TIME SERIES FORECASTING

PROBLEM STATEMENT – CODED PROJECT

By: BENITA MERLIN.E

PGP-Data Science and Business Analytics.

BATCH: PGP DSBA. O. MAY24.A

Contents

LIST OF FIGURES

A. DEFINE THE PROBLEM AND PERFORM EDA

1. INTRODUCTION

1.1 PROJECT OVERVIEW

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.

1.2 OBJECTIVE

The primary objective of this project is to analyze 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.

1.3 PROBLEM DEFINITION

- Analyze and forecast wine sales trends using 20th-century historical data.
- Identify seasonal patterns, trends, and anomalies in sales.
- Decompose the data to isolate trend, seasonality, and residual components for deeper insights.
- Identify factors influencing wine sales, including external events and market dynamics.
- Address data variability across wine varieties and time periods.
- Develop accurate sales forecasting models for strategic planning.
- Provide actionable insights to optimize future sales and maintain market competitiveness.

2. DATA DESCRIPTION

The data contains the different factors to analyze for the content. The detailed data dictionary is given below.

Dataset: Rose.csv

Dataset: Sparkling.csv

2.1 DATA DICTIONARY

SI.NO	VARIABLE	DESCRIPTION
1	yearmonth	Year and month of the respective wine sale. This appears to represent dates in a string format for both the datasets (e.g., "1980-01").
2	Rose/Sparkling	Distinct wine varieties

Figure1 DATA DICTIONARY

2.2 SAMPLE DATASET

Read the data as an appropriate time series data:

Head and tail dataset of rose wine:

	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

	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

(187, 2)

Figure 2 HEAD AND TAIL DATASET OF ROSE WINE SALES

Head and tail dataset of sparkling wine:

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471

	YearMonth	Sparkling
182	1995-03	1897
183	1995-04	1862
184	1995-05	1670
185	1995-06	1688
186	1995-07	2031

(187, 2)

Figure 3 HEAD AND TAIL DATASET OF SPARKLING WINE SALES

2.3 DATA INFORMATION

```
<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
<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   Sparkling   187 non-null    int64
dtypes: int64(1), object(1)
memory usage: 3.1+ KB
```

Figure 4 DATA INFORMATION

Observation:

- Both the dataset has two columns:
- YearMonth: This appears to represent dates in a string format for both the datasets (e.g., "1980-01").
- Rose: This contains sales data, but there are two missing values.
- Sparkling: This also contains sales data. There is no missing values.

2.4 FEATURE ENGINEERING

- Converting 'YearMonth' to datetime format for proper alignment
- Setting 'YearMonth' as the index for both datasets for alignment

```
Rose Data:
                Rose
YearMonth
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

Sparkling Data:
            Sparkling
YearMonth
1980-01-01       1686
1980-02-01       1591
1980-03-01       2304
1980-04-01       1712
1980-05-01       1471
```

Figure 5 DATETIME FORMAT

2.5 PLOT THE DATA

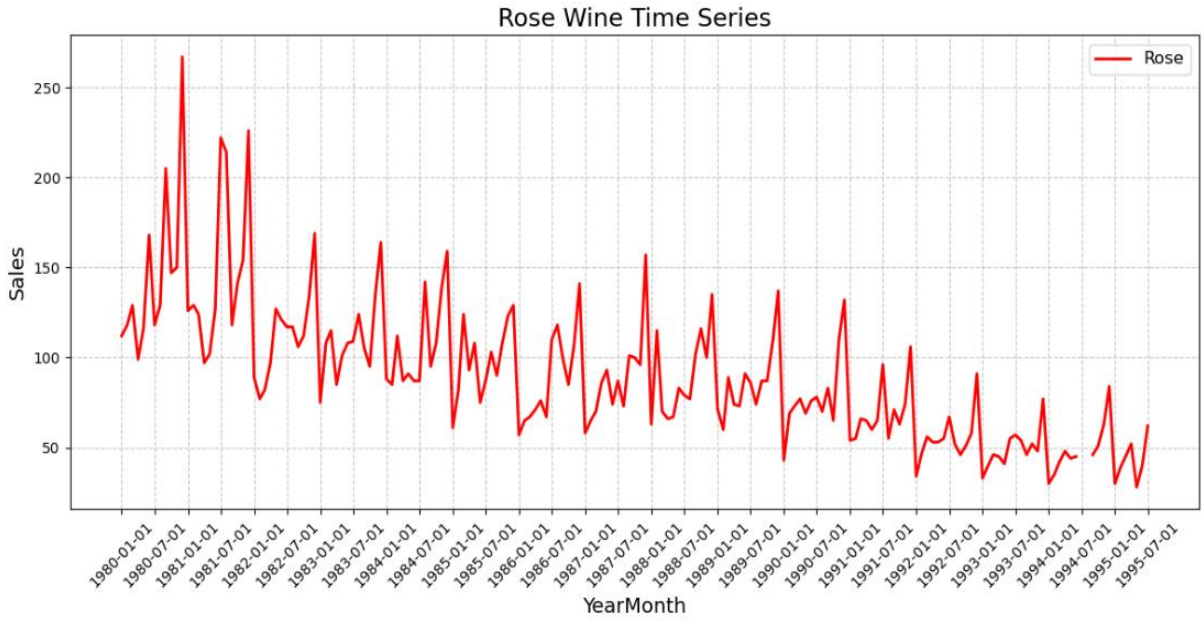


Figure 6 ROSE WINE TIME SERIES

Observation:

- Sales Trends: There is an overall declining trend in the sales of Rosé wine from 1980 to 1995.
- Seasonality: The sales exhibit a recurring pattern, suggesting the presence of seasonality. Peaks and troughs seem to repeat at regular intervals, possibly due to seasonal demand.
- High Variability in Early Years: Sales are more volatile with larger fluctuations between 1980 and 1985 compared to later years.
- Decreasing Peaks: The magnitude of the highest sales values diminishes over time, indicating reduced peak demand.
- Stabilization in Later Years: The sales variability reduces after 1990, with the data appearing more stable but with lower overall sales levels.
- Potential Anomalies: Some outliers might exist during early years with unusually high peaks, possibly indicating special events or external factors affecting sales.

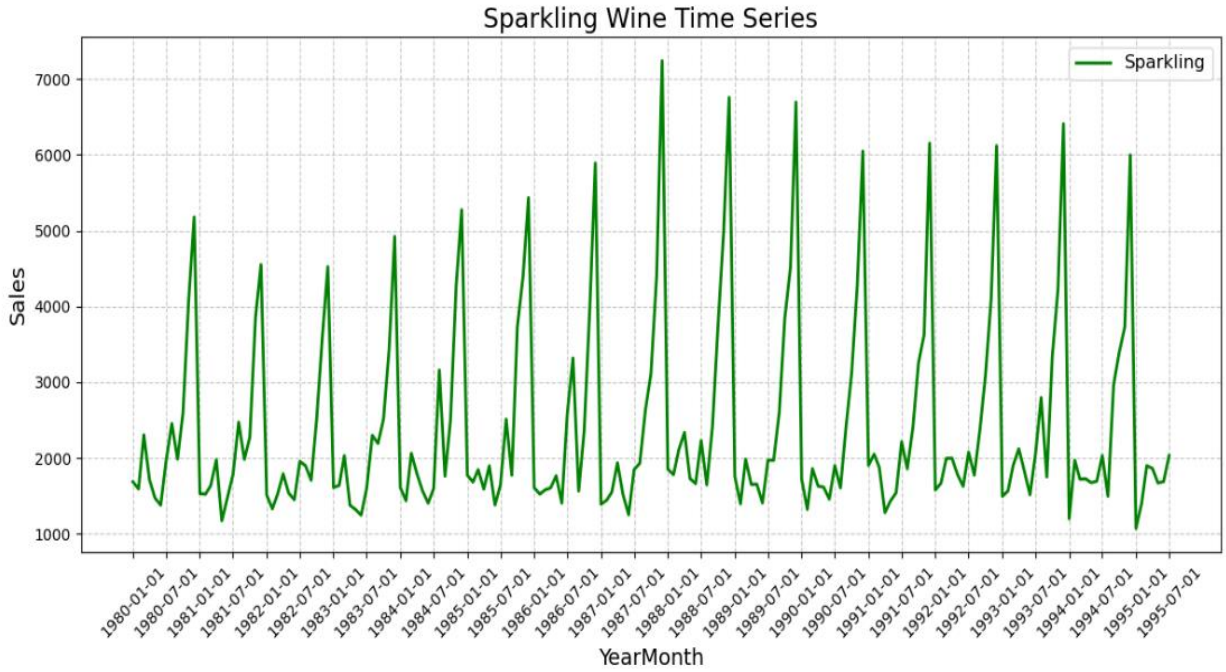


Figure 7 SPARKLING WINE TIME SERIES

Observation:

- From the time series plot for sparkling wine sales, here are the observations:

Strong Seasonality:

- The sales exhibit a clear seasonal pattern with prominent peaks recurring annually.
- Likely influenced by holidays, festivals, or specific events when sparkling wine is in higher demand.

Stable Trends:

- The overall sales trend appears stable, with consistent peak magnitudes over time.
- No significant long-term increase or decrease in sales is noticeable.

Peak Magnitudes:

- Peaks are significantly higher than off-peak periods, suggesting a highly seasonal product with concentrated demand.

Low Sales Between Peaks:

- The sales during off-peak periods remain relatively low, indicating minimal demand outside peak seasons.

Regular Patterns:

- The time intervals between peaks are uniform, reflecting a predictable annual demand cycle.

3. EXPLORATORY DATA ANALYSIS

3.1 STATISTICAL ANALYSIS

The statistical analysis provides a summary of the key metrics for the numerical columns in the dataset. This includes measures such as the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile, Q2), 75th percentile (Q3), and maximum values for each column

```
Summary statistics for Rose 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

Summary statistics for Sparkling dataset:
          Sparkling
count   187.000000
mean   2402.417112
std    1295.111540
min    1070.000000
25%    1605.000000
50%    1874.000000
75%    2549.000000
max    7242.000000
```

Figure 8 STATISTICAL SUMMARY OF ROSE AND SPARKLING

Observation:

Rose Dataset:

- Count: The dataset contains 185 observations, indicating 185 time periods.
- Mean Sales: The average sales for Rosé wine are approximately 90.39 units.
- Sales Variability:
- The standard deviation is 39.18, suggesting moderate variability in sales.
- Sales range from a minimum of 28 to a maximum of 267, indicating significant fluctuations.

Quartiles:

- 25% of sales are below 63.

- The median (50th percentile) is 86, showing the central tendency of sales.
- 75% of sales are below 112, with the remaining 25% contributing to higher sales figures.

Sparkling Dataset:

- Count: The dataset contains 187 observations, slightly more than the Rosé dataset.
- Mean Sales: The average sales for Sparkling wine are 2402.42 units, significantly higher than Rosé wine sales.
- Sales Variability:
- The standard deviation is 1295.11, indicating high variability in sales.
- Sales range from 1070 to 7242, showcasing much larger fluctuations than Rosé.

Quartiles:

- 25% of sales are below 1605.
- The median is 1874, meaning half the sales are below this value.
- 75% of sales are below 2549, with the remaining 25% showing higher sales, likely during seasonal peaks.

Comparison and Insights:

- Sparkling wine has significantly higher average sales and variability compared to Rosé wine.
- Rosé sales are more stable and less volatile, whereas Sparkling sales exhibit extreme peaks during high-demand periods.
- The higher maximum value and broader range in Sparkling wine sales highlight its strong seasonality and demand surges.
- Sparkling wine's demand is likely tied to specific events or seasons, while Rosé wine seems to have more consistent year-round demand.

3.2 CHECKING FOR MISSING VALUES:

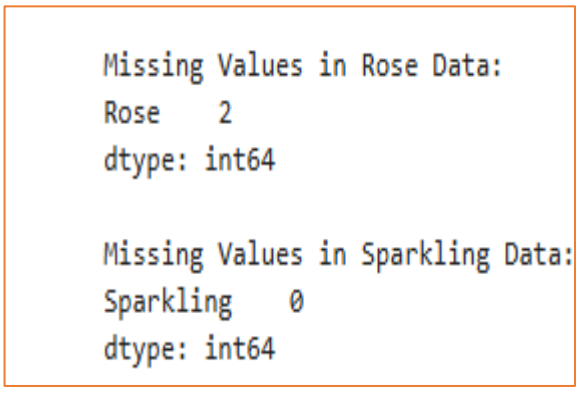


Figure 9 MISSING DATA

Observation:

- The dataset reveals two missing values in the Rose Sales column, while the Sparkling Sales column remains complete, ensuring a reliable basis for further analysis.

3.3 DISTRIBUTION OF PLOT

3.3.1 VISUALIZATION OF ROSE SALES DISTRIBUTION

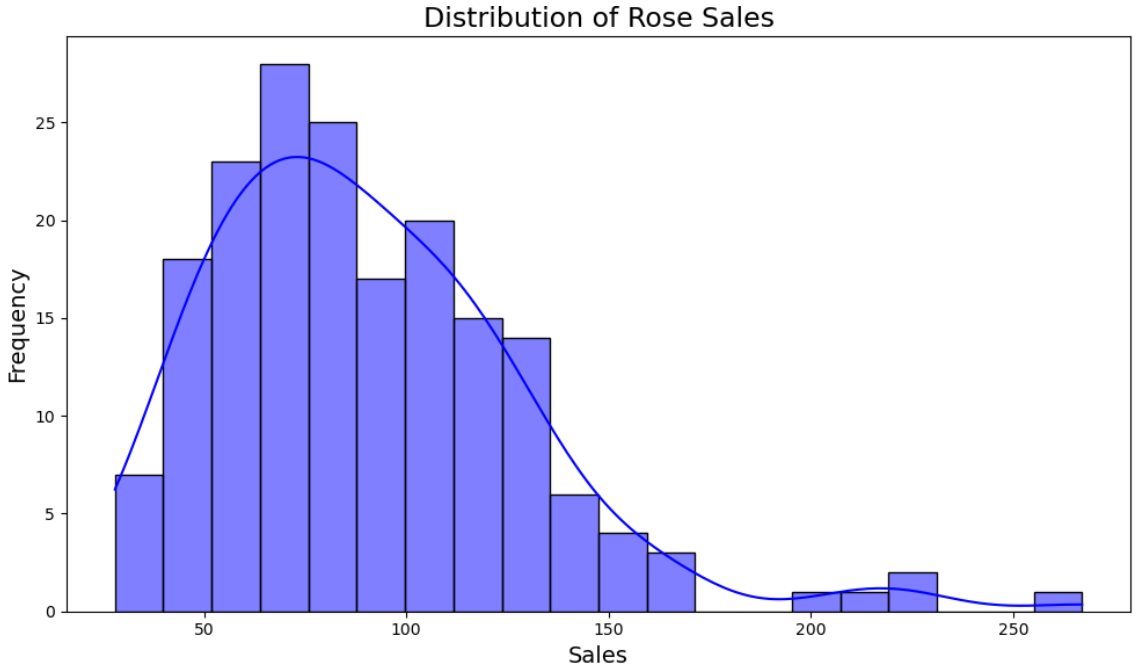


Figure 10 DISTRIBUTION PLOT OF ROSE SALES

Observations from the Distribution of Rosé Sales:

1. Right-Skewed Distribution:
    - The histogram shows a positively skewed distribution, with most sales concentrated on the lower end (50–100 units) and a long tail extending towards higher sales values.
2. Peak Frequency:
    - The highest frequency of sales lies between 60 and 100 units, indicating that these are the most common sales figures for Rosé wine.
3. Lower Frequency for High Sales:
    - Sales above 150 units are relatively rare, suggesting that very high sales are exceptional events.
4. Mean and Median Relationship:
    - Given the right-skewed nature, the mean sales (90.39 units) are likely higher than the median (86 units), influenced by a few high sales values.
5. Sales Range:

- Sales range from 28 to 267 units, showing a wide spread, but the majority of sales fall within a narrower range.

Insights:

- Most sales are concentrated around a specific range, reflecting a steady demand for Rosé wine.
- High sales events are infrequent, possibly linked to special occasions or seasonal promotions.
- This distribution suggests stable baseline demand with occasional demand spikes.

3.3.2 VISUALIZATION OF SPARKLING SALES DISTRIBUTION

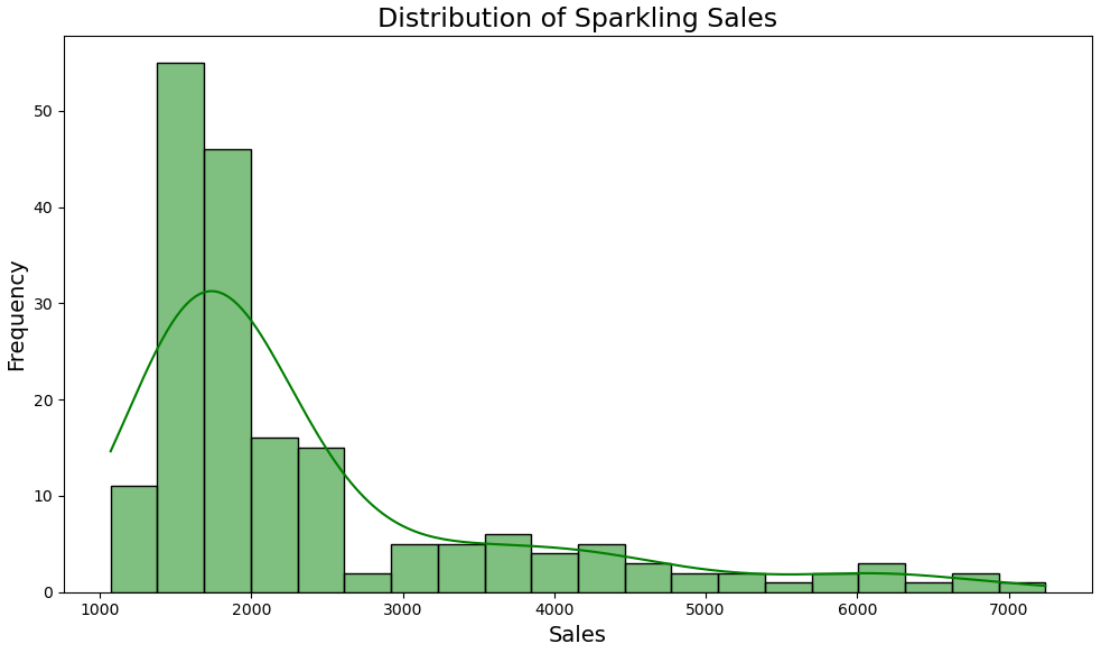


Figure 11 DISTRIBUTION PLOT OF SPARKLING SALES

Observation:

- Right-Skewed Distribution: The distribution appears to be right-skewed, as most sales are concentrated at the lower end of the range, and the frequency decreases as sales increase.
- Peak Frequency: The highest frequency of sales occurs in the 1000–2000 range, indicating this is the most common sales bracket.
- Long Tail: There is a long tail extending to higher sales values, suggesting that while high sales occur, they are relatively rare.
- Density Curve: The fitted density curve mirrors the distribution and highlights the skewness, showing a sharp rise and a gradual decline.

Insights:

- Sales Concentration: Most sales fall between 1000–2000, indicating this is the typical range.
- Right-Skewed Distribution: The data is skewed, with a few cases of very high sales driving up the mean.
- Long Tail: Sales above 5000 are rare but impactful.
- Potential Outliers: High sales values (>5000) might be outliers worth investigating.
- Business Focus: Strategies should target improving performance in the 1000–3000 range while learning from top performers (>5000).

3.4 VISUALISATION OF BOXPLOT FOR MONTHLY ROSE SALES

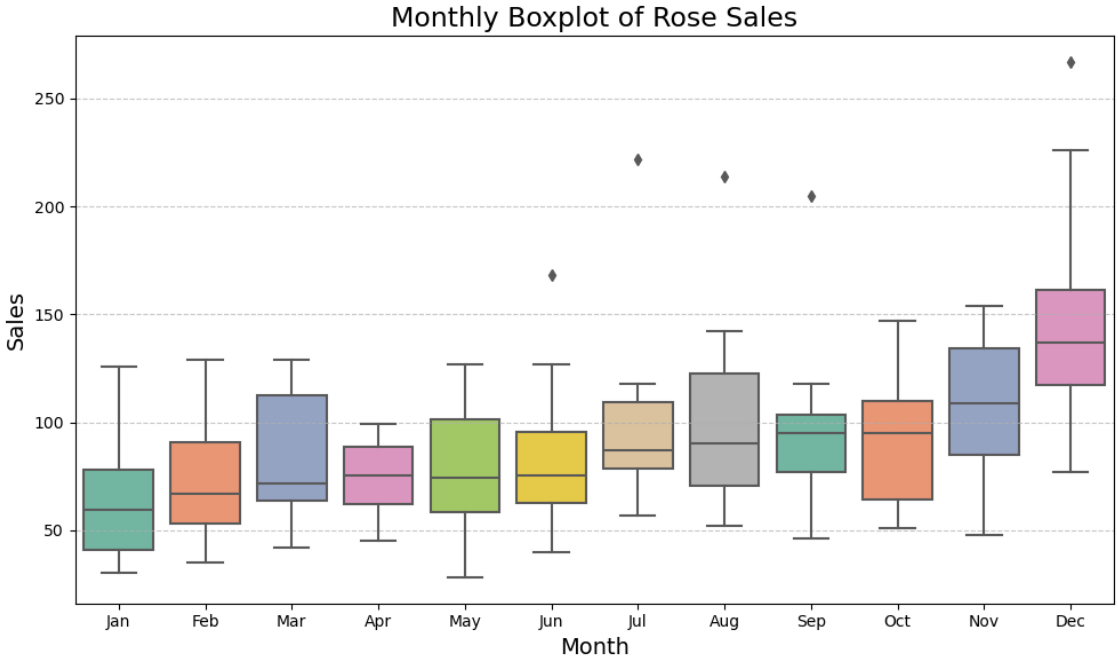


Figure 12 BOXPLOT OF MONTHLY ROSE SALES

Here are the observations based on the boxplot of Monthly Rose Sales:

1. Seasonality:
    - December has the highest median sales and the largest spread, indicating strong seasonal demand, possibly due to holidays.
    - Sales are generally higher in March, November, and December, compared to other months.
2. Outliers:
    - Outliers are noticeable in June, July, August, and December, where some sales significantly exceed typical values. These could represent special events or promotional impacts.
3. Sales Stability:
    - January, February, May, and September exhibit a smaller range (less variability), indicating more consistent sales during these months.

4. Median Sales:
    - The median gradually increases toward the end of the year, with December peaking.
    - Early months (e.g., January and February) have lower medians.
5. Interquartile Range (IQR):
    - March, August, and December have the largest IQR, suggesting greater variability in sales.
    - January, February, and May have smaller IQRs, indicating more predictable sales.

Insights:

1. Seasonal Peak:
    - December experiences the highest sales, likely due to holiday-related demand.
2. Consistent Months:
    - Sales are more stable in January, February, May, and September, with less variability and no outliers.
3. High Variability Months:
    - March, August, and December show significant variability, reflecting irregular events or factors driving occasional spikes.
4. Outliers Indicate Opportunities:
    - Outliers in June, July, August, and December suggest isolated instances of unusually high sales. Understanding these instances could help replicate success.
5. Rising Trend Toward Year-End:
    - Median sales increase steadily toward the last quarter, indicating a natural uptick in demand during the year-end.

3.5 VISUALISATION OF BOXPLOT FOR MONTHLY SPARKLING SALES

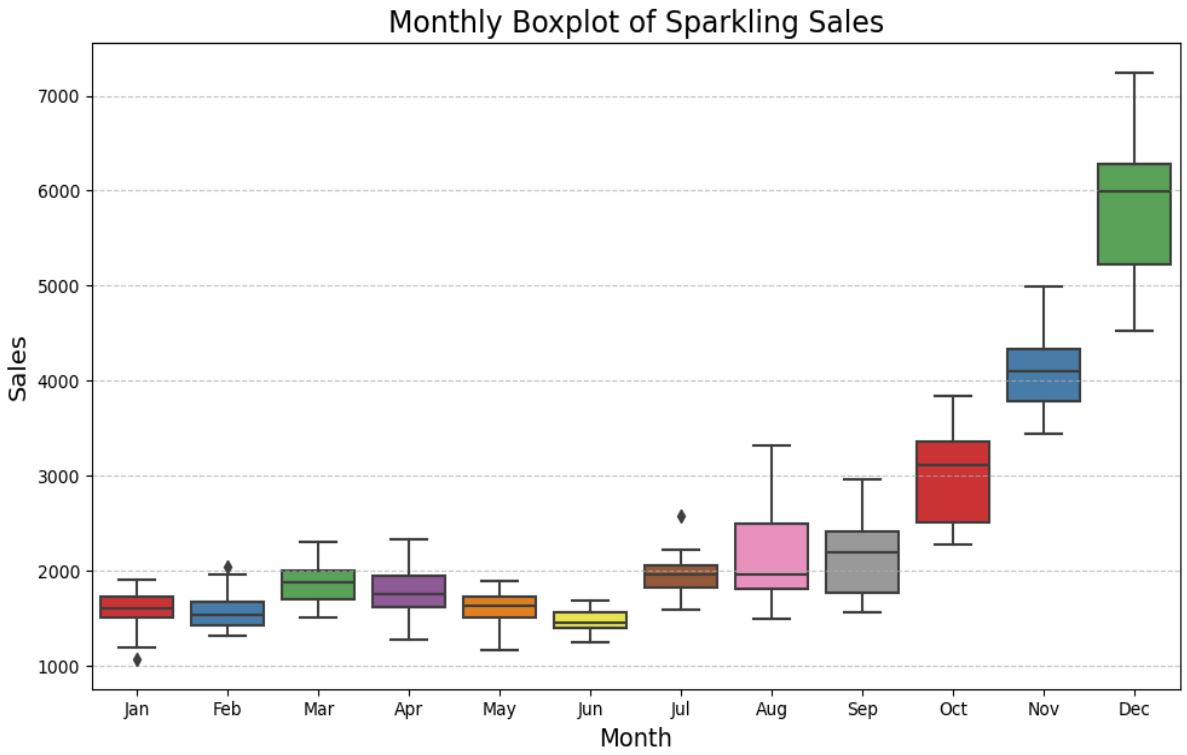


Figure 13 BOXPLOT OF MONTHLY SPARKLING SALES

Observation:

Here are the observations from the Monthly Boxplot of Sparkling Sales:

1. Strong Seasonal Trend:
    - December exhibits the highest sales with a significantly wide range, indicating a strong seasonal demand spike.
    - Sales increase steadily from September to December, reflecting a year-end sales surge.
2. Consistent Low Sales:
    - January to July show lower and more consistent sales, with median values clustering around 1000–2000.
3. Gradual Growth:
    - A gradual increase in sales is observed from August onwards, suggesting a ramp-up period leading to the holiday season.
4. Outliers:
    - Outliers are present in February and July, representing occasional sales spikes in otherwise stable months.

5. Largest Variability:
    - December has the largest variability, with sales ranging from just above 2000 to nearly 7000.

Insights:

- Year-End Focus: Optimize marketing and inventory for the last quarter, especially December.
- Early-Year Stability: Use January–July for building consistent sales strategies due to low variability.
- Outlier Analysis: Investigate February and July outliers to identify replicable opportunities for boosting sales in off-peak months.

3.6 CORRELATION ANALYSIS BETWEEN ROSE AND SPARKLING SALES

Correlation between Rose and Sparkling sales: 0.40457904770543324

4. PERFORM DECOMPOSITION

- Decomposition functions cannot handle missing values, so addressing these gaps is crucial for obtaining accurate and reliable results from the decomposition process. Here, Rose dataset has missing values.
- To perform the decomposition of Rose wine sales, it is essential to treat the missing values in the Rose dataset first.

Let’s perform decomposition of sparkling:

4.1 ADDITIVE DECOMPOSITION – SPARKLING

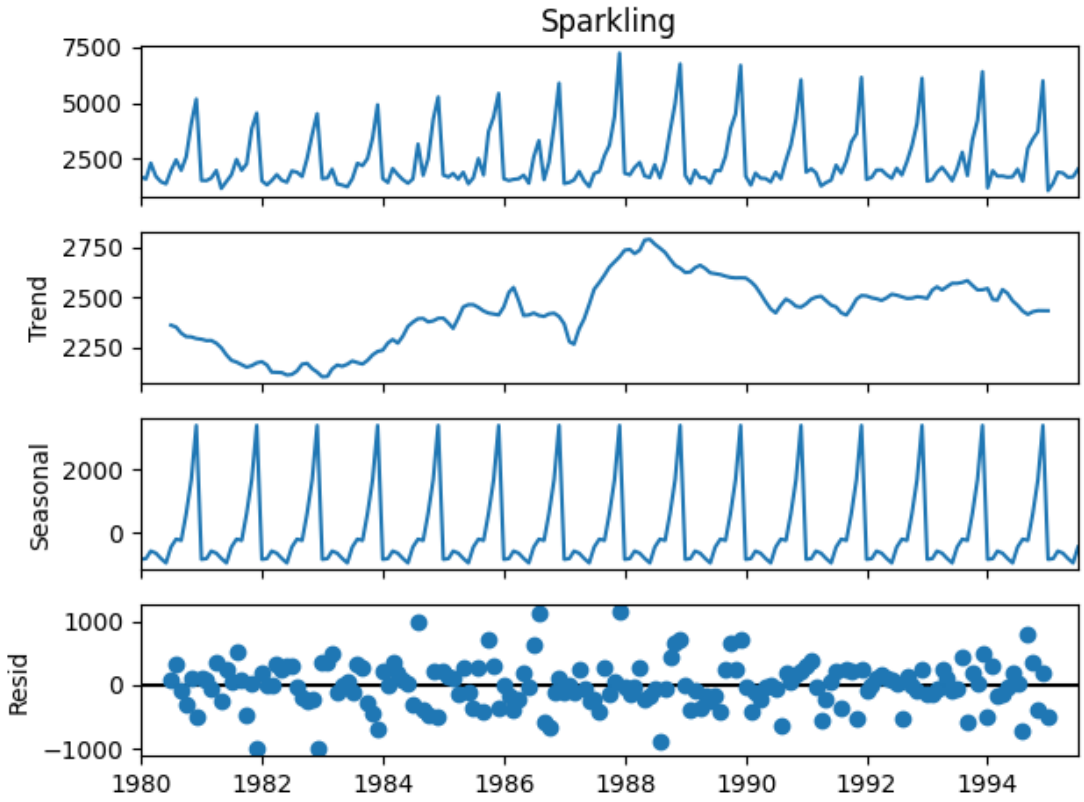


Figure 14 ADDITIVE DECOMPOSITION OF SPARKLING

The image shows a time series decomposition of "Sparkling," breaking it down into its key components: **Original**, **Trend**, **Seasonal**, and **Residual**. Here are the observations for each component:

1. **Original Series (First Plot)**:
    - The data exhibits a clear seasonal pattern, with recurring peaks and troughs.
    - There is an overall upward trend initially, followed by some stabilization or slight decline.
2. **Trend (Second Plot)**:
    - The trend component indicates an initial decline in values until around 1985, followed by a rise peaking around 1990.
    - After the peak, there appears to be a gradual decline or stabilization in the trend.
3. **Seasonal (Third Plot)**:
    - The seasonal component shows strong periodic fluctuations, which remain consistent throughout the time series.
    - This implies that the seasonal effect is stable over time.
4. **Residual (Fourth Plot)**:

- The residuals are scattered around zero with no visible pattern, indicating that the model captures the trend and seasonality well.
- There are some points of higher residual values, which could indicate outliers or unexplained variations.

Summary:

- The time series has both trend and strong seasonal components.
- The residuals suggest the decomposition is effective, with minimal noise or anomalies outside of expected patterns. Further analysis of outliers in the residuals may help refine the model.

4.1.1 UNDERSTANDING THE TREND, SEASONALITY AND RESIDUAL COMPONENT

```
Trend
 YearMonth
1980-01-01             NaN
1980-02-01             NaN
1980-03-01             NaN
1980-04-01             NaN
1980-05-01             NaN
1980-06-01             NaN
1980-07-01     2360.666667
1980-08-01     2351.333333
1980-09-01     2320.541667
1980-10-01     2303.583333
1980-11-01     2302.041667
1980-12-01     2293.791667
Name: trend, dtype: float64

Seasonality
 YearMonth
1980-01-01     -854.260599
1980-02-01     -830.350678
1980-03-01     -592.356630
1980-04-01     -658.490559
1980-05-01     -824.416154
1980-06-01     -967.434011
1980-07-01     -465.502265
1980-08-01     -214.332821
1980-09-01     -254.677265
1980-10-01      599.769957
1980-11-01     1675.067179
1980-12-01     3386.983846
Name: seasonal, dtype: float64

Residual
 YearMonth
1980-01-01            NaN
1980-02-01            NaN
1980-03-01            NaN
1980-04-01            NaN
1980-05-01            NaN
1980-06-01            NaN
1980-07-01      70.835599
1980-08-01     315.999487
1980-09-01     -81.864401
1980-10-01    -307.353290
1980-11-01     109.891154
1980-12-01    -501.775513
Name: resid, dtype: float64
```

Figure 15 TREND, SEASONALITY AND RESIDUAL COMPONENTS – ADDITIVE -SPARKLING

4.2 MULTIPLICATIVE DECOMPOSITION – SPARKLING

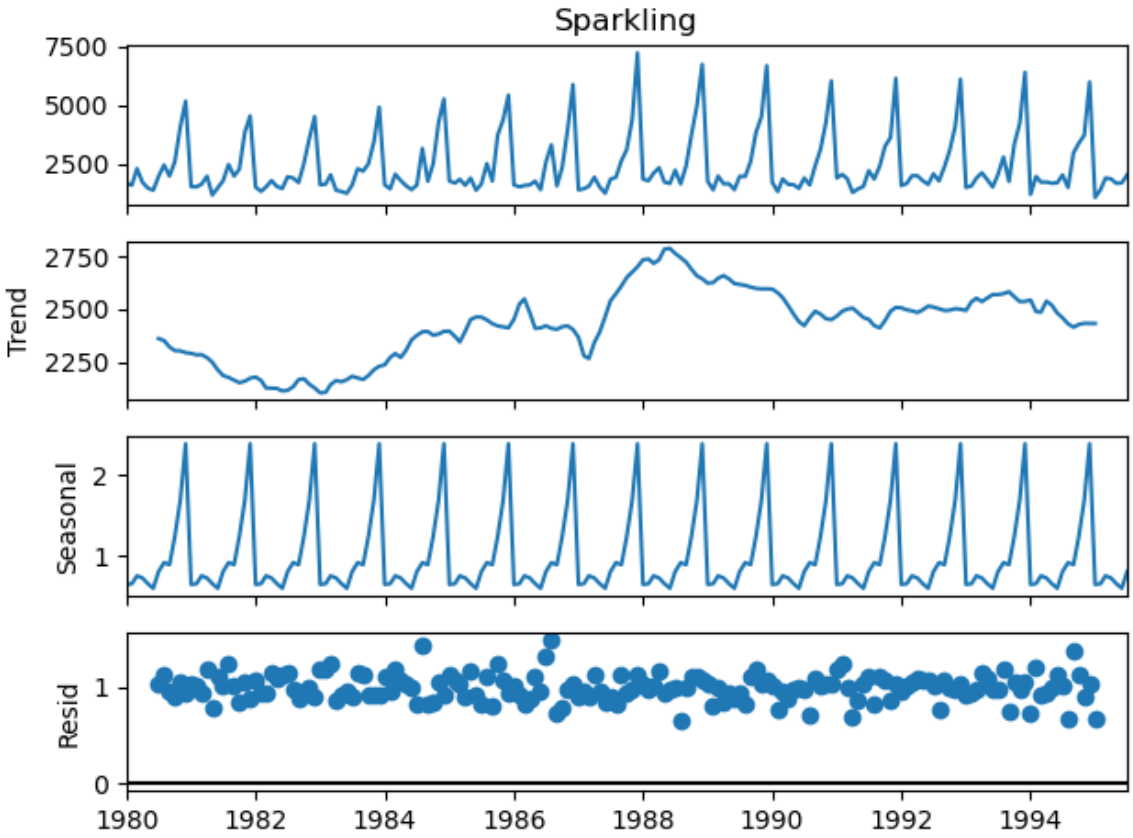


Figure 16 MULTIPLICATIVE DECOMPOSITION OF SPARKLING

- **Nature of Seasonal Component**:
  The seasonal fluctuations in the decomposition appear to be relatively consistent in amplitude across the time period. This suggests that **additive decomposition** might be more appropriate for this dataset, as the seasonal pattern does not seem to amplify or diminish based on the trend's level.
- **Trend Component**:
  The trend shows a steady rise and fall in sales over time, with no evident proportional changes in the seasonal pattern, further supporting the suitability of additive decomposition.
- **Residuals**:
  The residuals do not show any systematic increase or decrease in their range, aligning with the assumptions of an additive model.

