot of Forecasted Sales in TES with the additive trend and multiplicative seasonality iteration model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is approximately matching with the actual plot of the test set**. Hence, the triple exponential smoothing with the additive trend and multiplicative seasonality iteration model may be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in triple exponential smoothing with the additive trend and multiplicative seasonality iteration model is 9.26

Model 8: Triple Exponential Smoothing with Multiplicative Trend & Multiplicative Seasonality.

I. Optimized Model

A TES with the multiplicative trend and multiplicative seasonality model is built and fitted with optimum parameters.

	name	param	optimized
smoothing_level	alpha	0.055093	True
smoothing_trend	beta	0.031634	True
smoothing_seasonal	gamma	0.000334	True

Table 32. Smoothing Parameters in TES with the multiplicative trend and multiplicative seasonality optimized model.

YearMonth	Rose	forecast_tes_mult_mult_optimized
1991-01-01	54.0	55.7
1991-02-01	55.0	63.0
1991-03-01	66.0	68.7
1991-04-01	65.0	59.8
1991-05-01	60.0	67.1
1991-06-01	65.0	72.7
1991-07-01	96.0	79.7
1991-08-01	55.0	84.7
1991-09-01	71.0	80.0
1991-10-01	63.0	78.0

Table 33. Sample of Forecasted Sales in TES with the multiplicative trend and multiplicative seasonality optimized model.

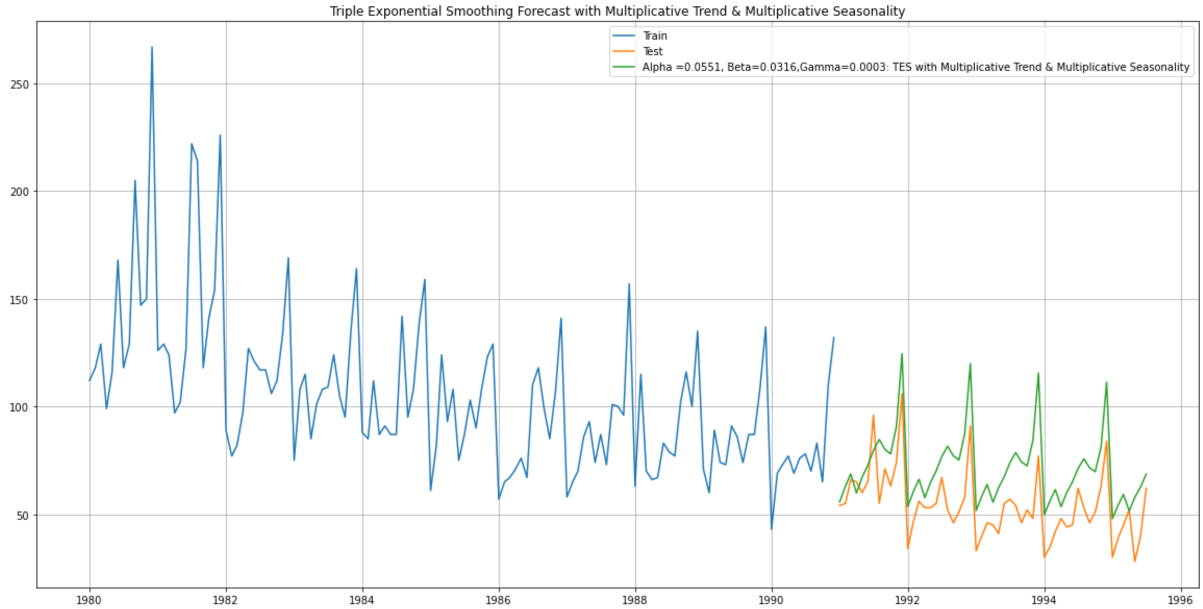


Figure 29. Plot of Forecasted Sales in TES with the multiplicative trend and multiplicative seasonality optimized model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is approximately matching with the actual plot of the test set**. Hence, the TES with the multiplicative trend and multiplicative seasonality optimized model may be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.

- RMSE in TES with the multiplicative trend and multiplicative seasonality optimized model is 19.54

II. Iteration Model – Finding best **α**, **β,** and **γ** to minimize RMSE on the test dataset.

- We will run a loop with different **α, β**, and **γ** values to understand which particular values work best for **α, β**, and **γ** on the test set.
- Different TES models are built and fitted with different **α, β**, and **γ** values (0.1 to 1), and RMSE values are calculated on the test dataset. Below are the top five models with low RMSE values on the test dataset.

	Alpha_Values	Beta Values	Gamma Values	RMSE_Train	RMSE_Test
136	0.2	0.7	0.2	24.042290	9.053203
176	0.3	0.2	0.6	26.940472	10.652112
9	0.1	0.2	0.1	19.647823	10.912264
10	0.1	0.2	0.2	20.172839	11.258336
11	0.1	0.2	0.3	20.828952	11.717807

Table 34. TES with the multiplicative trend and multiplicative seasonality models with low test RMSE values.

From the above table, it can be noticed that the TES model with smoothing constants α=0.2, β=0.7, and γ=0.2 can be considered as the best and final triple exponential smoothing model with multiplicative trend and multiplicative seasonality.

	Rose	forecast_tes_mult_mult_optimized	forecast_tes_mult_mult
YearMonth			
1991-01-01	54.0	55.7	49.0
1991-02-01	55.0	63.0	60.8
1991-03-01	66.0	68.7	64.5
1991-04-01	65.0	59.8	59.9
1991-05-01	60.0	67.1	60.0
1991-06-01	65.0	72.7	61.4
1991-07-01	96.0	79.7	66.4
1991-08-01	55.0	84.7	66.0
1991-09-01	71.0	80.0	72.7
1991-10-01	63.0	78.0	70.9

Table 35. Sample of Forecasted Sales in TES with the multiplicative trend and multiplicative seasonality Iteration Model.

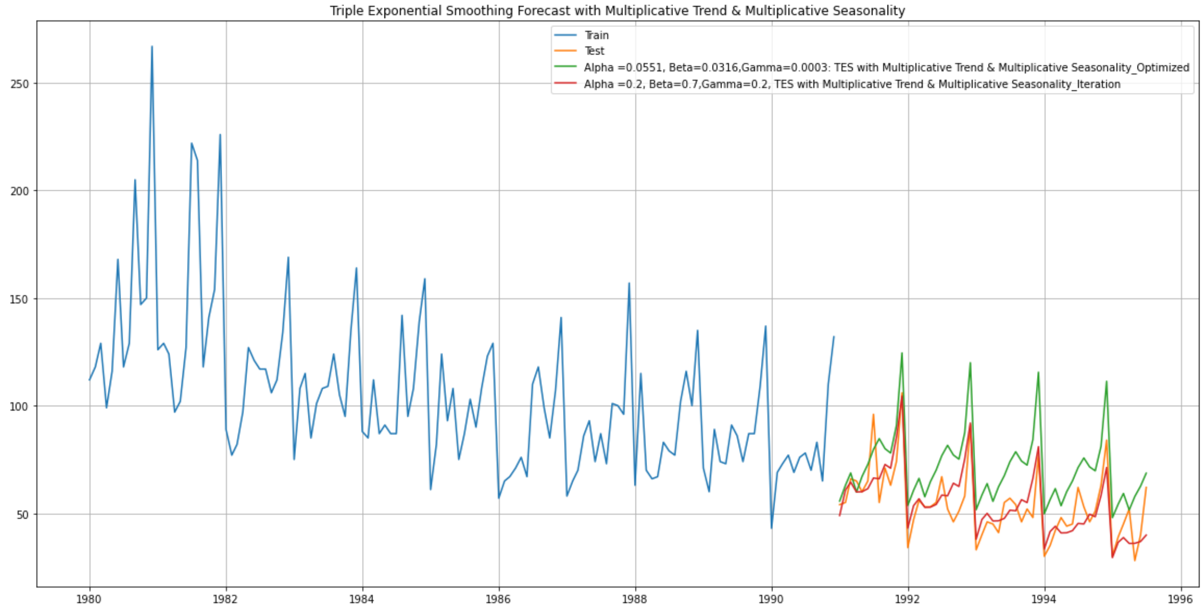


Figure 30. Plot of Forecasted Sales in TES with the multiplicative trend and multiplicative seasonality iteration model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is approximately matching with the actual plot of the test set**. Hence, the triple exponential smoothing with the multiplicative trend and multiplicative seasonality iteration model may be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in triple exponential smoothing with the multiplicative trend and multiplicative seasonality iteration model is 9.05

Model 9: Triple Exponential Smoothing with Multiplicative Trend & Additive Seasonality.

I. Optimized Model

A TES with the multiplicative trend and additive seasonality model is built and fitted with optimum parameters.

	name	param	optimized
smoothing_level	alpha	0.013450	True
smoothing_trend	beta	0.000060	True
smoothing_seasonal	gamma	0.000430	True

Table 36. Smoothing Parameters in TES with the multiplicative trend and additive seasonality optimized model.

	Rose	forecast_tes_mult_add_optimized
YearMonth		
1991-01-01	54.0	43.0
1991-02-01	55.0	55.1
1991-03-01	66.0	62.9
1991-04-01	65.0	52.0
1991-05-01	60.0	60.6
1991-06-01	65.0	65.8
1991-07-01	96.0	75.4
1991-08-01	55.0	81.2
1991-09-01	71.0	77.0
1991-10-01	63.0	74.8

Table 37. Sample of Forecasted Sales in TES with the multiplicative trend and additive seasonality optimized model.

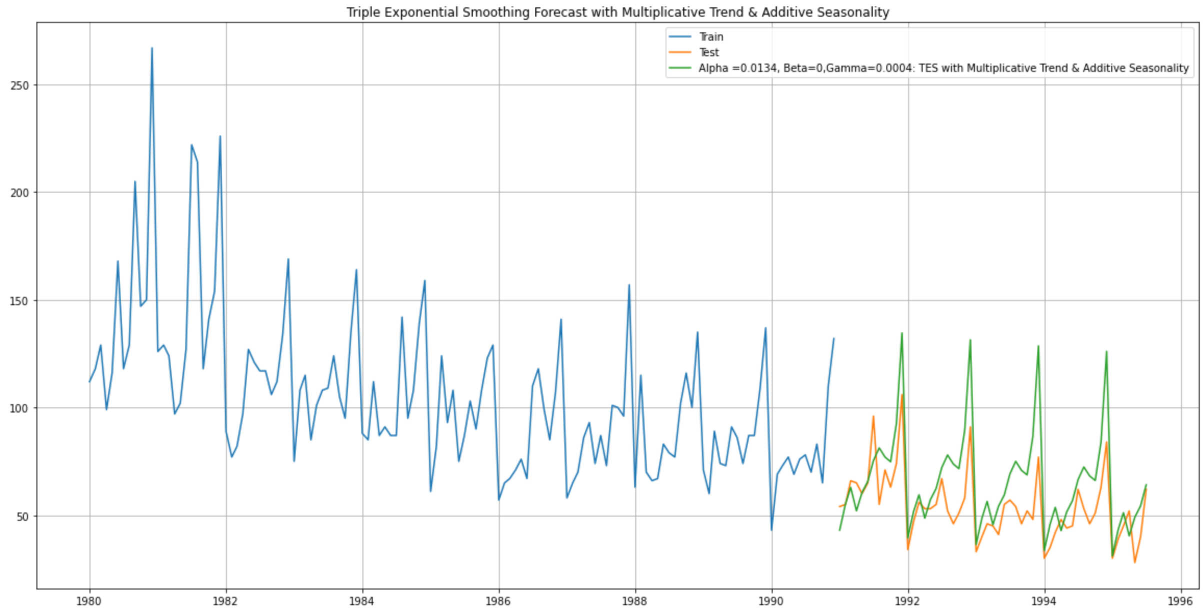


Figure 31. Plot of Forecasted Sales in TES with the multiplicative trend and additive seasonality optimized model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is approximately matching with the actual plot of the test set**. Hence, the TES with the multiplicative trend and additive seasonality optimized model may be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.

- RMSE in TES with the multiplicative trend and additive seasonality optimized model is 18.14

II. Iteration Model – Finding best **α**, **β,** and **γ** to minimize RMSE on the test dataset.

- We will run a loop with different **α, β**, and **γ** values to understand which particular values work best for **α, β**, and **γ** on the test set.
- Different TES models are built and fitted with different **α, β**, and **γ** values (0.1 to 1), and RMSE values are calculated on the test dataset. Below are the top five models with low RMSE values on the test dataset.

