Time Series Forecasting
ABC Estate Wines
Project - Coded

By: Agnes Raja Kumari. E

PGP-Data Science and Business Analytics

PGPDSBA.O.MAY24.A

Table of Contents

Contents

Table of Figure

Introduction

1.1 Context

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties. Our objective is to delve into the data, 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

The problem is to:

1. Identify patterns and trends in the historical wine sales data.
2. Analyze key factors that have influenced sales over time.
3. Develop accurate forecasting models to predict future sales trends.

Addressing these problems will enable ABC Estate Wines to make data-driven decisions, optimize sales strategies, and stay competitive in the evolving wine industry.

2. Data Background

2.1 Data Dictionary

- YearMonth: Represents the year and month of the sales data (e.g., "1980-01").
- Rose: Monthly sales quantity of Rose wine.
- Sparkling: Monthly sales quantity of Sparkling wine.

File Name: Rose.csv and Sparkling.csv

2.2 Overview of the Dataset

The initial steps to get an overview of any dataset is to:

- Observe the first few rows of the dataset, to check whether the dataset has been loaded properly or not
- Get information about the number of rows and columns in the dataset
- Convert 'YearMonth' to datetime format for proper alignment.
- Set 'YearMonth' as the index for both datasets for alignment.

- Check the statistical summary of the dataset to get an overview of the numerical columns of the data.

2.2.1 Displaying first 5 rows, last 5 rows and shape of the dataset

Displaying first 5 rows, last 5 rows and shape of the Rose dataset

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

	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 1 Displaying first 5 rows, last 5 rows and shape of the Rose dataset

Rows: 187

Columns: **YearMonth** (Object): A unique identifier for each time point in the format YYYY-MM.

Rose (Float64): Monthly sales of Rose wine.

Displaying first 5 rows, last 5 rows and shape of the Sparkling dataset

	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 2 Displaying first 5 rows, last 5 rows and shape of the Sparkling dataset

Rows: 187

Columns:

1. **YearMonth** (Object): A unique identifier for each time point in the format YYYY-MM.
2. **Sparkling** (Int64): Monthly sales of Sparkling wine. No missing values.

2.3 Find out the data types of the columns

```
<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 3 Data information of both Rose and Sparkling data

The dataset of Rose has two columns:

- YearMonth: This appears to represent dates in a string format (e.g., "1980-01").
- Rose: This contains sales data, but there are two missing values.

The dataset of Sparkling has two columns:

- YearMonth (Object): A unique identifier for each time point in the format YYYY-MM.
- Sparkling (Int64): Monthly sales of Sparkling wine. 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 4 Converting year month to date time series

2.5 Plotting the Data

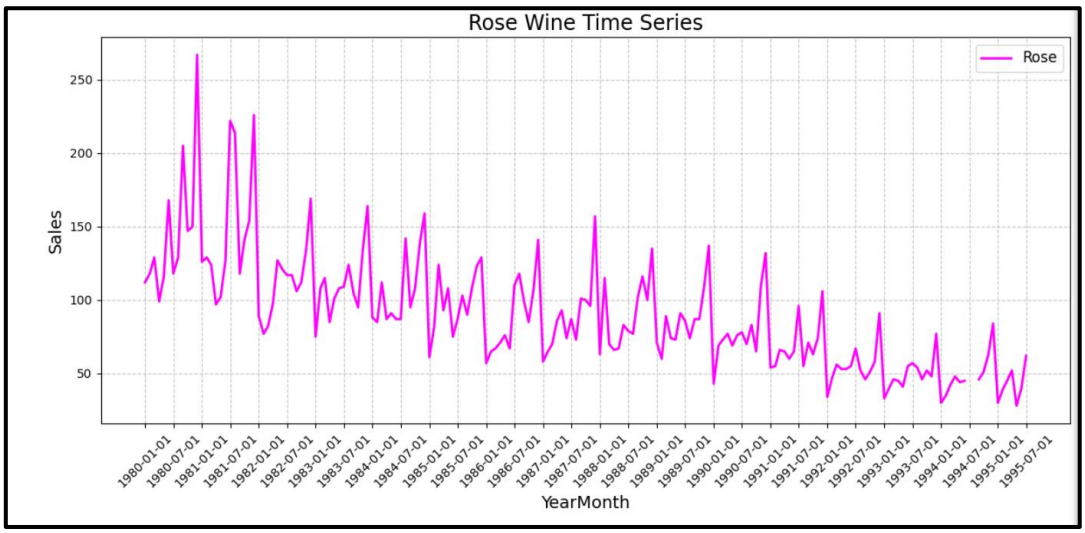


Figure 5 Rose wine Time series

Observations on Rose Wine Sales (1980–1995):

- Sales of Rose wine exhibit an overall downward trend over the 15-year period, indicating a gradual reduction in demand.
- A recurring pattern is evident, with regular peaks and troughs suggesting seasonal fluctuations in sales, likely driven by changes in consumer demand during specific times of the year.
- Between 1980 and 1985, sales display significant volatility with larger fluctuations, compared to the more stable patterns observed in later years.
- The highest sales values gradually decrease over time, reflecting a decline in peak demand for Rose wine.
- After 1990, sales variability reduces, with the data appearing more stable, albeit at consistently lower sales levels.
- Outliers in the early years, characterized by unusually high peaks, could be linked to special events or external factors impacting sales.

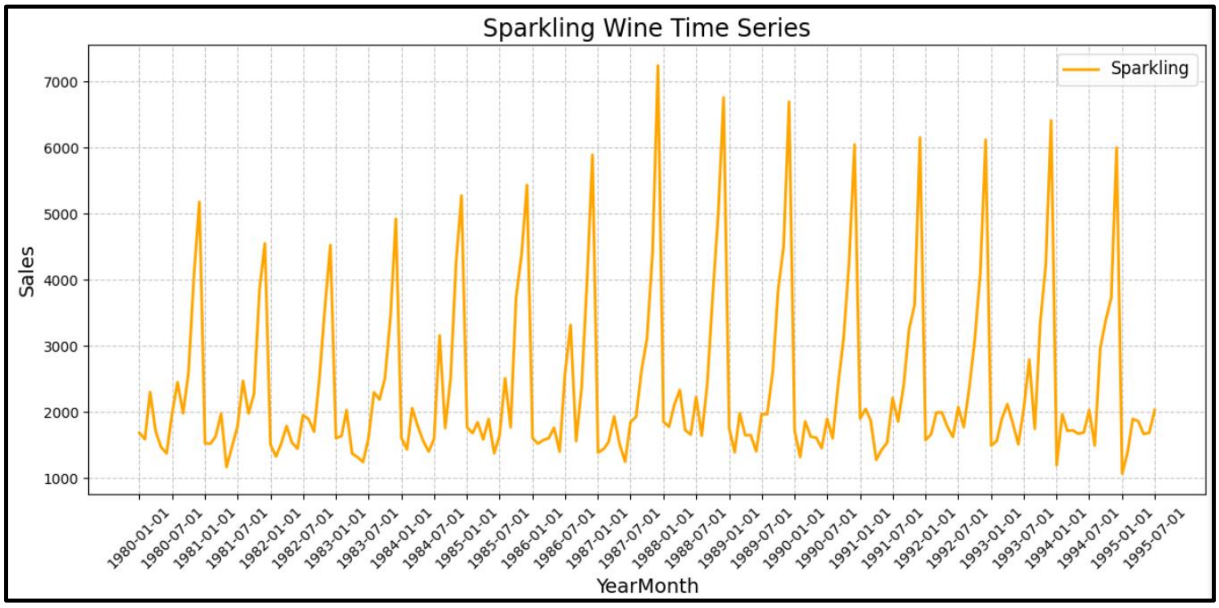


Figure 6 Sparkling Wine Time Series

Observations on Sparkling Wine Sales (Time Series Analysis):

- The sales display a distinct seasonal pattern, with prominent peaks recurring annually.
- These peaks are likely driven by holidays, festivals, or special events that boost demand for sparkling wine.
- The overall trend in sales remains stable over time, with peak magnitudes showing little variation.

- No significant long-term upward or downward trend in sales is evident.
- Sales during peak periods are substantially higher than off-peak periods, underscoring the highly seasonal nature of sparkling wine demand.
- Sales outside the peak seasons are relatively low, indicating limited demand during these periods.
- The intervals between peaks are consistent, reflecting a predictable annual demand cycle.

2.6 Performing EDA

2.6.1 Statistical Summary

```
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 7 Statistical Summary of Rose Wine data and Sparkling wine data

Observation for Summary Statistics for Rose dataset

- The dataset contains 185 observations, representing 185 time periods.
- The average sales for Rose wine are approximately **90.39 units**.
- The **standard deviation** is **39.18**, indicating moderate variability in sales.
- Sales range from a **minimum of 28** units to a **maximum of 267** units, showing significant fluctuations.
- 25% of sales are below **63** units.

- The median sales value is **86**, representing the central tendency.
- 75% of sales are below **112** units, with the top 25% contributing to higher sales figures.

Observation for Summary Statistics for Sparkling dataset

- The dataset contains **187 observations**, slightly more than the Rosé dataset.
- The average sales for Sparkling wine are **2402.42 units**, significantly higher than Rosé wine sales.
- The **standard deviation** is **1295.11**, indicating high variability in sales.
- Sales range from a **minimum of 1070** units to a **maximum of 7242** units, showcasing much larger fluctuations than Rosé wine.
- 25% of sales are below **1605** units.
- The median sales value is **1874**, meaning half of the sales are below this value.
- 75% of sales are below **2549** units, with the top 25% of sales likely reflecting seasonal peaks.

Comparison and Insights

- Sparkling wine has a significantly **higher average sales (2402.42 units)** compared to Rose wine (**90.39 units**).
- Sparkling wine sales exhibit **greater variability** (standard deviation of 1295.11 vs. 39.18 for Rosé).
- Sparkling wine has a **broader sales range** (1070–7242 units) compared to Rose wine (28–267 units).
- 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**, such as holidays or celebrations.
- Rose wine appears to have a **more consistent year-round demand**, reflecting **greater stability** and less volatility.

2.6.2 Checking the Missing Values

```
Missing Values in Rose Data:
Rose    2
dtype: int64

Missing Values in Sparkling Data:
Sparkling    0
dtype: int64
```

Figure 8 Finding the Missing Value

Observation:

- The dataset contains **2 missing values** in the Rose sales column.
- The Sparkling sales column has **no missing values**, ensuring a complete dataset for analysis.

2.6.3 Visualization of Distribution of Rose Sales and Distribution of Sparkling Sales

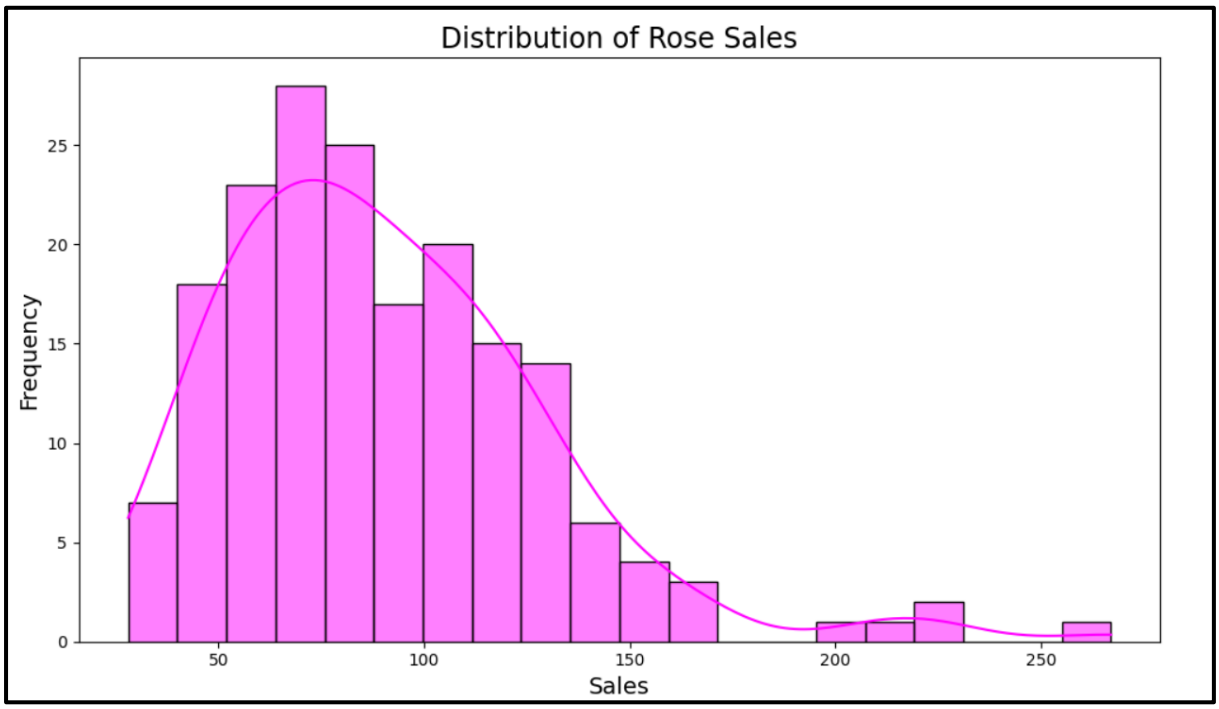


Figure 9 Visualization of Distribution of Rose Sales

Observations: Rose Wine Sales Distribution

- **Right-Skewed Distribution:** The sales distribution is **positively skewed**, with most sales concentrated in the range of **50–100 units** and a **long tail** extending towards higher values.
- **Common Sales Range:** The most frequent sales figures fall between **60 and 100 units**, indicating this as the **typical sales range** for Rosé wine.
- **Rare High Sales:** Sales above **150 units** are uncommon, suggesting that **very high sales are exceptional** rather than routine.
- **Mean vs. Median:** In a right-skewed distribution, the **mean sales** (90.39 units) exceed the **median sales** (86 units), influenced by a few **higher sales values**.
- **Overall Sales Range:** Sales span from **28 to 267 units**, reflecting a **broad range**, but most observations lie within a narrower band of sales.

Insights:

- **Stable Demand Pattern:** The concentration of sales around a specific range indicates a **consistent baseline demand** for Rosé wine.
- **Exceptional High Sales:** The rarity of sales exceeding 150 units suggests these are tied to **special occasions** or **promotional events**.
- **Strategic Implications:** The steady demand with occasional **demand spikes** provides opportunities to **optimize inventory**, **target peak sales periods**, and tailor **marketing efforts** to maximize profitability.

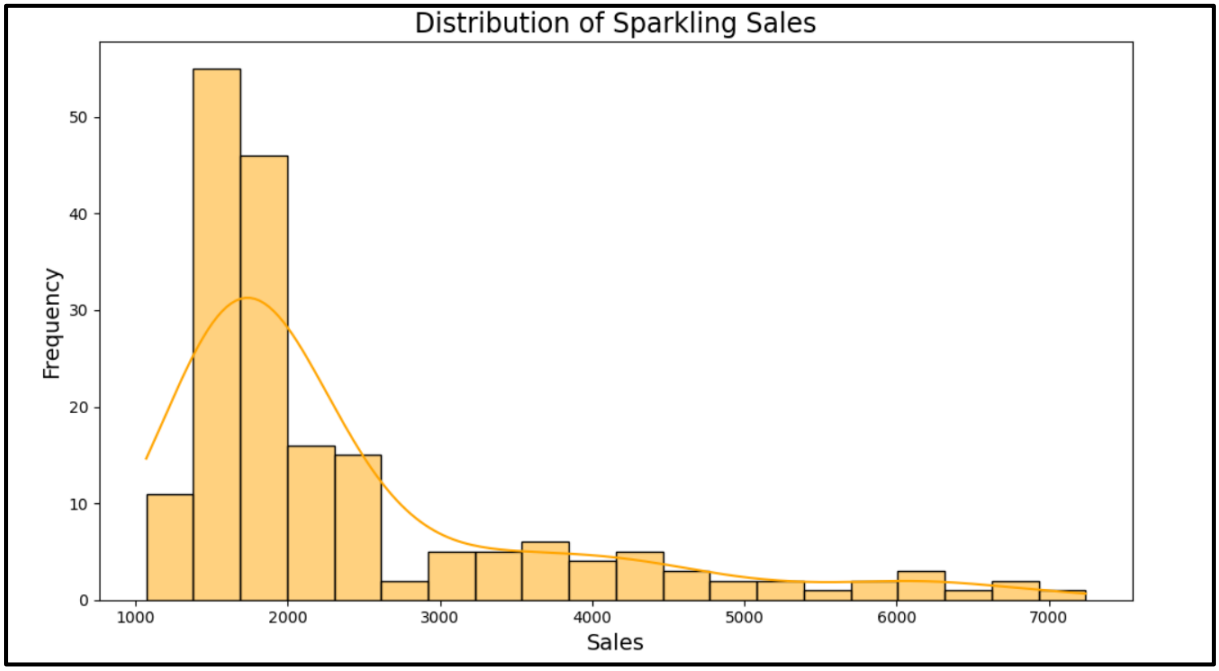


Figure 10 Visualization of Distribution of Sparkling Sales

Observations: Sales Distribution

- **Right-Skewed Distribution:** The sales data shows a **right-skewed distribution**, with most sales clustered at the lower end and a **declining frequency** as sales increase.
- **Common Sales Range:** The **1000–2000 range** has the **highest frequency** of sales, identifying it as the **most typical sales bracket**.
- **Long Tail:** A **long tail** stretches towards higher sales values, indicating that **high sales are rare but present**.
- **Density Curve:** The **density curve** illustrates the skewness, with a **steep rise** at lower sales values and a **gradual decline** as sales increase.