Based on these observations, an **additive decomposition model** appears to best fit the Sparkling Sales data.

4.2.1 UNDERSTANDING THE TREND, SEASONALITY AND RESIDUAL COMPONENT

```
Trend
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01    2360.666667
1980-08-01    2351.333333
1980-09-01    2320.541667
1980-10-01    2303.583333
1980-11-01    2302.041667
1980-12-01    2293.791667
Name: trend, dtype: float64

Seasonality
 YearMonth
1980-01-01    0.649843
1980-02-01    0.659214
1980-03-01    0.757440
1980-04-01    0.730351
1980-05-01    0.660609
1980-06-01    0.603468
1980-07-01    0.809164
1980-08-01    0.918822
1980-09-01    0.894367
1980-10-01    1.241789
1980-11-01    1.690158
1980-12-01    2.384776
Name: seasonal, dtype: float64

Residual
 YearMonth
1980-01-01         NaN
1980-02-01         NaN
1980-03-01         NaN
1980-04-01         NaN
1980-05-01         NaN
1980-06-01         NaN
1980-07-01    1.029230
1980-08-01    1.135407
1980-09-01    0.955954
1980-10-01    0.907513
1980-11-01    1.050423
1980-12-01    0.946770
Name: resid, dtype: float64
```

Figure 17 TREND, SEASONALITY AND RESIDUAL COMPONENTS -MULTIPLICATIVE-SPARKLING

- Additive is best when compare with multiplicative. Both residuals have no patterns.

B. DATA PRE-PROCESSING

1. MISSING VALUE TREATMENT

Before treatment:

- There are two missing values in rose sales.

```
Rose    2
dtype: int64
Missing values in Rose dataset before fillng:
              Rose
YearMonth
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
...           ...
1995-03-01   45.0
1995-04-01   52.0
1995-05-01   28.0
1995-06-01   40.0
1995-07-01   62.0

[187 rows x 1 columns]
Sparkling    0
dtype: int64

Missing values in Sparkling dataset:
             Sparkling
YearMonth
1980-01-01        1686
1980-02-01        1591
1980-03-01        2304
1980-04-01        1712
1980-05-01        1471
...                ...
1995-03-01        1897
1995-04-01        1862
1995-05-01        1670
1995-06-01        1688
1995-07-01        2031

[187 rows x 1 columns]
```

Figure 18 MISSING VALUES BEFORE TREATMENT

After treatment:

After applying the forward-fill method, all missing values in Rose Sales were imputed, leaving no missing data.

```
Missing values in Rose dataset:
 Rose    0
dtype: int64

Missing values in Sparkling dataset:
 Sparkling    0
dtype: int64
```

Figure 19 MISSING VALUES AFTER TREATMENT

CONTINUATION OF A.4 PERFORM DECOMPOSITION

4.3 ADDITIVE DECOMPOSITION-ROSE

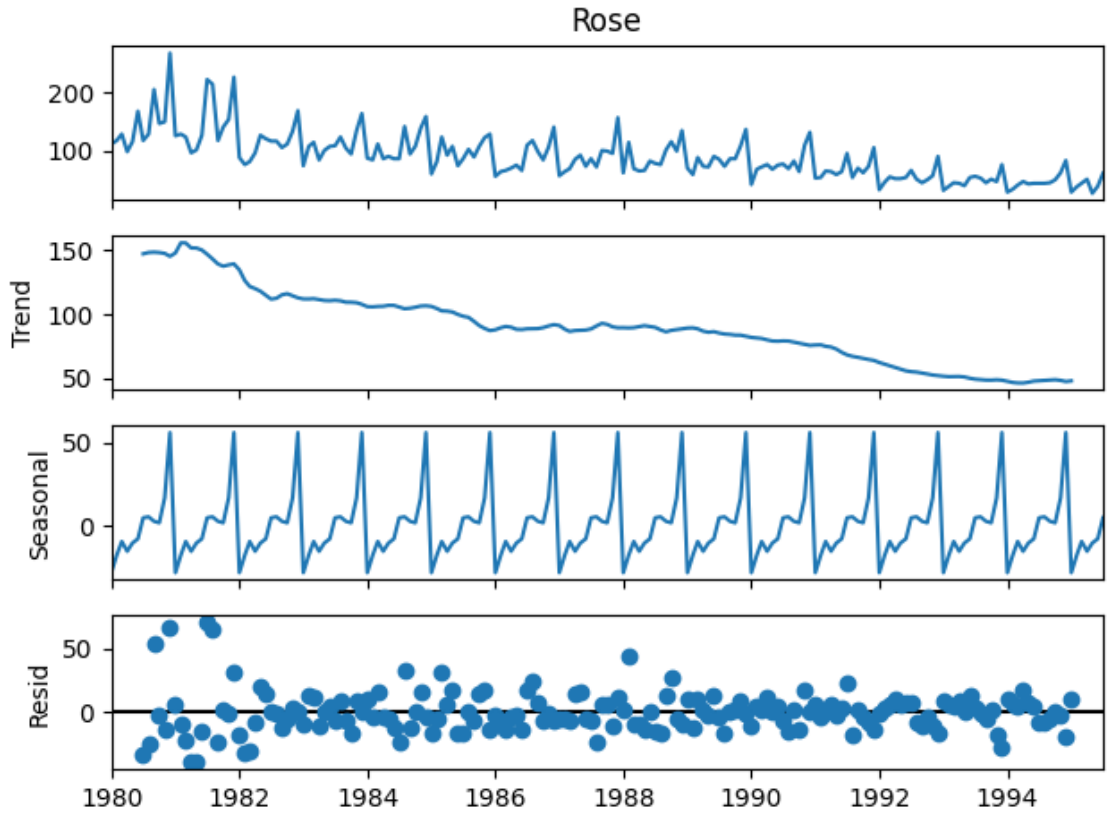


Figure 20 ADDITIVE DECOMPOSITION OF ROSE WINE

Observation:

- **Trend**:
  There is a clear downward trend in Rose Sales over time, indicating a decline in overall demand or popularity.

- **Seasonality**:
  A strong and consistent seasonal pattern is observed, with regular peaks and troughs at fixed intervals, suggesting periodic spikes in demand (likely influenced by seasonal events or promotions).
- **Residuals**:
  The residuals are randomly distributed without any discernible pattern, suggesting that the model has captured the key components (trend and seasonality) effectively.
- **Nature of Decomposition**:
  The seasonal component exhibits consistent amplitude, making **additive decomposition** suitable for this data.

4.3.1 UNDERSTANDING THE TREND, SEASONALITY AND RESIDUAL COMPONENT:

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

Seasonality
 YearMonth
1980-01-01   -27.903092
1980-02-01   -17.431663
1980-03-01    -9.279878
1980-04-01   -15.092378
1980-05-01   -10.190592
1980-06-01    -7.672735
1980-07-01     4.880241
1980-08-01     5.460797
1980-09-01     2.780241
1980-10-01     1.877464
1980-11-01    16.852464
1980-12-01    55.719130
Name: seasonal, dtype: float64

Residual
 YearMonth
1980-01-01          NaN
1980-02-01          NaN
1980-03-01          NaN
1980-04-01          NaN
1980-05-01          NaN
1980-06-01          NaN
1980-07-01   -33.963575
1980-08-01   -24.585797
1980-09-01    53.844759
1980-10-01    -2.960797
1980-11-01   -14.269130
1980-12-01    66.155870
Name: resid, dtype: float64
```

Figure 21 TREND,SEASONALITY AND RESIDUAL COMPONENT - ADDITIVE – ROSE

4.4 MULTPLICATIVE DECOMPOSITION-ROSE

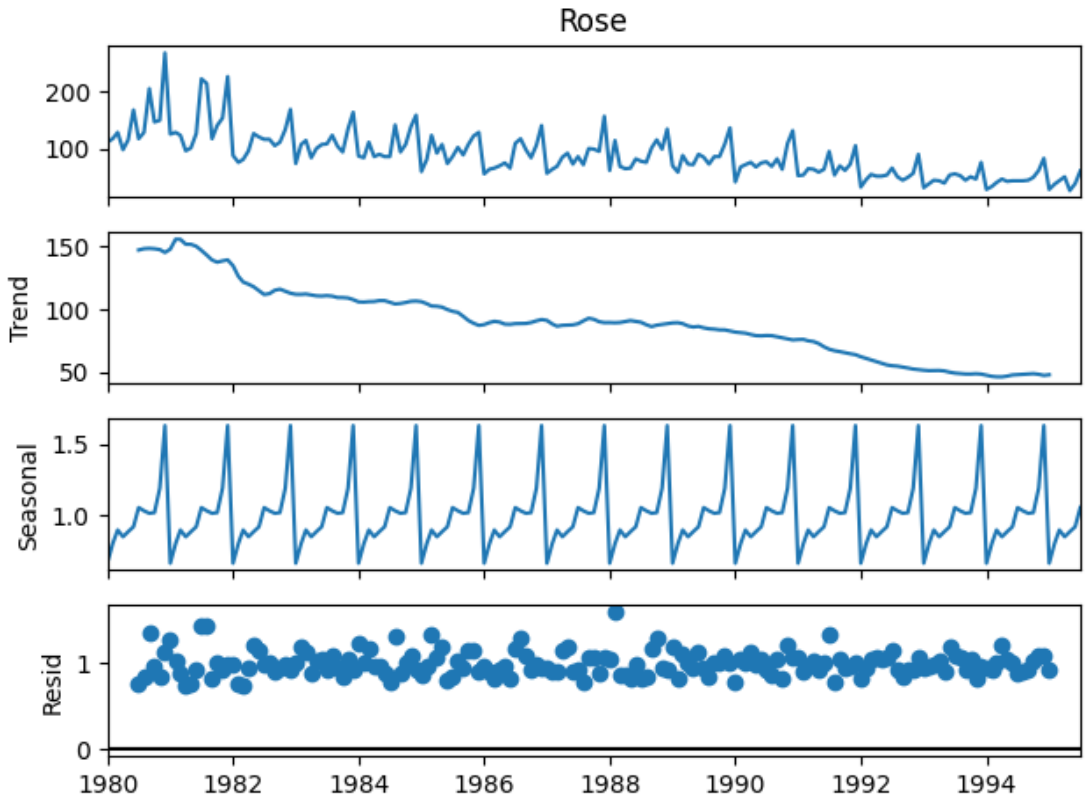


Figure 22 MULTIPLICATIVE DECOMPOSITION OF ROSE

Observation:

- **Nature of Seasonal Component**:
  The seasonal fluctuations in the decomposition appear to be relatively consistent in amplitude across the time period. This suggests that **additive decomposition** might be more appropriate for this dataset, as the seasonal pattern does not seem to amplify or diminish based on the trend's level.
- **Trend Component**:
  The trend shows a steady rise and fall in sales over time, with no evident proportional changes in the seasonal pattern, further supporting the suitability of additive decomposition.
- **Residuals**:
  The residuals do not show any systematic increase or decrease in their range, aligning with the assumptions of an additive model.

Based on these observations, an **additive decomposition model** appears to best fit the Sparkling Sales data.

4.4.1 UNDERSTANDING THE TREND, SEASONALITY AND RESIDUAL COMPONENT:

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

Seasonality
 YearMonth
1980-01-01    0.670182
1980-02-01    0.806224
1980-03-01    0.901278
1980-04-01    0.854154
1980-05-01    0.889531
1980-06-01    0.924099
1980-07-01    1.057682
1980-08-01    1.035066
1980-09-01    1.017753
1980-10-01    1.022688
1980-11-01    1.192494
1980-12-01    1.628848
Name: seasonal, dtype: float64

Residual
 YearMonth
1980-01-01         NaN
1980-02-01         NaN
1980-03-01         NaN
1980-04-01         NaN
1980-05-01         NaN
1980-06-01         NaN
1980-07-01    0.758514
1980-08-01    0.841382
1980-09-01    1.357534
1980-10-01    0.970661
1980-11-01    0.853274
1980-12-01    1.129506
Name: resid, dtype: float64
```

Figure 23 TREND, SEASONALITY AND RESIDUAL COMPONENT - MULTIPLICATIVE - ROSE

- Additive is best when compare with multiplicative. Both residuals have no patterns.

2. VISUALIZE THE PROCESSED DATA

2.1 ROSE WINE SALES OVER TIME

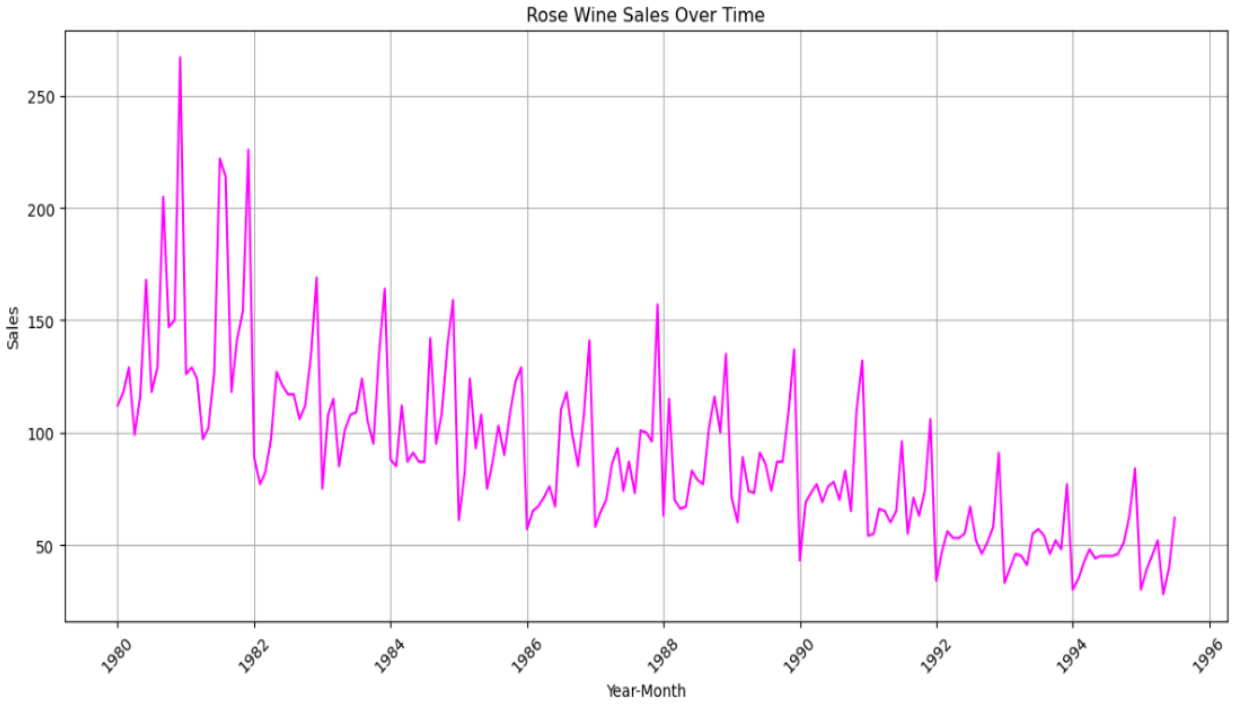


Figure 24 VISUALISATION OF PROCESSED DATA - ROSE WINE

Observation:

1. **Initial Peak**:
   Rose Wine Sales show a sharp peak around 1980-1981, indicating a period of high demand.
2. **Declining Trend**:
   After the initial peak, there is a noticeable decline in sales over the years, reflecting a downward trend in overall demand.
3. **Seasonal Fluctuations**:
   Despite the declining trend, periodic fluctuations with recurring peaks suggest a seasonal pattern in sales.
4. **Stabilization**:
   Towards the later years (1990-1995), sales appear to stabilize at lower levels, with fewer extreme spikes.

This suggests that while Rose Wine Sales are influenced by seasonal factors, their overall popularity has decreased over time.

2.2 SPARKLING WINE SALES OVER TIME

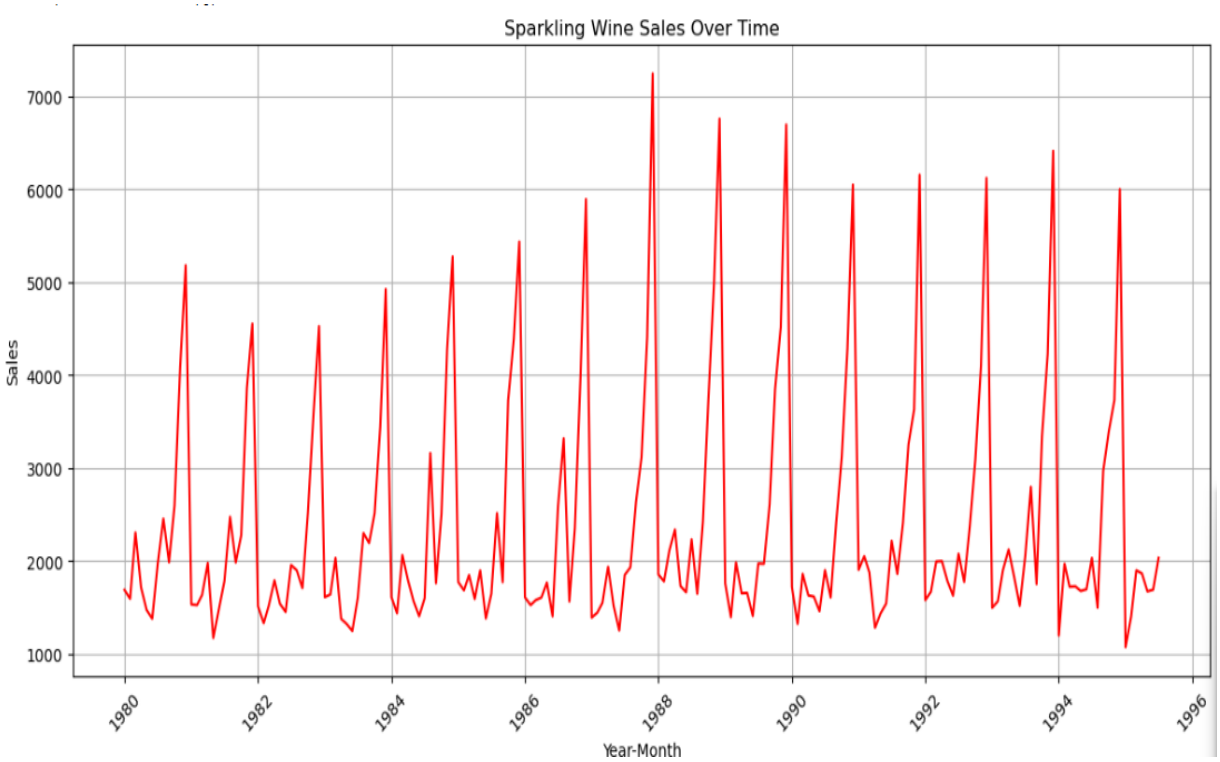


Figure 25 VISUALISATION OF PROCESSED DATA -SPARKLING WINE

This time-series plot shows the sales of sparkling wine over time, ranging from 1980 to 1996. Here are some key observations:

1. **Seasonality**: There is a clear yearly seasonal pattern, with sharp peaks occurring regularly, likely around the same months each year. This suggests higher sales during specific seasons, possibly holidays or festive periods.
2. **Trend**: While there is no significant long-term upward or downward trend, the peaks remain consistent over time, indicating stable demand patterns during those peak months.
3. **Magnitude of Peaks**: The highest sales during peak months are consistent over the years, often reaching above 6,000 units.
4. **Off-Peak Sales**: The sales during off-peak months show some variability but generally remain within the range of 1,000–2,000 units.

3. TRAIN-TEST SPLIT

```
Rose Train shape: (149, 2)
Rose Test shape: (38, 2)
Sparkling Train shape: (149, 2)
Sparkling Test shape: (38, 2)
```

Dataset Overview:

- The Rose dataset is divided into a training set (149 rows, 2 columns) and a testing set (38 rows, 2 columns).
- Similarly, the Sparkling dataset is split into a training set (149 rows, 2 columns) and a testing set (38 rows, 2 columns).
- Each dataset consists of 2 variables, likely representing features and/or target labels.

Insights:

- **Train/Test Split**: The datasets are split into approximately 80% for training and 20% for testing, which is a standard approach for achieving reliable model evaluation.
- **Dataset Comparison**: If the columns represent comparable features (e.g., time and sales of Rose and Sparkling wines), visualizations can be used to analyze and compare their trends, seasonality, or distributions, providing deeper insights into similarities and differences between the two datasets.

3.1 TRAIN AND TEST DATA – ROSE

First few rows of Rose Training Data

	YearMonth	Rose
0	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

Last few rows of Rose Training Data

	YearMonth	Rose
144	1992-01-01	34.0
145	1992-02-01	47.0
146	1992-03-01	56.0
147	1992-04-01	53.0
148	1992-05-01	53.0

First few rows of Rose Test Data

	YearMonth	Rose
149	1992-06-01	55.0
150	1992-07-01	67.0
151	1992-08-01	52.0
152	1992-09-01	46.0
153	1992-10-01	51.0

Last few rows of Rose Test Data

	YearMonth	Rose
182	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

Figure 26 FEW ROWS OF TRAIN AND TEST DATA - ROSE

3.2 TRAINING AND TEST DATA – SPARKLING

First few rows of Sparkling Training Data

	YearMonth	Sparkling
0	1980-01-01	1686
1	1980-02-01	1591
2	1980-03-01	2304
3	1980-04-01	1712
4	1980-05-01	1471

Last few rows of Sparkling Training Data

	YearMonth	Sparkling
144	1992-01-01	1577
145	1992-02-01	1667
146	1992-03-01	1993
147	1992-04-01	1997
148	1992-05-01	1783

First few rows of Sparkling Test Data

	YearMonth	Sparkling
149	1992-06-01	1625
150	1992-07-01	2076
151	1992-08-01	1773
152	1992-09-01	2377
153	1992-10-01	3088

Last few rows of Sparkling Test Data

	YearMonth	Sparkling
182	1995-03-01	1897
183	1995-04-01	1862
184	1995-05-01	1670
185	1995-06-01	1688
186	1995-07-01	2031

Figure 27 FEW ROWS OF TRAIN AND TEST DATA - SPARKLING

3.3 PLOTTING ROSE DATA TRAIN AND TEST SPLIT

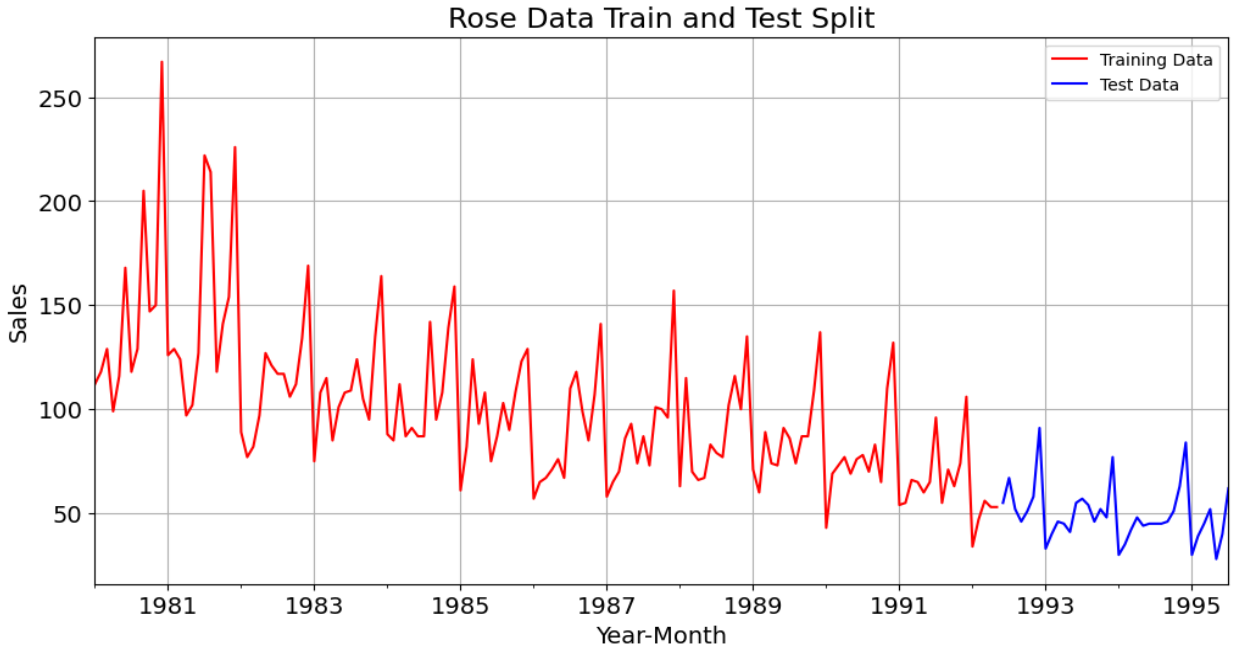


Figure 28 PLOTTING ROSE DATA TRAIN AND TEST SPLIT

Observations:

1. **Declining Trend**: Rosé wine sales show a downward trend from 1980 to 1995, indicating reduced consumer demand.
2. **High Volatility**: Sales fluctuate significantly, with peaks over 250 units and troughs near 50 units, reflecting inconsistent purchasing behavior.
3. **Stabilization in Test Data**: Sales stabilize in later years, suggesting potential predictability for future trends.
4. **Market Insights**: The decline and volatility likely stem from shifting consumer preferences, economic factors, or increased competition.
5. **Strategic Opportunities**: Stabilizing sales offer a chance for recovery through targeted marketing and improved inventory management.

3.4 PLOTTING SPARKLING DATA TRAIN AND TEST SPLIT

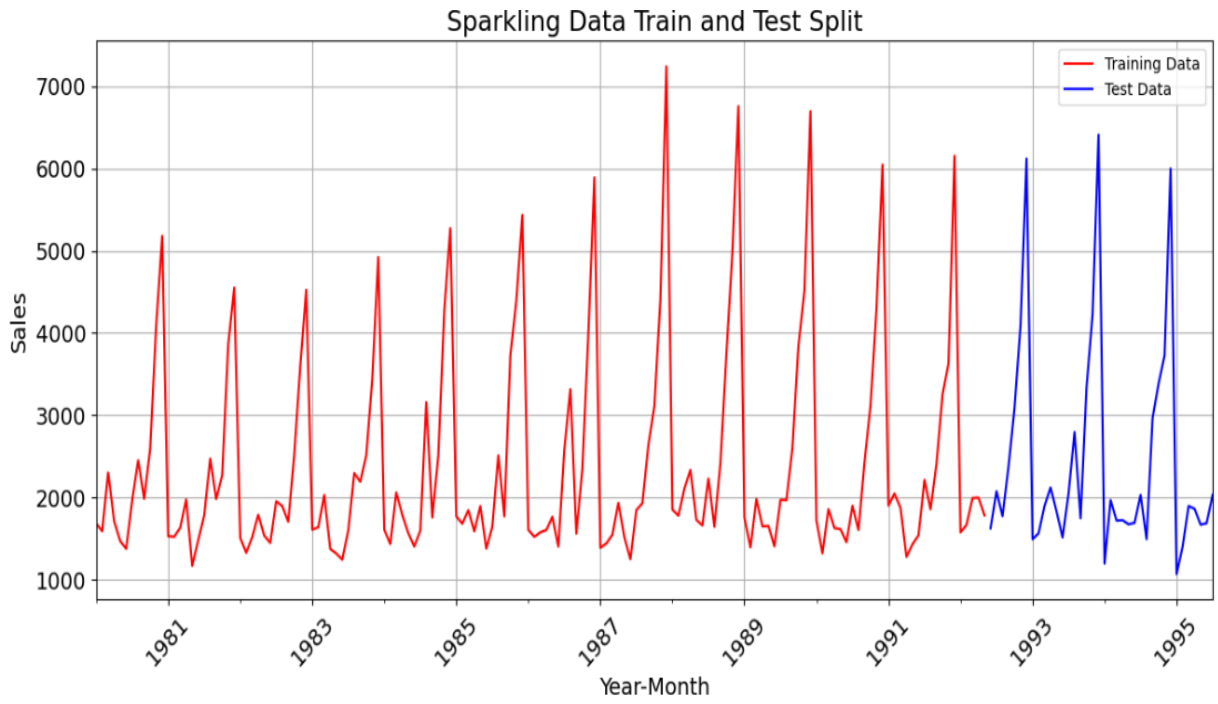


Figure 29 PLOTTING SPARKLING DATA TRAIN AND TEST SPLIT

OBSERVATION:

- **Seasonal Pattern**: Sales consistently peak annually, likely during holidays or celebrations.
- **High Variability**: Sales fluctuate between peaks above 6,000 units and troughs near 2,000 units.
- **Stable Trend**: Overall sales remain steady from 1980 to 1995, with no significant growth or decline.
- **Test Data Consistency**: The test data mirrors the training data's patterns, suggesting predictable future trends.
- **Actionable Insight**: Seasonal trends highlight opportunities for targeted marketing and inventory optimization.

C. MODEL BUILDING - ORIGINAL DATA

Forecasting Models Overview: We will implement and evaluate the following forecasting models:

- Linear Regression
- Simple Average
- Moving Average
- Exponential Smoothing Models (Single, Double, Triple)

MODEL 1- LINEAR REGRESSION

1.1 LINEAR REGRESSION - ROSE

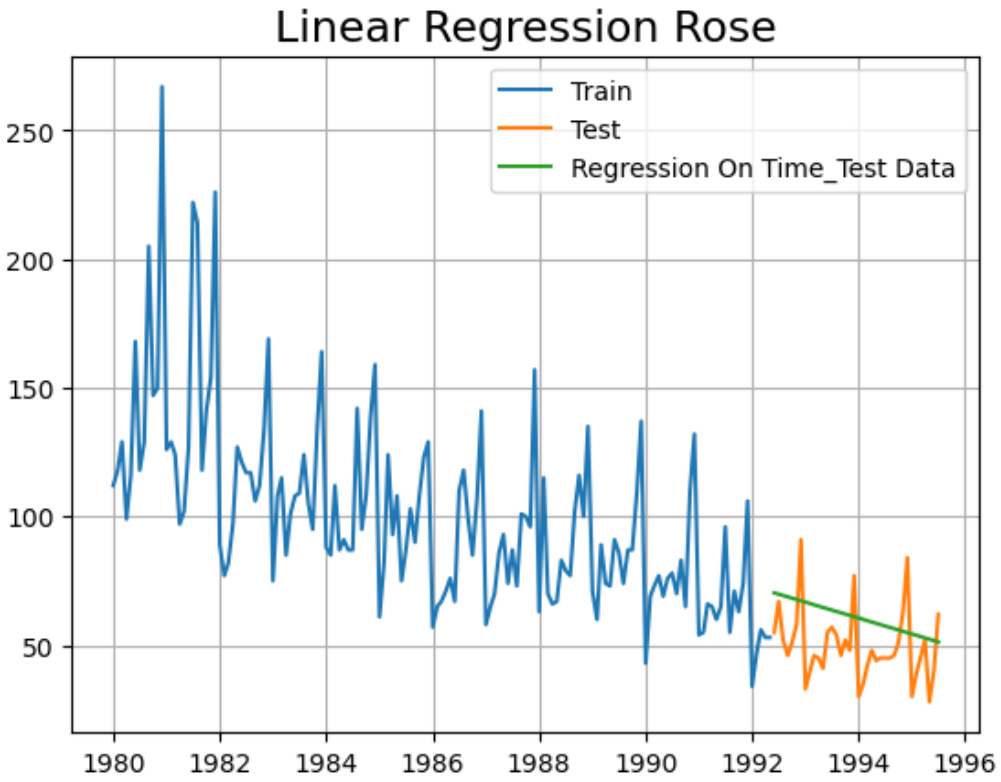


Figure 30 LINEAR REGRESSION - ROSE

Observations:

1. **Fluctuating Sales**: Rosé wine sales (blue line) show noticeable ups and downs from 1980 to 1995.
2. **Declining Trend**: Both training (blue) and test data (orange) indicate a steady decrease in sales over time.
3. **Stabilization**: Test data (orange) shows less variation, suggesting sales are stabilizing in later years despite the overall decline.
4. **Future Outlook**: The green regression line confirms a gradual decline in sales if current trends continue.

	Test RMSE Rose
RegressionOnTime	17.510241

Figure 31 TEST RMSE ROSE - LINEAR REGRESSION

Observations:

1. **Model Accuracy**: An RMSE of 17.51 indicates the model's average prediction error is 17.51 units in the test dataset.
2. **Performance Evaluation**: Compare the RMSE to the data range (e.g., 50–250 units for Rose sales) to determine if the error is acceptable. A relatively low RMSE suggests good performance, while a high RMSE indicates room for improvement.
3. **Enhancement Options**: Improve the model by:
    - Adding relevant features (e.g., seasonality or economic factors).
    - Exploring advanced time-series methods like ARIMA, SARIMA, or LSTMs.
    - Analyzing residuals to detect patterns, systematic errors, or outliers.

1.2 LINEAR REGRESSION - SPARKLING

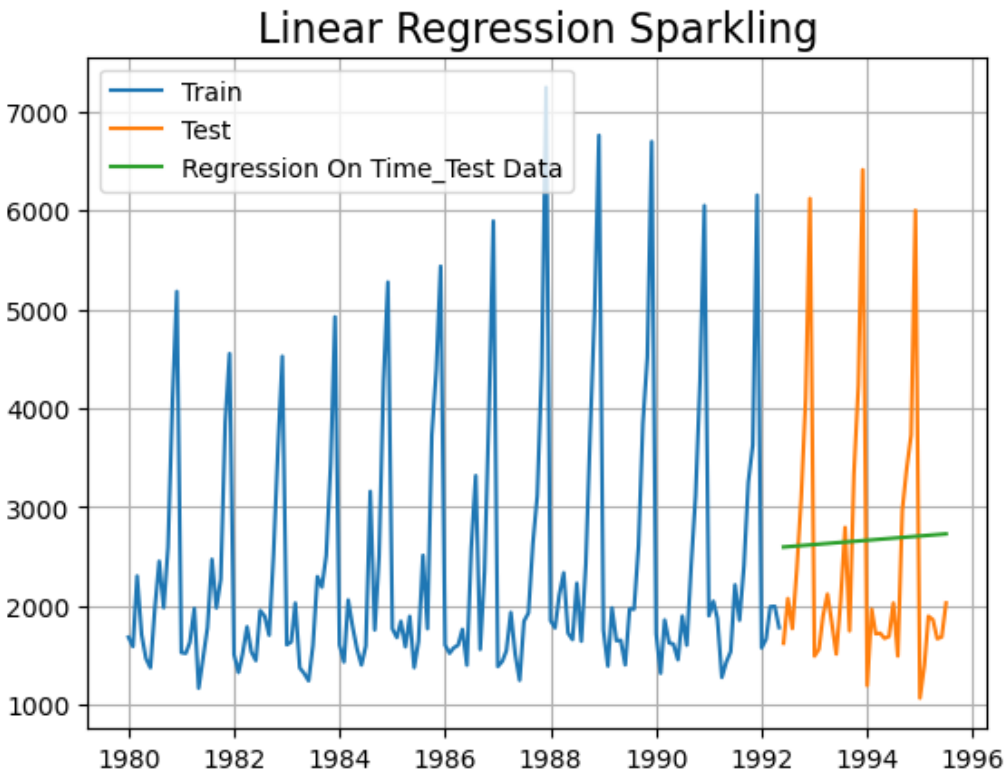


Figure 32 LINEAR REGRESSION -SPARKLING

Observations:

1. **Data Overview**:
    - **Train Data (Blue)**: Sparkling wine sales (1980–1995) fluctuate significantly, ranging between 1,000 and 7,000 units.

- **Test Data (Orange)**: Similar fluctuations continue, with higher peaks observed in later years.
- **Regression Line (Green)**: Represents the overall trend, showing a simplified analysis of sales.

2. **Seasonal Trends**:
   - Sales follow a consistent yearly pattern, peaking during festive or holiday seasons.
3. **Fluctuations**:
   - Both datasets show high variability, with strong seasonal peaks and low off-season engagement.
4. **Stabilization**:
   - Test data peaks are more pronounced in later years, suggesting renewed interest or increased consumption.
5. **Trend Analysis**:
   - The regression line shows a slight upward slope, indicating possible stabilization or growth in sales.
6. **Key Takeaways**:
   - Focus on marketing during peak seasons to boost sales.
   - Use insights on fluctuations to optimize inventory.
   - The upward trend signals potential for future growth with strategic efforts.