	Alpha_Values	Beta Values	Gamma Values	RMSE_Train	RMSE_Test
64	0.1	0.8	0.2	23.635989	12.157265
37	0.1	0.5	0.2	22.408154	12.223765
55	0.1	0.7	0.2	23.272614	12.255784
28	0.1	0.4	0.2	21.982924	12.270506
10	0.1	0.2	0.2	21.514299	12.426141

Table 38. TES with the multiplicative trend and additive seasonality models with low test RMSE values.

From the above table, it can be noticed that the TES model with smoothing constants α=0.1, β=0.8, and γ=0.2 can be considered as the best and final triple exponential smoothing model with multiplicative trend and additive seasonality.

	Rose	forecast_tes_mult_add_optimized	forecast_tes_mult_add
YearMonth			
1991-01-01	54.0	43.0	45.4
1991-02-01	55.0	55.1	60.1
1991-03-01	66.0	62.9	65.3
1991-04-01	65.0	52.0	57.6
1991-05-01	60.0	60.6	59.6
1991-06-01	65.0	65.8	61.9
1991-07-01	96.0	75.4	68.3
1991-08-01	55.0	81.2	68.4
1991-09-01	71.0	77.0	72.5
1991-10-01	63.0	74.8	69.7

Table 39. Sample of Forecasted Sales in TES with the multiplicative trend and additive seasonality Iteration Model.

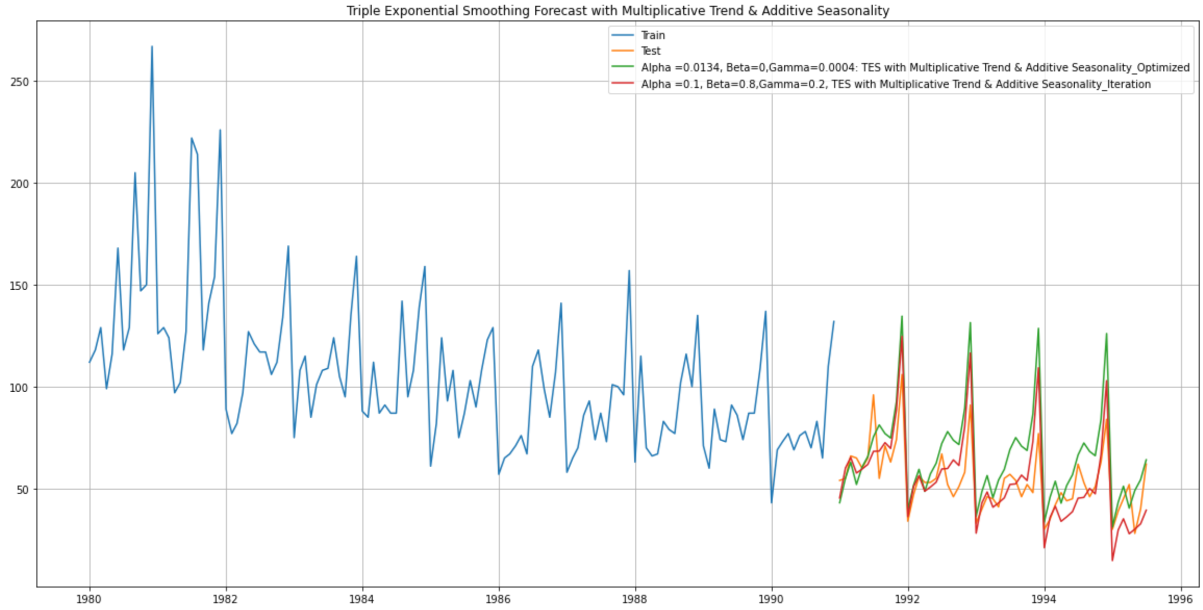


Figure 32. Plot of Forecasted Sales in TES with the multiplicative trend and additive seasonality iteration model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is approximately matching with the actual plot of the test set**. Hence, the triple exponential smoothing with the multiplicative trend and additive seasonality iteration model may be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in triple exponential smoothing with the multiplicative trend and additive seasonality iteration model is 12.15.

Q5. Check for the stationarity of the data on which the model is being built 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.

Check for Stationarity of Whole Time Series

The Augmented Dickey-Fuller test is a unit root test that determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis for the ADF test is

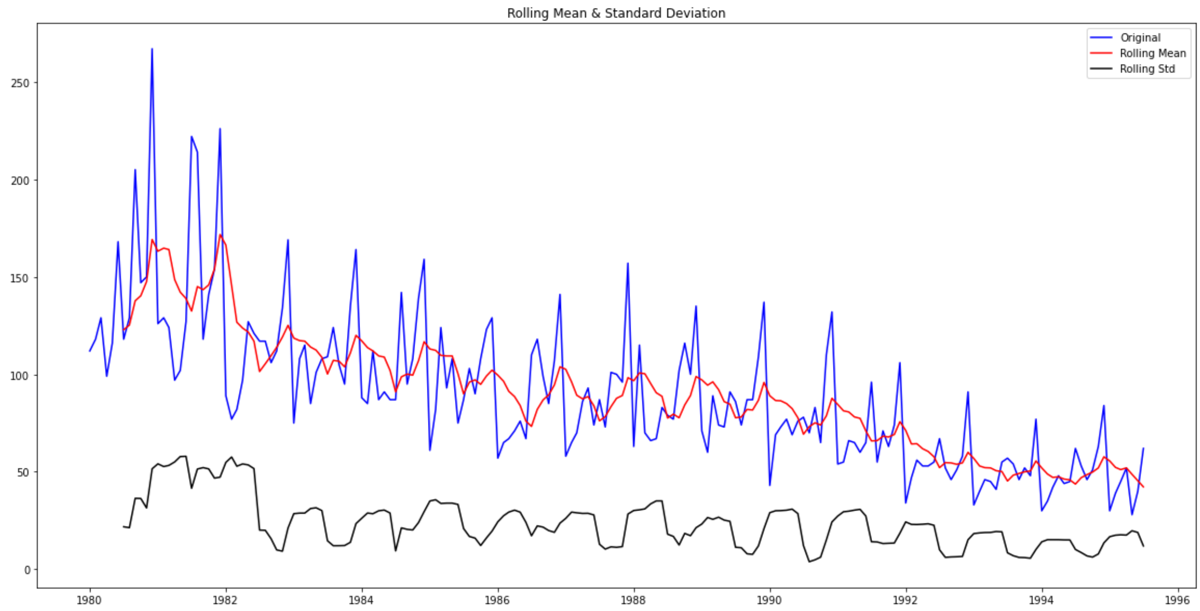
H0: The Time Series has a unit root and is thus non-stationary

H1: The Time Series does not have a unit root and is thus stationary

Significance level $\alpha = 0.05$

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value (0.05).

Results of Dickey-Fuller Test on Whole Time Series

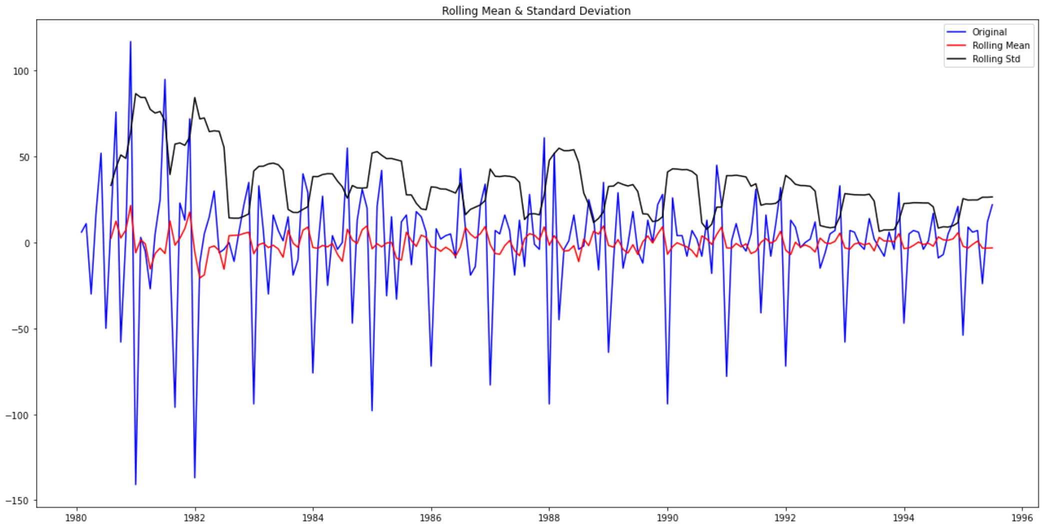



```
Results of Dickey-Fuller Test:
Test Statistic                  -1.887760
p-value                          0.337841
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

As the P-value in Dicky-Fuller test is 0.33 which is more than the significance level (0.05), we failed to reject the null hypothesis. Hence, the given **time series is non-stationary**.

To make time-series stationary, let us take first differencing on whole time series and perform Dicky-Fuller test one again.

Results of Dickey-Fuller Test on Differenced Whole Time Series