Insights:

- **Typical Sales Range:** Most sales fall between **1000 and 2000**, which represents the **core operational focus** for the business.
- **Impact of Skewness:** The **right skewness** means that a small number of high sales values are **increasing the mean**, emphasizing the importance of examining both median and mean sales.
- **High Sales Events:** Sales exceeding **5000 units** are **infrequent but significant**, contributing disproportionately to overall revenue.
- **Potential Outliers:** High sales values over **5000 units** may indicate **outliers** or exceptional demand events, deserving further exploration to uncover driving factors.
- **Strategic Focus:**
    - Prioritize **enhancing performance** in the 1000–3000 range, as it represents the bulk of sales.
    - Investigate **high-performing cases** (>5000) to extract insights for scaling strategies.

2.6.4 Visualization of Monthly Boxplot of Rose Sales and Monthly Boxplot of Sparkling Sales

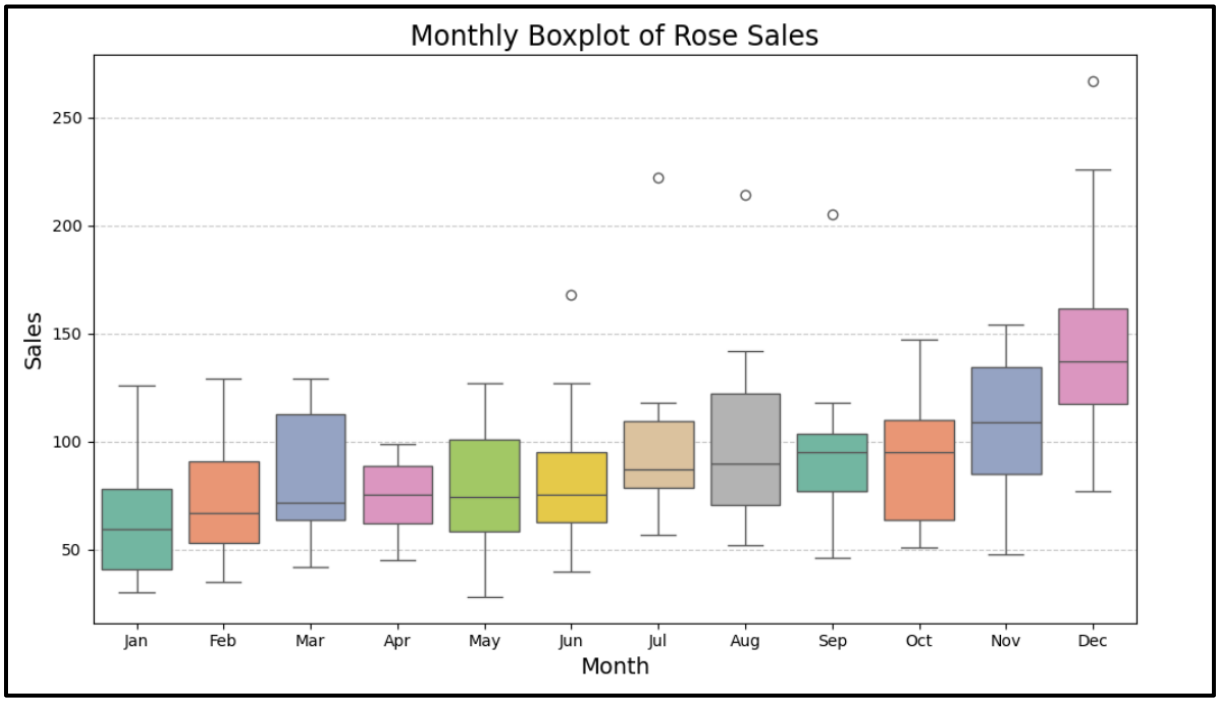


Figure 11 Visualization of Monthly Boxplot of Rose sales

Observations Based on the Boxplot of Monthly Rose Sales

- **Seasonality**:
  - December shows the highest median sales and the largest spread, reflecting strong seasonal demand, likely due to holidays.
  - Sales are generally higher during March, November, and December compared to other months.
- **Outliers**:
  - Outliers are prominent in June, July, August, and December, with some sales significantly exceeding typical values. These could be attributed to special events, promotions, or other irregular factors.
- **Sales Stability**:
  - Months like January, February, May, and September exhibit smaller ranges (less variability), indicating more stable and predictable sales.
- **Median Sales**:
  - The median sales gradually increase as the year progresses, peaking in December.

- Early months, such as January and February, have notably lower median sales compared to later months.

Interquartile Range (IQR):

- March, August, and December have the largest IQR, indicating greater variability in sales during these months.
- January, February, and May have smaller IQRs, reflecting more consistent and predictable sales patterns.

Insights:

- **Seasonal Peaks**:
  - December experiences the highest sales, driven by holiday-related demand and year-end activities.
- **Stable Sales Periods**:
  - Sales are notably stable during January, February, May, and September, with minimal variability and absence of outliers.
- **High Variability Months**:
  - March, August, and December exhibit significant variability in sales, suggesting the impact of irregular events, promotions, or seasonal fluctuations.
- **Opportunities from Outliers**:
  - The outliers in June, July, August, and December indicate isolated instances of unusually high sales. Analyzing these outliers could uncover opportunities to replicate these successes.
- **Rising Trend Toward Year-End**:
  - The steady increase in median sales toward the year's end highlights a natural demand surge during the last quarter, which businesses can leverage for strategic planning and inventory management.

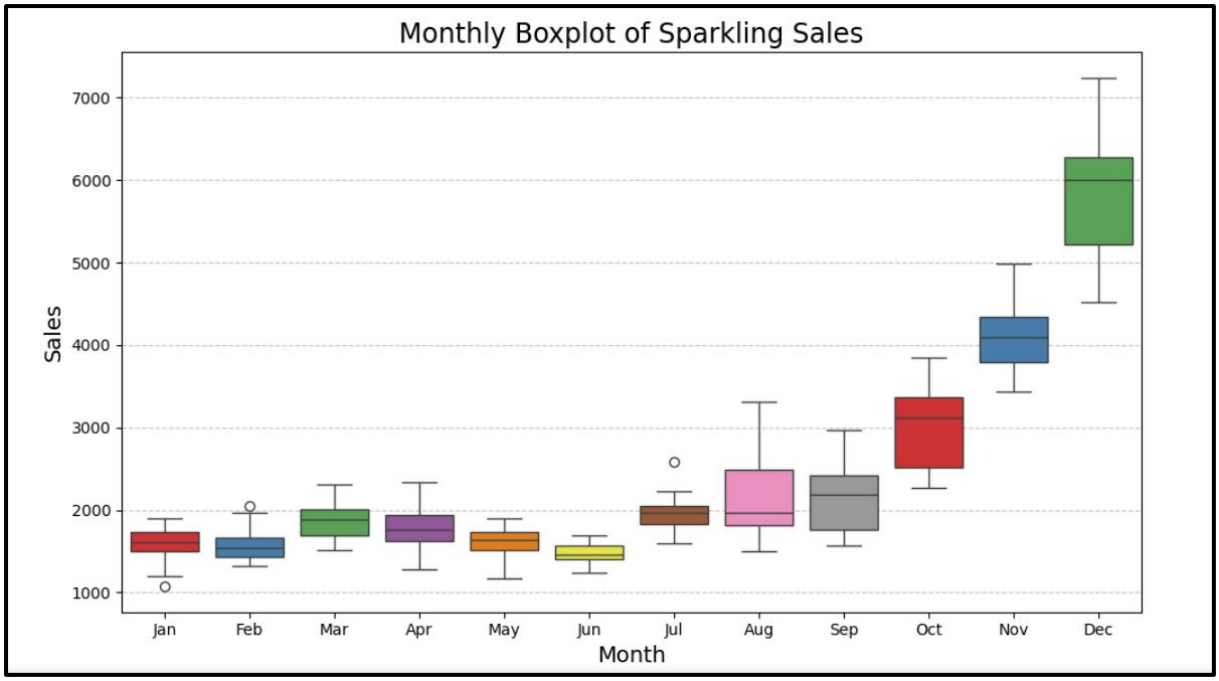


Figure 12 Visualization of Monthly Boxplot of Sparkling sales

Observations from the Monthly Boxplot of Sparkling Sales:

- **Strong Seasonal Trend**:
  - December stands out with the highest sales and a significantly wide range, highlighting a strong seasonal demand spike.
  - Sales show a steady increase from September to December, reflecting a year-end surge.
- **Consistent Low Sales**:
  - From January to July, sales are lower and more stable, with median values clustering between 1,000 and 2,000.
- **Gradual Growth**:
  - A gradual rise in sales is evident from August onward, likely as part of a ramp-up toward the holiday season.
- **Outliers**:
  - Outliers are observed in February and July, indicating occasional sales spikes in otherwise stable months.
- **Largest Variability**:
  - December exhibits the highest variability, with sales ranging from just over 2,000 to nearly 7,000.

Insights:

- **Year-End Optimization**:
  - Focus marketing and inventory strategies on the last quarter, particularly December, to capitalize on the seasonal demand.
- **Early-Year Stability**:
  - Utilize the stable sales period from January to July to develop and test consistent strategies, as variability during these months is low.
- **Growth Preparation**:
  - Plan for a gradual sales increase starting in August to ensure readiness for the holiday season demand.
- **Outlier Opportunities**:
  - Investigate the sales spikes in February and July to uncover potential opportunities for boosting sales during traditionally off-peak months.

2.6.5 Correlation Analysis between Rose and Sparkling Sales

Correlation between Rose and Sparkling sales: 0.40457904770543324

Observation:

Moderate Positive Correlation:

A correlation of 0.404 indicates a moderate positive relationship between Rose and Sparkling sales. This suggests that when Rose sales increase, Sparkling sales tend to increase as well, but the relationship is not very strong.

3.Perform Decomposition

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

3.1 Data Pre-processing

3.1.1 Missing value treatment

- Fill missing values in Rose dataset using forward filling.
- Check for missing values in Rose dataset after forward filling.

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

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

Figure 13 Treating Value missing value in Rose wine data

Observation:

No Missing Values:

- Both the "Rose" and "Sparkling" datasets now have zero missing values, as confirmed by the isnull().sum() output.
- This indicates that any prior missing values in these datasets were successfully handled, ensuring completeness.

Now will do decomposition for Rose wine sales and Sparkling wine sales.

3.2 Decompose the time series to understand trends, seasonality, and residuals.

3.2.1 Rose wine sales.

Additive Decomposition - Rose

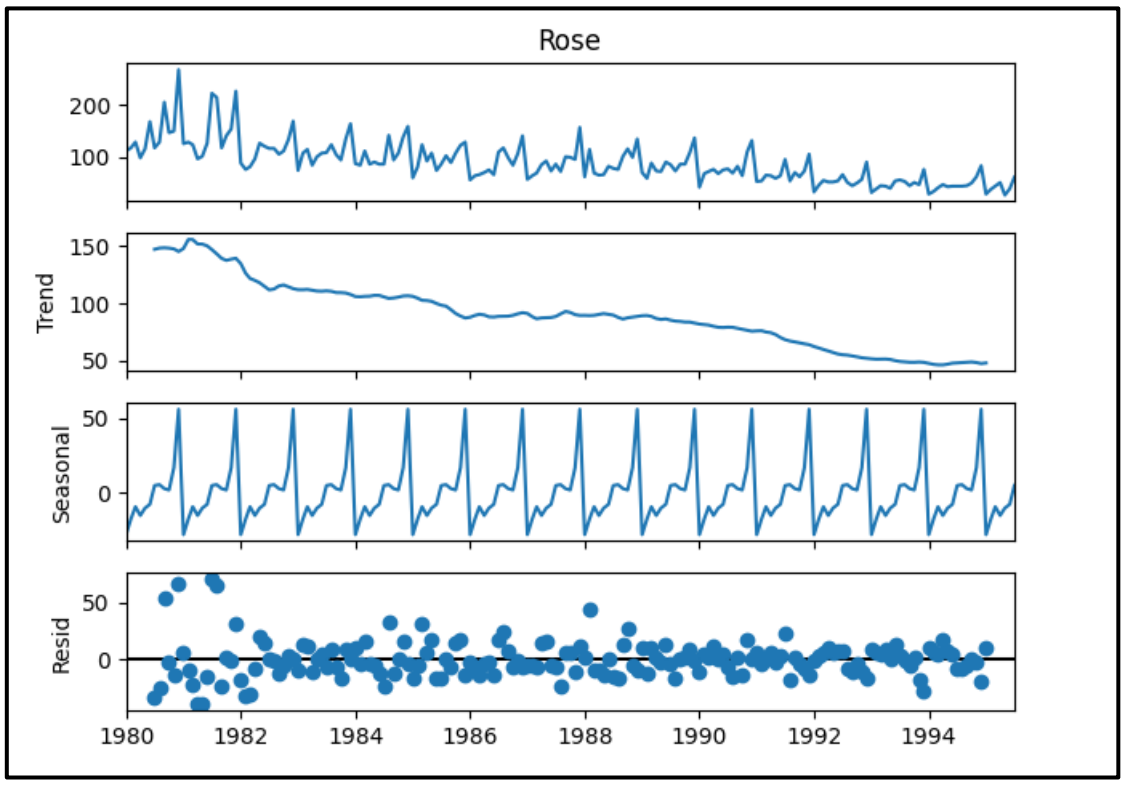


Figure 14 Additive Decomposition - Rose

We see that the residuals are located around 0 from the plot of the residuals in the decomposition.

Observation:

- **Residuals Centered Around Zero**:
    - The residuals from the decomposition plot are distributed around 0, indicating that the model effectively captures the primary trend and seasonal components of the time series.