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	17.510241	1349.042457

Figure 33 TEST RMSE ROSE & SPARKLING - LINEAR REGRESSION

Observation:

- **High Prediction Error**: An RMSE of 1349.042 is extremely high compared to the sales range of 1500–3000 units, indicating poor model accuracy and significant deviation from actual values.

MODEL 2- SIMPLE AVERAGE

2.1 SIMPLE AVERAGE -ROSE

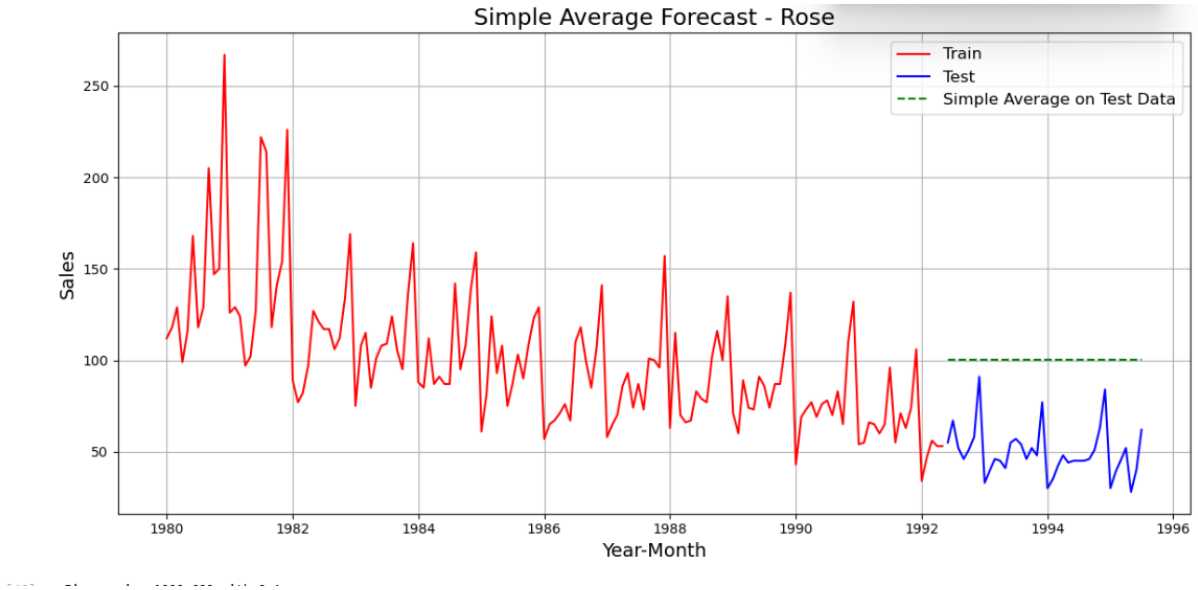


Figure 34 SIMPLE AVERAGE -ROSE

Observations:

1. **Declining Trend**: Rosé wine sales peaked early (around 250 units) but steadily declined over time, showing reduced demand.
2. **Lower Popularity**: Monthly fluctuations are present, but peaks are much smaller compared to sparkling wine, indicating less appeal.
3. **Stable Test Data**: Test data (blue line) is flat, with low peaks around 100 units, reflecting steady but declining sales.
4. **Consistently Low Sales**: The green dashed line shows consistently low average sales, highlighting ongoing challenges in increasing demand.
5. **No Seasonal Spikes**: Unlike sparkling wine, rosé sales lack distinct seasonal patterns, suggesting it's less tied to celebratory occasions.

2.1 SIMPLE AVERAGE -SPARKLING

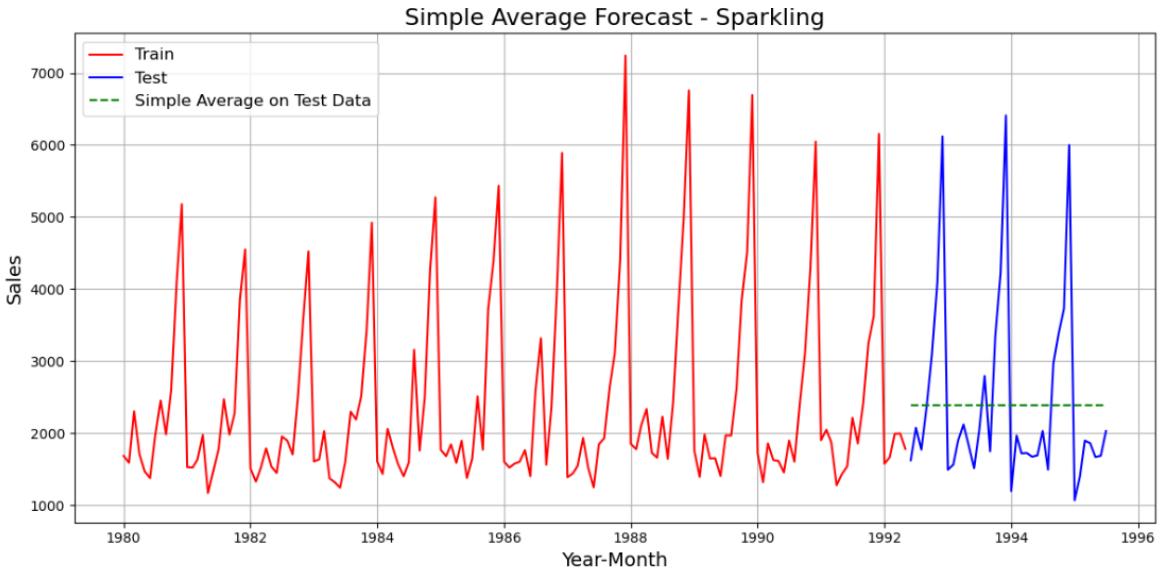


Figure 35 SIMPLE AVERAGE - SPARKLING

1. **Sales Fluctuations:**
   - The training data (red line) shows big changes in sales, with peaks around 6000-7000 units, indicating strong seasonal demand.
2. **Seasonal Peaks:**
   - Sales increase at certain times, likely during holidays or special events.
3. **Test Data Consistency:**
   - The test data (blue line) has similar fluctuations but appears more stable, showing continued consumer interest.
4. **Average Sales Trend:**
   - The green dashed line represents the average sales, around 2000 units, serving as a baseline for overall performance.
5. **Trend Observation:**
   - While there are occasional spikes, the sales trend remains steady over time, with no major yearly changes in the test data.

	Test RMSE Rose	Test RMSE Sparkling
SimpleAverageModel	52.239499	1331.037637

Figure 36 TEST RMSE FOR ROSE & SPARKLING - SIMPLE AVERAGE

Simple Average Model Performance:

- For Rose, the Test RMSE is 52.24, much higher than the RegressionOnTime model (17.51), showing it performs worse.
- For Sparkling, the Test RMSE is 1331.04, slightly better than RegressionOnTime (1349.04), but still shows large prediction errors.
- **Model Suitability:**
- The Simple Average Model doesn't work well for Rose, as it can't capture the downward trend or fluctuations.
- For Sparkling, it performs a bit better, suggesting there may be some patterns it can detect.
- **Error Magnitude:**
- Both models have high RMSEs for Sparkling, indicating the need for more advanced methods to handle the variability and trends in this data.
- **Insights for Future Models:**
- A model that considers trends or seasonality would likely do better than both RegressionOnTime and Simple Average Model, especially for Sparkling.

MODEL 3- MOVING AVERAGE

3.1 MOVING AVERAGE -ROSE

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	114.5	NaN	NaN
1980-05-01	116.0	107.5	115.5	NaN	NaN

Figure 37 RAW SALES DATA FOR ROSÉ WINE ALONG WITH CALCULATED TRAILING AVERAGES

OBSERVATION:

1. **Sales Data and Moving Averages:**
    - The table shows raw sales data for rosé wine along with calculated trailing averages for 2, 4, and 9 periods.

- Early months show sales data, but trailing averages can't be calculated until there are enough data points (e.g., 4, 6, and 9-period averages need more data).

3.1.1 PLOTTING ON THE WHOLE DATA

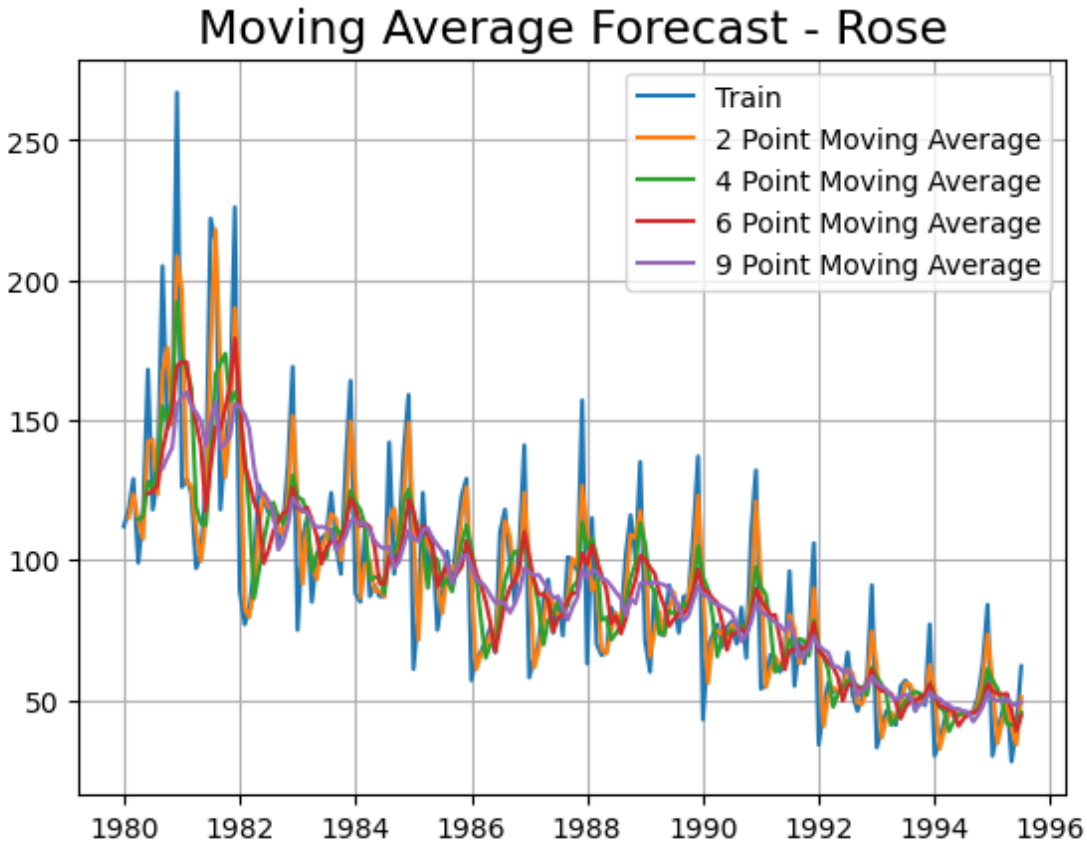


Figure 38 MOVING AVERAGE FORECAST - ROSE ON WHOLE DATA

OBSERVATION:

1. **Comparison of Moving Averages:**

- The graph shows the training data (blue line) along with various moving averages (2, 4, 6, and 9 periods).
- These moving averages follow the training data closely but show a decreasing trend over time.

2. **Visual Stability:**

- The moving averages start to converge, especially in the later years, suggesting that sales may stabilize at lower average levels.

3. **Peak Visibility:**

- The sales peaks are more noticeable in the training data, while the moving averages smooth out these fluctuations, showing the difficulty of maintaining consistent sales.

3.1.2 PLOTTING ON BOTH THE TRAINING AND TEST DATA

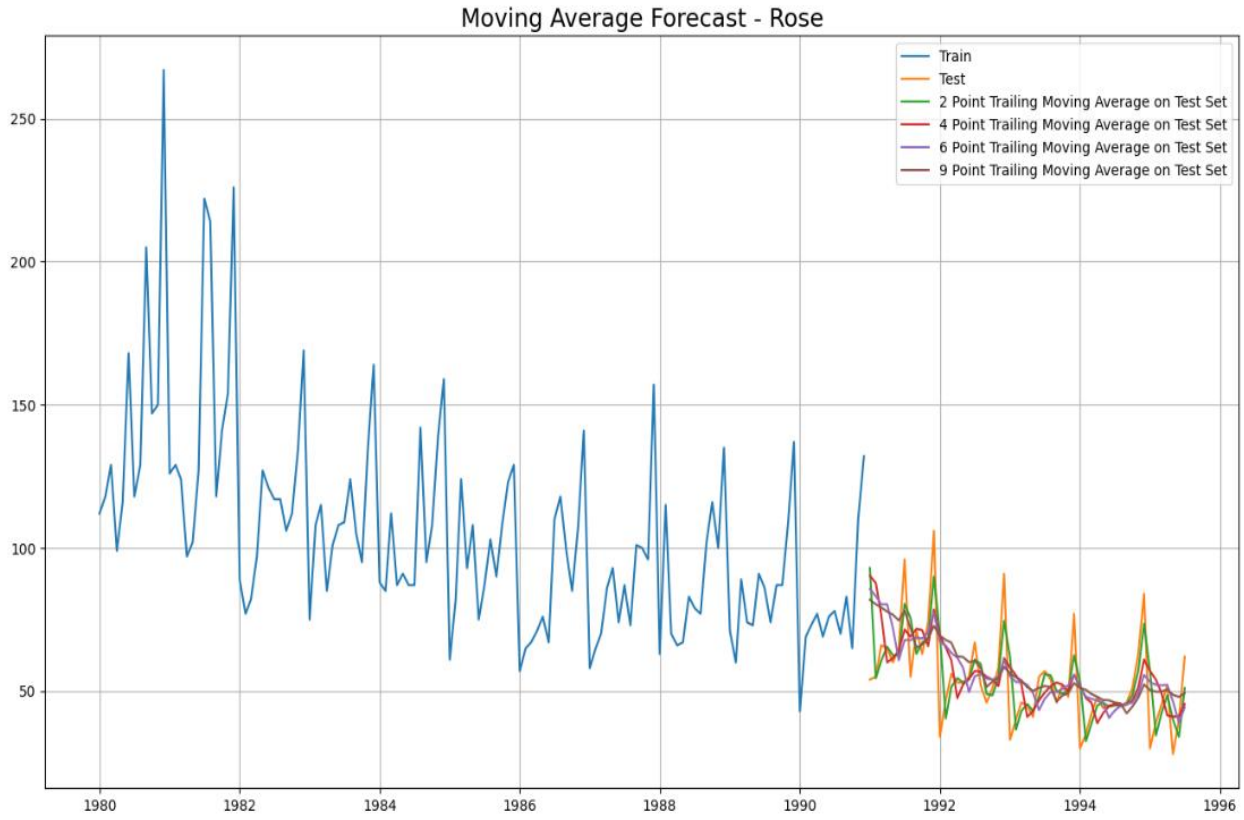


Figure 39 MOVING AVERAGE FORECAST-ROSE ON BOTH TRAIN AND TEST DATA

OBSERVATION:

1. **Training Data:**
    - The blue line shows the training data, reflecting sales trends from 1980 to 1996. It has noticeable peaks and valleys, indicating fluctuating sales.
2. **Test Data:**
    - The test data (orange line) follows the same trend as the training data but with less variability, suggesting more stable sales in the later years.
3. **Moving Averages:**
    - The moving averages (2, 4, 6, and 9 periods) smooth out the data, making overall trends easier to see compared to the raw sales data.
    - The 9-point moving average (purple line) is more stable, highlighting long-term trends more accurately than shorter averages.
4. **Trend Analysis:**
    - The moving averages show a general downward trend, especially in the recent years, which may indicate a decline in interest for rosé wine.

	Test RMSE Rose
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209

Figure 40 TEST RMSE - ROSE FOR DIFFERENT TRAILING MOVING AVERAGE

OBSERVATION:

1. **2-Point Moving Average:**
    - The 2-point moving average has the lowest RMSE (11.53), making it the most accurate model for this dataset. This suggests that the most recent two months' data are the best predictor for future sales.
2. **4-Point Moving Average:**
    - The 4-point moving average has a slightly higher RMSE (14.46) but still performs well. The small increase in error might mean that using a wider window (four points) introduces more noise or doesn't capture short-term trends as effectively.
3. **6-Point and 9-Point Moving Averages:**
    - The 6-point and 9-point moving averages have similar RMSE values (14.57 and 14.73), showing that as the window size increases beyond 4 points, performance plateaus or even worsens.
    - Larger windows may smooth out important short-term fluctuations, leading to higher prediction errors.
4. **Trend Sensitivity:**
    - The 2-point moving average is most responsive to recent trends, offering better short-term accuracy. In contrast, longer averages are less sensitive and may miss recent fluctuations.
5. **Implications:**
    - The 2-point moving average is likely the best choice, as it provides the lowest error and captures recent patterns effectively. Larger windows (4, 6, or 9 points) don't significantly improve accuracy and may even increase error slightly.

3.2 MOVING AVERAGE – SPARKLING

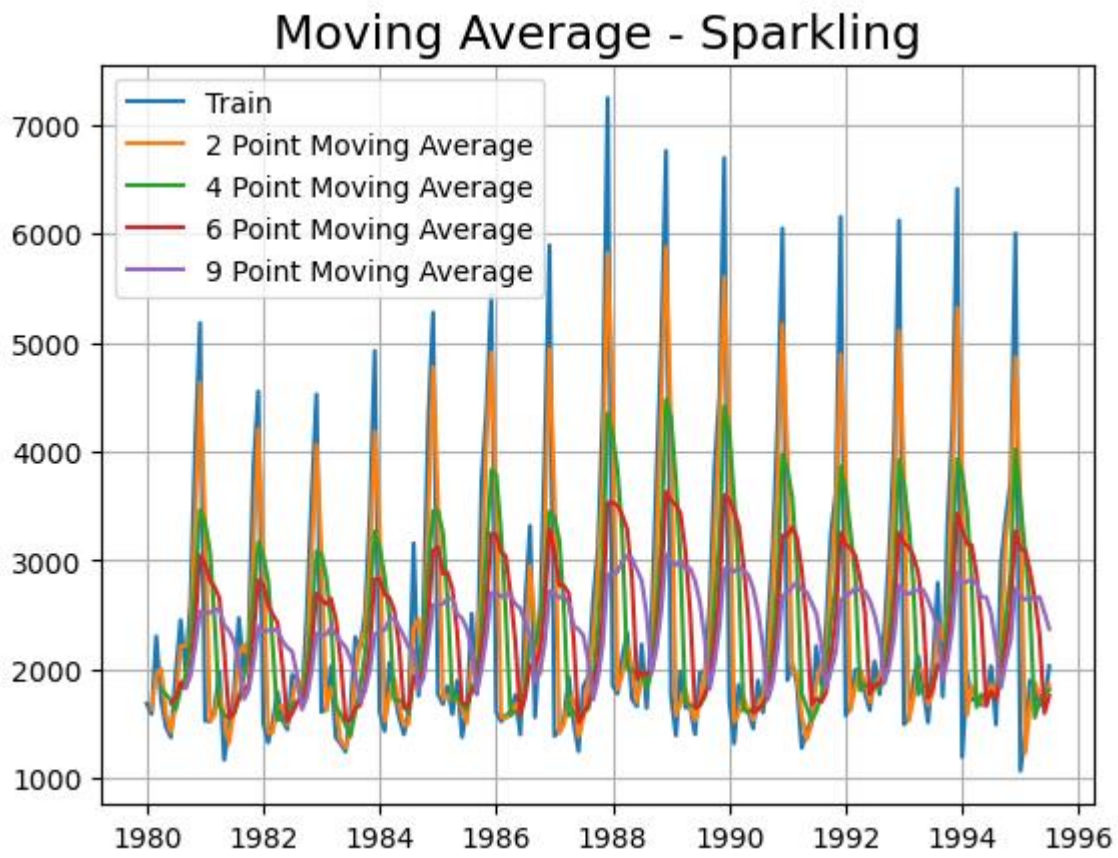
YearMonth	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN

Figure 41 RAW SALES DATA FOR SPARKLING WINE ALONG WITH CALCULATED TRAILING AVERAGES

Trailing Moving Averages:

- **Trailing_2:** The 2-point trailing average is calculated starting from the second row, using the current and previous month's data (e.g., for 1980-02-01, the average of 1686 and 1591 is 1638.5).
- **Trailing_4, Trailing_6, Trailing_9:** These columns show "NaN" (Not a Number) where there isn't enough previous data to calculate the moving average. For example, the Trailing_4 average needs 4 months of data, but by 1980-04-01, only 3 months are available, so the value remains NaN.

3.2.1 PLOTTING ON THE WHOLE DATA



MOVING AVERAGE FORECAST - SPARKLING ON WHOLE DATA

Observations:

1. **Seasonal Sales Patterns:**
    - Sales spike during certain months, likely around holidays, indicating strong seasonal demand for targeted marketing.
2. **Fluctuations:**
    - The training data shows significant volatility, with peaks at 7000 units and troughs around 1000 units, highlighting unpredictable demand.
3. **Moving Averages:**
    - The 9-point moving average provides the most stable view, showing long-term trends while minimizing short-term fluctuations.
4. **Stabilizing Trend:**
    - Despite seasonal peaks, sales remain stable at 2000-3000 units in later years, suggesting no drastic year-to-year decline.
5. **Average Sales Context:**

- The baseline of 2000-3000 units remains stable, indicating reliability for future sales forecasts.

6. **Strategic Opportunities:**
    - Peaks suggest prime months for promotions, allowing for better inventory and marketing planning.
7. **Market Dynamics:**
    - Stable sales levels show a solid consumer base despite competition and volatility.

3.2.2 PLOTTING ON BOTH THE TRAINING AND TEST DATA

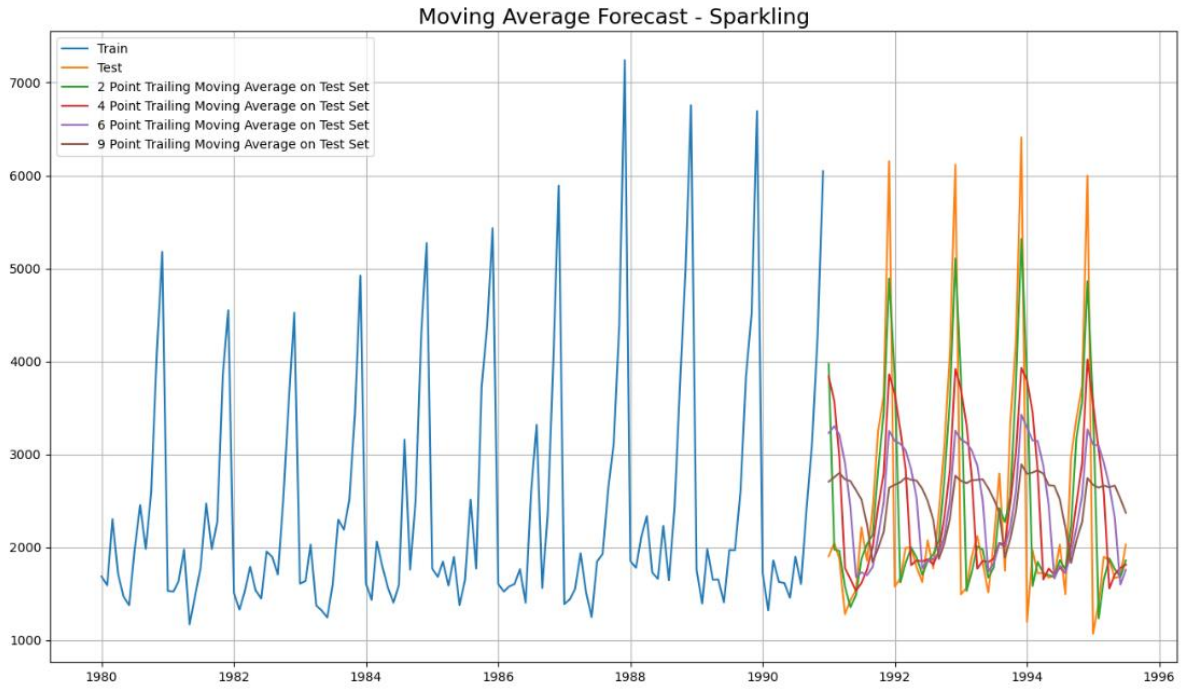


Figure 42 MOVING AVERAGE FORECAST - SPARKLING ON BOTH TRAIN AND TEST DATA

Observations:

1. **Sales Variability:**
    - The training data shows significant fluctuations in sparkling wine sales, with peaks over 6000 units and drops to around 1000 units, suggesting high variability likely driven by seasonal factors.
2. **Seasonal Peaks:**
    - Defined sales peaks indicate higher demand during specific months, likely holidays or celebrations, which can be targeted through focused marketing strategies.
3. **Moving Averages:**

- The 9-point moving average smooths short-term fluctuations, providing a clearer long-term trend, while shorter averages react more sensitively to variations.

4. **Decreasing Volatility:**
    - The test data shows less volatility, suggesting recent market stabilization while following similar trends to the training data.
5. **Trend Identification:**
    - A slight downward trend in the moving averages may indicate challenges in maintaining high sales, signaling the need for strategic marketing.
6. **Sales Forecasting:**
    - The moving averages offer useful insights for forecasting future sales, aiding businesses in production and inventory planning based on expected demand.
7. **Consumer Behavior Insights:**
    - The fluctuations reflect changing consumer behavior, potentially influenced by marketing, preferences, or economic conditions.

	Test RMSE Rose	Test RMSE Sparkling
2pointTrailingMovingAverage	11.529409	813.400684
4pointTrailingMovingAverage	14.455221	1156.589694
6pointTrailingMovingAverage	14.572009	1283.927428
9pointTrailingMovingAverage	14.731209	1346.278315

Figure 43 TEST RMSE FOR ROSE AND SPARKLING FOR 2,4,6,9 POINT TRAILING MOVING AVERAGE

Observations:

1. **2-Point Trailing Moving Average:**
    - The 2-point moving average has the lowest RMSE (813.40), indicating it offers the best predictive performance, as it reacts quickly to sales fluctuations.
2. **4-Point Trailing Moving Average:**
    - The 4-point moving average has a higher RMSE (1156.59), suggesting reduced accuracy as it smooths out more data points.
3. **6-Point and 9-Point Trailing Moving Averages:**
    - The 6-point (1283.93) and 9-point (1346.28) averages show progressively higher RMSEs, indicating that longer windows smooth the data too much, increasing prediction errors and failing to capture sudden sales changes.

4. **Trend:**
    - Shorter windows (2-point) provide more accurate, responsive forecasts, while longer windows (4, 6, 9 points) reduce sensitivity to short-term changes, leading to higher errors.

COMPARISON OF THE THREE ANALYSED MODELS

FOR ROSE WINE

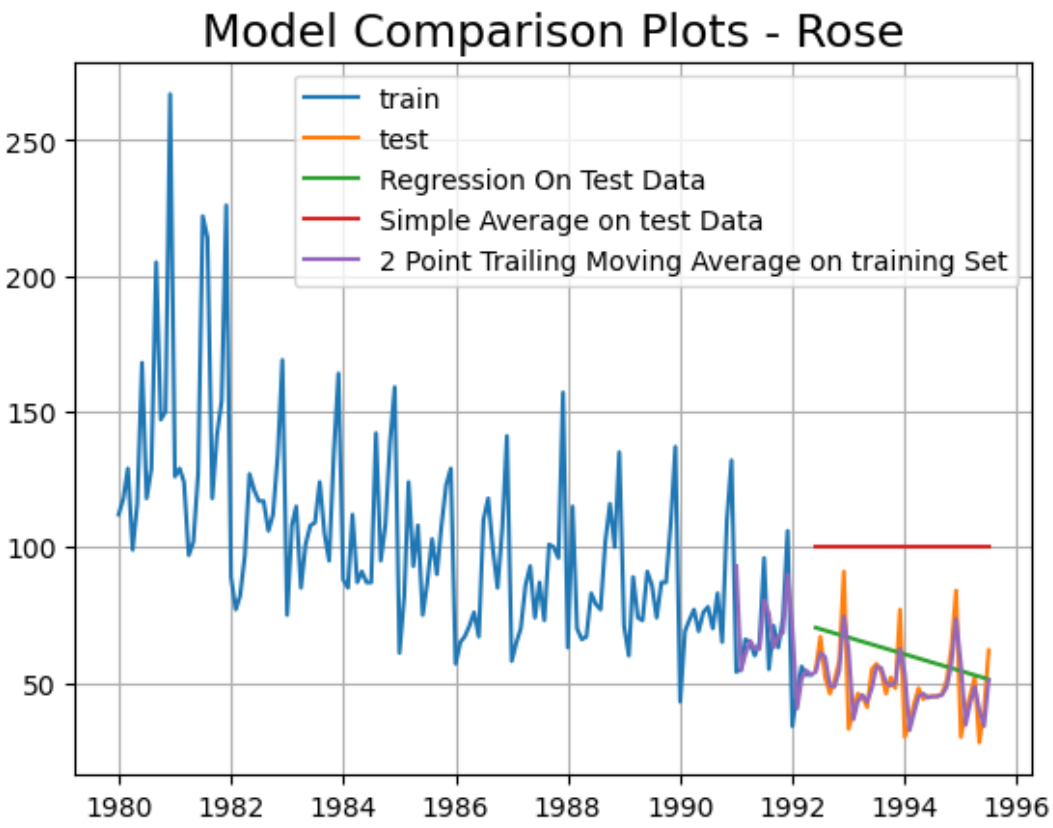


Figure 44 MODEL COMPARISON PLOT FOR ROSE

Rosé wine sales insights:

1. **Overall Sales Decline:**
    - Rosé wine sales show variability but peak around 250 units, with a general decline over time, indicating decreasing consumer interest.
2. **Lower Peaks and Stability:**
    - Unlike sparkling wine, rosé lacks sharp sales spikes, showing more stable but consistently lower sales.
3. **Moving Averages and Trends:**
    - The 2-point moving average highlights a downward trend, confirming the decline in rosé sales and suggesting a need for strategic marketing or rebranding.

4. **Limited Response to Seasonality:**
    - The stable, declining sales pattern suggests rosé doesn't benefit from seasonal marketing opportunities like sparkling wine.
5. **Test Data Insights:**
    - The test data stabilizes at a lower level, reinforcing the need for new strategies to stimulate sales growth.

FOR SPARKLING WINE:

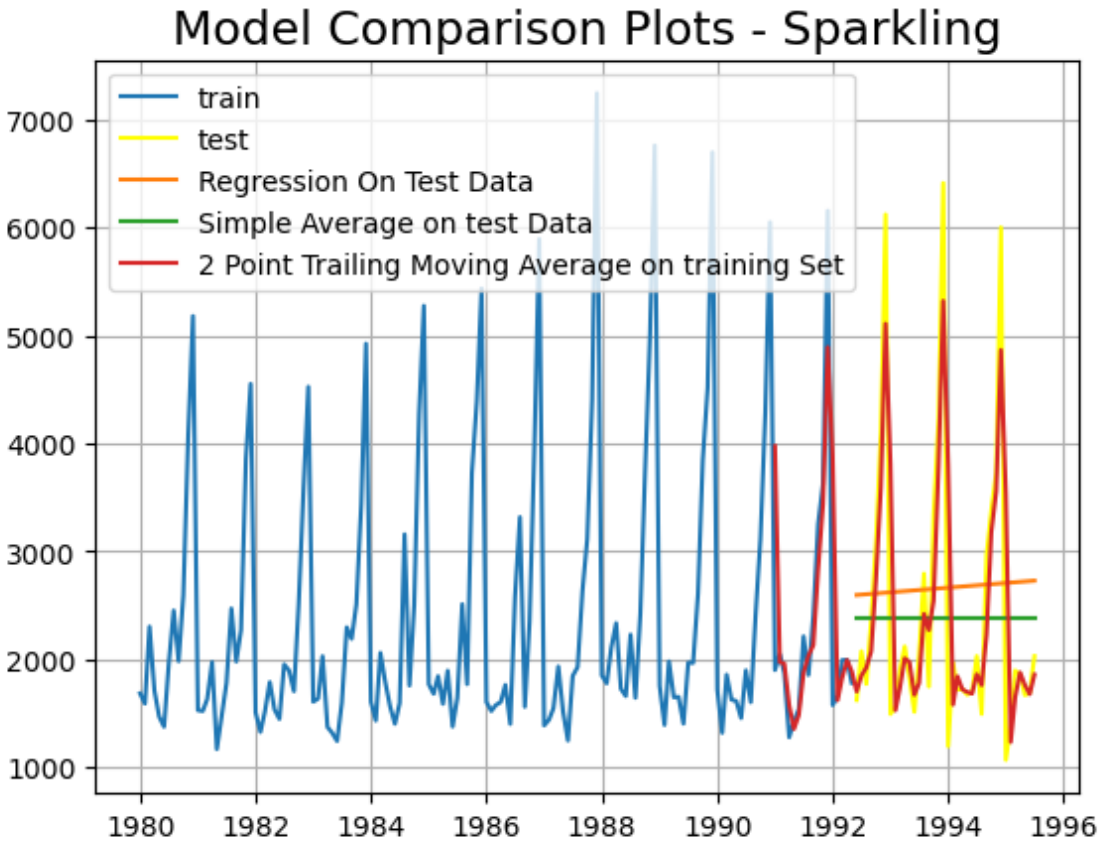


Figure 45 MODEL COMPARISON PLOT FOR SPARKLING

Sparkling wine sales insights:

1. **Sales Trends:**
    - The training data shows significant variability, with sales peaks often exceeding 6000 units, reflecting fluctuating consumer interest in sparkling wine.
2. **Seasonal Peaks:**
    - Regular sales spikes likely align with holidays and events, indicating seasonal demand that can be leveraged for targeted marketing.
3. **Moving Averages:**

- The 2-point moving average smooths short-term fluctuations, offering a clearer view of general sales trends over time.

4. **Test Data Stability:**
   - The test data shows less volatility, indicating potential stabilization in recent years while following the trends from the training data.
5. **Regression Analysis:**
   - The regression line suggests future sales trends, offering valuable insights for forecasting and strategic planning.

MODEL 4- EXPONENTIAL MODELS (SINGLE, DOUBLE, TRIPLE)

4.1 SIMPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS – ROSE

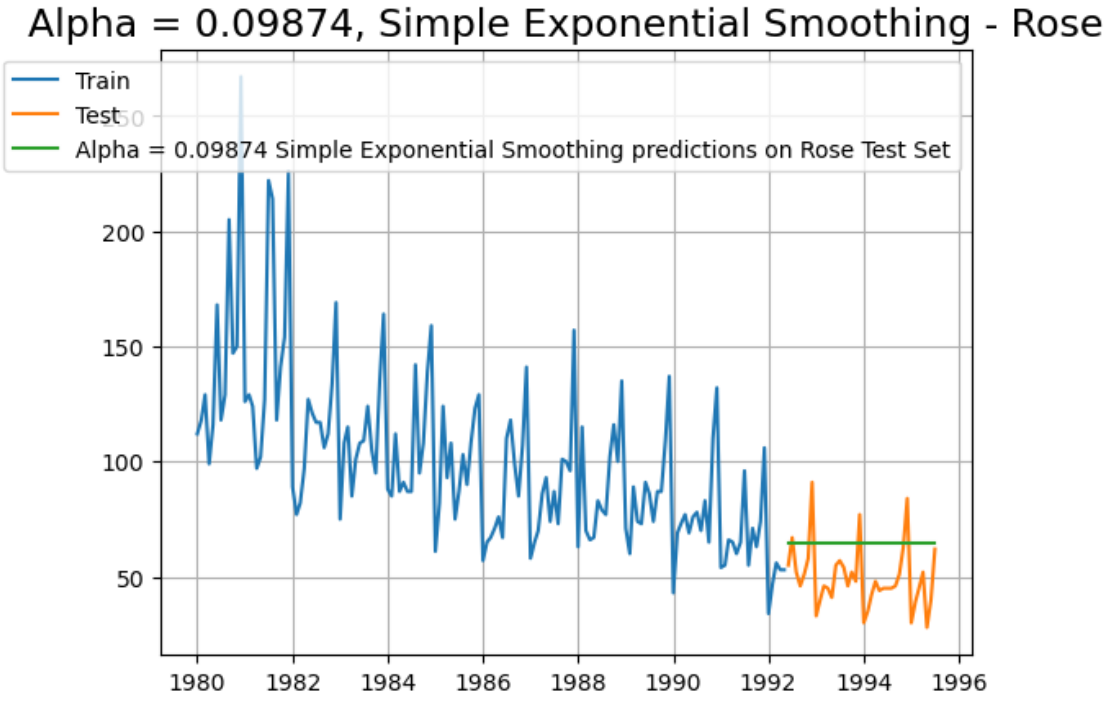


Figure 46 SES – ROSE

OBSERVATION:

1.Declining Sales:

- The training data shows a peak around 200 units, with an overall downward trend, indicating declining interest in rosé wine.