```
Results of Dickey-Fuller Test:
Test Statistic                -8.028518e+00
p-value                        1.987167e-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
```

As the P-value in Dicky-Fuller test is 0 which is less than the significance level (0.05), we can reject the null hypothesis. Hence, **one level differenced time series is stationary**.

Plotting Differenced Time Series

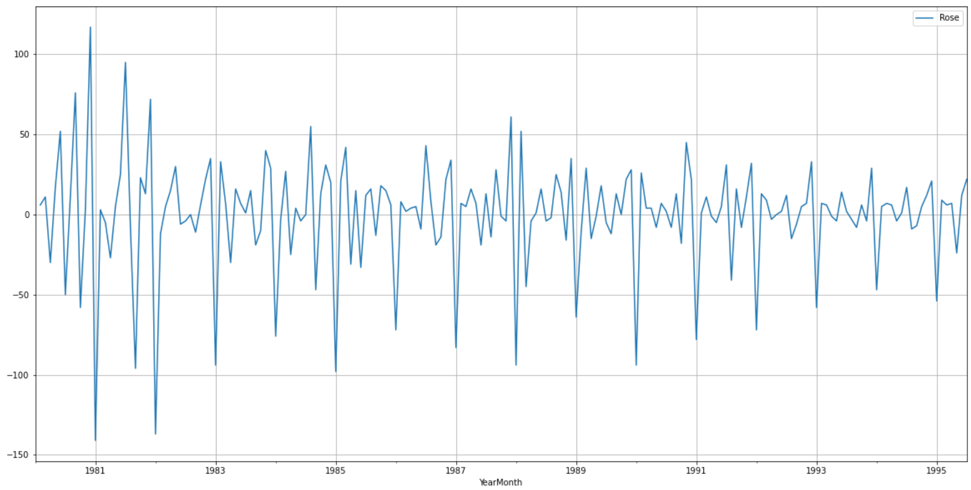


Figure 33. Plot of Differenced Time Series

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

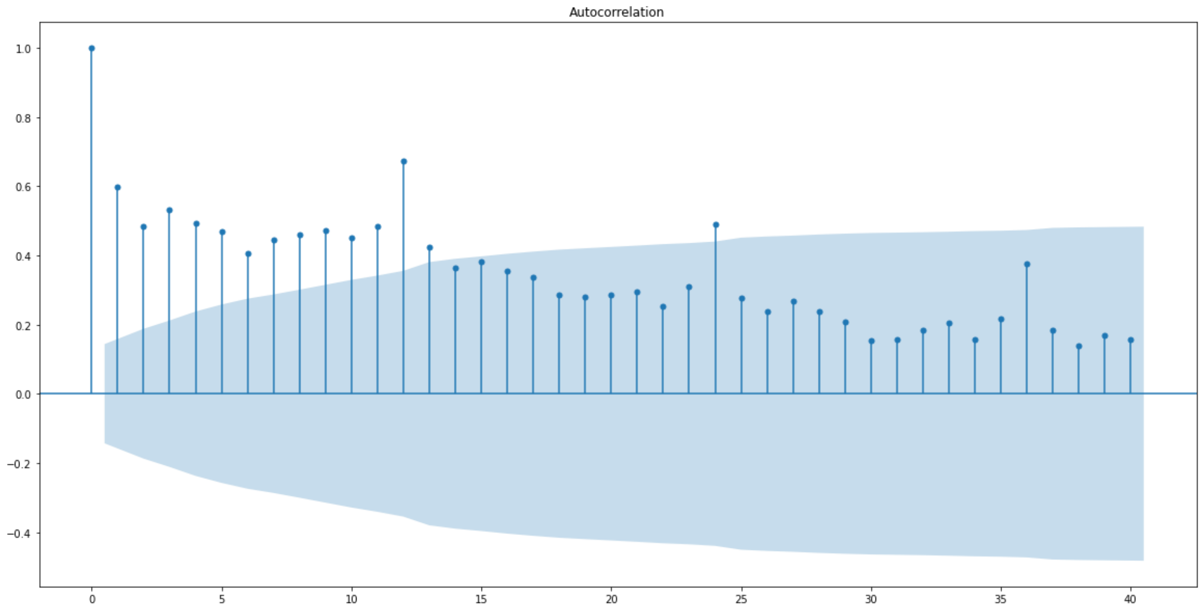


Figure 34. Autocorrelation Plot of Whole Time Series

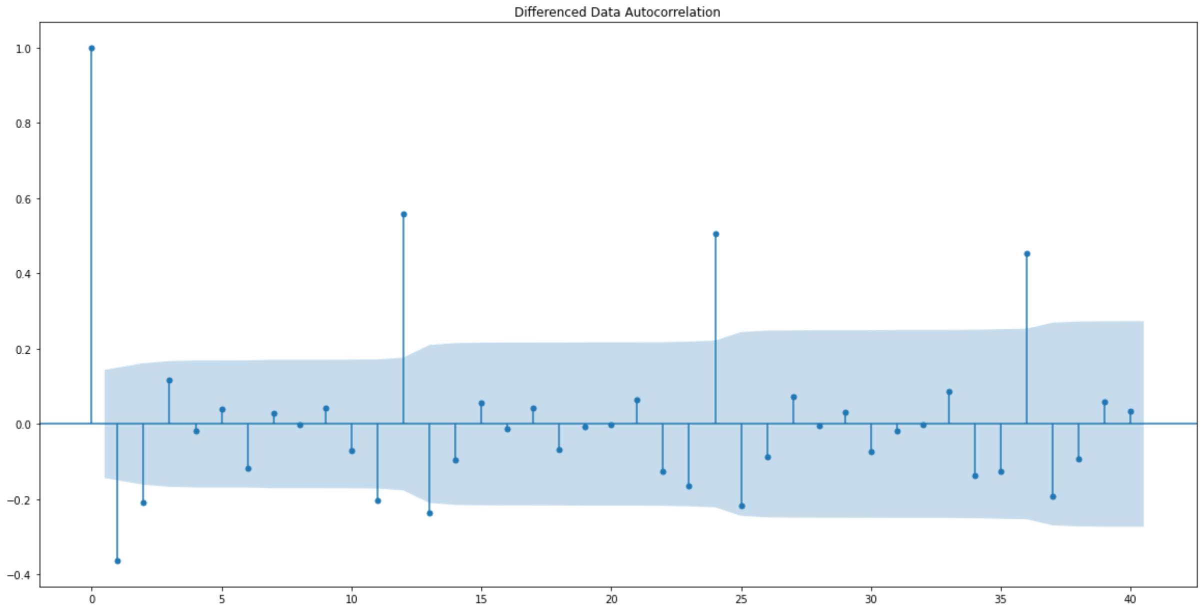


Figure 35. Autocorrelation Plot of Differenced Whole Time Series

From the above plot, it can be noticed that there is a seasonality in the time series.

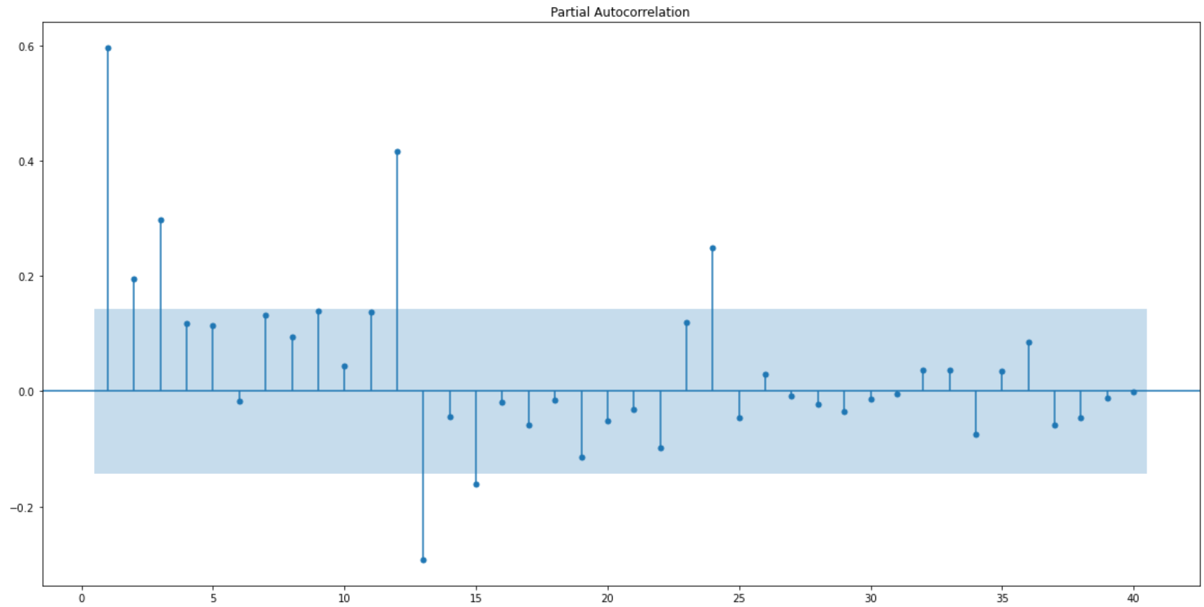


Figure 36. Partial Autocorrelation Plot of Whole Time Series

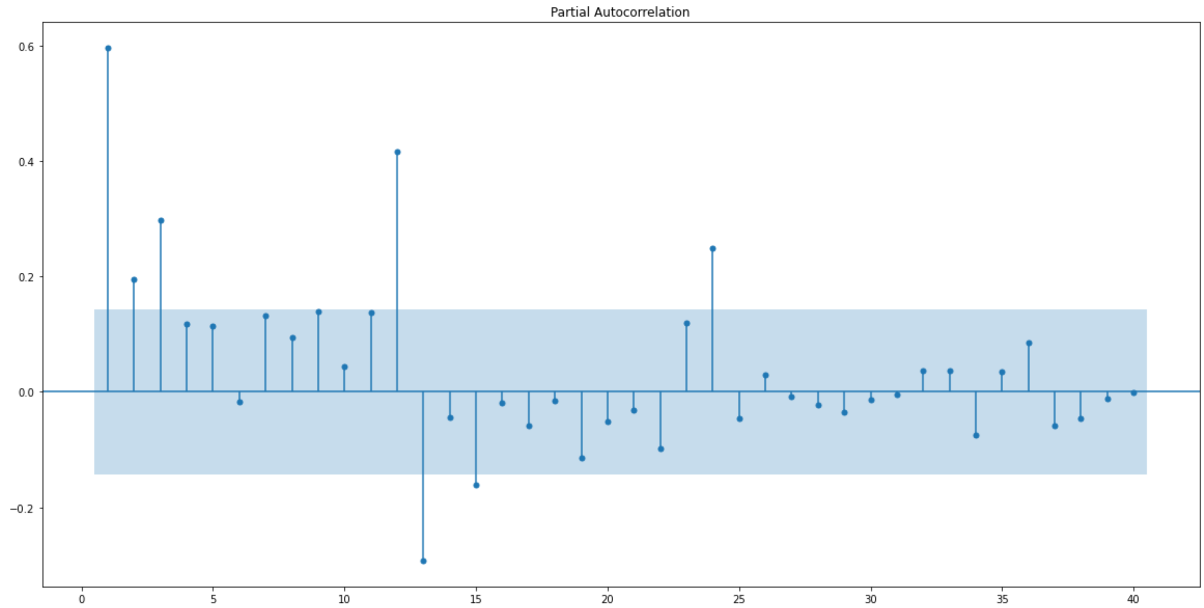


Figure 37. Partial Autocorrelation Plot of Differenced Whole Time Series

Check for Stationarity of Train Dataset

Results of Dickey-Fuller Test on Train Dataset

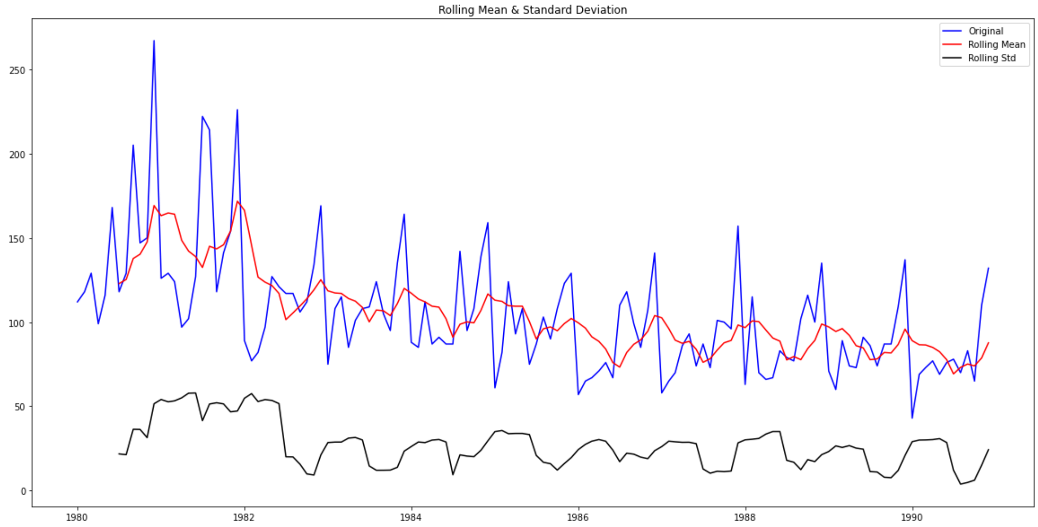
The Augmented Dickey-Fuller test is a unit root test that determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis 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

Significance level, α = 0.05

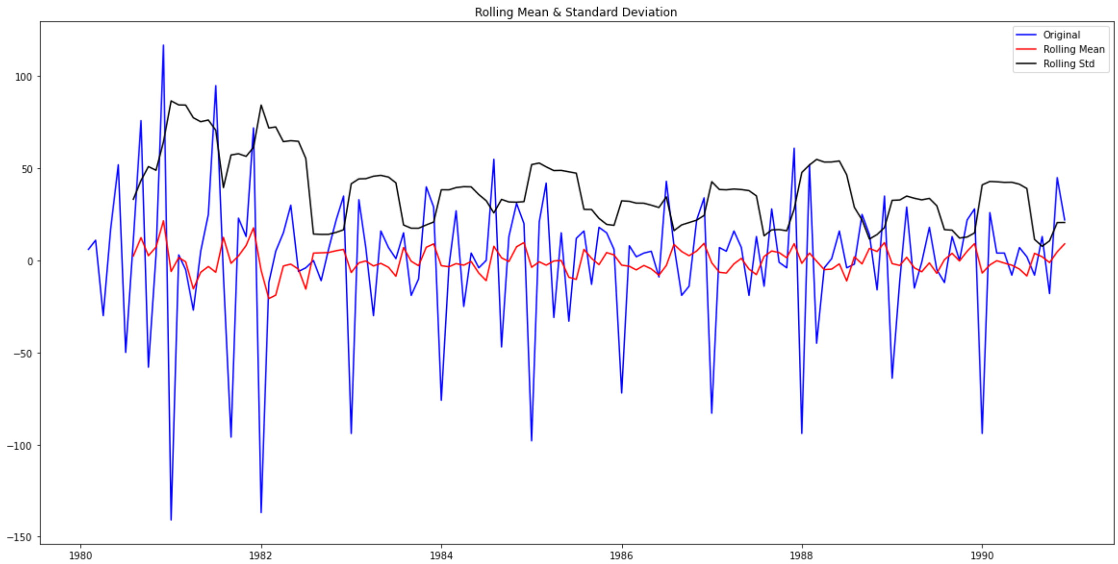



```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value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

As the P-value in Dicky-Fuller test is 0.219 which is more than the significance level (0.05), we failed to reject the null hypothesis. Hence, the training dataset **is non-stationary**.

To make time-series stationary, let us take first differencing on the training dataset and perform Dicky-Fuller test once again.

Results of Dickey-Fuller Test on Differenced Training Dataset



```
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
```

As the P-value in Dicky-Fuller test is 0 which is less than the significance level (0.05), we can reject the null hypothesis. Hence, **one level differenced training dataset is stationary**.