Understanding trends, Seasonality and residuals

```
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 15 Understanding trends, Seasonality and residuals – Rose Additive

Observations Based on Decomposed Components:

Trend Component

- **Initial NaNs**:
    - The trend values for the initial months (January to June 1980) are NaN, which is typical in decomposition due to the need for a centered moving average to calculate the trend.
- **General Pattern**:
    - Starting in July 1980, the trend begins around 147 and shows minor fluctuations, peaking in September 1980 (148.375) before declining slightly by December 1980 (145.125).
    - This suggests a slow downward trend toward the end of the year.

Seasonality Component

- **Recurring Pattern**:
    - Strong seasonal effects are evident, with values ranging from a low of -27.903 (January) to a high of 55.719 (December).
    - December exhibits the most significant positive seasonality, indicating high sales during the holiday season, while January shows the largest negative seasonality.

Residual Component

- **Random Variations**:
    - Residuals for July to December 1980 are scattered around zero, ranging from -33.963 (July) to 66.156 (December).
    - High positive residuals in September (53.845) and December (66.156) suggest unaccounted spikes, while negative residuals in July and November indicate lower-than-expected values.

Insights:

- **Trend Insights**:
    - The gradual decline in the trend toward the year's end indicates potential saturation or a natural stabilization of long-term growth.
- **Seasonality Insights**:
    - Seasonal effects align with expectations, showing a significant demand increase during the holiday season (December) and lower demand in early months (January to March).

- **Residual Insights**:
    - High residuals in September and December may indicate special events or factors driving additional sales beyond the expected trend and seasonality.
    - Negative residuals in July and November suggest unexpected dips that may require further analysis to identify contributing factors.

Multiplicative Decomposition - Rose

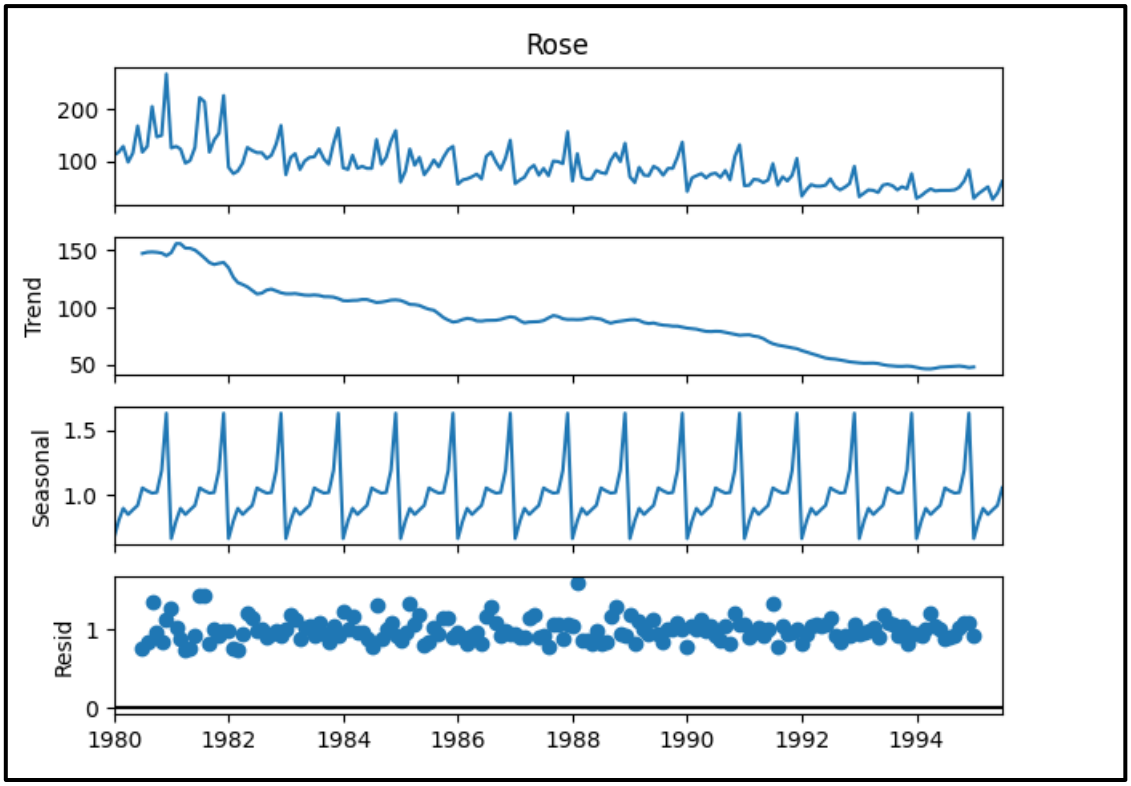


Figure 16 Multiplicative Decomposition - Rose

For the multiplicative series, we see that a lot of residuals are located around 1. Thus Multiplicative Decomposition is the right way to decompose the time series.

Also, it is evident that there is a 6 months seasonality in the data from the above plots.

Observation

- **Residuals Centered Around 1:**
    - For the multiplicative decomposition, residuals being centered around 1 indicates that the model effectively explains the observed data, with most variations accounted for by the trend and seasonal components.
- **Suitability of Multiplicative Decomposition:**
    - The multiplicative model is appropriate when the data exhibits proportional relationships between components, i.e., seasonal fluctuations increase or decrease with the level of the trend.

- The residuals' distribution around 1 supports this choice, as it aligns with the multiplicative assumption where the observed data is a product of its components.

Understanding trends, Seasonality and residuals

```
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 17 Understanding trends, Seasonality and residuals – Rose Additive

Observations Based on Decomposed Components:

Trend Component

- **NaNs in the Beginning:**
    - As expected, the trend values for the first six months (1980-01 to 1980-06) are NaN due to the nature of moving averages in trend decomposition.
- **General Pattern:**
    - The trend fluctuates between 145 and 148 from July to December 1980, suggesting a relatively stable but slightly declining trend towards the end of the year.

Seasonality Component

- **Fluctuating Seasonal Values:**
    - The seasonal component varies from 0.670 in January to 1.629 in December.
    - December shows the highest seasonal multiplier (1.629), indicating a significant seasonal peak during the holiday period.
    - The seasonal pattern reflects the cyclicality of sales, peaking around December and showing some variability across the months.

Residual Component

- **Residuals Distributed Around 1:**
    - Residuals range from 0.758 to 1.357, with the majority of values hovering around 1.
    - This suggests that the model has captured most of the trends and seasonal effects, and the residuals represent random noise or minor fluctuations.
    - The presence of positive residuals implies occasional sales spikes or variations not fully explained by the trend and seasonality.

Insights:

- **Stable Trend:**
    - The trend component suggests a gradual decline or stabilization in sales, especially from September to December, signaling that the growth rate may be slowing or stabilizing towards the end of the year.
- **Seasonal Peaks in December:**
    - The strong seasonality in December (1.629) is consistent with high sales due to holiday-related demand. The smaller seasonal multipliers in the earlier months indicate lower demand or fewer promotional activities.

- **Residual Behavior:**
    - The residuals being mostly around 1 with some fluctuation indicate that the multiplicative decomposition model is well-suited, capturing the primary cyclical behavior of the series. The few instances of higher residuals (e.g., in September and December) suggest occasional spikes that may be due to special events, promotions, or external factors.

3.2.2 Sparkling wine sales.

Additive Decomposition - Sparkling

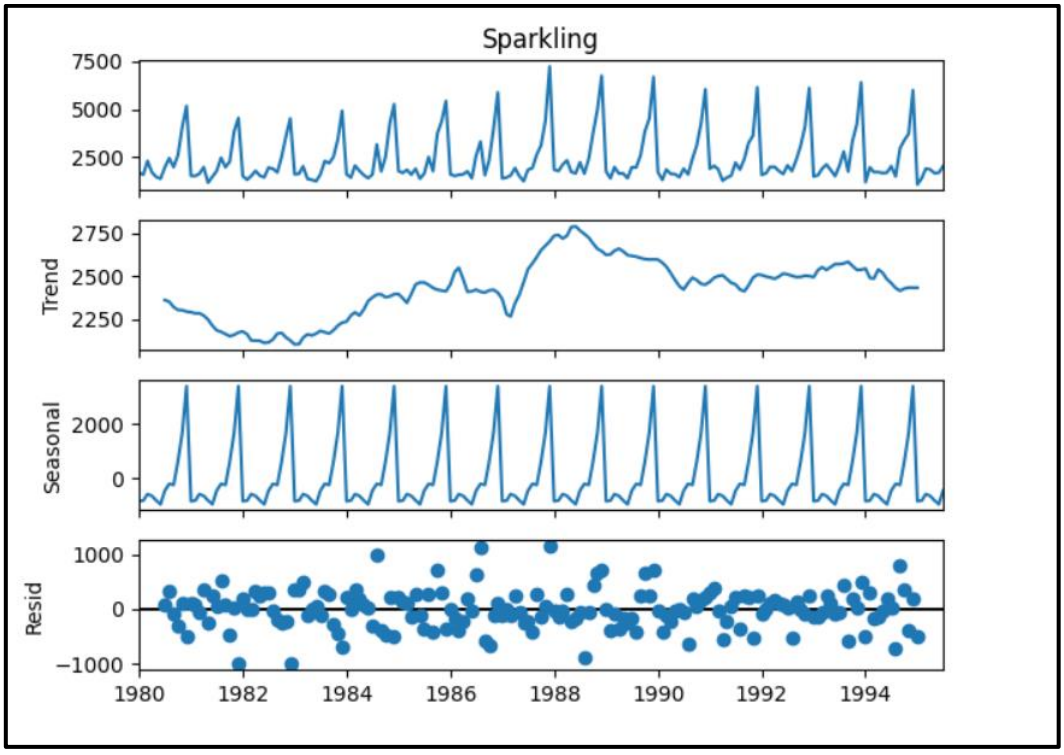


Figure 18 Additive Decomposition - Sparkling

We see that the residuals are located around 0 from the plot of the residuals in the decomposition.

Observation:

- **Residuals Centered Around 0**:
    - If the residuals are centered around 0, this suggests that the decomposition model (multiplicative or additive) has effectively captured the trend and seasonality components of the time series. In an ideal decomposition, the residuals should reflect random noise or small fluctuations that are not explained by the trend or seasonality.

Understanding trends, Seasonality and residuals

```
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 19 Understanding trends, Seasonality and residuals- Sparkling Additive

Observations Based on Decomposed Components:

Trend Component

- **NaNs in the Beginning:**
    - As expected, the first six months (1980-01 to 1980-06) have NaN values for the trend due to the nature of trend decomposition.

- **Steady Decline:**
    - The trend component shows a steady decline from around 2360 in July to 2293 in December, suggesting that sales are gradually decreasing during this period.

Seasonality Component

- Large Negative and Positive Seasonal Values:
    - The seasonal component fluctuates significantly, with negative values early in the year (e.g., January to June) and positive values starting from October to December.
    - December shows the highest seasonal effect (3386.98), indicating a strong seasonal spike at the end of the year, likely due to holiday-related demand.
- Mid-Year Dip:
    - The seasonal effect is negative from January to June, indicating weaker sales or external factors leading to lower-than-expected values during the first half of the year.

Residual Component

- Residuals Represent Fluctuations Around 0:
    - The residuals show values ranging from around 70.84 in July to -501.78 in December. These deviations are not large, but they suggest some unexplained variation beyond the trend and seasonal components.
- December Residual Spike:
    - A significant negative residual in December (-501.77) may indicate a mismatch between the expected and actual sales during the peak season, potentially due to factors not captured by the model (e.g., promotions, external events, or market shifts).

Insights:

- **Declining Trend:**
    - The steady decline in the trend component suggests that the sales may be on a downward trajectory. It's important to investigate if this decline is due to seasonality, market factors, or underlying changes in demand over the year.
- **Strong Seasonal Peak in December:**
    - The seasonal spike in December confirms that holiday-related demand drives up sales. It's critical to plan for this peak period in terms of inventory, marketing, and resource allocation to maximize sales.

- **Residuals Analysis:**
    - The residuals indicate some unexplained fluctuations, particularly in December. This could point to anomalies or unforeseen events affecting sales during this peak period, warranting further investigation to understand the causes of these discrepancies.

Multiplicative Decomposition – Sparkling

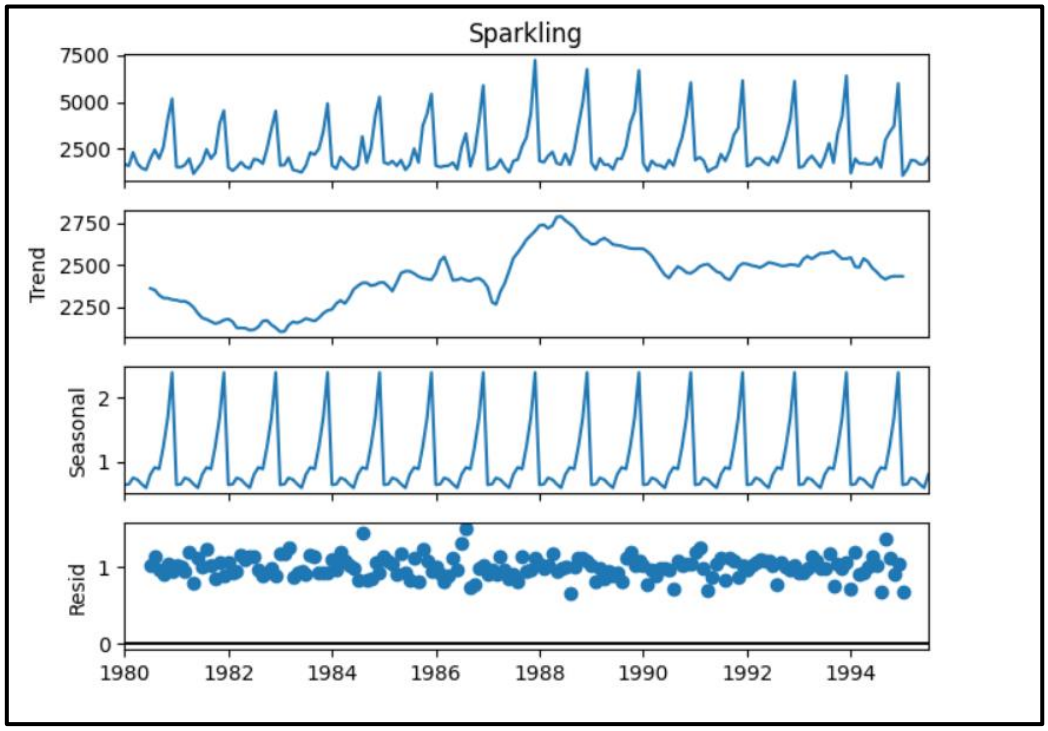


Figure 20 Multiplicative Decomposition - Saprkling

For the multiplicative series, we see that a lot of residuals are located around 1. Thus Multiplicative Decomposition is the right way to decompose the time series .

Also it is evident that there is a 6 month seasonality in the data from the above plots .

Observation:

- **Multiplicative Decomposition:** Since many residuals are clustered around 1, this suggests that the multiplicative model is a good fit. In a multiplicative decomposition, the seasonal and trend components interact in a way that can better capture proportional variations in the data (i.e., percentage-based fluctuations). Residuals close to 1 indicate that the model has effectively accounted for the variability in the data, leaving little unexplained variation.
- **6-Month Seasonality:** The seasonal component shows a recurring pattern every 6 months, with significant fluctuations observed. This strong 6-month seasonality

implies that the data exhibits biannual trends or periodic effects, likely tied to external factors such as market cycles, consumer behavior, or seasonal demand.

Understanding trends, seasonality, and residuals

```
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 21 Understanding trends, Seasonality and residuals- Sparkling Mulitiplicative

Observations:

Trend:

- The trend shows a gradual decline from 2360.67 in July to 2293.79 in December. This suggests a downward trend over the second half of the year, with a steady decrease in the value.

Seasonality:

The seasonal component shows fluctuations with values that vary significantly across the months:

- Early months (January to June) have a mild seasonal variation, ranging from 0.6 to 0.75.
- The second half of the year (July to December) exhibits a stronger seasonal variation, with values peaking in December at 2.38. This suggests that the seasonal effect is stronger towards the end of the year, possibly due to factors like holidays or year-end demand surges.

Residuals:

- The residuals are relatively stable around 1, with values ranging from 0.91 to 1.14. This suggests that the multiplicative decomposition model fits the data well, with minimal unexplained variation. Any fluctuations in the residuals are small, indicating that the model captures the underlying trend and seasonality effectively.

Insights

- **Stable Trend with Mild Decline:**
    - The trend is showing a mild downward slope, which may indicate a slight reduction in sales or a natural decline in demand over the period. Monitoring this trend could help in forecasting future periods and adjusting strategies if the decline continues.
- **Stronger Seasonal Effects in the Second Half of the Year:**
    - The seasonal component indicates a stronger demand surge towards the end of the year, particularly in December. This is common in many industries, driven by holiday-related demand. Strategies like promotional campaigns and inventory ramp-ups can be timed to align with these seasonal peaks.

- **Effective Model Fit:**
    - The residuals' consistency around 1 suggests that the multiplicative decomposition model provides a good fit to the data, with minimal noise or anomalies. This suggests that further modeling or adjustments are unnecessary at the moment.