2. **Minimal Fluctuations:**

- Sales are more stable but lower compared to sparkling wine, reflecting a smaller, more consistent consumer base.

3. **Exponential Smoothing Results:**

- The predicted line (alpha = 0.09874) shows higher sensitivity to recent sales, reinforcing the decline and suggesting the need for more aggressive strategies to reverse this trend.

4. **Future Sales Outlook:**

- The test data remains flat at around 50 units, indicating a stagnating market and the need for intervention to boost sales.

	Test RMSE Rose
Simple Exponential Smoothing	20.313625

Figure 47 SES TEST RMSE -ROSE

4.2 SIMPLE EXPONENTIAL SMOOTHING WITH ADDITIVE ERRORS – SPARKLING

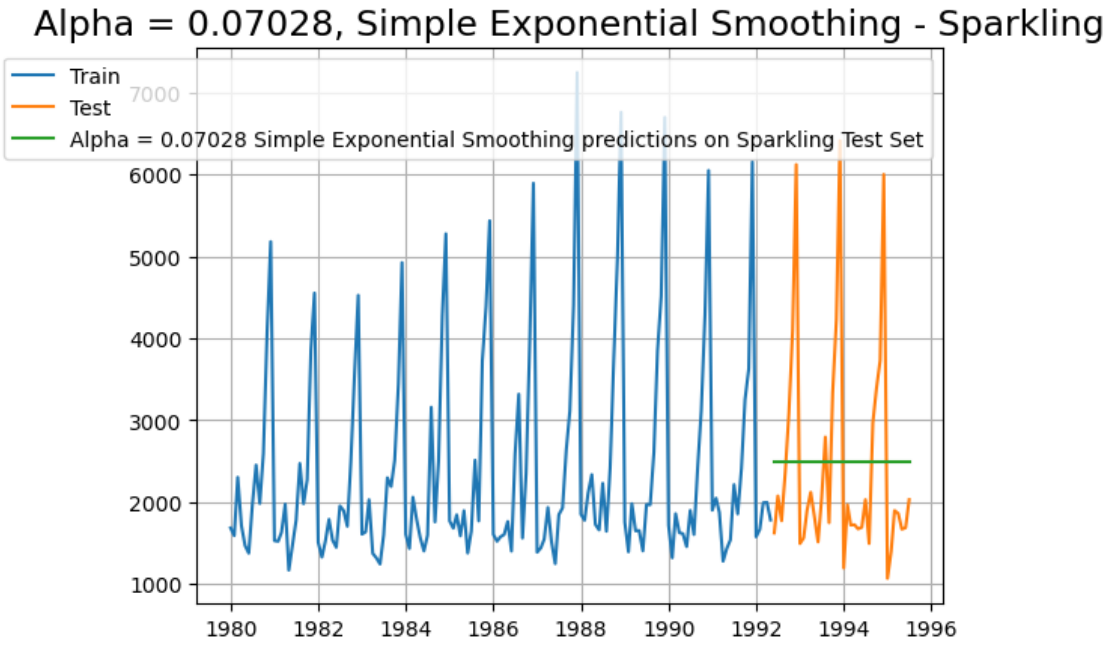


Figure 48 SES -SPARKLING

OBSERVATION:

1. **Sales Trends:**

- The training data shows significant variability, with sales peaks over 6000 units, indicating strong demand during specific periods.

2. **Seasonal Patterns:**

- The spikes align with festive occasions, suggesting that seasons strongly influence consumer purchasing behavior.

3. **Exponential Smoothing:**

- The predicted sales (green line, alpha = 0.07028) show a moderate adjustment to recent data, balancing responsiveness with stability.

4. **Testing Period and Stability:**
   - The test data shows less extreme fluctuations, indicating potential sales stabilization in recent years, with future sales expected to remain near an average level.

	Test RMSE Sparkling
Simple Exponential Smoothing	1329.926189

Figure 49 SES TEST RMSE - SPARKLING

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	17.510241	1349.042457
Simple Exponential Smoothing	20.313625	1329.926189

Figure 50 TEST RMSE FOR ROSE & SPARKLING FOR LINEAR REGRESSION AND SES

Observation:

1. **Simple Exponential Smoothing for Rosé:**
   - The model's RMSE of 20.31 indicates relatively low error, capturing the underlying trend and level well without significant bias.
2. **Comparison with Other Models:**
   - With an RMSE of 20.31, it outperforms the Simple Average Model (52.24) but is less accurate than the Regression On Time model (17.51), balancing simplicity and accuracy.
3. **Model Suitability:**
   - Simple Exponential Smoothing works well for data with minimal trends or seasonality, which likely applies to Rosé sales in your dataset.
4. **Error Significance:**
   - The RMSE suggests reasonable accuracy, but there's room for improvement, particularly if hidden seasonality or trends exist that the model doesn't capture.

4.3 DOUBLE EXPONENTIAL SMOOTHING WITH ADDITION ERRORS – ROSE

- A limitation of simple exponential smoothing is its poor performance in handling trended data.

- An enhanced version of SES, called the Double Exponential model, addresses this issue by incorporating two smoothing parameters.
- It is suitable for datasets exhibiting trend but lacking seasonality.
- The model distinguishes between two main components: Level and Trend.
- The Level represents the local mean, while the Trend captures directional changes.
- The smoothing parameter α governs the Level series, while β controls the Trend series.

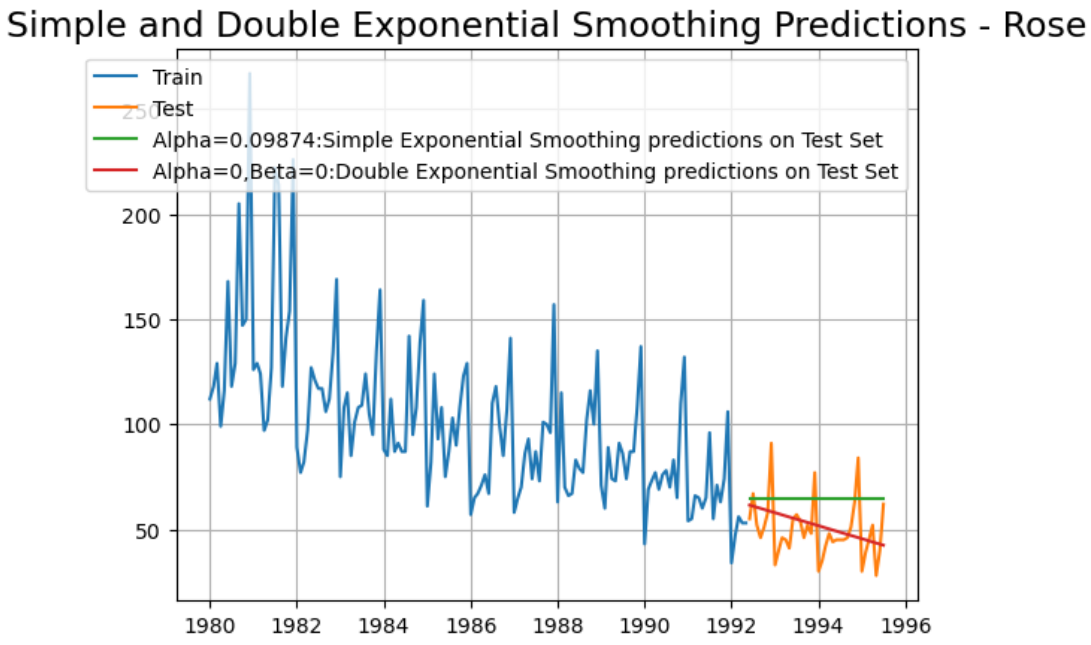


Figure 51 SES & DES -ROSE

OBSERVATION:

1. **Overall Decline:**
    - The training data shows a flat trend with occasional peaks around 250 units, but overall, sales have been declining over time.
2. **Minimal Sales Fluctuations:**
    - Rosé wine sales show less variability compared to sparkling wine, indicating stable but lower consumer interest.
3. **Simple Exponential Smoothing:**
    - The green line (alpha = 0.09874) predicts a slight downward trend, reflecting ongoing challenges in the rosé wine market.
4. **Double Exponential Smoothing Predictions:**
    - The red line for double exponential smoothing is less effective due to the flat trend, showing minimal growth or decline.
5. **Market Challenges:**

- Predictions suggest stagnant sales for rosé wine, requiring strategic marketing or product innovations to boost demand.

	Test RMSE Rose
Double Exponential Smoothing	14.623742

Figure 52 DES -TEST RMSE- ROSE

4.4 DOUBLE EXPONENTIAL SMOOTHING WITH ADDITION ERRORS - SPARKLING

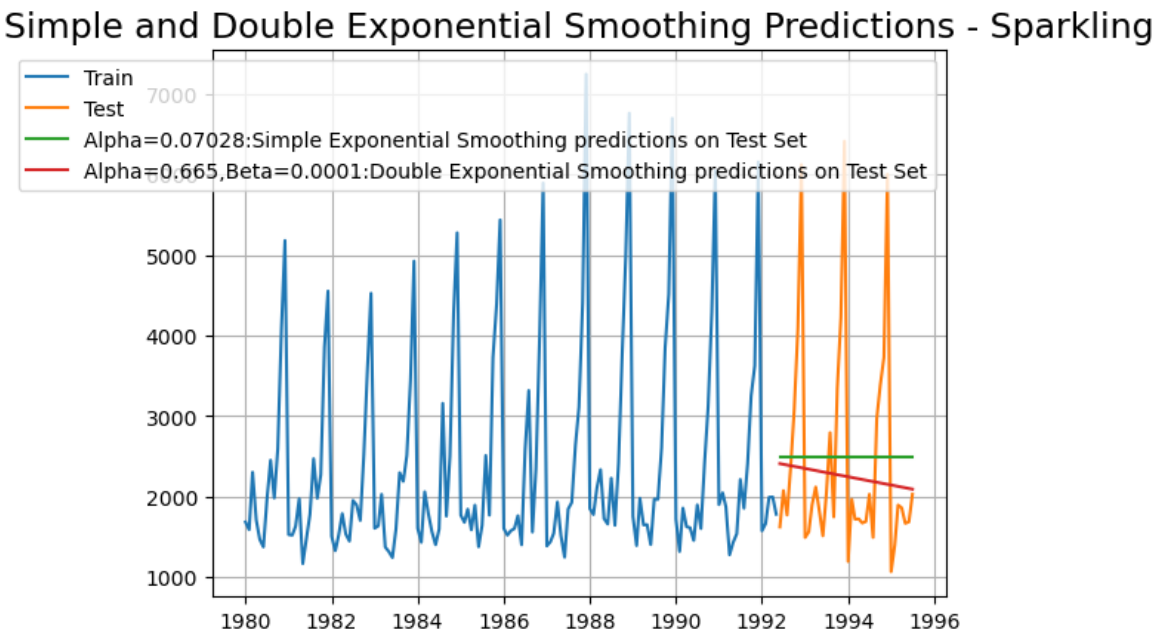


Figure 53 SES & DES FOR SPARKLING

OBSERVATION:

1. **Sales Trends:**

- The training data shows significant fluctuations, with peaks exceeding 6000 units, reflecting strong demand during certain periods.

2.**Seasonal Demand:**

- Seasonal spikes suggest ties to holidays and events, indicating opportunities for targeted marketing during these times.

3.**Simple Exponential Smoothing:**

- The green line (alpha = 0.07028) provides predictions that smooth out fluctuations, offering a moderately stable outlook despite variations.

4.**Double Exponential Smoothing:**

- The red line (alpha = 0.665, beta = 0.0001) offers a better fit, accounting for trends and providing a more responsive prediction.

5.**Future Outlook:**

- Double exponential smoothing suggests a clearer stabilization trend, while the simple model reacts more to short-term fluctuations, offering insights for future sales forecasting.

	Test RMSE Rose	Test RMSE Sparkling
Double Exponential Smoothing	14.623742	1340.452791

observations:

1. **Double Exponential Smoothing for Rosé:**
    - The RMSE for Rosé (14.62) is lower than both Simple Exponential Smoothing (20.31) and RegressionOnTime (17.51), indicating better performance due to its ability to capture both level and trend.
2. **Double Exponential Smoothing for Sparkling:**
    - The RMSE for Sparkling (1340.45) is slightly lower than RegressionOnTime (1349.04) and Simple Average Model (1331.04), but still high, suggesting the model doesn't effectively capture the volatility and complexity of Sparkling sales.
3. **Model Performance:**
    - Double Exponential Smoothing works well for Rosé due to its linear trend, but for Sparkling, more advanced models may be needed to account for seasonality and other complexities.
4. **Trend and Seasonality:**
    - Double Exponential Smoothing is suited for linear trends (as seen in Rosé), but may not handle non-linear patterns or seasonality, which might explain its limited effectiveness for Sparkling.

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	17.510241	1349.042457
Simple Exponential Smoothing	20.313631	1329.402402
Double Exponential Smoothing	14.623742	1340.452791

Figure 54 COMPARISON OF TEST RMSE FOR ROSE AND SPARKLING AMONG LINEAER REGRESSION, SES & DES

Inference

In this observation, it's evident that Double Exponential Smoothing outperforms Simple Exponential Smoothing. This superiority stems from Double Exponential Smoothing's capability to capture the trend component effectively.

4.5 TRIPLE EXPONENTIAL SMOOTHING WITH ADDITION ERRORS - ROSE

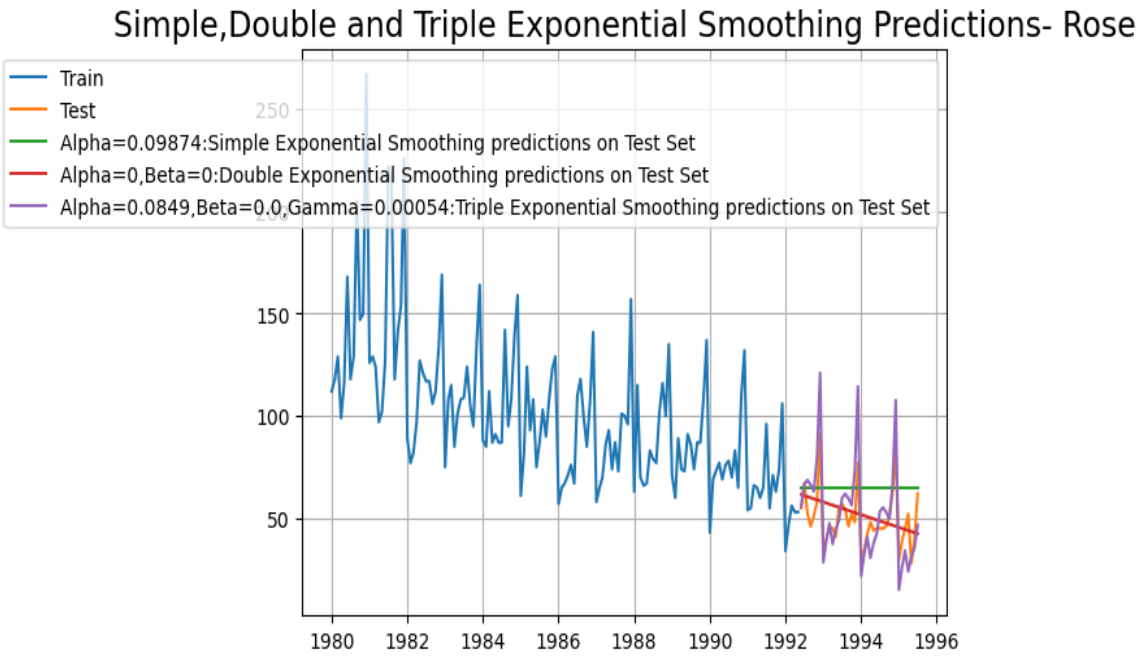


Figure 55 SES, DES & TES - ROSE

Rosé wine insights:

1. **Sales Decline:**
    - Rosé wine sales show a downward trend, peaking at around 250 units but generally declining over time.
2. **Smoothing Models:**
    - **Simple Exponential Smoothing** (green line, alpha = 0.09874): Captures the general sales decline.
    - **Double Exponential Smoothing** (red line, alpha = 0.0, beta = 0.0): Produces flat predictions due to lack of trend responsiveness.
    - **Triple Exponential Smoothing** (purple line, alpha = 0.0849, beta = 0.0, gamma = 0.00054): Shows minimal seasonal adjustment and reflects a stable but declining trend.
3. **Future Sales Outlook:**
    - Predictions suggest sales will remain low (around 50-70 units), indicating the need for strategic marketing changes.

4. **Strategic Opportunities:**
    - Innovation and promotions may help revive rosé sales, as the market appears stagnant.

	Test RMSE Rose
Triple Exponential Smoothing (Additive Season)	13.877335

Figure 56 TES ADDITIVE SEASON- TEST RMSE-ROSE

Triple Exponential Smoothing has performed the best on the test as expected since the data had both trend and seasonality.

4.6 TRIPLE EXPONENTIAL SMOOTHING WITH ADDITION ERRORS – SPARKLING

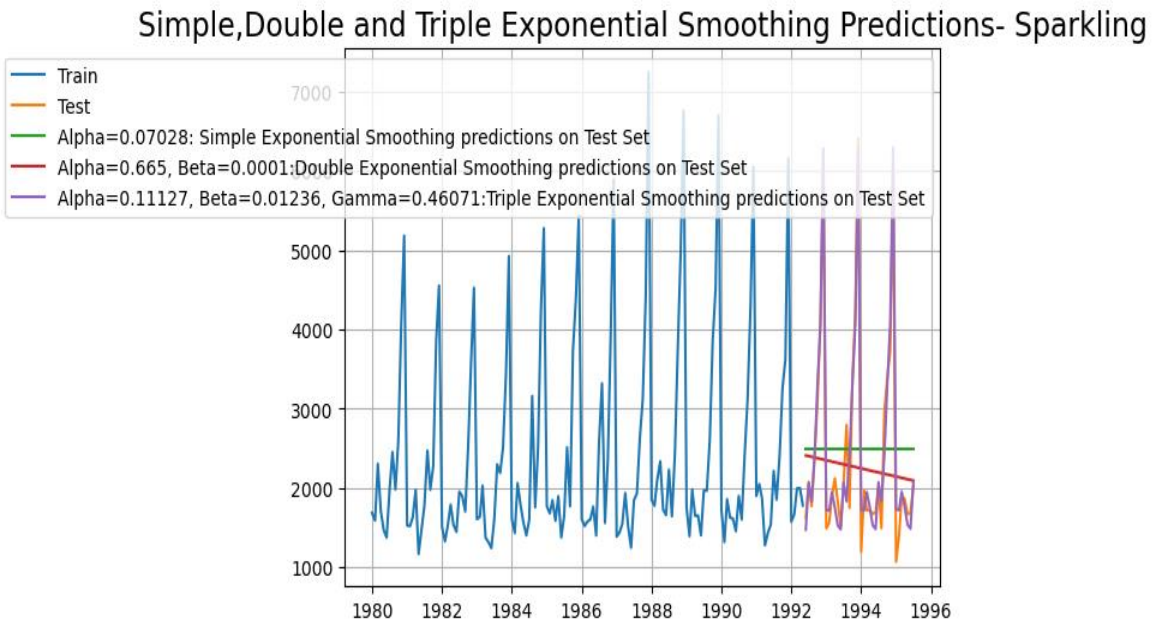


Figure 57 SES, DES, AND TES ON TEST SET -SPARKLING

Sparkling wine insights:

1. **Sales Variability:**
    - Sales show considerable fluctuations, with peaks often surpassing 6000 units, indicating strong seasonal demand.
2. **Exponential Smoothing Models:**
    - **Simple Exponential Smoothing** (green line, alpha = 0.07028): Smooths data but may not capture trends well.

- **Double Exponential Smoothing** (red line, alpha = 0.665, beta = 0.0001): Accounts for trends and shows gradual market stabilization.
- **Triple Exponential Smoothing** (purple line, alpha = 0.11127, beta = 0.01236, gamma = 0.46071): Best for capturing seasonal effects, offering the most accurate predictions.

3. **Future Sales Stability:**
   - Triple Exponential Smoothing suggests sales will stabilize around 2500-3000 units, aiding in forecasting.
4. **Seasonal Effects:**
   - Sales peaks align with holidays and events, indicating strong opportunities for targeted marketing.

	Test RMSE Sparkling
Triple Exponential Smoothing (Additive Season)	304.247029

Figure 58 TES ADDITIVE SEASON - TEST RMSE -SPARKLING

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	17.510241	1349.042457
Simple Exponential Smoothing	20.313631	1329.402402
Double Exponential Smoothing	14.623742	1340.452791
Triple Exponential Smoothing (Additive Season)	13.877335	304.247029

Figure 59 COMPARISON OF TEST RMSE FOR ROSE AND SPARKLING AMONG LINEAER REGRESSION, SES ,DES & TES (ADDITIVE SEASON)

Observation for triple exponential smoothing additive season:

1. **Double Exponential Smoothing for Rosé:**
   - RMSE of 14.62 indicates good performance, outshining Simple Exponential Smoothing (20.31) and RegressionOnTime (17.51). This suggests the model captures both trend and level effectively.
2. **Double Exponential Smoothing for Sparkling:**
   - RMSE of 1340.45, slightly better than RegressionOnTime (1349.04), but still high, suggesting that the model struggles with the volatility and seasonality in Sparkling sales.
3. **Performance Comparison:**
   - For Rosé, Double Exponential Smoothing significantly improves forecast accuracy by adjusting both trend and level.

- For Sparkling, the model provides limited improvements, indicating that more complex models like Holt-Winters or ARIMA may be needed to handle seasonality and volatility.

4. **Trend Sensitivity:**
    - Double Exponential Smoothing works well for linear trends (Rosé) but struggles with datasets featuring volatility or non-linear patterns (Sparkling).

4.7 TAKING MULTIPLICATIVE SEASONALITY- ROSE

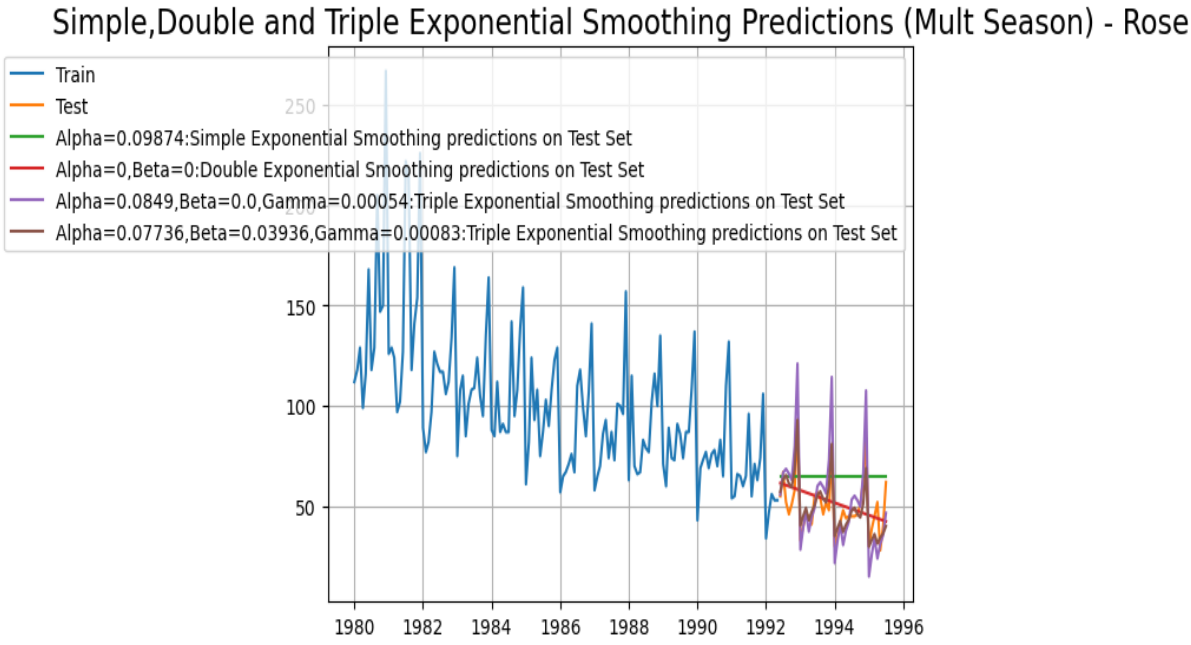


Figure 60 SES, DES & TES (MULTIPLICATIVE SEASONALITY) -ROSE

Rosé Wine Insights

1. Sales Decline:
    - Rosé wine sales exhibit a downward trend over time, with peaks reaching approximately 250 units, signaling waning consumer interest.
2. Prediction Models:
    - Simple Exponential Smoothing (green line):
        - Captures the downward trend but lacks sensitivity to fluctuations.
    - Double Exponential Smoothing (red line):
        - With no adjustments for trend (Alpha = 0, Beta = 0), it reflects a stagnant market.
    - Triple Exponential Smoothing (purple line):

- Incorporates minor seasonal adjustments (Alpha = 0.07736, Beta = 0.03936, Gamma = 0.00083), but reinforces the declining sales pattern.

3. Market Stagnation:
    - Flat predictions toward the end of the test period suggest a plateau, underscoring challenges in reviving interest.
4. Call for Revitalization:
    - Reversing the declining trend requires innovative marketing strategies and product differentiation to reinvigorate consumer demand.

	Test RMSE Rose
Triple Exponential Smoothing (Multiplicative Season)	8.405441

Figure 61 TEST RMSE - ROSE FOR TES(MULTIPLICATIVE)

We see that the multiplicative seasonality model has not done that well when compared to the additive seasonality Triple Exponential Smoothing model.

4.8 TAKING MULTIPLICATIVE SEASONALITY- SPARKLING

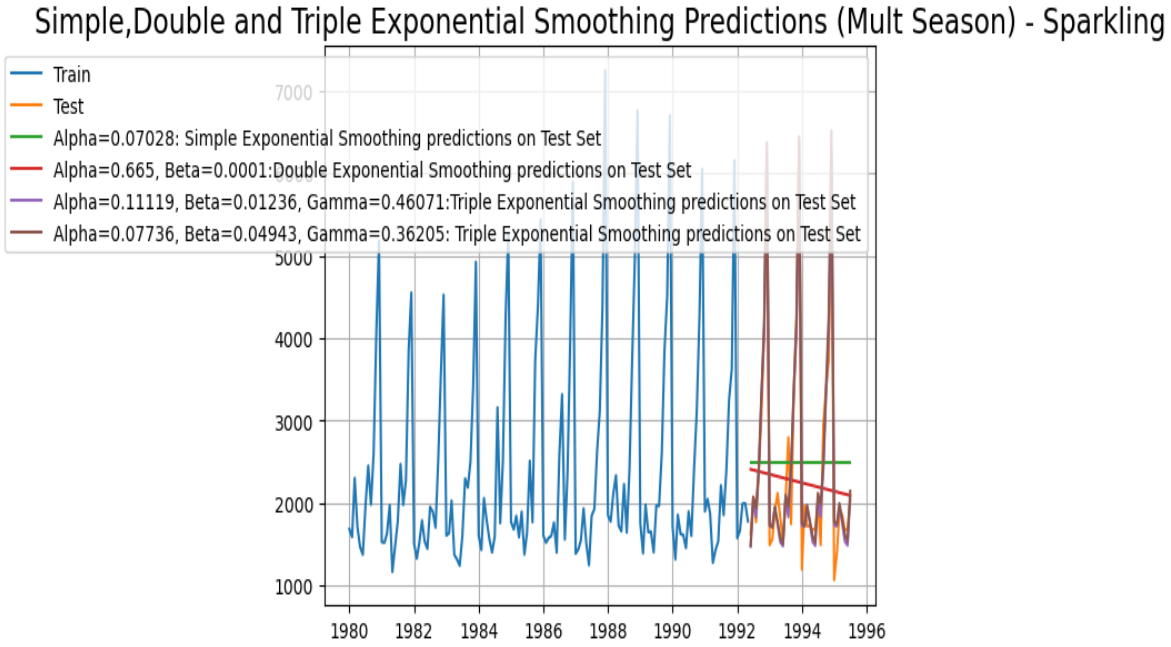


Figure 62 SES, DES, TES(MULTIPLICATIVE)-SPARKLING

Sparkling Wine Insights

1. Sales Fluctuations:
    - The data reveals significant seasonal fluctuations, with peaks exceeding 6,000 units, highlighting strong demand during specific periods.

2. Prediction Models:
    - Simple Exponential Smoothing (green line):
        - Smooths the data but lacks the ability to capture evolving trends (Alpha = 0.07028).
    - Double Exponential Smoothing (red line):
        - Incorporates trends (Alpha = 0.665, Beta = 0.0001), offering a more responsive and stable prediction.
    - Triple Exponential Smoothing (purple line):
        - Effectively models seasonal variations (Alpha = 0.11119, Beta = 0.01236, Gamma = 0.46071), providing the most detailed and accurate future outlook.
3. Future Sales Trends:
    - Predictions indicate potential stabilization around 2,500–3,000 units as the test period progresses.
4. Strategic Marketing Opportunities:
    - Strong seasonal patterns present opportunities for targeted marketing campaigns during peak demand periods to maximize sales impact.

	Test RMSE Rose	Test RMSE Sparkling
Triple Exponential Smoothing (Multiplicative Season)	8.405441	318.695471

Figure 63 TEST RMSE FOR ROSE AND SPARKLING - TES (MULTIPLICATIVE)

Observations

1. Triple Exponential Smoothing (Multiplicative Seasonality) for Rosé:
    - The model achieves the lowest RMSE (8.41) compared to other approaches, including:
        - Simple Exponential Smoothing (RMSE: 20.31)
        - Double Exponential Smoothing (RMSE: 14.62)
        - Triple Exponential Smoothing with Additive Seasonality (RMSE: 13.88).
    - This indicates that Multiplicative Seasonality effectively captures proportional sales fluctuations, making it the best model for Rosé.
2. Triple Exponential Smoothing (Multiplicative Seasonality) for Sparkling:

- The RMSE (318.70) is slightly higher than Additive Seasonality (RMSE: 304.25), but still significantly better than:
    - Regression on Time (RMSE: 1349.04)
    - Double Exponential Smoothing (RMSE: 1340.45).
- The model demonstrates good performance, capturing the scaling nature of seasonal fluctuations in Sparkling sales.

3. Performance Comparison:
    - Rosé: Multiplicative seasonality outperforms all other models due to its ability to account for proportional seasonal fluctuations.
    - Sparkling: While multiplicative seasonality improves seasonal modeling compared to additive seasonality, the higher RMSE suggests that Sparkling data exhibits more complex seasonal behaviors.
4. Multiplicative vs. Additive Seasonality:
    - Multiplicative Seasonality: Best suited for datasets like Rosé and Sparkling, where seasonal fluctuations grow with the data's level.
    - Additive Seasonality: Suitable for cases with constant seasonal variations regardless of data level, which is less relevant for both Rosé and Sparkling in this analysis.

Key Insight:

Multiplicative seasonality is the preferred approach for both Rosé and Sparkling wine datasets, particularly for Rosé, as it better models the proportional and scaling nature of seasonal variations. However, further refinement may be needed for Sparkling due to its greater complexity.

4.9 CHECK THE PERFORMANCE OF THE MODELS BUILT

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	17.510241	1349.042457
Simple Exponential Smoothing	20.313631	1329.402402
Double Exponential Smoothing	14.623742	1340.452791
Triple Exponential Smoothing (Additive Season)	13.877335	304.247029
SimpleAverageModel	52.239499	1331.037637
2pointTrailingMovingAverage	11.529409	813.400684
4pointTrailingMovingAverage	14.455221	1156.589694
6pointTrailingMovingAverage	14.572009	1283.927428
9pointTrailingMovingAverage	14.731209	1346.278315
Triple Exponential Smoothing (Multiplicative Season)	8.405441	318.695471

Figure 64 PERFORMANCE OF THE MODEL BUILT

Best Models

1. Rosé Wine:
    - Triple Exponential Smoothing (Multiplicative Seasonality): Lowest RMSE (8.41), making it the best model for forecasting. Captures proportional seasonal fluctuations and trend effectively.
    - 2-Point Trailing Moving Average: A competitive alternative for short-term forecasting with an RMSE of 11.53.
2. Sparkling Wine:
    - Triple Exponential Smoothing (Additive Seasonality): Lowest RMSE (304.25), indicating it's the most accurate model for capturing seasonality and trend despite the data's complexity.
    - 2-Point Trailing Moving Average: A viable alternative for more volatile trends with an RMSE of 813.40.

Models with High RMSE

3. Simple Average Model:
    - Produces the worst RMSE (52.24 for Rosé, 1331.04 for Sparkling), failing to capture trends, seasonality, or data fluctuations.
4. Trailing Moving Averages:

- Sparkling Wine: As the window size increases (e.g., 2-Point to 9-Point), RMSE rises, smoothing out critical fluctuations.
- Rosé Wine: 4-Point Trailing Moving Average performs well (RMSE: 14.46), balancing smoothness and responsiveness.

Intermediate Model Performance

5. Double Exponential Smoothing:
    - Performs moderately for both datasets (RMSE: 14.62 for Rosé, 1340.45 for Sparkling), capturing trends but less effective with seasonality compared to Triple Exponential models.

Summary

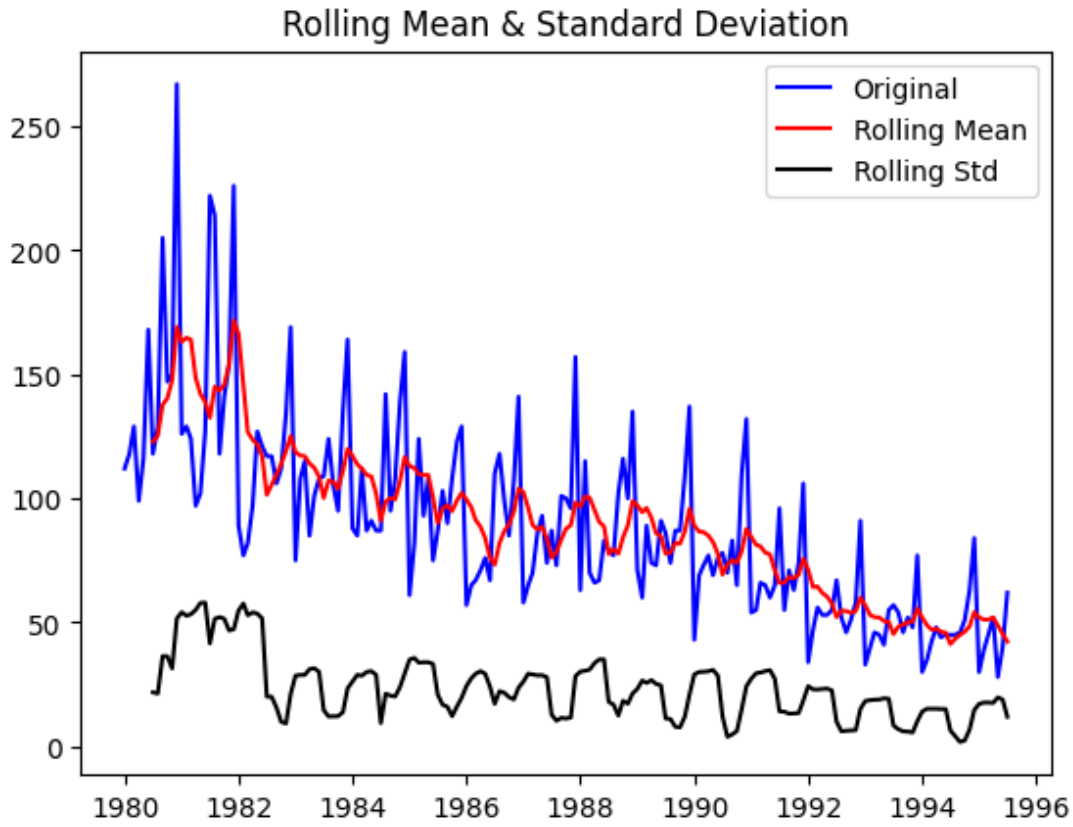
- Rosé: Triple Exponential Smoothing (Multiplicative Seasonality) is the best model, with 2-Point Trailing Moving Average as a secondary option for short-term forecasts.
- Sparkling: Triple Exponential Smoothing (Additive Seasonality) provides the best accuracy, with 2-Point Trailing Moving Average as a competitive alternative for handling volatility.