Plotting Differenced Time Series

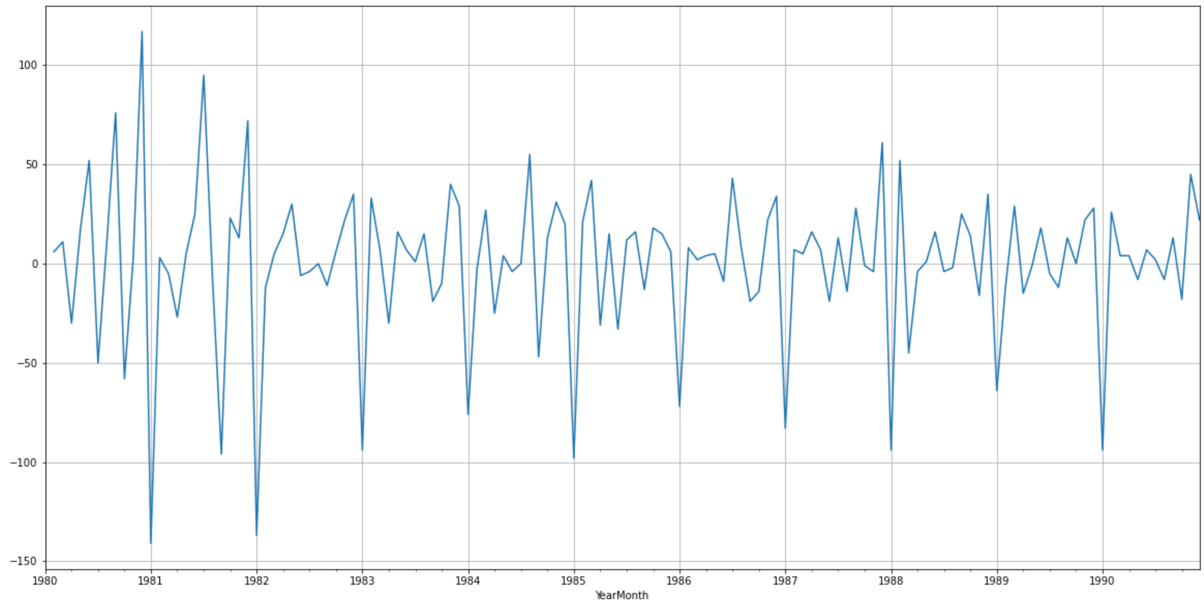


Figure 38. Plot of Differenced Training Dataset

Q6. 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.

Automated Version of ARIMA Model

The following values are considered for the parameters p, d, and q to build to various ARIMA models.

Autoregressive Component, p = 0, 1, 2, 3

Moving average component, q = 0, 1, 2, 3

Differencing Component, d = 1

order = [(0, 1, 0), (0, 1, 1), (0, 1, 2), (0, 1, 3), (1, 1, 0), (1, 1, 1), (1, 1, 2), (1, 1, 3), (2, 1, 0), (2, 1, 1), (2, 1, 2), (2, 1, 3), (3, 1, 0), (3, 1, 1), (3, 1, 2), (3, 1, 3)]

Various ARIMA models are built and fitted by using above order values and for every model AIC value is calculated. The ARIMA model with lowest AIC value is considered as the best model. Again, the best ARIMA model is built by using corresponding order parameters for forecasting purpose.

	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

Table 40. Top five ARIMA models with low AIC values.

ARIMA model with order (2, 1, 3)

Summary

Dep. Variable:	Rose	**No. Observations:**	132
Model:	ARIMA(2, 1, 3)	**Log Likelihood**	-631.348
Date:	Sun, 16 Jan 2022	**AIC**	1274.695
Time:	22:59:41	**BIC**	1291.947
Sample:	01-01-1980	**HQIC**	1281.705
	- 12-01-1990		
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 41. Summary of an automated ARIMA model.

Diagnostics Plots

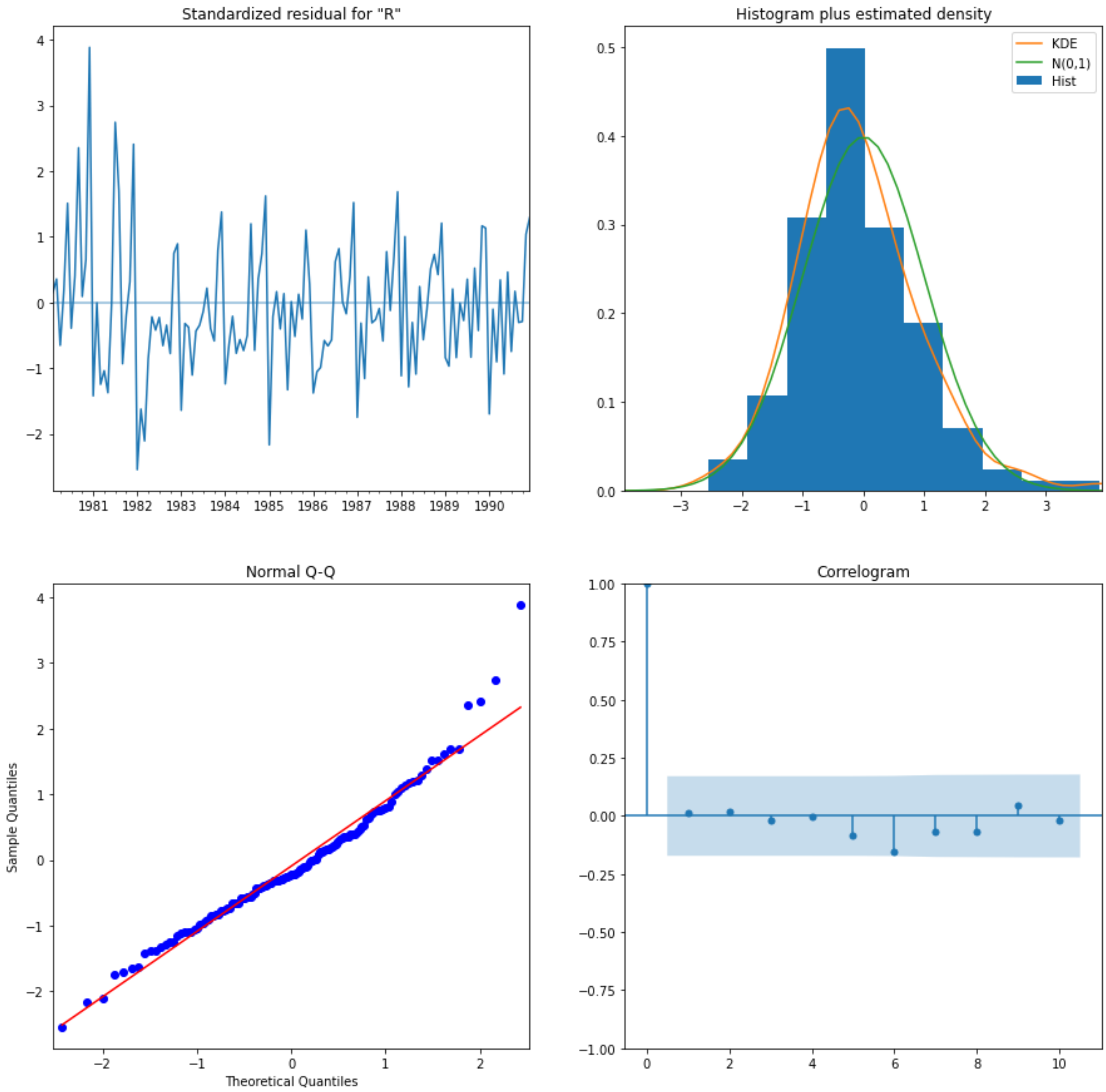


Figure 39. Diagnostics Plots of an automated ARIMA model.

Insights

- Autoregressive model - Lag 1 series has Z value (-19.999). It means that the forecast for this month is largely influenced by last month value.

YearMonth	Rose	forecast_ARIMA_auto
1991-01-01	54.0	85.6
1991-02-01	55.0	90.5
1991-03-01	66.0	82.0
1991-04-01	65.0	92.7
1991-05-01	60.0	80.9
1991-06-01	65.0	92.9
1991-07-01	96.0	81.4
1991-08-01	55.0	92.0
1991-09-01	71.0	82.6
1991-10-01	63.0	90.6

Table 42. Sample of Forecasted Sales in an automated ARIMA Model.

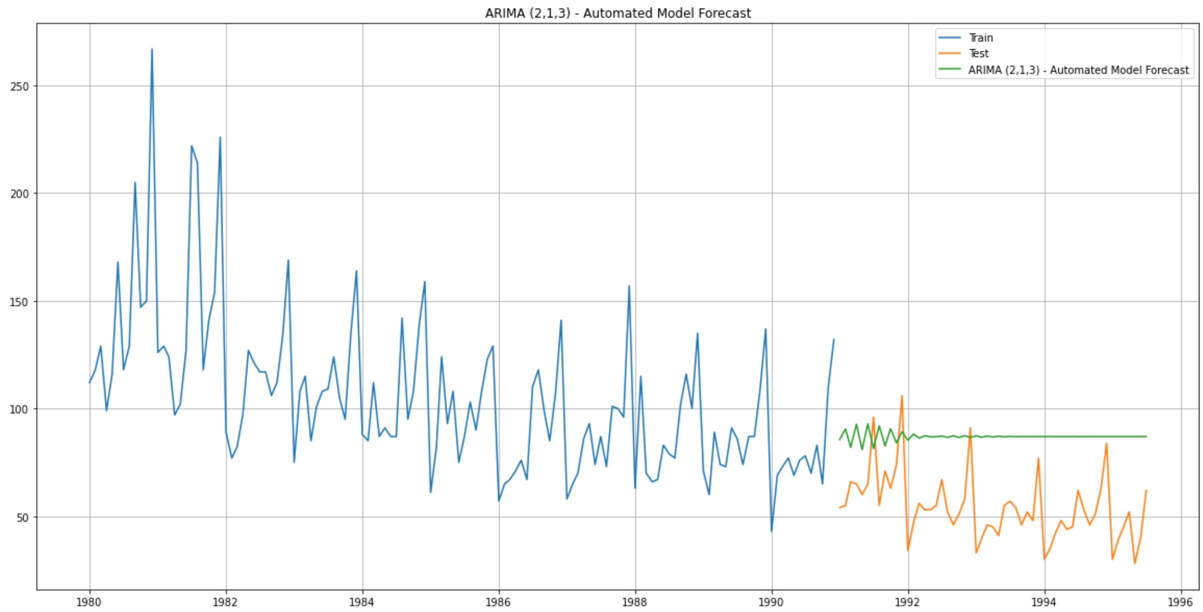


Figure 40. Plot of Forecasted Sales in an automated ARIMA model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the automated ARIMA model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in an automated ARIMA model is 36.4

Automated Version of SARIMA Model

The following values are considered for the parameters (p, d, q), (P, D, Q, s) to build to various ARIMA models.

Autoregressive Component, p = 0, 1, 2, 3

Moving average component, q = 0, 1, 2, 3

Differencing Component, d = 1

Seasonal Autoregressive Component, P = 0, 1, 2, 3

Seasonal Moving average component, Q = 0, 1, 2, 3

Seasonal Differencing Component, D = 0, 1

Periodicity, s = 6

order = [(0, 1, 0), (0, 1, 1), (0, 1, 2), (0, 1, 3), (1, 1, 0), (1, 1, 1), (1, 1, 2), (1, 1, 3), (2, 1, 0), (2, 1, 1), (2, 1, 2), (2, 1, 3), (3, 1, 0), (3, 1, 1), (3, 1, 2), (3, 1, 3)]

Seasonal order = [(0, 0, 0, 6), (0, 0, 1, 6), (0, 0, 2, 6), (0, 0, 3, 6), (0, 1, 0, 6), (0, 1, 1, 6), (0, 1, 2, 6), (0, 1, 3, 6), (1, 0, 0, 6), (1, 0, 1, 6), (1, 0, 2, 6), (1, 0, 3, 6), (1, 1, 0, 6), (1, 1, 1, 6), (1, 1, 2, 6), (1, 1, 3, 6), (2, 0, 0, 6), (2, 0, 1, 6), (2, 0, 2, 6), (2, 0, 3, 6), (2, 1, 0, 6), (2, 1, 1, 6), (2, 1, 2, 6), (2, 1, 3, 6), (3, 0, 0, 6), (3, 0, 1, 6), (3, 0, 2, 6), (3, 0, 3, 6), (3, 1, 0, 6), (3, 1, 1, 6), (3, 1, 2, 6), (3, 1, 3, 6)]

Various SARIMA models are built and fitted by using the above order values and for every model, the AIC value is calculated. The SARIMA model with the lowest AIC value is considered the best model. Again, the best SARIMA model is built by using corresponding order parameters for forecasting purposes.

	param	param_seasonal	AIC
375	(2, 1, 3)	(2, 1, 3, 6)	889.189817
503	(3, 1, 3)	(2, 1, 3, 6)	891.125985
511	(3, 1, 3)	(3, 1, 3, 6)	893.125640
367	(2, 1, 3)	(1, 1, 3, 6)	894.757072
127	(0, 1, 3)	(3, 1, 3, 6)	894.905687

Table 43. Top five SARIMA models with low AIC values.