3.4 Visualize the Processed Data

3.4.1 Rose wine sales over time

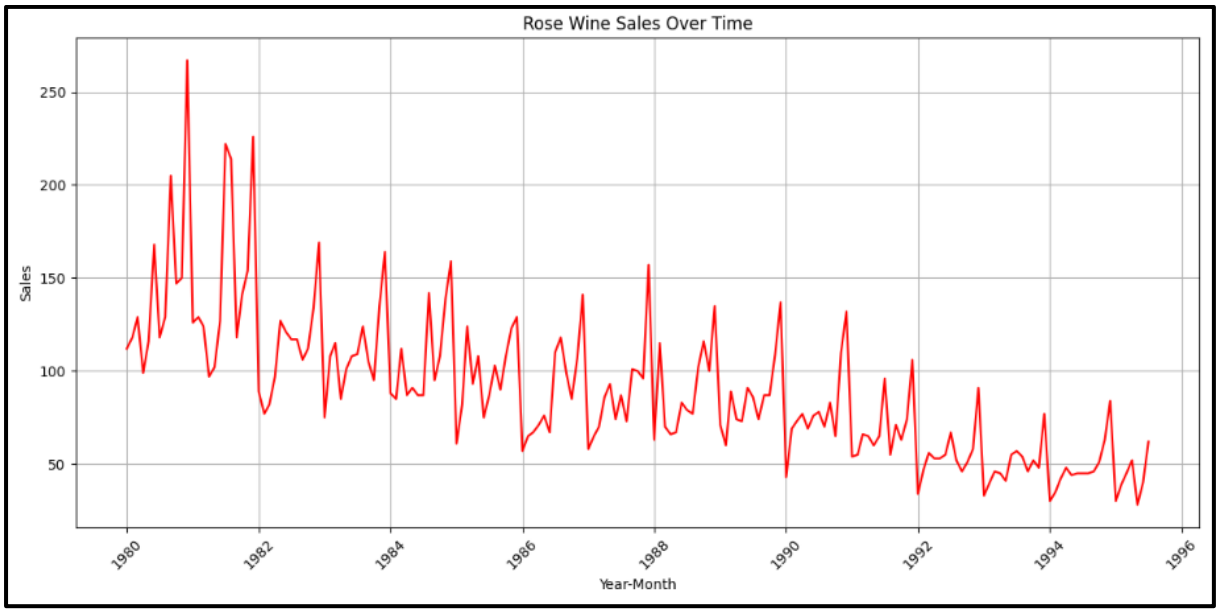


Figure 22 Visualize the Rose wine sales over time

Overall Trend:

- There is a clear downward trend in rose wine sales from 1980 to 1996
- The highest peaks occur in the early 1980s, with sales reaching around 250 units
- By the mid-1990s, sales had declined to around 50 units

Volatility:

- The data shows high volatility, especially in the early years (1980-1985)
- The amplitude of fluctuations decreases over time
- More stable but lower sales levels are seen in the later years

Seasonal Patterns:

- There appear to be regular seasonal fluctuations throughout the time series
- Peaks and troughs occur multiple times within each year
- The seasonal variation becomes less pronounced in later years

Notable Features:

- Several major spikes in sales occurred between 1980-1983
- A gradual smoothing of the curve occurs as time progresses
- The baseline sales level drops from ~150 units to ~50 units over the 16-year period

Business Implications:

- This declining trend might indicate:
    - Changing consumer preferences
    - Increased competition
    - Market saturation
    - Possible shift in marketing strategy needed
- The reduced volatility might suggest more stable but smaller market share

3.4.2 Sparkling wine sales over time

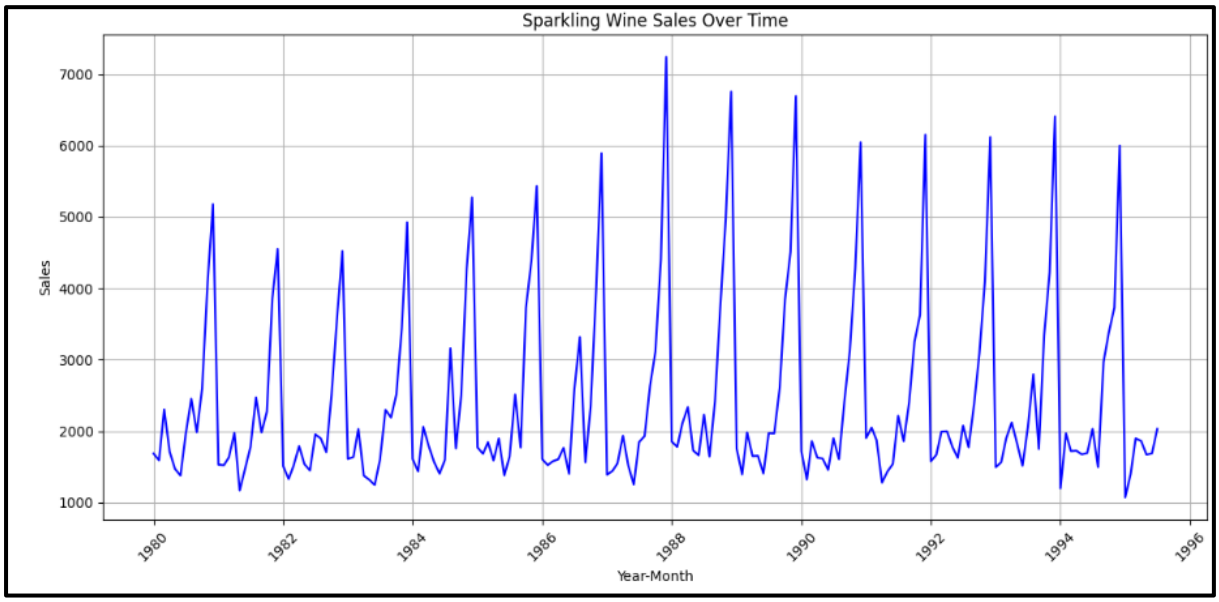


Figure 23 Visualize the Sparkling wine sales over time

- **Seasonality**:
    - There is a distinct seasonal pattern in the sales data. Peaks occur consistently during specific periods, suggesting strong sales during festive occasions, particularly around New Year's and holiday celebrations.

- **Sales Peaks and Troughs**:
    - Each year shows multiple peaks with sales typically rising significantly at certain points, indicating heightened consumer interest during these times. Conversely, troughs suggest lower sales months, likely during off-peak periods.
- **Sales Volume Fluctuations**:
    - Sales volumes fluctuate between approximately 1000 and 7000 units. The significant variation indicates that while there is a core level of sales, there are certain times of the year when sparkling wine is particularly popular.
- **Stability Over Time**:
    - Over the entire analyzed period (1980-1996), the overall sales trend seems stable, with no dramatic increases or decreases. This stability might suggest a consistent demand for sparkling wine, likely influenced by cultural factors and marketing efforts.
- **Potential Influences**:
    - Factors such as economic conditions, advertising campaigns, and changes in consumer preferences may impact sales but are not specifically detailed in the graph. External events (like social shifts or wine import/export regulations) could also play a role in certain years.
- **Year-Round Interest**:
    - Despite seasonal spikes, there is a baseline level of interest throughout the year. This suggests that while some consumers predominantly purchase during celebrations, there remains a steady demand for sparkling wine during other times as well.

3.5 Train-test split

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

Figure 24 Train-test split

Dataset Overview:

- **Rose Train and Test**: The Rose dataset is split into training (149 rows, 2 columns) and testing (38 rows, 2 columns).
- **Sparkling Train and Test**: The Sparkling dataset is similarly split into training (149 rows, 2 columns) and testing (38 rows, 2 columns).
- **2 Columns**: Indicates that each dataset has two variables (likely features and/or labels).

Potential Insights:

- **Train/Test Split**: The proportions seem typical, with ~80% data in training and ~20% in testing. This is a good practice for ensuring robust model evaluation.
- **Comparison**: If the columns represent similar features (e.g., characteristics of Rose and Sparkling wines), visualizations could compare the trends or distributions across the datasets.

3.5.1 Rose training and testing sets

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

Figure 25 First few and Last few rows of training data

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 First few and Last few rows of Test data

3.5.2 Sparkling training and testing sets

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

Figure 27 First few and Last few Sparkling of training data

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 28 First few and Last few Sparkling of testing data

Understanding the Data:

- **Rose and Sparkling Data**:
  - Each dataset has two columns:
    - ✓ YearMonth: A time-series index (monthly data from 1980 to 1995).
    - ✓ Rose or Sparkling: Represents the sales or quantity for each category.
  - Data is split into training (1980–1992) and testing (1992–1995).

3.5.3 Plotting Rose data train and test split

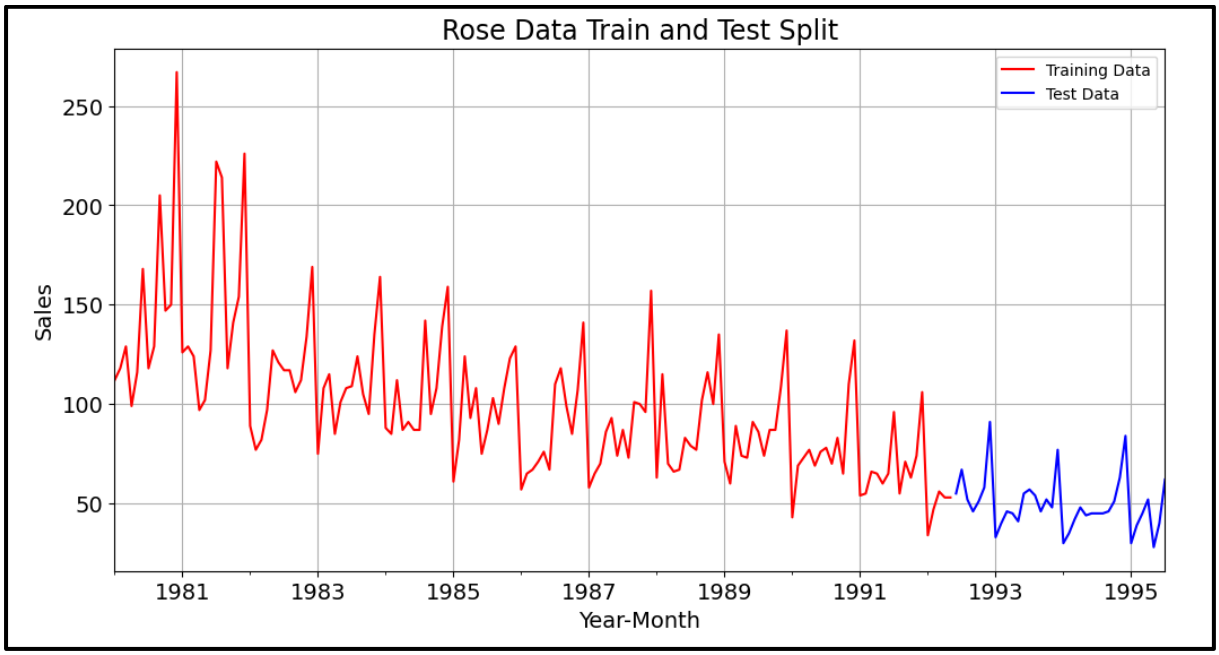


Figure 29 Plotting Rose data train and test split

Observations

- A downward trend in sales from 1980 to 1995 suggests a decrease in consumer interest or market demand for rosé wine.
- Sales demonstrate significant fluctuations, with peaks exceeding 250 units and troughs as low as 50 units, indicating variable consumer purchasing behavior.
- Sales in the test data show a more stable trend compared to the training data, particularly in later years, hinting at potential predictability in future trends.
- The overall decline and fluctuations in sales could reflect changes in consumer preferences, economic conditions, or increased competition in the wine market.
- These patterns highlight the need for targeted marketing strategies, optimized inventory management, and consumer segmentation to adapt to market dynamics effectively.
- Stabilization in sales during the later years suggests opportunities for recovery or growth with the implementation of suitable strategies.

3.5.4 Plotting Sparkling data train and test split

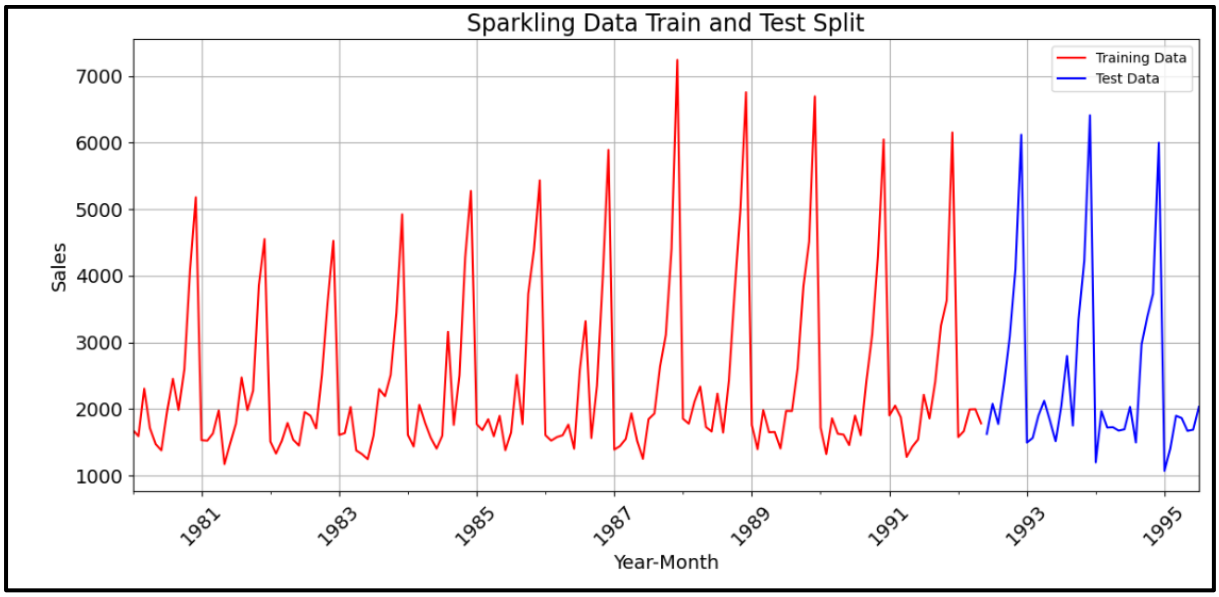


Figure 30 Plotting Soarkling Data Train and Test Data

- **Sales Pattern**:

  Sales go up and down in a regular pattern each year, suggesting people buy more sparkling wine during certain months, likely around holidays or celebrations.

- **Peaks and Valleys**:

  Sales often spike, reaching over 6000 units, and then drop down to around 2000 units. This fluctuation shows that buyer interest can vary a lot from month to month.

- **Overall Trend**:

  Over the years, the sales appear to stay relatively stable, suggesting no big increase or decrease in popularity from 1980 to 1995.

- **Test Data Stability**:

  The test data (blue line) shows a similar pattern to the training data (red line) but seems more consistent in the later years, indicating that trends might continue.

- **Market Insights**:

  Understanding these sales trends can help in planning marketing efforts and stocking products at the right times of the year.

3. 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)

4.1 Linear regression

4.1.1 Rose

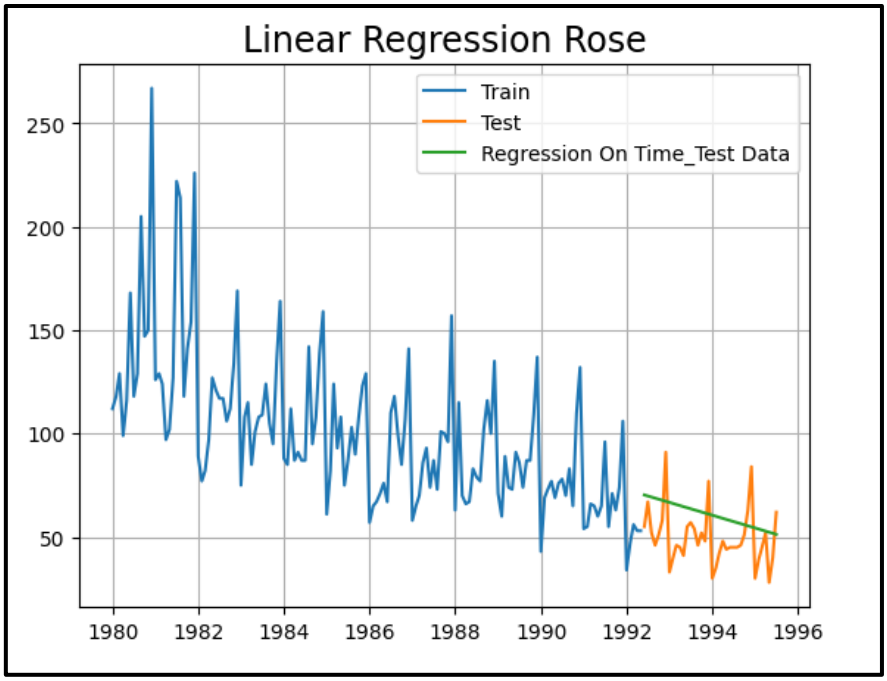


Figure 31 Linear Regression Rose

- **Sales Data**:

  The blue line represents the training data, which shows that rosé wine sales have fluctuated significantly from 1980 to 1995, with noticeable peaks and troughs.

