Time Series Forecasting

Coded Project

SRINIVASAN T

Table of Content:

List of Tables:

List of Figures:

Scoring guide (Rubric) - Some Rubric

Criteria	Points
Define the problem and perform Exploratory Data Analysis - Read the data as an appropriate time series data - Plot the data - Perform EDA - Perform Decomposition	9
Data Pre-processing - Missing value treatment - Visualize the processed data - Train-test split	4
Model Building - Original Data - Build forecasting models - Linear regression - Simple Average - Moving Average - Exponential Models (Single, Double, Triple) - Check the performance of the models built	15
Check for Stationarity - Check for stationarity - Make the data stationary (if needed)	4
Model Building - Stationary Data - Generate ACF & PACF Plot and find the AR, MA values. - Build different ARIMA models - Auto ARIMA - Manual ARIMA - Build different SARIMA models - Auto SARIMA - Manual SARIMA - Check the performance of the models built	12
Compare the performance of the models - Compare the performance of all the models built - Choose the best model with proper rationale - Rebuild the best model using the entire data - Make a forecast for the next 12 months	6
Actionable Insights & Recommendations - Conclude with the key takeaways (actionable insights and recommendations) for the business	4
Business Report Quality - Adhere to the business report checklist	6
Points	60

1. Define the problem and perform Exploratory Data Analysis

Problem Statement - TSF Project

Context

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties. Our objective is to delve into the data, analyse trends, patterns, and factors influencing wine sales over the course of the century. By leveraging data analytics and forecasting techniques, we aim to gain actionable insights that can inform strategic decision-making and optimize sales strategies for the future.

Objective

The primary objective of this project is to analyse and forecast wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

1.1 Read the data as an appropriate time series data

	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

Table 1.1: Reading the dataset (top 5 rows)

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

Table 1.2: Reading the dataset (bottom 5 rows)

Checking the Datatypes:

```
YearMonth     object
Rose         float64
dtype: object
```

Need to change the datatype to datatime64[ns] format using the parse_date function

```
YearMonth    datetime64[ns]
Rose                float64
dtype: object
```

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

Table 1.3: Final dataset with appropriate time series format

1.2 Plot the data

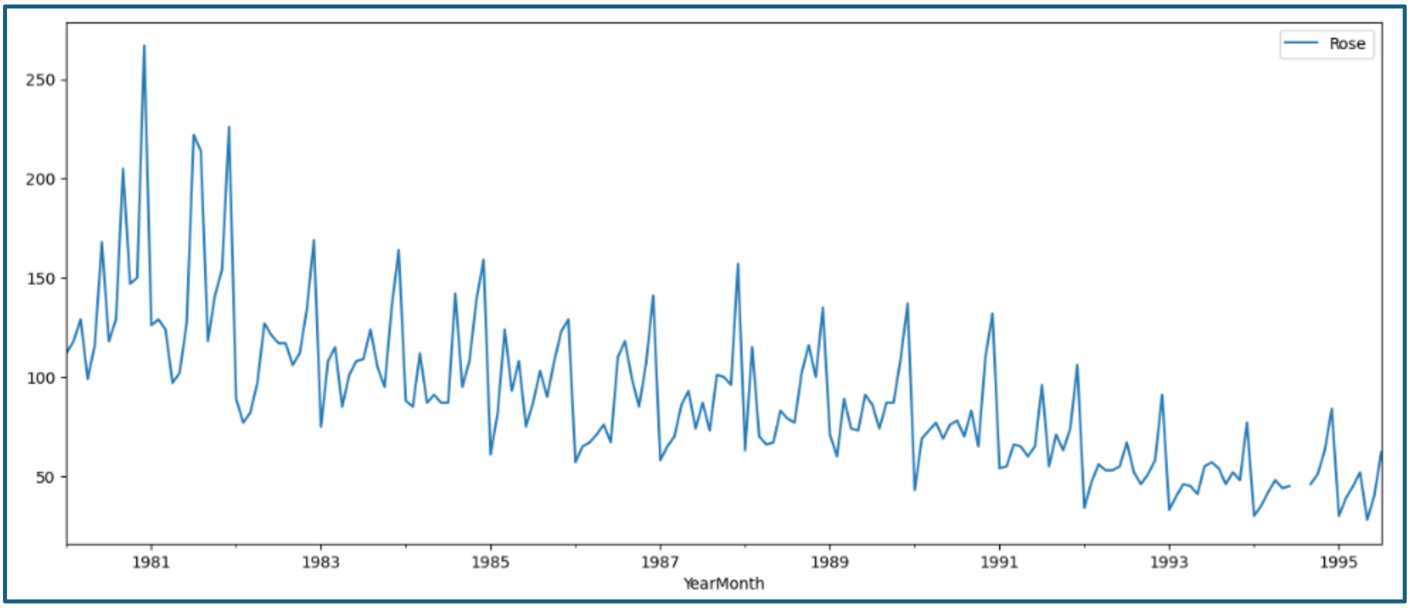


Figure 1.1: Rose Dataset plot

The data seems to have seasonality but not trend. 1987 - 19990 seems to have the highest sales across the given date range.

1.3 Perform EDA

1.3.1 Yearly Boxplot:

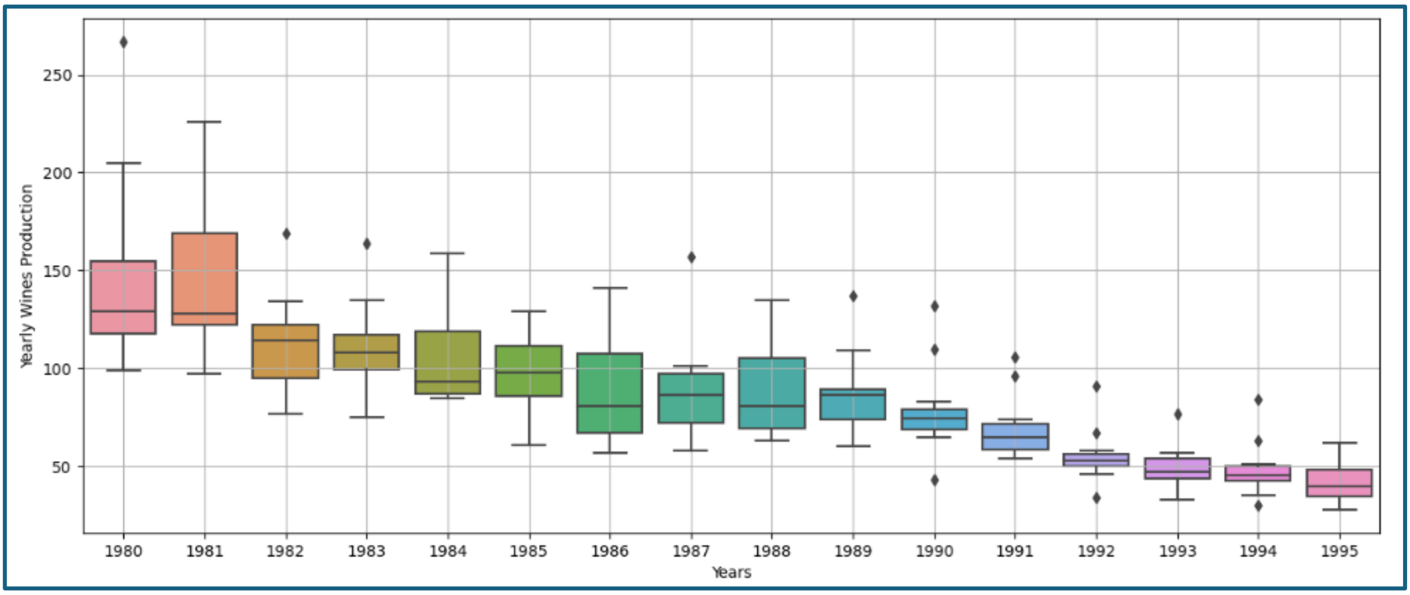


Figure 1.2: Yearly Boxplot

The yearly trend is decreasing with increase in years of 1985 & 1987. Year on year the sales of Rose wines goes down

1.3.2 Monthly Boxplot:

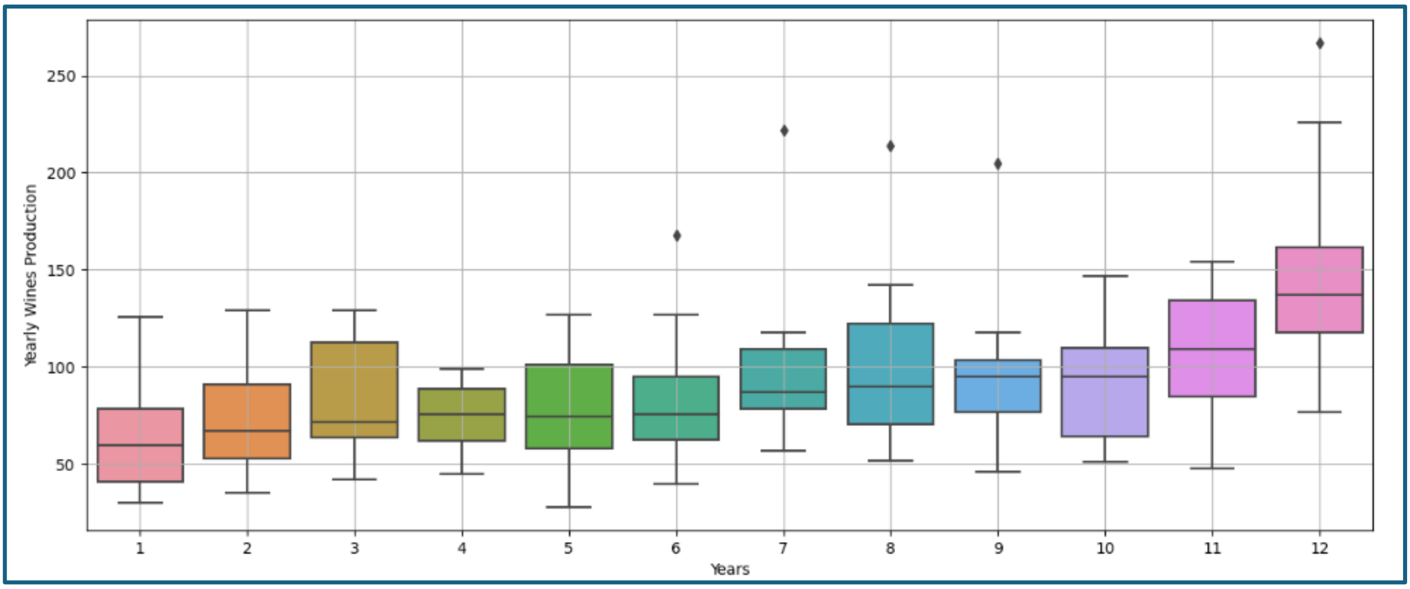


Figure 1.3: Monthly Boxplot

The monthly boxplot clearly shows that there is an increasing trend from Jan to Apr then steady trend from May & Jun and then increasing trend

1.3.3 Quarterly Line plot:

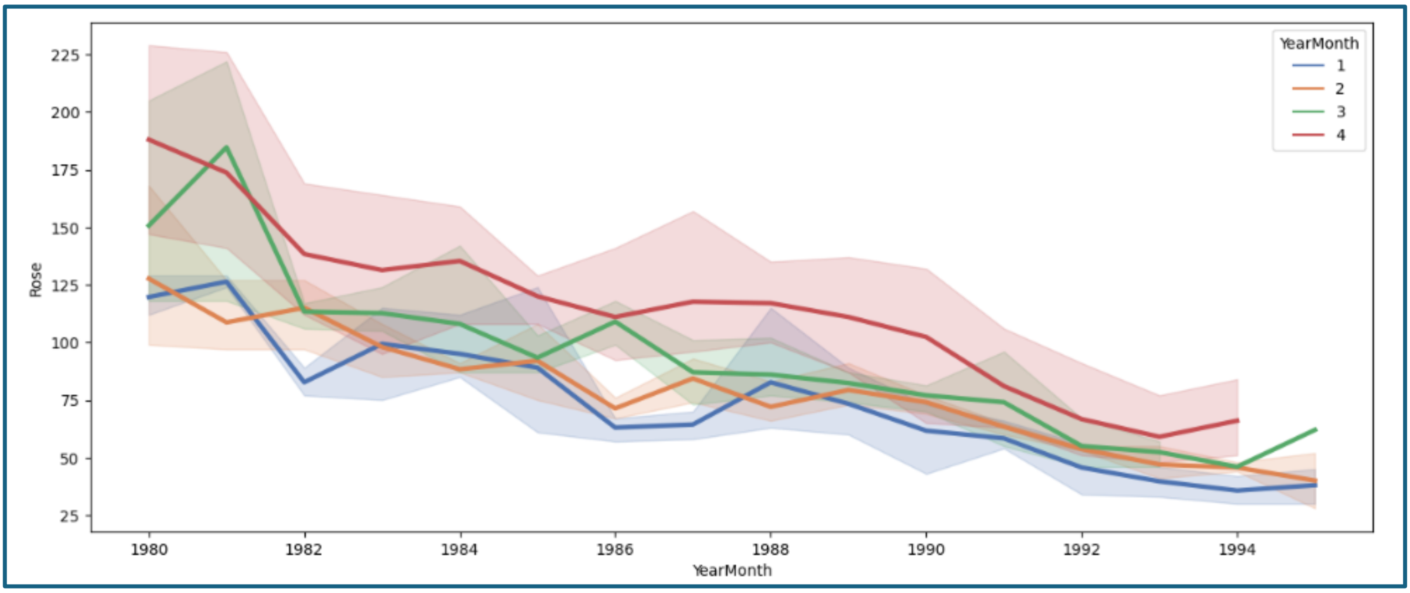


Figure 1.4: Quarterly Line plot

The Q4 having the highest sales across all the quarters.

1.3.4 Descriptive plot:

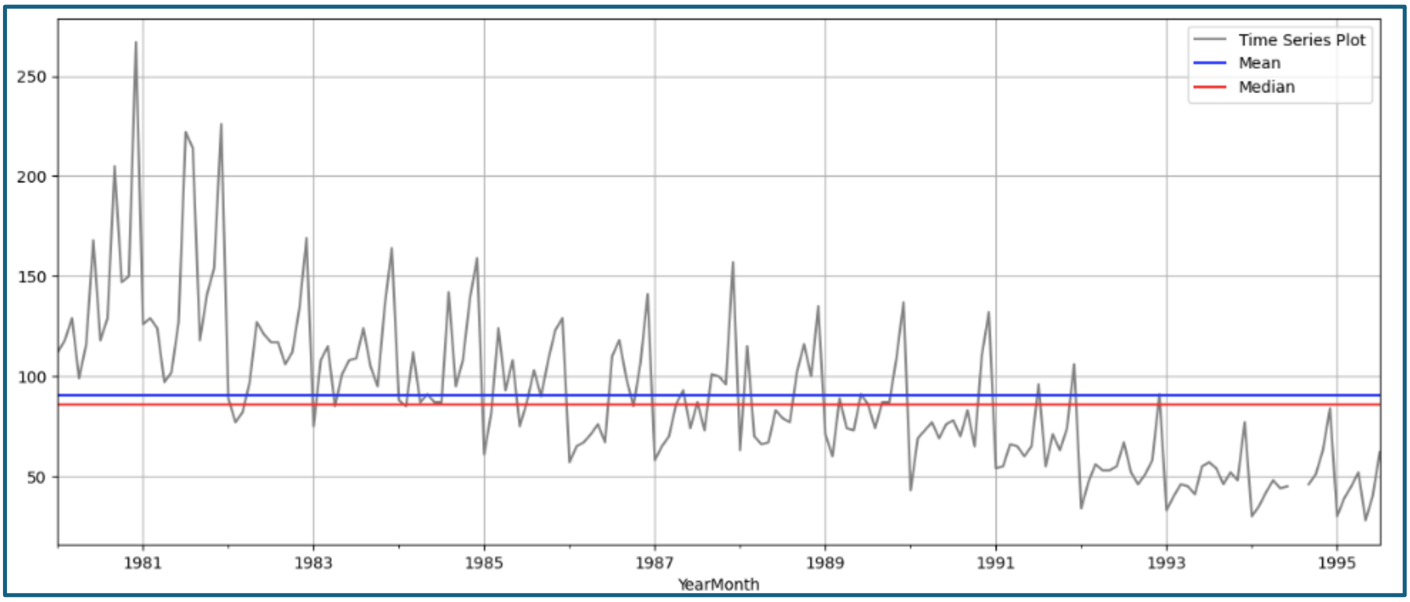


Figure 1.5: Descriptive plot

The mean is greater than median, obviously shows a right skewed plot.

1.4 Perform Decomposition

1.4.1 Additive Model

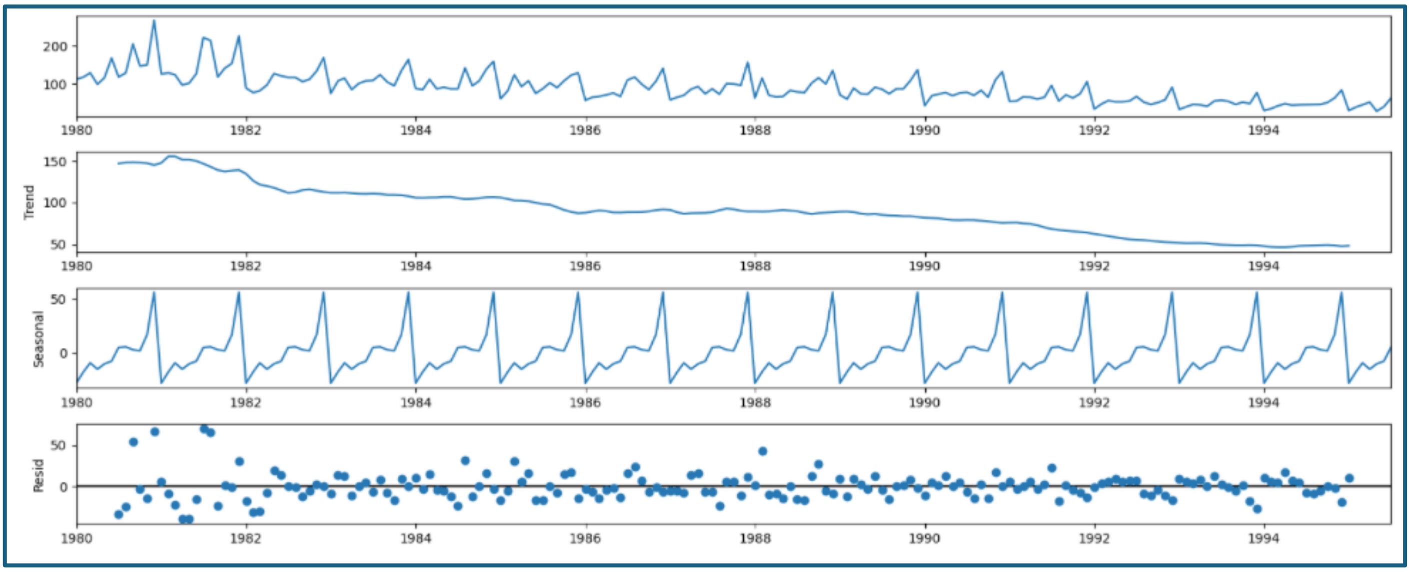


Figure 1.6: Additive Decomposition Model

The additive model shows that the original data does have a trend and have seasonality and have residual around 0 with some pattern in it.