D. Model Building - Stationary Data

1. CHECK STATIONARITY OF ROSE DATA

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

***H*0 : The Time Series has a unit root and is thus non-stationary.; *H*1 : The Time Series does not have a unit root and is thus stationary.**



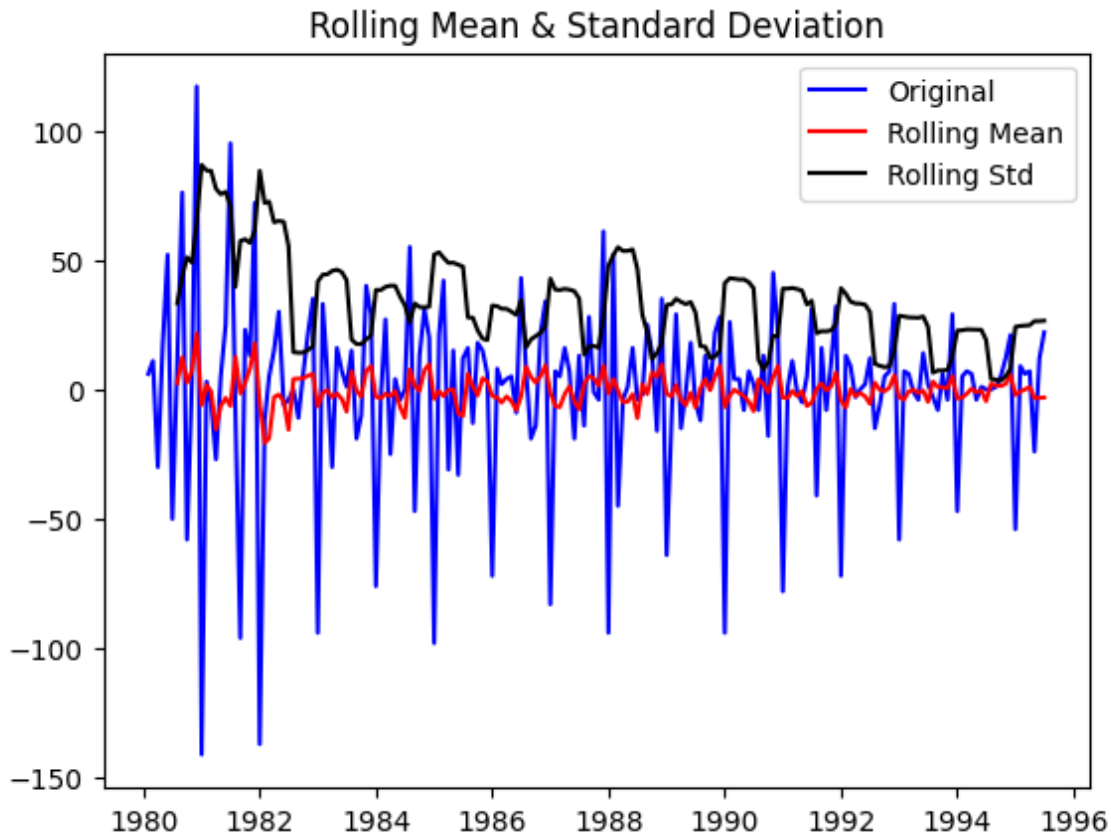

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.874856
p-value                          0.343981
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

Figure 65 DICKEY- FULLER TEST - ROSE

ADF Test Results - Interpretation

1. Test Statistic: -1.875 (greater than critical values at 1%, 5%, and 10% significance levels).
2. p-value: 0.344 (greater than 0.05), indicating we fail to reject the null hypothesis (H0H_0H0).
3. Conclusion: The series is non-stationary and likely has a unit root. Transformation (e.g., differencing) is required to achieve stationarity.

To solve this issue, we'll apply a single level of difference to determine, if the series becomes stationary.



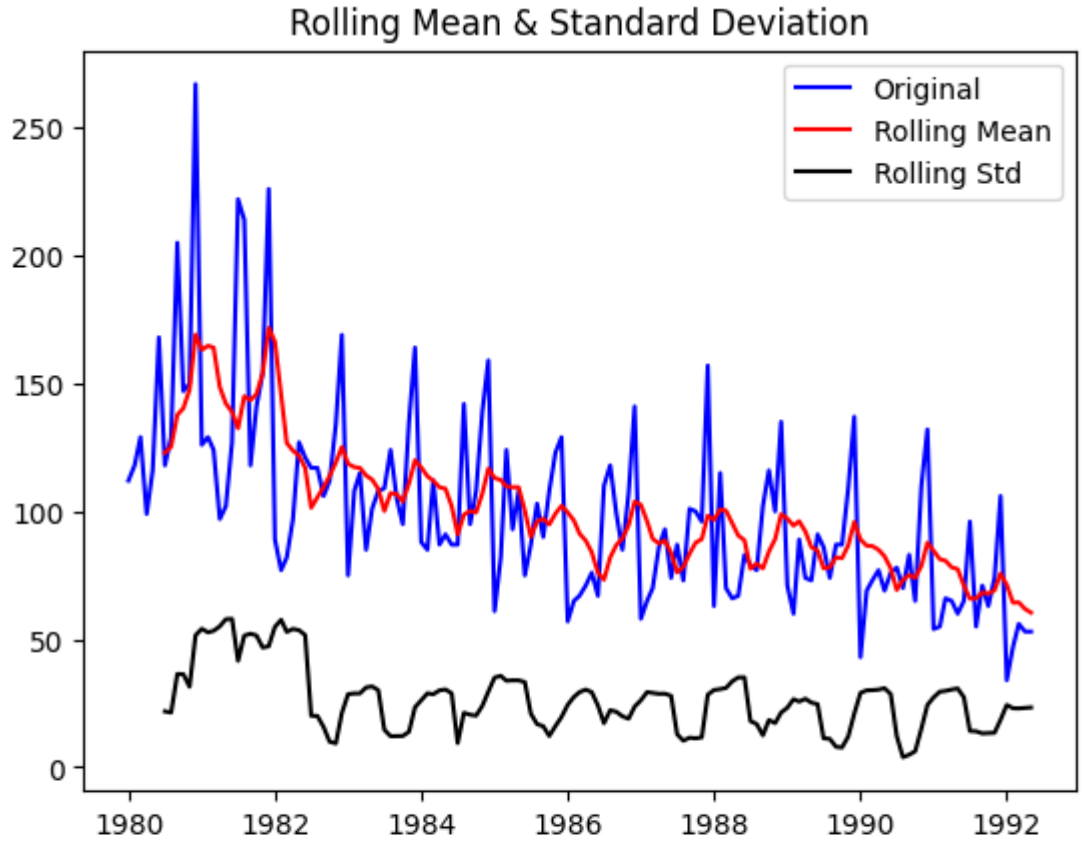

```
Results of Dickey-Fuller Test:
Test Statistic                -8.044139e+00
p-value                        1.813580e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

Figure 66 DIFFERENCED (1) ADF TEST - ROSE

Updated ADF Test Results - Interpretation

1. Test Statistic: -8.0441 (much lower than all critical values at 1%, 5%, and 10% significance levels).
2. p-value: 1.81e-12 (much smaller than 0.05), allowing us to reject the null hypothesis (H0H_0H0).
3. Conclusion: The series is now stationary, does not have a unit root, and is ready for modeling or forecasting without additional transformations.

1.1 CHECK FOR STATIONARITY OF THE TRAINING DATA TIME SERIES-ROSE



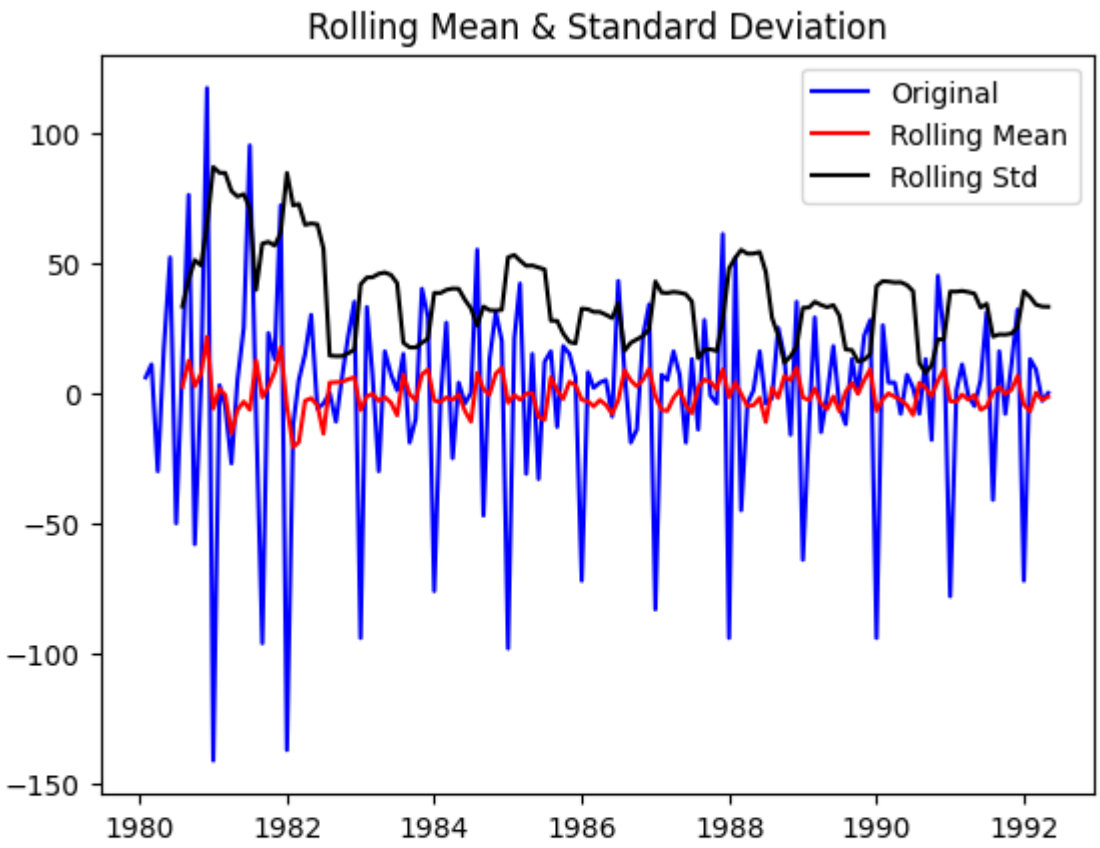

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.649379
p-value                          0.457364
#Lags Used                      13.000000
Number of Observations Used    135.000000
Critical Value (1%)             -3.479743
Critical Value (5%)             -2.883198
Critical Value (10%)            -2.578320
dtype: float64
```

Figure 67 ADF TEST FOR TRAINING DATA TIME SERIES-ROSE

We see that the series is not stationary at α= 0.05

ADF Test Results - Interpretation

1. Test Statistic: -1.6494 (greater than critical values at 1%, 5%, and 10% significance levels).
2. p-value: 0.4574 (greater than 0.05), indicating we fail to reject the null hypothesis (H0H_0H0).
3. Conclusion: The time series is non-stationary and likely contains a unit root.




```
Results of Dickey-Fuller Test:
Test Statistic                -7.132106e+00
p-value                        3.496102e-10
#Lags Used                     1.200000e+01
Number of Observations Used    1.350000e+02
Critical Value (1%)           -3.479743e+00
Critical Value (5%)           -2.883198e+00
Critical Value (10%)          -2.578320e+00
dtype: float64
```

Figure 68 DIFFERENCED (1) ADF TEST - TRAINING DATA TIME SERIES- ROSE

The ADF test results indicate that the time series is stationary:

- **Test Statistic**: -7.1321, which is much lower than the critical values at all significance levels.
- **p-value**: 3.50e-10, which is extremely small and well below 0.05 or 0.01, leading to strong rejection of the null hypothesis.
- **Critical Values**: The test statistic is significantly lower than the critical values at 1%, 5%, and 10% significance levels.

Conclusion: With the p-value and test statistic both supporting it, we reject the null hypothesis and confirm that the time series is stationary.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 149 entries, 1980-01-01 to 1992-05-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    149 non-null    float64
dtypes: float64(1)
memory usage: 2.3 KB
```

Figure 69 TRAINED DATA INFORMATION -ROSE

Key Points:

- **DatetimeIndex**: The data is time series with a DatetimeIndex.
- **149 Entries**: There are 149 data points.
- **Non-null Count**: No missing values in the "Rose" column.
- **Column Type**: The "Rose" column is of type float64 (numeric).
- **Memory Usage**: The dataset is small (2.3 KB), making it easy to process.

2 CHECK THE ACF AND PACF OF THE TRAINING DATA-ROSE

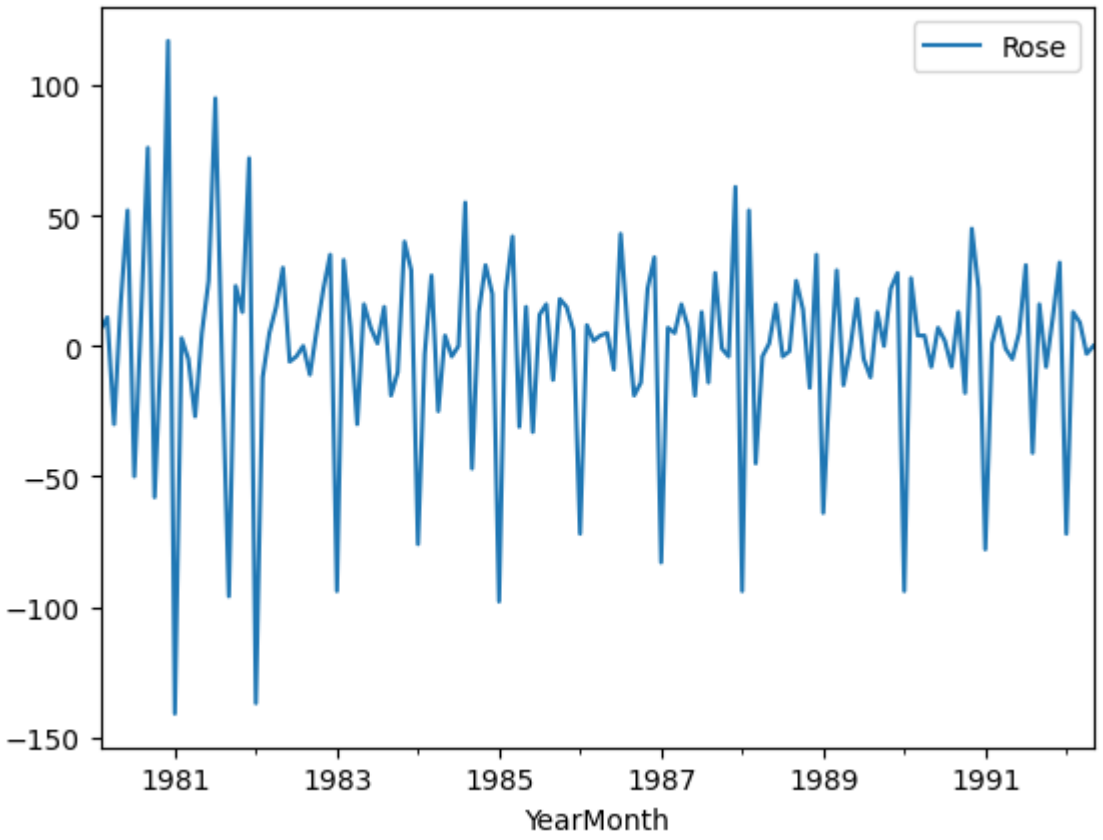


Figure 70 TRAINED DATA PLOT -ROSE

Insights from Rosé Wine Sales Data:

1. **Sales Variability**: Significant fluctuations in sales from 1981 to 1991, indicating inconsistent consumer interest.
2. **Cycle Characteristics**: Peaks and troughs suggest cyclical behavior, possibly linked to seasonal or event-driven factors.
3. **Negative Values**: Negative sales values may indicate stock issues, returns, or market disruptions.
4. **Lack of Clear Trend**: No clear upward or downward trend, implying a stable but unexciting market.
5. **Potential for Improvement**: Fluctuations present opportunities to stabilize sales through targeted marketing or increased product visibility.
6. **Market Dynamics**: Variability could reflect changes in consumer preferences or market competition, highlighting the need for deeper market analysis.

2.1 GENERATE ACF & PACF PLOT and FINDING THE AR, MA VALUES:

Use Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to identify potential values for:

p (AR order): Based on PACF plot.

d (Differencing order): Based on stationarity checks.

q (MA order): Based on ACF plot.

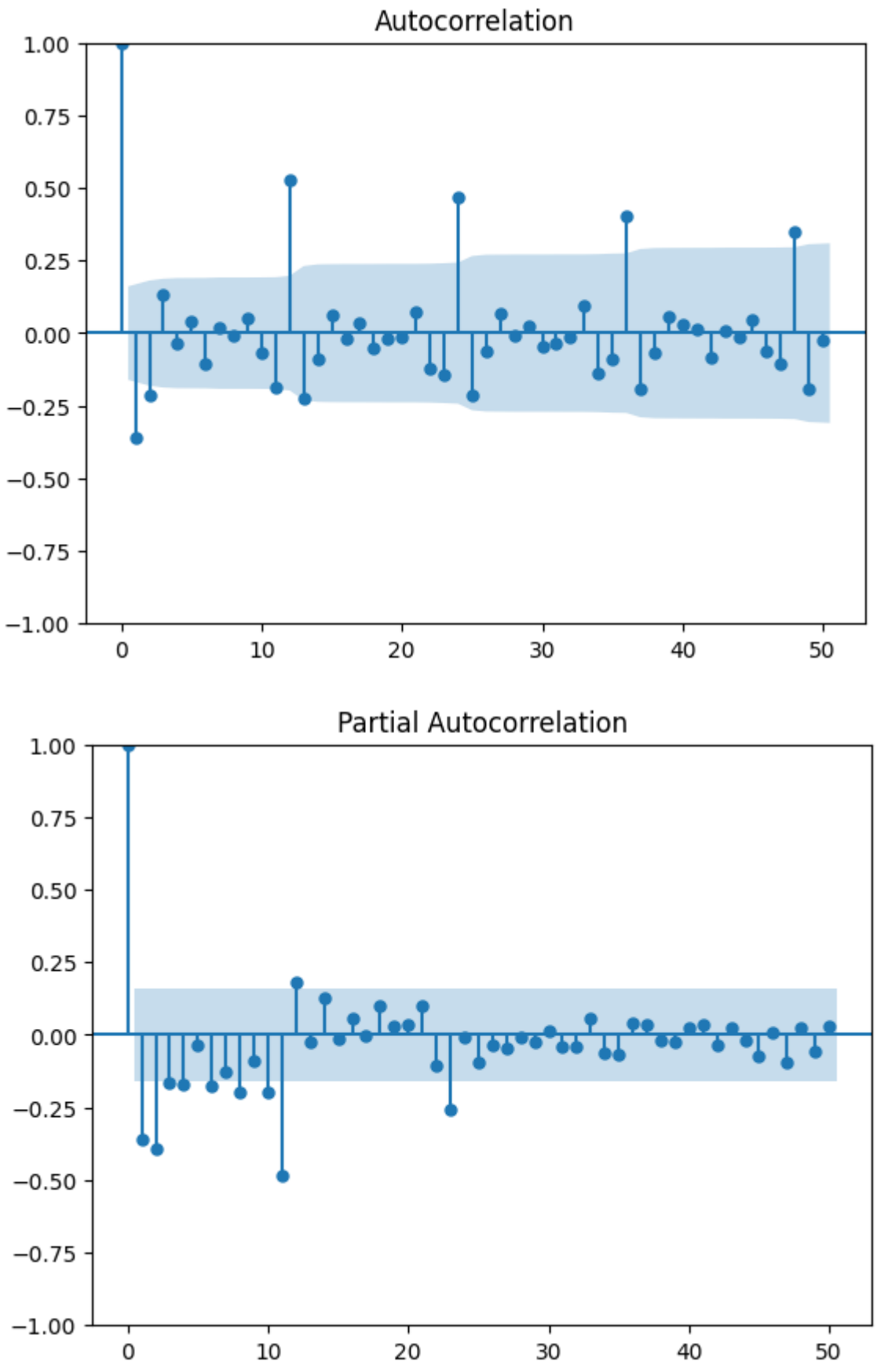


Figure 71 ACF & PACF PLOT

Model Selection Insights:

- AR(p) Model: PACF suggests AR(1) or AR(2), as lag 1 shows significant correlation.
    - p = 1 or 2 (due to significant decay at lag 1).
- MA(q) Model: ACF shows strong autocorrelation at lag 1, with significant peaks at lags 3, 6, 12, and 24.
    - q = 1 or 2 (for capturing autocorrelation at these lags).
- Differencing (d): Based on ADF test, if stationary, d = 0; if non-stationary, d = 1.
- Seasonal Components (P, Q, D): ACF indicates seasonal patterns at lags 12 and 24.
    - P = 1 (for seasonal autoregressive component at lag 12).
    - Q = 2 (to capture seasonal moving averages at lags 12 and 24).
    - D = 1 (for seasonal differencing if needed).

suggested Model Parameters:

- ARIMA(1,0,1) or ARIMA(2,0,2) (non-seasonal)
- SARIMA(1,0,1)(1,0,2,12) or SARIMA(2,0,2)(1,0,2,12) (seasonal with yearly periodicity).

3. BUILDS DIFFERENT ARIMA MODELS

- Auto ARIMA

- Manual ARIMA

Build an Automated version of an ARMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC)- ROSE:

Acknowledging the seasonality in the data, it's indeed prudent to consider a SARIMA model. However, before proceeding, let's test whether an ARIMA or SARIMA model better fits the data by comparing their Akaike Information Criteria (AIC) values. We'll choose the model with the lowest AIC as the preferred option.

	param	AIC
8	(2, 0, 2)	1448.974415
5	(1, 0, 2)	1449.505125
7	(2, 0, 1)	1450.457007
4	(1, 0, 1)	1451.902141
6	(2, 0, 0)	1466.844586
3	(1, 0, 0)	1467.277579
1	(0, 0, 1)	1474.472163
2	(0, 0, 2)	1474.768373
0	(0, 0, 0)	1501.769377

Figure 72 SORTED PARAMETER BASED ON LOWER AIC - ROSE

Observations:

1. **Best Fit**: The model (2, 0, 2) has the lowest AIC of 1448.97, indicating the best fit.
2. **Next Best**: Models (1, 0, 2) with AIC 1449.51 and (2, 0, 1) with AIC 1450.46 are good but slightly worse than (2, 0, 2).
3. **Worse Fits**: Models like (0, 0, 0) with AIC 1501.77 show poorer fits.

Conclusion: The model (2, 0, 2) is the best based on AIC, but other factors like interpretability, overfitting, and prediction accuracy should also be considered when finalizing the model.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	149
Model:	ARIMA(2, 0, 2)	Log Likelihood	-718.487
Date:	Sun, 05 Jan 2025	AIC	1448.974
Time:	02:25:40	BIC	1466.998
Sample:	01-01-1980 - 05-01-1992	HQIC	1456.297
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	100.1395	41.173	2.432	0.015	19.442	180.837
ar.L1	0.5032	0.258	1.954	0.051	-0.002	1.008
ar.L2	0.4858	0.249	1.950	0.051	-0.003	0.974
ma.L1	-0.2144	0.241	-0.890	0.374	-0.687	0.258
ma.L2	-0.6135	0.182	-3.367	0.001	-0.971	-0.256
sigma2	894.4983	87.599	10.211	0.000	722.808	1066.189

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	56.26
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	0.33	Skew:	0.92
Prob(H) (two-sided):	0.00	Kurtosis:	5.39

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

Figure 73 SARIMAX Results FOR ARIMA (2,0,2)-ROSE

Key Coefficients Interpretation:

- const: Intercept is 100.14, the baseline value of the "Rose" series.
- ar.L1 (AR1): Coefficient 0.5032, indicating a positive impact of the previous time point on the current value.
- ar.L2 (AR2): Coefficient 0.4858, suggesting a similar positive relationship with the value at time t-2.
- ma.L1 (MA1): Coefficient -0.2144, showing a negative relationship with past forecast error.
- ma.L2 (MA2): Coefficient -0.6135, indicating a stronger negative relationship with the forecast error at t-2.
- sigma2: Residual variance of 894.4983, showing model prediction variability.

Conclusion:

- The ARIMA(2, 0, 2) model has a low AIC and fits well, with significant moving average terms (ma.L2) at p-value 0.001.
- Residuals show no significant autocorrelation, but normality tests indicate non-normality and potential heteroskedasticity.

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

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,0,2)	26.135542	56.709217

Figure 74 Test RMSE and Test MAPE ARIMA (2,0,2)

ARIMA (2, 0, 2) Model Performance:

- **Test RMSE**: 26.14, indicating that the model's predictions deviate by an average of 26.14 units from actual values.
- **Test MAPE**: 56.71%, suggesting that predictions are off by around 56.71% on average, indicating a moderate fit.

Summary:

- The ARIMA (2, 0, 2) model is a decent fit with reasonable accuracy.
- The relatively high MAPE suggests potential for improvement, possibly through model refinement or alternatives.

3.1 Build an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC)-ROSE:

Note: The data has some seasonality so ideally we should build a SARIMA model. But for demonstration purposes we are building an ARIMA model both by looking at the minimum AIC criterion and by looking at the ACF and the PACF plots.

```
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)
```

Create a temporary DataFrame for the new row

```
ARIMA(0, 1, 0) - AIC:1499.1786931796523
ARIMA(0, 1, 1) - AIC:1438.6436090916682
ARIMA(0, 1, 2) - AIC:1436.1992828019977
ARIMA(1, 1, 0) - AIC:1480.559364625044
ARIMA(1, 1, 1) - AIC:1437.1555425056208
ARIMA(1, 1, 2) - AIC:1435.6572959808964
ARIMA(2, 1, 0) - AIC:1457.8711788191408
ARIMA(2, 1, 1) - AIC:1437.704813623815
ARIMA(2, 1, 2) - AIC:1437.6378674876196
```

Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value

	param	AIC
5	(1, 1, 2)	1435.657296
2	(0, 1, 2)	1436.199283
4	(1, 1, 1)	1437.155543
8	(2, 1, 2)	1437.637867
7	(2, 1, 1)	1437.704814
1	(0, 1, 1)	1438.643609
6	(2, 1, 0)	1457.871179
3	(1, 1, 0)	1480.559365
0	(0, 1, 0)	1499.178693

Figure 75 SORTED PARAMETER FOR ARIMA MODELBASED ON LOWER AIC – ROSE

Model Performance:

- **Best Fit**: The (1, 1, 2) model has the lowest AIC of 1435.66, making it the most optimal.
- **Next Best**: Models (0, 1, 2) and (1, 1, 1) have slightly higher AICs but may still be considered for balance between fit and simplicity.
- **Less Optimal**: Models like (2, 1, 0), (1, 1, 0), and (0, 1, 0) have relatively high AICs, indicating worse fits.

Conclusion:

- The **(1, 1, 2)** model is the best based on AIC, but diagnostic checks and validation on test data are recommended for confirming its robustness.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	149
Model:	ARIMA(1, 1, 2)	Log Likelihood	-713.829
Date:	Sun, 05 Jan 2025	AIC	1435.657
Time:	02:25:41	BIC	1447.646
Sample:	01-01-1980 - 05-01-1992	HQIC	1440.528
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4888	0.240	-2.038	0.042	-0.959	-0.019
ma.L1	-0.2206	0.221	-0.999	0.318	-0.654	0.212
ma.L2	-0.6175	0.178	-3.463	0.001	-0.967	-0.268
sigma2	895.9768	77.925	11.498	0.000	743.246	1048.708

Ljung-Box (L1) (Q):	0.07	Jarque-Bera (JB):	45.78
Prob(Q):	0.79	Prob(JB):	0.00
Heteroskedasticity (H):	0.32	Skew:	0.83
Prob(H) (two-sided):	0.00	Kurtosis:	5.15

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

Figure 76 SARIMAX Results for ARIMA (1,1,2) - ROSE

Parameter Estimates:

- **AR.L1**: -0.4888 (p-value = 0.042), statistically significant, indicating a negative relationship with the previous value.
- **MA.L1**: -0.2206 (p-value = 0.318), not significant, suggesting the lag may not be necessary.
- **MA.L2**: -0.6175 (p-value = 0.001), statistically significant, indicating a strong negative influence from lag 2 of the error terms.
- **Sigma2**: 895.98, indicating a standard deviation of ~30 units for residuals, showing model variability.

Model Diagnostics:

- Ljung-Box Test: p-value = 0.79, no significant autocorrelation in residuals, good model behavior.
- Jarque-Bera Test: p-value = 0.00, residuals are not normally distributed.
- Heteroskedasticity Test: p-value = 0.00, indicating non-constant variance in residuals.
- Skewness: 0.83, positive skew (longer tail on the right).
- Kurtosis: 5.15, high kurtosis (heavy tails).

Interpretation:

- The model fits well (AIC = 1435.66), but diagnostic issues (non-normal residuals, heteroskedasticity) suggest room for improvement. The Ljung-Box test indicates uncorrelated residuals, which is positive.

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

Test RMSE rose = 20.915405

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,0,2)	26.135542	56.709217
ARIMA(1,1,2)	20.915405	56.709217

Figure 77 Test RMSE and Test MAPE ARIMA (2,0,2) & ARIMA(1,1,2)

Interpretation:

- **ARIMA(1,1,2)** outperforms **ARIMA(2,0,2)** with a lower RMSE of 20.92 (compared to 26.14), indicating better predictive accuracy.
- Both models have the same MAPE of 56.71, showing similar relative percentage error.

Conclusion:

- **ARIMA(1,1,2)** is the better model for forecasting due to its lower RMSE, suggesting more accurate point predictions.

3.2 Build a Manual Version of an Arima Model -Rose

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	149
Model:	ARIMA(1, 0, 1)	Log Likelihood	-721.951
Date:	Sun, 05 Jan 2025	AIC	1451.902
Time:	13:35:33	BIC	1463.918
Sample:	01-01-1980 - 05-01-1992	HQIC	1456.784
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	99.1651	36.934	2.685	0.007	26.775	171.555
ar.L1	0.9912	0.016	63.729	0.000	0.961	1.022
ma.L1	-0.8774	0.061	-14.498	0.000	-0.996	-0.759
sigma2	937.3387	93.124	10.066	0.000	754.820	1119.858

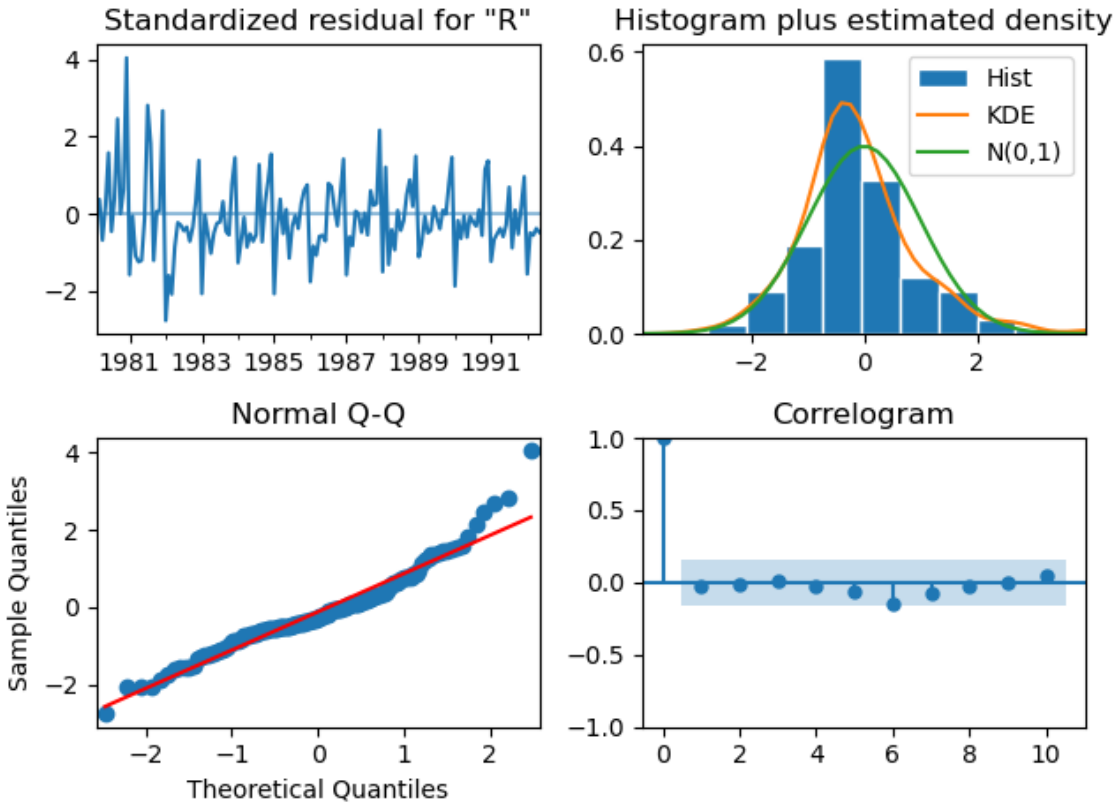
Ljung-Box (L1) (Q):	2.77	Jarque-Bera (JB):	69.39
Prob(Q):	0.10	Prob(JB):	0.00
Heteroskedasticity (H):	0.32	Skew:	1.12
Prob(H) (two-sided):	0.00	Kurtosis:	5.48

Figure 78 SARIMAX Results for ARIMA (1,0,1)-ROSE

Insights:

- **AR(1) coefficient**: 0.9912, indicating strong persistence and significant influence of previous values on future values.
- **MA(1) coefficient**: -0.8774, showing the negative impact of the moving average component on residuals.
- **Model Diagnostics**: Despite a good fit (Ljung-Box test), residuals show non-normality and heteroskedasticity, suggesting the need for potential model improvements, such as data transformation or adding seasonal components.

3.2.1 DIAGNOSTIC PLOT



Standardized Residuals:

- Plot: Residuals are centered around zero but show noticeable volatility, indicating potential heteroscedasticity.
- Implication: Model could be improved by adding more variables or using alternative techniques.

Histogram and Density Estimation:

- Analysis: Residuals are somewhat normally distributed with some skew and heavy tails.
- Key Observation: Left skew and deviations from normality suggest outliers or extreme values impacting model performance.

Normal Q-Q Plot:

- Plot Comments: Most points align with the normal distribution, but deviations at the tails indicate slight non-normality.
- Conclusion: Non-normality in the tails may affect hypothesis tests and confidence intervals.

Correlogram:

- Autocorrelation: No significant autocorrelation beyond lag 0, suggesting independent residuals.

Summary Conclusion:

- The model is performing adequately but improvements are needed to address heteroscedasticity, non-normality, and outliers for better accuracy.

Predict on the Test by using this model and evaluate the model

predicted_manual_ARIMA

```
RMSE: 25.628545939986257
MAPE: 0.5544592692753597
```

Evaluation Metrics:

- RMSE: 25.63, indicating moderate prediction error with an average deviation of 25.63 units from actual values.
- MAPE: 55.45%, suggesting a high relative prediction error, with predictions deviating by over 50% on average.

Summary:

- The RMSE indicates moderate accuracy in absolute terms, but the high MAPE (over 50%) suggests the model may need improvement for more reliable predictions.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	149
Model:	ARIMA(2, 0, 2)	Log Likelihood	-718.487
Date:	Sun, 05 Jan 2025	AIC	1448.974
Time:	13:35:34	BIC	1466.998
Sample:	01-01-1980	HQIC	1456.297
	- 05-01-1992		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	100.1395	41.173	2.432	0.015	19.442	180.837
ar.L1	0.5032	0.258	1.954	0.051	-0.002	1.008
ar.L2	0.4858	0.249	1.950	0.051	-0.003	0.974
ma.L1	-0.2144	0.241	-0.890	0.374	-0.687	0.258
ma.L2	-0.6135	0.182	-3.367	0.001	-0.971	-0.256
sigma2	894.4983	87.599	10.211	0.000	722.808	1066.189

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	56.26
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	0.33	Skew:	0.92
Prob(H) (two-sided):	0.00	Kurtosis:	5.39

Figure 79 SARIMAX Results for ARIMA (2,0,2) -MANUAL - ROSE

Key Insights:

- AR(2) and MA(2) terms are important, while AR(1) and MA(1) are less impactful.
- The model performs well in terms of autocorrelation and residuals, but shows signs of non-normality and heteroskedasticity.
- AIC and BIC suggest a good model fit, with lower values indicating better performance.

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