- **Declining Trend**:

  Both the training data and the orange test data indicate an overall decreasing trend in sales over the years, suggesting that interest in rosé wine may be declining.

- **Test Data Stability**:

  The orange line (test data) has less variability compared to the training data, indicating that sales are becoming more stable in the later years, but still reflect the overall downward trend.

- **Regression Line**:

  The green regression line applied to the test data demonstrates a slight downward slope, affirming the expectation of continued decline in sales if current trends persist.

	Test RMSE Rose
RegressionOnTime	17.510241

Figure 32 Test RMSE For Rose

Observation:

Model Performance:

- The RMSE value of 17.51 indicates the average deviation between the predicted and actual sales values is approximately 17.51 units in the test dataset. This reflects the model's prediction accuracy.

Acceptability:

- To assess whether this RMSE is acceptable, compare it to:
  - The range of Rose sales (e.g., from 50 to 250 units).
  - Alternative models or benchmarks.
- If 17.51 is relatively low compared to the data range, the model performs well; otherwise, there may be room for improvement.

Improvement Opportunities:

- Consider improving the model by:
  - Adding more features (e.g., seasonality, economic indicators).
  - Using advanced time-series models like ARIMA, SARIMA, or LSTMs.
  - Analyzing residuals to identify systematic errors or outliers.

4.1.2 Sparkling

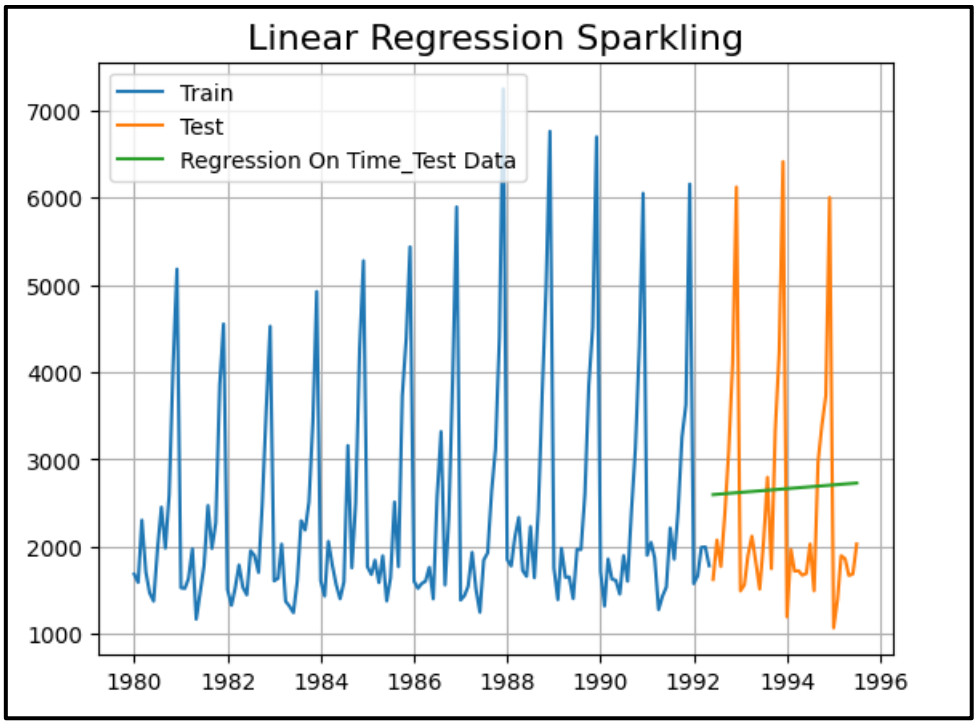


Figure 33 Linear Regression Sparkling

Data Segmentation:

- **Blue Line (Train)**: This represents the training data, showing sparkling wine sales from 1980 to 1995. It displays significant fluctuations but generally remains within the range of 1000 to 7000 units.
- **Orange Line (Test)**: This line indicates the test data, which continues beyond the training data. It shows a similar level of fluctuation but often appears to have higher peaks in later years.
- **Green Line (Regression)**: This line represents the linear regression analysis applied to the test data. It provides a simplified view of the trend over time.

Sales Trends:

- The **training data exhibits consistent seasonal patterns** with notable peaks, especially during certain months, likely coinciding with holidays or celebrations, implying these times drive higher sales.

Fluctuations:

- Both training and test sales data show considerable variability, with peaks suggesting robust interest during seasonal spikes but also representing moments of less engagement.

Test Data Stability:

- The test data has more pronounced peaks in later years, which could indicate a renewed interest or increased consumption of sparkling wine compared to earlier periods.

Regression Line Analysis:

- The green regression line slopes slightly upward, suggesting that while there are fluctuations, the overall trend in sales may be stabilizing or improving in the later years of the test data. This is a positive sign, indicating potential growth or recovery in sales.

Key Implications:

- The trends indicate that timely marketing strategies during peak seasons can be effective in driving sales. Additionally, understanding the reasons behind fluctuations can inform better inventory management.
- The upward trend in the regression line suggests the possibility of future growth in sparkling wine sales if marketing efforts are aligned effectively.

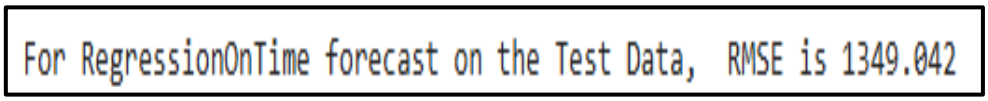


Figure 34 Regression on time for Rose

Observations:

High Prediction Error:

- An RMSE of **1349.042** is very high, especially when Sparkling sales range from approximately 1500 to 3000 units.
- This indicates that the model's predictions deviate significantly from the actual values, implying poor forecasting accuracy.

Linear Regression for Rose and Sparkling wine dataset

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	17.510241	1349.042457

Figure 35 Linear Regression for Rose and Sparkling wine dataset

Observations:

- **Model Accuracy**:

  The Regression on Time model performs significantly better on the **Rose** dataset (RMSE: 17.51) than on the **Sparkling** dataset (RMSE: 1349.04).

- **Error Comparison**:

  The RMSE for Sparkling sales is substantially higher, indicating much larger prediction errors relative to the Rose data.

- **Data Complexity**:

  The high RMSE for Sparkling suggests greater variability, complexity, or unmodeled factors in the data compared to the more stable Rose dataset.

- **Model Suitability**:

  The model may not adequately capture Sparkling sales patterns (e.g., seasonality, trends) and may need enhancement or replacement with a more sophisticated approach.

- **Impact**:

  Predictions for Rose are more reliable, while Sparkling forecasts may lead to significant inaccuracies if used for decision-making.

4.2 Simple Average

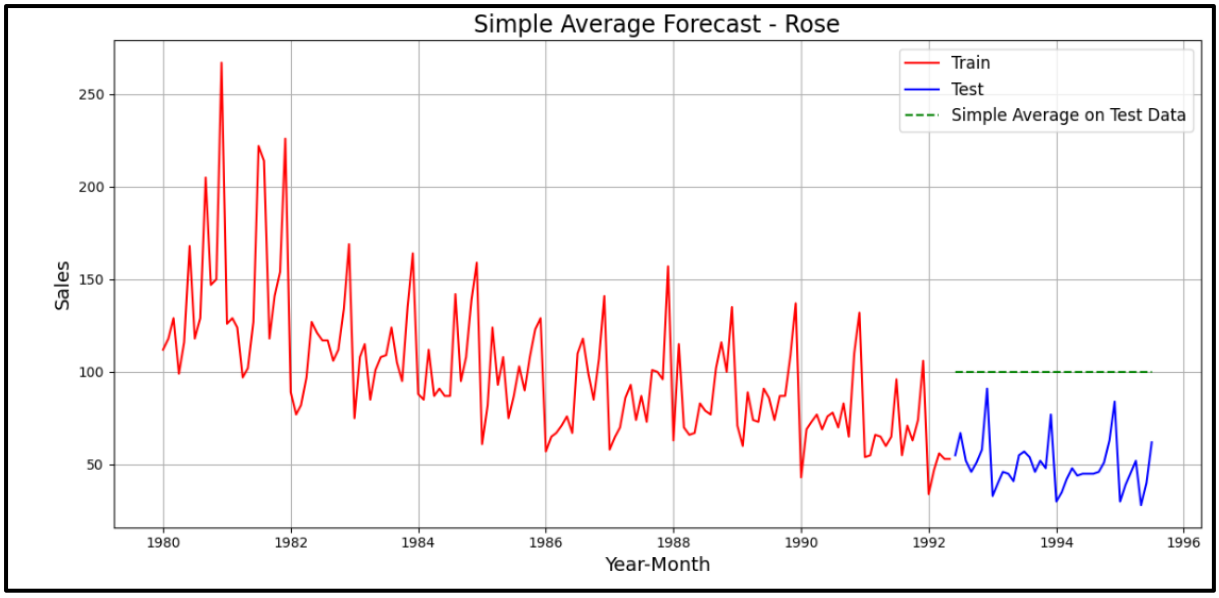


Figure 36 Simple Average – Rose

- **Overall Decline**:

  The training data shows a general peak at the beginning of the period (around 250 units) but declines over time, indicating waning interest or demand for rosé wine.
- **Sales Variability**:

  The red line exhibits monthly fluctuations, but the peaks are lower than those seen in the sparkling wine graph, suggesting less overall popularity.
- **Stable Test Data**:

  The test data (blue line) remains relatively flat and low, peaking around 100 units, clearly reflecting a stable but declining interest in rosé sales.
- **Average Sales Trend**:

  The green dashed line for the simple average indicates that sales remain consistently low over time, highlighting ongoing challenges in boosting demand.
- **Lack of Seasonal Trends**:

  Unlike sparkling wine, there are fewer distinct seasonal spikes, implying rosé wine doesn't experience the same celebratory sales periods.

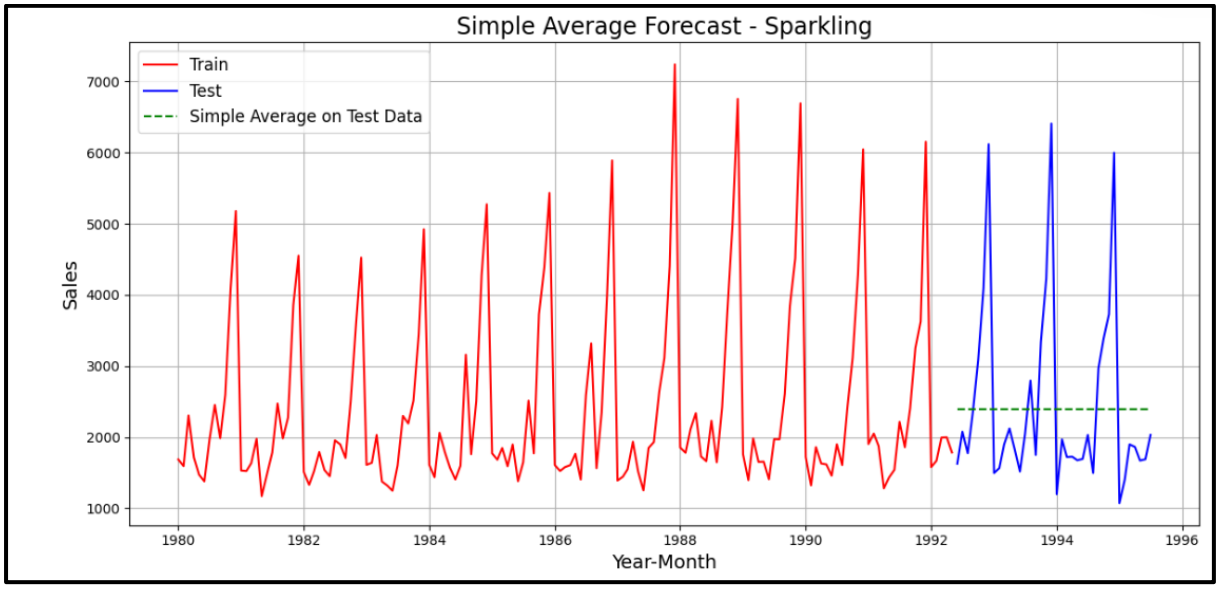


Figure 37 Simple Average - Sparkling

Sparkling Wine Sales Graph

- **Sales Fluctuations**:

  The training data (red line) shows significant fluctuations in sales, with peaks often reaching around 6000-7000 units, particularly during specific months, suggesting strong seasonal demand.

- **Seasonal Peaks**:

  There are clear seasonal spikes in sales, likely correlating with holidays and celebratory events throughout the year.
- **Test Data Consistency**:

  The test data (blue line) shows a mix of fluctuations similar to the training data but appears slightly more stable, suggesting ongoing consumer interest.
- **Average Sales Trend**:

  The green dashed line indicates the simple average on test data, which suggests that the average sales level is around 2000 units, providing a baseline that reflects overall performance.
- **Trend Observation**:

  While there are spikes, the overall trend indicates stable sales, with no drastic changes from year to year in the test data.

Simple Average for Rose and Sparkling wine dataset

	Test RMSE Rose	Test RMSE Sparkling
SimpleAverageModel	52.239499	1331.037637

Figure 38 Simple Average for Rose and Sparkling wine dataset

Observations:

- **SimpleAverageModel Performance**:
  - For **Rose**, the Test RMSE is **52.24**, significantly higher than the **RegressionOnTime** model (17.51), indicating poorer performance.
  - For **Sparkling**, the Test RMSE is **1331.04**, slightly better than the **RegressionOnTime** model (1349.04), but still indicative of large prediction errors.
- **Model Suitability**:
  - The **SimpleAverageModel** performs poorly for **Rose**, likely due to its inability to capture the downward trend or fluctuations effectively.
  - For **Sparkling**, it performs marginally better, suggesting that Sparkling sales might have patterns that are partially captured by a simple average.
- **Error Magnitude**:
  - Both models yield high RMSEs for **Sparkling**, indicating a need for more sophisticated methods to address the variability or trends in this dataset.

Insights for Future Models:

A trend-aware or seasonality-based model would likely outperform both **RegressionOnTime** and **SimpleAverageModel**, particularly for **Sparkling**.

4.3 Moving Average

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 39 Moving Average Rose

- **Sales Data and Moving Averages**:
    - The table provides raw sales figures for rosé wine alongside calculated trailing averages (2, 4, and 9 periods).
    - Early months in the table show initial sales data but no trailing averages available until enough data points are collected (e.g., trailing averages for 4, 6, and 9 require more data points).