1.4.2 Multiplicative Model

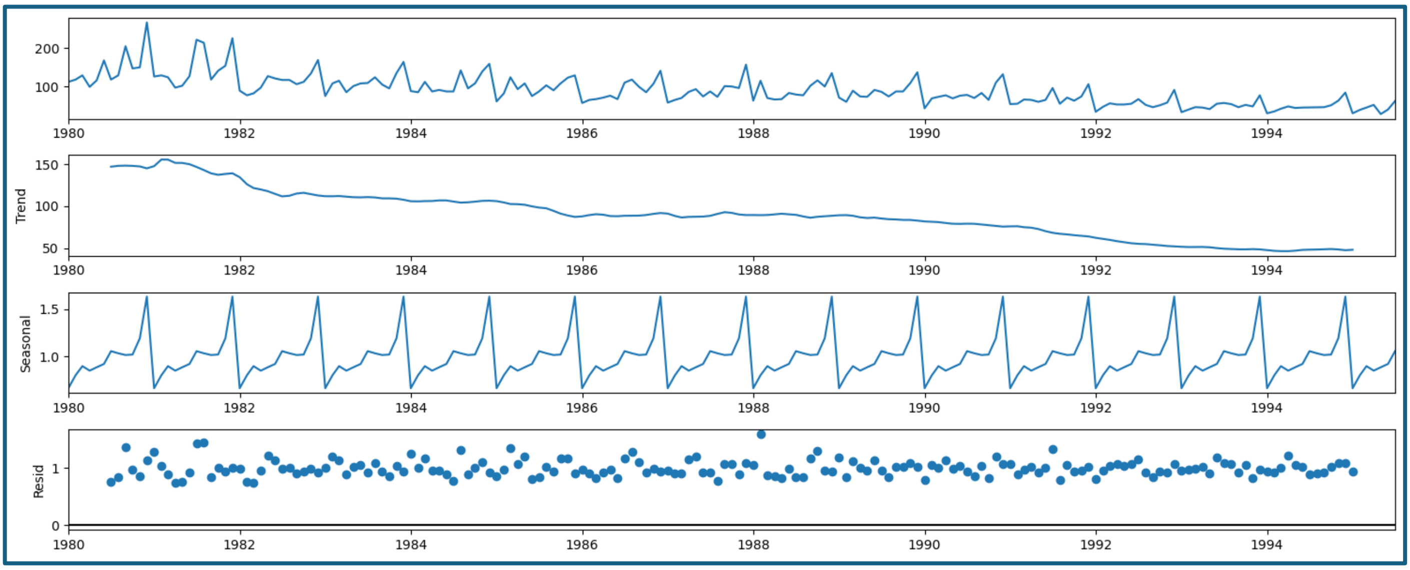


Figure 1.7: Multiplicative Decomposition Model

The multiplicative model shows that the original data does have trend and seasonality and have residual around -1 without any pattern in it.

2. Data Pre-processing

2.1 Missing value treatment

The missing values can be identified using the isnull function on the dataset.

```
Rose     2
dtype: int64
```

There are 2 null values

The Null values are treated using the interpolate function

2.2 Visualize the processed data

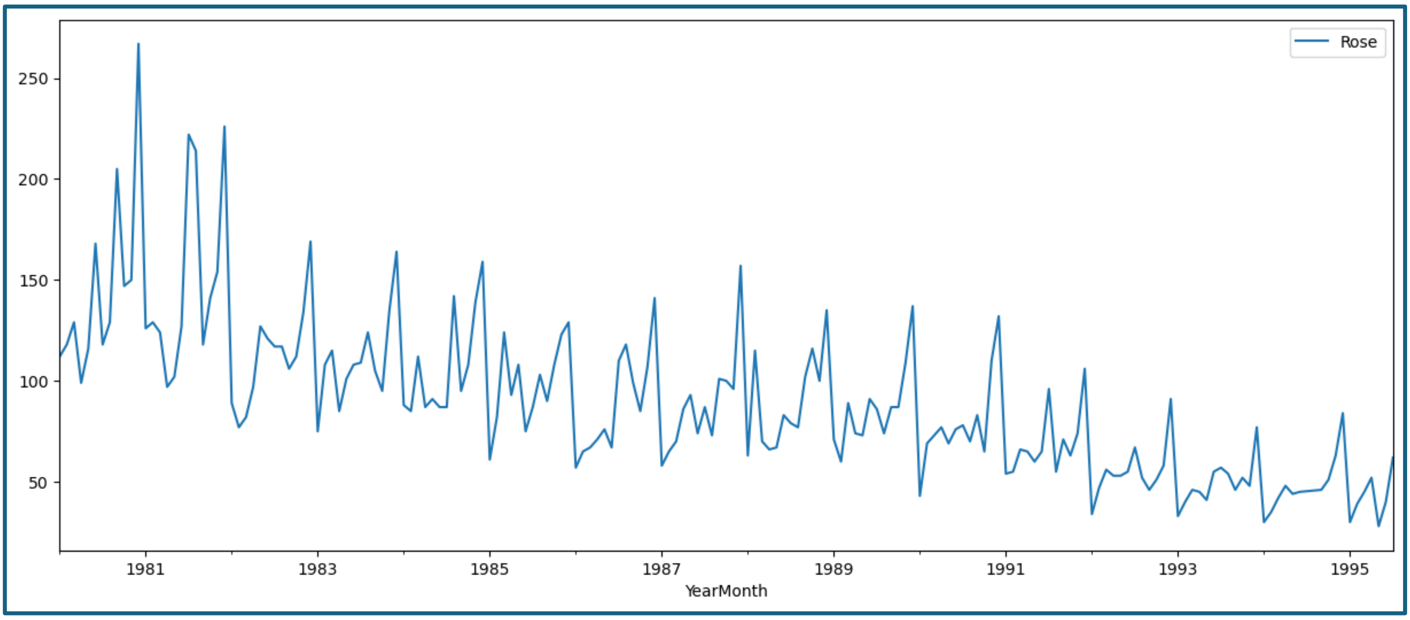


Figure 1.8: Processed Data

2.3 Train-test split

The train-test split happens in sequential manner, such that the first 70% of the data is considered as train in the same times series order and the last 30% again in the same sequential times series order is considered and test data. The train-test split cannot random as it is a time series format.

```
Train Shape - (130, 1) [train has 130 rows]
Test Shape - (57, 1)   [test has 57 rows]
```

2.3.1 Train-test split

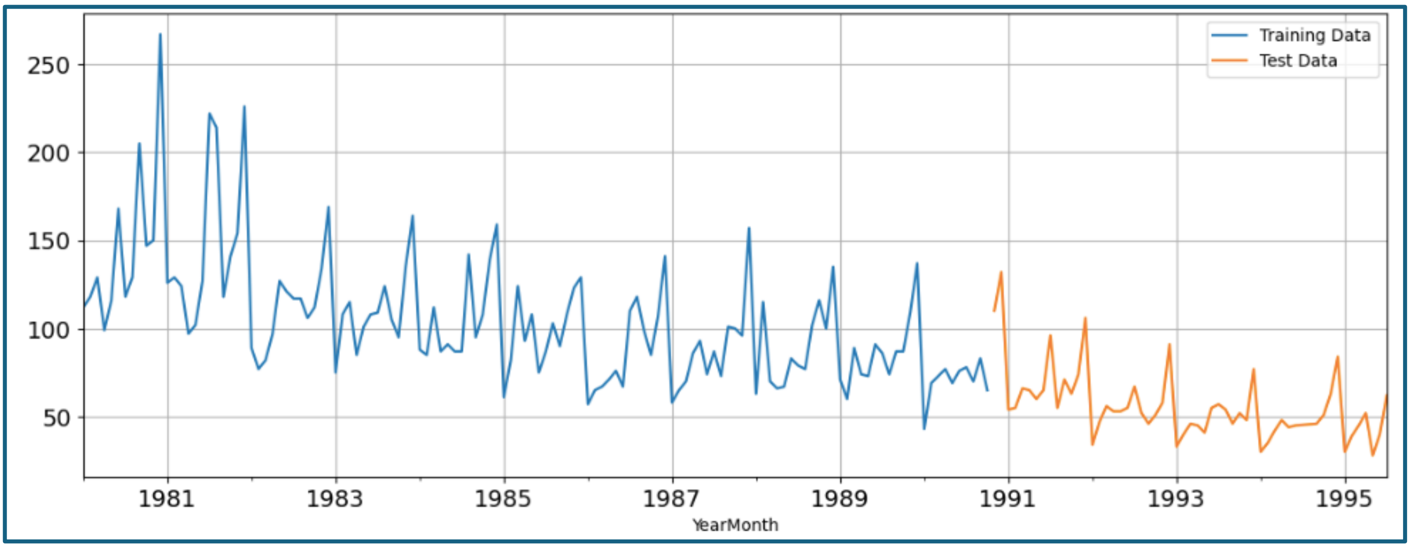


Figure 1.9: Train-test Data Plot

The plot clearly shows that the time series train-train split in a sequential order.

3. Model Building - Original Data

3.1 Build forecasting models

3.1.1 Linear regression

Before Linear Regression we will fitting the numerical time instance of the sales (sparkling) and adding the same to the DataFrame.

Training Time instance

```
 [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, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 4
4, 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, 10
5, 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]
```

Test Time instance

```
 [131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 1
47, 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, 183, 184, 185, 186, 187]
```

```
First few rows of Training Data
              Rose  time
YearMonth
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

Last few rows of Training Data
             Rose  time
YearMonth
1990-06-01   76.0   126
1990-07-01   78.0   127
1990-08-01   70.0   128
1990-09-01   83.0   129
1990-10-01   65.0   130

First few rows of Test Data
              Rose  time
YearMonth
1990-11-01   110.0   131
1990-12-01   132.0   132
1991-01-01    54.0   133
1991-02-01    55.0   134
1991-03-01    66.0   135

Last few rows of Test Data
             Rose  time
YearMonth
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 1.4: Time instance for Regression

Running the Linear Regression Model:

```
▾ LinearRegression
LinearRegression()
```

Figure 1.10: Linear Regression

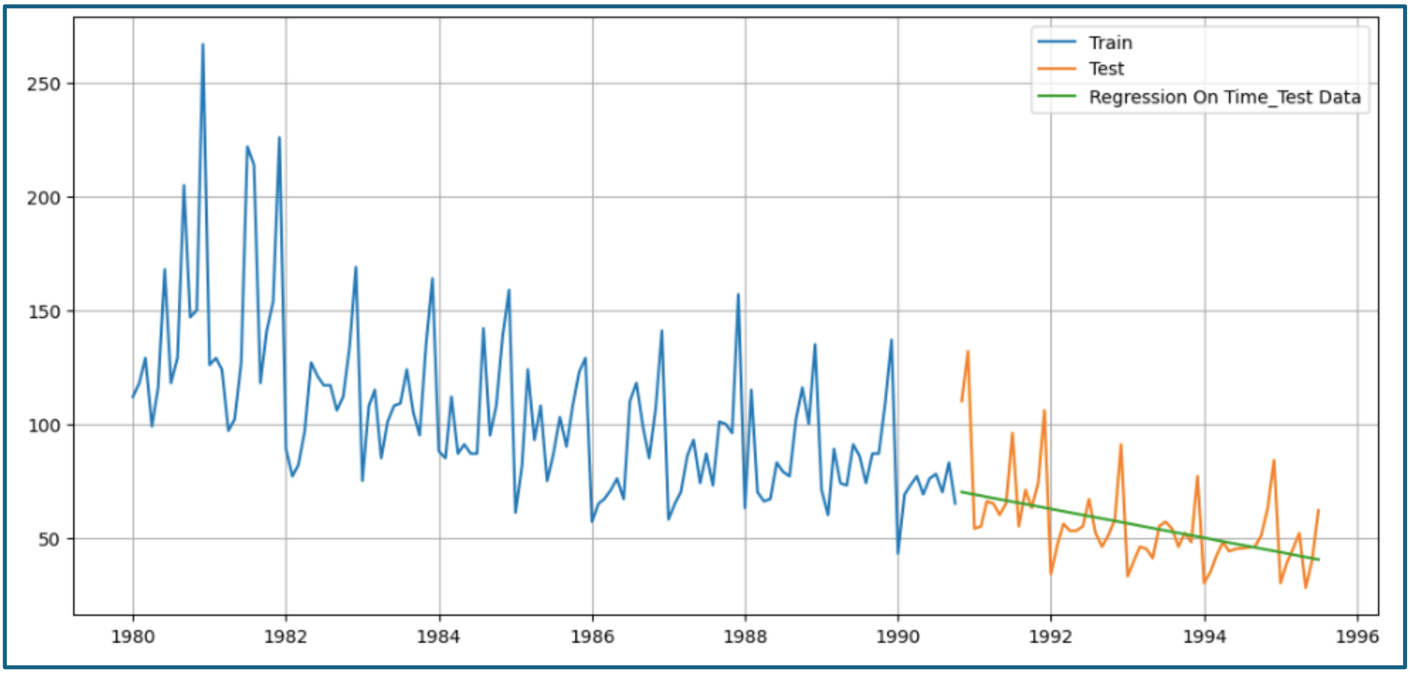


Figure 1.11: Linear Regression Plot

The green line on the test data represents the regression on time.

Model Evaluation:

The model is evaluated using the Root Mean Squared Error value on the regression model and added to a new DataFrame for comparison purpose of different models.

	Test RMSE
RegressionOnTime	17.355796

Table 1.5: Linear Regression

3.3 Simple Average

YearMonth	Rose	forecast
1990-11-01	110.0	104.692308
1990-12-01	132.0	104.692308
1991-01-01	54.0	104.692308
1991-02-01	55.0	104.692308
1991-03-01	66.0	104.692308

Table 1.5: Simple Average

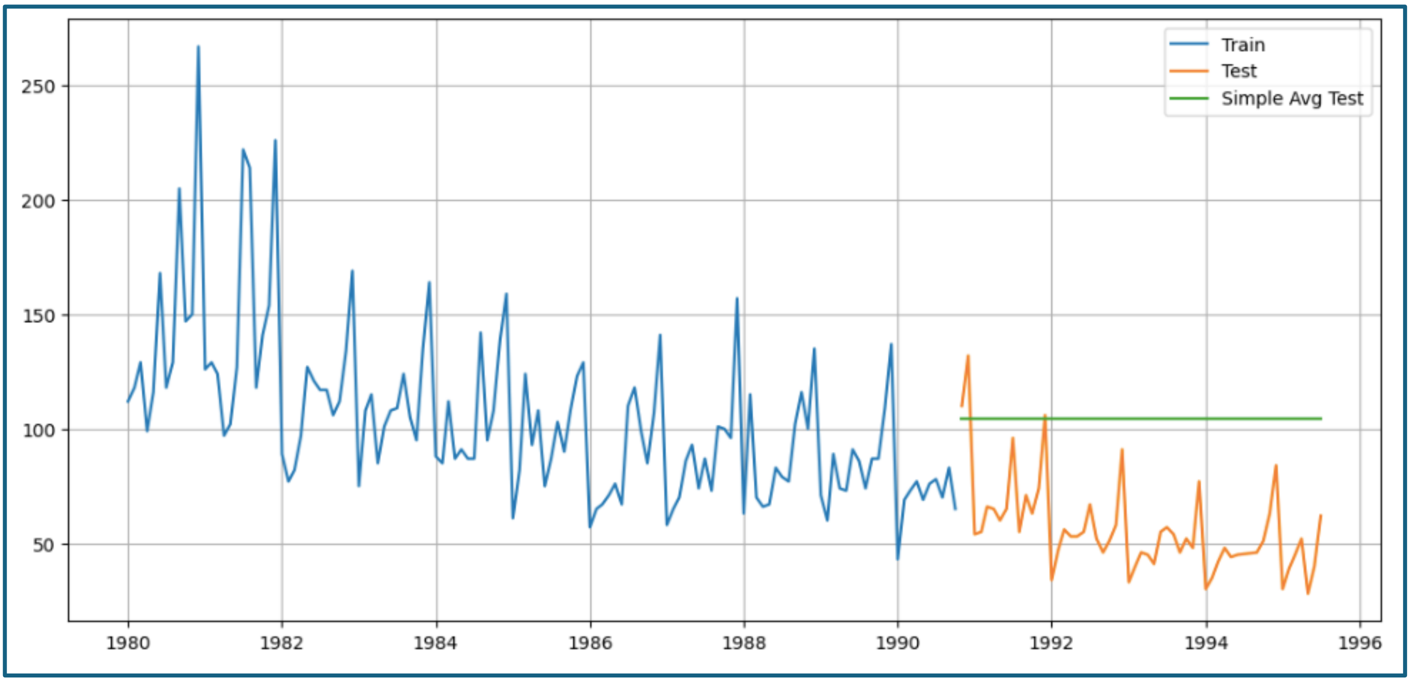


Figure 1.11: Simple Average Plot

3.4 Moving Average

For the moving average model rolling means (or moving averages) are calculated for different intervals. The best average can be determined by min. error value.

The trailing values for the moving average can be considered as 2, 4, 6 & 9

YearMonth	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	112.0	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.50	NaN	NaN
1980-05-01	116.0	107.5	115.50	NaN	NaN
...	...	...	...	...	...
1995-03-01	45.0	42.0	49.50	52.000000	49.888889
1995-04-01	52.0	48.5	41.50	52.166667	50.629630
1995-05-01	28.0	40.0	41.00	46.333333	48.666667
1995-06-01	40.0	34.0	41.25	39.000000	48.000000
1995-07-01	62.0	51.0	45.50	44.333333	49.222222

Table 1.6: Moving Average

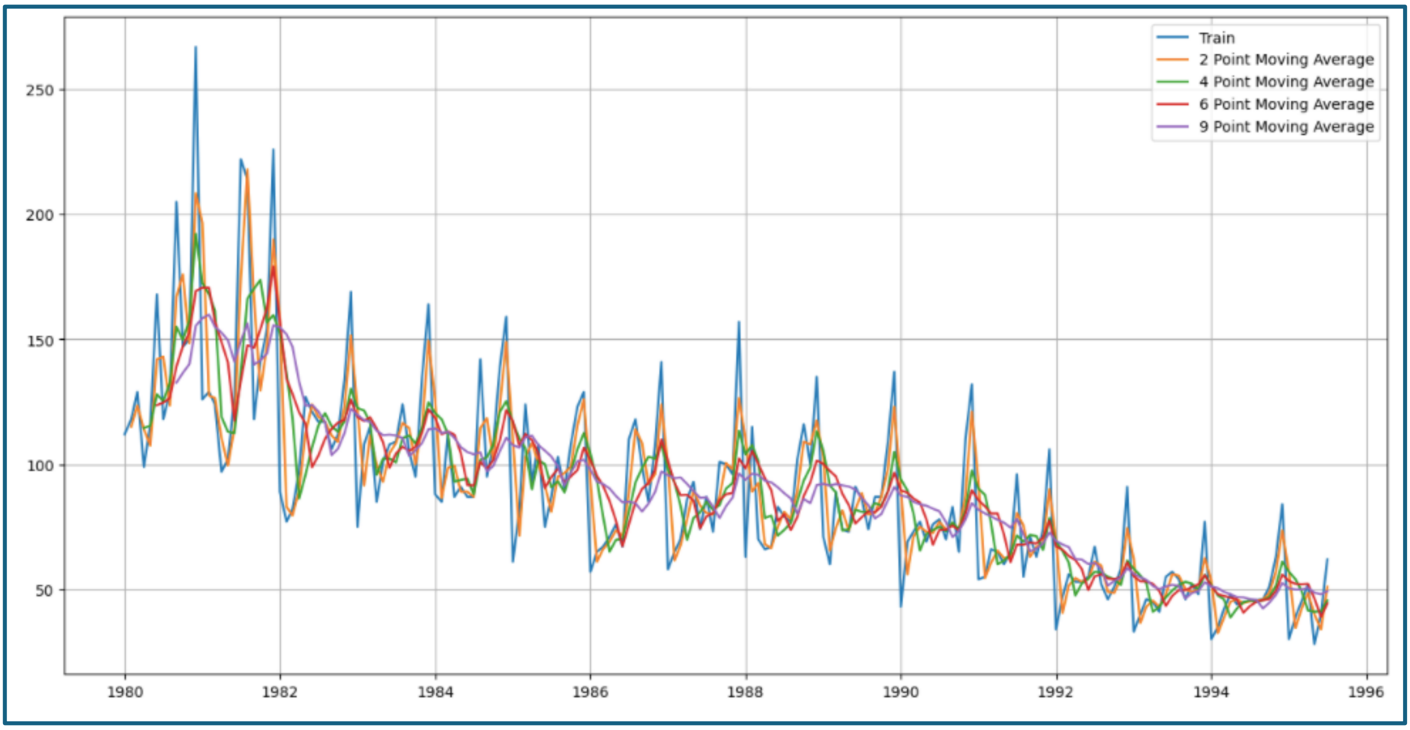


Figure 1.12: Moving Average Plot

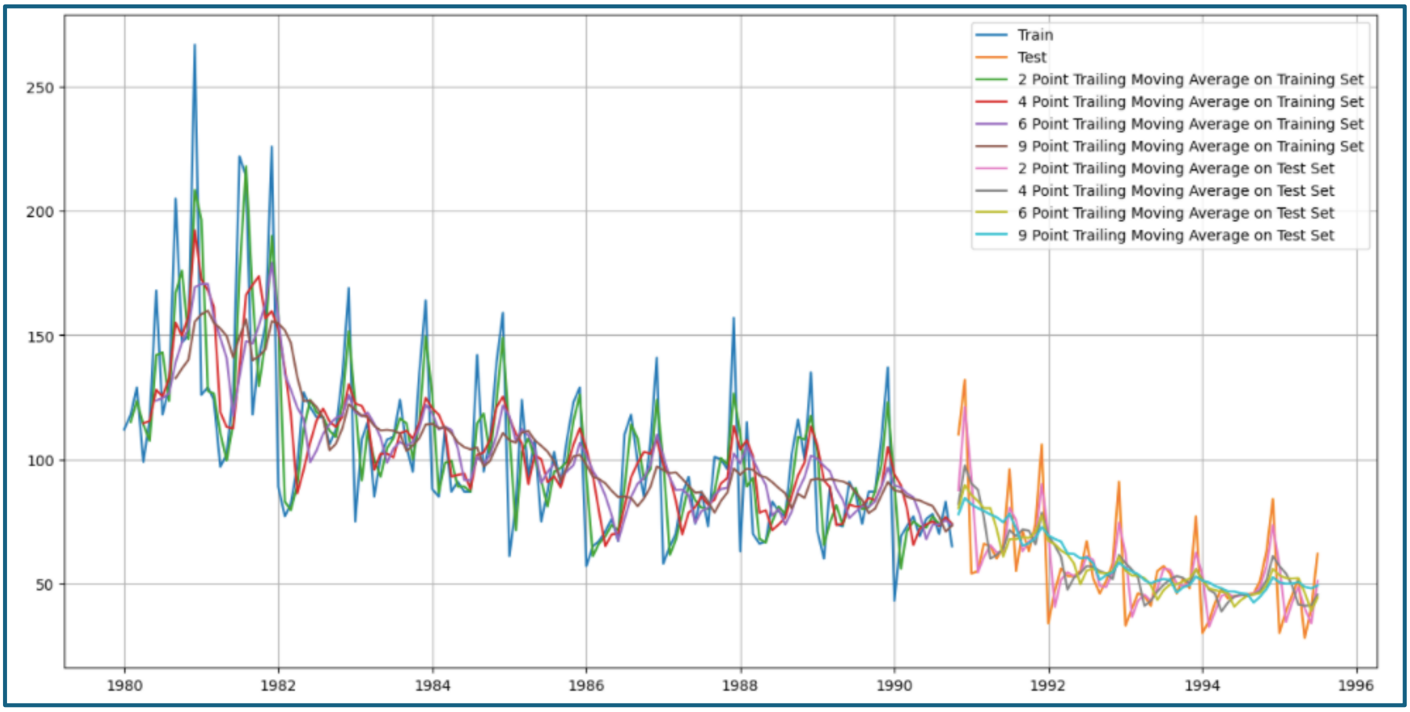


Figure 1.13: Moving Average Train-Test Split Plot

Model Evaluation:

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.801
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 15.367
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 15.862
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 16.342
```