SARIMA model with order (2, 1, 3) (2, 1, 3, 6)

Summary

Dep. Variable:	Rose	**No. Observations:**	132
Model:	SARIMAX(2, 1, 3)x(2, 1, 3, 6)	**Log Likelihood**	-433.595
Date:	Sun, 16 Jan 2022	**AIC**	889.190
Time:	22:37:46	**BIC**	918.172
Sample:	01-01-1980	**HQIC**	900.929
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5746	0.023	25.067	0.000	0.530	0.620
ar.L2	-0.9162	0.021	-43.468	0.000	-0.958	-0.875
ma.L1	-1.4571	31.539	-0.046	0.963	-63.273	60.359
ma.L2	1.5182	128.954	0.012	0.991	-251.228	254.264
ma.L3	-0.8409	86.128	-0.010	0.992	-169.648	167.966
ar.S.L6	-0.4347	0.106	-4.087	0.000	-0.643	-0.226
ar.S.L12	0.4834	0.102	4.734	0.000	0.283	0.684
ma.S.L6	-1.6588	3.472	-0.478	0.633	-8.464	5.147
ma.S.L12	-1.0800	9.161	-0.118	0.906	-19.036	16.876
ma.S.L18	1.5946	5.609	0.284	0.776	-9.398	12.588
sigma2	68.3643	7026.947	0.010	0.992	-1.37e+04	1.38e+04

Ljung-Box (L1) (Q):	0.03	**Jarque-Bera (JB):**	5.80
Prob(Q):	0.86	**Prob(JB):**	0.05
Heteroskedasticity (H):	0.45	**Skew:**	0.56
Prob(H) (two-sided):	0.02	**Kurtosis:**	3.29

Table 44. Summary of an automated SARIMA model.

Insights

- Autoregressive model - Lag 2 series has the highest Z value (-43.468). It means that forecast for this month is largely influenced by two months before value.
- Autoregressive model - Lag 1 series has the second-highest Z value (25.067). It means that forecast for this month is also influenced by last month value.

Diagnostics Plots

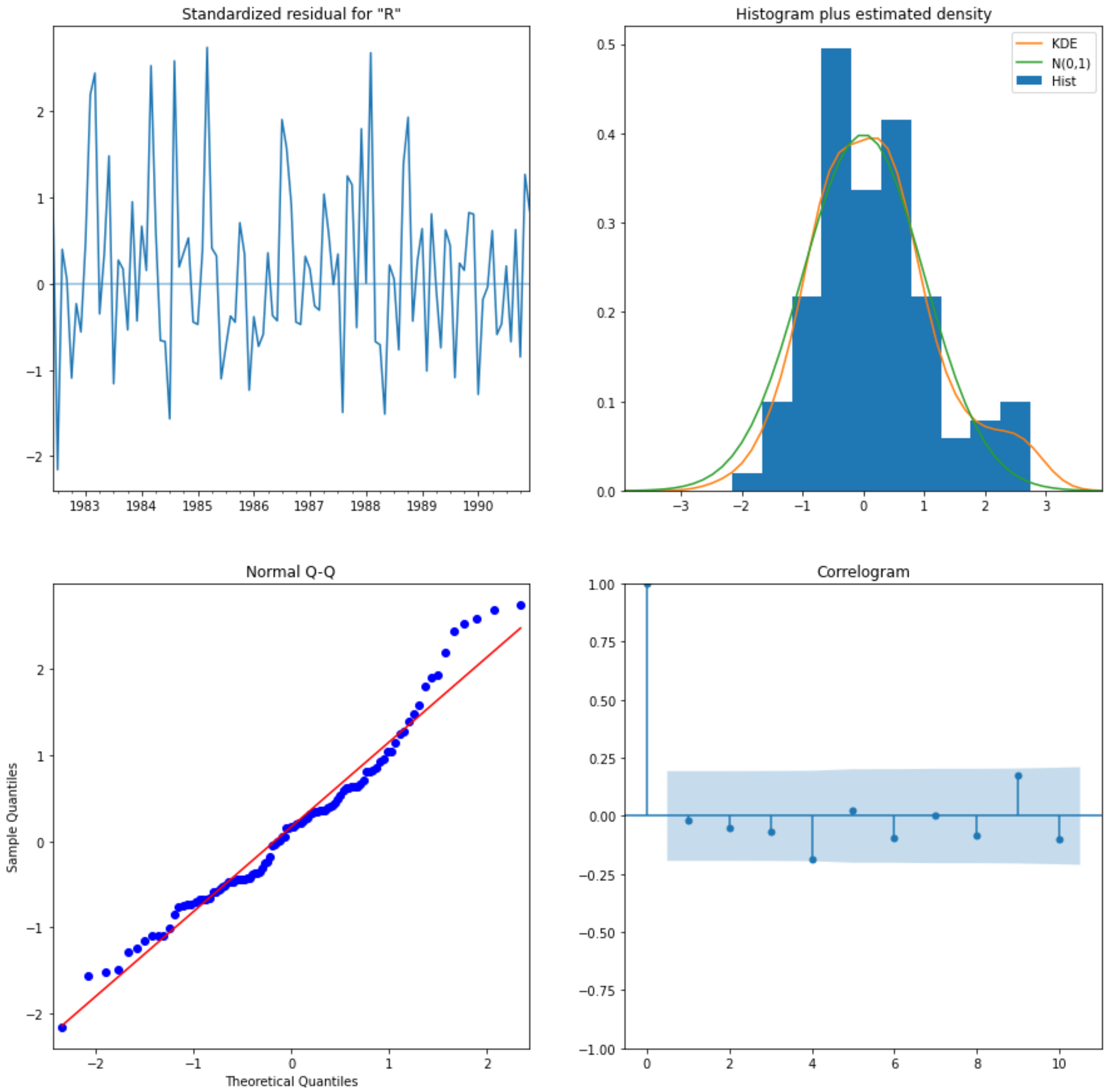


Figure 41. Diagnostics Plots of an automated SARIMA model.

YearMonth	Rose	forecast_ARIMA_auto	forecast_SARIMA_auto
1991-01-01	54.0	85.6	54.9
1991-02-01	55.0	90.5	64.0
1991-03-01	66.0	82.0	71.8
1991-04-01	65.0	92.7	69.1
1991-05-01	60.0	80.9	77.0
1991-06-01	65.0	92.9	80.3
1991-07-01	96.0	81.4	68.5
1991-08-01	55.0	92.0	77.2
1991-09-01	71.0	82.6	76.5
1991-10-01	63.0	90.6	76.5

Table 45. Sample of Forecasted Sales in an automated SARIMA Model.

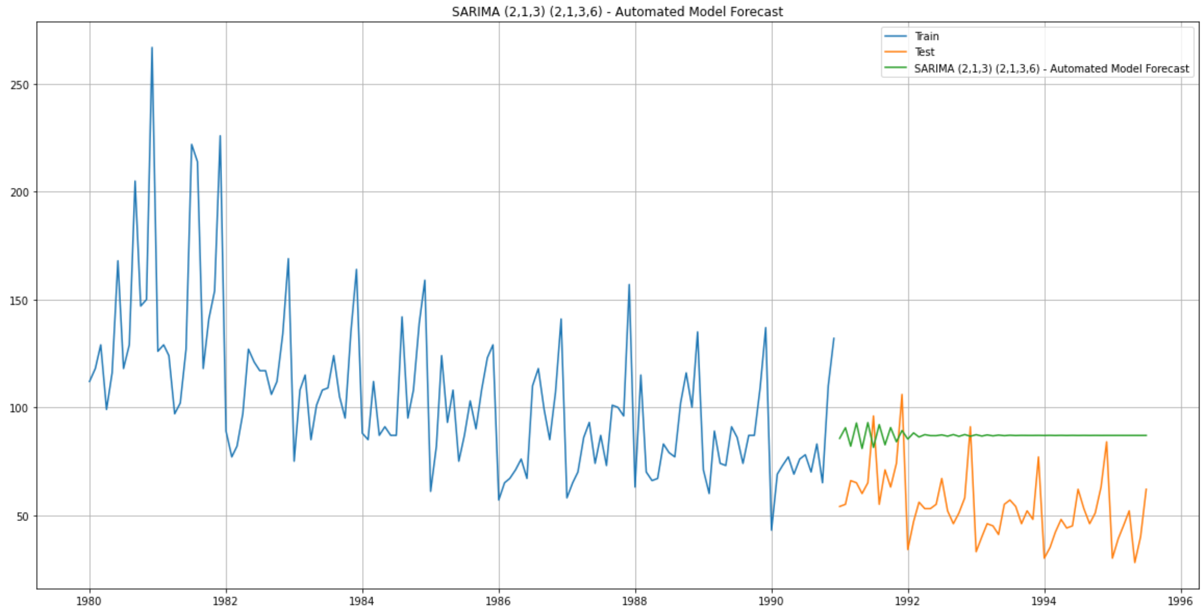


Figure 42. Plot of Forecasted Sales in an automated SARIMA model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the automated SARIMA model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in an automated SARIMA model is 16.68

Q7. 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.

Manual Version of ARIMA Model

The values for the parameters p and q are found by reading PACF and ACF plots of training data respectively.