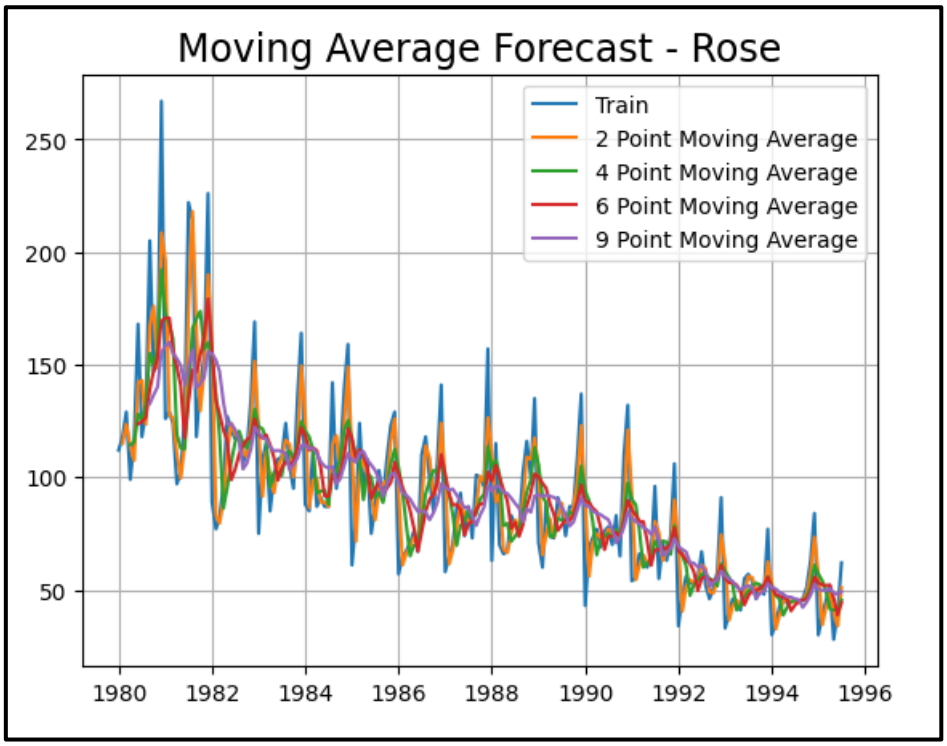


Figure 40 Moving Average Forecast - Rose

- **Comparison of Moving Averages**:
    - This graph again shows training data (blue line) and various moving averages.
    - The moving averages (2, 4, 6, and 9) closely follow the training data but mean that all moving averages show a decreasing trend over time.
- **Visual Stability**:
    - The moving averages tend to converge, particularly around the later years, suggesting that sales may stabilize around lower average values.
- **Peak Visibility**:
    - The peaks in sales are more pronounced in the training data compared to the moving averages, which smooth out these fluctuations, highlighting the challenges in consistent sales.

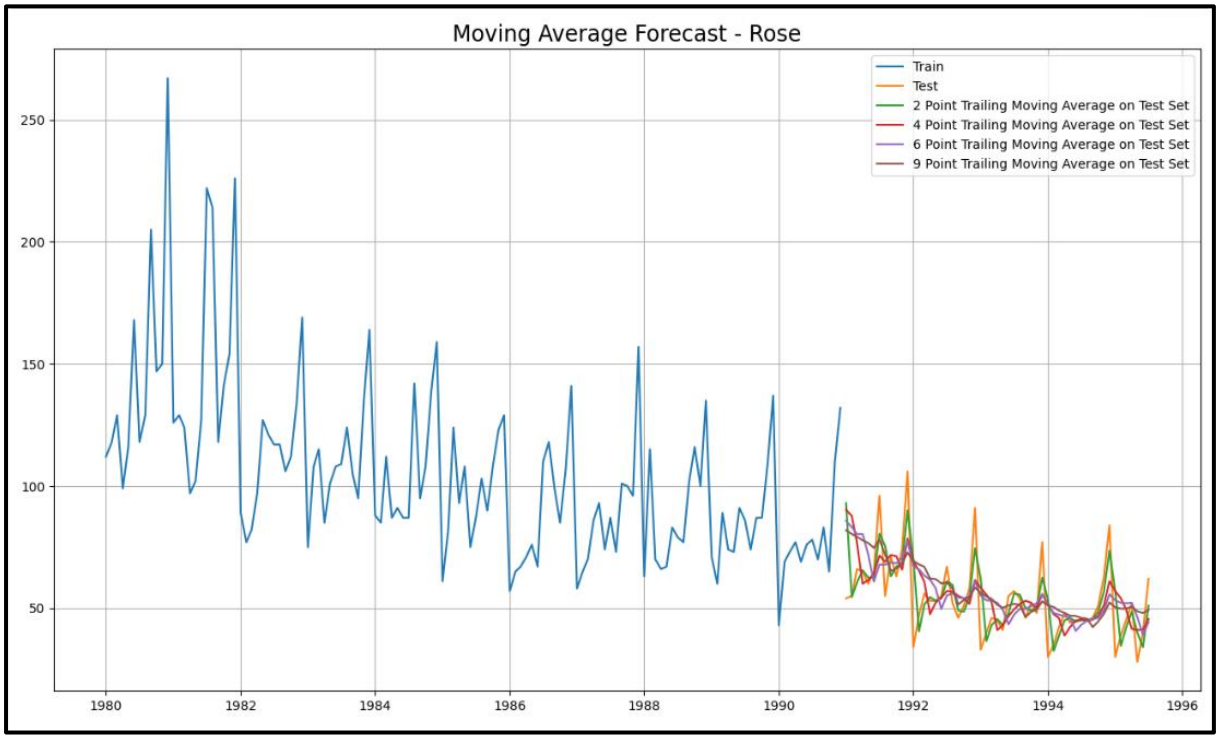


Figure 41 Moving Average for overall Training set

- **Training Data**:
    - The blue line represents the training data, showing overall sales trends from 1980 to 1996. It has significant peaks and valleys, indicating fluctuating sales.
- **Test Data**:
    - The test data (orange line) largely follows the trend established in the training data but appears to have less variability, suggesting a more stable sales environment in later years.

- **Moving Averages**:
    - Multiple moving averages (2, 4, 6, and 9 points) are represented by different colored lines. They smooth out the sales data, indicating overall trends clearer than the irregularities of the actual sales data.
    - The 9-point moving average (purple line) appears to be more stable, reflecting long-term trends more reliably than the shorter averages.
- **Trend Analysis**:
    - The moving averages show a general downward trend, particularly in more recent years, which may indicate declining interest in rosé wine.

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

Figure 42 Moving Average Test RMSE Rose

Observations:

- **2-Point Moving Average**:
    - The **2-point moving average** model has the lowest RMSE (**11.53**), suggesting it is the most accurate among the moving average models for this dataset. This indicates that the most recent two months' data are the best predictor for future values in this case.
- **4-Point Moving Average**:
    - The **4-point moving average** model shows a slightly higher RMSE (**14.46**), but still performs well. The slight increase in error may suggest that using a wider window (four points) introduces more noise or fails to capture short-term trends effectively.
- **6-Point and 9-Point Moving Averages**:
    - Both the **6-point** and **9-point moving averages** have similar RMSE values (**14.57** and **14.73**, respectively), indicating that as the window size increases beyond 4 points, the model's performance plateaus or even worsens.
    - Larger windows might smooth out important short-term fluctuations, leading to higher prediction errors.

- **Trend Sensitivity**:
    - The **2-point moving average** seems to be most sensitive to recent trends, offering better accuracy in the short term. In contrast, the longer window averages are less sensitive and may struggle with capturing recent fluctuations.

Implications:

- The **2-point moving average** is likely the best choice for this dataset, as it provides the lowest error while still capturing recent patterns effectively. Larger windows (4-point, 6-point, or 9-point) do not significantly improve accuracy and might even introduce slight additional error.

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 43 Moving Average for Sparkling data

Observations:

Trailing Moving Averages:

- **Trailing_2**: The 2-point trailing average is calculated for the second row onward, using the current and previous month's data (e.g., for 1980-02-01, the average of 1686 and 1591 results in 1638.5).
- **Trailing_4, Trailing_6, Trailing_9**: These columns are showing NaN (Not a Number) for rows where there aren't enough previous data points to compute the moving average. For instance, the **Trailing_4** average requires 4 months of data, but only 3 months are available by 1980-04-01, so it's still NaN.

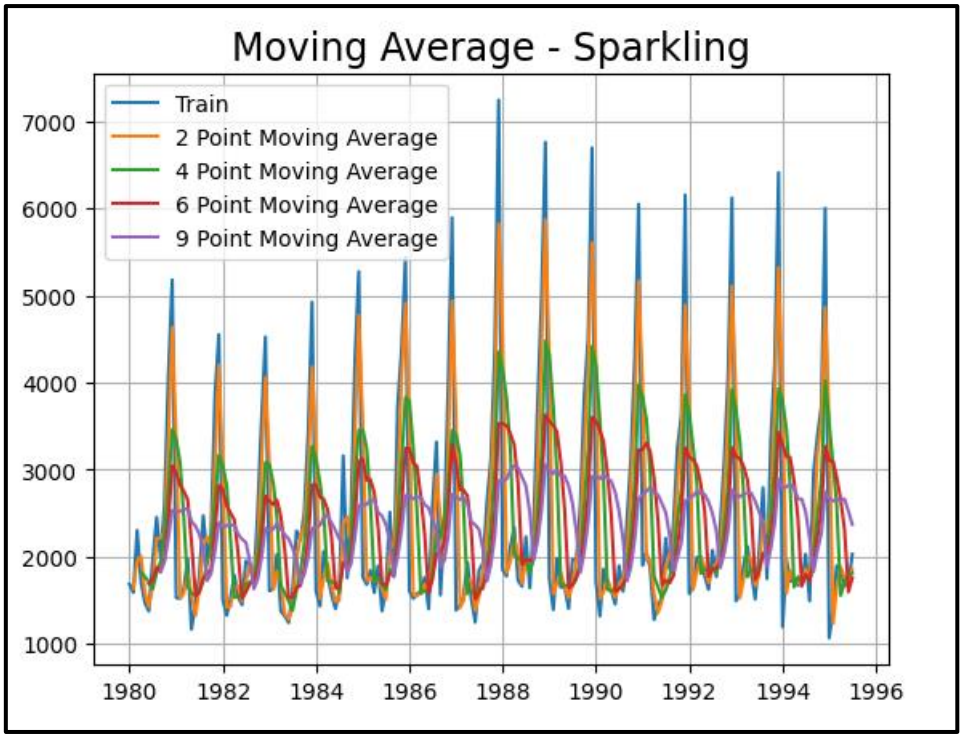


Figure 44 Moving Average - Sparkling For train data

- **Seasonal Sales Patterns**:

  The data reveals distinct seasonal spikes in sparkling wine sales, peaking particularly during certain months (likely around holidays and celebrations). This pattern indicates strong seasonal demand that businesses can target with marketing campaigns.

- **Fluctuations**:

  The training data (blue line) shows considerable fluctuations, with peaks reaching up to 7000 units and troughs dropping to around 1000 units. This volatility highlights the unpredictable nature of consumer demand within the sparkling wine market.

- **Moving Averages**:

  The different moving averages (2-point, 4-point, 6-point, and 9-point) help smooth out the data. The 9-point moving average (purple line) presents the most stable view of the trend, indicating long-term sales performance while minimizing the impact of short-term fluctuations.

- **Stabilizing Trend**:

  Although there are seasonal peaks, the moving averages suggest a general level of stability in the later years, with average sales remaining consistently around 2000 to 3000 units. This may imply that while seasonal demand varies, overall sales do not decline drastically year-on-year.

- **Average Sales Context**:

  The average sales, depicted by higher moving averages, indicate that while there are fluctuations, the baseline level of around 2000 to 3000 units is relatively stable, which is a positive indicator for forecasting future sales.

- **Strategic Opportunities**:

  The observed peaks could suggest specific months as prime opportunities for promotional activities. Understanding these patterns allows wineries and retailers to optimize inventory and marketing efforts to coincide with anticipated sales spikes.

- **Market Dynamics**:

  The consistent sales levels illustrated by the moving averages suggest that while sparkling wine faces competition and volatility, it maintains a solid consumer base.

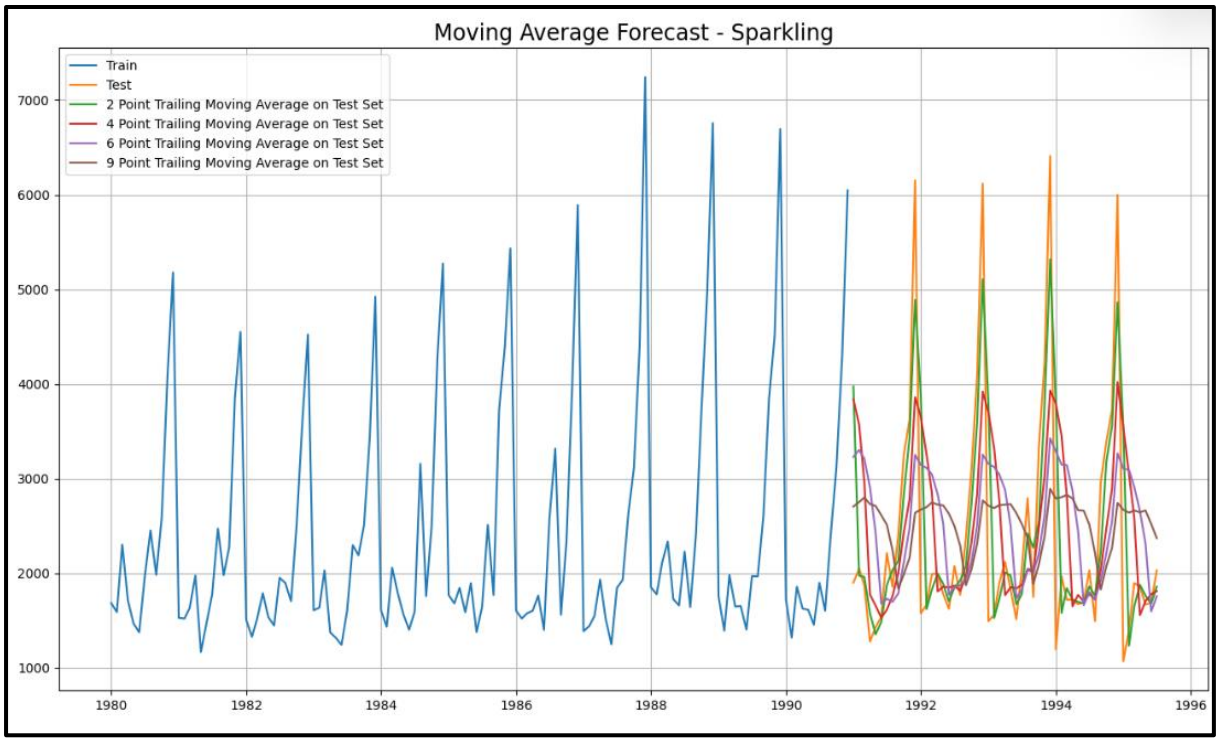


Figure 45 Moving Average Forecast - Sparkling train data

- **Sales Variability**:
  - The training data (blue line) shows significant fluctuations in sparkling wine sales from 1980 to 1996, with peaks often exceeding 6000 units and occasional drops to around 1000 units. This variability indicates that consumer interest can dramatically change over the months, likely influenced by seasonal factors.

- **Seasonal Peaks**:
    - Clearly defined peaks in the sales data suggest that certain months experience significantly higher sales—likely during holidays, celebrations, or special occasions. Retailers can capitalize on these peaks through targeted marketing strategies during these times.
- **Moving Averages**:
    - The various moving averages (2-point, 4-point, 6-point, and 9-point) provide different perspectives on sales trends.
        - The **9-point moving average** (purple line) effectively smooths out short-term fluctuations, providing a clearer picture of the long-term trend, while other shorter moving averages react more sensitively to sales variations.
- **Decreasing Volatility**:
    - The test data (orange line) exhibits less volatility compared to the training data, which suggests that while overall sales trends may follow a similar pattern, the more recent years show a slight stabilization in the market.
- **Trend Identification**:
    - The slight downward trend observed from the moving averages may suggest potential challenges in sustaining high sales levels over time, signaling a need for strategic marketing efforts to reinvigorate interest.
- **Sales Forecasting**:
    - The smooth lines of the moving averages indicate potential tools for forecasting future sales patterns. This can aid businesses in planning production and inventory management based on expected seasonal demands.
- **Consumer Behavior Insights**:
    - The significant fluctuations in sales indicate varying consumer behavior year on year, suggesting that factors such as marketing influences, changing consumer preferences, or economic conditions could play a role.

