TSF Project - Coded

Rose

Isha Shukla

25 May 2024

INDEX

SL No.	Title	Page No.
1	Define the problem and perform Exploratory Data Analysis	3
2	Data Pre-processing	15
3	Model Building - Original Data	16
4	Check for Stationarity	25
5	Model Building - Stationary Data	29
6	Compare the performance of the models	44
7	Actionable Insights & Recommendations	46

Plots

1. Line plot of dataset
2. Boxplot of dataset
3. Lineplot of sales
4. Boxplot of yearly data
5. Boxplot of monthly data
6. Weekly boxplot
7. Graph of monthly sales over the year
8. Correlation
9. ECDF plot
10. Decomposition addictive
11. Decomposition multiplicative
12. Train and test dataset
13. Linear regression
14. Moving average
15. Simple exponential smoothing
16. Double exponential smoothing
17. Naive approach
18. Simple average
19. Triple exponential smoothing
20. Dickey fuller test
21. Dickey fuller test after diff
22. Auto ARIMA plot
23. Auto SARIMA plots
24. Manual ARIMA
25. Manual SARIMA
26. PACF and ACF plot
27. PACF and ACF plot train dataset
28. Manual ARIMA plot
29. Manual SARIMA plot
30. Prediction plot

Problem Statement

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, analyze 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.

The main goal of this project is to study and predict wine sales trends from the 20th century using historical data from ABC Estate Wines. We want to give ABC Estate Wines useful insights to improve sales, take advantage of new market opportunities, and stay competitive in the wine industry.

1. Define the problem and perform Exploratory Data Analysis

(I) Read the data

- There are two columns in the dataset Rose.csv
- The dataset has 187 rows.
- Columns - YearMonth(datatype as object) and Rose (datatype as float)
- Rose column has 2 null values.

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

Null values in 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: 3.1+ KB
```

Information about the dataset

Table 1 - Rows of the dataset

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

Top 5 rows of the dataset

	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

Last 5 rows of the dataset

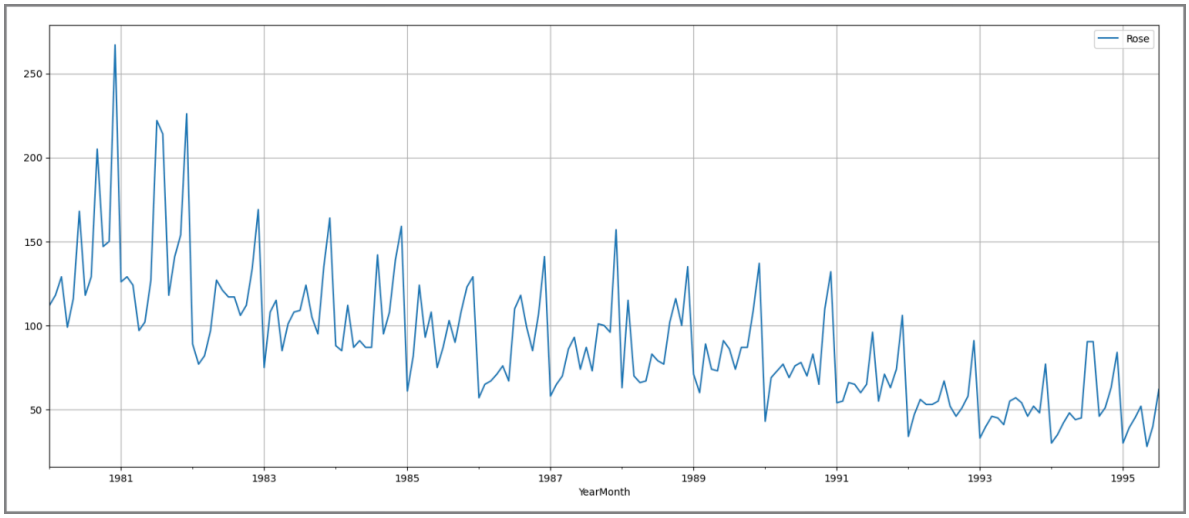
Table 2 - Statistical Summary of the dataset.

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

Comprehensive Summary of the dataset.

II. Plot the data

Plot 1 - Plot the data



Plot of the dataset - Rose vs YearMonth

For enhanced analysis of the dataset, we have segmented it further by extracting the month and year components from the 'YearMonth' column. This division allows for more granular examination of the data based on month-to-month and year-to-year trends. Now, there are 3 columns and 187 rows.

Table 3 - After extraction of Year and Month.

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

Top 5 rows after extraction of Year and Month

IV. Perform Exploratory Data Analysis (EDA)

- There are 2 null values.
- Replace missing values in the Rose column with the calculated mean.

Treating missing values is crucial for maintaining data integrity, ensuring accurate analyses, and deriving reliable insights, thereby enabling informed decision-making and valid conclusions.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    187 non-null    float64
 1   Year    187 non-null    int32
 2   Month   187 non-null    int32
dtypes: float64(1), int32(2)
memory usage: 4.4 KB
```

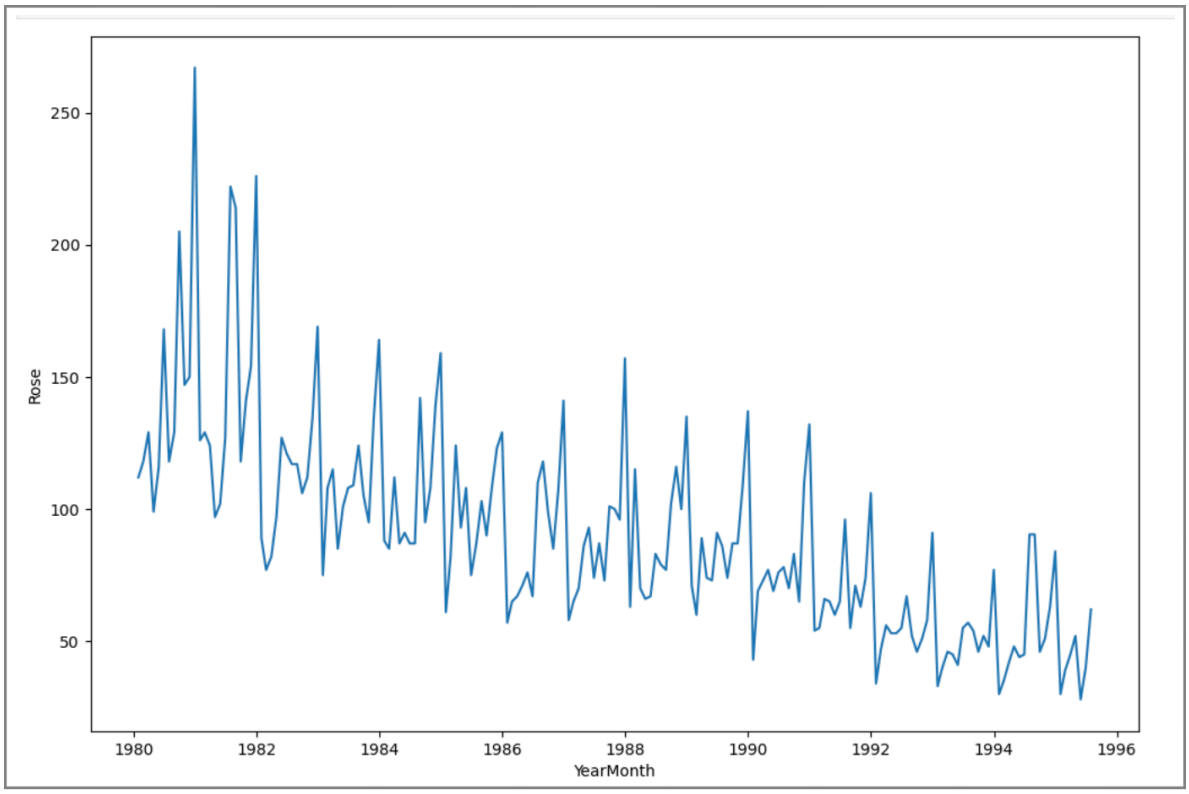
After treatment of missing value

- We'll resample the data to aggregate values at a monthly level from the daily-level data, computing the average for each month.

YearMonth	Rose	Year	Month
1980-01-31	112.0	1980.0	1.0
1980-02-29	118.0	1980.0	2.0
1980-03-31	129.0	1980.0	3.0
1980-04-30	99.0	1980.0	4.0
1980-05-31	116.0	1980.0	5.0

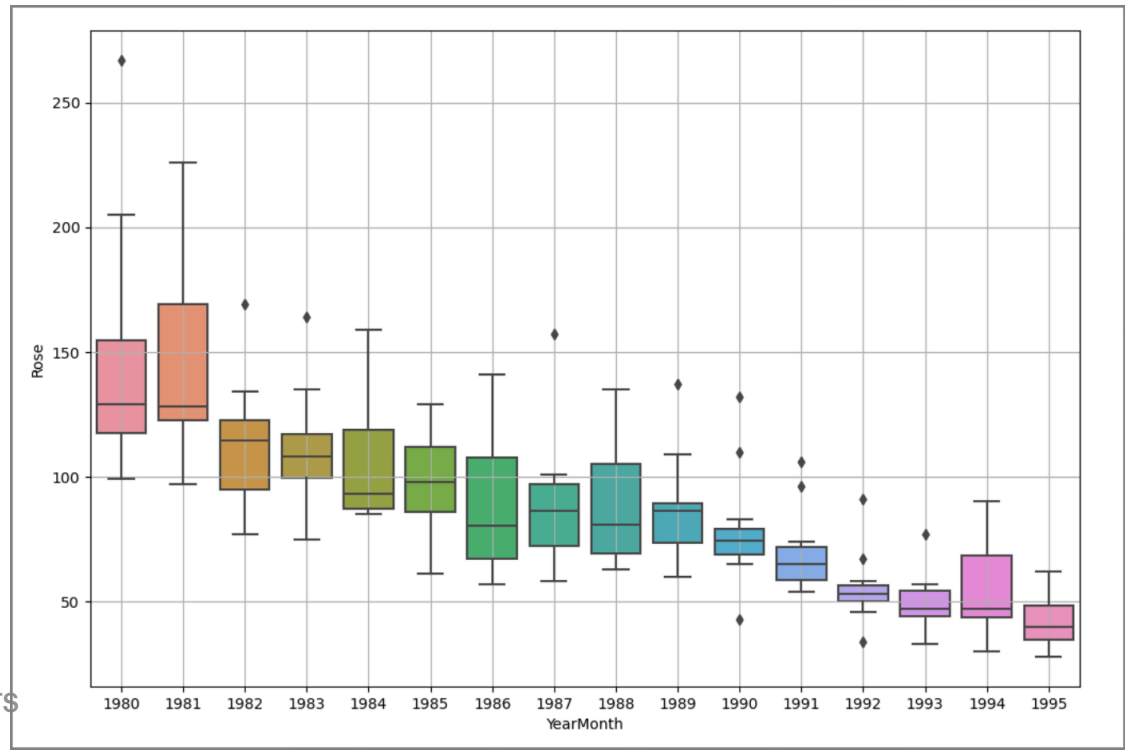
After resampling of the dataset

Plot 2 - The trend of Rose at year level



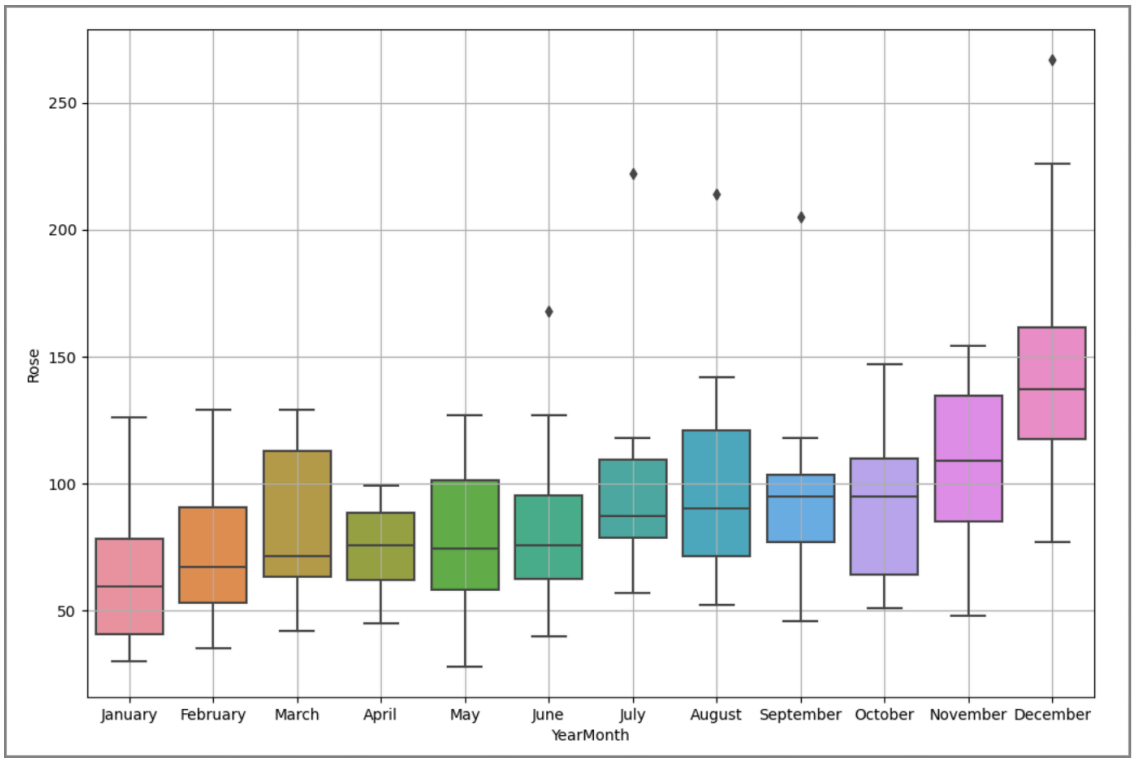
- There was a peak in 1981. The plot shows that there is trend and seasonality.