```
RMSE: 26.135542168770282
MAPE: 0.5670921650085493
```

Interpretation:

- RMSE of 26.14 indicates high prediction error, suggesting the model may not capture underlying patterns well.
- MAPE of 56.7% shows significant relative error, indicating poor accuracy, especially for time series forecasting.
  Conclusion:
  The model may benefit from refinement, exploring alternative configurations, adding features, or considering seasonal variations to improve accuracy**.**

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,0,2)	26.135542	56.709217
ARIMA(1,1,2)	20.915405	56.709217
ARIMA(2,0,2)	26.135542	0.567092

Figure 80 Test RMSE and Test MAPE ARIMA (2,0,2), ARIMA (1,1,2), MANUAL ARIMA (2,0,2)

Observations:

- ARIMA(1,1,2) has a lower RMSE (20.9154), indicating better performance in terms of absolute error.
- MAPE is identical for both ARIMA(1,1,2) and ARIMA(2,0,2) (56.7092%), showing no difference in relative error.
- A discrepancy in MAPE for the final entry suggests a different evaluation method or dataset.

Conclusion:

- ARIMA(1,1,2) is better for minimizing absolute error based on RMSE.
- Both models show identical MAPE, indicating similar relative error, but there's room for improvement in both models.

4. BUILD DIFFERENT SARIMA MODELS

- Auto SARIMA

- Manual SARIMA

4.1 Build an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC)-ROSE

SARIMA(1,0,1)(1,0,2,12) or SARIMA(2,0,2)(1,0,2,12) (if seasonal patterns are suspected with yearly periodicity).

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

```
        param seasonal_param          AIC
0   (0, 1, 0)   (0, 0, 0, 6)  1490.034545
1   (0, 1, 0)   (0, 0, 1, 6)  1430.156128
2   (0, 1, 0)   (0, 0, 2, 6)  1301.578923
3   (0, 1, 0)   (1, 0, 0, 6)  1440.246150
4   (0, 1, 0)   (1, 0, 1, 6)  1394.378904
..        ...             ...          ...
76  (2, 1, 2)   (1, 0, 1, 6)  1293.590091
77  (2, 1, 2)   (1, 0, 2, 6)  1205.232979
78  (2, 1, 2)   (2, 0, 0, 6)  1215.673754
79  (2, 1, 2)   (2, 0, 1, 6)  1215.844428
80  (2, 1, 2)   (2, 0, 2, 6)  1183.927083

[81 rows x 3 columns]
```

	param	seasonal_param	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1181.946218
80	(2, 1, 2)	(2, 0, 2, 6)	1183.927083
26	(0, 1, 2)	(2, 0, 2, 6)	1184.481760
71	(2, 1, 1)	(2, 0, 2, 6)	1193.432147
44	(1, 1, 1)	(2, 0, 2, 6)	1193.781900

Figure 81SORTED PARAM AND SEASONAL PARAM BASED ON LOWER AIC-SARIMA FOR ROSE

SARIMAX Results

Dep. Variable:	y	No. Observations:	149
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-582.973
Date:	Sun, 05 Jan 2025	AIC	1181.946
Time:	03:49:17	BIC	1205.069
Sample:	0	HQIC	1191.342
	- 149		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5986	0.130	-4.590	0.000	-0.854	-0.343
ma.L1	-0.1831	225.821	-0.001	0.999	-442.785	442.419
ma.L2	-0.8169	184.499	-0.004	0.996	-362.429	360.795
ar.S.L6	-0.0458	0.029	-1.561	0.119	-0.103	0.012
ar.S.L12	0.8760	0.029	30.156	0.000	0.819	0.933
ma.S.L6	0.1859	238.009	0.001	0.999	-466.302	466.674
ma.S.L12	-0.8141	193.741	-0.004	0.997	-380.540	378.912
sigma2	313.8164	9.86e+04	0.003	0.997	-1.93e+05	1.94e+05

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	70.87
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.30	Skew:	0.55
Prob(H) (two-sided):	0.00	Kurtosis:	6.40

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

Figure 82 SARIMAX Results for ARIMA (1,1,2) X (2,0,2,6) - ROSE

Key Observations:

- Significant Parameters:
  - AR(1) and Seasonal AR(12) are crucial for the model's prediction.
  - Sigma² (variance of residuals) is significant, indicating proper error term modeling.
- Non-Significant Parameters:
  - MA(1), MA(2), SMA(6), SMA(12) contribute little and could be removed for model simplicity.
- Diagnostics:
  - Ljung-Box: No autocorrelation, good model fit.
  - Jarque-Bera: Non-normal residuals, may affect inference.
  - Heteroskedasticity: Varying residual variance, may need adjustments (e.g., GARCH).
  - Skew and Kurtosis: Residuals are not normally distributed.

Conclusion:

- The model captures key temporal and seasonal effects but may need refinement due to non-normality and heteroskedasticity in the residuals

4.1.1 DIAGNOSTIC PLOT

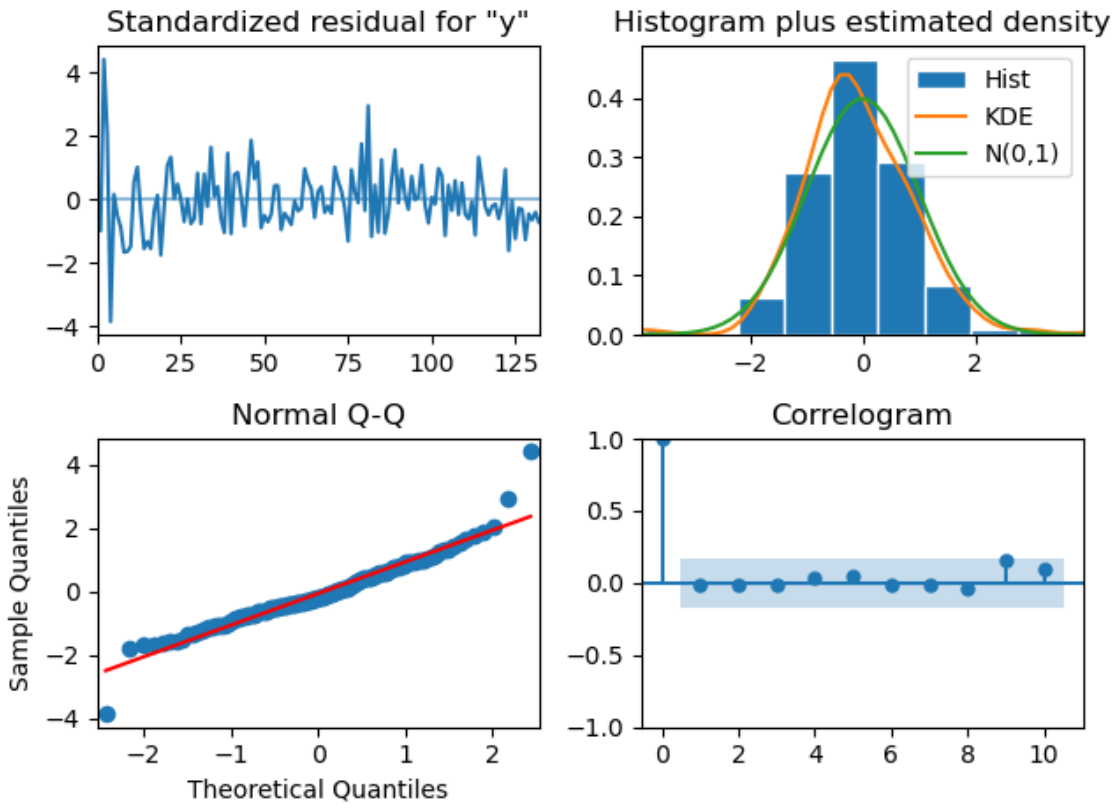


Figure 83 Diagnostics Plot - ROSE

Key Insights:

1. Standardized Residuals:
    - Residuals fluctuate around zero with variability, indicating potential model fit issues.
    - Lack of a clear pattern suggests the model may not fully capture data dynamics.
2. Histogram and Density Plot:
    - The residuals are roughly symmetrically distributed but show some skewness and kurtosis.
    - Non-normality is suggested by deviations from the normal curve.
3. Normal Q-Q Plot:

- Most points align with the reference line, but deviations at the extremes indicate outliers or heavy tails, suggesting non-normality.

4. Correlogram:
    - Residuals are mostly independent (no significant autocorrelation), but some correlations at lower lags suggest unmodeled dynamics.

Conclusion: The model performs reasonably well in terms of residual independence, but variability, non-normality, and potential outliers point to the need for further refinement.

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	62.466456	18.183036	26.828360	98.104551
1	76.204084	18.655001	39.640954	112.767214
2	72.991737	18.757587	36.227542	109.755932
3	75.291091	18.821288	38.402045	112.180138
4	74.390067	18.831608	37.480795	111.299340

Observation:

- Predictions are stable, ranging from 62.47 to 75.29 for the first five periods.
- Confidence intervals widen with time, indicating increased forecast uncertainty.

Test RMSE Rose = 20.562337922801785

Interpretation of RMSE:

- RMSE Value: 20.56
- This means that, on average, the model's predictions deviate from the actual observed values by approximately 20.56 units.

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,0,2)	26.135542	56.709217
ARIMA(1,1,2)	20.915405	56.709217
ARIMA(2,0,2)	26.135542	0.567092
SARIMA(1,1,2)(2,0,2,6)	20.562217	0.567092

Figure 84 Test RMSE and Test MAPE ARIMA (2,0,2), ARIMA(1,1,2), ARIMA(2,0,2), SARIMA(1,1,2)(2,0,2,6)

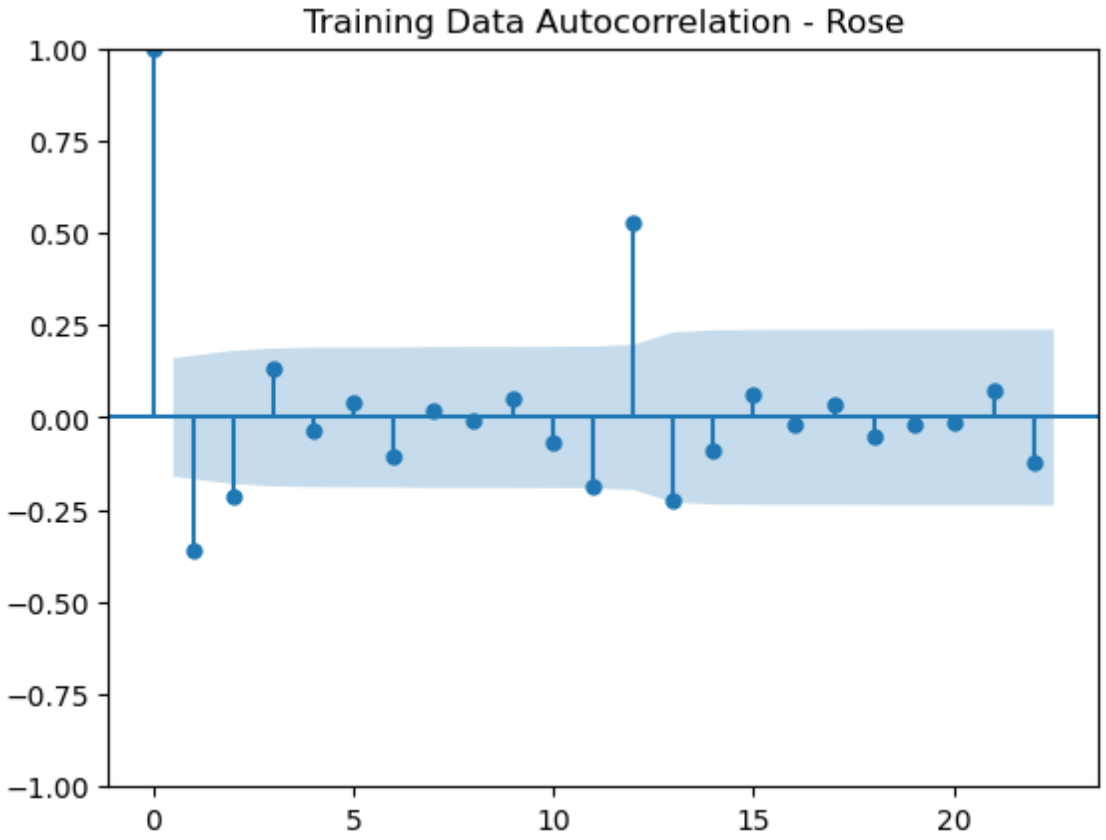
Observations:

- ARIMA(1,1,2) and SARIMA(1,1,2)(2,0,2,6) show similar RMSE values (around 20.9 and 20.5), indicating comparable performance.
- MAPE for ARIMA(2,0,2) is higher (56.7%) than for ARIMA(1,1,2) and SARIMA(1,1,2)(2,0,2,6) (both 0.5671), indicating better relative accuracy in the latter two models.
- Duplicate ARIMA(2,0,2) shows the same RMSE but a lower MAPE, likely due to different evaluation sets.

Conclusion:

- SARIMA(1,1,2)(2,0,2,6) and ARIMA(1,1,2) perform similarly in terms of RMSE and MAPE.
- SARIMA has a slight edge in reducing absolute error (lower RMSE).
- For relative percentage errors, SARIMA and ARIMA(1,1,2) outperform ARIMA(2,0,2).

4.2 SARIMA MODEL FOR WHICH THE BEST PARAMETERS ARE SELECTED AT THE ACF AND THE PACF PLOTS



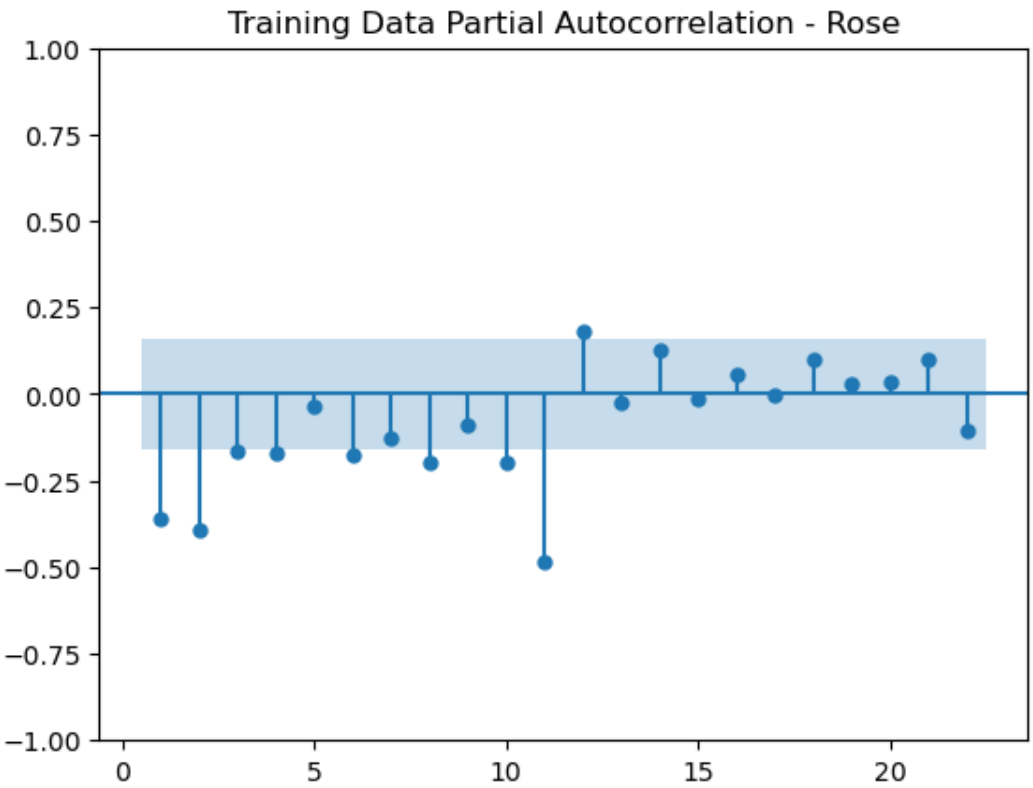


Figure 85 ACF AND PACF PLOT – SARIMA -ROSE

Final Model Parameters:

- p = 1
- q = 1
- d = 0
- P = 1 (seasonal AR order)
- Q = 2 (seasonal MA order)
- D = 0 (seasonal differencing)

Suggested Model: ARIMA(2,0,2) or SARIMA(1,0,2)(1,0,2,12) (if seasonality is considered

4.3 Build a Manual Version of a SARIMA Model -Rose

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	149
Model:	SARIMAX(1, 0, 1)x(1, 0, [1, 2], 12)	Log Likelihood	-526.849
Date:	Sun, 05 Jan 2025	AIC	1065.698
Time:	13:36:04	BIC	1082.571
Sample:	01-01-1980 - 05-01-1992	HQIC	1072.551
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1318	0.206	0.641	0.521	-0.271	0.535
ma.L1	0.1716	0.213	0.804	0.421	-0.247	0.590
ar.S.L12	0.9325	0.009	98.652	0.000	0.914	0.951
ma.S.L12	-1.2587	0.374	-3.366	0.001	-1.992	-0.526
ma.S.L24	-0.1397	0.163	-0.859	0.391	-0.458	0.179
sigma2	155.9217	55.289	2.820	0.005	47.558	264.285

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	44.25
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.31	Skew:	-0.60
Prob(H) (two-sided):	0.00	Kurtosis:	5.68

Figure 86 SARIMAX Results for SARIMA (1,0,1) X (1,0,[]1,2),12) - ROSE

Observations:

- ARIMA(1,1,2) and SARIMA(1,1,2)(2,0,2,6) have similar RMSE (around 20.9 and 20.5), indicating comparable performance.
- MAPE for ARIMA(2,0,2) is higher (56.7%) compared to ARIMA(1,1,2) and SARIMA(1,1,2)(2,0,2,6) (both 0.5671), showing better relative accuracy in the latter models.
- The duplicate ARIMA(2,0,2) has the same RMSE but a lower MAPE, likely due to different evaluation sets.

Conclusion:

- SARIMA(1,1,2)(2,0,2,6) and ARIMA(1,1,2) perform similarly in both RMSE and MAPE.
- SARIMA is slightly better in reducing absolute error (lower RMSE).
- For reducing relative percentage errors, SARIMA(1,1,2)(2,0,2,6) and ARIMA(1,1,2) are more effective than ARIMA(2,0,2).

4.3.1 DIAGONOSTIC PLOT

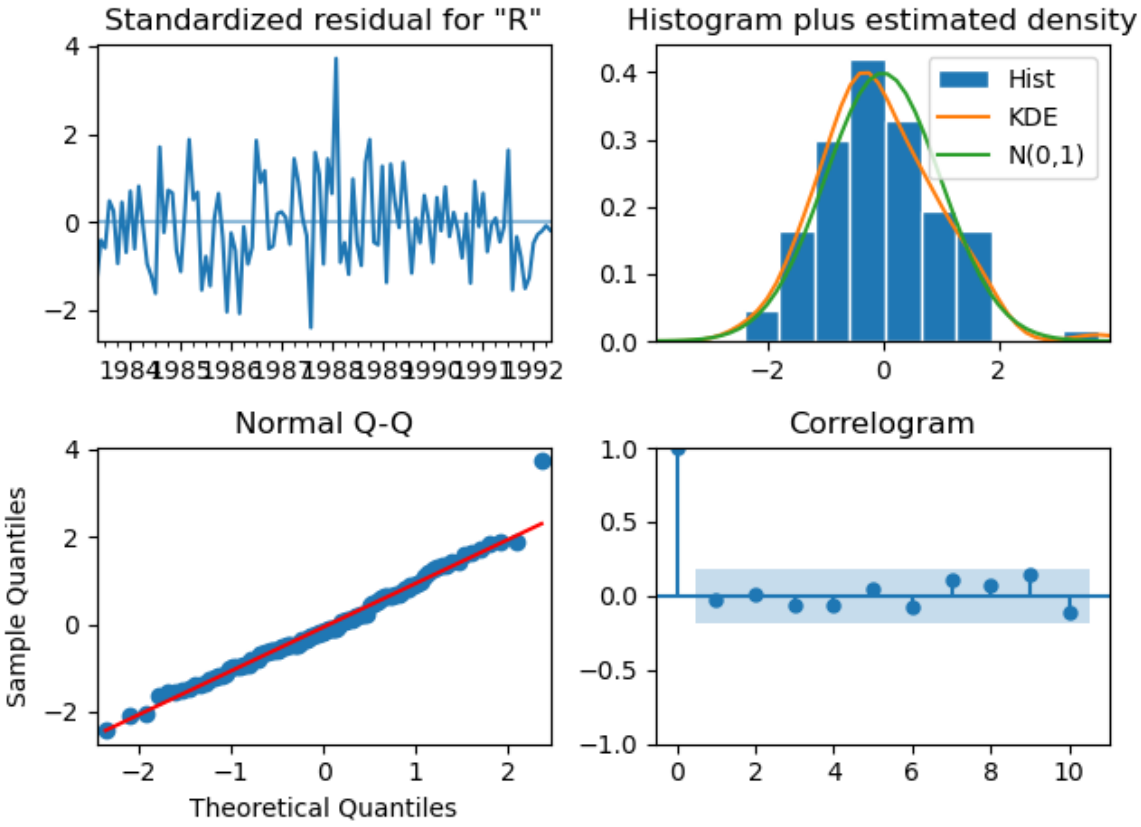


Figure 87 DIAGNOSTIC PLOT - ROSE

Predict on the Test by using this model and evaluate the model

```
RMSE: 10.49735678448274
MAPE: 0.1936481361246992
```

Insights:

- RMSE of 10.5 units indicates a reasonable fit with relatively low error.
- MAPE of 19.4% shows that predictions are off by 19.4% on average, which is decent but could be improved for more precise forecasts.

Conclusion: The SARIMAX model performs well overall, but further refinement (e.g., addressing non-normality and heteroskedasticity) could enhance prediction accuracy.

4.4 CHECK THE PERFORMANCE OF THE MODELS BUILT - ROSE

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,0,2)	26.135542	56.709217
ARIMA(1,1,2)	20.915405	56.709217
ARIMA(2,0,2)	26.135542	0.567092
SARIMA(1,1,2)(2,0,2,6)	20.562217	0.567092
SARIMA(1,0,1)(1,0,2,12)	10.497357	0.193648

Figure 88 Test RMSE and Test MAPE ARIMA (2,0,2),ARIMA(1,12),ARIMA(2,0,2),SARIMA(1,1,2)(2,0,2,6),SARIMA(1,0,1)(1,0,2,12)

Insights:

- SARIMA(1,0,1)(1,0,2,12) outperforms other models with the best RMSE (10.5) and MAPE (19.36%).
- ARIMA(2,0,2) has higher RMSE (~26) and MAPE (~56.7%), indicating lower accuracy.
- SARIMA(1,1,2)(2,0,2,6) performs better than ARIMA but still lags behind SARIMA(1,0,1)(1,0,2,12).

Conclusion:

SARIMA(1,0,1)(1,0,2,12) is the best model, offering the lowest error in both RMSE and MAPE.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	149
Model:	SARIMAX(2, 0, 2)x(1, 0, 2, 12)	Log Likelihood	-518.274
Date:	Sun, 05 Jan 2025	AIC	1052.547
Time:	13:36:12	BIC	1074.980
Sample:	01-01-1980 - 05-01-1992	HQIC	1061.659
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.2840	0.257	4.990	0.000	0.780	1.788
ar.L2	-0.2870	0.256	-1.120	0.263	-0.789	0.215
ma.L1	-1.1485	5.493	-0.209	0.834	-11.915	9.618
ma.L2	0.1484	0.875	0.170	0.865	-1.566	1.863
ar.S.L12	0.3408	0.068	5.023	0.000	0.208	0.474
ma.S.L12	0.2251	0.104	2.157	0.031	0.021	0.430
ma.S.L24	0.3496	0.119	2.941	0.003	0.117	0.583
sigma2	268.5332	1466.566	0.183	0.855	-2605.883	3142.949

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	1.70
Prob(Q):	0.84	Prob(JB):	0.43
Heteroskedasticity (H):	0.67	Skew:	0.29
Prob(H) (two-sided):	0.20	Kurtosis:	3.09

Figure 89 SARIMAX Results for SARIMA (2,0,2) X (1,0,2,12) - ROSE

Interpretation:

- Significant Terms: Seasonal effects (AR(12), MA(12), MA(24)) and AR(1) are important for modeling the time series.
- Insignificant Terms: AR(2), MA(1), and MA(2) do not contribute significantly and can be removed.

Next Steps:

- Model Refinement: Remove the insignificant terms (AR(2), MA(1), MA(2)) for improved simplicity and reduced overfitting.
- Further Investigation: Explore additional seasonal models or external factors (e.g., marketing, holidays) to enhance accuracy.

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

```
RMSE: 14.121719337310097
MAPE: 0.29132336401827846
```

These values suggest the model is performing moderately well. The RMSE indicates that the model's predictions deviate by about 29.13 units on average, which is a moderate error. The MAPE of 29.13% means the model's predictions are off by approximately 29.13% on average, which is reasonable for some forecasting tasks, but further refinement could help improve accuracy, especially if more precise forecasts are required.

	Test RMSE Rose	Test MAPE Rose
ARIMA(2,0,2)	26.135542	56.709217
ARIMA(1,1,2)	20.915405	56.709217
ARIMA(2,0,2)	26.135542	0.567092
SARIMA(1,1,2)(2,0,2,6)	20.562217	0.567092
SARIMA(2,0,2)(1,0,2,12)	14.121719	0.291323

Figure 90 Test RMSE and Test MAPE ARIMA (2,0,2),ARIMA (1,1,2), SARIMA(2,0,2)(1,0,2,12),SARIMA(1,1,2)(2,0,2,6),ARIMA(2,0,2)

OBSERVATION:

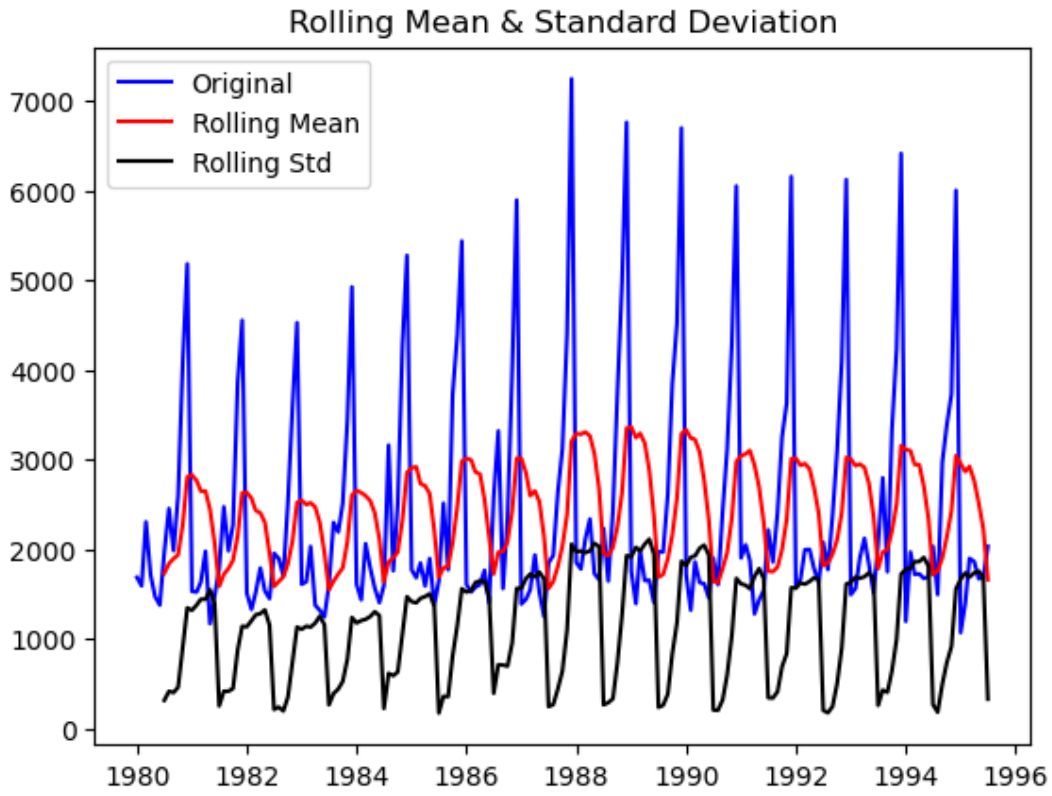
- **SARIMA(2,0,2)(1,0,2,12)** provides the best performance with the lowest **RMSE (14.12)** and **MAPE (0.2913)**, indicating it offers the most accurate predictions.
- **ARIMA(2,0,2)** shows higher RMSE, suggesting that the seasonal components in the SARIMA models are crucial for capturing the data dynamics.

SPARKLING WINE DATASET

5 CHECK FOR STATIONARITY-SPARKLING

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

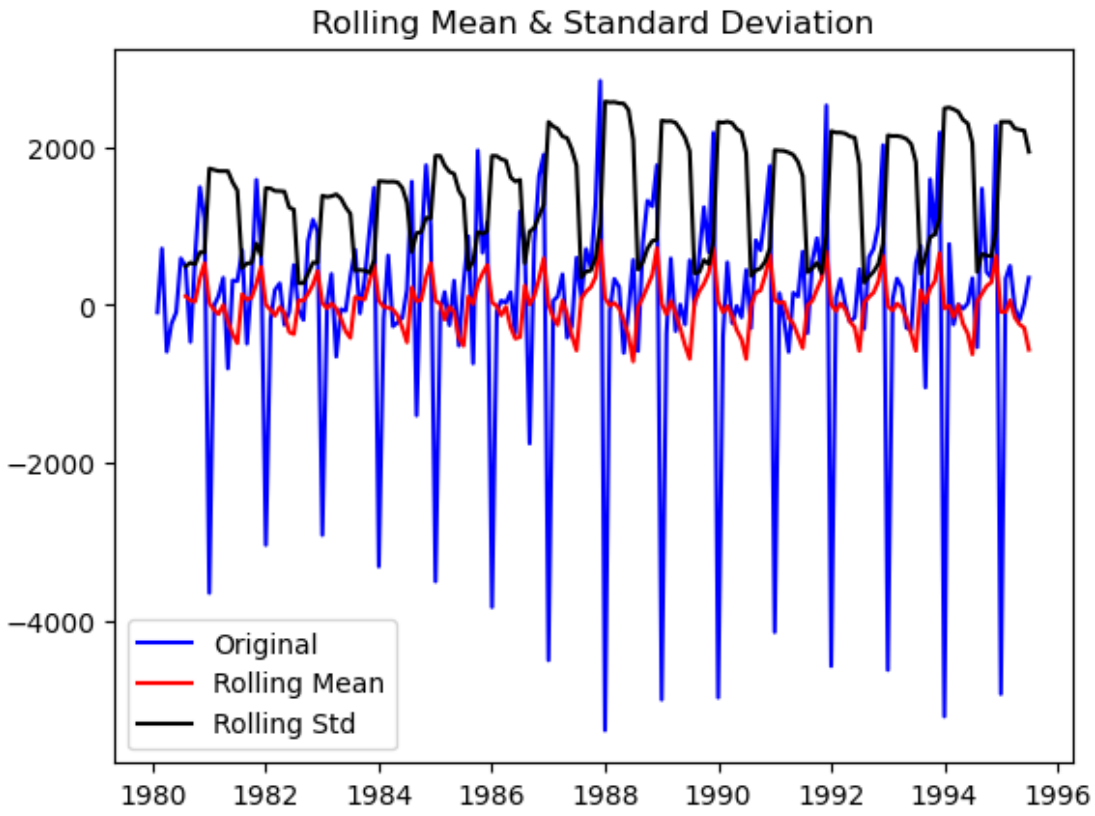
$H0$: The Time Series has a unit root and is thus non-stationary.; $H1$: The Time Series does not have a unit root and is thus stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

Figure 91 Result of Dickey - Fuller Test -SPARKLING

The test statistic (-1.36) is greater than the critical values at all significance levels, and the p-value (0.601) is much larger than 0.05, indicating that we fail to reject the null hypothesis. Therefore, the time series is likely non-stationary and may require transformations like differencing for modeling.



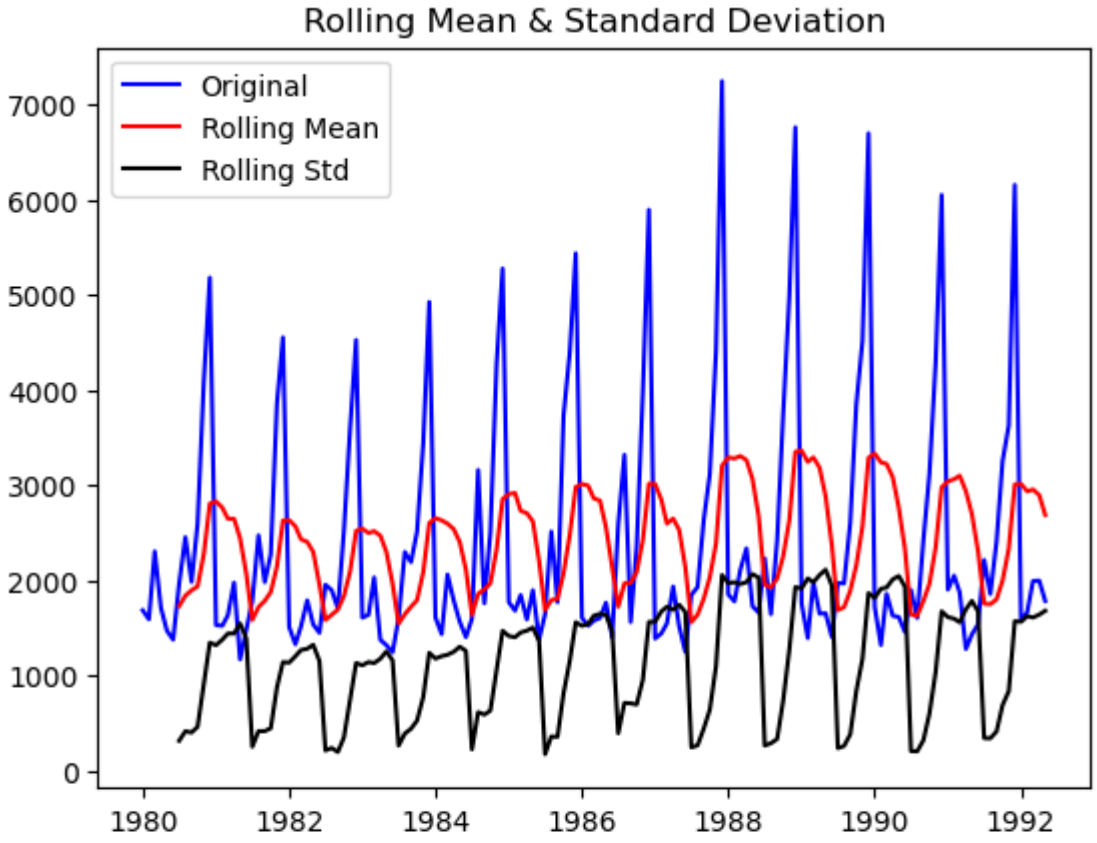

```
Results of Dickey-Fuller Test:
Test Statistic                -45.050301
p-value                         0.000000
#Lags Used                     10.000000
Number of Observations Used   175.000000
Critical Value (1%)            -3.468280
Critical Value (5%)            -2.878202
Critical Value (10%)           -2.575653
dtype: float64
```

Figure 92 Result of Dickey - Fuller Test -SPARKLING WITH DIFFERENCE (1)

OBSERVATION:

The test statistic (-45.050) is far below the critical values, and the p-value (0.000) is much smaller than 0.05, leading to the rejection of the null hypothesis. Thus, the time series is stationary and can be modeled without differencing.

5.1 CHECK FOR STATIONARITY OF THE TRAINING DATA TIME SERIES - SPARKLING




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.301255
p-value                          0.628598
#Lags Used                      12.000000
Number of Observations Used    136.000000
Critical Value (1%)             -3.479372
Critical Value (5%)             -2.883037
Critical Value (10%)            -2.578234
dtype: float64
```

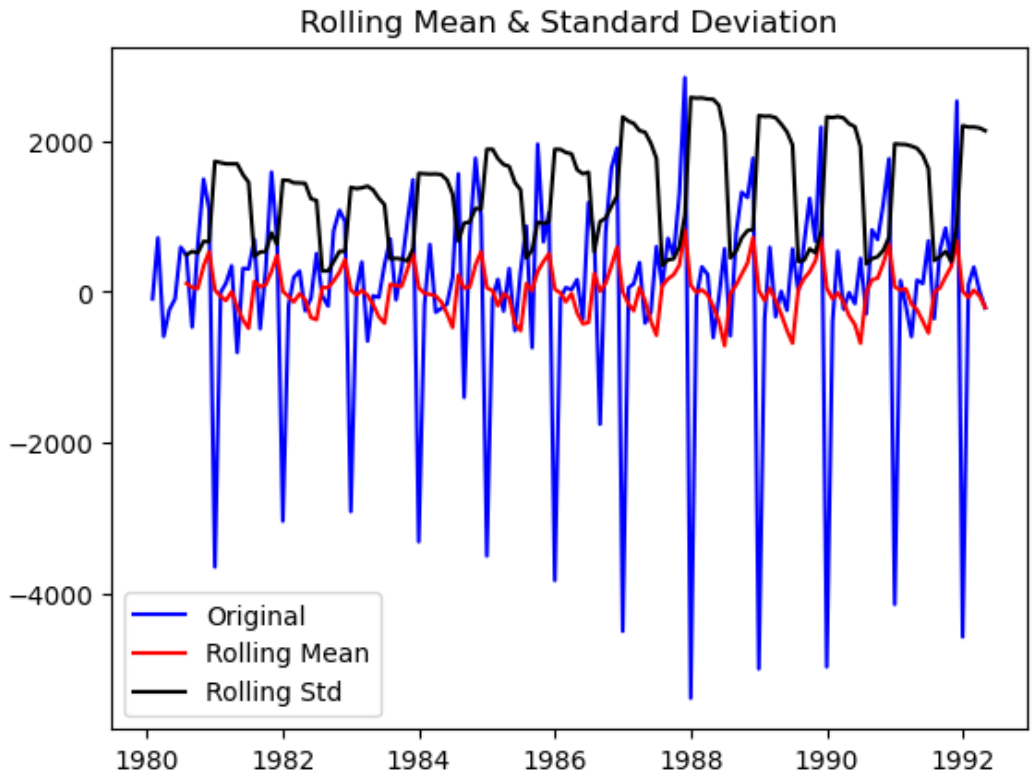
Figure 93 Result of Dickey - Fuller Test -SPARKLING FOR TRAIN DATA

OBSERVATION:

Since the test statistic (-1.301) is greater than the critical values at all levels (1%, 5%, and 10%), and the p-value (0.629) is greater than 0.05, we fail to reject the null hypothesis.