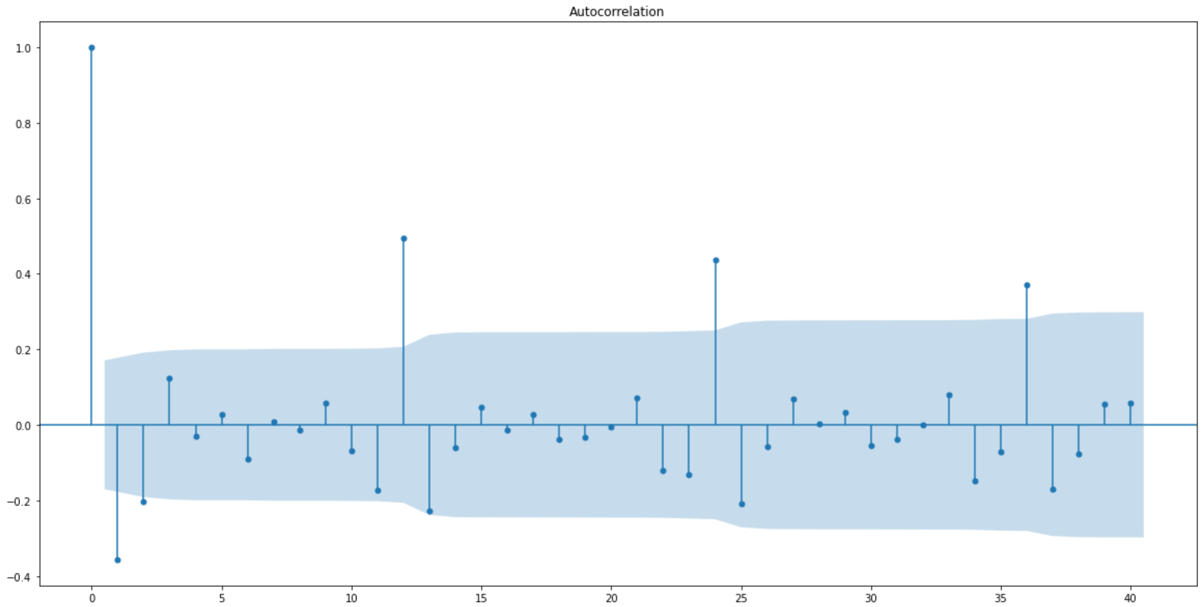


Figure 43. Autocorrelation Plot of Differenced Training Dataset

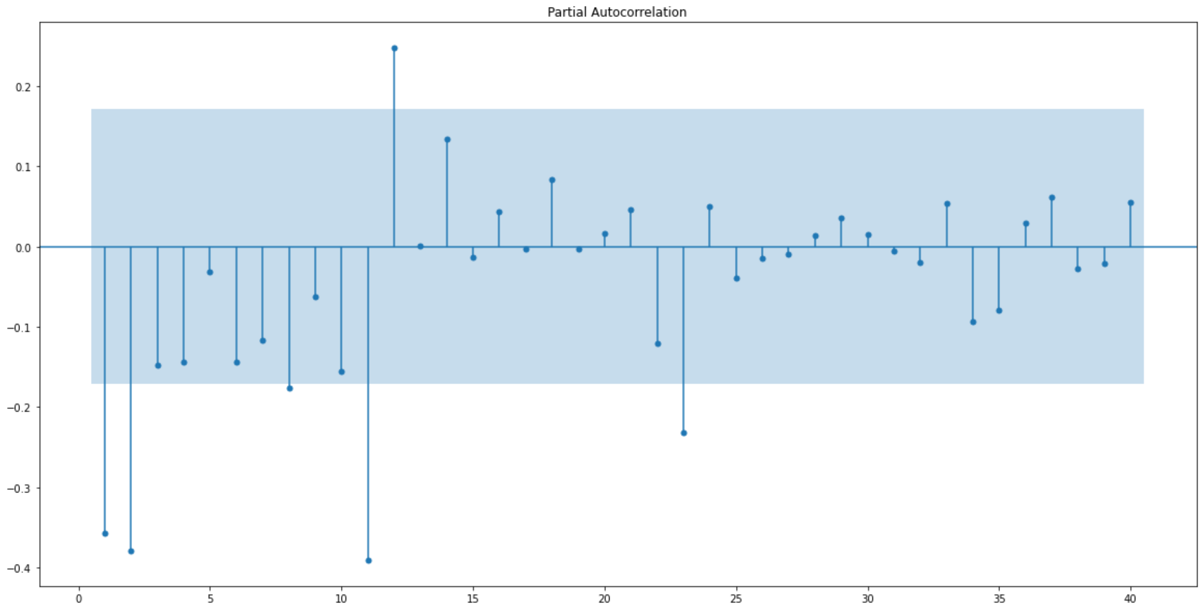


Figure 44. Partial Autocorrelation Plot of Differenced Training Dataset

Insights

- 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.
- Differencing Component, d is taken as 1 to make the time series stationarity.
- Final order to be considered is (2, 1, 2) to build the manual ARIMA model

Manual ARIMA model with order (2, 1, 2)

Summary

Dep. Variable:	Rose	**No. Observations:**	132
Model:	ARIMA(2, 1, 2)	**Log Likelihood**	-635.935
Date:	Sun, 16 Jan 2022	**AIC**	1281.871
Time:	22:37:49	**BIC**	1296.247
Sample:	01-01-1980	**HQIC**	1287.712
	- 12-01-1990		
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

Table 46. Summary of a Manual ARIMA model.

Diagnostics Plots

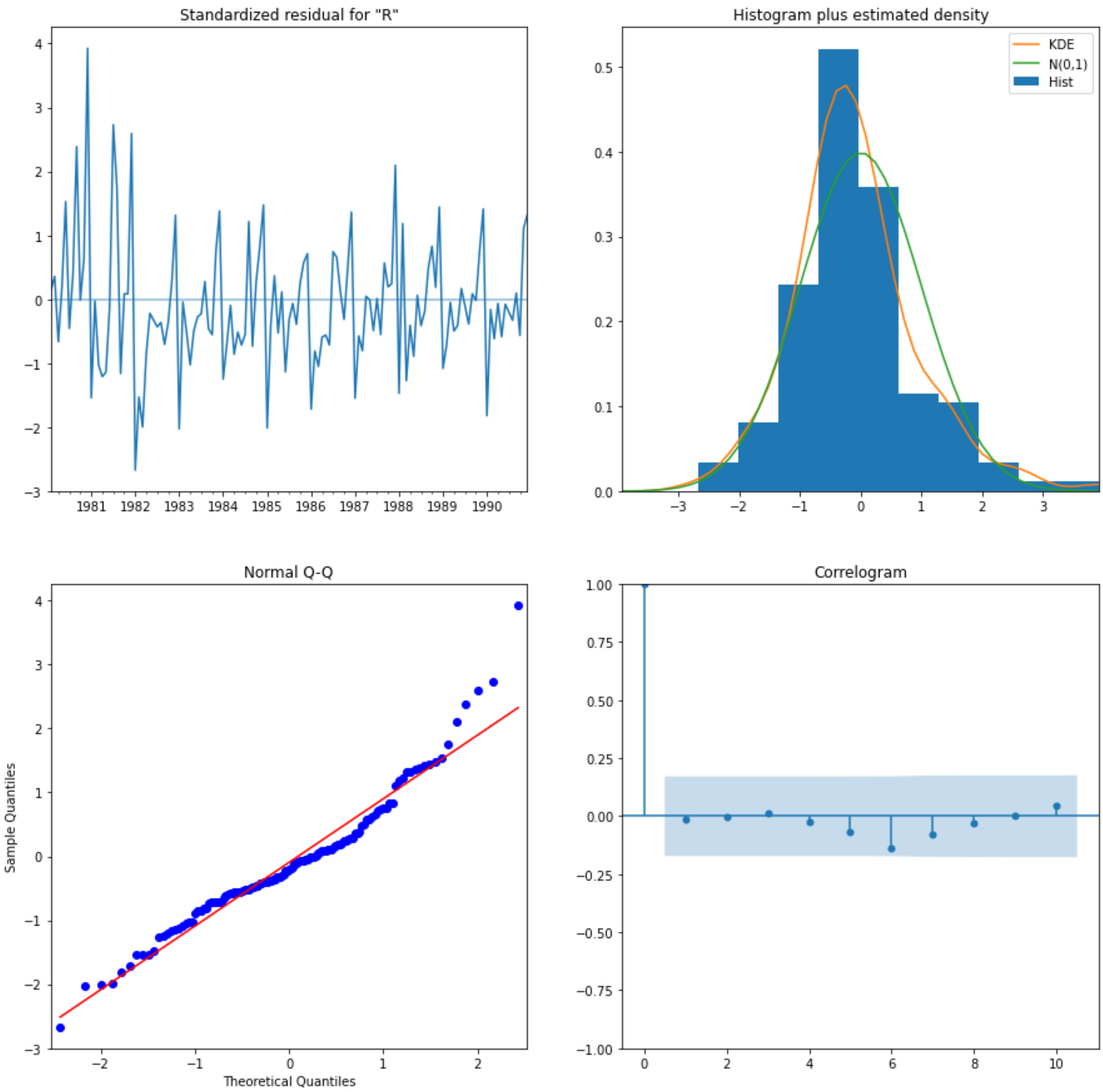


Figure 45. Diagnostics Plots of a Manual ARIMA model.

YearMonth	Rose	forecast_ARIMA_manual
1991-01-01	54.0	91.2
1991-02-01	55.0	85.3
1991-03-01	66.0	88.0
1991-04-01	65.0	86.8
1991-05-01	60.0	87.3
1991-06-01	65.0	87.1
1991-07-01	96.0	87.2
1991-08-01	55.0	87.1
1991-09-01	71.0	87.1
1991-10-01	63.0	87.1

Table 47. Sample of Forecasted Sales in a Manual ARIMA Model.

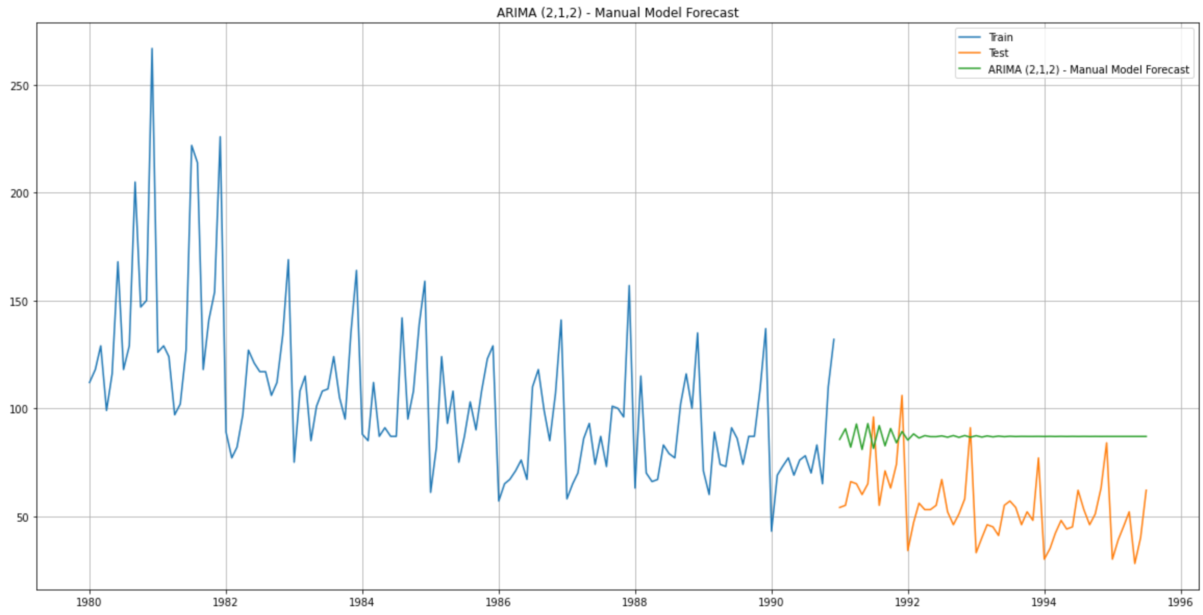


Figure 46. Plot of Forecasted Sales in a Manual ARIMA model.

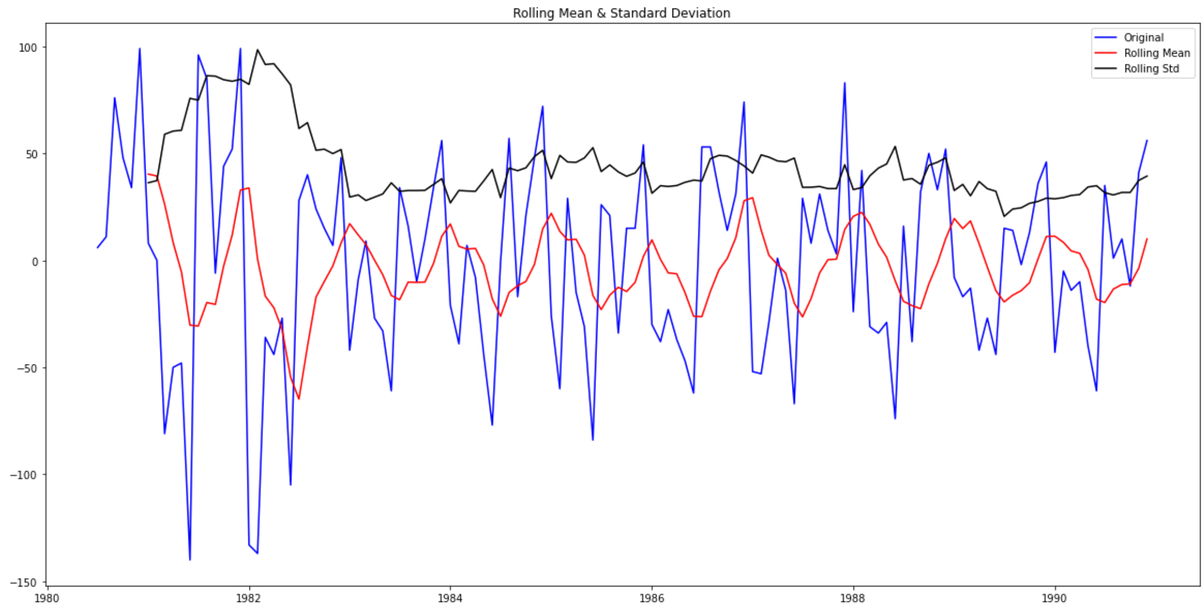
Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the Manual ARIMA model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in a Manual ARIMA model is 36.45

Manual Version of SARIMA Model

- The values for the parameters p and q are found by reading PACF and ACF plots of training data respectively.
- The values for the parameters P and Q are found by reading PACF and ACF plots of seasonally differenced training data respectively.

Results of Dickey-Fuller Test on Seasonal Differenced Training Dataset




```
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
```

As the P-value in Dicky-Fuller test is 0 which is less than the significance level (0.05), we can reject the null hypothesis. Hence, the **seasonal differenced training dataset is stationary**.