Plot 3 - Yearly Box-plot



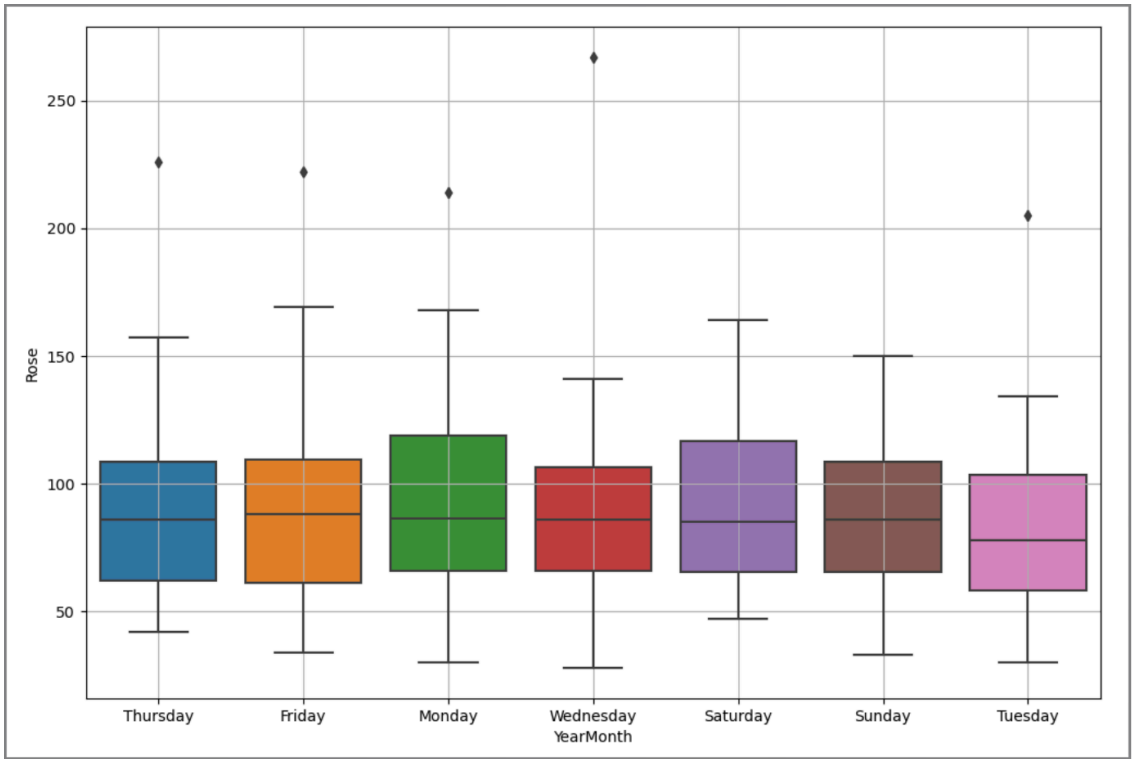
Yearly Box-plot

Plot 4 - Monthly Boxplot



Monthly Boxplot

Plot 5 - Weekly Box-plot



Weekly Box-plot

Outliers Observation

- **Yearly Boxplot** - Outliers persist across nearly all years. There was a peek in 1981.
- **Monthly Boxplot** - The graph indicates that wine sales peak in December and hit their lowest point in January. Sales remain steady from January to June, but then begin to rise steadily from July onwards. However, there are some outliers in June, July, August, September, and December.
- **Weekly Boxplot -** There are some outliers in Thursday, Friday, Monday, Wednesday and Tuesday.

Table 4 - Pivot table displays monthly price across years

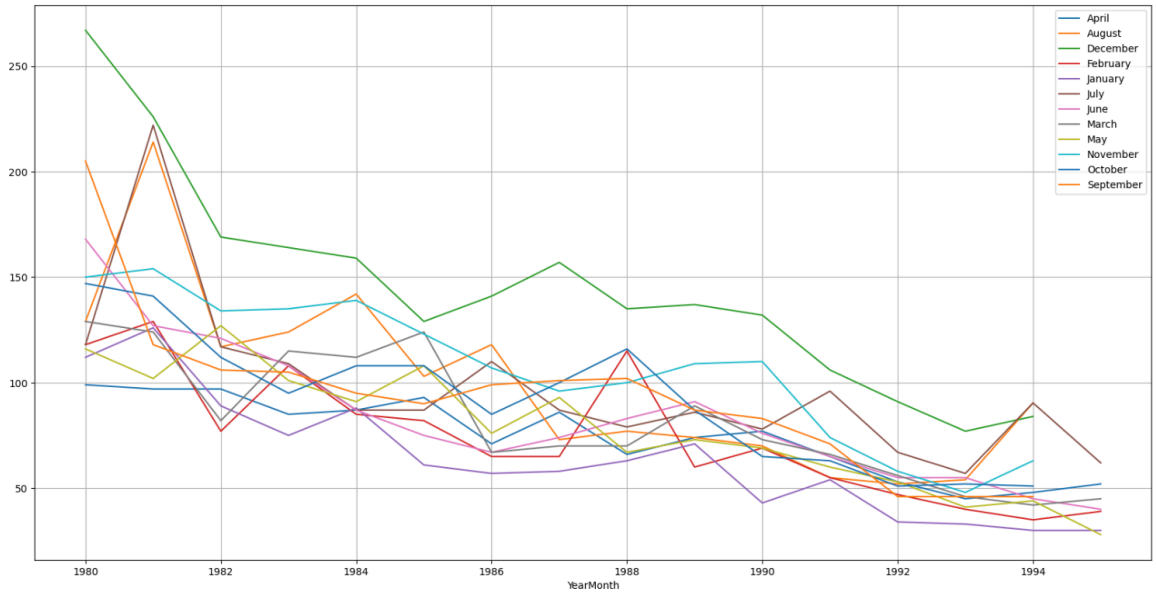
YearMonth	April	August	December	February	January	July	June	March	May	November	October	September
YearMonth												
1980	99.0	129.000000	267.0	118.0	112.0	118.000000	168.0	129.0	116.0	150.0	147.0	205.0
1981	97.0	214.000000	226.0	129.0	126.0	222.000000	127.0	124.0	102.0	154.0	141.0	118.0
1982	97.0	117.000000	169.0	77.0	89.0	117.000000	121.0	82.0	127.0	134.0	112.0	106.0
1983	85.0	124.000000	164.0	108.0	75.0	109.000000	108.0	115.0	101.0	135.0	95.0	105.0
1984	87.0	142.000000	159.0	85.0	88.0	87.000000	87.0	112.0	91.0	139.0	108.0	95.0
1985	93.0	103.000000	129.0	82.0	61.0	87.000000	75.0	124.0	108.0	123.0	108.0	90.0
1986	71.0	118.000000	141.0	65.0	57.0	110.000000	67.0	67.0	76.0	107.0	85.0	99.0
1987	86.0	73.000000	157.0	65.0	58.0	87.000000	74.0	70.0	93.0	96.0	100.0	101.0
1988	66.0	77.000000	135.0	115.0	63.0	79.000000	83.0	70.0	67.0	100.0	116.0	102.0
1989	74.0	74.000000	137.0	60.0	71.0	86.000000	91.0	89.0	73.0	109.0	87.0	87.0
1990	77.0	70.000000	132.0	69.0	43.0	78.000000	76.0	73.0	69.0	110.0	65.0	83.0
1991	65.0	55.000000	106.0	55.0	54.0	96.000000	65.0	66.0	60.0	74.0	63.0	71.0
1992	53.0	52.000000	91.0	47.0	34.0	67.000000	55.0	56.0	53.0	58.0	51.0	46.0
1993	45.0	54.000000	77.0	40.0	33.0	57.000000	55.0	46.0	41.0	48.0	52.0	46.0
1994	48.0	90.394595	84.0	35.0	30.0	90.394595	45.0	42.0	44.0	63.0	51.0	46.0
1995	52.0	NaN	NaN	39.0	30.0	62.000000	40.0	45.0	28.0	NaN	NaN	NaN

Pivot Table

Here are observations from the pivot table:

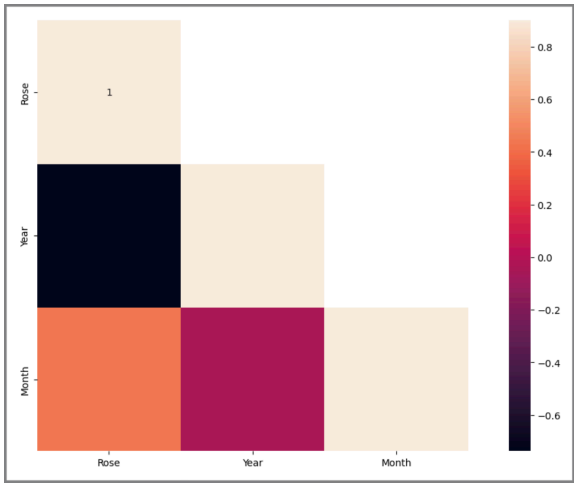
- There are missing values for August, October, September, December, and November in 1995.

Plot 6 - Pivot table plot for monthly wine sale across year



Graph to display monthly price across years

Plot 7 - Correlation Plot



Correlation Plot

	Rose	Year	Month
Rose	1.000000	-0.737184	0.439426
Year	-0.737184	1.000000	-0.046502
Month	0.439426	-0.046502	1.000000

Rose Wine sales Correlation with respect to year and month

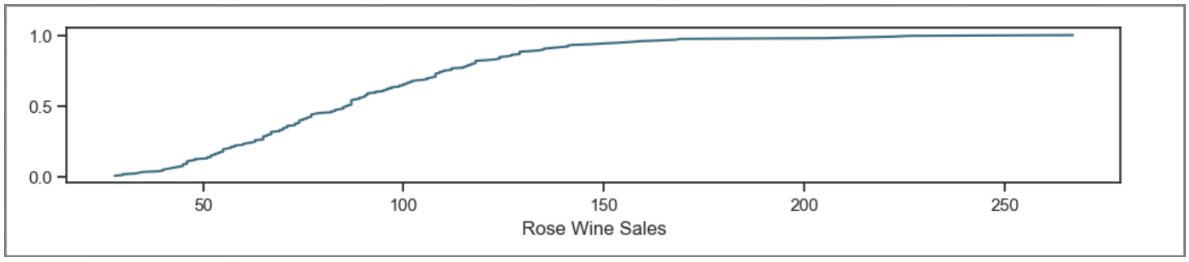
Observations from the correlation table and plot:

●The strong negative correlation between Rose Wine sales and Year suggests a clear trend over time. This

indicates that there is a consistent decrease in Rose wine sales values as the years progress, implying a long-term downward trend in Rose sales.

- The moderate positive correlation between Rose Wine sales and Month indicates some seasonality in Rose sales. This suggests that there is a tendency for Rose sales to increase as the month progresses, implying a seasonal pattern within each year. However, this correlation is not as strong as the trend observed over the years.
- Overall, these observations suggest that while there is both a long-term trend of decreasing Rose Wine sales over the years and a seasonal pattern of increasing sales within each year, the trend effect is more pronounced than the seasonal effect.

Plot 8 - Empirical Cumulative Distribution Function (ECDF)



Graph for ECDF

From the ECDF plot of wine sales observations, we can observe the following:

- The x-axis represents the range of wine sales observations, while the y-axis represents the cumulative probability.
- The plot shows how the cumulative probability increases as we move along the sorted wine sales values.
- By examining the slope of the curve, we can infer the density of the observations at different values. Steeper slopes indicate higher density of observations, while flatter slopes indicate lower density.

- The ECDF plot provides a comprehensive overview of the distribution of wine sales observations, allowing us to assess characteristics such as central tendency, spread, and percentiles.
- Overall, the ECDF plot helps us understand the empirical distribution of wine sales observations and can provide insights into the underlying patterns and variability in the data.

```
<statsmodels.distributions.empirical_distribution.ECDF object at 0x15932b350>
[        -inf  28.          30.          30.          33.
  34.          35.          39.          40.          40.
  41.          42.          43.          44.          45.
  45.          45.          46.          46.          46.
  46.          47.          48.          48.          51.
  51.          52.          52.          52.          53.
  53.          54.          54.          55.          55.
  55.          55.          56.          57.          57.
  58.          58.          60.          60.          61.
  62.          63.          63.          63.          65.
  65.          65.          65.          65.          66.
  66.          67.          67.          67.          67.
  69.          69.          70.          70.          70.
  71.          71.          71.          73.          73.
  73.          74.          74.          74.          74.
  75.          75.          76.          76.          77.
  77.          77.          77.          78.          79.
  82.          82.          83.          83.          84.
  85.          85.          85.          86.          86.
  87.          87.          87.          87.          87.
  87.          87.          88.          89.          89.
  90.          90.39459459  90.39459459  91.          91.
  91.          93.          93.          95.          95.
  96.          96.          97.          97.          99.
  99.         100.         100.         101.         101.
 102.         102.         103.         105.         106.
 106.         107.         108.         108.         108.
 108.         108.         109.         109.         110.
 110.         112.         112.         112.         115.
 115.         116.         116.         117.         117.
 118.         118.         118.         118.         121.
 123.         124.         124.         124.         126.
 127.         127.         129.         129.         129.
 129.         132.         134.         135.         135.
 137.         139.         141.         141.         142.
 147.         150.         154.         157.         159.
 164.         168.         169.         205.         214.
 222.         226.         267.        ]
```

ECDF for X- axis(Wine sales)

```
[0.         0.00534759 0.01069519 0.01604278 0.02139037 0.02673797
 0.03208556 0.03743316 0.04278075 0.04812834 0.05347594 0.05882353
 0.06417112 0.06951872 0.07486631 0.0802139  0.0855615  0.09090909
 0.09625668 0.10160428 0.10695187 0.11229947 0.11764706 0.12299465
 0.12834225 0.13368984 0.13903743 0.14438503 0.14973262 0.15508021
 0.16042781 0.1657754  0.17112299 0.17647059 0.18181818 0.18716578
 0.19251337 0.19786096 0.20320856 0.20855615 0.21390374 0.21925134
 0.22459893 0.22994652 0.23529412 0.24064171 0.2459893  0.2513369
 0.25668449 0.26203209 0.26737968 0.27272727 0.27807487 0.28342246
 0.28877005 0.29411765 0.29946524 0.30481283 0.31016043 0.31550802
 0.32085561 0.32620321 0.3315508  0.3368984  0.34224599 0.34759358
 0.35294118 0.35828877 0.36363636 0.36898396 0.37433155 0.37967914
 0.38502674 0.39037433 0.39572193 0.40106952 0.40641711 0.41176471
 0.4171123  0.42245989 0.42780749 0.43315508 0.43850267 0.44385027
 0.44919786 0.45454545 0.45989305 0.46524064 0.47058824 0.47593583
 0.48128342 0.48663102 0.49197861 0.4973262  0.5026738  0.50802139
 0.51336898 0.51871658 0.52406417 0.52941176 0.53475936 0.54010695
 0.54545455 0.55080214 0.55614973 0.56149733 0.56684492 0.57219251
 0.57754011 0.5828877  0.58823529 0.59358289 0.59893048 0.60427807
 0.60962567 0.61497326 0.62032086 0.62566845 0.63101604 0.63636364
 0.64171123 0.64705882 0.65240642 0.65775401 0.6631016  0.6684492
 0.67379679 0.67914439 0.68449198 0.68983957 0.69518717 0.70053476
 0.70588235 0.71122995 0.71657754 0.72192513 0.72727273 0.73262032
 0.73796791 0.74331551 0.7486631  0.7540107  0.75935829 0.76470588
 0.77005348 0.77540107 0.78074866 0.78609626 0.79144385 0.79679144
 0.80213904 0.80748663 0.81283422 0.81818182 0.82352941 0.82887701
 0.8342246  0.83957219 0.84491979 0.85026738 0.85561497 0.86096257
 0.86631016 0.87165775 0.87700535 0.88235294 0.88770053 0.89304813
 0.89839572 0.90374332 0.90909091 0.9144385  0.9197861  0.92513369
 0.93048128 0.93582888 0.94117647 0.94652406 0.95187166 0.95721925
 0.96256684 0.96791444 0.97326203 0.97860963 0.98395722 0.98930481
 0.99465241 1.        ]
```