The 2 point moving average model has the lowest RMSE score compared to others.

	Test RMSE
RegressionOnTime	17.355796
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919

Table 1.7: RMSE of different models

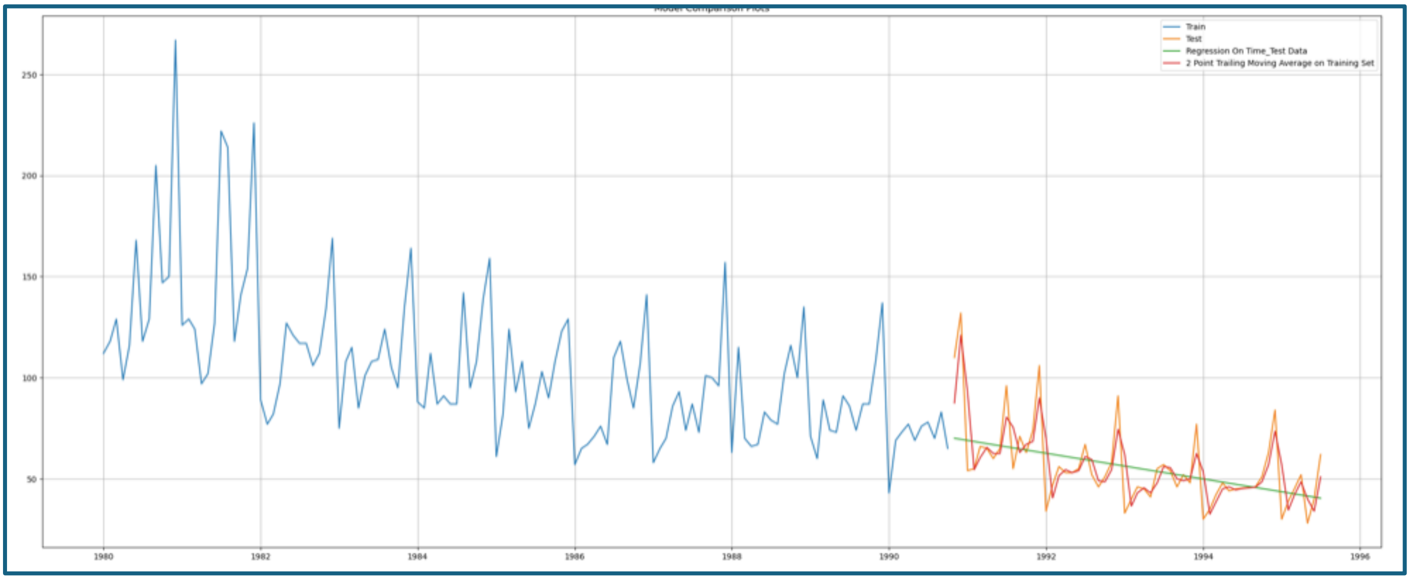


Figure 1.14: Train-Test Split – Moving Average - Regression Plots

3.5 Exponential Models (Single, Double, Triple)

3.5.1 Single Exponential Smoothing

Running the single exponential smoothing using the SimpleExpSmoothing function with train data and fitting it.