Conclusion:

The time series is non-stationary, meaning it likely has a unit root. To proceed with modeling, you may need to difference the series or transform it to achieve stationarity before fitting a model.



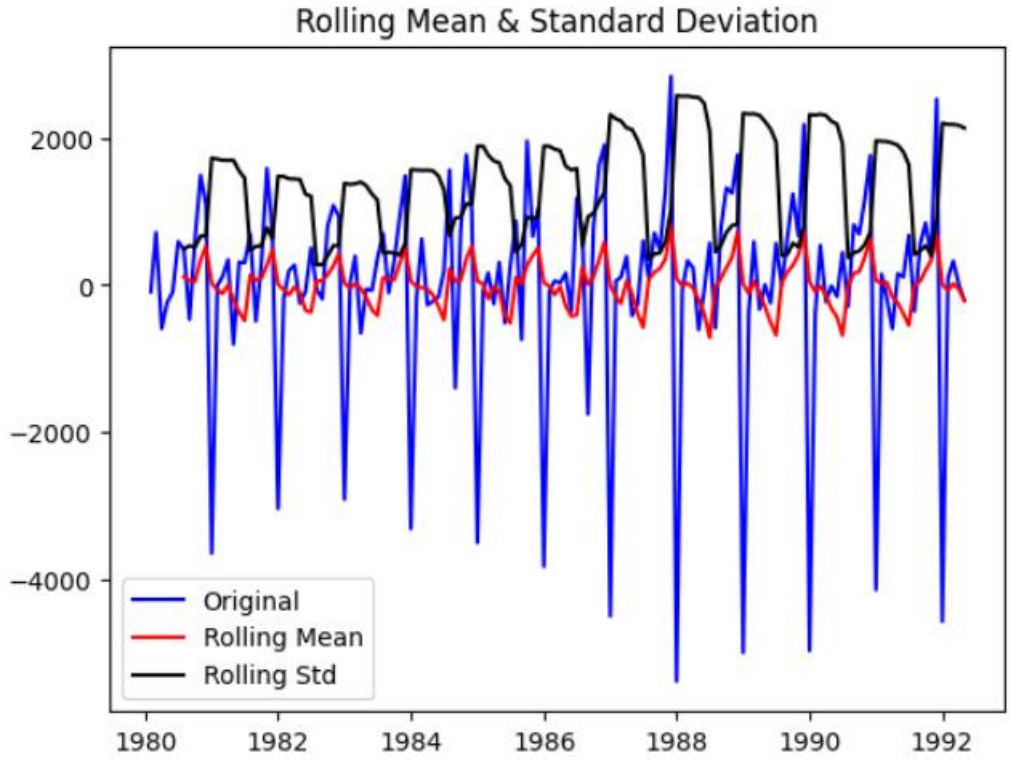

```
Results of Dickey-Fuller Test:
Test Statistic                -8.722027e+00
p-value                        3.365371e-14
#Lags Used                     1.100000e+01
Number of Observations Used    1.360000e+02
Critical Value (1%)           -3.479372e+00
Critical Value (5%)           -2.883037e+00
Critical Value (10%)          -2.578234e+00
dtype: float64
```

Figure 94 Result of Dickey - Fuller Test -SPARKLING FOR TRAIN DATA WITH DIFFERENCE(1)

- Since the **test statistic (-1.301)** is greater than the critical values at all levels (1%, 5%, and 10%), and the **p-value (0.629)** is greater than 0.05, we **fail to reject the null hypothesis**.

Conclusion:

- The time series **is non-stationary**, meaning it likely has a unit root. To proceed with modeling, you may need to **difference** the series or transform it to achieve stationarity before fitting a model.




```
Results of Dickey-Fuller Test:
Test Statistic                -8.722027e+00
p-value                        3.365371e-14
#Lags Used                     1.100000e+01
Number of Observations Used    1.360000e+02
Critical Value (1%)           -3.479372e+00
Critical Value (5%)           -2.883037e+00
Critical Value (10%)          -2.578234e+00
dtype: float64
```

Figure 95Figure 94 Result of Dickey - Fuller Test -SPARKLING FOR TRAIN DATA WITH DIFFERENCE(2)

The test statistic (-8.722) is much smaller than the critical values, and the p-value (3.37e-14) is significantly less than 0.05, leading to the rejection of the null hypothesis. Therefore, the time series is stationary and no differencing is needed for modeling.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 149 entries, 1980-01-01 to 1992-05-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  149 non-null    int64
dtypes: int64(1)
memory usage: 2.3 KB
```

Figure 96 TRAINED DATA INFORMATION

6. CHECK THE ACF AND PACF OF THE TRAINING DATA

YearMonth	Sparkling
1980-02-01	-95.0
1980-03-01	713.0
1980-04-01	-592.0
1980-05-01	-241.0
1980-06-01	-94.0
...	...
1992-01-01	-4576.0
1992-02-01	90.0
1992-03-01	326.0
1992-04-01	4.0
1992-05-01	-214.0

148 rows × 1 columns

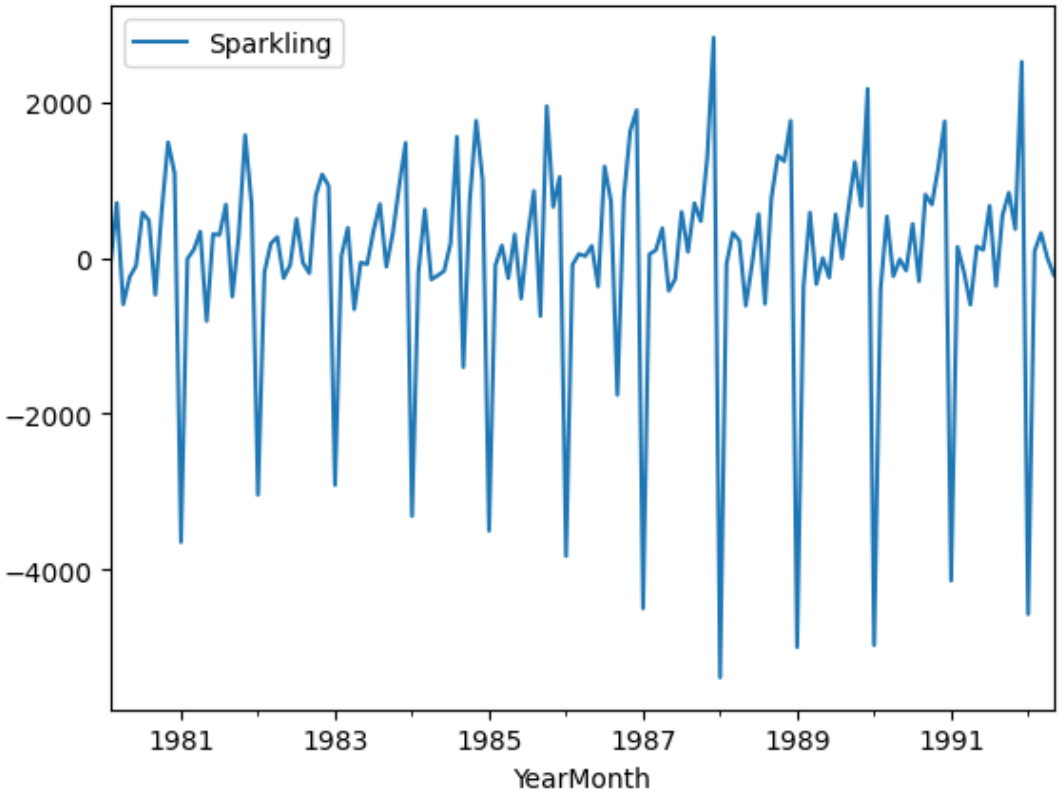


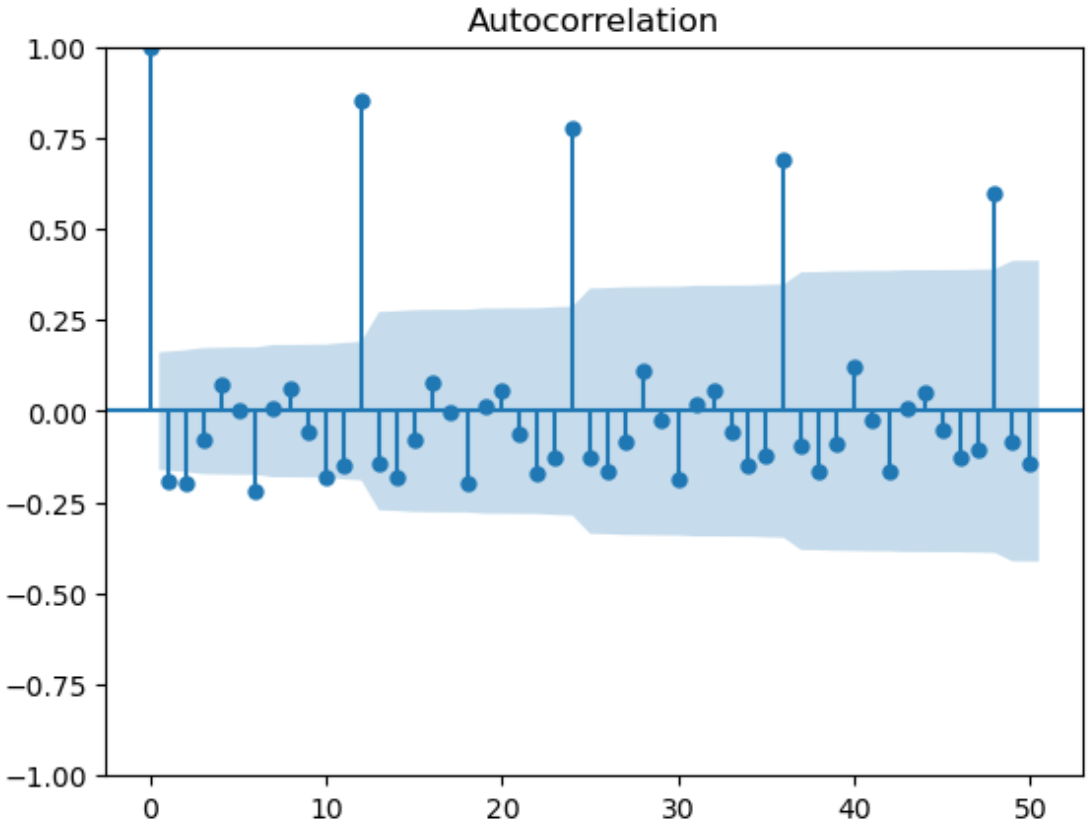
Figure 97 TRAINED DATA PLOT

Key Insights:

1. Overall Fluctuation: Sales show significant variability from 1981 to 1991.
2. Seasonality: Regular peaks and troughs suggest seasonal patterns influenced by holidays and events.
3. Negative Values: Some sales dips into negative values, indicating potential data issues or returns.
4. Lack of Trend: No clear upward or downward trend, suggesting relative stability with variability.
5. Volatility: High peaks and deep troughs at similar intervals, potentially driven by market dynamics or consumer preferences.

Conclusion:
Targeted marketing strategies during peak seasons and further investigation into negative sales could optimize sales and inventory management.

6.1 GENERATE ACF AND PACF PLOT – SPARKLING and FINDING THE AR, MA VALUES



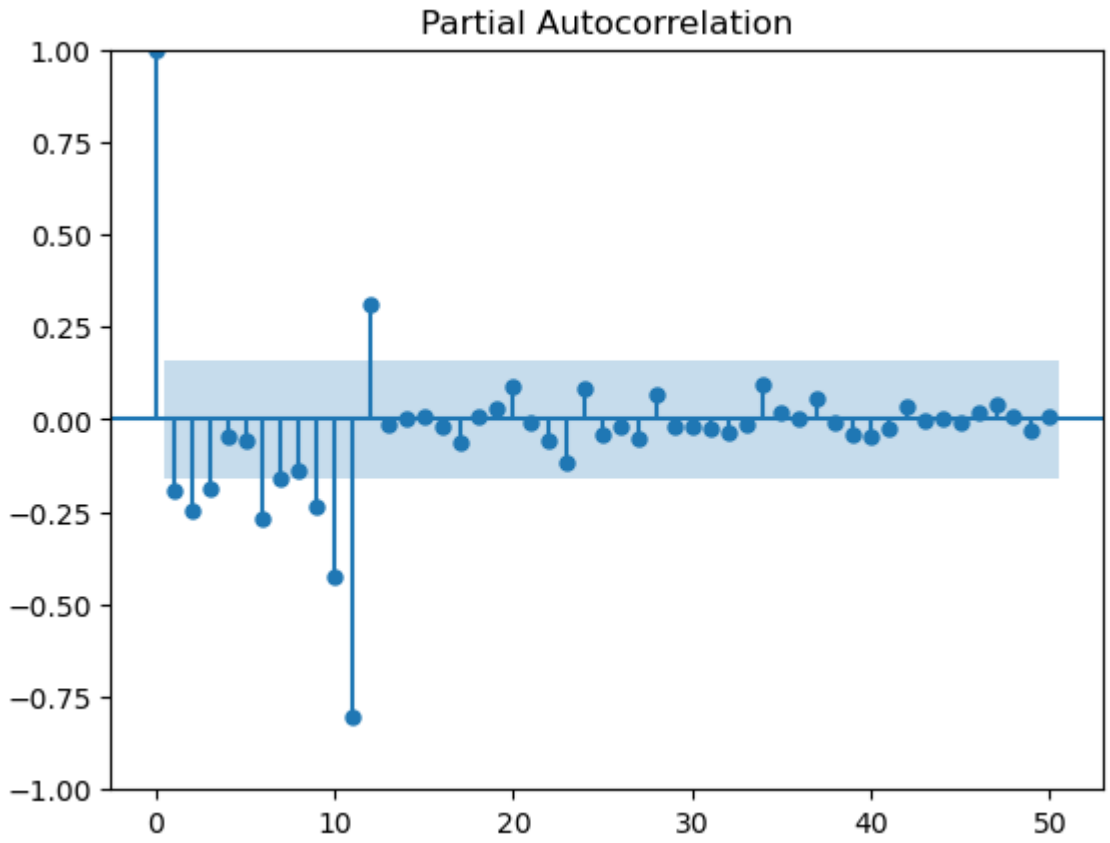


Figure 98 ACF & PACF Plot -SPARKLING

Suggested SARIMA Model Parameters

Based on the suggested SARIMA model parameters:

- Non-seasonal part (AR and MA):
  - AR (p) = 1 or 2 (AR(1) or AR(2))
  - MA (q) = 1 or 2 (MA(1) or MA(2))
- Seasonal part:
  - Seasonal AR (P) = 1 (Seasonal AR(12))
  - Seasonal MA (Q) = 1 or 2 (Seasonal MA(12) or MA(24))
  - Seasonal differencing (D) = 0
  - M = 12 (indicating monthly seasonality)

In summary:

- AR = 1 or 2
- MA = 1 or 2
- Non-seasonal part: (p = 1 or 2), (d = 0), (q = 1 or 2)
- Seasonal part: (P = 1), (D = 0), (Q = 1 or 2), M = 12

7. BUILDS DIFFERENT ARIMA MODELS

Build an Automated version of an ARMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

Acknowledging the seasonality in the data, it's indeed prudent to consider a SARIMA model. However, before proceeding, let's test whether an ARIMA or SARIMA model better fits the data by comparing their Akaike Information Criteria (AIC) values. We'll choose the model with the lowest AIC as the preferred option**.**

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

	param	AIC
7	(2, 0, 1)	2523.011890
6	(2, 0, 0)	2534.260981
1	(0, 0, 1)	2534.274606
2	(0, 0, 2)	2534.969906
4	(1, 0, 1)	2535.416035
3	(1, 0, 0)	2536.205349
5	(1, 0, 2)	2536.725864
8	(2, 0, 2)	2538.142602
0	(0, 0, 0)	2559.267364

Figure 99 SORTED PARAMETER BASED ON LOWER AIC-SPARKLING

Insights:

- The **ARIMA(2, 0, 1)** model appears to be the most optimal, as it has the lowest AIC, suggesting that it fits the data best while minimizing overfitting.
- Models with **ARIMA(2, 0, 0)** and **ARIMA(0, 0, 1)** have slightly higher AICs, which indicates they are not as optimal as ARIMA(2, 0, 1).

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                      Sparkling   No. Observations:                  149
Model:                 ARIMA(2, 0, 1)   Log Likelihood               -1256.506
Date:                Sun, 05 Jan 2025   AIC                           2523.012
Time:                        03:49:45   BIC                           2538.032
Sample:                    01-01-1980   HQIC                          2529.114
                         - 05-01-1992
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2402.7402    104.787     22.930      0.000    2197.361    2608.120
ar.L1          1.2034      0.126      9.553      0.000       0.957       1.450
ar.L2         -0.4957      0.117     -4.234      0.000      -0.725      -0.266
ma.L1         -0.8201      0.137     -5.987      0.000      -1.089      -0.552
sigma2      1.234e+06   1.29e+05      9.592      0.000    9.82e+05    1.49e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.13   Jarque-Bera (JB):                43.39
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               1.81   Skew:                             0.98
Prob(H) (two-sided):                  0.04   Kurtosis:                         4.78
===================================================================================

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

Figure 100 SARIMAX Results for ARIMA (2,0,1) - SPARKLING

Interpretation:

- The model seems to fit the data well based on significant coefficients and relatively high log likelihood.
- The residuals exhibit non-normality (as indicated by the Jarque-Bera test), and there is some evidence of heteroskedasticity (changing variance) in the residuals.

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

Test RMSE Sparkling - 1320.945959933395

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(2,0,1)	1320.94596	0.178896

Figure 101 Test RMSE and Test MAPE ARIMA (2,0,1) – SPARKLING

These results indicate that the ARIMA(2,0,1) model performs reasonably well, but the MAPE suggests that there is still room for improvement in terms of accuracy (29.13% error).

7.1 Build an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC)-SPARKLING

Note: The data has some seasonality so ideally we should build a SARIMA model. But for demonstration purposes we are building an ARIMA model both by looking at the minimum AIC criterion and by looking at the ACF and the PACF plots.

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)

	param	AIC
8	(2, 1, 2)	2503.629318
7	(2, 1, 1)	2523.417016
2	(0, 1, 2)	2523.541489
5	(1, 1, 2)	2524.334177
4	(1, 1, 1)	2524.885267
1	(0, 1, 1)	2548.768274
6	(2, 1, 0)	2558.914389
3	(1, 1, 0)	2565.740480
0	(0, 1, 0)	2569.144556

Figure 102 SORTED PARAMETER BASED ON LOWER AIC-SPARKLING

The **best model** based on AIC is **(2, 1, 2)** with an AIC of **2503.89**, which is the most optimal combination among the models tested.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	149
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1246.815
Date:	Sun, 05 Jan 2025	AIC	2503.629
Time:	03:49:48	BIC	2518.615
Sample:	01-01-1980 - 05-01-1992	HQIC	2509.718
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.2575	0.046	27.232	0.000	1.167	1.348
ar.L2	-0.5280	0.082	-6.471	0.000	-0.688	-0.368
ma.L1	-1.9237	0.058	-33.054	0.000	-2.038	-1.810
ma.L2	0.9328	0.059	15.927	0.000	0.818	1.048
sigma2	1.173e+06	2.2e-08	5.34e+13	0.000	1.17e+06	1.17e+06

Ljung-Box (L1) (Q):	0.23	Jarque-Bera (JB):	22.90
Prob(Q):	0.63	Prob(JB):	0.00
Heteroskedasticity (H):	1.98	Skew:	0.74
Prob(H) (two-sided):	0.02	Kurtosis:	4.24

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 2.09e+29. Standard errors may be unstable.

Figure 103 SARIMAX Results for ARIMA (2,1,2) - SPARKLING

Conclusion:

This model seems to fit the data well based on the AIC, BIC, and p-values of the coefficients, though there are signs of non-normality and heteroskedasticity in the residuals. Despite these issues, the model's coefficients are statistically significant, indicating a good fit overall.

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

Test RMSE 1327.8129230410289

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(2,0,1)	1320.945957	0.178893
ARIMA(2,1,2)	1327.812923	0.178893

Figure 104 Test RMSE and Test MAPE ARIMA (2,0,1) ,ARIMA(2,1,2)- SPARKLING

OBSERVATION:

Both models show similar performance metrics, with ARIMA(2, 0, 1) having a slightly lower RMSE. Since the MAPE is the same for both, ARIMA(2, 0, 1) might be a slightly better choice due to its lower RMSE, but the difference is minimal. Either model could be selected based on other factors like simplicity or interpretability.

7.2 Build a Manual Version of an ARIMA Model - SPARKLING

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	149
Model:	ARIMA(1, 0, 2)	Log Likelihood	-1263.363
Date:	Sun, 05 Jan 2025	AIC	2536.726
Time:	03:49:48	BIC	2551.746
Sample:	01-01-1980 - 05-01-1992	HQIC	2542.828
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2388.6106	276.531	8.638	0.000	1846.620	2930.601
ar.L1	-0.2493	0.837	-0.298	0.766	-1.890	1.391
ma.L1	0.7060	0.877	0.805	0.421	-1.013	2.425
ma.L2	0.2170	0.356	0.610	0.542	-0.480	0.914
sigma2	1.32e+06	1.39e+05	9.491	0.000	1.05e+06	1.59e+06

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	30.25
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	2.06	Skew:	0.83
Prob(H) (two-sided):	0.01	Kurtosis:	4.45

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

Figure 105 SARIMAX Results for MANUAL ARIMA (1,0,2) - SPARKLING

7.2.1 DIAGONOSTIC PLOT

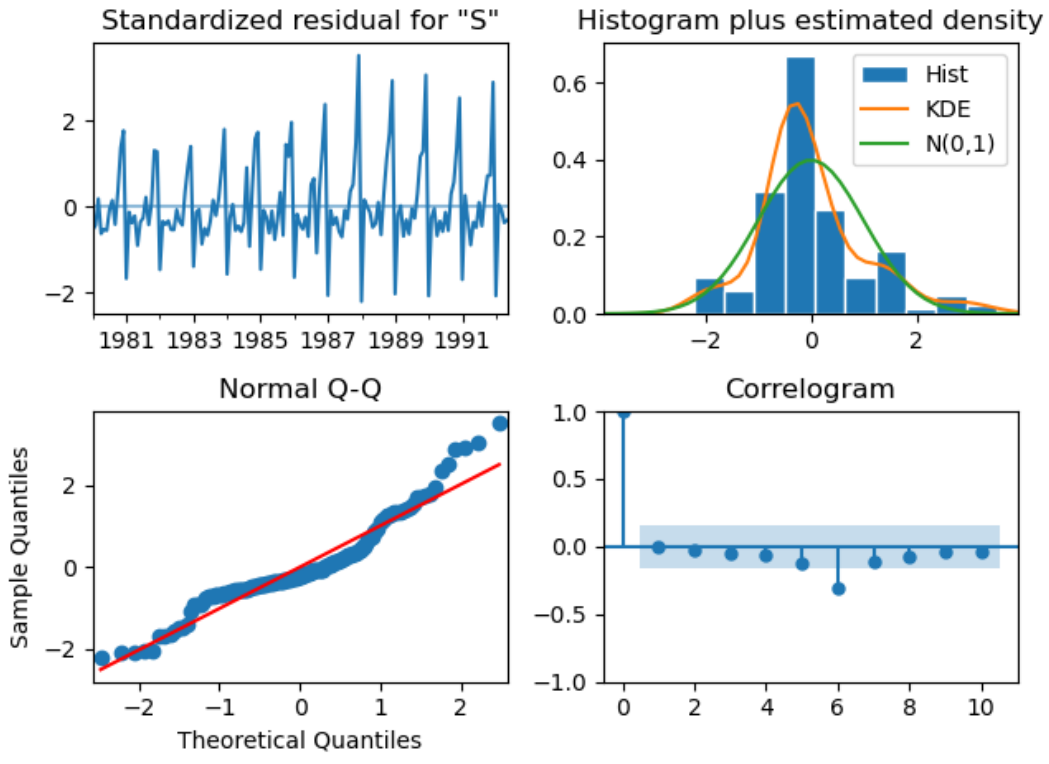


Figure 106 Diagnostics Plot

Model Performance:

The model seems to capture some of the fundamental dynamics but may not account adequately for all importance aspects, particularly outliers or non-normality in the residuals, suggesting potential improvements

Predict on the Test by using this model and evaluate the model

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(1,0,2)	1327.812923	41.651064

These values indicate that the model's predictions deviate, on average, by 1327.81 units, and the MAPE suggests that the error is about **41.65%** of the true values.

The high MAPE indicates that the model might not be performing as well for certain periods.

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(2,0,1)	1320.945957	0.291323
ARIMA(2,1,2)	1327.812923	0.291323
ARIMA(1,0,2)	1327.812923	41.651064

Figure 107 Test RMSE and Test MAPE ARIMA (2,0,1),ARIMA(2,1,2)& ARIMA(1,0,2) - SPARKLING

Insights:

- **ARIMA(2, 0, 1)** and **ARIMA(2, 1, 2)** models have nearly identical RMSE and MAPE values, both around **1320** for RMSE and **29.13%** for MAPE. These models seem to perform well overall.
- **ARIMA(1, 0, 2)** has the same RMSE as **ARIMA(2, 1, 2)** but a much higher **MAPE** (41.65%), which indicates that the model has large prediction errors relative to the true values.

Given that **ARIMA(2, 0, 1)** and **ARIMA(2, 1, 2)** perform similarly with lower errors, they might be the better choices for modeling.

8. BUILD DIFFERENT SARIMA MODELS

- Auto SARIMA

- Manual SARIMA

8.1 Build an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC)- SPARKLING

Let us look at the ACF plot once more to understand the seasonal parameter for the SARIMA model.

We see that there can be a seasonality of 6 as well as 12. But from the decompostion at the start we ascertained that visually it looks like the seasonality =6 and thus using the same

Setting the seasonality as 6 to estimate parametrs using auto SARIMA model.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

```
        param seasonal_param          AIC
0   (0, 1, 0)   (0, 0, 0, 6)  2552.801194
1   (0, 1, 0)   (0, 0, 1, 6)  2453.406173
2   (0, 1, 0)   (0, 0, 2, 6)  2245.286123
3   (0, 1, 0)   (1, 0, 0, 6)  2464.717657
4   (0, 1, 0)   (1, 0, 1, 6)  2391.043981
..        ...            ...          ...
76  (2, 1, 2)   (1, 0, 1, 6)  2226.843661
77  (2, 1, 2)   (1, 0, 2, 6)  2099.648985
78  (2, 1, 2)   (2, 0, 0, 6)  2007.233827
79  (2, 1, 2)   (2, 0, 1, 6)  2009.079677
80  (2, 1, 2)   (2, 0, 2, 6)  1978.620667