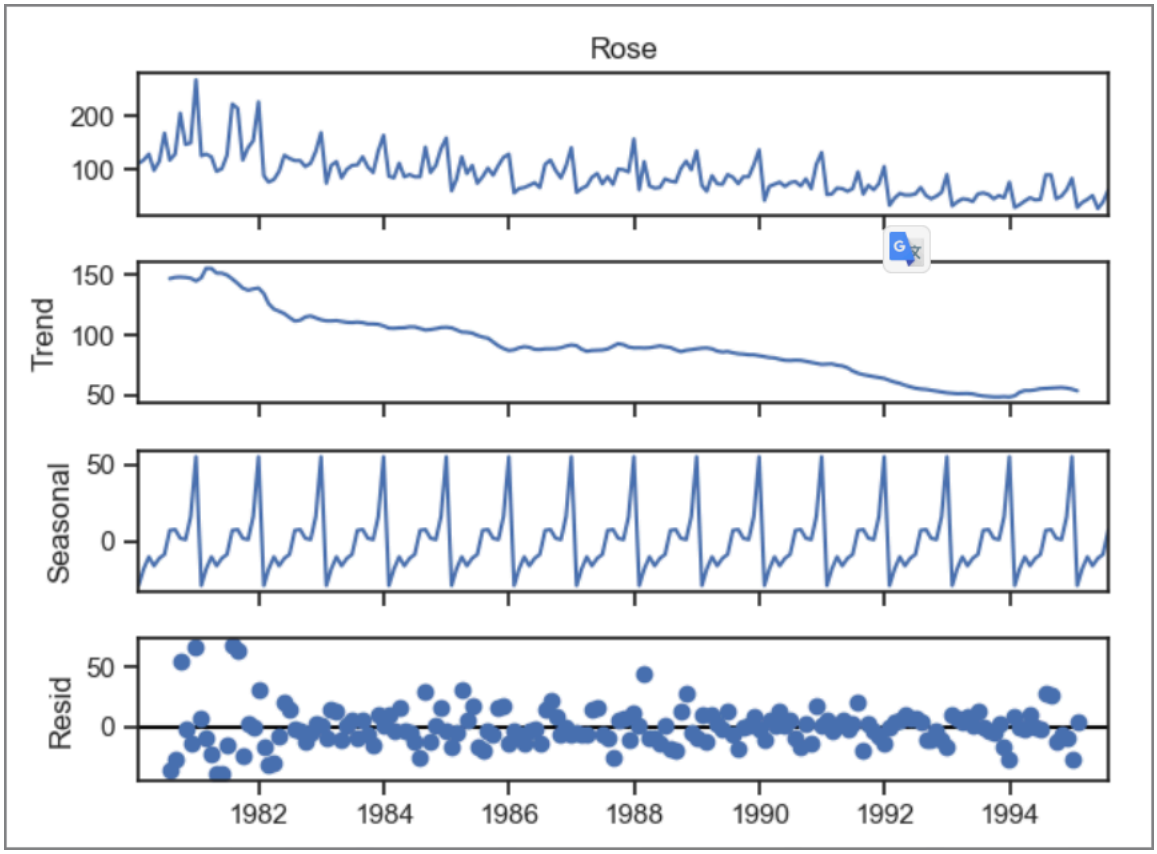
ECDF for Y-axis(Cumulative Probability)

- Highest Wine sale is 267.
- Lowest wine sale is 28.
- More than 75% has sales less than 150.

IV. Decomposition

a. Additive

Plot 9 - Additive Decomposition



Decomposition - Additive

- Trend and Seasonality is present.
- Residue/Noise is also there.
- Trend is decreasing with respect to year.
- Rose wine sales increases as the month progresses, implying a seasonal pattern within each year.
- Peak year was 1981. Afterward sales is decreasing over the time.

```
Trend
 YearMonth
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
1980-11-30    147.416667
1980-12-31    145.125000
Freq: M, Name: trend, dtype: float64

Seasonality
 YearMonth
1980-01-31   -28.403723
1980-02-29   -17.833219
1980-03-31    -9.816537
1980-04-30   -15.629037
1980-05-31   -10.727251
1980-06-30    -8.209394
1980-07-31     7.405916
1980-08-31     7.986472
1980-09-30     2.279610
1980-10-31     1.376832
1980-11-30    16.351832
1980-12-31    55.218499
Freq: M, Name: seasonal, dtype: float64

Residual
 YearMonth
1980-01-31          NaN
1980-02-29          NaN
1980-03-31          NaN
1980-04-30          NaN
1980-05-31          NaN
1980-06-30          NaN
1980-07-31   -36.489250
1980-08-31   -27.111472
1980-09-30    54.345390
1980-10-31    -2.460165
1980-11-30   -13.768499
1980-12-31    66.656501
Freq: M, Name: resid, dtype: float64
```

Additive Decomposition - Trend, Seasonality and Residual in year 1980

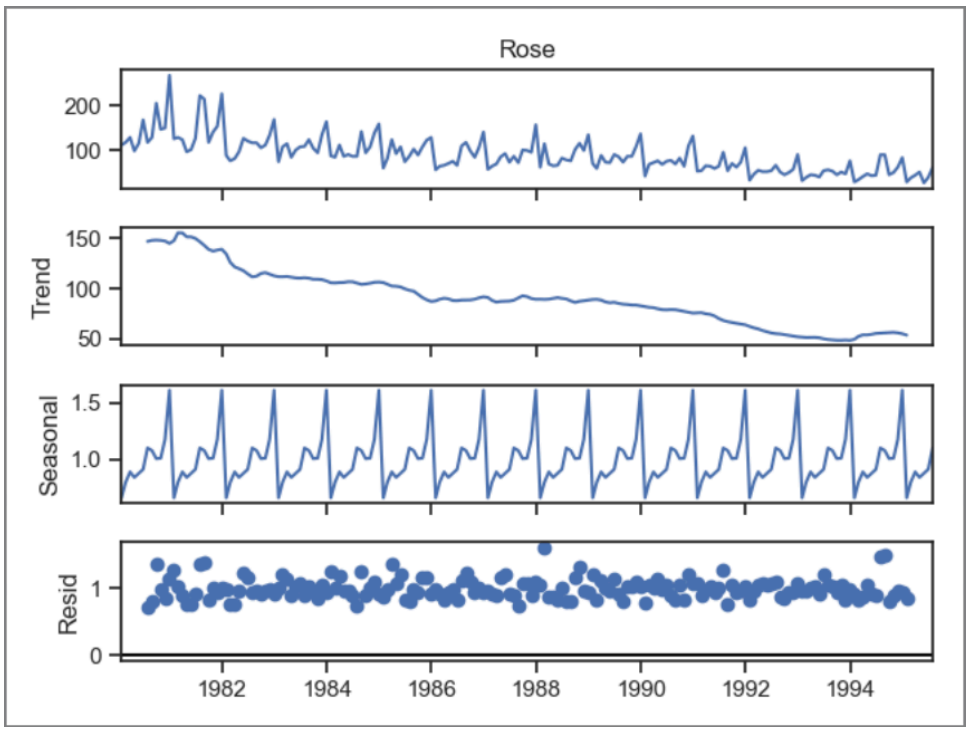
```
Trend
 YearMonth
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
1980-11-30    147.416667
1980-12-31    145.125000
Freq: M, Name: trend, dtype: float64

Seasonality
 YearMonth
1980-01-31    0.664388
1980-02-29    0.800694
1980-03-31    0.892495
1980-04-30    0.844044
1980-05-31    0.880516
1980-06-30    0.915220
1980-07-31    1.103899
1980-08-31    1.081169
1980-09-30    1.009574
1980-10-31    1.013692
1980-11-30    1.181135
1980-12-31    1.613174
Freq: M, Name: seasonal, dtype: float64

Residual
 YearMonth
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    0.726757
1980-08-31    0.805504
1980-09-30    1.368532
1980-10-31    0.979276
1980-11-30    0.861480
1980-12-31    1.140480
Freq: M, Name: resid, dtype: float64
```

Multiplicative Decomposition - Trend, Seasonality and Residual in year 1980

b. **Multiplicative -** Plot 10



Decomposition - Multiplicative

- Trend and Seasonality is present.
- Residue/Noise ranges from 0 to 1, whereas additive noise ranges from 0 to 50.
- Trend is decreasing with respect to year.
- Rose wine sales increases as the month progresses, implying a seasonal pattern within each year.
- Peak year was 1981. Afterward sales is decreasing over the time.
- The multiplicative model is preferred over the additive model for decomposing rose wine sales because of its narrower residual range.

2. Data Pre-processing

I. Train-test split

- The data from 1980 to 1990 is used as the training set, while the data from 1991 to 1995 is used as the testing set. This separation allows us to use the earlier data for training models and the later data for testing their performance.

Table 5 - Train and Test rows and columns

First few rows of Training Data

YearMonth	Rose	Year	Month
1980-01-31	112.0	1980.0	1.0
1980-02-29	118.0	1980.0	2.0
1980-03-31	129.0	1980.0	3.0
1980-04-30	99.0	1980.0	4.0
1980-05-31	116.0	1980.0	5.0

Last few rows of Training Data

YearMonth	Rose	Year	Month
1990-08-31	70.0	1990.0	8.0
1990-09-30	83.0	1990.0	9.0
1990-10-31	65.0	1990.0	10.0
1990-11-30	110.0	1990.0	11.0
1990-12-31	132.0	1990.0	12.0

Train Dataset

```
Train:  (132, 3)
Test:  (55, 3)
```

Shape of Test(55 rows and 3 columns) and Train (132 rows and 3 columns)

First few rows of Test Data

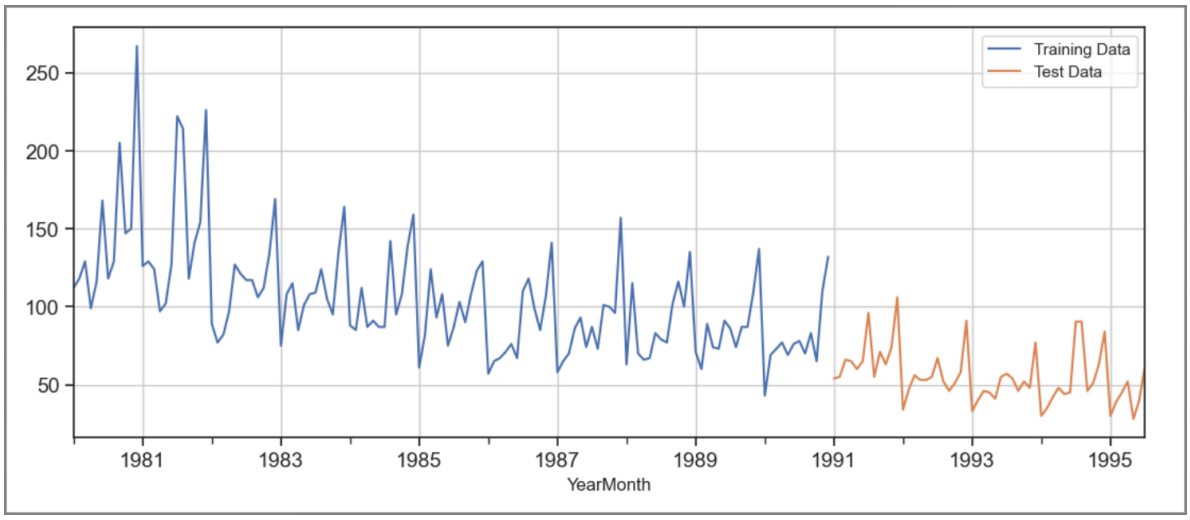
YearMonth	Rose	Year	Month
1991-01-31	54.0	1991.0	1.0
1991-02-28	55.0	1991.0	2.0
1991-03-31	66.0	1991.0	3.0
1991-04-30	65.0	1991.0	4.0
1991-05-31	60.0	1991.0	5.0

Last few rows of Test Data

YearMonth	Rose	Year	Month
1995-03-31	45.0	1995.0	3.0
1995-04-30	52.0	1995.0	4.0
1995-05-31	28.0	1995.0	5.0
1995-06-30	40.0	1995.0	6.0
1995-07-31	62.0	1995.0	7.0

Test Dataset

Plot 11 - Plot of train and test

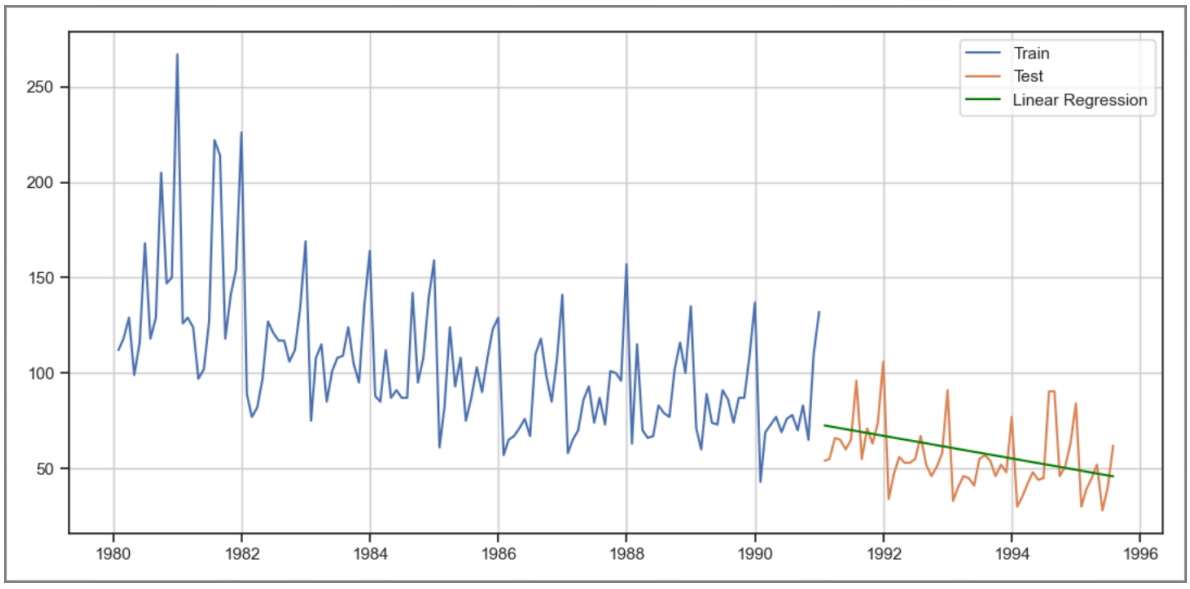


Plot of train and test

3. Model Building

(1) Linear Regression

Plot 12 - Linear Regression



Green Line indicates Linear Regression Prediction

RMSE calculated for Linear Regression: ***17.08***

```
For RegressionOnTime forecast on the Test Data,  RMSE is 17.08
```