Model Parameters:

```
{'smoothing_level': 0.1027210902642377,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.26272138895413,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Prediction:

YearMonth	Rose	predict
1990-11-01	110.0	78.89952
1990-12-01	132.0	78.89952
1991-01-01	54.0	78.89952
1991-02-01	55.0	78.89952
1991-03-01	66.0	78.89952

Table 1.8: Single Exponential Smoothing Predictions

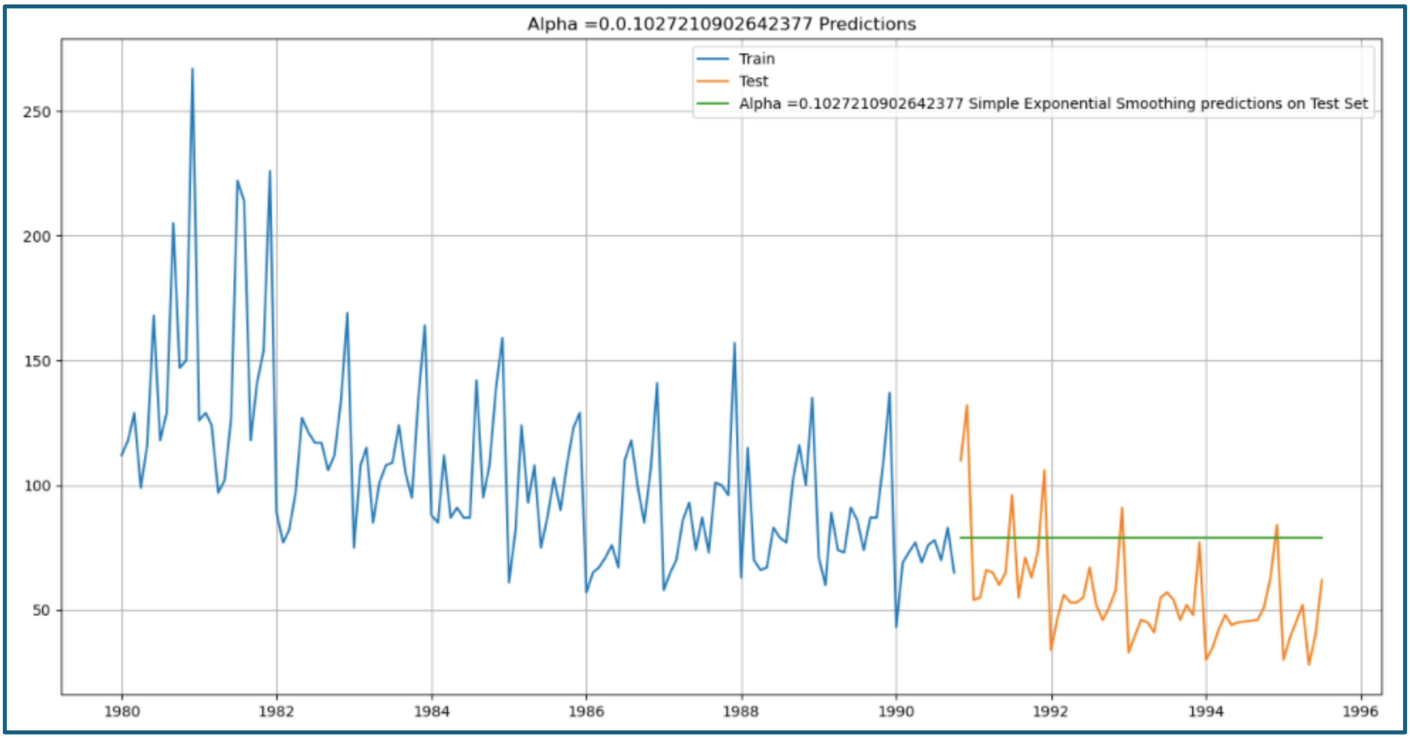


Figure 1.15: Single Exponential Smoothing with Alpha – 0.038 value Plot

Model Evaluation for α = 0.1027210902642377 : Simple Exponential Smoothing

	Test RMSE
RegressionOnTime	17.355796
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919
Alpha=0.1027210902642377,SimpleExponentialSmoothing	30.188322

Table 1.9: RMSE after Single Exponential Smoothing Model

Still the RMSE of Single Exponential Smoothing Model is higher than the 2 Point Moving Average Value

Different RMSE values for different alpha values range from 0.1 to 1

	Alpha Values	Train RMSE	Test RMSE
8	0.9	37.507371	22.496819
7	0.8	36.330954	23.212777
6	0.7	35.288467	23.894929
5	0.6	34.372651	24.529811
4	0.5	33.578304	25.109604
3	0.4	32.893017	25.657764
2	0.3	32.292266	26.310348
1	0.2	31.779467	27.482131
0	0.1	31.643829	30.310589

Table 1.10: Different RMSE for Alpha 0.1 - 1

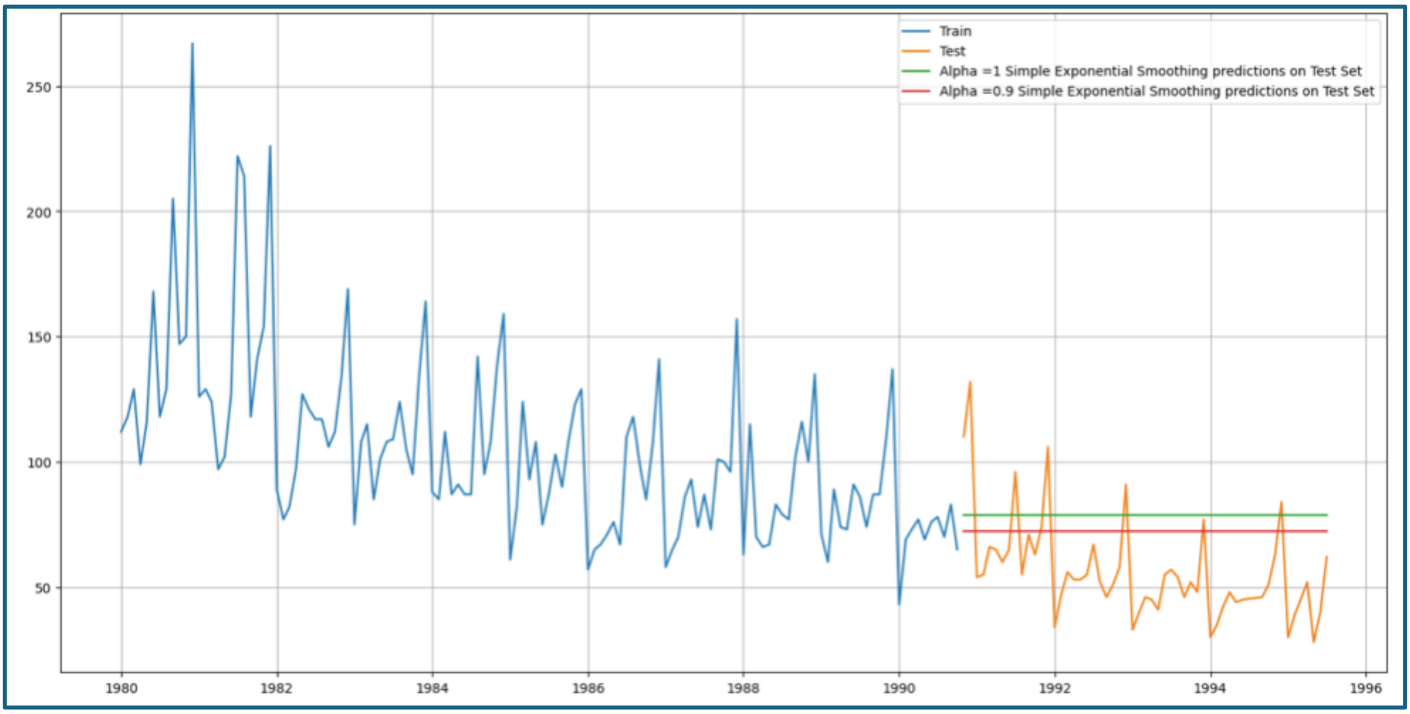


Figure 1.16: Alpha – 0.9 value Plot on Single Exponential Smoothing

	Test RMSE
RegressionOnTime	17.355796
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919
Alpha=0.1027210902642377,SimpleExponentialSmoothing	30.188322
Alpha=0.9,SimpleExponentialSmoothing	22.496819

Table 1.11: Different RMSE for different models

3.5.2 Double Exponential Smoothing

Running the single exponential smoothing using the Holt function with train data and fitting it.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	35.792345	26.531373
1	0.3	0.4	37.287813	18.343250
2	0.3	0.5	38.841090	26.672378
3	0.3	0.6	40.299159	59.403849
4	0.3	0.7	41.486887	95.600325
...	...	...	...	...
59	1.0	0.6	52.018949	264.855344
60	1.0	0.7	54.697098	331.622574
61	1.0	0.8	57.575818	406.506557
62	1.0	0.9	60.691028	491.247514
63	1.0	1.0	64.093561	587.897332

Table 1.12: Different RMSE for different models

	Alpha Values	Beta Values	Train RMSE	Test RMSE
1	0.3	0.4	37.287813	18.343250
12	0.4	0.7	40.744796	18.975318
9	0.4	0.4	37.990913	19.133156
17	0.5	0.4	38.598226	19.197151
8	0.4	0.3	36.682435	19.769770

Table 1.13: Different RMSE for different alpha & beta sorted

The Model with Alpha – 0.3 & Beta – 0.4 has the lowest RMSE value in both train and test data

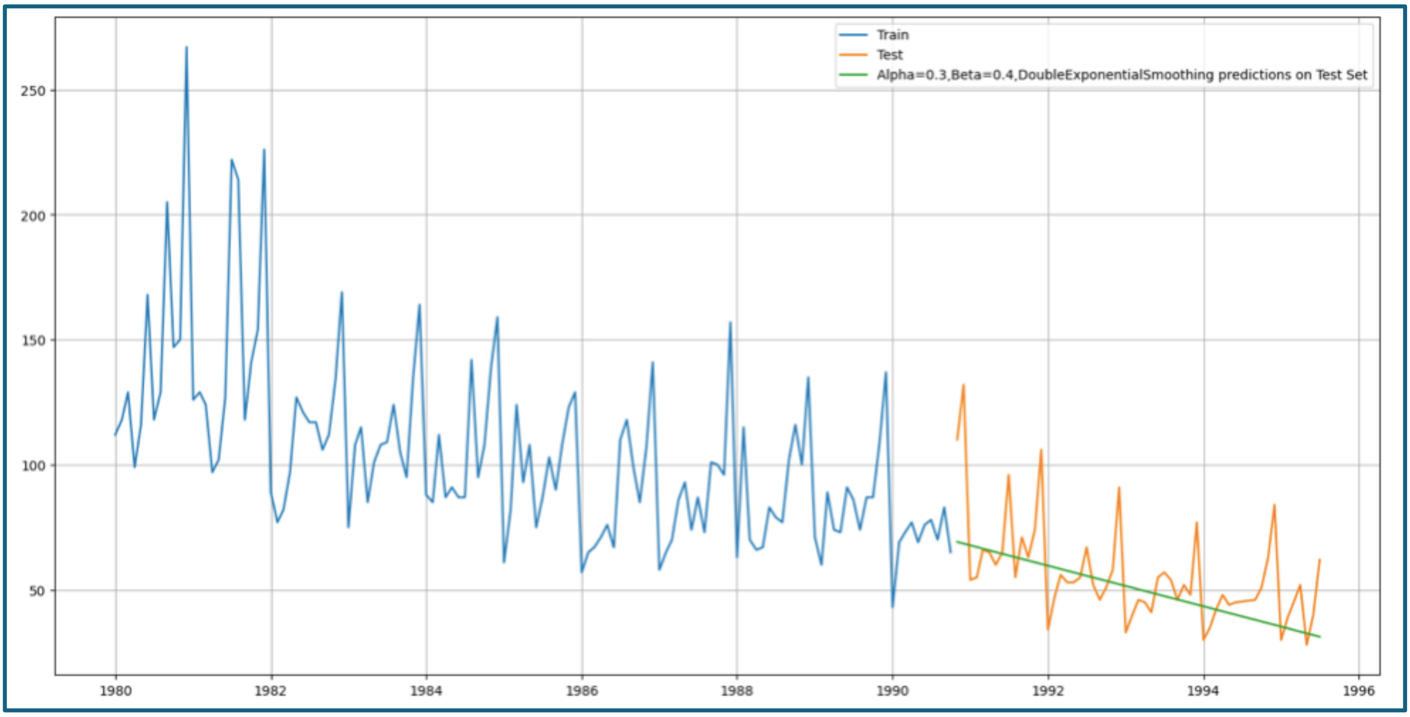


Figure 1.17: Alpha – 0.3 & Beta – 0.4 value Plot on Double Exponential Smoothing

	Test RMSE
RegressionOnTime	17.355796
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919
Alpha=0.1027210902642377,SimpleExponentialSmoothing	30.188322
Alpha=0.9,SimpleExponentialSmoothing	22.496819
Alpha=0.3,Beta=0.4,DoubleExponentialSmoothing	18.343250

Table 1.14: Different RMSE for different models

Till the double exponential smoothing the 2Point MA has the lowest RMSE value.

3.5.3 Triple Exponential Smoothing

Three parameters α , β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

YearMonth	Rose	auto_predict
1990-11-01	110.0	86.291902
1990-12-01	132.0	117.979447
1991-01-01	54.0	51.933830
1991-02-01	55.0	58.193935
1991-03-01	66.0	63.075288

Table 1.15: Prediction Triple Exponential Smoothing model

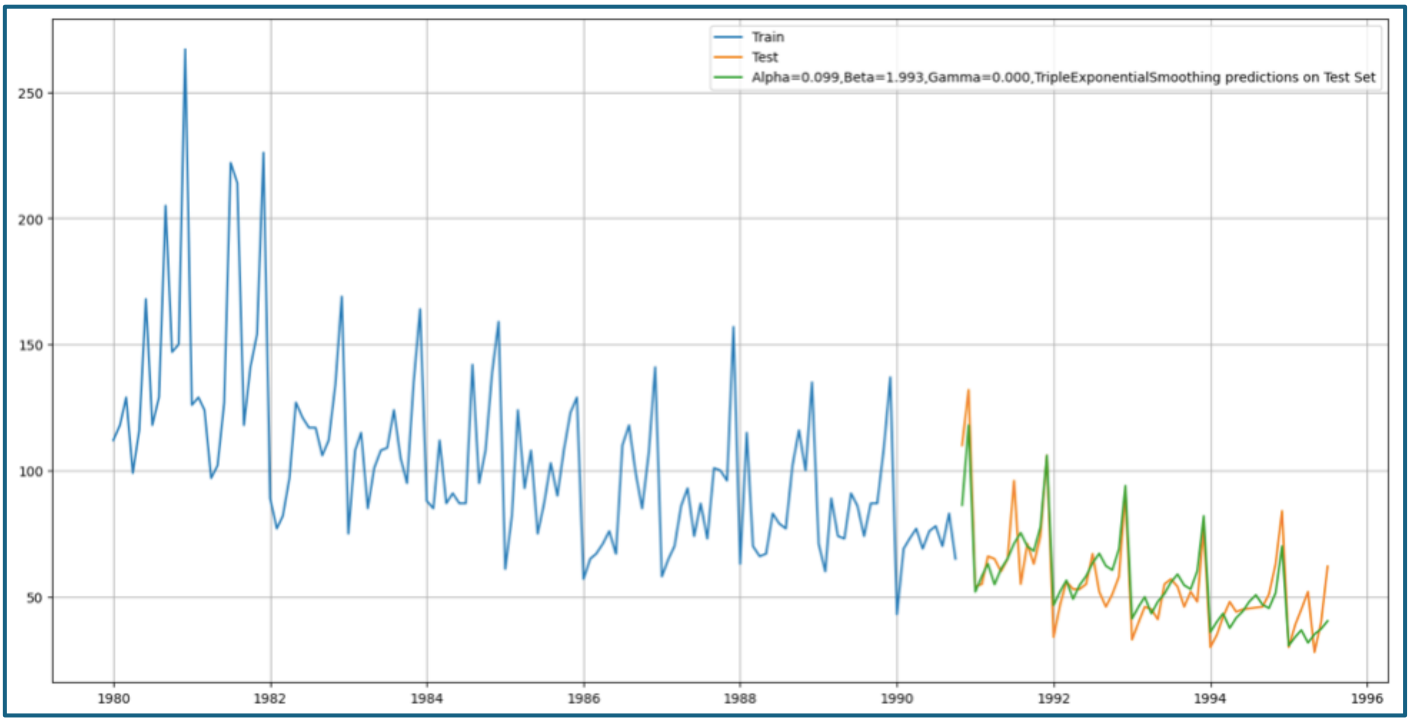


Figure 1.18: Alpha – 0.099, Beta – 1.993 & Gamma – 0.000 value Plot on Triple Exponential Smoothing

The predicted values are almost similar to the test data in the plot for the Triple Exponential Smoothing Model with Alpha – 0.099, Beta – 1.993 & Gamma – 0.000

	Test RMSE
RegressionOnTime	17.355796
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919
Alpha=0.1027210902642377,SimpleExponentialSmoothing	30.188322
Alpha=0.9,SimpleExponentialSmoothing	22.496819
Alpha=0.3,Beta=0.4,DoubleExponentialSmoothing	18.343250
Alpha=0.099,Beta=1.993,Gamma=0.000,TripleExponentialSmoothing	9.328733

Table 1.16: Different RMSE for different models

After all the Models the TES model has the lowest RMSE value.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
0	0.1	0.1	0.1	19.589725	9.584356
1	0.1	0.1	0.2	20.134187	9.531308
2	0.1	0.1	0.3	20.793342	9.425275
3	0.1	0.1	0.4	21.552201	9.288651
4	0.1	0.1	0.5	22.415877	9.212838
...	...	...	...	...	...
995	1.0	1.0	0.6	2156.016082	1764.935972
996	1.0	1.0	0.7	3114.343611	5623.268754
997	1.0	1.0	0.8	2455.974802	4317.157984
998	1.0	1.0	0.9	29056.690525	30910.588575
999	1.0	1.0	1.0	1588.932809	8368.146310

Table 1.17: Different RMSE for different α , β and γ values

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
91	0.1	1.0	0.2	23.140673	9.129075
4	0.1	0.1	0.5	22.415877	9.212838
5	0.1	0.1	0.6	23.396696	9.282132
3	0.1	0.1	0.4	21.552201	9.288651
106	0.2	0.1	0.7	25.623069	9.331747

Table 1.18: Different RMSE for different α , β and γ values sorted

While sorting the different values of RMSE on train and test data of TES model, the $\alpha - 0.1$, $\beta - 1.0$ and $\gamma - 0.2$ has the lowest RMSE values across all the values

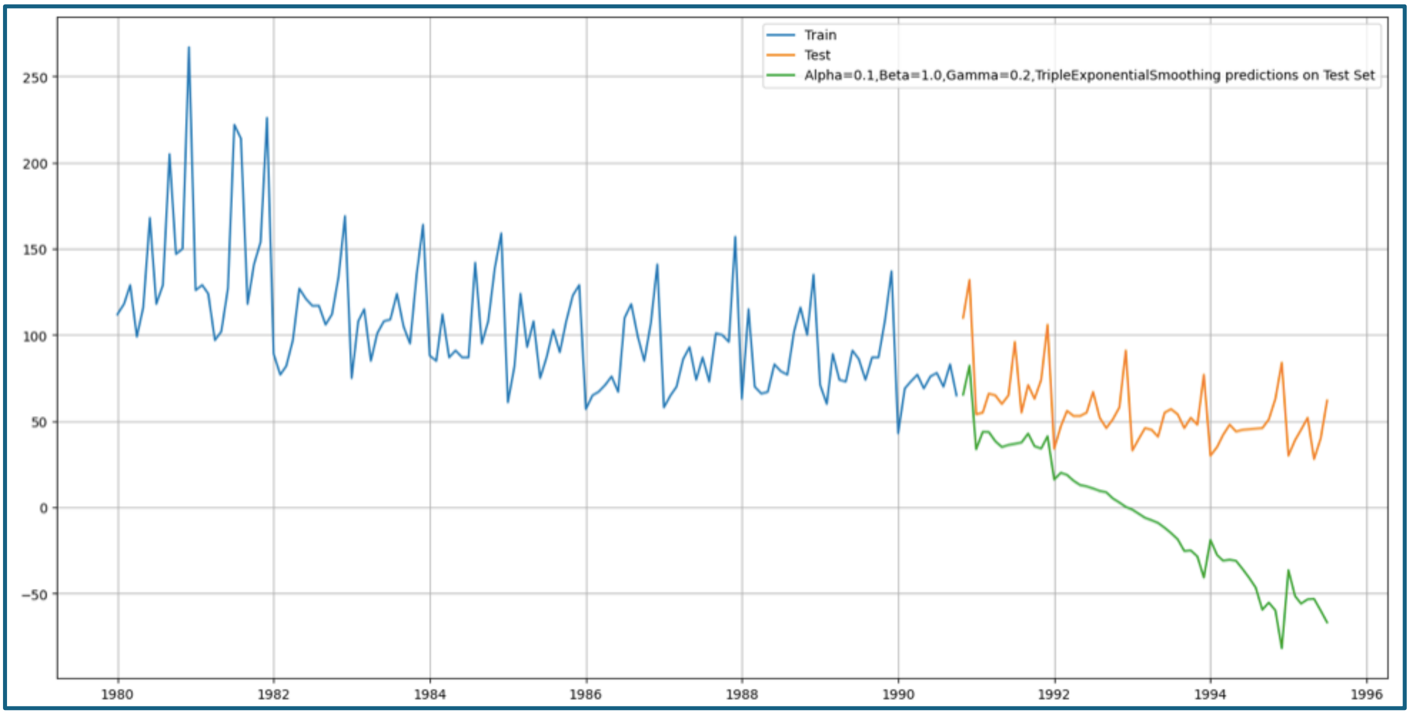


Figure 1.19: Alpha – 0.1, Beta – 1.0 & Gamma – 0.2 value Plot on Triple Exponential Smoothing

RMSE

	Test RMSE
RegressionOnTime	17.355796
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919
Alpha=0.1027210902642377,SimpleExponentialSmoothing	30.188322
Alpha=0.9,SimpleExponentialSmoothing	22.496819
Alpha=0.3,Beta=0.4,DoubleExponentialSmoothing	18.343250
Alpha=0.099,Beta=1.993,Gamma=0.000,TripleExponentialSmoothing	9.328733
Alpha=0.1,Beta=1.0,Gamma=0.2,TripleExponentialSmoothing	9.129075

Table 1.19: Different RMSE for different models

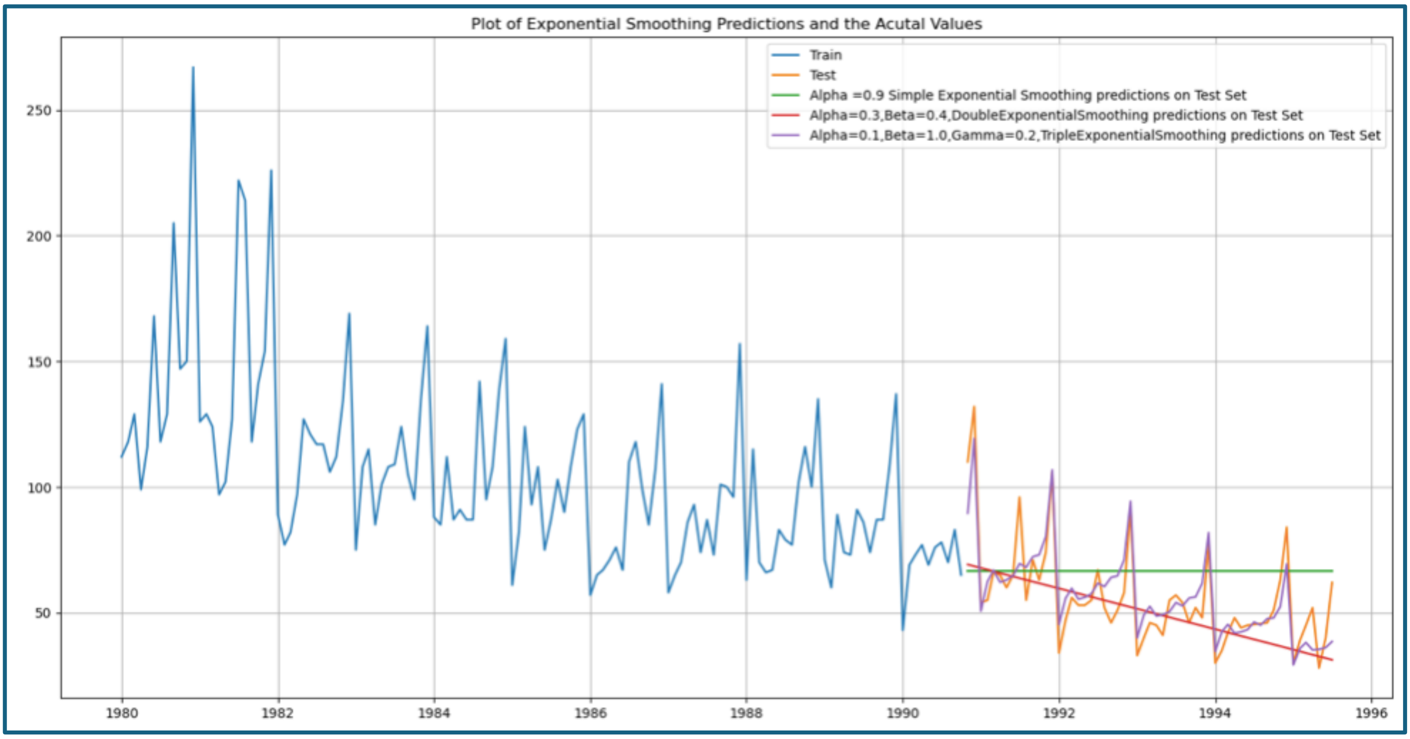


Figure 1.20: Exponential Smoothing predictions plot and Actual values

Full Model

The predictions on the full model values are based on the TES model with $\alpha - 0.1$, $\beta - 1.0$ and $\gamma - 0.2$ and RMSE of 342.322

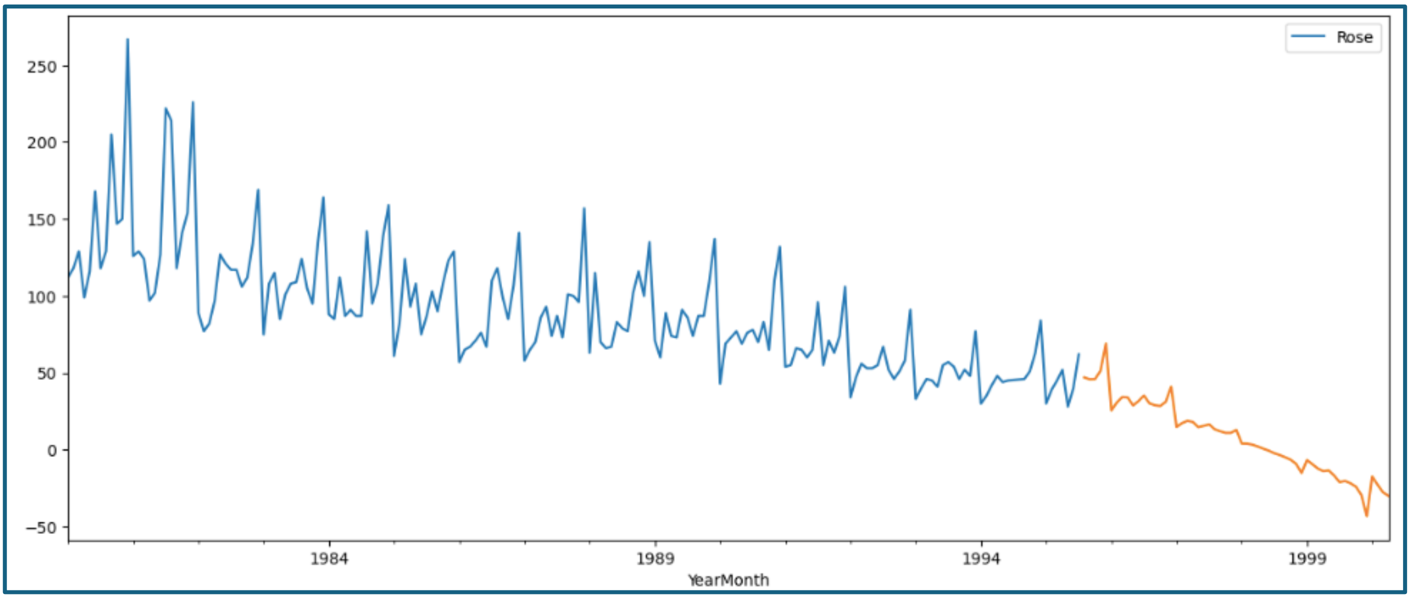


Figure 1.21: Predictions plot and Actual values on the full model

	lower_CI	prediction	upper_ci
1995-08-01	8.742476	47.019262	85.296048
1995-09-01	7.562921	45.839707	84.116493
1995-10-01	7.681913	45.958699	84.235485
1995-11-01	13.026824	51.303610	89.580396
1995-12-01	30.694471	68.971257	107.248043

Table 1.20: First 5 rows of the predicted values on the full model

	lower_CI	prediction	upper_ci
1999-12-01	-81.438896	-43.162110	-4.885324
2000-01-01	-55.718198	-17.441412	20.835374
2000-02-01	-61.098043	-22.821257	15.455529
2000-03-01	-66.019526	-27.742740	10.534046
2000-04-01	-68.244240	-29.967453	8.309333

Table 1.21: Last 5 rows of the predicted values on the full model

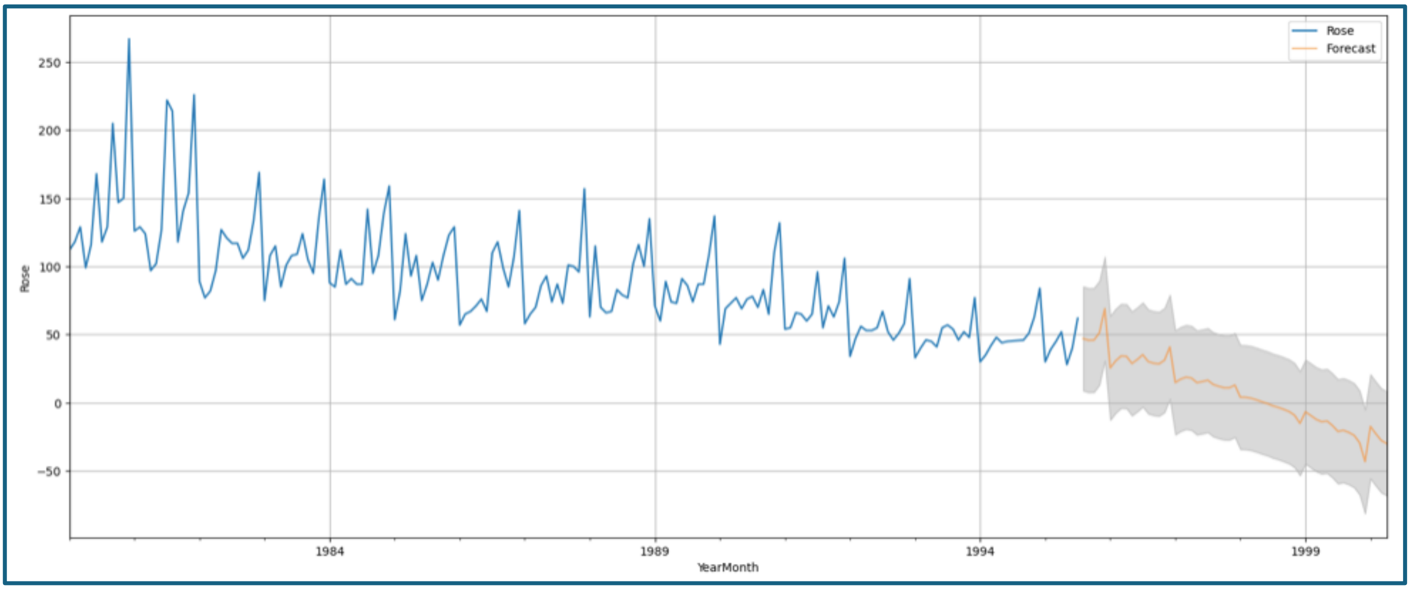


Figure 1.22: Forecast with Lower and Upper Limit and Actual values on the full model

The final plot shows the forecast values using TES model with 95% interval level.

3.6 Check the performance of the models built

	Test RMSE
Alpha=0.1,Beta=1.0,Gamma=0.2,TripleExponentialSmoothing	9.129075
Alpha=0.099,Beta=1.993,Gamma=0.000,TripleExponentialSmoothing	9.328733
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919
RegressionOnTime	17.355796
Alpha=0.3,Beta=0.4,DoubleExponentialSmoothing	18.343250
Alpha=0.9,SimpleExponentialSmoothing	22.496819
Alpha=0.1027210902642377,SimpleExponentialSmoothing	30.188322

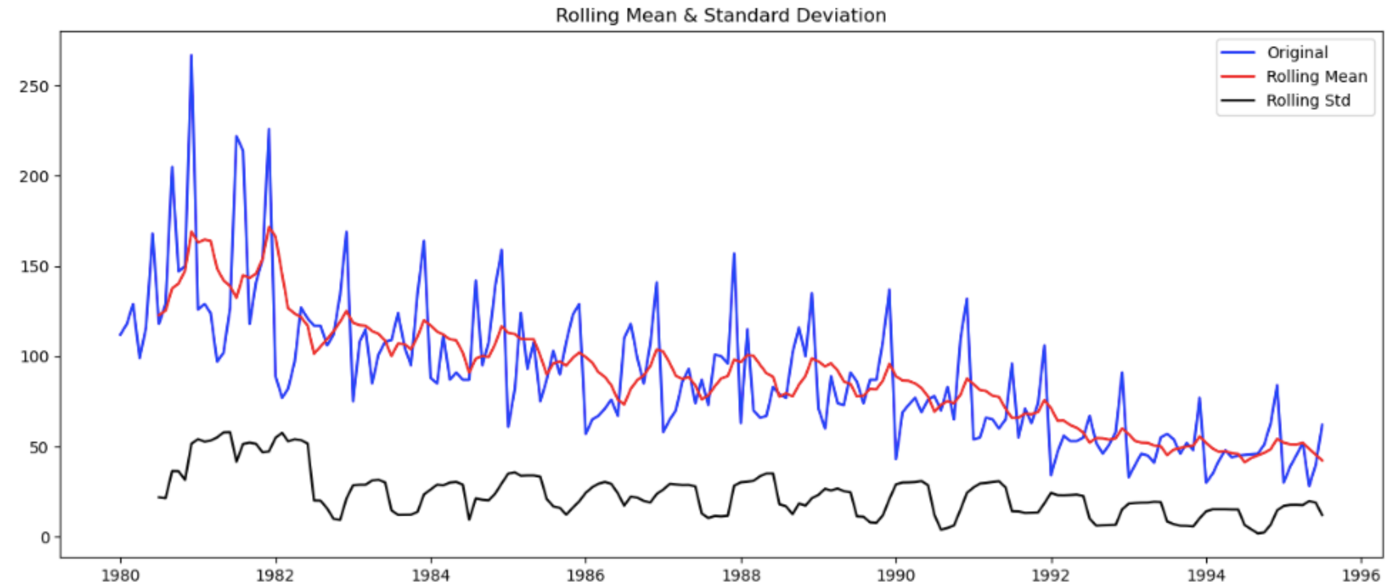
Table 1.22: Different RMSE for different models

Across all the models built the TES with best params Alpha – 0.1, Beta – 1.0 & Gamma – 0.2 value has the lowest RMSE value

4. Check for Stationarity

4.1 Check for stationarity

AD Fuller Test if p value > 0.05 then We fail to reject null hypothesis (H0 is true). We go for differencing if p value < 0.05 then the data is stationary



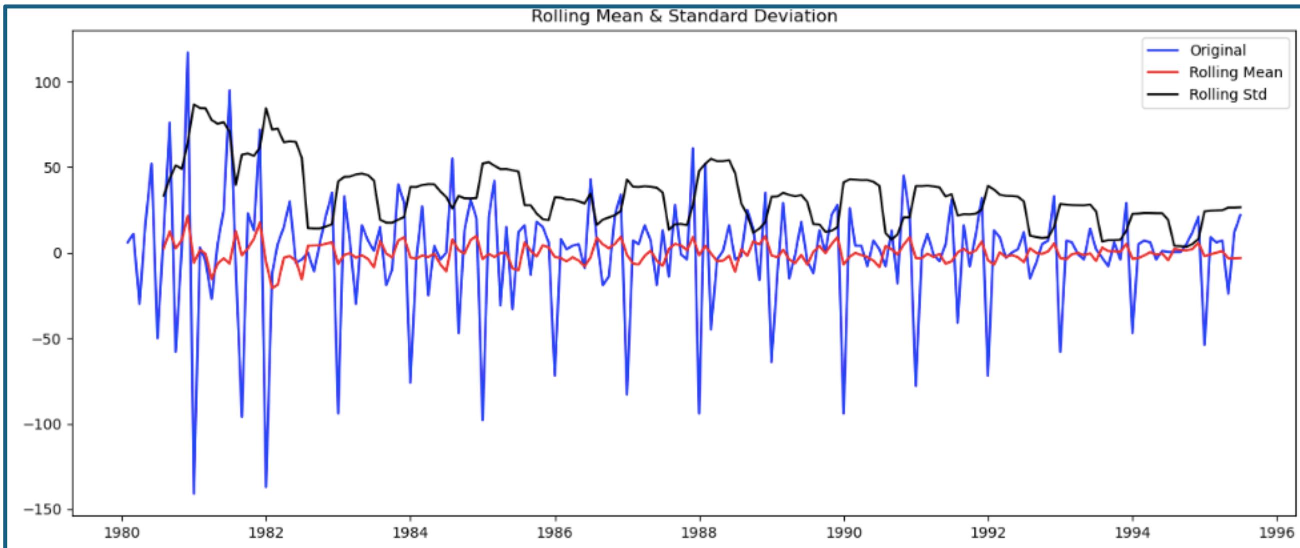

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value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

Figure 1.23: AD Fuller Test

We see that at 5% significant level the Time Series is non-stationary. Let us take a difference of order 1 and check whether the Time Series is stationary or not.

4.2 Make the data stationary (if needed)

The p-value is less than 0.05 and hence the data is stationary