Moving Average for Rose and Sparkling wine dataset

	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 46 Moving Average for Rose and Sparkling wine dataset

Observations:

- **2-Point Trailing Moving Average**:
    - The **2-point trailing moving average** model has the **lowest RMSE** (**813.40**), indicating the best predictive performance among the trailing averages. This aligns with the earlier observation that shorter windows react more swiftly to fluctuations in sales.
- **4-Point Trailing Moving Average**:
    - The **4-point trailing moving average** has a higher RMSE (**1156.59**) compared to the 2-point model, suggesting that the model becomes less accurate as it tries to smooth out more data points.
- **6-Point and 9-Point Trailing Moving Averages**:
    - The **6-point** and **9-point** moving averages show progressively higher RMSE values (**1283.93** and **1346.28**, respectively), indicating that longer windows smooth out the data too much, leading to greater prediction error.
    - These longer windows are less responsive to recent trends and may fail to capture sudden changes in sales.
- **Trend**:
    - As expected, **shorter windows** (2-point) provide more accurate, responsive forecasts, whereas **longer windows** (4, 6, and 9 points) introduce more smoothing, which reduces sensitivity to short-term fluctuations, leading to higher errors.

4.4 Model Comparison

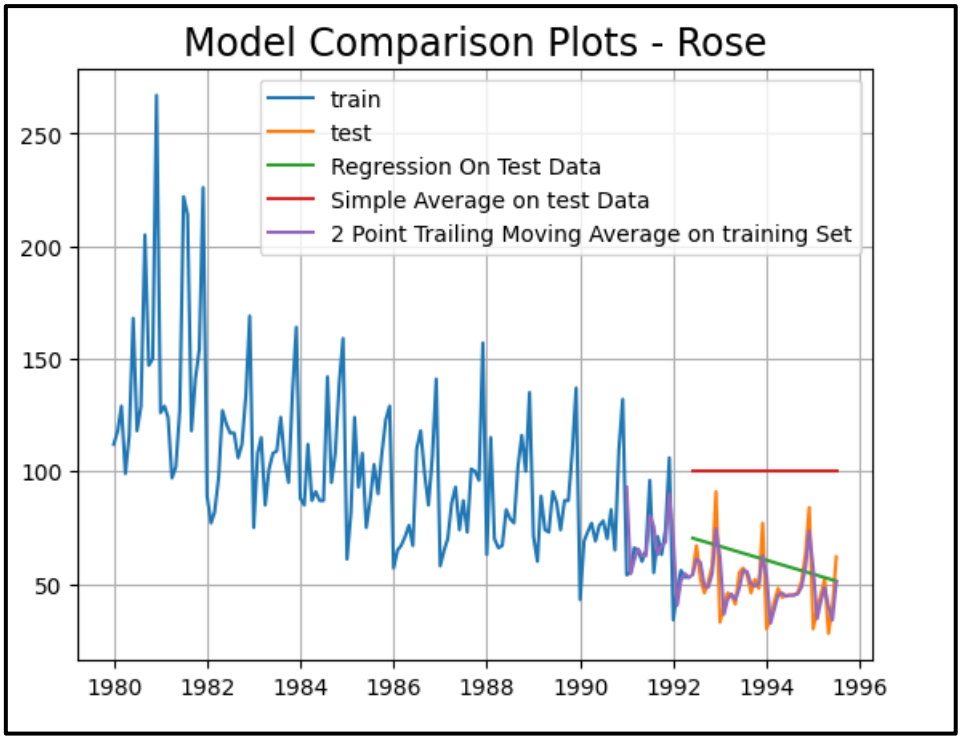


Figure 47 Model Comparison plots - Rose

Rosé Wine Sales Insights

- **Overall Sales Decline**:
    - The training data for rosé wine also shows variability, but the scale of sales is much lower, peaking around 250 units. The trend demonstrates a general decline over time, highlighting decreasing interest.
- **Lower peaks and Stability**:
    - The absence of sharp spikes indicates rosé wine does not benefit from seasonal peaks like sparkling wine. The sales appear more stable but at a consistently lower level.
- **Moving Averages and Trends**:
    - The 2-point moving average (purple line) illustrates a downward trend, confirming the overall decline in interest and sales for rosé wine. This emphasizes a potential need for strategic marketing or product rebranding.
- **Limited Response to Seasonality**:
    - The more stable yet declining pattern in sales indicates that rosé may not have the same strategic seasonal marketing opportunities available to sparkling wine.

Test Data Insights:

- The test data (orange line) remains low and stabilizes at a lower range, further emphasizing the need for investigating new strategies to drive sales growth.

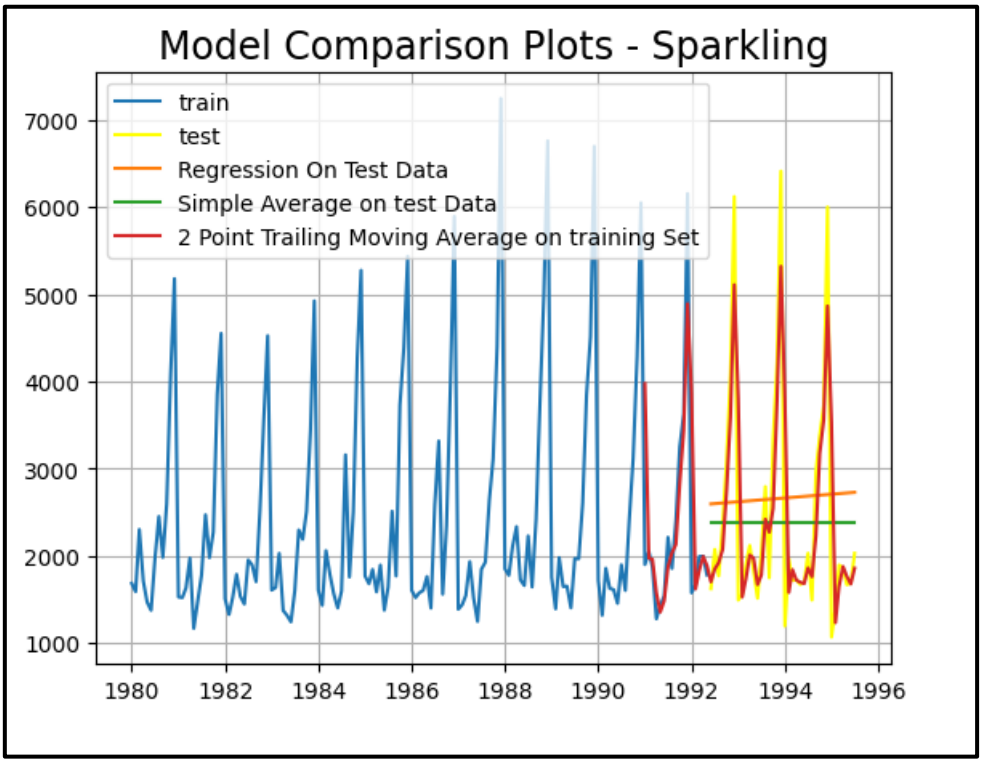


Figure 48 Model Comparison plots - Sparkling

Sparkling Wine Sales Insights

- **Sales Trends**:
  - The training data (blue line) shows significant variability in sales from 1980 to 1996, with peaks commonly exceeding 6000 units. This indicates a strong but fluctuating consumer interest in sparkling wine.
- **Seasonal Peaks**:
  - The regular peaks likely correlate with holiday seasons and special events, indicating that sparkling wine sales are heavily influenced by seasonal demand. Marketing efforts can be optimized around these times.
- **Moving Averages**:
  - The **2-point trailing moving average** (red line) smooths out short-term fluctuations but still reflects the general trends in the data. This average helps identify a more stable view of performance over time.

- **Test Data Stability**:
    - The test data (orange line) exhibits some instability but generally follows the trends seen in the training data. Notably, this data does not show as much volatility, suggesting potential stabilization in more recent years.
- **Regression Analysis**:
    - The regression line on the test data (green line) suggests potential future sales trends, providing businesses insights for forecasting and strategy.

4.5 Exponential Models (Single, Double, Triple)

4.5.1 SImple Exponential Smoothing with additive errors – ROSE

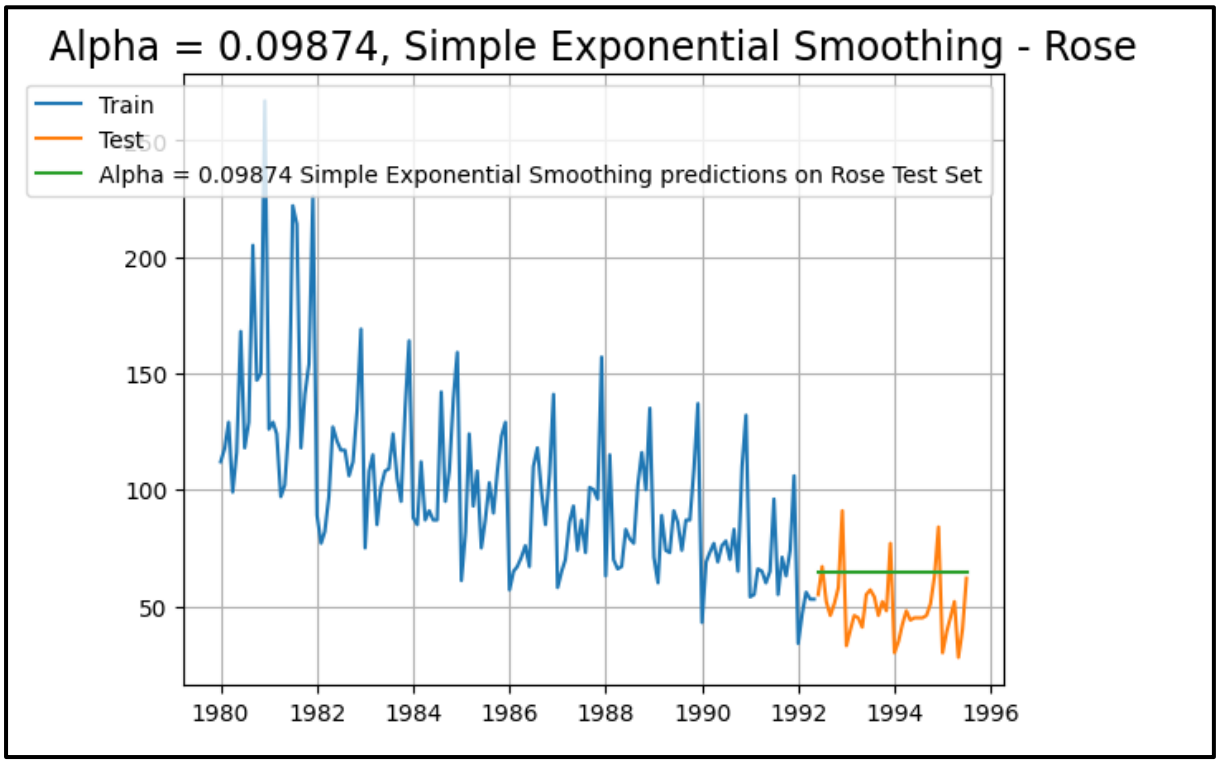


Figure 49 SImple Exponential Smoothing with additive errors - ROSE

Rosé Wine Insights

- **Declining Sales**:
    - The training data indicates a peak around 200 units, but the overall trend shows a slow decline, suggesting decreasing interest in rosé wine over time.
- **Minimal Fluctuations**:
    - The sales data presents less variability compared to sparkling wine, indicating a more stable but lower consumer base.
- **Exponential Smoothing Results**:

- The predicted line (alpha = 0.09874) for rosé suggests a higher sensitivity to recent sales data, reflecting the declining trend. This indicates the need for more vigorous strategies to reverse this trend.

- **Future Sales Outlook**:
  - The test data remains flat at lower levels of around 50 units, highlighting an ongoing challenge in revitalizing the rosé wine market and implying that without intervention, sales may stagnate.

4.5.2 SImple Exponential Smoothing with additive errors – SPARKLING

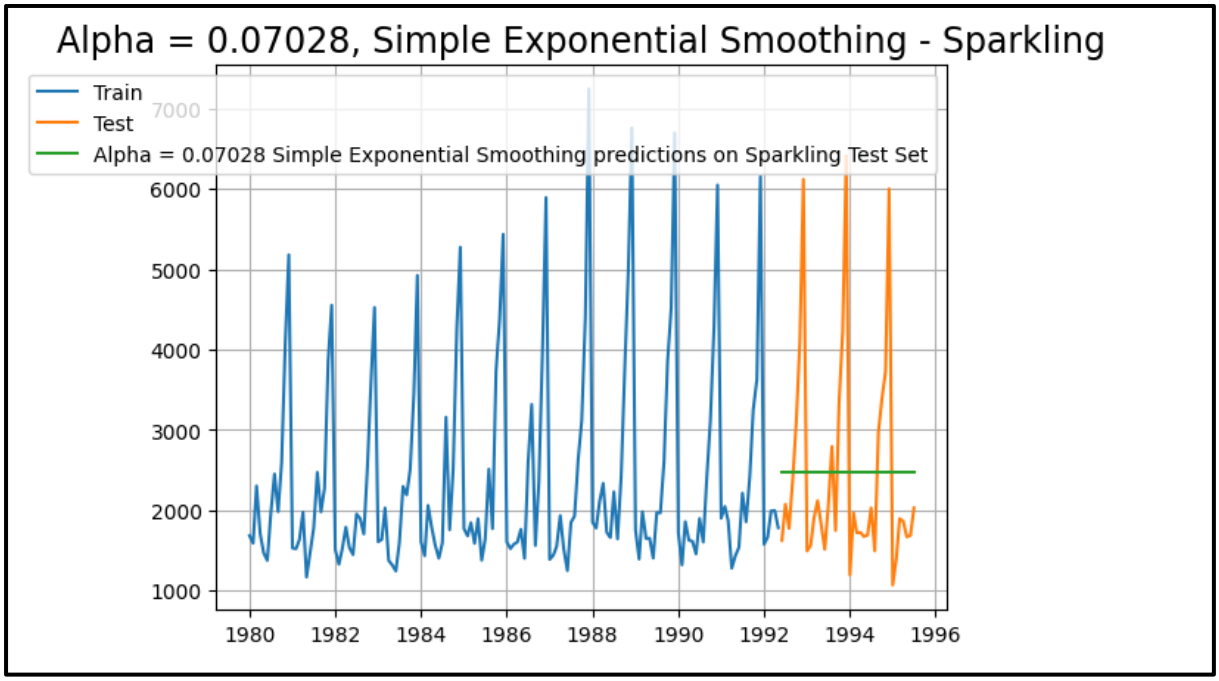


Figure 50 SImple Exponential Smoothing with additive errors - SPARKLING

Sparkling Wine Insights

- **Sales Trends**:
  - The training data reveals substantial variability, with sales peaks reaching over 6000 units, indicating strong demand during certain periods.

- **Seasonal Patterns**:
  - The pronounced spikes in sales typically correlate with festive occasions, suggesting that seasons have a significant impact on consumer purchasing behavior.

- **Exponential Smoothing**:
    - The green line, representing the predicted sales using a smoothing factor (alpha = 0.07028), indicates a moderate rate of adjustment to recent sales data. This suggests a balance between responsiveness to recent trends and stability.
- **Testing Period and Stability**:
    - The test data demonstrates some fluctuations, though not as extreme as the training data, implying a potential stabilization of sales in more recent years. The smooth prediction line suggests that future sales may continue to hover around an approximate average.

Simple Exponential Smoothing for Rose and Sparkling wine dataset

	Test RMSE Rose	Test RMSE Sparkling
Simple Exponential Smoothing	20.313631	1329.402402

Figure 51 Simple Exponential Smoothing for Rose and Sparkling wine dataset

Observation:

- **Simple Exponential Smoothing for Rose**:
    - The **Simple Exponential Smoothing** model gives an **RMSE of 20.31**, indicating relatively low error in predicting Rose sales, suggesting the model captures the underlying trend and level well without introducing significant bias or large deviations.
- **Comparison with Other Models**:
    - The RMSE of 20.31 is higher than the **RegressionOnTime** model (17.51) but lower than the **Simple Average Model** (52.24), indicating that Simple Exponential Smoothing strikes a balance between simplicity and accuracy, outperforming the simple average model.
- **Model Suitability**:
    - Simple Exponential Smoothing works well for data with little to no trend or seasonality (or very subtle trends), which might be the case with Rose sales over the time period in your dataset.

- **Error Significance**:
    - The RMSE of **20.31** suggests that while the model is reasonably accurate, there's still some room for improvement, especially if there are hidden seasonality patterns or trends that the model doesn't capture fully.