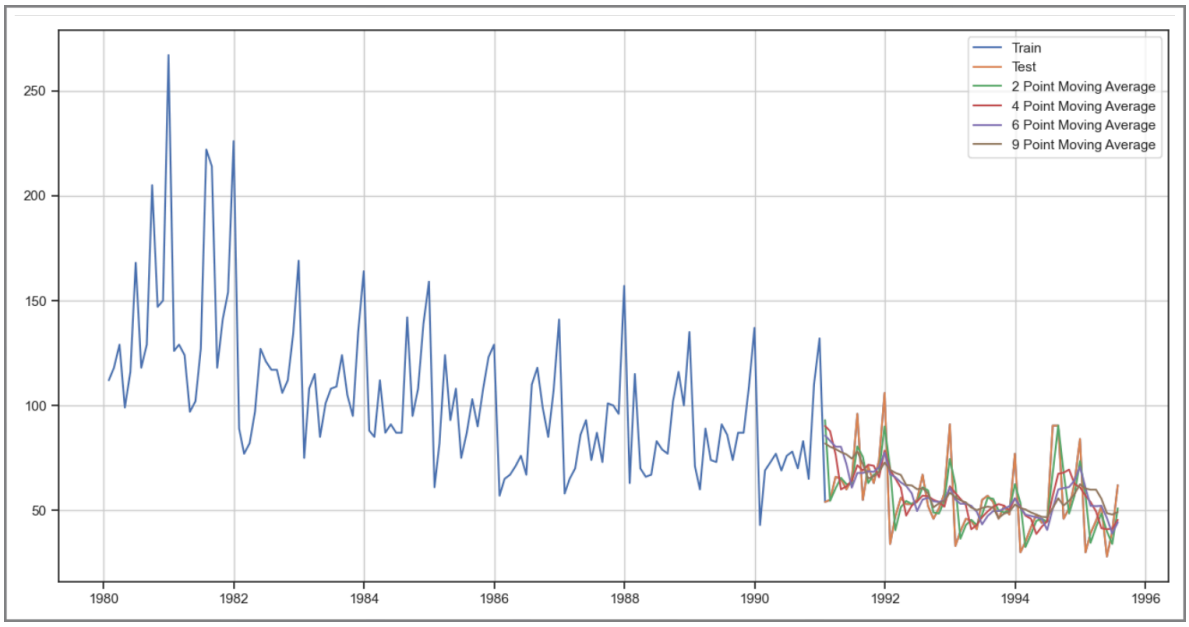
(2) **Moving Average (MA)**

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here. For Moving Average, we are going to average over the entire data.

YearMonth	Rose	Year	Month	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	1980.0	1.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	1980.0	2.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	1980.0	3.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	1980.0	4.0	114.0	114.5	NaN	NaN
1980-05-31	116.0	1980.0	5.0	107.5	115.5	NaN	NaN

Top 5 rows for Trailing Moving average for 2, 4, 6 and 9

Plot 13 -Moving Average (MA)

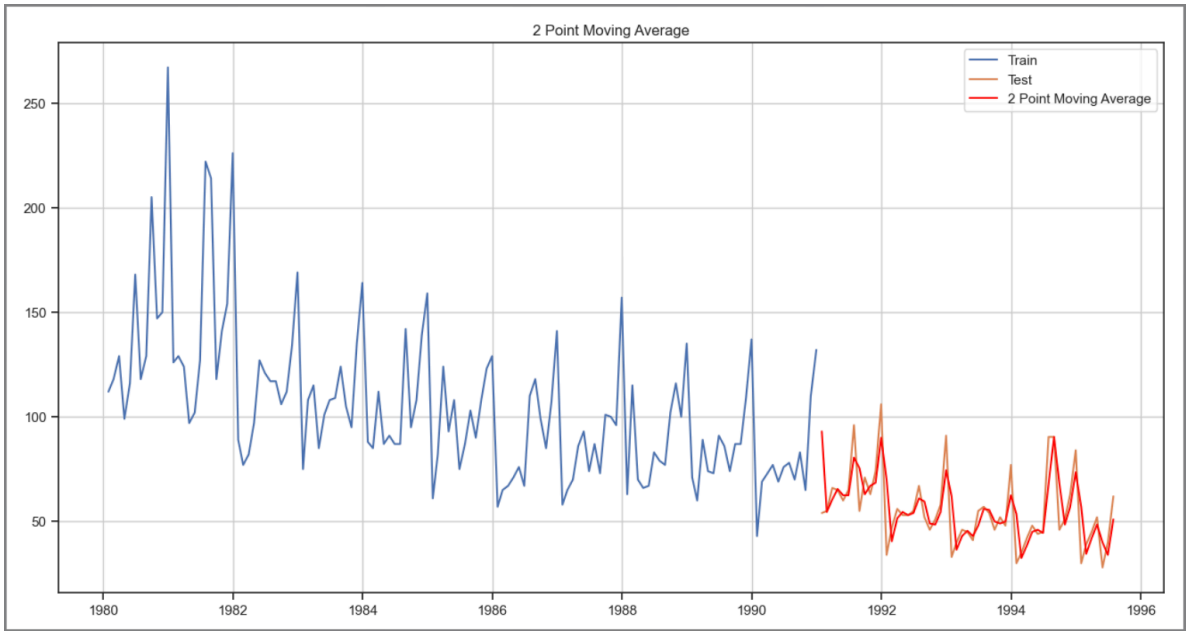


Moving Average plot - 2 point Moving Average is best

RMSE calculated for Moving Average:

```
For 2 point Moving Average Model forecast on the Testing Data,  RMSE is 12.298
For 4 point Moving Average Model forecast on the Testing Data,  RMSE is 15.846
For 6 point Moving Average Model forecast on the Testing Data,  RMSE is 15.986
For 9 point Moving Average Model forecast on the Testing Data,  RMSE is 16.501
```

We created several moving average models with rolling windows ranging from 2 to 9 points. The best model was the 2-point moving average, with a RMSE value of 12.29.



Plot for the best Moving Average that is rolling window 2.

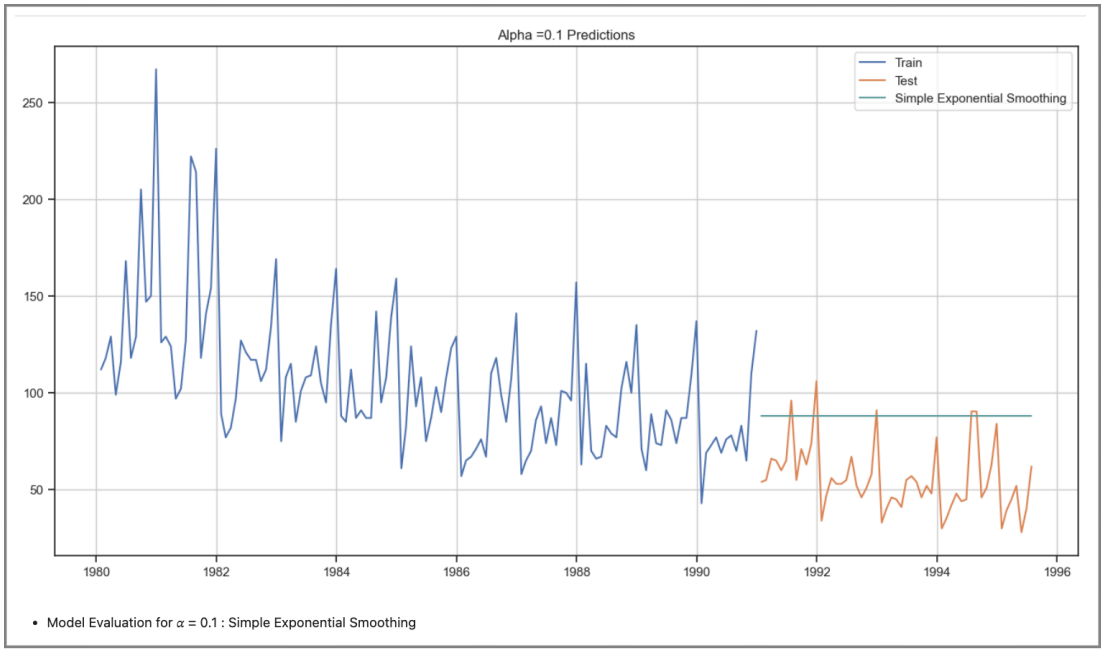
(3) **Simple Exponential Smoothening Model - Plot 14**

A Simple Exponential Smoothing (SES) model is a time series forecasting technique that applies weighted averages of past observations to make future predictions. In SES, more recent observations are given exponentially more weight compared to older observations, allowing the model to adapt quickly to changes in the data.

The SES model is particularly useful for data with no clear trend or seasonal pattern, as it effectively smooths out short-term fluctuations to reveal longer-term trends or patterns.

YearMonth	Rose	Year	Month	predict
1991-01-31	54.0	1991.0	1.0	87.983766
1991-02-28	55.0	1991.0	2.0	87.983766
1991-03-31	66.0	1991.0	3.0	87.983766
1991-04-30	65.0	1991.0	4.0	87.983766
1991-05-31	60.0	1991.0	5.0	87.983766

Table for forecast of the model



RMSE

```
For Alpha =0.1 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.712
```

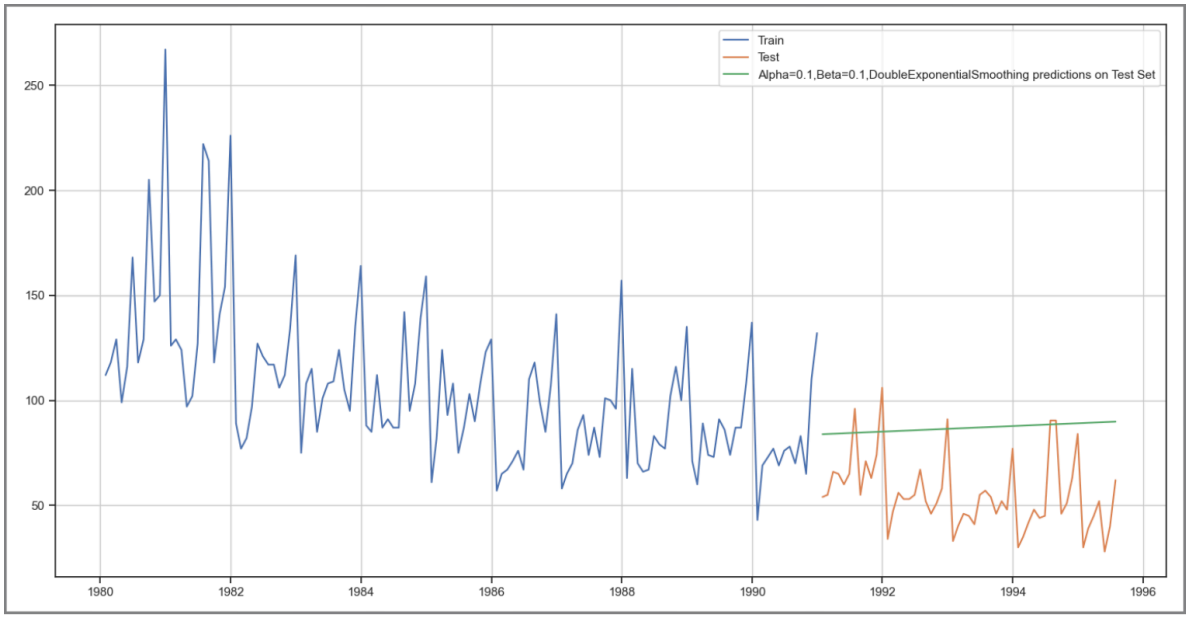
(4) **Double Exponential Smoothening (Holt's Model)**

Double Exponential Smoothing (DES), also known as Holt's Exponential Smoothing, is an extension of Simple Exponential Smoothing that incorporates both level and trend components to handle time series data with trends.

The DES method helps in capturing both the level and the trend in the time series, making it suitable for datasets where trends are present, thus providing more accurate forecasts compared to Simple Exponential Smoothing when trends exist in the data.

- Two parameters α and β are estimated in this model. Level and Trend are accounted for in this model.

Plot 15 - Holt's Model



Holt's Model

RMSE

α–0.1 and β–0.1 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 36.000

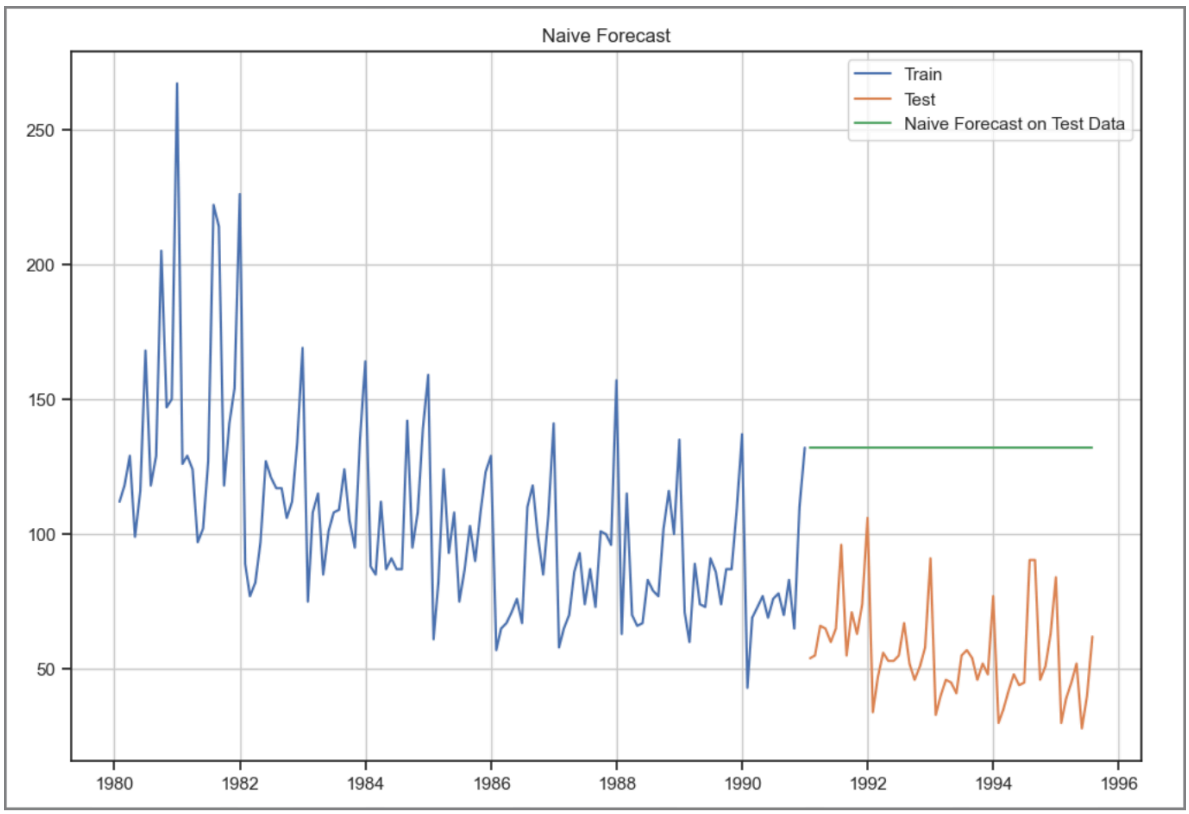
(5) **Naive Approach**

The Naive Approach is a simple and straightforward time series forecasting method where the forecast for any future period is assumed to be equal to the most recent actual observation.

```
YearMonth
1991-01-31    132.0
1991-02-28    132.0
1991-03-31    132.0
1991-04-30    132.0
1991-05-31    132.0
Freq: M, Name: naive, dtype: float64
```

forecast for test prediction is assumed 132

Plot 16 Naive Approach



Plot for Naive Approach

RMSE for Naive Approach

```
Naive Model forecast on the Test Data, RMSE is 78.396
```