```
Results of Dickey-Fuller Test:
Test Statistic                -8.044392e+00
p-value                        1.810895e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

Figure 1.24: AD Fuller Test after differencing of order 1

5. Model Building - Stationary Data

5.1 Generate ACF & PACF Plot and find the AR, MA values

ACF

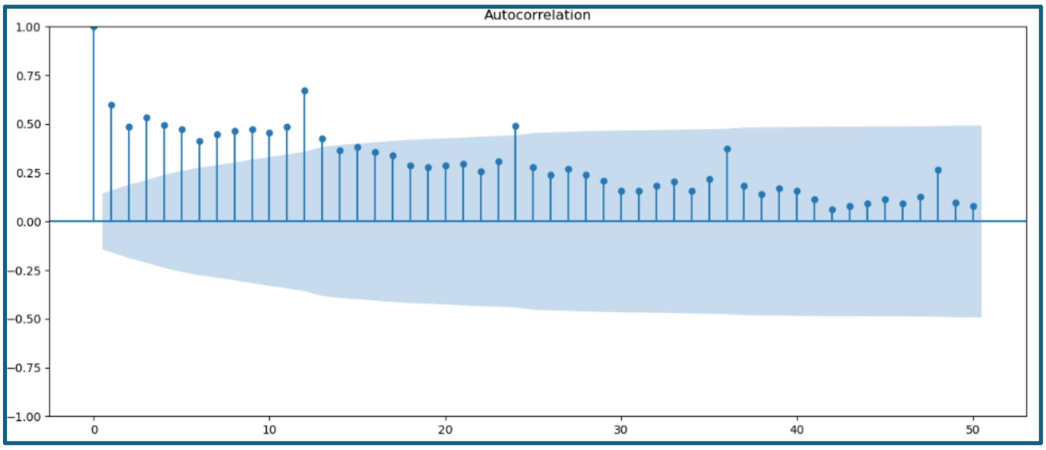


Figure 1.25: ACF

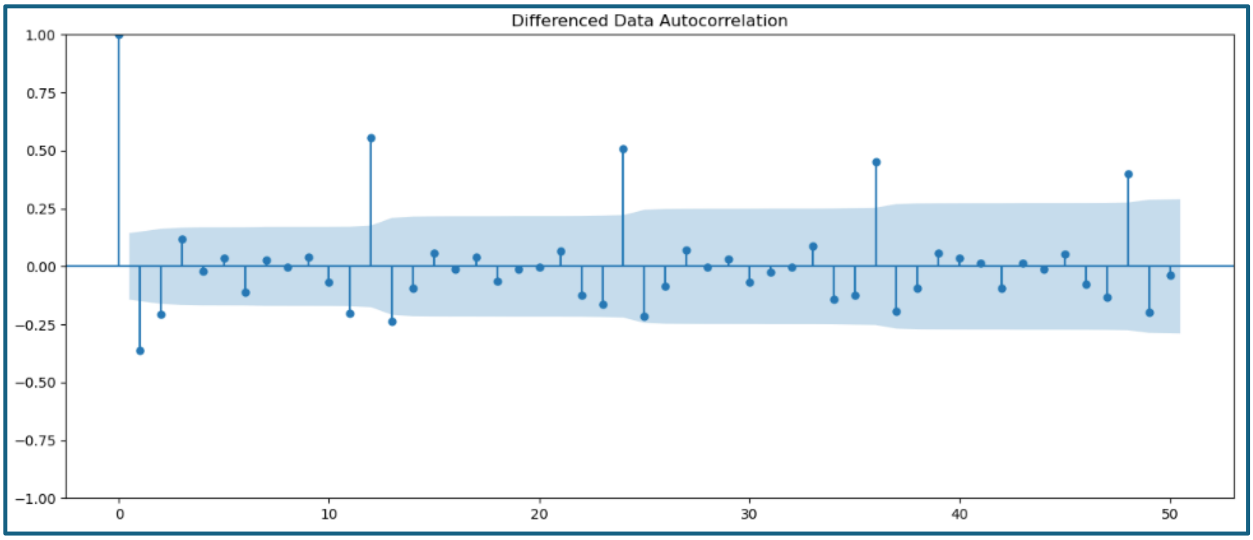


Figure 1.26: ACF with order 1 differencing

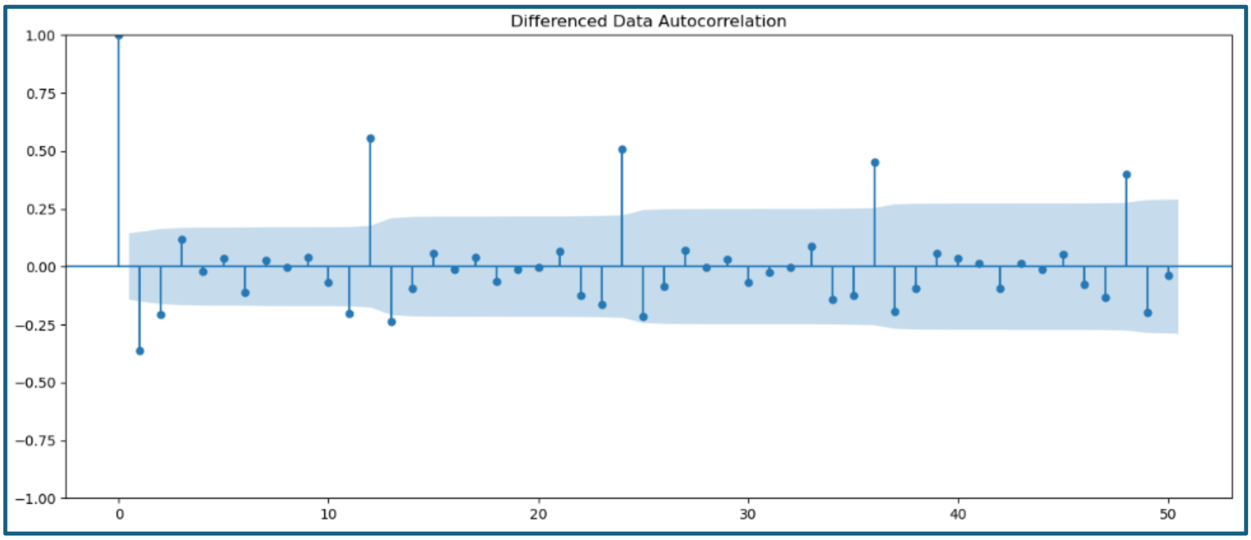


Figure 1.27: PACF

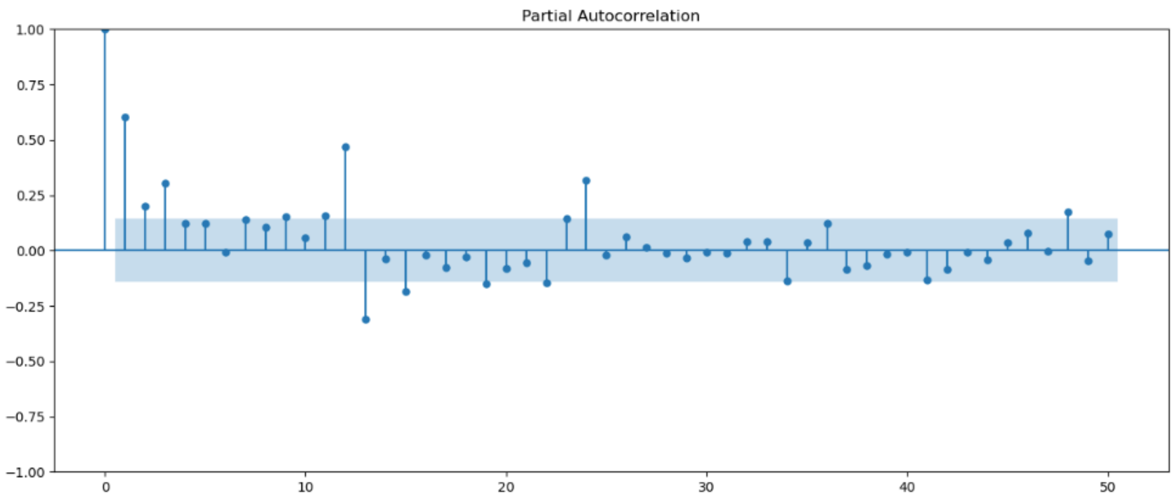


Figure 1.28: PACF with order 1 differencing

5.2 Build different ARIMA models

5.2.1 Auto ARIMA

Getting a combination of different parameters of p and q in the range of 0 and 2 the value of d as 1 as we need to take a difference of the series to make it stationary.

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

Sorted AIC values for different (p,d,q) combinations are derived using the ARIMA function and fitting the same.

	param	AIC
2	(0, 1, 2)	1259.247780
5	(1, 1, 2)	1259.473205
4	(1, 1, 1)	1260.036763
7	(2, 1, 1)	1261.014076
1	(0, 1, 1)	1261.327444
8	(2, 1, 2)	1261.472001
6	(2, 1, 0)	1278.135281
3	(1, 1, 0)	1297.077294
0	(0, 1, 0)	1313.175861

Table 1.23: Sorted AIC values in ARIMA Model

Getting the SARIMAX results by using the ARIMA function with the best params of p,d,q of (0,1,2)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  130
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -626.624
Date:                Sun, 14 Apr 2024   AIC                           1259.248
Time:                        20:15:40   BIC                           1267.827
Sample:                    01-01-1980   HQIC                          1262.734
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7059      0.072     -9.851      0.000      -0.846      -0.565
ma.L2         -0.1915      0.074     -2.574      0.010      -0.337      -0.046
sigma2       958.5998     86.875     11.034      0.000     788.328    1128.872
===================================================================================
Ljung-Box (L1) (Q):                   0.15   Jarque-Bera (JB):                45.85
Prob(Q):                              0.70   Prob(JB):                         0.00
Heteroskedasticity (H):               0.32   Skew:                             0.88
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.34
===================================================================================
```

Table 1.24: SARIMAX results in ARIMA Model

Predict on the Test Set using this model and evaluate the model.