Plotting Seasonal Differenced Time Series

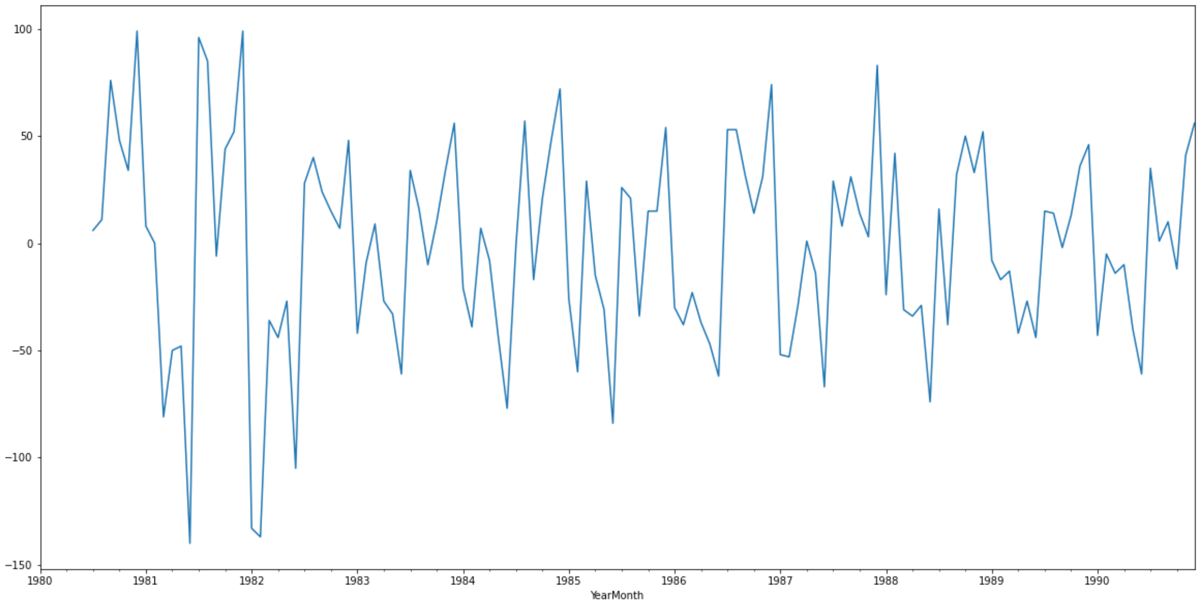


Figure 47. Plot of Seasonal Differenced Training Dataset

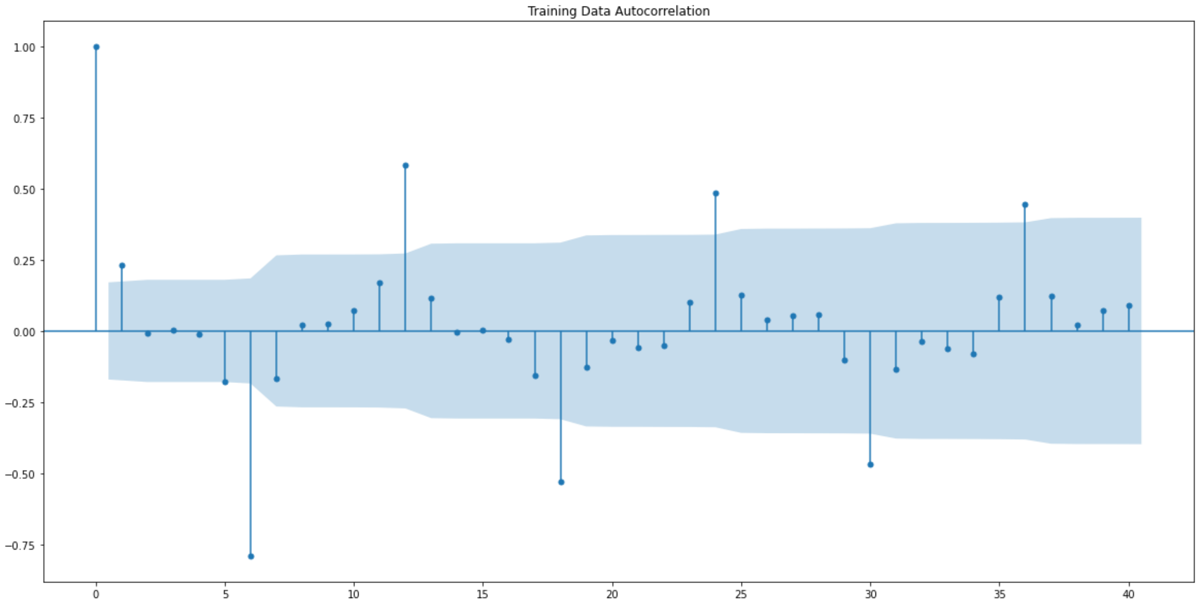


Figure 48. Autocorrelation Plot of Seasonal Differenced Training Dataset

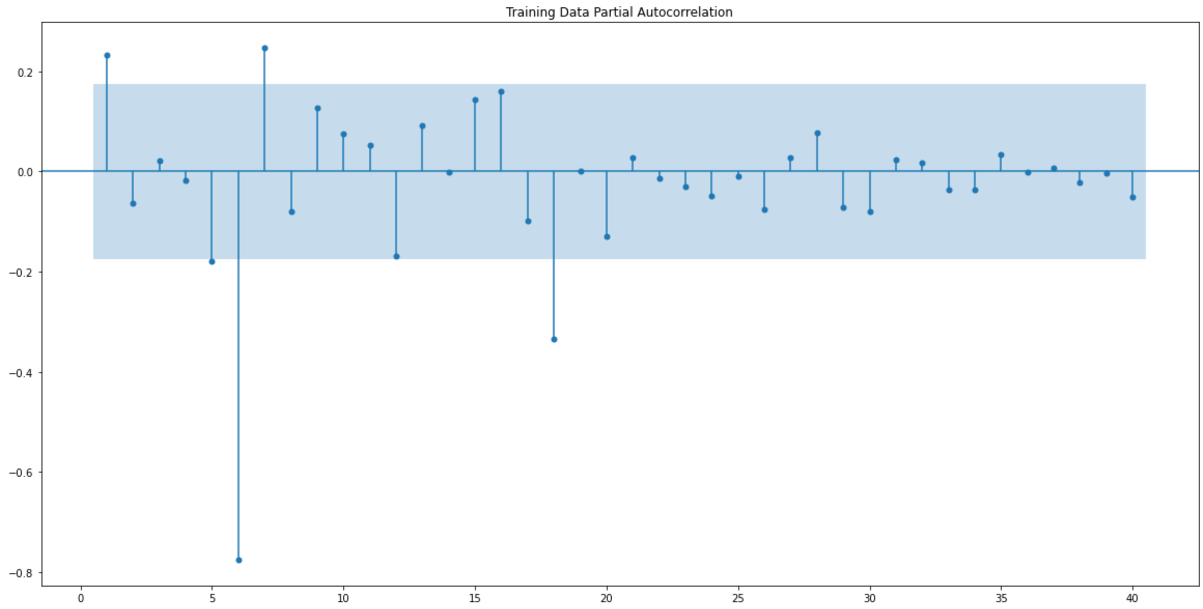


Figure 49. Partial Autocorrelation Plot of Seasonal Differenced Training Dataset

Insights

- We are taking the p-value to be 2 and the q value also to be 2 as the parameters same as the ARIMA model.
- Differencing Component, d is taken as 1 to make the time series stationarity.
- The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts off to 1.
- The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts off to 1.
- Periodicity, s is taken as 6.
- Final order to be considered is (2, 1, 2) (1, 0, 1, 6) to build the manual SARIMA model

Manual SARIMA model with order (2, 1, 2) (1, 0, 1, 6)

Summary

Dep. Variable:	Rose	**No. Observations:**	132
Model:	SARIMAX(2, 1, 2)x(1, 0, [1], 6)	**Log Likelihood**	-568.364
Date:	Sun, 16 Jan 2022	**AIC**	1150.728
Time:	22:37:52	**BIC**	1170.356
Sample:	01-01-1980	**HQIC**	1158.701
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.0205	0.076	13.370	0.000	0.871	1.170
ar.L2	-0.0423	0.077	-0.546	0.585	-0.194	0.109
ma.L1	-1.9988	967.983	-0.002	0.998	-1899.210	1895.213
ma.L2	1.0000	968.541	0.001	0.999	-1897.306	1899.306
ar.S.L6	-0.9492	0.020	-48.632	0.000	-0.987	-0.911
ma.S.L6	1.0000	968.540	0.001	0.999	-1897.304	1899.304
sigma2	516.6778	1.257	410.959	0.000	514.214	519.142

Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	6.37
Prob(Q):	0.94	**Prob(JB):**	0.04
Heteroskedasticity (H):	0.35	**Skew:**	0.13
Prob(H) (two-sided):	0.00	**Kurtosis:**	4.09

Table 48. Summary of a Manual SARIMA model.

Insights

- Seasonal Autoregressive model – Lag 6 series has the highest Z value (-48.632). It means that forecast for this month is largely influenced by the 6 months before value.
- Autoregressive model – Lag 1 series has the second-highest Z value (13.37). It means that forecast for this month is largely influenced by the previous month value.

Diagnostics Plots

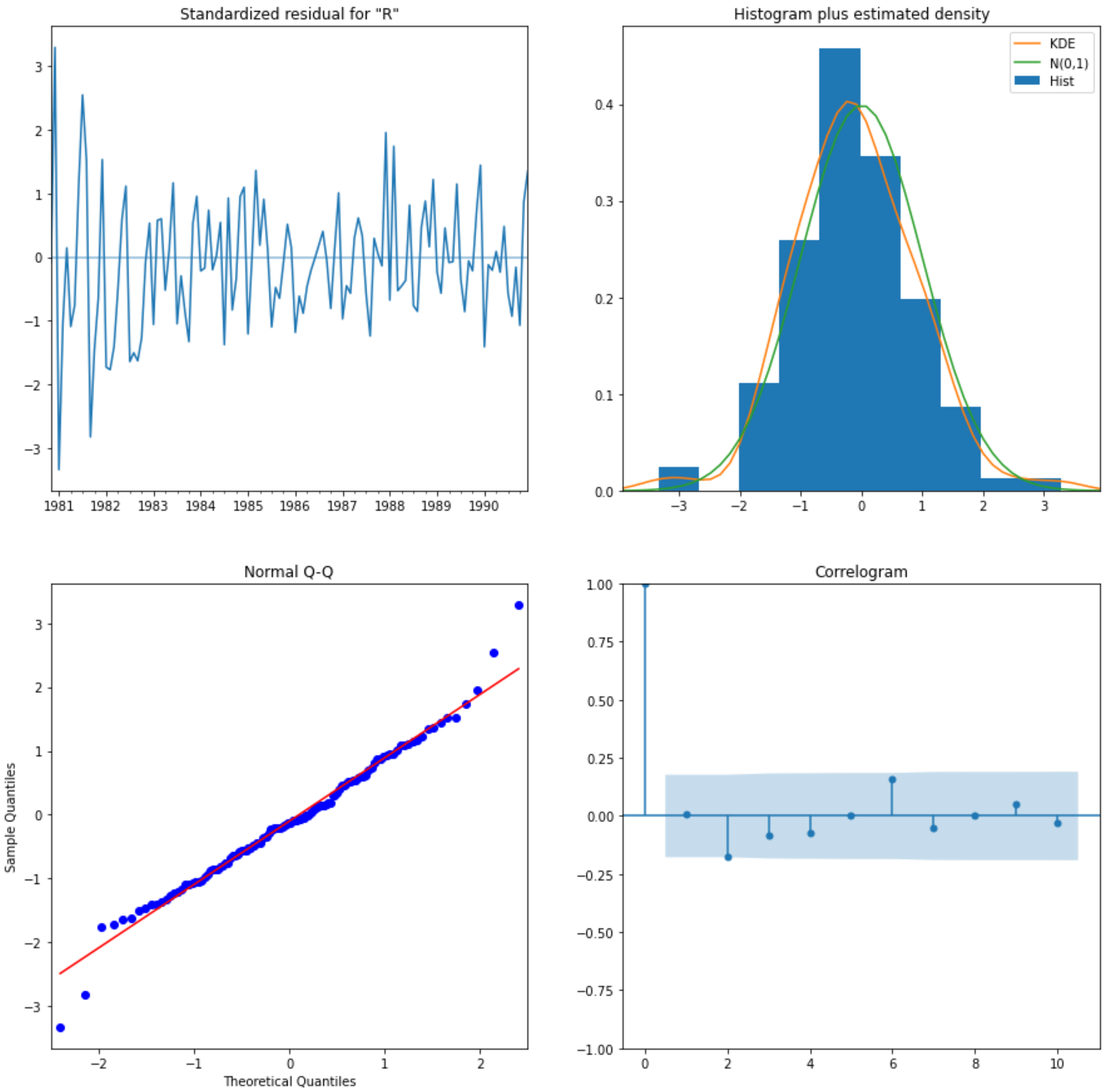


Figure 50. Diagnostics Plots of a Manual SARIMA model.