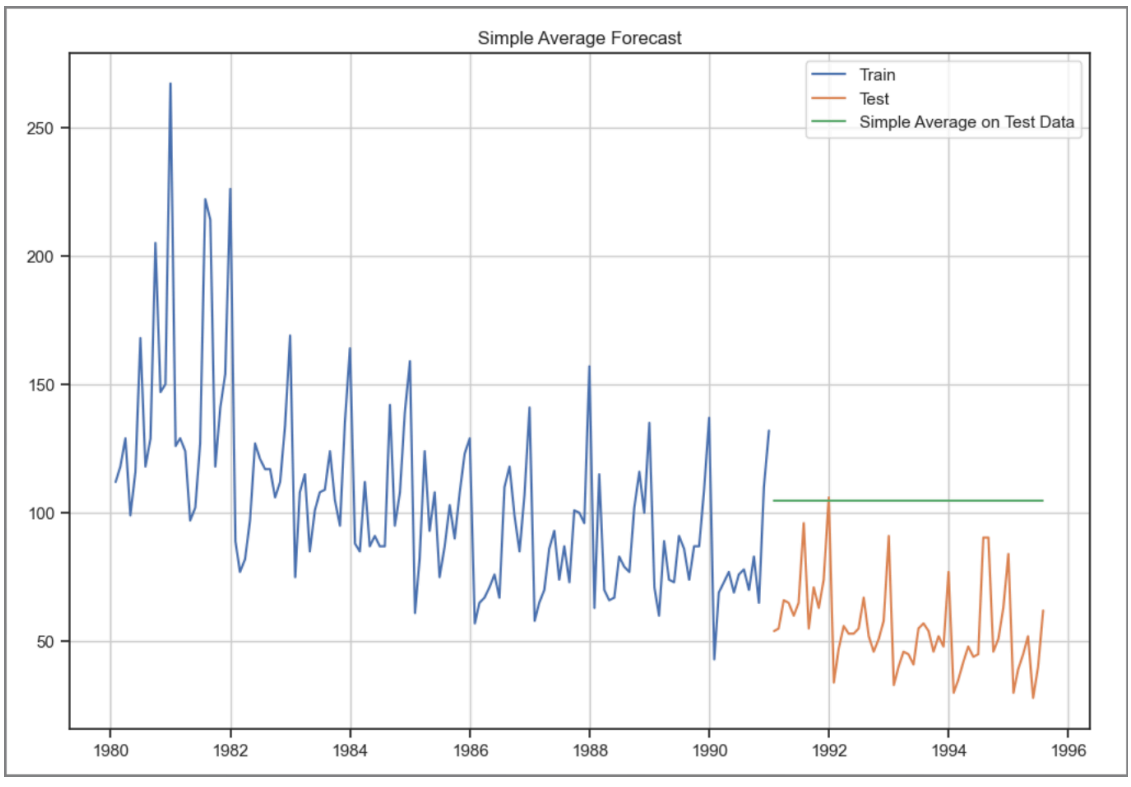
(6) **Simple Average**

The Simple Average Time Series Forecasting (TSF) model is a basic yet effective method for predicting future values in a time series. It operates on the principle that the forecasted value for a given period is the simple average (arithmetic mean) of all previous observations. This model is particularly useful for data with a consistent level over time and without significant trends or seasonal patterns.

YearMonth	Rose	Year	Month	mean_forecast
1991-01-31	54.0	1991.0	1.0	104.939394
1991-02-28	55.0	1991.0	2.0	104.939394
1991-03-31	66.0	1991.0	3.0	104.939394
1991-04-30	65.0	1991.0	4.0	104.939394
1991-05-31	60.0	1991.0	5.0	104.939394

Simple Average Test mean forecast predicted values

Plot 17 Simple Average



Simple Average Plot

RMSE for Simple Average

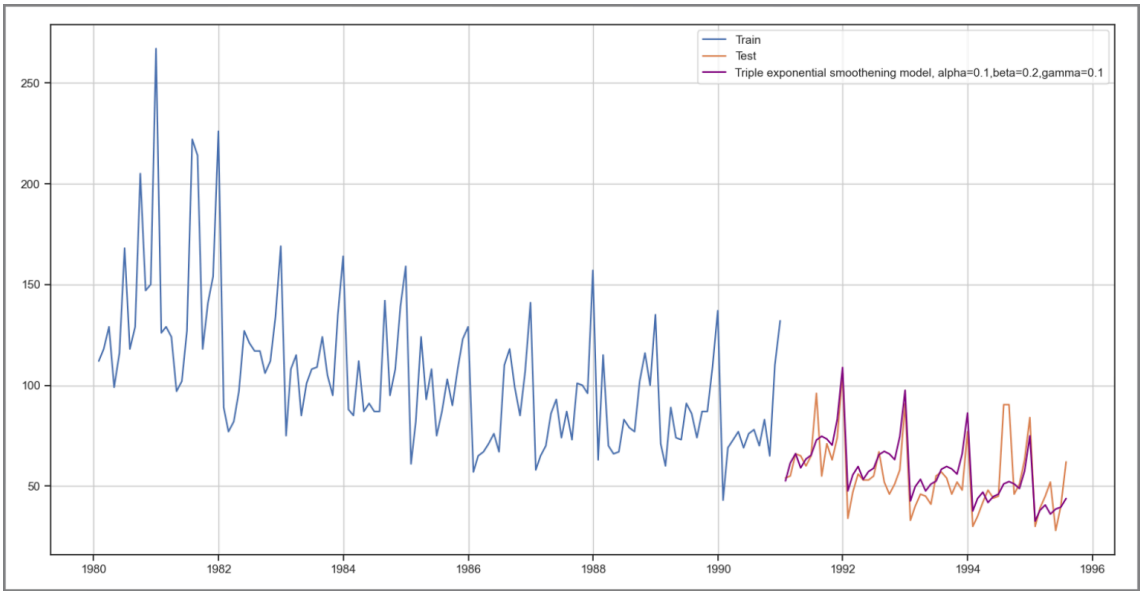
```
Simple Average forecast on the Test Data, RMSE is 52.319
```

(7) Triple Exponential Smoothing (Holt - Winter's Model)

Triple Exponential Smoothing, also known as the Holt-Winters method, is an extension of Exponential Smoothing that accounts for trends and seasonality in time series data. This method involves three components: level, trend, and seasonality. There are two main variations of the Holt-Winters method: additive and multiplicative. The additive model is used when seasonal variations are roughly constant over time, while the multiplicative model is used when seasonal variations change proportionally to the level of the time series.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE	Method
1010	0.1	0.2	0.1	19.770392	11.757610	ta_sm
1011	0.1	0.2	0.2	20.253487	12.158822	ta_sm
2136	0.2	0.7	0.2	24.042290	12.253129	tm_sm
2009	0.1	0.2	0.1	19.647823	12.476510	tm_sm
1012	0.1	0.2	0.3	20.871304	12.558595	ta_sm

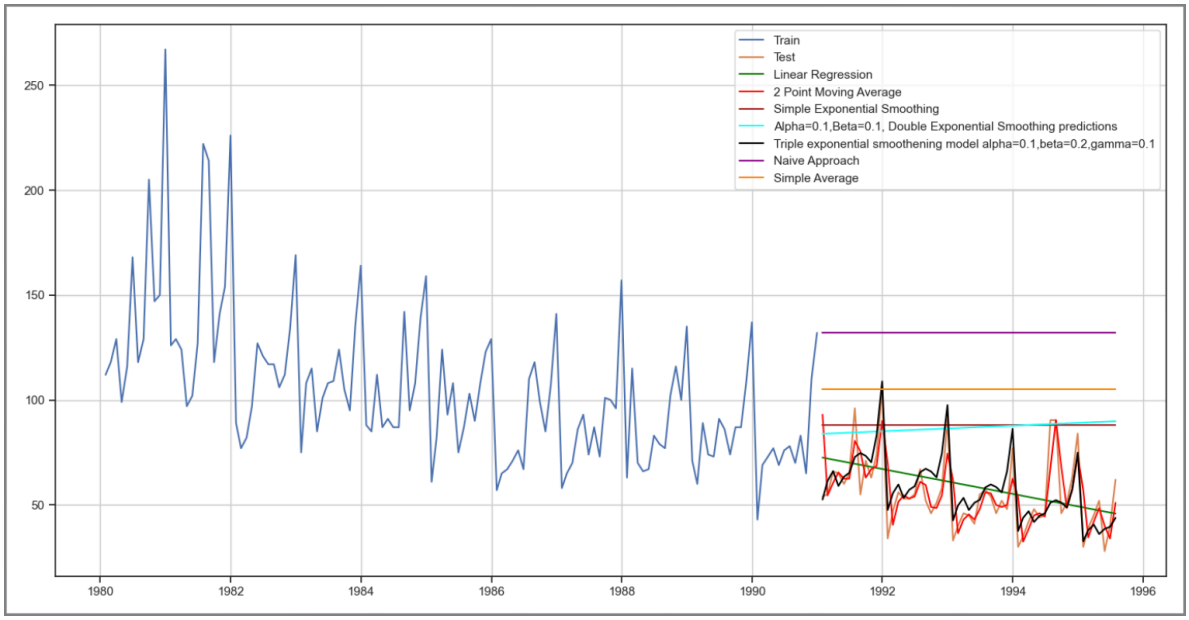
Plot 18 - Triple Exponential Smoothing



RMSE = 11.76

The optimal Triple Exponential Smoothing (Holt-Winters) model, featuring an additive trend and multiplicative seasonality, has been identified. The best smoothing parameters alpha, beta and gamma have been determined, making this the most effective model so far.

Plot 19 - Model Building for all forecast



Plot of all the forecasting models

```
Triple exponential smoothening model, alpha=0.1,beta=0.2,gamma=0.1 is 11.757610431857731
```

Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	11.757610
2pointTrailingMovingAverage	12.298291
4pointTrailingMovingAverage	15.845558
6pointTrailingMovingAverage	15.986163
9pointTrailingMovingAverage	16.500823
Linear Regression	17.080298
Alpha=0.1,Beta=0.1, Double Exponential Smoothing prediction	35.999680
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	35.999680
Alpha=0.995,SimpleExponentialSmoothing	36.711757
Simple Average	52.318735
Naive Model	78.396083

RMSE Value in sorted way for all the building

3. Check for Stationarity

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

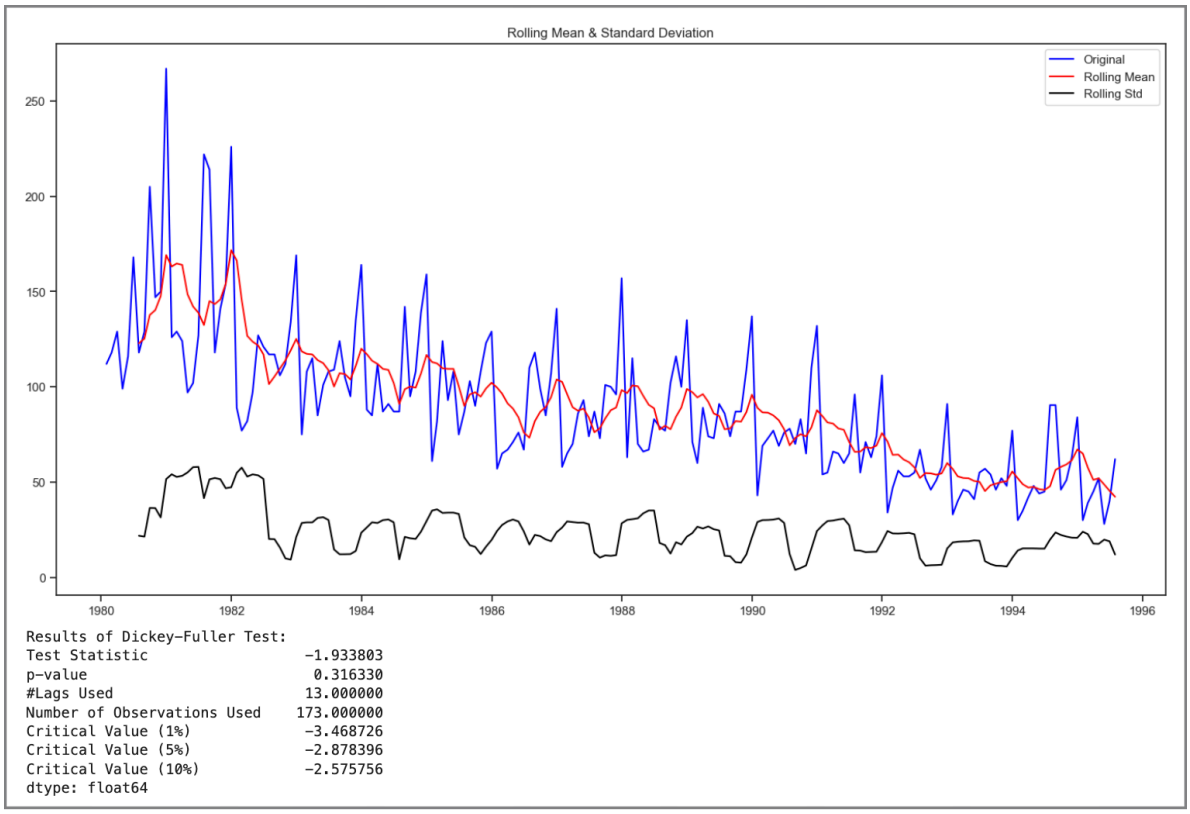
The hypothesis in a simple form for the ADF test is:

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

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

We see that at 5% significant level the Time Series is non-stationary.