```
1990-11-01    76.445779
1990-12-01    79.764438
1991-01-01    79.764438
1991-02-01    79.764438
1991-03-01    79.764438
Freq: MS, Name: predicted_mean, dtype: float64
```

RMSE value is calculated using the best params of (0,1,2) and added to a results DataFrame.

	RMSE
ARIMA(0,1,2)	30.903804

Table 1.25: RMSE of ARIMA(0,1,2)

5.2.2 Manual ARIMA

Differentiated ACF & PACF plot to get the p & q value for the model building

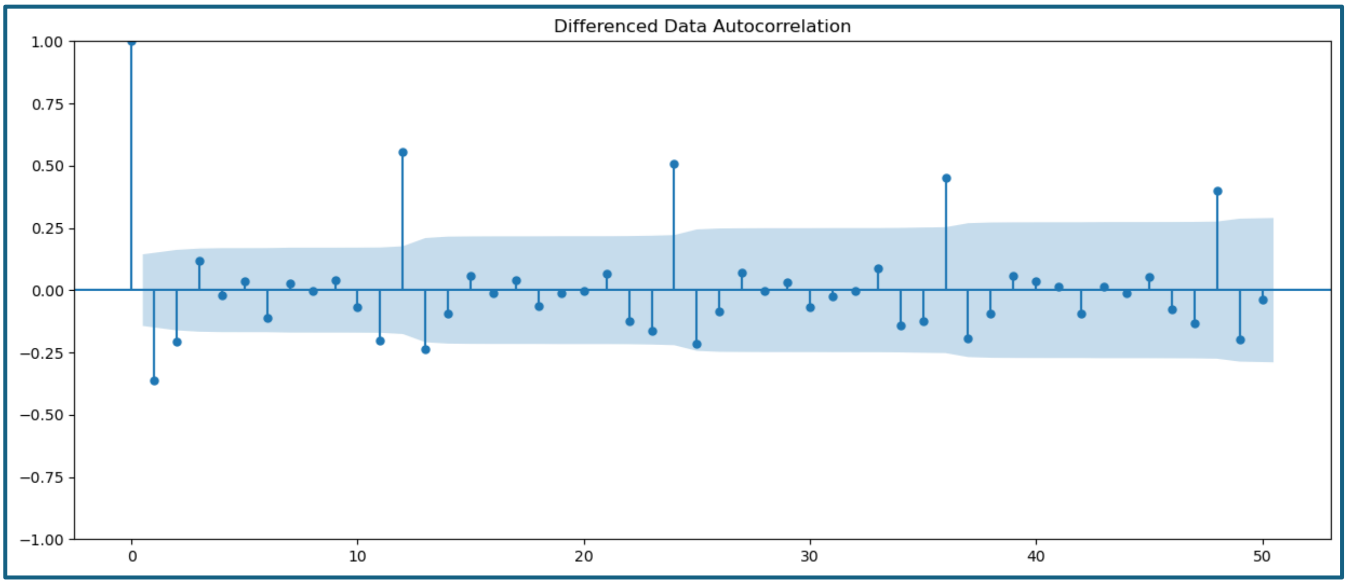


Figure 1.29: Differentiated ACF plot

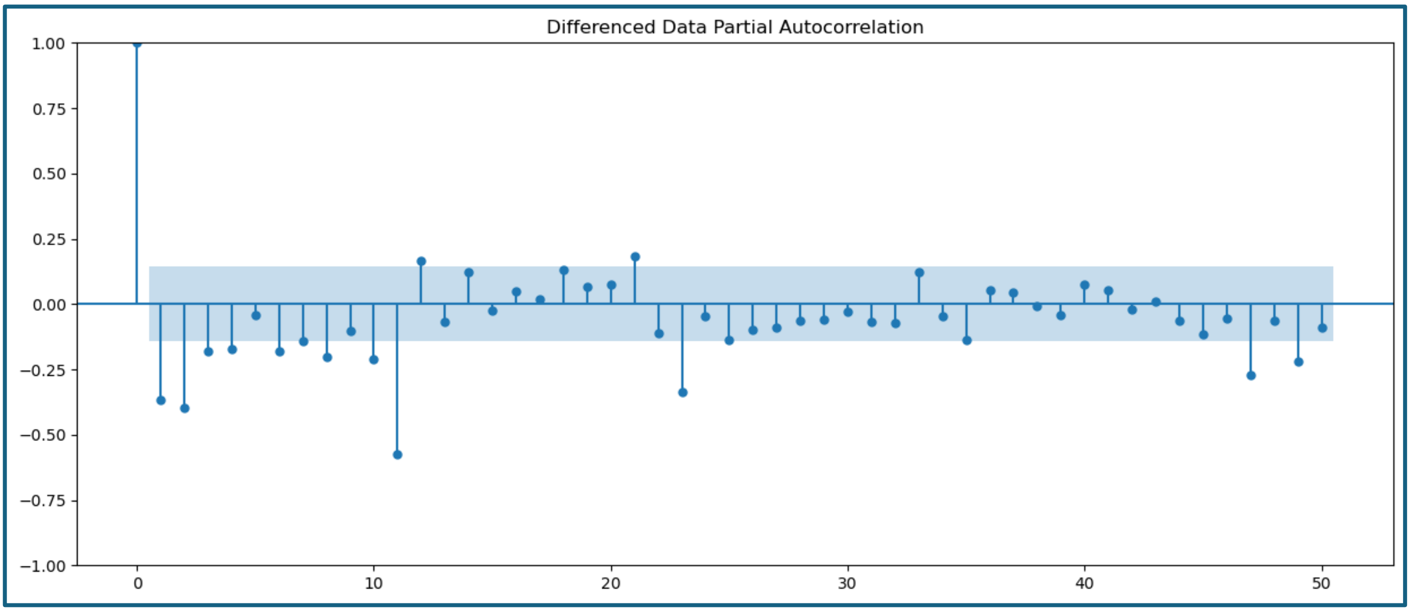


Figure 1.30: Differentiated PACF plot

Here, we have taken alpha=0.05.

The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 4.

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

By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  130
Model:                 ARIMA(4, 1, 2)   Log Likelihood                -625.665
Date:                Sun, 14 Apr 2024   AIC                           1265.331
Time:                        20:15:41   BIC                           1285.349
Sample:                    01-01-1980   HQIC                          1273.465
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3929      0.970     -0.405      0.685      -2.294       1.508
ar.L2          0.0012      0.260      0.005      0.996      -0.508       0.511
ar.L3          0.0405      0.113      0.357      0.721      -0.182       0.263
ar.L4         -0.0038      0.179     -0.021      0.983      -0.355       0.347
ma.L1         -0.3211      0.980     -0.328      0.743      -2.242       1.600
ma.L2         -0.5370      0.915     -0.587      0.557      -2.330       1.256
sigma2       944.0427     92.372     10.220      0.000     762.997    1125.088
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                38.44
Prob(Q):                              0.86   Prob(JB):                         0.00
Heteroskedasticity (H):               0.33   Skew:                             0.83
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.10
===================================================================================
```

Table 1.26: SARIMAX results in Manual ARIMA Model

Predict on the Test Set using this model and evaluate the model.

RMSE value is calculated using the best params of (4,1,2) and added to a results DataFrame.

	RMSE
ARIMA(0,1,2)	30.903804
ARIMA(4,1,2)	30.607788

Table 1.27: RMSE of both Auto & Manual ARIMA models

The Auto ARIMA(4,1,2) has the lowest RMSE compared to the Manual ARIMA(0,1,2)

5.3 Build different SARIMA models

5.3.1 Auto SARIMA

Checking the ACF plot once again

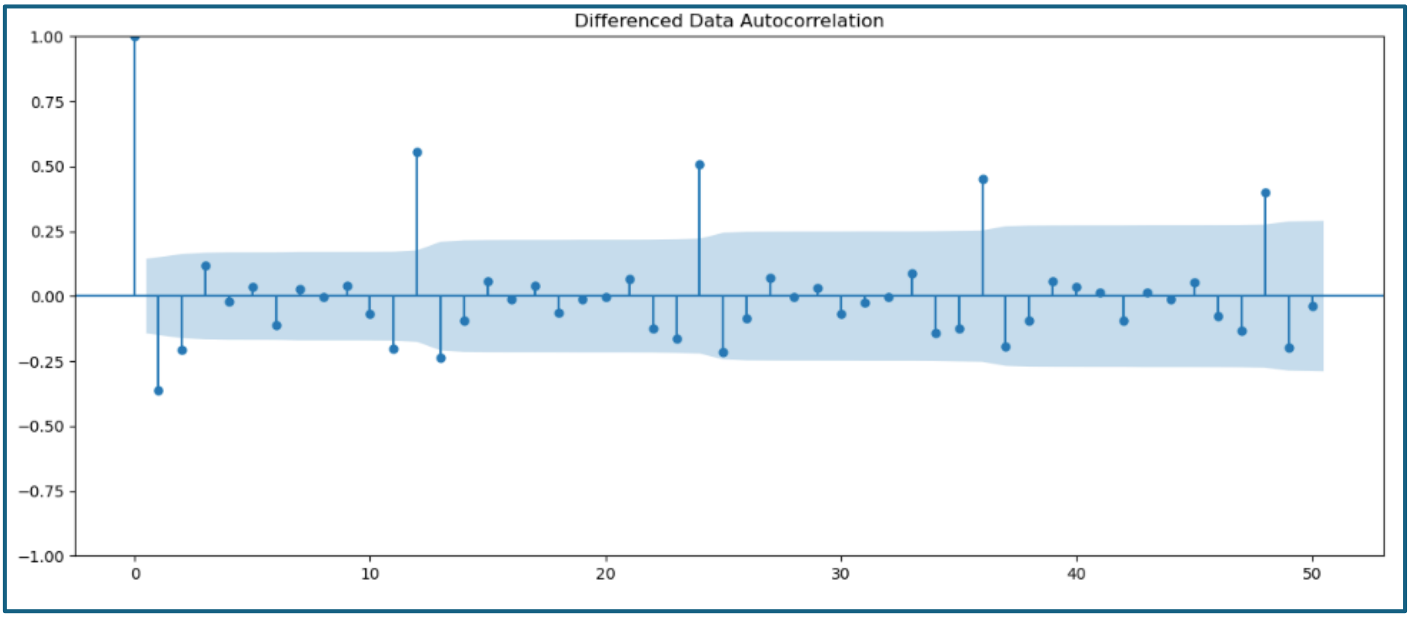


Figure 1.31: Differentiated ACF plot

We see that there can be a seasonality of 12 as well as 24 . We will run our auto SARIMA models by setting seasonality both as 6 and 12.

Setting the seasonality as 6 for the first iteration of the auto SARIMA model

Getting a combination of different parameters of p and q in the range of 0 and 2 the value of d as 1 & 2 as we need to take a difference of the series to make it stationary with P,D,Q and seasonality as 12

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

Running the SARIMA model for all the combination of (p,d,q) & (P,D,Q,S)

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	871.075238
53	(1, 1, 2)	(2, 0, 2, 12)	873.003875
80	(2, 1, 2)	(2, 0, 2, 12)	874.213960
69	(2, 1, 1)	(2, 0, 0, 12)	879.792363
78	(2, 1, 2)	(2, 0, 0, 12)	880.763857

Table 1.28: AIC value of SARIMA model sorted

Building the model again with the best params (0,1,2) (2,0,2,12)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  130
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -428.538
Date:                            Sun, 14 Apr 2024   AIC                            871.075
Time:                                    20:16:33   BIC                            889.450
Sample:                                         0   HQIC                           878.516
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8367    239.113     -0.003      0.997    -469.490     467.817
ma.L2         -0.1633     39.028     -0.004      0.997     -76.656      76.329
ar.S.L12       0.3494      0.079      4.408      0.000       0.194       0.505
ar.S.L24       0.3067      0.075      4.103      0.000       0.160       0.453
ma.S.L12       0.0454      0.134      0.338      0.735      -0.218       0.309
ma.S.L24      -0.0912      0.145     -0.628      0.530      -0.376       0.193
sigma2       250.7786      6e+04      0.004      0.997   -1.17e+05    1.18e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 3.10
Prob(Q):                              0.76   Prob(JB):                         0.21
Heteroskedasticity (H):               0.88   Skew:                             0.43
Prob(H) (two-sided):                  0.71   Kurtosis:                         3.05
===================================================================================
```

Table 1.29: SARIMAX results in Manual SARIMA Model with best params

Model Diagnostic Plot

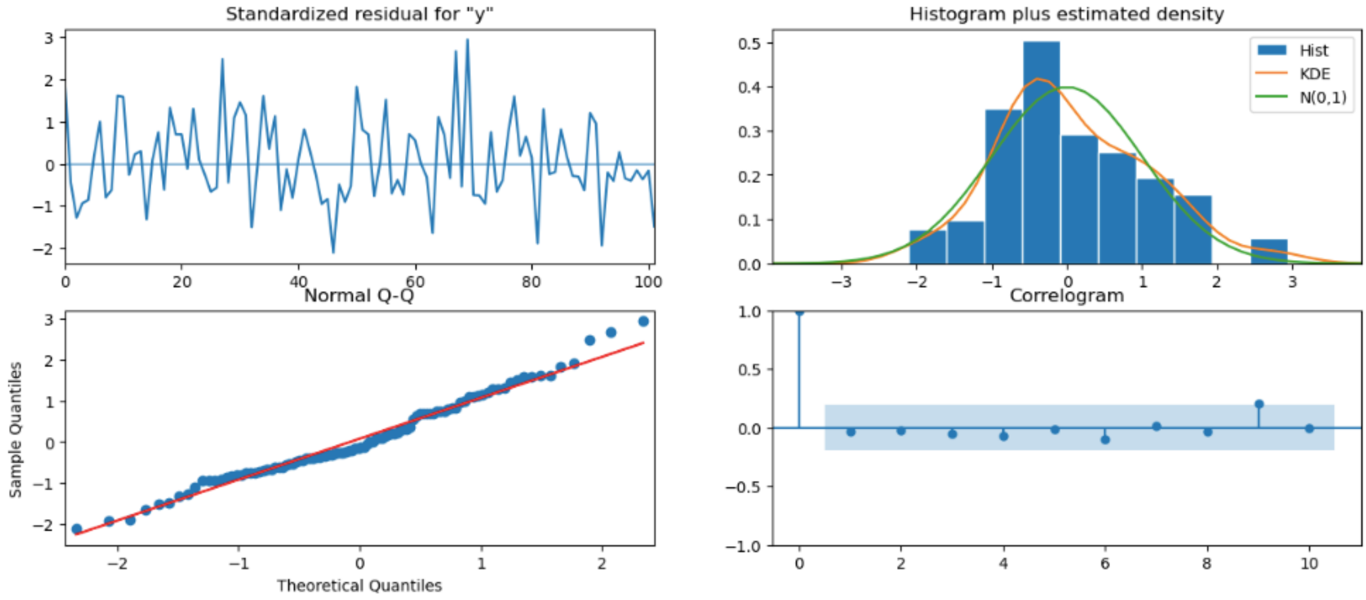


Figure 1.32: SARIMA Model Diagnostic Plot

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

Predict on the Test Set using this model and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	90.849106	15.914360	59.657533	122.040679
1	114.913416	16.150397	83.259220	146.567613
2	60.936673	16.150397	29.282477	92.590869
3	70.599288	16.150396	38.945093	102.253484
4	76.843515	16.150393	45.189325	108.497704

Table 1.30: Prediction with SARIMA Model of seasonality 12

	RMSE
ARIMA(0,1,2)	30.903804
ARIMA(4,1,2)	30.607788
SARIMA(0,1,2)(2,0,2,12)	25.343324

Table 1.31: Different model RSME value

The Auto SARIMA model with seasonality 12 has the lowest RSME value till now.

Setting the seasonality as 24 for the second iteration of the auto SARIMA model.

Getting a combination of different parameters of p and q in the range of 0 and 2 the value of d as 1 & 2 as we need to take a difference of the series to make it stationary with P,D,Q and seasonality as 24

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 24)
Model: (0, 1, 2)(0, 0, 2, 24)
Model: (1, 1, 0)(1, 0, 0, 24)
Model: (1, 1, 1)(1, 0, 1, 24)
Model: (1, 1, 2)(1, 0, 2, 24)
Model: (2, 1, 0)(2, 0, 0, 24)
Model: (2, 1, 1)(2, 0, 1, 24)
Model: (2, 1, 2)(2, 0, 2, 24)
```

Running the SARIMA model for all the combination of (p,d,q) & (P,D,Q,S)

	param	seasonal	AIC
23	(0, 1, 2)	(1, 0, 2, 24)	670.027549
50	(1, 1, 2)	(1, 0, 2, 24)	671.642194
26	(0, 1, 2)	(2, 0, 2, 24)	671.665143
53	(1, 1, 2)	(2, 0, 2, 24)	673.271274
77	(2, 1, 2)	(1, 0, 2, 24)	673.274340