YearMonth	Rose	forecast_ARIMA_manual	forecast_SARIMA_manual
1991-01-01	54.0	91.2	75.8
1991-02-01	55.0	85.3	72.9
1991-03-01	66.0	88.0	77.3
1991-04-01	65.0	86.8	74.5
1991-05-01	60.0	87.3	74.2
1991-06-01	65.0	87.1	65.6
1991-07-01	96.0	87.2	88.9
1991-08-01	55.0	87.1	90.3
1991-09-01	71.0	87.1	85.9
1991-10-01	63.0	87.1	88.4

Table 49. Sample of Forecasted Sales in a Manual SARIMA Model.

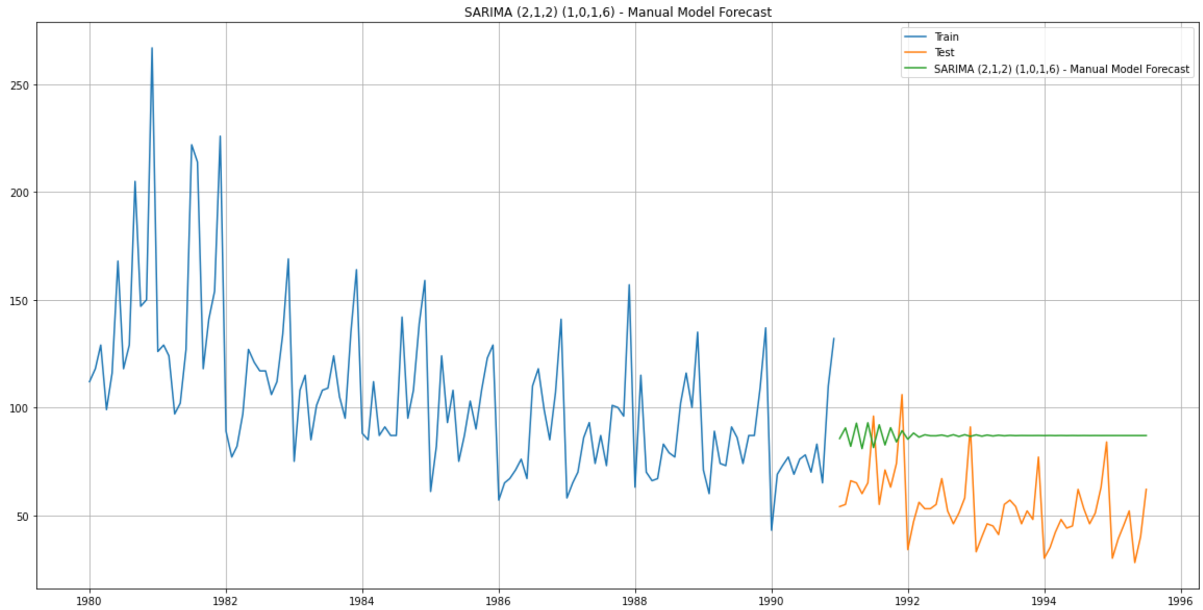


Figure 51. Plot of Forecasted Sales in a Manual SARIMA model.

Insights

- From the above plot, it can be noticed that **the plot of forecast sales is not matching with the actual plot of the test set**. Hence, the Manual SARIMA model may not be an appropriate model to forecast sales in this project. It will be concluded after comparing RMSE's of different forecast models in subsequent sections.
- RMSE in a Manual SARIMA model is 28.45

Q8. 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.

Model	RMSE
Linear_Regression	15.3
Naive_Forecast	79.3
Simple_Average	53.0
Simple_Exponential_Smoothing_optimized, Alpha =0.0987,	36.4
Simple_Exponential_Smoothing_iteration, Alpha =0.07	36.4
Double_Exponential_Smoothing_optimized, Alpha=0, Beta=0	15.3
Double_Exponential_Smoothing_iteration, Alpha=0.1, Beta=0.1	36.4
Triple Exponential_Smoothing_with additive trend & additive seasonality_optimized, Alpha=0.0895, Beta=0.0002, Gamma=0.0035	14.2
Triple Exponential_Smoothing_with additive trend & additive seasonality_iteration, Alpha=0.1, Beta=0.4, Gamma=0.3	12.3
Triple Exponential_Smoothing_with additive trend & multiplicative seasonality_optimized, Alpha=0.0715, Beta=0.0453, Gamma=0	19.7
Triple Exponential_Smoothing_with additive trend & multiplicative seasonality_iteration, Alpha=0.1, Beta=0.2, Gamma=0.1	9.3
Triple Exponential_Smoothing_with multiplicative trend & multiplicative seasonality_optimized, Alpha=0.0551, Beta=0.0316, Gamma=0.0003	19.5
Triple Exponential_Smoothing_with multiplicative trend & multiplicative seasonality_iteration, Alpha=0.2, Beta=0.7, Gamma=0.2	9.1
Triple Exponential_Smoothing_with multiplicative trend & additive seasonality_optimized, Alpha=0.0134, Beta=0, Gamma=0.0004	18.1
Triple Exponential_Smoothing_with multiplicative trend & additive seasonality_iteration, Alpha=0.1, Beta=0.8, Gamma=0.2	12.2
ARIMA_Automated (2,1,3)	36.4
ARIMA_ACF Plot (2,1,2)	36.5
SARIMA_Automated (2,1,3)(2,1,3,6)	16.7
SARIMA_ACF Plot (2,1,2)(1,0,1,6)	28.5

Let us sort the data frame by RMSE values to find the best model

Model	RMSE
Triple Exponential_Smoothing_with multiplicative trend & multiplicative seasonality_iteration, Alpha=0.2, Beta=0.7, Gamma=0.2	9.1
Triple Exponential_Smoothing_with additive trend & multiplicative seasonality_iteration, Alpha=0.1, Beta=0.2, Gamma=0.1	9.3
Triple Exponential_Smoothing_with multiplicative trend & additive seasonality_iteration, Alpha=0.1, Beta=0.8, Gamma=0.2	12.2
Triple Exponential_Smoothing_with additive trend & additive seasonality_iteration, Alpha=0.1, Beta=0.4, Gamma=0.3	12.3
Triple Exponential_Smoothing_with additive trend & additive seasonality_optimized, Alpha=0.0895, Beta=0.0002, Gamma=0.0035	14.2
Double_Exponential_Smoothing_optimized, Alpha=0, Beta=0	15.3
Linear_Regression	15.3
SARIMA_Automated (2,1,3)(2,1,3,6)	16.7
Triple Exponential_Smoothing_with multiplicative trend & additive seasonality_optimized, Alpha=0.0134, Beta=0, Gamma=0.0004	18.1
Triple Exponential_Smoothing_with multiplicative trend & multiplicative seasonality_optimized, Alpha=0.0551, Beta=0.0316, Gamma=0.0003	19.5
Triple Exponential_Smoothing_with additive trend & multiplicative seasonality_optimized, Alpha=0.0715, Beta=0.0453, Gamma=0	19.7
SARIMA_ACF Plot (2,1,2)(1,0,1,6)	28.5
Simple_Exponential_Smoothing_optimized, Alpha =0.0987,	36.4
ARIMA_Automated (2,1,3)	36.4
Simple_Exponential_Smoothing_iteration, Alpha =0.07	36.4
Double_Exponential_Smoothing_iteration, Alpha=0.1, Beta=0.1	36.4
ARIMA_ACF Plot (2,1,2)	36.5
Simple_Average	53.0
Naive_Forecast	79.3

Table 50. All Forecast Models with Respective Parameters and RMSE's

From the above table, it is evident that the best or most optimum model is the Triple Exponential Smoothing with Multiplicative Trend & Multiplicative seasonality with the Parameters $\alpha = 0.2$, $\beta = 0.7$ and $\gamma = 0.2$.

Q9. 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.

- Let us build Triple Exponential Smoothing with Multiplicative Trend & Multiplicative seasonality with the Parameters $\alpha = 0.2$, $\beta = 0.7$ and $\gamma = 0.2$ (most optimum model) on full data.
- RMSE obtained in full model is 20.73

	lower_CI	Forecast	upper_ci
1995-08-01	-3.3	37.3	78.0
1995-09-01	-5.2	35.4	76.0
1995-10-01	-4.8	35.8	76.4
1995-11-01	0.3	40.9	81.5
1995-12-01	15.9	56.5	97.2
1996-01-01	-19.0	21.6	62.2
1996-02-01	-13.7	26.9	67.5
1996-03-01	-9.4	31.2	71.8
1996-04-01	-8.6	32.0	72.6
1996-05-01	-13.8	26.8	67.4
1996-06-01	-10.4	30.2	70.8
1996-07-01	-6.7	33.9	74.5

Table 51. Forecasted Sales with 95% Confidence Interval.

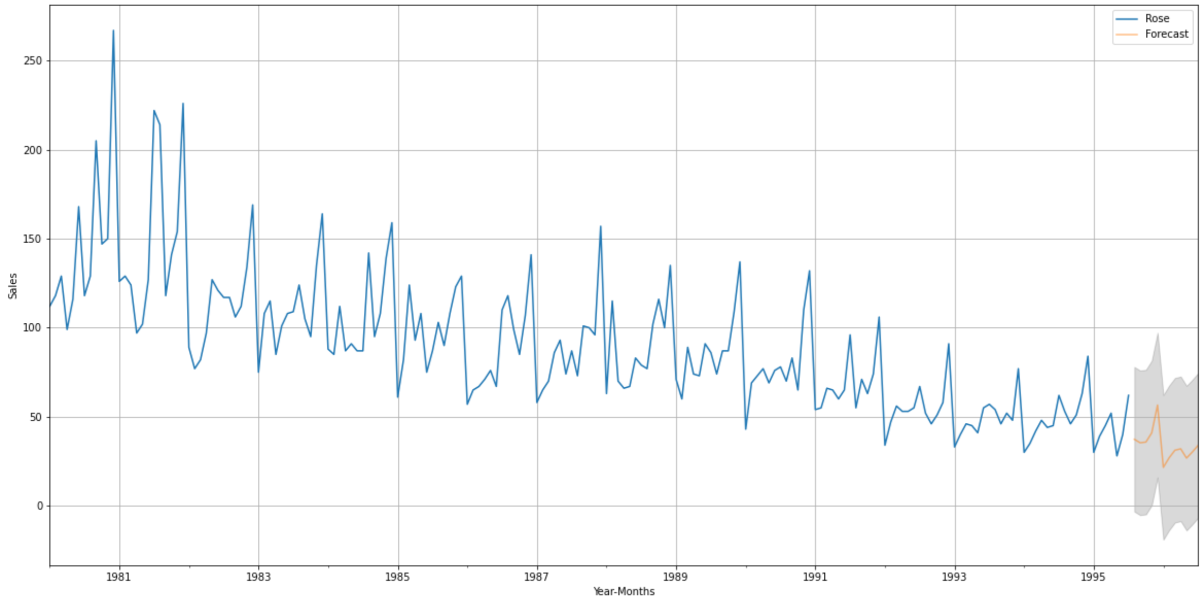


Figure 52. Plot of Forecasted Sales with 95% Confidence Interval.

Q10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

- The best **or most optimum model is the Triple Exponential Smoothing with Multiplicative Trend & Multiplicative seasonality** with the Parameters $\alpha = 0.2$, $\beta = 0.7$ and $\gamma = 0.2$.
- RMSE obtained in the most optimum model on the test data set is 9.1
- RMSE obtained when this model is fitted on the full dataset is 20.73
- The plot of 12 months forecasted sales is following the approximately same pattern as that of the original time series.
- It is expected that every December wines sales reach to maximum. Hence, the company should be ready with production to meet the demand at end of the year.
- The sales of wine have some outliers for certain years
- Mean wine sales in a year are decreasing continuously
- Total wine sales in a year are following almost the same pattern as that of the mean wine sales in a year.
- Mean sales are increasing gradually from quarter1 to quarter 4.
- Total wine sales in a quarter are following almost the same pattern as that of the mean wine sales in a quarter.
- Mean sales in a month are increasing gradually from January to December.

- Total sales in a month are following almost the same pattern as that of the mean sales in a month.