[81 rows x 3 columns]
```

	param	seasonal_param	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1976.643202
26	(0, 1, 2)	(2, 0, 2, 6)	1976.665612
80	(2, 1, 2)	(2, 0, 2, 6)	1978.620667
44	(1, 1, 1)	(2, 0, 2, 6)	1993.143115
17	(0, 1, 1)	(2, 0, 2, 6)	1993.922040

Figure 108 SORTED PARAMETER AND SEASONAL PARAM BASED ON LOWER AIC-SPARKLING

Insights:

- (1, 1, 2) with seasonal parameters (2, 0, 2, 6) has the lowest AIC (1976.64), suggesting it is the best model in terms of fit.
- (0, 1, 2) with seasonal parameters (2, 0, 2, 6) is also a close contender with a slightly higher AIC.
- Models with higher AIC values, such as (2, 1, 1) and (1, 1, 1), are less optimal compared to the first two.

SARIMAX Results

Dep. Variable:	y	No. Observations:	149
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-980.322
Date:	Sun, 05 Jan 2025	AIC	1976.643
Time:	03:50:45	BIC	1999.766
Sample:	0	HQIC	1986.039
	- 149		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5628	0.237	-2.371	0.018	-1.028	-0.098
ma.L1	-0.3236	0.190	-1.705	0.088	-0.696	0.048
ma.L2	-0.8108	0.211	-3.844	0.000	-1.224	-0.397
ar.S.L6	-0.0036	0.026	-0.141	0.888	-0.054	0.047
ar.S.L12	1.0161	0.017	60.404	0.000	0.983	1.049
ma.S.L6	-0.0027	0.118	-0.023	0.982	-0.234	0.228
ma.S.L12	-0.5463	0.085	-6.441	0.000	-0.713	-0.380
sigma2	1.24e+05	1.68e+04	7.357	0.000	9.09e+04	1.57e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	19.24
Prob(Q):	0.91	Prob(JB):	0.00
Heteroskedasticity (H):	1.24	Skew:	0.40
Prob(H) (two-sided):	0.48	Kurtosis:	4.69

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

Figure 109 SARIMAX Results for SARIMA (1,1,2) X (2,0,2,6) - SPARKLING

Conclusion:

The model (1, 1, 2) x (2, 0, 2, 6) seems to capture the seasonal and non-seasonal dynamics well, with significant seasonal AR and MA components at lags 12 and 6. However, the residual diagnostics suggest the presence of some non-normality and potential outliers.

8.1.1 DIAGONOSTIC PLOT

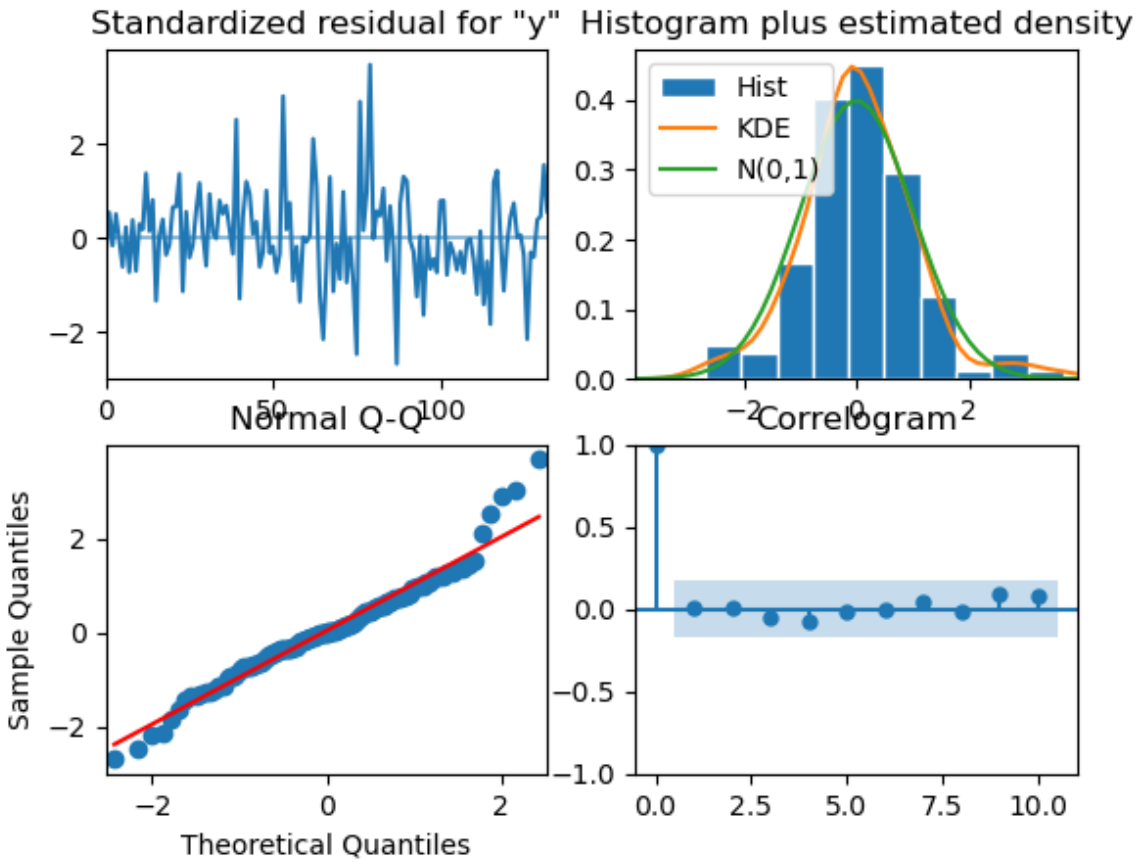


Figure 110 DIAGNOSTIC PLOT

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	1384.122018	379.058245	641.181509	2127.062527
1	2066.330213	391.806304	1298.403967	2834.256458
2	1803.477193	391.897824	1035.371572	2571.582813
3	2399.761089	395.343810	1624.901459	3174.620718
4	3314.973682	395.753585	2539.310908	4090.636455

- Mean: Predicted values for upcoming months, showing expected trends.
- Standard Error: Indicates prediction uncertainty; smaller values suggest higher accuracy.

- Confidence Interval: Likely range for true values, providing a 95% assurance of accuracy.

The forecasts show a rising trend in values over the next few periods, with consistent uncertainty levels.

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(2,0,1)	1320.945957	0.291323
ARIMA(2,1,2)	1327.812923	0.291323
ARIMA(1,0,2)	1327.812923	41.651064
SARIMA(1,1,2)(2,0,2,6)	316.997579	41.651064

Figure 111 Test RMSE and Test MAPE ARIMA (2,0,1),ARIMA(2,1,2), ARIMA(1,0,2), SARIMA(1,1,2)(2,0,2,6) -SPARKLING

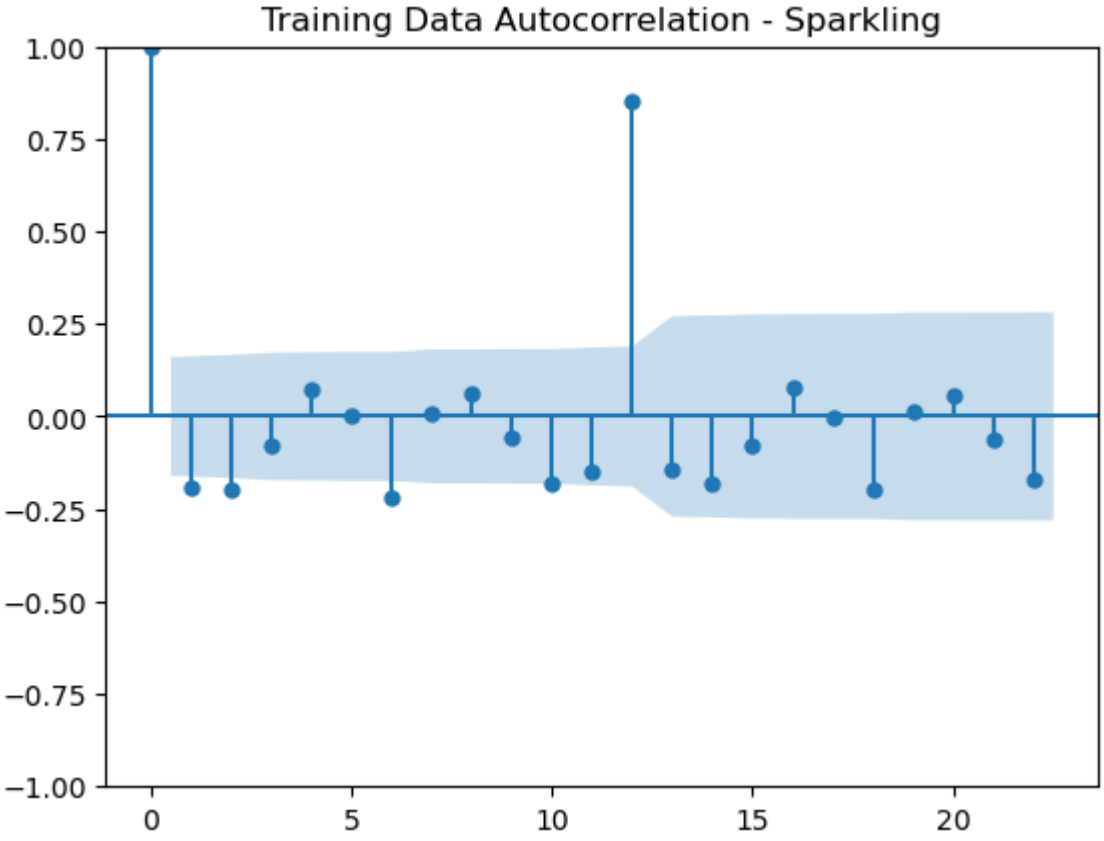
Model Comparison Insights :

1. ARIMA(2, 0, 1)
    - Test RMSE: 1320.95
    - Test MAPE: 29.13%
2. ARIMA(2, 1, 2)
    - Test RMSE: 1327.81
    - Test MAPE: 29.13%
3. SARIMA(1, 1, 2) x (2, 0, 2, 6)
    - Test RMSE: 316.99 (lowest, best fit)
    - Test MAPE: 41.65% (highest, less accurate forecasting)

Conclusion:

- SARIMA offers better fit (low RMSE) but lower forecasting accuracy (high MAPE).
- ARIMA(2, 0, 1) and ARIMA(2, 1, 2) provide balanced forecasting with lower MAPE.
- Choose SARIMA if minimizing RMSE is critical, but opt for ARIMA models for more reliable percentage-based forecasting.

8.2 SARIMA MODEL FOR WHICH THE BEST PARAMETERS ARE SELECTED AT THE ACF AND THE PACF PLOTS



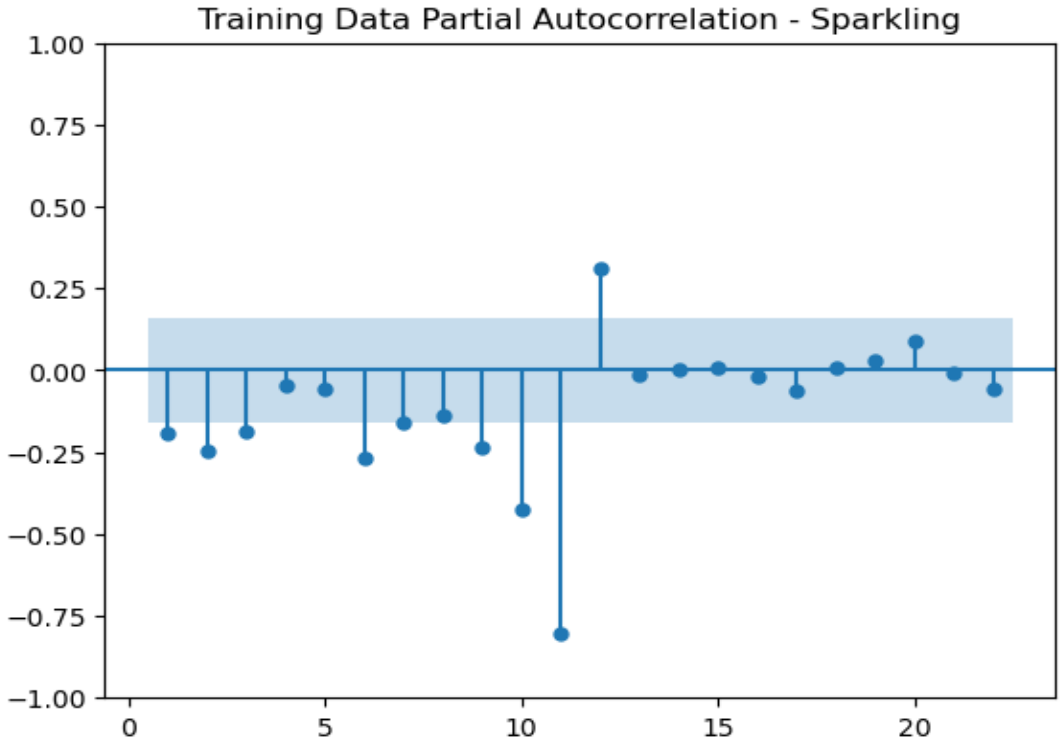


Figure 112 ACF AND PACF PLOT

OBSERVATION

Moving Average (MA):

- Seasonal MA order (Q): 2
- Non-seasonal MA order (q): 2

Autoregressive (AR):

- Seasonal AR order (P): 1
- Non-seasonal AR order (p): 1
- P = 1 (Seasonal AR order)
- Q = 2 (Seasonal MA order)
- D = 1 (Seasonal differencing)
- M = 12 (Seasonal period)
- p = 1 (Non-seasonal AR order)
- q = 2 (Non-seasonal MA order)
- d = 1 (Non-seasonal differencing)

These values indicate that the model includes both seasonal and non-seasonal MA and AR terms for capturing patterns effectively over a 12-month seasonal period.

This suggests a SARIMA(1, 1, 2)(1, 1, 2, 12) model as a potential candidate based on the ACF and PACF analysis

8.3 Build a Manual Version of a SARIMA Model - SPARKLING

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	149
Model:	SARIMAX(1, 1, 2)x(1, 1, 2, 12)	Log Likelihood	-808.032
Date:	Sun, 05 Jan 2025	AIC	1630.065
Time:	15:20:10	BIC	1648.904
Sample:	01-01-1980	HQIC	1637.705
	- 05-01-1992		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5246	0.228	-2.304	0.021	-0.971	-0.078
ma.L1	-0.2049	0.187	-1.098	0.272	-0.571	0.161
ma.L2	-0.7348	0.159	-4.628	0.000	-1.046	-0.424
ar.S.L12	-0.1071	0.812	-0.132	0.895	-1.699	1.485
ma.S.L12	-0.3252	0.825	-0.394	0.693	-1.942	1.292
ma.S.L24	-0.0942	0.389	-0.242	0.809	-0.857	0.669
sigma2	1.585e+05	2.07e+04	7.668	0.000	1.18e+05	1.99e+05

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	12.45
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	0.75	Skew:	0.49
Prob(H) (two-sided):	0.38	Kurtosis:	4.33

Figure 113 SARIMAX Results for ARIMA (1,1,2) X (1,1,2,12) - SPARKLING

Observations:

- **AR(1)** is significant, indicating that the most recent value in the time series has a significant influence on the current value.
- **MA(2)** is also significant, capturing the effect of past errors.
- Seasonal components (especially at lags 12 and 24) do not show significance, suggesting that seasonal factors may not strongly influence the data after differencing.

8.3.1 DIAGONOSTIC PLOT

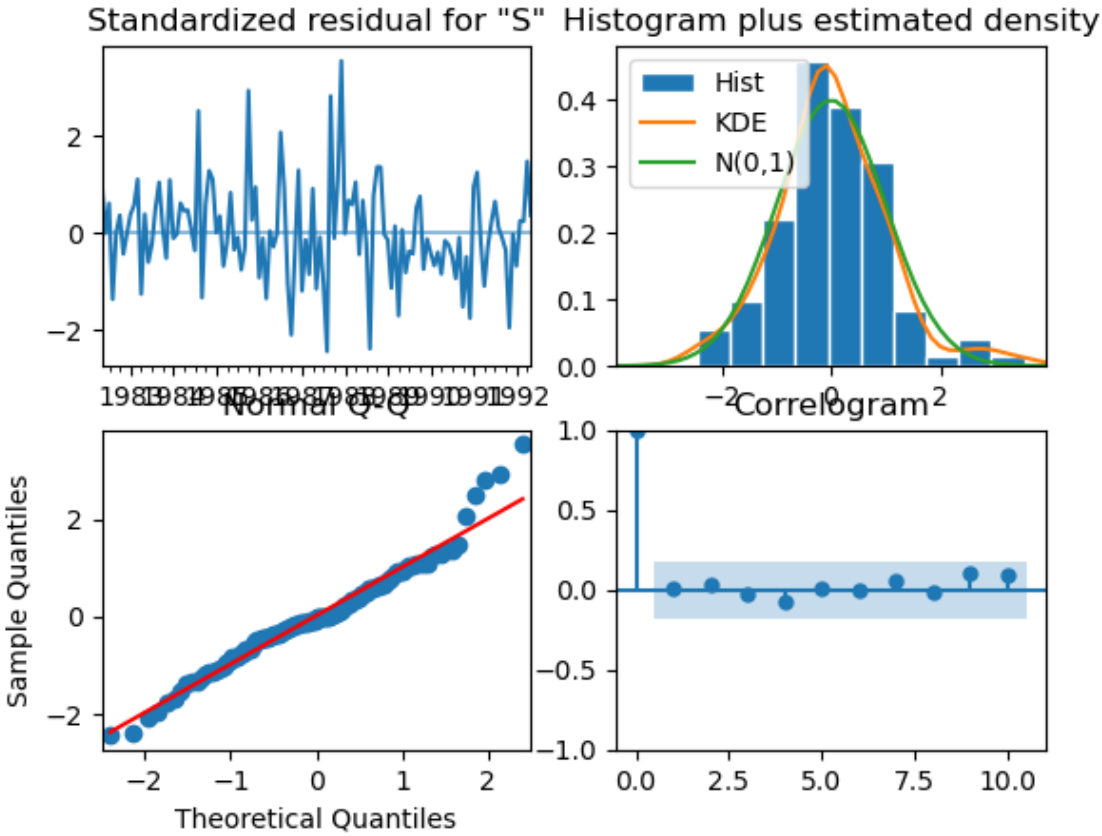


Figure 114 DIAGNOSTIC PLOT

Predict on the Test by using this model and evaluate the model

RMSE: 294.1666272635208

MAPE: 10.922985045468526

8.4 CHECK THE PERFORMANCE OF THE MODELS BUILT

	Test RMSE Sparkling	Test MAPE Sparkling
ARIMA(2,0,1)	1320.945960	0.282300
ARIMA(2,1,2)	1326.416273	0.282300
ARIMA(1,0,2)	1326.416273	41.351452
SARIMA(1,1,2)(2,0,2,6)	317.016897	41.351452
SARIMA(1,1,2)(1,1,2,12)	294.166627	10.922985

Figure 115 Test RMSE and Test MAPE ARIMA (2,0,1) , ARIMA(2,1,2), ARIMA (1,O,2),SARIMA(1,1,2)(2,0,2,6),SARIMA(1,1,2)(1,1,2,12) - SPARKLING

From the results, the **best performing model** for forecasting Sparkling wine sales is:

- **SARIMA(1,1,2)(1,1,2,12)** with an **RMSE of 294.1666** and a **MAPE of 10.9230**.

This model has the lowest RMSE and MAPE, which suggests it provides the most accurate forecast. It effectively captures the seasonal and non-seasonal patterns in the data, making it the optimal choice among the models tested.

F. COMPARE THE PERFORMANCE OF ALL THE MODELS BUILT

1. COMPARE THE PERFORMANCE OF ALL THE MODELS BUILT – ROSE

1.1 CHOOSE THE BEST MODEL WITH PROPER RATIONALE – ROSE

	Test RMSE Rose	Test RMSE Sparkling	Test MAPE Rose
RegressionOnTime	17.510241	1349.042457	NaN
Simple Exponential Smoothing	20.313631	1329.402402	NaN
Double Exponential Smoothing	14.623742	1340.452791	NaN
Triple Exponential Smoothing (Additive Season)	13.877335	304.247029	NaN
SimpleAverageModel	52.239499	1331.037637	NaN
2pointTrailingMovingAverage	11.529409	813.400684	NaN
4pointTrailingMovingAverage	14.455221	1156.589694	NaN
6pointTrailingMovingAverage	14.572009	1283.927428	NaN
9pointTrailingMovingAverage	14.731209	1346.278315	NaN
Triple Exponential Smoothing (Multiplicative Season)	8.405441	318.695471	NaN
ARIMA(2,0,2)	26.135542	NaN	56.709217
ARIMA(1,1,2)	20.915405	NaN	56.709217
ARIMA(2,0,2)	26.135542	NaN	0.567092
SARIMA(1,1,2)(2,0,2,6)	20.562217	NaN	0.567092
SARIMA(2,0,2)(1,0,2,12)	14.121719	NaN	0.291323

Figure 116 OVERALL TEST RMSE AND TEST MAPE TO COMPARE THE PERFORMANCE OF MODEL BUILT FOR ROSE

OBSERVATION:

- Triple Exponential Smoothing (Multiplicative Season) and SARIMA(2,0,2)(1,0,2,12) for Rose have the best performance in terms of RMSE (8.41 and 14.12 respectively) and MAPE (0.2913).
- RegressionOnTime and Simple Average Model have much higher RMSE values, indicating they don't perform as well compared to other models.

1.2. BUILDING THE MOST OPTIMUM MODEL ON THE FULL DATA-ROSE

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(2, 0, 2)x(1, 0, 2, 12)	Log Likelihood	-664.014
Date:	Sun, 05 Jan 2025	AIC	1344.028
Time:	14:06:00	BIC	1368.630
Sample:	01-01-1980 - 07-01-1995	HQIC	1354.018
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3233	0.196	6.736	0.000	0.938	1.708
ar.L2	-0.3268	0.196	-1.671	0.095	-0.710	0.057
ma.L1	-1.1681	2.586	-0.452	0.651	-6.236	3.900
ma.L2	0.1683	0.486	0.346	0.729	-0.785	1.121
ar.S.L12	0.3494	0.056	6.295	0.000	0.241	0.458
ma.S.L12	0.2486	0.084	2.974	0.003	0.085	0.412
ma.S.L24	0.3460	0.097	3.563	0.000	0.156	0.536
sigma2	224.0524	578.166	0.388	0.698	-909.133	1357.237

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	5.82
Prob(Q):	0.84	Prob(JB):	0.05
Heteroskedasticity (H):	0.29	Skew:	0.38
Prob(H) (two-sided):	0.00	Kurtosis:	3.55

Figure 117 SARIMAX RESULTS ON THE FULL DATA-ROSE

- **AR(1)** and **seasonal AR(12)** are highly significant, capturing the autoregressive dynamics and yearly seasonality.
- **MA(1)** and **MA(2)** are not significant, indicating limited contribution to the model.
- Significant seasonal components (AR and MA at lags 12 and 24) highlight yearly seasonal effects.
- Residual diagnostics show no autocorrelation but some heteroskedasticity.

The model effectively captures the seasonal pattern in Rose sales, but improvements could address heteroskedasticity and refine the MA terms.

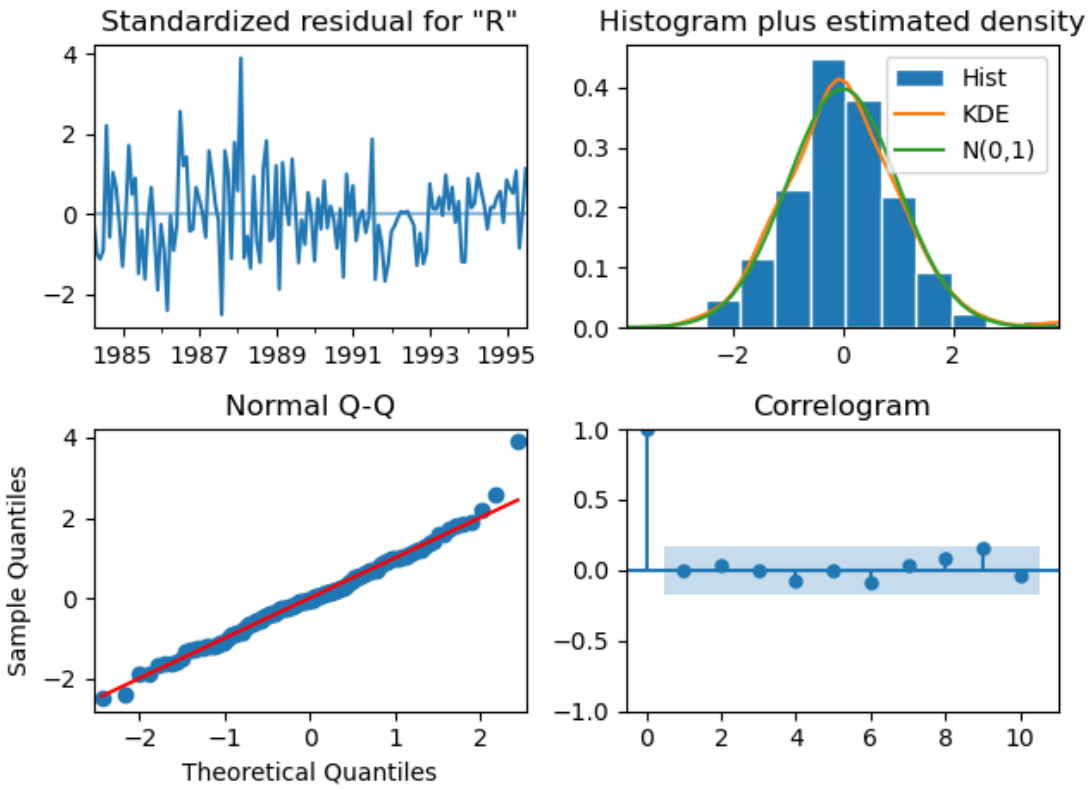


Figure 118 Diagnostics Plot

1.3. MAKE A FORECAST FOR THE NEXT 12 MONTHSPOINTS – ROSE

Evaluate the model on the whole and predict 12 months into the future (till the end of next year)

	Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	50.112544	15.010045	20.693396	79.531692
1995-09-01	49.426759	15.202513	19.630382	79.223136
1995-10-01	49.509797	15.223225	19.672823	79.346770
1995-11-01	57.808321	15.225277	27.967327	87.649315
1995-12-01	68.839497	15.225228	38.998598	98.680397

OBSERVATION:

- Seasonal Trends: Sales show a steady increase, peaking in December, aligning with holiday demand.
- Sales Variability: Confidence intervals show forecast uncertainty, especially for August-October, suggesting market factors may impact demand.
- Inventory Management: Prepare for high sales in November and December by increasing production and stock. Align inventory with forecasted growth.
- Marketing Strategy: Focus marketing efforts on the holiday season, particularly December, to capitalize on rising demand.

- Sales Risk: Uncertainty in early months highlights the need for real-time updates to forecasts and flexible production schedules.

Conclusion: Focus on seasonal demand, adapt to forecast uncertainty, and strategize for the holiday peak to maximize revenue.

```
RMSE of the Full Model 31.43756133070178
```

- Model Performance: RMSE of 31.44 indicates reasonable accuracy, but some forecast variability exists.
- Forecast Confidence: Moderate confidence in forecasts, useful for general trends but not precise for fluctuations.
- Improvement Areas: To reduce RMSE, refine the model with additional features or seasonal adjustments.
- Impact on Business: Factor in forecast error for inventory planning, especially for volatile demand or perishable goods.

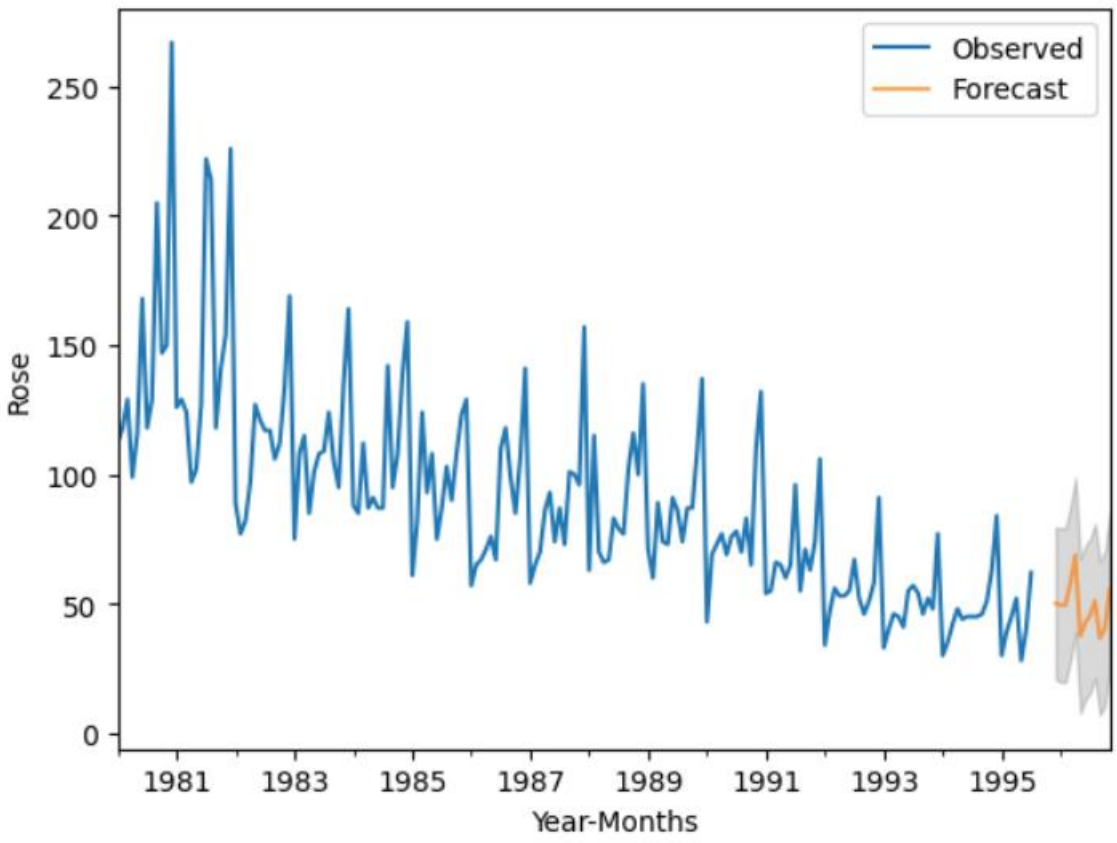


Figure 119 FORECAST FOR THE NEXT 12 MONTHS - ROSE

Observation:

- Sales Trend: Declining trend from 250 units in the early 1980s to under 50 units by the 1990s, with smaller fluctuations over time.
- Forecast: Predicts stabilized sales around 50 units, closely following the declining trend, with minimal uncertainty.
- Seasonality: Early years show high variability and seasonal spikes, which decrease over time.

2. COMPARE THE PERFORMANCE OF ALL THE MODELS BUILT – SPARKLING

2.1 CHOOSE THE BEST MODEL WITH PROPER RATIONALE – SPARKLING

	Test RMSE Rose	Test RMSE Sparkling	Test MAPE Sparkling
RegressionOnTime	17.510241	1349.042457	NaN
Simple Exponential Smoothing	20.313631	1329.402402	NaN
Double Exponential Smoothing	14.623742	1340.452791	NaN
Triple Exponential Smoothing (Additive Season)	13.877335	304.247029	NaN
SimpleAverageModel	52.239499	1331.037637	NaN
2pointTrailingMovingAverage	11.529409	813.400684	NaN
4pointTrailingMovingAverage	14.455221	1156.589694	NaN
6pointTrailingMovingAverage	14.572009	1283.927428	NaN
9pointTrailingMovingAverage	14.731209	1346.278315	NaN
Triple Exponential Smoothing (Multiplicative Season)	8.405441	318.695471	NaN
ARIMA(2,0,1)	NaN	1320.945957	0.291323
ARIMA(2,1,2)	NaN	1327.812923	0.291323
ARIMA(1,0,2)	NaN	1327.812923	41.651064
SARIMA(1,1,2)(2,0,2,6)	NaN	316.997579	41.651064
SARIMA(1,1,2)(1,1,2,12)	NaN	294.166627	10.922985

Figure 120 OVERALL TEST RMSE AND TEST MAPE TO COMPARE THE PERFORMANCE OF MODEL BUILT FOR SPARKLING

Based on the performance metrics for both **RMSE** and **MAPE**, the **best-performing model** for **Sparkling** wine sales is:

- **SARIMA(1,1,2)(1,1,2,12)**
  - RMSE: **294.1666**
  - MAPE: **10.9230**

- This model has the lowest **RMSE** and relatively low **MAPE**, making it the most accurate model for forecasting **Sparkling** wine sales compared to the other models listed.

2.2 BUILDING THE MOST OPTIMUM MODEL ON THE FULL DATA-SPARKLING

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(1, 1, 2)x(1, 1, 2, 12)	Log Likelihood	-1086.479
Date:	Sun, 05 Jan 2025	AIC	2186.959
Time:	16:02:23	BIC	2207.892
Sample:	01-01-1980 - 07-01-1995	HQIC	2195.464
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5964	0.337	-1.772	0.076	-1.256	0.063
ma.L1	-0.2476	0.307	-0.807	0.420	-0.849	0.354
ma.L2	-0.6873	0.281	-2.447	0.014	-1.238	-0.137
ar.S.L12	-0.1349	0.757	-0.178	0.859	-1.618	1.348
ma.S.L12	-0.4133	0.768	-0.538	0.591	-1.919	1.092
ma.S.L24	-0.0863	0.430	-0.201	0.841	-0.928	0.756
sigma2	1.517e+05	1.61e+04	9.436	0.000	1.2e+05	1.83e+05

Figure 121SARIMAX RESULTS ON THE FULL DATA-SPARKLING

Insights:

- The **AR(1)** and **MA(2)** terms are important in modeling, with **MA(2)** being statistically significant.
- The seasonal components (**AR(12)**, **MA(12)**, **MA(24)**) do not appear significant.
- The residuals exhibit non-normality, which could suggest further improvements or transformations might be necessary.

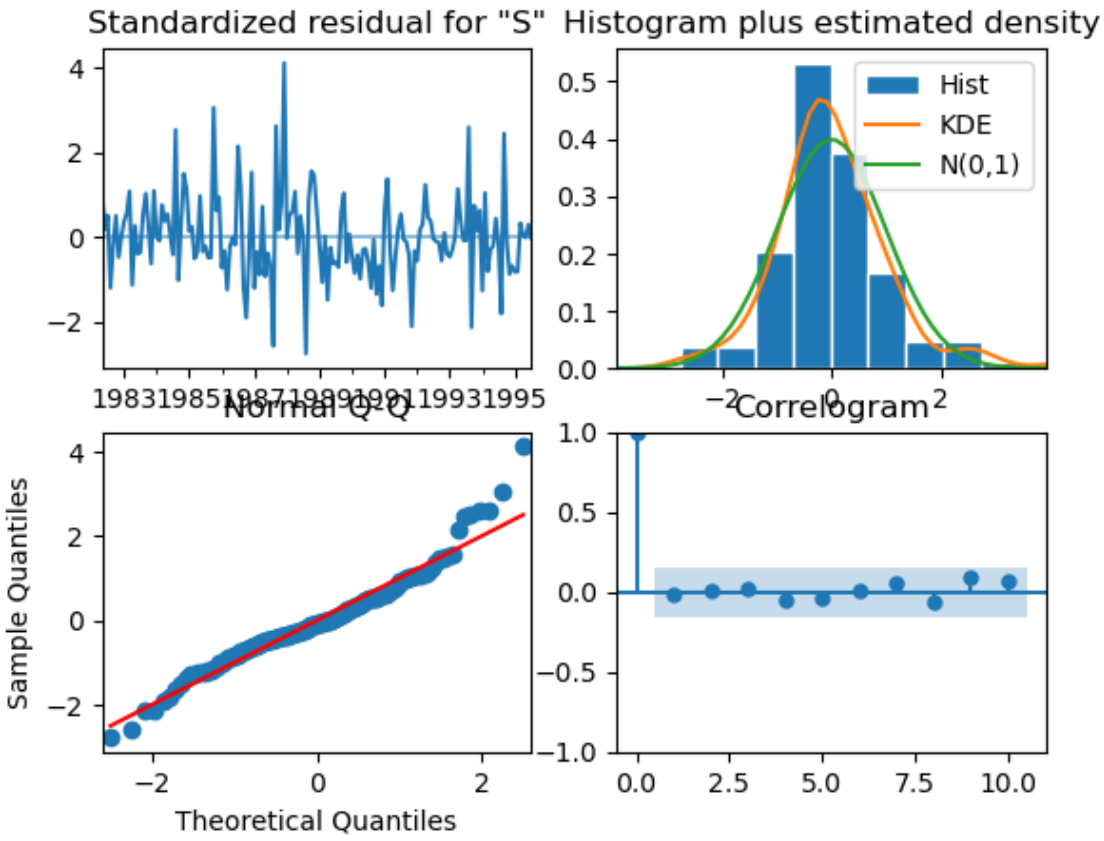


Figure 122 DIAGNOSTIC PLOT

2.3. MAKE A FORECAST FOR THE NEXT 12 MONTHSPOINTS-SPARKLING

Evaluate the model on the whole and predict 12 months into the future (till the end of next year)

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	1876.060470	389.457895	1112.737023	2639.383917
1995-09-01	2478.217287	394.168593	1705.661040	3250.773534
1995-10-01	3293.555600	394.318087	2520.706350	4066.404849
1995-11-01	3933.325253	395.602487	3157.958626	4708.691879
1995-12-01	6132.783697	395.653821	5357.316457	6908.250936

Trend of Increasing Sales:

- The forecasted mean values for each month show a clear upward trend, starting at 1876.06 in August 1995 and reaching **613

```
RMSE of the Full Model 554.1755396135478
```

RMSE Analysis:

1. Model Accuracy:

- RMSE of 554.18 implies the model's predictions deviate by about 554 units of sparkling wine sales, indicating moderate accuracy.

2. Error Magnitude:
   - Lower RMSE suggests better performance. This value shows room for improvement in prediction accuracy.
3. Model Comparison:
   - If 554.18 is higher than other models, the current model may be overfitting or missing key data patterns. A simpler or alternate model might perform better.

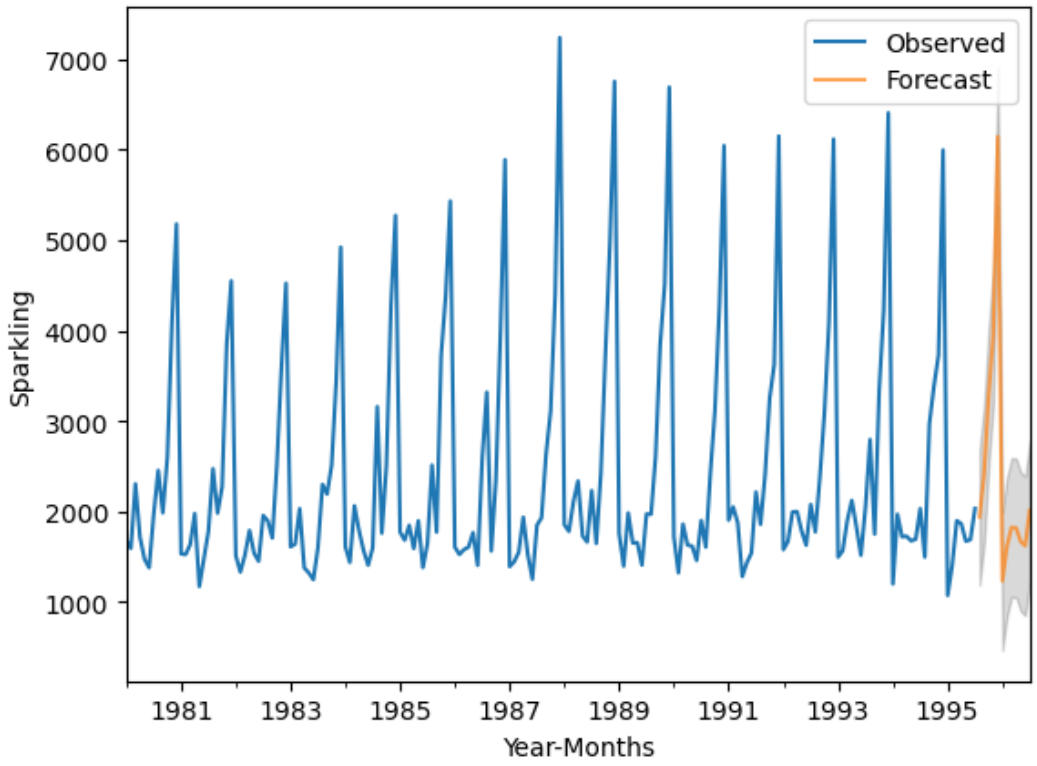


Figure 123 FORECAST FOR THE NEXT 12 MONTHSPOINTS – SPARKLING

Time-Series Analysis

1. Seasonal Pattern:
   - Clear, consistent peaks and troughs indicate a strong seasonal cycle.
2. Trend:
   - Peaks show a gradual upward trend, suggesting growth in the "Sparkling" variable over time.
3. Forecasting:
   - Forecast (orange line) follows the observed seasonal and trend patterns but slightly underestimates peak values.

4. Uncertainty Band:
    - Prediction interval widens over time, reflecting higher uncertainty for long-term forecasts.
5. Outliers/Irregularities:
    - Minor deviations from the seasonal pattern, but trend and seasonality remain dominant.

G. ACTIONABLE INSIGHTS & RECOMMENDATIONS:

Understanding Seasonal Demand:

- Insight: Sparkling wine sales peak during holiday months, especially December.
- Action: Adjust production, marketing, and distribution efforts for Q4.
- Recommendation: Launch targeted marketing campaigns and stock up before November and December.

Monitor Sales Growth and Stability:

- Insight: Upward sales trend since the mid-1990s suggests growing demand.
- Action: Plan for production and distribution expansion.
- Recommendation: Conduct market research to identify growth drivers (e.g., demographic or preference shifts).

Model Improvement & Forecast Accuracy:

- Insight: RMSE of 554.18 indicates room for better prediction accuracy.
- Action: Incorporate granular data like regional sales and promotional impacts.
- Recommendation: Refine models with external factors (e.g., economic data) and explore advanced machine learning methods.

Addressing Fluctuations and Volatility:

- Insight: High sales volatility during peak months (e.g., November and December).
- Action: Ensure scalable production to meet unexpected demand.
- Recommendation: Develop flexible inventory and production strategies using automated forecasting tools.

Improving Forecasting Models:

- Insight: Seasonal models like SARIMA and SARIMAX outperform simpler models.
- Action: Incorporate external regressors for improved accuracy (e.g., promotional spend, macroeconomic factors).
- Recommendation: Use hybrid models combining SARIMAX and external variables.

Strategic Inventory Management:

- Insight: Demand spikes during holiday months require inventory optimization.
- Action: Avoid stockouts or overstocking during peak/off-peak periods.
- Recommendation: Implement just-in-time inventory systems based on demand forecasting.

Customer Insights for Targeted Marketing:

- Insight: Peak month sales indicate purchases for special occasions.
- Action: Tailor marketing to emphasize holiday and celebration themes.
- Recommendation: Strengthen digital marketing, use personalized offers, and partner with event organizers.

Conclusion:

- Leverage forecast insights to optimize marketing, production, and distribution strategies.
- Align with demand forecasts to improve resource allocation and customer satisfaction.
- Continuously improve forecasting accuracy and adapt to changing market dynamics.

1.KEY TAKEAWAYS (ACTIONABLE INSIGHTS AND RECOMMENDATIONS) FOR THE BUSINESS:

1. Leverage Seasonal Demand
    - Boost production, marketing, and inventory for peak Q4 sales.
2. Optimize Inventory
    - Use just-in-time inventory to avoid stockouts or overstocking.
3. Expand for Growth
    - Scale production and explore new markets to meet rising demand.
4. Improve Forecast Accuracy
    - Refine models with regional, promotional, and economic data.
5. Targeted Marketing
    - Focus holiday campaigns on celebrations; personalize offers.
6. Build Flexibility
    - Ensure scalable production to handle demand spikes.
7. Understand Growth Drivers
    - Conduct market research to target key customer segments.
8. Enhance Customer Experience
    - Use loyalty programs and targeted promotions to boost retention.