Table 1.32: AIC value of SARIMA model sorted with seasonality 24

Building the model again with the best params (0,1,2) (1,0,2,24)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  130
Model:             SARIMAX(0, 1, 2)x(1, 0, 2, 24)   Log Likelihood                -329.014
Date:                            Sun, 14 Apr 2024   AIC                            670.028
Time:                                    20:19:04   BIC                            684.168
Sample:                                         0   HQIC                           675.688
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9344    115.542     -0.008      0.994    -227.393     225.524
ma.L2         -0.0656      7.518     -0.009      0.993     -14.801      14.670
ar.S.L24       0.8436      0.060     13.975      0.000       0.725       0.962
ma.S.L24      -0.9178   8441.665     -0.000      1.000   -1.65e+04    1.65e+04
ma.S.L48      -0.0822    693.628     -0.000      1.000   -1359.568    1359.404
sigma2       170.3708   1.44e+06      0.000      1.000   -2.83e+06    2.83e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 5.16
Prob(Q):                              0.98   Prob(JB):                         0.08
Heteroskedasticity (H):               0.77   Skew:                             0.61
Prob(H) (two-sided):                  0.52   Kurtosis:                         3.28
===================================================================================
```

Table 1.33: SARIMAX results in Manual SARIMA Model with best params of seasonality 24

Model Diagnostic Plot

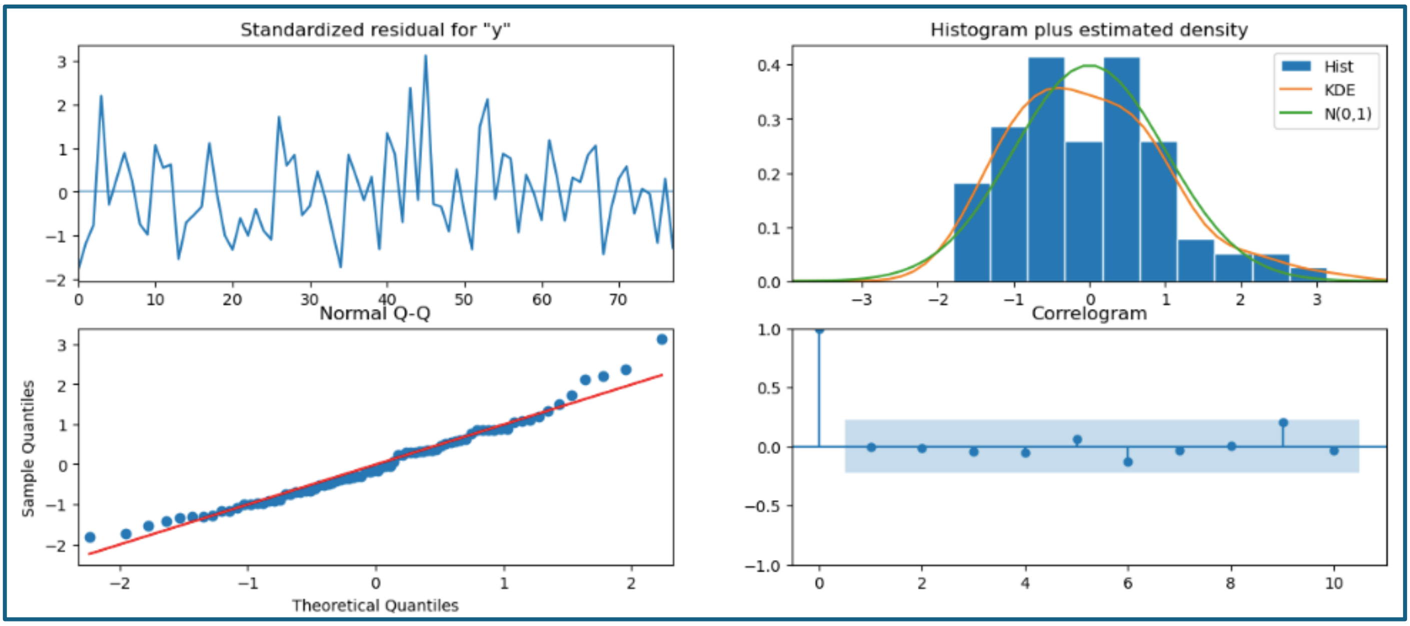


Figure 1.33: SARIMA Model Diagnostic Plot with seasonality 24

Similar to the last iteration of the model where the seasonality parameter was taken as 24, here also we see that the model diagnostics plot does not indicate any remaining information that we can get.

Predict on the Test Set using this model and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	90.727227	14.757082	61.803877	119.650576
1	111.564533	14.816957	82.523831	140.605236
2	58.147405	14.816958	29.106701	87.188108
3	66.914807	14.816958	37.874103	95.955511
4	79.986971	14.816958	50.946268	109.027675

Table 1.34: Prediction with SARIMA Model of seasonality 24

	RMSE
ARIMA(0,1,2)	30.903804
ARIMA(4,1,2)	30.607788
SARIMA(0,1,2)(2,0,2,12)	25.343324
SARIMA(0,1,2)(1,0,2,24)	19.581159

Table 1.35: Different model RSME value

Still Auto SARIMA model with seasonality 24 has the lowest RSME value till now.

5.3.2 Manual SARIMA

Differentiated ACF & PACF plot to get the p & q value for the model building

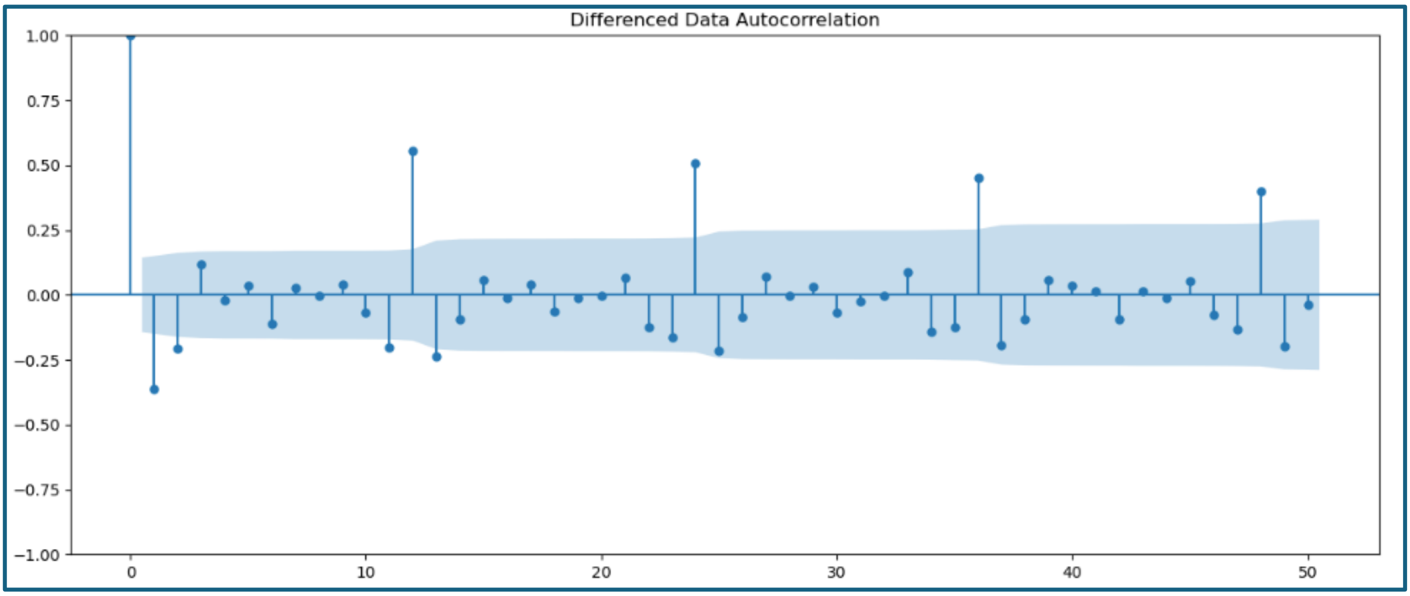


Figure 1.34: Differentiated ACF plot

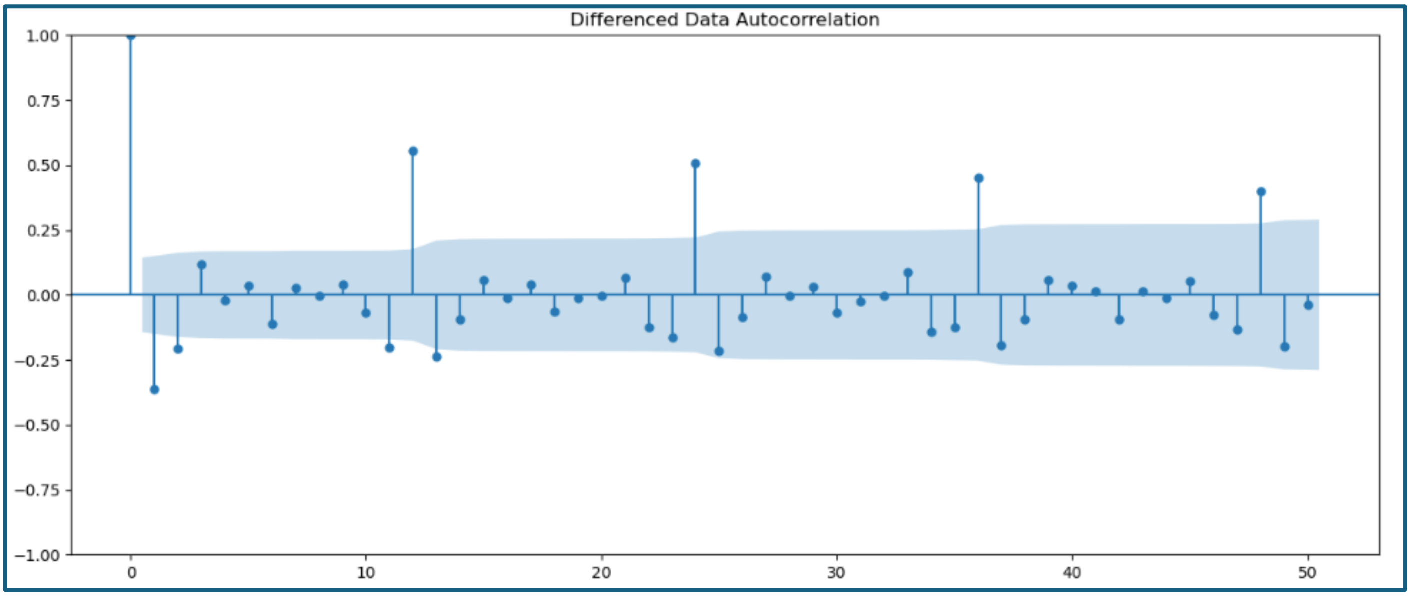


Figure 1.35: Differentiated PACF plot

We see that our ACF plot at the seasonal interval (12) does not taper off. So, we go ahead and take a seasonal differencing of the original series.

Original Data

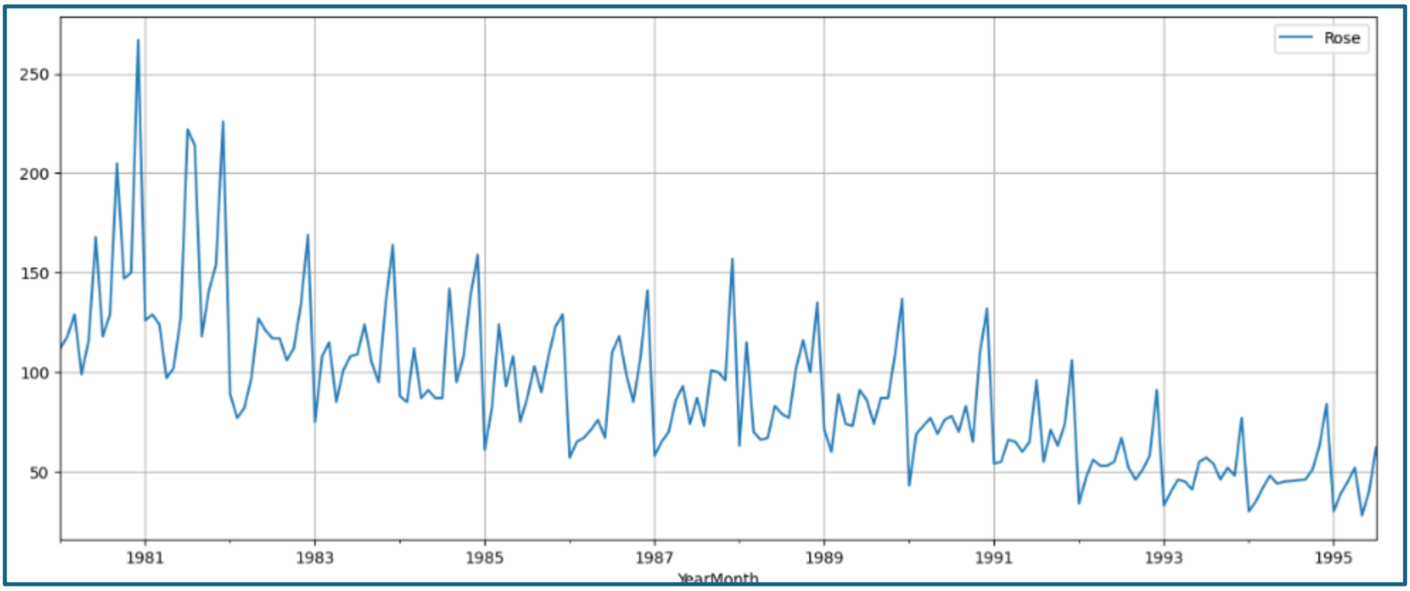


Figure 1.36: Original Data Plot

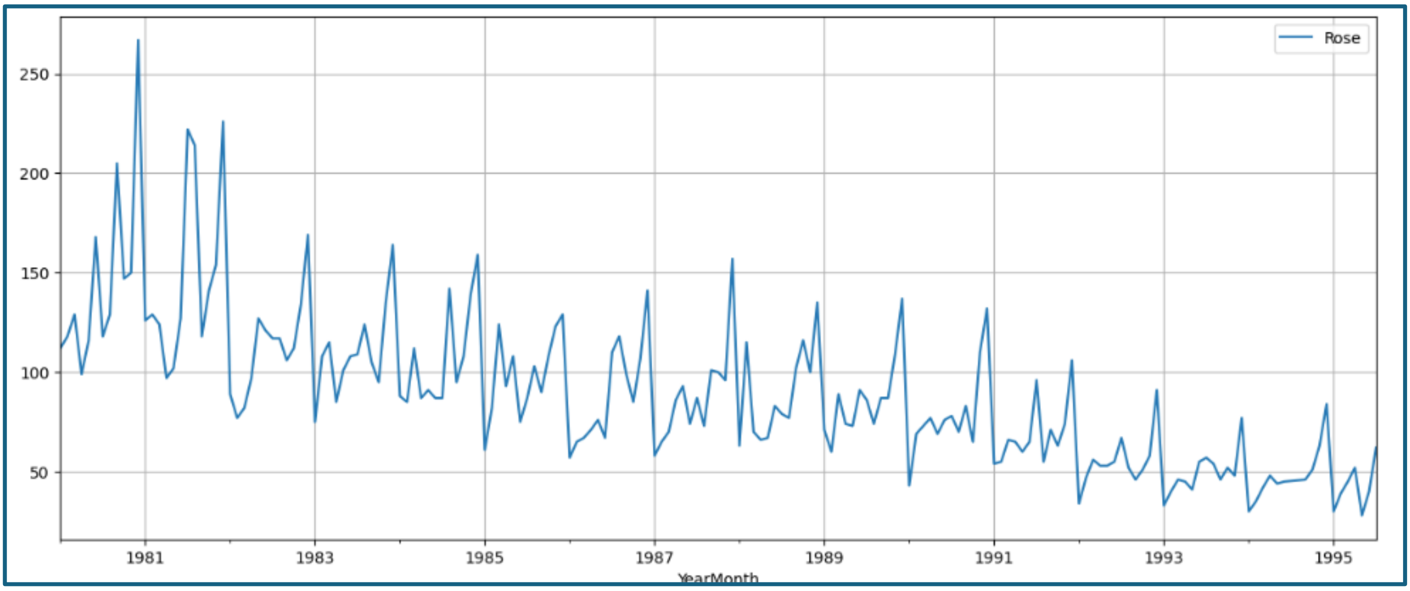


Figure 1.37: Differentiated Plot with Seasonality 24

We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.

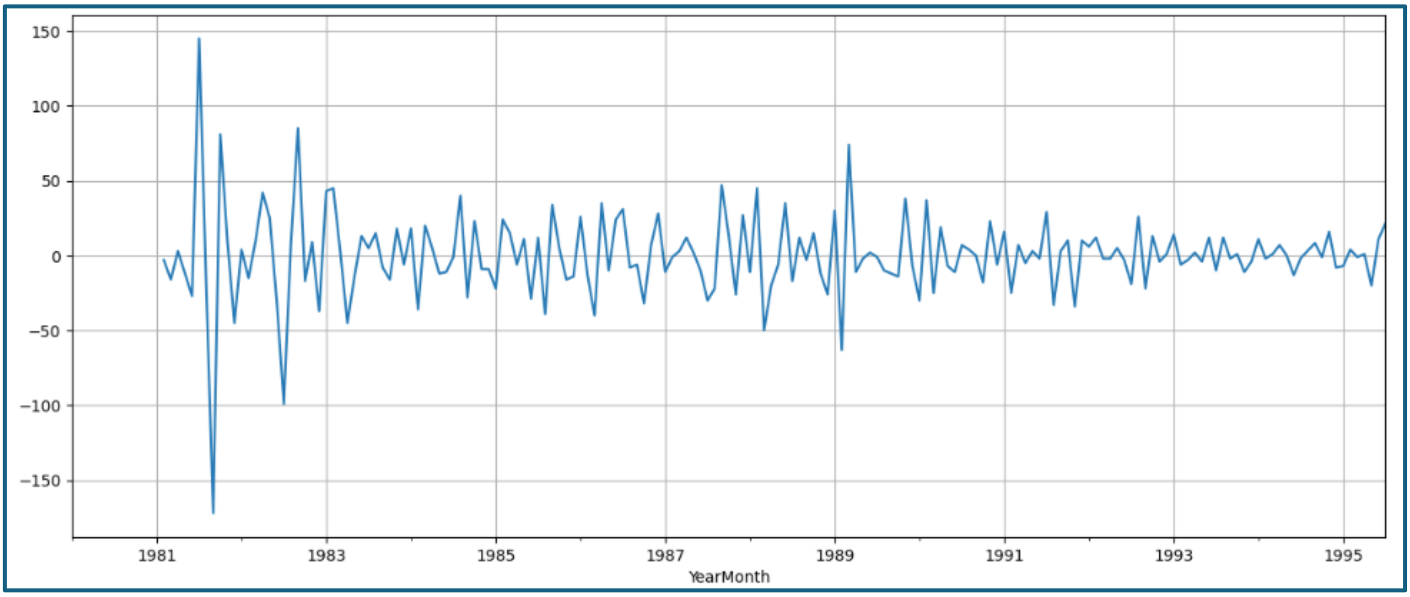
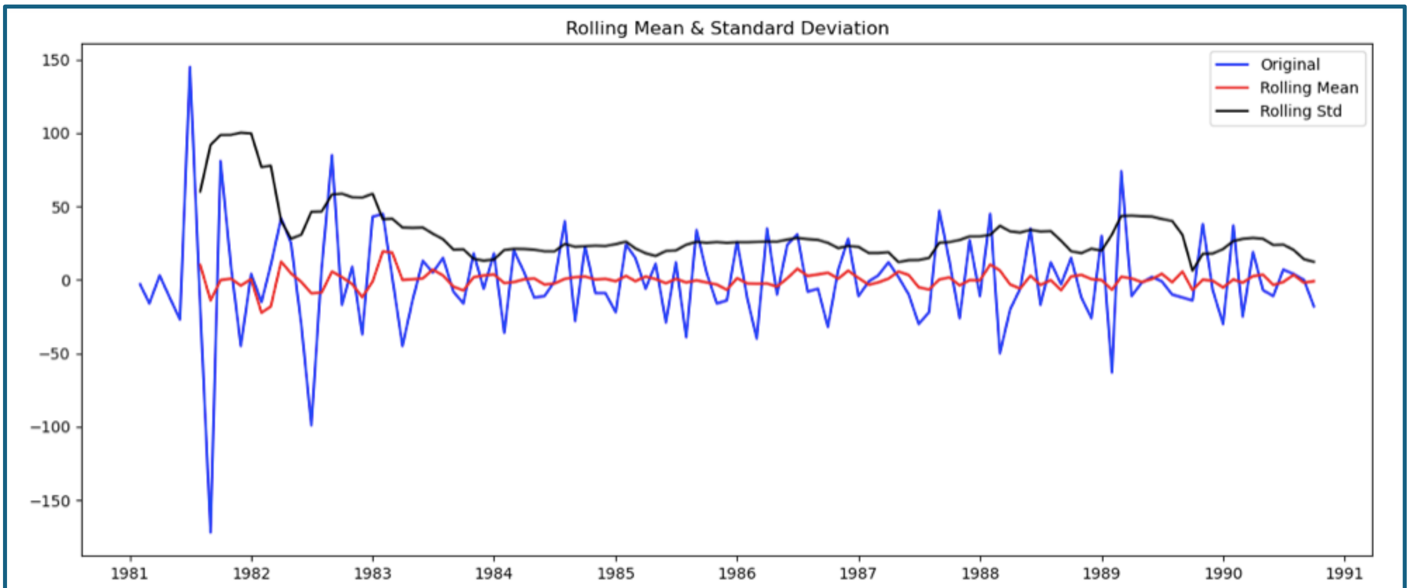


Figure 1.38: Differentiated 1 order after seasonality 24

Now we see that there is almost no trend present in the data. Seasonality is only present in the data. Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.