Plot 20 - The Augmented Dickey-Fuller test



Dickey-Fuller Test

Here are the results of the Dickey-Fuller test presented in points:

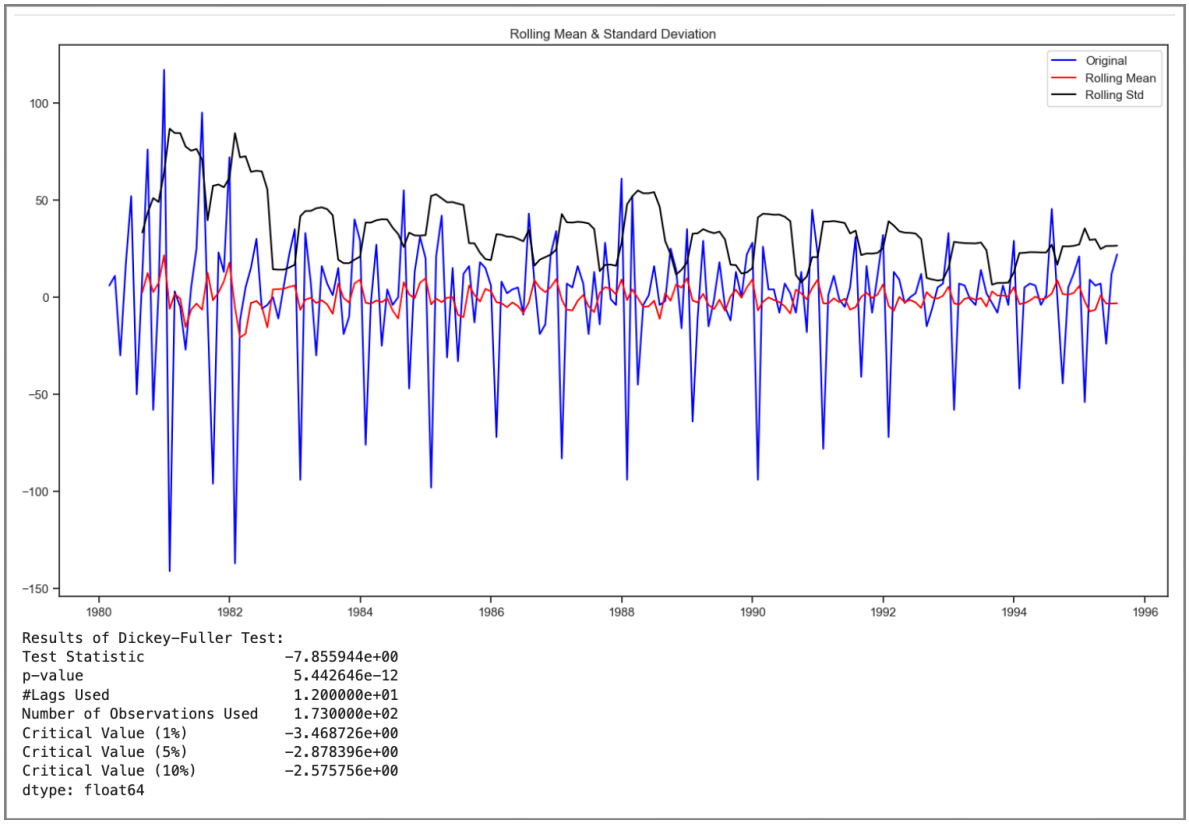
- Test Statistic: -1.933803
- p-value: 0.316330
- Lags Used 13
- Number of Observations Used: 173
- Critical Values - 1%: -3.468726, 5%: -2.878396 and 10%: -2.575756

Conclusion:

- The test statistic (-1.933803) is higher than the critical values at the 1%, 5%, and 10% significance levels.
- The p-value (0.316330) is greater than typical significance thresholds (e.g., 0.01, 0.05, 0.10).

- As a result, we fail to reject the null hypothesis that the time series has a unit root (i.e., it is non-stationary).
- This suggests that the time series is likely non-stationary, meaning its statistical properties such as mean and variance may change over time.

Plot 21 - The Augmented Dickey-Fuller test after difference



AdF test after difference

The Dickey-Fuller test, performed after differencing the data, is used to test the null hypothesis that a unit root is present in the time series sample. Here are the key points regarding the null hypothesis and the interpretation of the test results:

1. Null Hypothesis (H0): The time series has a unit root (i.e., it is non-stationary).
2. Alternative Hypothesis (H1): The time series does not have a unit root (i.e., it is stationary).

Interpretation of the Results:

Test Statistic: -7.855944

- This value is the computed test statistic for the Dickey-Fuller test. It is compared against the critical values to determine whether to reject the null hypothesis.
- The p-value is extremely small, much less than typical significance levels (e.g., 0.01, 0.05, 0.10). This indicates strong evidence against the null hypothesis.

- Critical Values: 1%: -3.468726, 5%: -2.878396 and 10%: -2.575756

These values represent the thresholds for rejecting the null hypothesis at the 1%, 5%, and 10% significance levels.

Comparison:

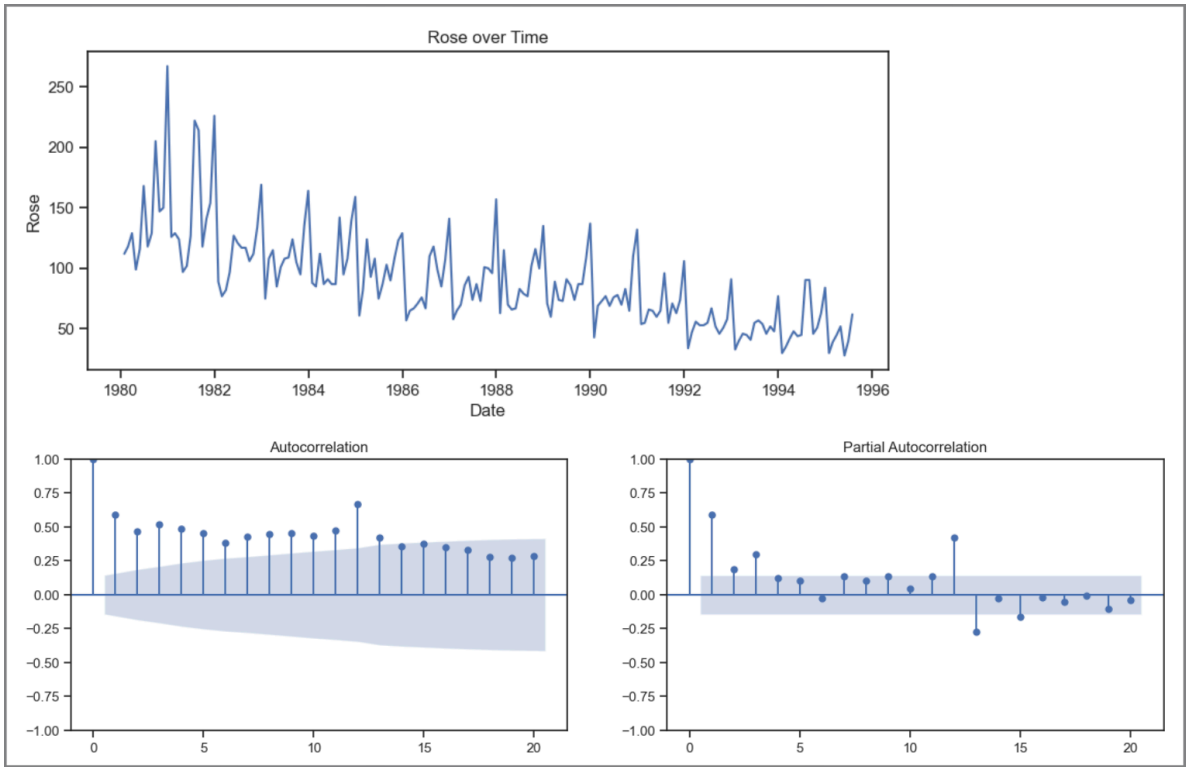
- The test statistic (-7.855944) is more negative than all the critical values at the 1%, 5%, and 10% levels.
- Since the test statistic is much lower (more negative) than the critical values, and the p-value is extremely small, we reject the null hypothesis.

Conclusion:

- Given that the test statistic (-7.855944) is much lower than the critical values and the p-value is significantly small, we reject the null hypothesis that the time series has a unit root.
- This indicates that the time series is stationary, meaning its statistical properties such as mean and variance remain constant over time.

5. Model Building - Stationary Data

Plot 22 - Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.



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

(1) Auto ARIMA (Auto-Regressive Integrated Moving Average)

The Auto ARIMA (Auto-Regressive Integrated Moving Average) model is a statistical analysis technique used for time series forecasting that automatically selects the best-fitting ARIMA model by optimizing the parameters. ARIMA models are widely used for forecasting data that show evidence of non-stationarity and require differencing to achieve stationarity. The ARIMA model comprises three components: the auto-regressive (AR) part, which regresses the variable on its own lagged values; the integrated (I) part, which involves differencing the observations to make the time series stationary; and the moving average (MA) part, which models the error term as a linear combination of past error terms. The Auto ARIMA model automates the identification of optimal values for these parameters (p, d, q) by evaluating multiple ARIMA models with different combinations and selecting the best one based on information criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). This process includes differencing the series to achieve stationarity, exploring a range of p and q values, and evaluating each model to find the one with the lowest criterion value.

For Rose wine sales analysis, the parameter d represents the differencing required to render the series stationary. The for loop iterates over p and q values ranging from 0 to 3, while a fixed value of 1 is assigned to d. This choice is made because we had previously determined through the Augmented Dickey-Fuller (ADF) test that a differencing order of 1 was necessary to achieve stationarity.

Some parameter combinations for the Model:

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

Parameters for Auto ARIMA

To obtain the parameters corresponding to the minimum AIC value, we need to sort the AIC values in ascending order and then select the parameters associated with the lowest AIC value.

	param	AIC
11	(2, 1, 3)	1274.694973
15	(3, 1, 3)	1278.668346
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.574230
9	(2, 1, 1)	1281.507862
10	(2, 1, 2)	1281.870722
7	(1, 1, 3)	1281.870722
1	(0, 1, 1)	1282.309832
13	(3, 1, 1)	1282.419278
14	(3, 1, 2)	1283.720741
12	(3, 1, 0)	1297.481092
8	(2, 1, 0)	1298.611034
4	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

Arranged the AIC values in ascending order to identify the set of parameters that yield the minimum AIC value.

p=2, d=1 and q=3 has the minimum AIC value of 1274.695.

We will generate the summary report for this.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	-631.347
Date:	Sun, 26 May 2024	AIC	1274.695
Time:	13:40:32	BIC	1291.946
Sample:	01-31-1980	HQIC	1281.705
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6781	0.084	-20.038	0.000	-1.842	-1.514
ar.L2	-0.7289	0.084	-8.704	0.000	-0.893	-0.565
ma.L1	1.0449	0.673	1.554	0.120	-0.273	2.363
ma.L2	-0.7716	0.136	-5.675	0.000	-1.038	-0.505
ma.L3	-0.9046	0.611	-1.481	0.138	-2.101	0.292
sigma2	858.3757	566.506	1.515	0.130	-251.955	1968.706

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.44
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	4.57

Auto ARIMA Summary Report for p=2, d=1 and q=3

The summary report for the Auto ARIMA model offers a detailed overview of the model's performance and diagnostics. It begins by identifying the dependent variable, labeled as "Rose," and specifies that 132 observations were utilized in the analysis. The chosen ARIMA model is denoted as ARIMA(2, 1, 3), indicating auto-regressive and moving average orders of 2 and 3, respectively, with a differencing order of 1. The log likelihood, AIC (Akaike Information Criterion) - 1274.695 , and BIC (Bayesian Information Criterion) - 1291.946 values provide measures of model fit, with lower AIC and BIC values indicating better fit. Additionally, the report includes parameter estimates for the model coefficients, standard errors, and statistical significance. Diagnostic tests such as the Ljung-Box (Q) - 0.02 and Jarque-Bera (JB) - 24.44

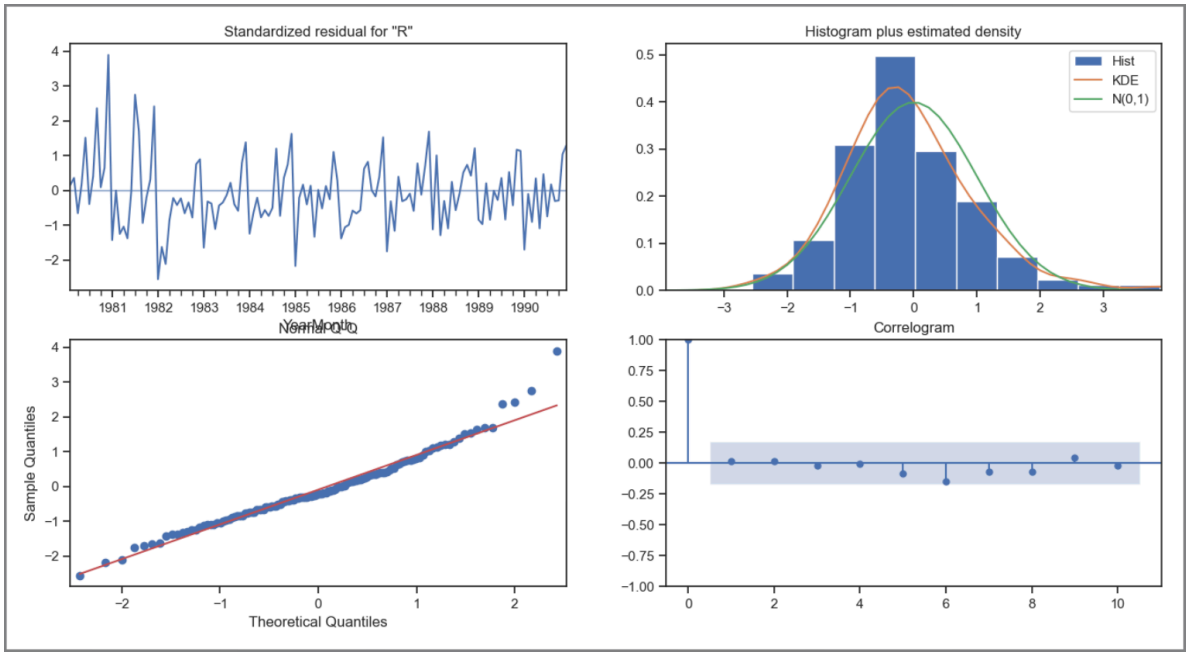
tests assess the goodness of fit, while the Heteroskedasticity (H) - 0.4 test evaluates the constancy of residual variance. Skewness and kurtosis measures provide insights into the distributional properties of the residuals. Overall, this comprehensive summary aids in the interpretation and evaluation of the Auto ARIMA model, helping to understand its effectiveness in capturing the underlying patterns in the time series data.