```
Results of Dickey-Fuller Test:
Test Statistic                 -3.664525
p-value                         0.004641
#Lags Used                     11.000000
Number of Observations Used   105.000000
Critical Value (1%)            -3.494220
Critical Value (5%)            -2.889485
Critical Value (10%)           -2.581676
dtype: float64
```

Figure 1.39: Differentiated PACF plot

The p-value is less than 0.05 and the data is stationary

Checking the ACF and the PACF plots for the new modified Time Series.

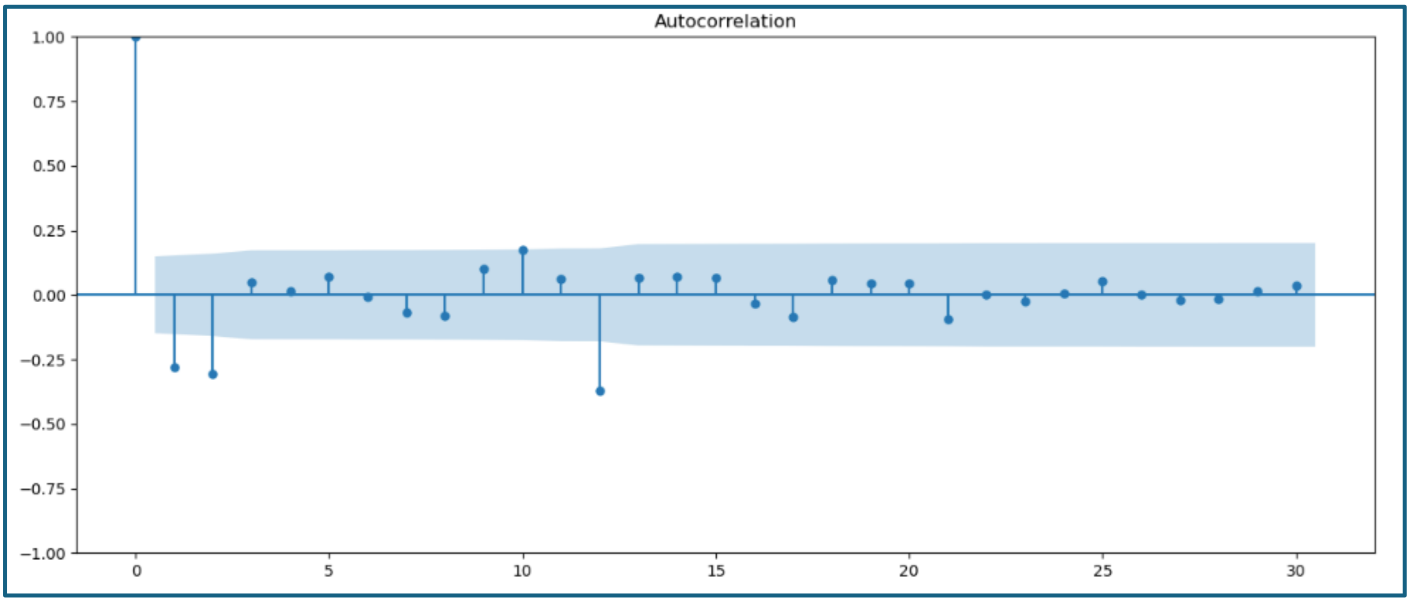


Figure 1.40: Modified ACF plot

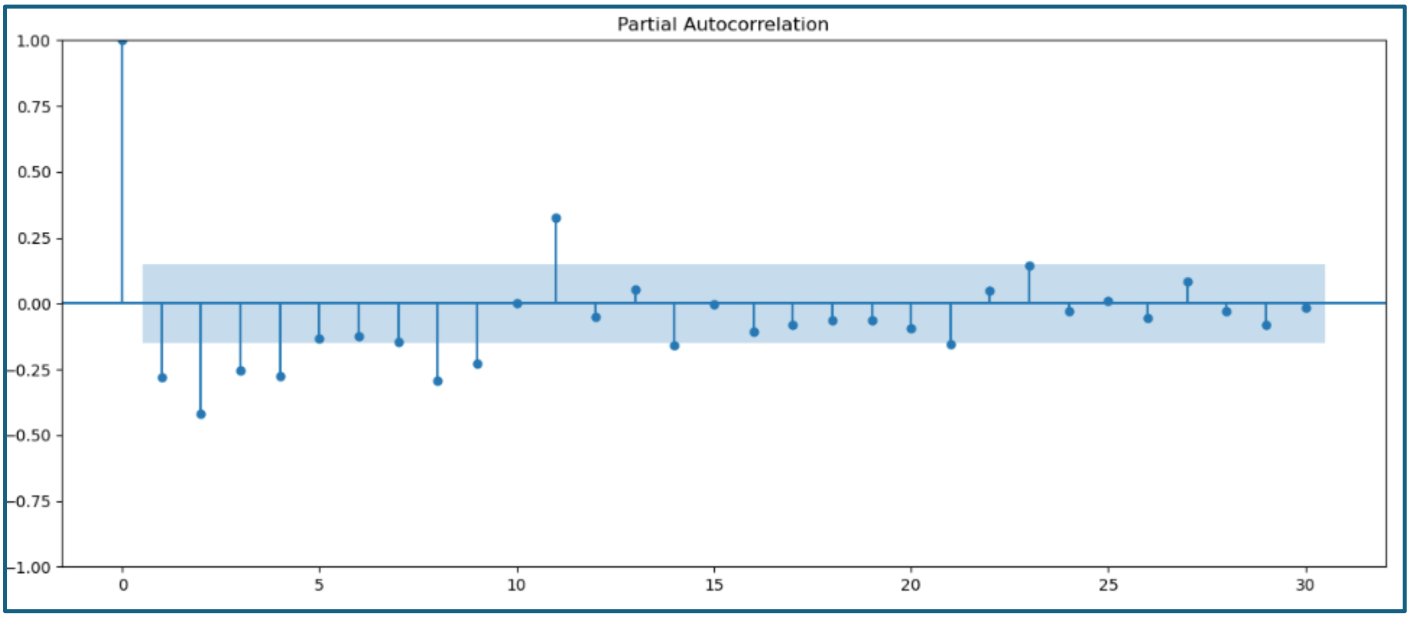


Figure 1.41: Modified PACF plot

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 6. We will keep the p(1) and q(1) parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 4.

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

Model Diagnostic Plot

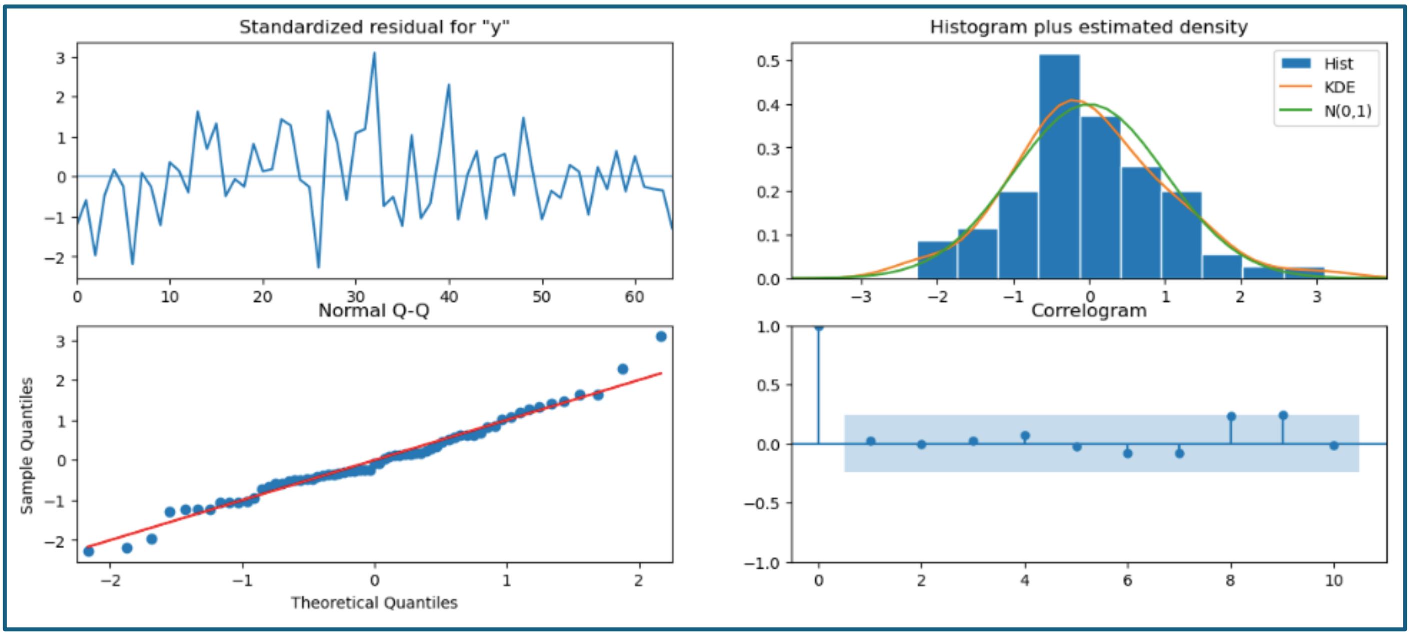


Figure 1.42: Manual SARIMA Model Diagnostic Plot with seasonality 24

Predict on the Test Set using this model and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	98.795028	14.961886	69.470271	128.119786
1	125.362230	15.208538	95.554043	155.170417
2	44.890262	15.226452	15.046965	74.733560
3	60.659579	15.431733	30.413938	90.905221
4	62.626596	15.434874	32.374800	92.878393

Table 1.36: Prediction with Manual SARIMA Model of seasonality 24

	RMSE
ARIMA(0,1,2)	30.903804
ARIMA(4,1,2)	30.607788
SARIMA(0,1,2)(2,0,2,12)	25.343324
SARIMA(0,1,2)(1,0,2,24)	19.581159
SARIMA(4,1,2)(4,1,2,12)	13.960344

Table 1.37: Different model RSME value

The Manual SARIMA model with seasonality 12 has the lowest RSME value till now.

Manual SARIMA with seasonality 12

Building the model again with the best params (4,1,2) (4,0,2,12)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  130
Model:             SARIMAX(4, 1, 2)x(1, 0, 2, 24)   Log Likelihood                -327.906
Date:                            Sun, 14 Apr 2024   AIC                            675.813
Time:                                    20:19:25   BIC                            699.380
Sample:                                         0   HQIC                           685.247
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1055      0.474      0.223      0.824      -0.823       1.034
ar.L2         -0.0966      0.128     -0.753      0.451      -0.348       0.155
ar.L3         -0.1056      0.122     -0.867      0.386      -0.344       0.133
ar.L4         -0.1279      0.154     -0.830      0.406      -0.430       0.174
ma.L1         -1.2322      0.597     -2.064      0.039      -2.402      -0.062
ma.L2          0.1255      0.619      0.203      0.839      -1.087       1.339
ar.S.L24       0.8632      0.059     14.740      0.000       0.748       0.978
ma.S.L24      -0.9301      0.278     -3.350      0.001      -1.474      -0.386
ma.S.L48      -0.0700      0.193     -0.362      0.717      -0.449       0.309
sigma2       137.0170      0.002   6.79e+04      0.000     137.013     137.021
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 6.02
Prob(Q):                              0.97   Prob(JB):                         0.05
Heteroskedasticity (H):               0.78   Skew:                             0.64
Prob(H) (two-sided):                  0.53   Kurtosis:                         3.44
===================================================================================
```

Table 1.38: SARIMAX results Manual SARIMA with seasonality 12

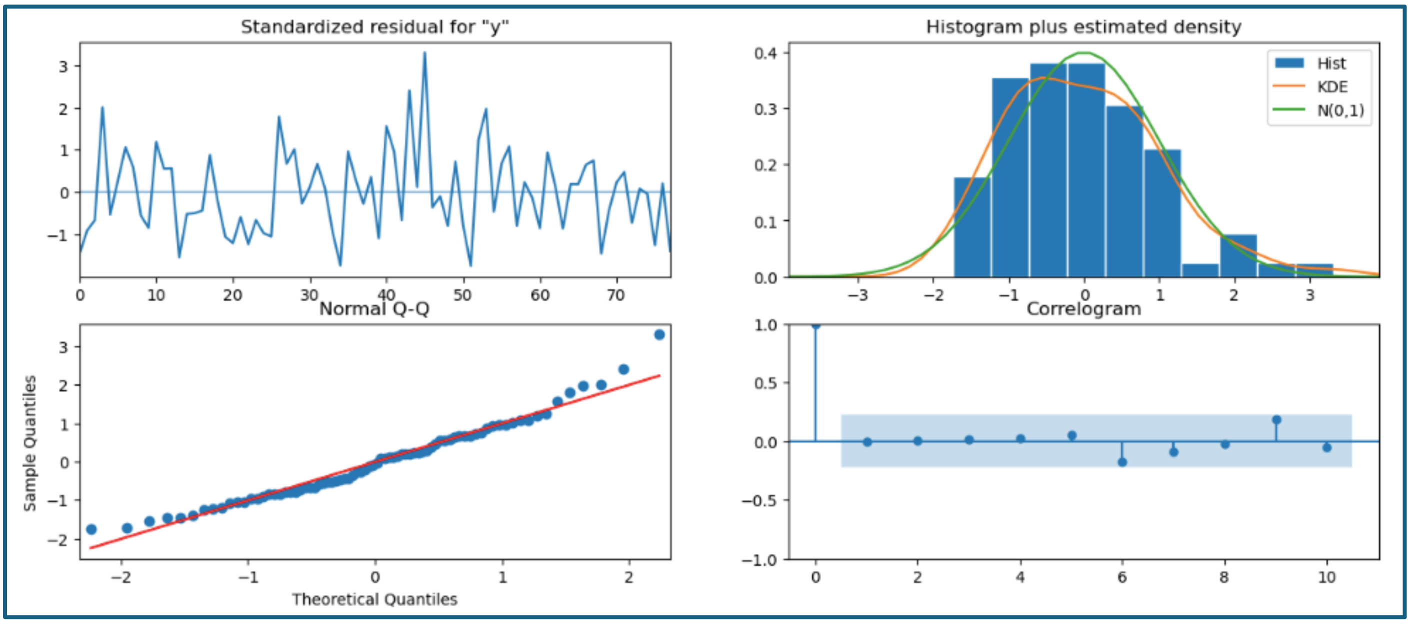


Figure 1.43: Manual SARIMA Model Diagnostic Plot with seasonality 12

Predict on the Test Set using this model and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	93.012882	14.631062	64.336527	121.689236
1	116.070103	14.704675	87.249470	144.890737
2	56.989493	14.702018	28.174067	85.804920
3	66.703424	14.701215	37.889572	95.517276
4	78.321689	14.713372	49.484010	107.159369

Table 1.39: Prediction with Manual SARIMA Model of seasonality 12

5.4 Check the performance of the models built

	RMSE
ARIMA(0,1,2)	30.903804
ARIMA(4,1,2)	30.607788
SARIMA(0,1,2)(2,0,2,12)	25.343324
SARIMA(0,1,2)(1,0,2,24)	19.581159
SARIMA(4,1,2)(4,1,2,12)	13.960344
SARIMA(4,1,2)(1,2,0,24)	17.700300

Table 1.40: Different model RSME value

In the models build using ARIMA & SARIMA the best model is Manual SARIMA with params of (4,1,2) (4,1,2,12) if seasonality 12 has the lowest RMSE among the other models.

6. Compare the performance of the models

6.1 Compare the performance of all the models built

The models with RMSE values are

	Test RMSE
RegressionOnTime	17.355796
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919
Alpha=0.1027210902642377,SimpleExponentialSmoothing	30.188322
Alpha=0.9,SimpleExponentialSmoothing	22.496819
Alpha=0.3,Beta=0.4,DoubleExponentialSmoothing	18.343250
Alpha=0.099,Beta=1.993,Gamma=0.000,TripleExponentialSmoothing	9.328733
Alpha=0.1,Beta=1.0,Gamma=0.2,TripleExponentialSmoothing	9.129075

Table 1.41: All model RSME value

	RMSE
ARIMA(0,1,2)	30.903804
ARIMA(4,1,2)	30.607788
SARIMA(0,1,2)(2,0,2,12)	25.343324
SARIMA(0,1,2)(1,0,2,24)	19.581159
SARIMA(4,1,2)(4,1,2,12)	13.960344
SARIMA(4,1,2)(1,2,0,24)	17.700300

6.2 Choose the best model with proper rationale

The best model to choose is the Manual SARIMA model with best params of (4,1,2) (4,1,2,12) and seasonality 12 has the lowest RMSE values across all the model.

We can build optimum model on the full data with this params.

6.3 Rebuild the best model using the entire data

Building the most optimum model on the Full Data.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:               187
Model:             SARIMAX(4, 1, 2)x(4, 1, 2, 12)   Log Likelihood              -484.502
Date:                            Sun, 14 Apr 2024   AIC                          995.005
Time:                                    20:19:45   BIC                         1031.457
Sample:                                01-01-1980   HQIC                        1009.810
                                     - 07-01-1995
Covariance Type:                              opg
==================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
ar.L1         -0.9669      0.129     -7.501      0.000      -1.220      -0.714
ar.L2         -0.0276      0.185     -0.149      0.881      -0.390       0.335
ar.L3          0.0180      0.153      0.117      0.907      -0.283       0.319
ar.L4         -0.0205      0.091     -0.226      0.821      -0.199       0.158
ma.L1          0.1386      4.255      0.033      0.974      -8.202       8.479
ma.L2         -0.8618      3.678     -0.234      0.815      -8.071       6.348
ar.S.L12      -0.6692      0.187     -3.586      0.000      -1.035      -0.303
ar.S.L24      -0.1392      0.169     -0.823      0.410      -0.471       0.192
ar.S.L36      -0.1891      0.081     -2.334      0.020      -0.348      -0.030
ar.S.L48      -0.1751      0.045     -3.852      0.000      -0.264      -0.086
ma.S.L12       0.1236      0.217      0.569      0.570      -0.303       0.550
ma.S.L24      -0.3106      0.187     -1.664      0.096      -0.676       0.055
sigma2       156.8750    674.915      0.232      0.816   -1165.934    1479.684
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 7.85
Prob(Q):                              0.92   Prob(JB):                         0.02
Heteroskedasticity (H):               0.20   Skew:                             0.26
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.13
===================================================================================
```

Table 1.42: SARIMAX results Manual SARIMA with seasonality 12

The SARIMAX results of SARIMA model with seasonality 12 along with the best params (4,1,2) (4,1,2,12)

6.4 Make a forecast for the next 12 months

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	44.23	12.58	19.57	68.89
1995-09-30	45.91	12.75	20.93	70.89
1995-10-31	47.57	12.80	22.49	72.65
1995-11-30	59.55	13.05	33.98	85.12
1995-12-31	86.41	13.07	60.80	112.02
1996-01-31	25.02	13.33	-1.11	51.15
1996-02-29	32.09	13.36	5.90	58.28
1996-03-31	39.84	13.59	13.21	66.48
1996-04-30	44.15	13.63	17.44	70.87
1996-05-31	30.29	13.83	3.19	57.39
1996-06-30	39.43	13.88	12.22	66.63
1996-07-31	55.18	14.05	27.64	82.72

Table 1.43: Predicted values for 12 months

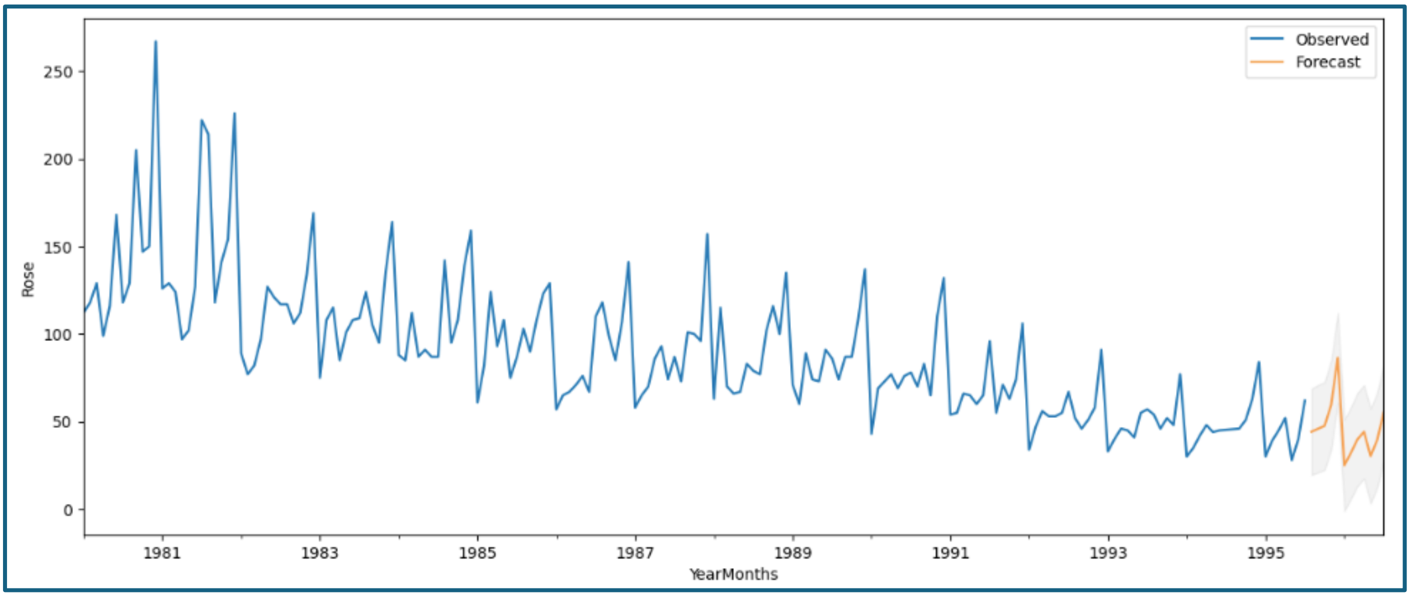


Figure 1.44: Final forecasted plot on the entire data

7. Actionable Insights & Recommendations

7.1 Key takeaways (actionable insights and recommendations) for the business

Insights:

- There is decreasing trend in the data series
- There are outliers in years data except 5 years
- In monthly trend there was increase, steady and increase in trend
- December month across years has the highest sales
- On quarterly basis the 4th quarter has the highest sales with decreasing trend
- The data is right skewed as the mean is greater than median
- In decomposition, multiplicative model is the suited model as there no specific pattern in the residual
- The data has decreasing trend and seasonality
- Across LR, SA, MA, SES, DES & TES – TES has the lowest RMSE value
- The DataFrame is not stationary and AD fuller method is used to make it stationary
- Among Auto ARIMA & SARIMA, Manual ARIMA & SARIMA – Manual SARIMA with seasonality 12 has the lowest RMSE value

Recommendations:

- The Rose wines variety has the decreasing or lower sales in the forecast also
- The company must review the quality and the price of the wine compared to the competitors
- The marketing spend has to be increased in order to increase the sales for a shorter period
- If there is no ROI that the company can stop or withdraw the Rose variety
- The wine can be made available only on the winter season where the sales are higher