Forecast for Auto ARIMA model before we evaluate the RMSE

```
1991-01-31    85.615048
1991-02-28    90.533523
1991-03-31    81.973917
1991-04-30    92.752339
1991-05-31    80.904628
1991-06-30    92.929380
1991-07-31    81.386943
1991-08-31    91.990985
1991-09-30    82.610062
1991-10-31    90.622518
1991-11-30    84.014898
1991-12-31    89.262597
1992-01-31    85.272940
1992-02-29    88.142776
1992-03-31    86.235079
1992-04-30    87.344487
1992-05-31    86.873352
1992-06-30    86.855300
1992-07-31    87.229000
1992-08-31    86.615069
1992-09-30    87.372889
1992-10-31    86.548718
1992-11-30    87.379350
1992-12-31    86.586240
1993-01-31    87.311678
1993-02-28    86.672448
1993-03-31    87.216341
1993-04-30    86.769591
1993-05-31    87.122821
1993-06-30    86.855716
1993-07-31    87.046464
1993-08-31    86.921070
1993-09-30    86.992453
1993-10-31    86.964068
1993-11-30    86.959669
1993-12-31    86.987740
1994-01-31    86.943842
1994-02-28    86.997045
1994-03-31    86.939764
1994-04-30    86.997105
1994-05-31    86.942636
1994-06-30    86.992242
1994-07-31    86.948703
1994-08-31    86.985607
1994-09-30    86.955415
1994-10-31    86.979179
1994-11-30    86.961309
1994-12-31    86.973975
1995-01-31    86.965746
1995-02-28    86.970322
1995-03-31    86.968641
1995-04-30    86.968126
1995-05-31    86.970215
1995-06-30    86.967085
1995-07-31    86.970815
```

RMSE value for Auto ARIMA = 35.96

```
RSME for Auto ARIMA model for test data:  35.96414703874144
```



Diagnostic Plot for auto ARIMA

Plot 23 - Diagnostic plot for auto ARIMA for the best auto ARIMA model

(2) Auto SARIMA (Seasonal Auto-Regressive Integrated Moving Average)

SARIMA, which stands for Seasonal Auto-Regressive Integrated Moving Average, extends the ARIMA model to account for seasonal patterns in the data.

Components of SARIMA

1. **Auto-Regressive (AR) part**: Represents the correlation between the current observation and a lagged (past) observation within the same series.
2. **Integrated (I) part**: Involves differencing the raw observations to make the time series stationary. This accounts for trends present in the data.

3. **Moving Average (MA) part**: Represents the correlation between the current observation and a residual error from a moving average model applied to lagged observations.

Additionally, SARIMA includes seasonal components:

1. **Seasonal Auto-Regressive (SAR) part**: Represents the correlation between the current observation and a lagged observation within the same series, but over seasonal intervals.
2. **Seasonal Integrated (SI) part**: Involves seasonal differencing to remove seasonal trends from the data.
3. **Seasonal Moving Average (SMA) part**: Represents the correlation between the current observation and a residual error from a moving average model applied to lagged observations over seasonal intervals.

Overall, Auto SARIMA helps you forecast future values in your data easily and accurately by automatically finding the best way to do it.

For Rose wine sales analysis, the parameter d represents the differencing required to render the series stationary. The for loop iterates over p, d and q values ranging from 0 to 2. The parameter m represent number of seasonals months. We are keeping seasonal month as 12. This choice is made because we had previously determined through the Augmented Dickey-Fuller (ADF) test that a differencing order of 1 was necessary to achieve stationarity.

To obtain the parameters corresponding to the minimum AIC value, we need to sort the AIC values in ascending order and then select the parameters associated with the lowest AIC value.

p=0, d=1 and q=2 has the minimum AIC value of 716.793
Seasonal p=2, d=2 ,q=2 and m=12.

	param	seasonal	AIC
161	(0, 1, 2)	(2, 2, 2, 12)	716.792983
143	(0, 1, 2)	(0, 2, 2, 12)	718.350980
404	(1, 1, 2)	(2, 2, 2, 12)	718.768293
485	(1, 2, 2)	(2, 2, 2, 12)	719.164876
476	(1, 2, 2)	(1, 2, 2, 12)	719.481009

Top 5 rows for Auto SARIMA based on the minimum AIC value

The summary report for Auto SARIMA

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 2, 2, 12)   Log Likelihood                -351.396
Date:                            Sun, 26 May 2024   AIC                            716.793
Time:                                    14:24:27   BIC                            733.467
Sample:                                01-31-1980   HQIC                           723.478
                                     - 12-31-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9186   1379.546     -0.001      0.999   -2704.780    2702.943
ma.L2         -0.0814    112.275     -0.001      0.999    -220.137     219.974
ar.S.L12      -0.4301      0.197     -2.180      0.029      -0.817      -0.043
ar.S.L24      -0.2034      0.097     -2.099      0.036      -0.393      -0.013
ma.S.L12      -0.9798   1379.559     -0.001      0.999   -2704.866    2702.906
ma.S.L24      -0.0202     27.816     -0.001      0.999     -54.538      54.498
sigma2       274.1713      5.531     49.568      0.000     263.330     285.012
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 5.21
Prob(Q):                              0.92   Prob(JB):                         0.07
Heteroskedasticity (H):               0.67   Skew:                            -0.23
Prob(H) (two-sided):                  0.30   Kurtosis:                         4.16
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.24e+21. Standard errors may be unstable.
```

Summary Report for Auto SARIMA
p=0, d=1 and q=2 has the minimum AIC value of 716.793
Seasonal p=2, d=2 ,q=2 and m=12

The summary report provides a detailed overview of observations derived from a SARIMAX model. The analyzed data comprises 132 observations of the dependent variable labeled as "Rose." The SARIMAX model, characterized as SARIMAX(0, 1, 2)x(2, 2, 2, 12), encompasses seasonal and exogenous factors in its formulation. Evaluating the model's fit, the log likelihood is reported as

-351.396, indicating how well the model aligns with the data, while the AIC and BIC stand at 716.793 and 733.467, respectively, serving as measures of model fit and complexity. The temporal span of the dataset extends from January 31, 1980, to December 31, 1990. Covariance estimation of the model is identified as "opg." Parameter estimates offer insights into the coefficients of model terms, alongside their associated standard errors and statistical significance. The variance of residuals, denoted as Sigma2, is recorded at 274.1713. Diagnostic tests encompass the Ljung-Box (Q) test, Jarque-Bera (JB) test, and a test for heteroskedasticity (H), assessing various assumptions underlying the model. Additionally, skewness and kurtosis measures provide further characterization of the distributional properties of residuals. Overall, the summary furnishes a comprehensive assessment of the SARIMAX model's performance, encompassing its alignment with data, parameter significance, and adherence to underlying assumptions.

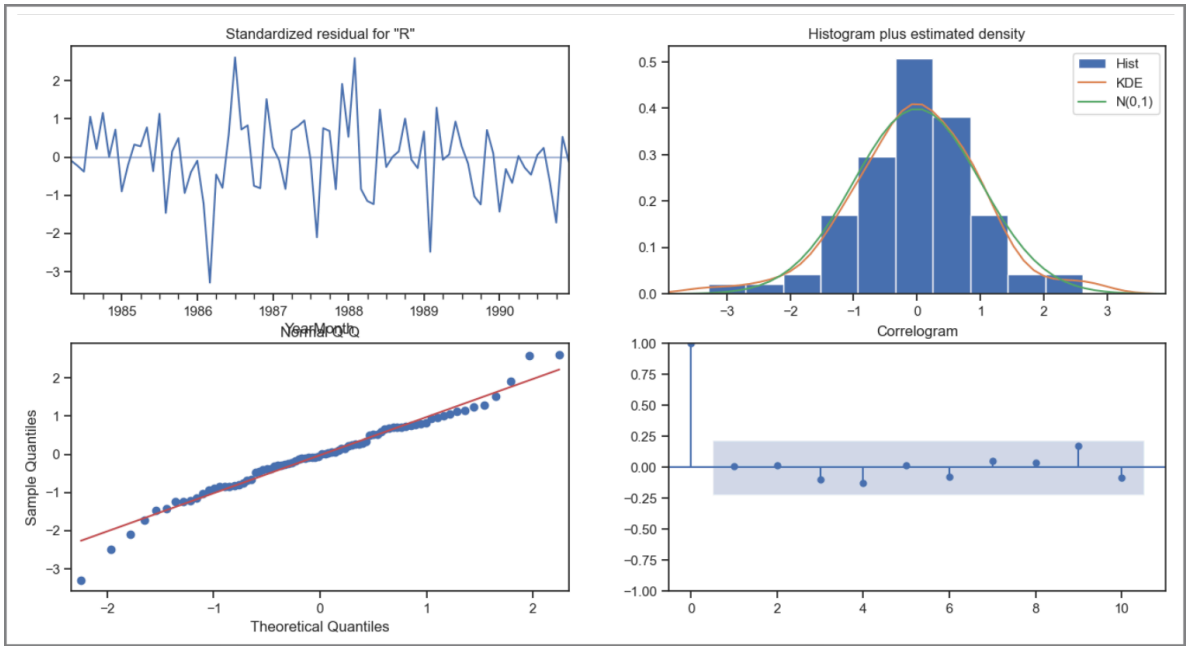
Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-31	50.314012	17.870027	15.289404	85.338621
1991-02-28	79.003497	17.970444	43.782075	114.224920
1991-03-31	78.689729	17.968049	43.473001	113.906458
1991-04-30	77.266513	17.966148	42.053510	112.479517
1991-05-31	67.468825	17.966086	32.255943	102.681706

Auto SARIMA forecast values

```
RMSE for Auto SARIMA model for test data:  34.2675171795747
```

RMSE

Plot 24 - Diagnostic plot for auto SARIMA for the best auto SARIMA model



Diagnostic Plot for Auto SARIMA

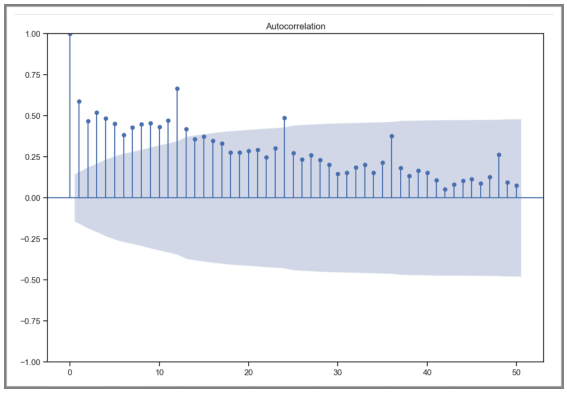
(3) Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE

- **Manual ARIMA**

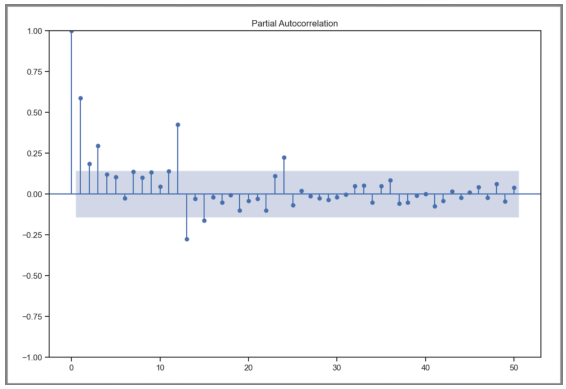
In manual ARIMA, the user manually selects appropriate values for p, d and q based on prior knowledge, domain expertise, or iterative testing.

This approach requires a deep understanding of the data and the underlying patterns to choose the most suitable parameters. Manual ARIMA is often used when automated methods like Auto ARIMA or Auto SARIMA are not available or when users prefer a more hands-on approach to model selection. However, it can be time-consuming and may not always yield the best results compared to automated approaches.

Plot 25 - ACF Plot on train data

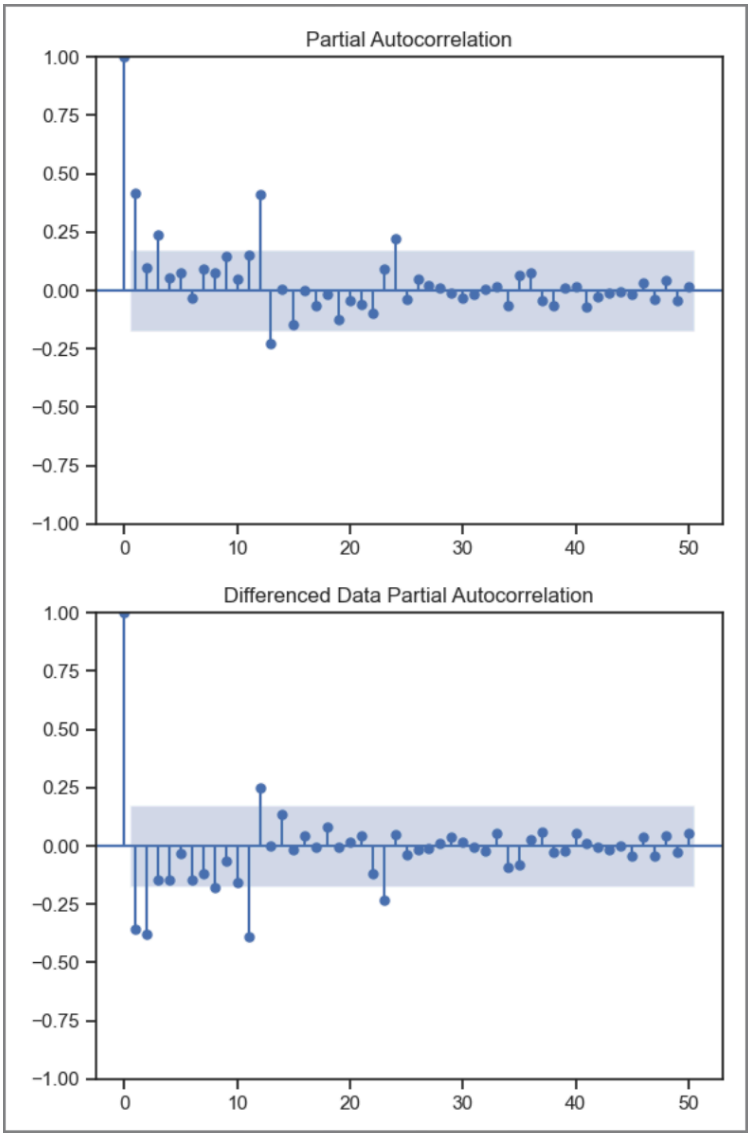


ACF Plot on train data



ACF Plot on Partial Auto-correlation

Plot 26 - PACF Plot on train data



PACF Plot on train data

Value selected for manual ARIMA: p=1, q=1 and d=1

```
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -637.287
Date:                Sun, 26 May 2024   AIC                           1280.574
Time:                        16:00:28   BIC                           1289.200
Sample:                    01-31-1980   HQIC                          1284.079
                         - 12-31-1990
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1814      0.076      2.396      0.017       0.033       0.330
ma.L1         -0.9192      0.053    -17.362      0.000      -1.023      -0.815
sigma2       972.5964     88.768     10.957      0.000     798.614    1146.579
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                41.59
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               0.35   Skew:                             0.87
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.14
===================================================================================
```

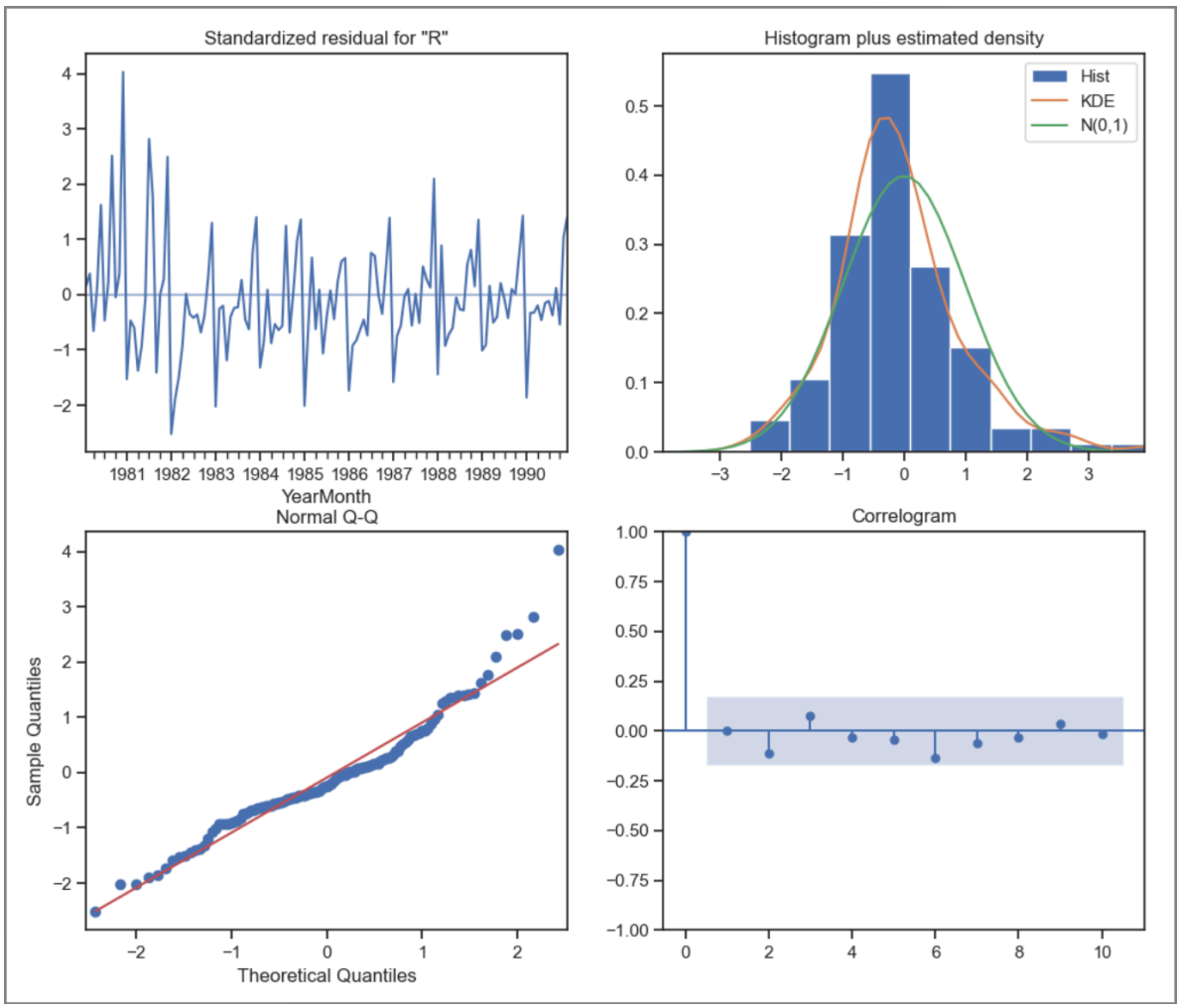
Manual ARIMA p=1, q=1 and d=1

The manual ARIMA model, denoted as ARIMA(1, 1, 1), was applied to analyze the "Rose" dataset comprising 132 observations. The model suggests a first-order auto-regressive component p=1 and a first-order moving average component q=1, along with first-order differencing d=1. The log likelihood of the model is reported as -637.287, with corresponding AIC and BIC values of 1280.574 and 1289.200, respectively. The HQIC value stands at 1284.079. Parameter estimates indicate a coefficient of 0.1814 for the auto-regressive term and -0.9192 for the moving average term. The variance of residuals (sigma2) is calculated as 972.5964. Diagnostic tests include the Ljung-Box test for autocorrelation, with a p-value of 0.98, indicating no significant autocorrelation, and the Jarque-Bera test for normality, yielding a p-value of 0.00. Additionally, the model's heteroskedasticity test returns a p-value of 0.00, suggesting heteroskedasticity is present. Overall, the manual ARIMA model provides insights into the relationships between the variables and their predictive capabilities within the dataset.

RSME value for Manual ARIMA

```
RMSE for manual ARIMA model: 36.57508705767662
```

Plot 27 - Manual ARIMA diagnostic



Manual ARIMA - p-1, q-1 and q-1

- **Manual SARIMA**

The manual SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model is a technique for time series forecasting where the user manually selects the values of the SARIMA parameters to capture both the seasonal and non-seasonal patterns present in the data.

Once the parameters are selected, the manual SARIMA model is fitted to the data, and forecasts can be generated for future time points. Diagnostic tests and evaluation metrics are then used to assess the model's performance and determine if adjustments to the parameter values are necessary.

Manual SARIMA offers flexibility and control over the modeling process, but it requires expertise in time series analysis and a deep understanding of the data to select appropriate parameter values that result in accurate forecasts.

Value selected for manual SARIMA: p=1, q=1 and d=1

Seasonal p=1, q=1 m= 12 and d=1

The summary report for manual SARIMA

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood                -458.646
Date:                            Sun, 26 May 2024   AIC                            927.292
Time:                                    16:14:29   BIC                            940.562
Sample:                                         0   HQIC                           932.669
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2139      0.117      1.829      0.067      -0.015       0.443
ma.L1         -0.9289      0.055    -16.951      0.000      -1.036      -0.821
ar.S.L12      -0.4113      0.065     -6.323      0.000      -0.539      -0.284
ma.S.L12       0.0061      0.134      0.046      0.964      -0.257       0.270
sigma2       361.3677     50.392      7.171      0.000     262.602     460.134
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 0.25
Prob(Q):                              0.96   Prob(JB):                         0.88
Heteroskedasticity (H):               0.59   Skew:                             0.11
Prob(H) (two-sided):                  0.12   Kurtosis:                         3.07
===================================================================================

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

The manual SARIMA model, specified as SARIMAX(1, 1, 1)x(1, 1, 1, 12), was applied to the dataset "y," consisting of 132 observations. This model configuration includes an auto-regressive order of 1 (p=1), a differencing order of 1 (d=1), and a moving average order of 1 (q=1) for the non-seasonal components. For the seasonal components, there is an auto-regressive order of 1 (P=1), a differencing order of 1 (D=1), a moving average order of 1 (Q=1), and a seasonal period of 12 months.

The log likelihood of the model is reported as -458.646, with corresponding AIC and BIC values of 927.292 and 940.562, respectively. The HQIC value stands at 932.669. Parameter estimates indicate a coefficient of 0.2139 for the auto-regressive term and -0.9289 for the moving average term. The seasonal auto-regressive term has a coefficient of -0.4113, and the seasonal moving average term has a coefficient of 0.0061.

The variance of the residuals (sigma2) is calculated as 361.3677. Diagnostic tests include the Ljung-Box test for autocorrelation, with a p-value of 0.96 indicating no significant autocorrelation, and the Jarque-Bera test for normality, yielding a p-value of 0.88. Additionally, the model's heteroskedasticity test returns a p-value of 0.12, suggesting the presence of heteroskedasticity. Overall, the manual SARIMA model provides insights into the relationships between the variables and their predictive capabilities within the dataset.

RSME for Manual SARIMA

```
RMSE for manual SARIMA model: 18.23298592199148
```

6. Compare the performance of the models

- **Compare the performance of the models**

We can see that Alpha = 0.1, Beta = 0.2 and Gamma = 0.1 Triple Exponential Smoothing has the lowest RSME value. So this will be considered as the best mode

Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	11.757610
2pointTrailingMovingAverage	12.298291
4pointTrailingMovingAverage	15.845558
6pointTrailingMovingAverage	15.986163
9pointTrailingMovingAverage	16.500823
Linear Regression	17.080298
order=(1, 1, 1),seasonal_order=(1, 1, 1, 12),Manual SARIMA	18.232986
order=(0,1,2),seasonal_order=(2, 2, 2, 12),Auto_SARIMA	34.267517
Auto ARIMA	35.964147
Alpha=0.1,Beta=0.1, Double Exponential Smoothing prediction	35.999680
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	35.999680
order=(1, 1, 1) - Manual ARIMA	36.575087
Alpha=0.995,SimpleExponentialSmoothing	36.711757
Simple Average	52.318735
Naive Model	78.396083

Sorted by RMSE values on the Test Data

- **Rebuild the best model using the entire data - Make a forecast for the next 12 months**

After comparing all the models we constructed, it's evident that the triple exponential smoothing or Holt-Winters model yields the lowest RMSE. Therefore, it emerges as the most optimal choice. We will rebuild the best model using triple exponential smoothing for the next 12 months prediction. Forecasts and confidence intervals into a DataFrame.

```
DatetimeIndex(['1995-08-01', '1995-09-01', '1995-10-01', '1995-11-01',
               '1995-12-01', '1996-01-01', '1996-02-01', '1996-03-01',
               '1996-04-01', '1996-05-01', '1996-06-01', '1996-07-01'],
              dtype='datetime64[ns]', freq='MS')
```

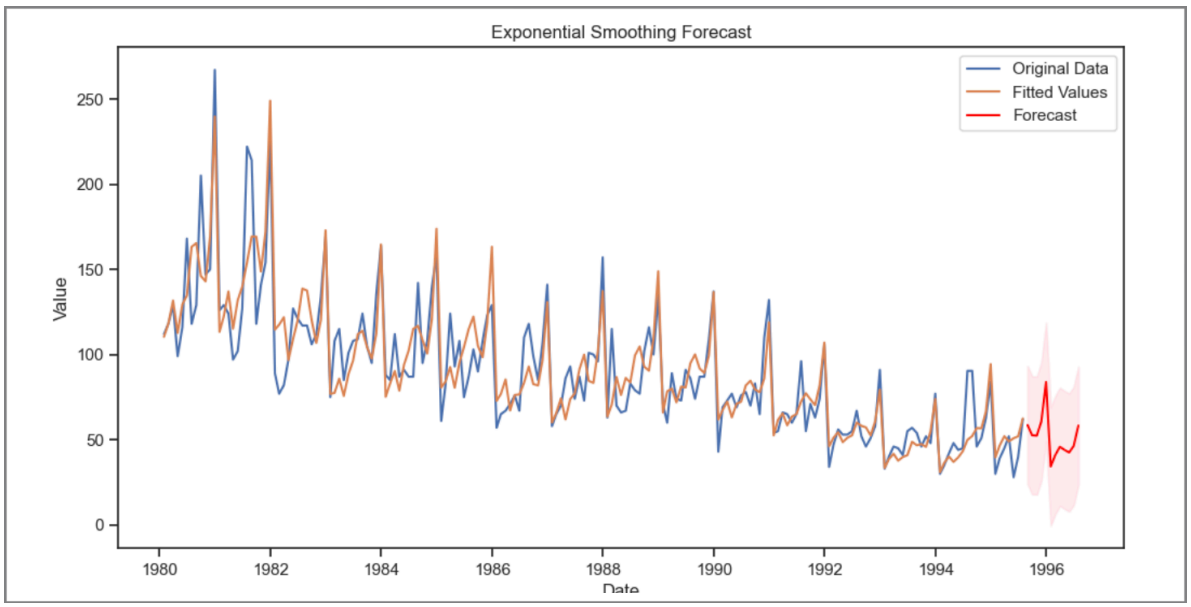
Next 12 months dates that we will be using for prediction

	Forecast	Lower Bound	Upper Bound
1995-08-31	58.607416	23.995426	93.219406
1995-09-30	52.609129	17.997138	87.221119
1995-10-31	52.404658	17.792667	87.016648
1995-11-30	60.737040	26.125050	95.349031
1995-12-31	83.883256	49.271266	118.495247
1996-01-31	34.281140	-0.330850	68.893131
1996-02-29	40.862360	6.250370	75.474351
1996-03-31	45.851970	11.239980	80.463961
1996-04-30	44.043241	9.431250	78.655231
1996-05-31	42.522405	7.910414	77.134395
1996-06-30	46.305973	11.693983	80.917964
1996-07-31	58.274895	23.662905	92.886885

Forecasted value for next 12 months

```
1995-08-31    58.607416
1995-09-30    52.609129
1995-10-31    52.404658
1995-11-30    60.737040
1995-12-31    83.883256
1996-01-31    34.281140
1996-02-29    40.862360
1996-03-31    45.851970
1996-04-30    44.043241
1996-05-31    42.522405
1996-06-30    46.305973
1996-07-31    58.274895
Freq: M, dtype: float64
```

Rose Wine Sales predicted values made with Triple Exponential Smoothing



Future predicted plot

Plot 29 - Prediction for future 12 months

7. Actionable Insights & Recommendations

Actionable Insights

1. **Trend and Seasonality Observations:**

- Long-term Trend: There is a strong negative correlation between Rose Wine sales and Year, indicating a consistent decrease in sales over time. This suggests a long-term downward trend in Rose wine sales.
- Seasonal Pattern: A moderate positive correlation between Rose Wine sales and Month suggests some seasonality, with sales tending to increase as the month progresses. However, this seasonal pattern is less pronounced than the long-term downward trend.

2. **Monthly Sales Insights:**

- December: Sales are consistently highest across all years.
- January: Sales tend to be the lowest.
- January to June: Sales remain relatively stable.

- July Onwards: Sales begin to increase.
- The highest wine sale was recorded in the year 1981.

3. **Seasonal Influence:**

Wine sales are significantly influenced by seasonal changes, with an increase during the festival season and a drop during peak winter (January).

Recommendations

- Focus on marketing campaigns from April to June, when sales are low, to boost overall annual performance.
- Consider running campaigns throughout the year to encourage wine consumption during typically low-sales periods.
- Running campaigns during peak periods might not significantly impact sales, as they are already high.
- Due to low purchase tendencies during peak winter (January), campaigns may not be effective.
- Explore reasons behind the decline in Rose wine popularity and adjust production and marketing strategies as needed to regain market share.
- If necessary, modify the production and marketing strategies to address the long-term downward trend in sales and improve market performance.