4.5.3 Double Exponential Smoothing with Addition Errors – Rose

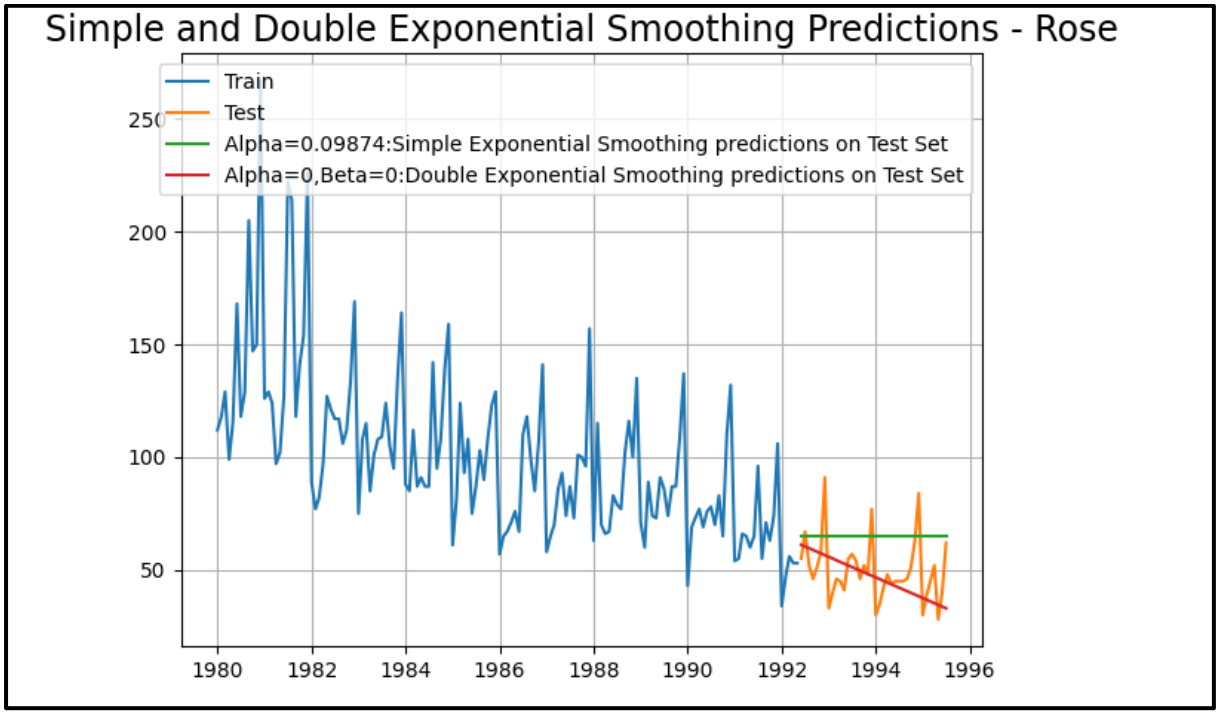


Figure 52 Simple and Double Exponential Smoothing with Addition Errors - Rose

Rose Wine Insights

- **Overall Decline**:
    - The training data for rosé wine (blue line) reveals a relatively flat trend with some peaks, but overall, the sales have been declining over time, with peaks around 250 units.
- **Minimal Sales Fluctuations**:
    - Compared to sparkling wine, rosé wine sales show less variability, indicating a stable but lower level of consumer interest.
- **Simple Exponential Smoothing**:
    - The green line indicates the predictions made using simple exponential smoothing (alpha = 0.09874). This model suggests a slight downward trend, reflecting ongoing challenges in the rosé wine market.

- **Double Exponential Smoothing Predictions**:
    - The red line for double exponential smoothing is less relevant here, as its effectiveness diminishes due to the flat trend, indicating little to no significant growth or decline.
- **Market Challenges**:
    - The predictions suggest stagnant sales for rosé wine unless significant marketing changes or product innovations are pursued to stimulate demand.

4.5.4 Double Exponential Smoothing with Addition Errors – Sparkling

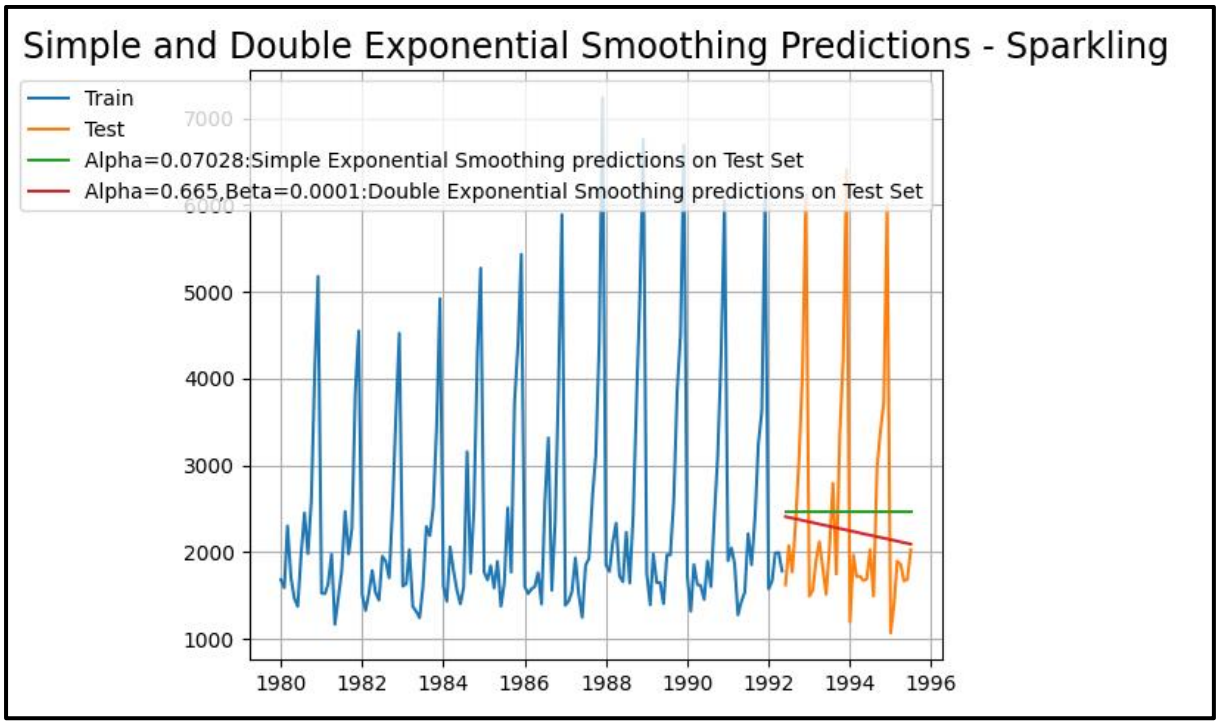


Figure 53 Simple and Double Exponential Smoothing with Addition Errors - Sparkling

Sparkling Wine Insights

- **Sales Trends**:
    - The training data (blue line) shows significant fluctuations, with peaks often exceeding 6000 units, highlighting strong consumer demand during certain periods.
- **Seasonal Demand**:
    - The seasonal spikes indicate likely connections to holidays and occasions that drive increased sparkling wine sales, suggesting marketing strategies could effectively leverage these periods.

- **Simple Exponential Smoothing**:
    - The green line represents predictions using simple exponential smoothing (alpha = 0.07028). This approach captures some of the underlying trend while smoothing out noise, indicating a moderately stable outlook despite variations.
- **Double Exponential Smoothing**:
    - The red line reflects double exponential smoothing (alpha = 0.665, beta = 0.0001), which accounts for trends. This method appears to provide a better fit for the sales data, indicating a clearer trajectory and a more responsive prediction model.
- **Future Outlook**:
    - Both the smoothed predictions provide insights for future sales, with double exponential smoothing suggesting a clearer stabilization trend compared to the more reactive simple model.

Double Exponential Smoothing for Rose and Sparkling wine dataset

	Test RMSE Rose	Test RMSE Sparkling
Double Exponential Smoothing	14.623742	1340.452791

Figure 54 Double Exponential Smoothing for Rose and Sparkling wine dataset

Observations:

- **Double Exponential Smoothing for Rose**:
    - The **Double Exponential Smoothing** model has an **RMSE of 14.62** for Rose, which is **lower** than both **Simple Exponential Smoothing** (20.31) and **RegressionOnTime** (17.51), indicating that Double Exponential Smoothing performs better for this dataset by capturing both level and trend more effectively.
- **Double Exponential Smoothing for Sparkling**:
    - The **RMSE for Sparkling** is **1340.45**, which is slightly lower than the **RegressionOnTime** RMSE (1349.04) and the **Simple Average Model** (1331.04). However, it still represents a high error, suggesting that the model is not very effective at forecasting Sparkling sales and may not fully capture the volatility and complexity in the data.

- **Model Performance**:
    - The **Rose** dataset benefits from **Double Exponential Smoothing**, which accounts for both trend and level in the data. However, for **Sparkling** sales, the model does not significantly outperform simpler models, suggesting that the data may require more advanced methods that also account for seasonality and other dynamics.
- **Trend and Seasonality**:
    - **Double Exponential Smoothing** is useful when there is a linear trend in the data, which appears to be the case for **Rose**. For **Sparkling**, the lack of improvement might indicate that the dataset has non-linear patterns or seasonality that the model doesn't address.

4.5.6 Triple Exponential Smoothing with Addition Errors – Rose

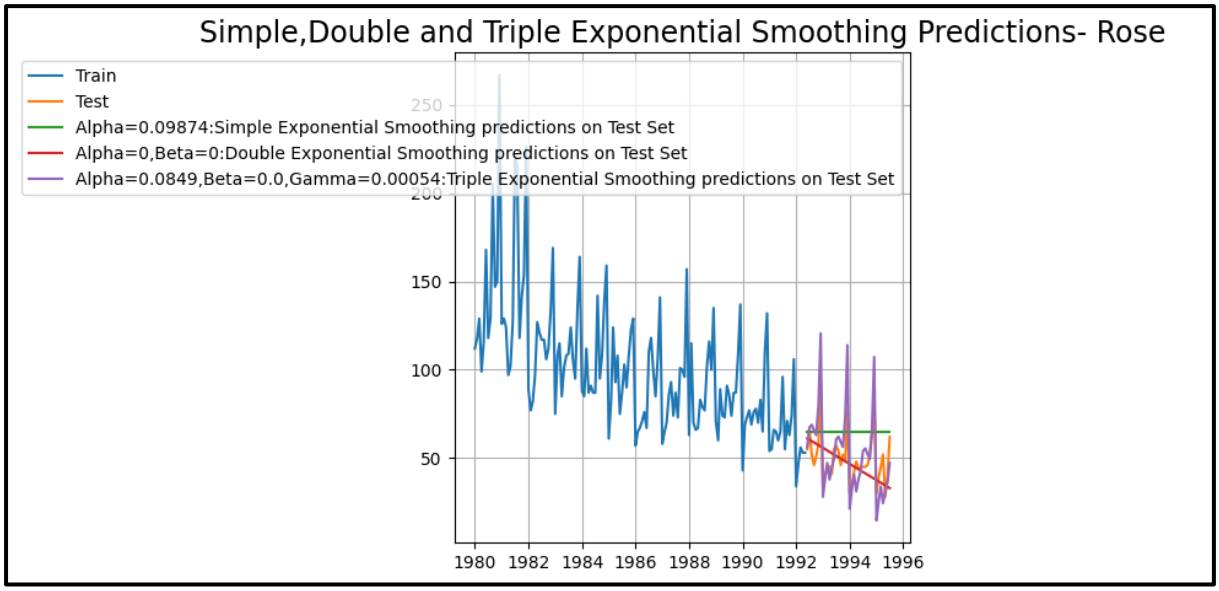


Figure 55 Simple, Double and Triple Exponential Smoothing with Addition Errors - Rose

Rose Wine Insights

Sales Decline:

- The rose wine sales graph suggests a trend of declining consumer interest, with sales peaking around 250 units but generally trending downward over the observed years.

Smoothing Models:

- **Simple Exponential Smoothing** (green line, alpha = 0.09874): This generates a basic trend line that captures the general decline in sales.

- **Double Exponential Smoothing** (red line, alpha = 0.0, beta = 0.0): The lack of trend responsiveness might limit its effectiveness, leading to flat predictions.
- **Triple Exponential Smoothing** (purple line, alpha = 0.0849, beta = 0.0, gamma = 0.00054): This model indicates minimal seasonal adjustment and reflects a stable yet declining trend, suggesting ongoing challenges for growth.

Future Sales Outlook:

- The predictions indicate that assuming current trends continue, rosé wine sales will remain low, around 50-70 units, emphasizing the need for strategic marketing shifts.

Strategic Opportunities:

- Rose wine's performance suggests the potential benefit of leveraging innovation and promotional efforts to stimulate sales, particularly since the existing market appears stagnant.

4.5.7 Triple Exponential Smoothing with Addition Errors – Sparkling

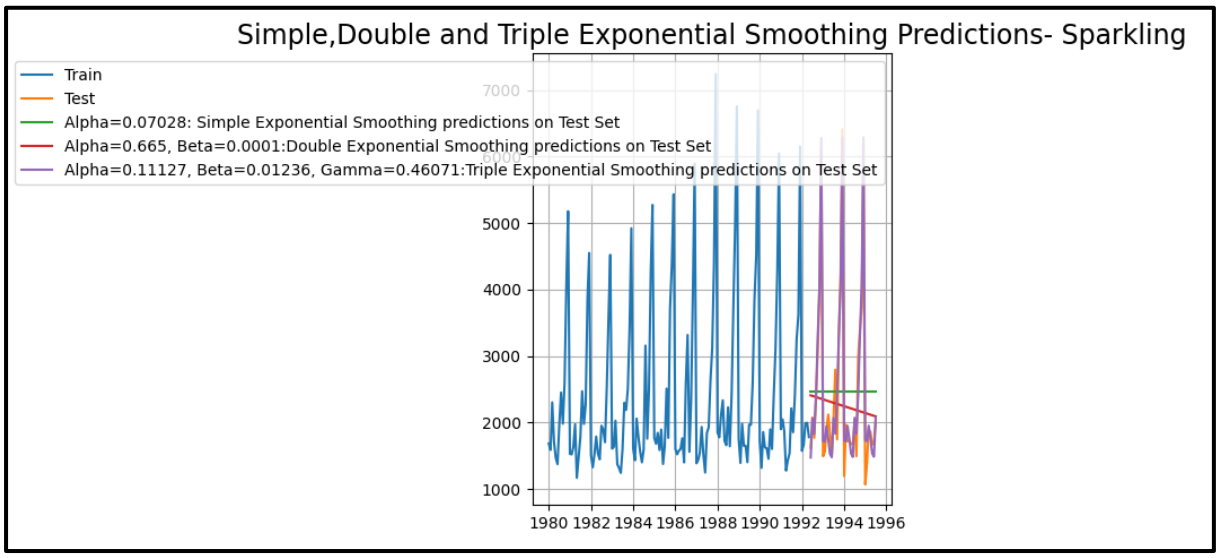


Figure 56 Simple, Double and Triple Exponential Smoothing with Addition Errors - Sparkling

Sparkling Wine Insights

- **Sales Variability**:
  - The training data shows considerable variability with peaks often exceeding 6000 units. This indicates strong consumer demand, particularly influenced by seasonal purchasing behavior.

- **Exponential Smoothing Models**:
    - **Simple Exponential Smoothing** (green line, alpha = 0.07028): Provides a smoothed prediction but might not fully capture trends due to lower responsiveness to recent changes.
    - **Double Exponential Smoothing** (red line, alpha = 0.665, beta = 0.0001): This model accounts for trends and is more responsive to sales patterns, suggesting a gradual stabilization in the market.
    - **Triple Exponential Smoothing** (purple line, alpha = 0.11127, beta = 0.01236, gamma = 0.46071): This model is particularly useful if there are seasonal variations, providing the most nuanced predictions and potentially a better fit for the data.
- **Future Sales Stability**:
    - The triple smoothing predictions suggest a tendency for sales levels to stabilize around 2500-3000 units, providing a clearer outlook for inventory management and marketing strategies.
- **Seasonal Effects**:
    - The pronounced peaks align with holidays and special events, highlighting opportunities for targeted marketing campaigns.

Triple Exponential Smoothing - Additive for Rose and Sparkling wine dataset

	Test RMSE Rose	Test RMSE Sparkling
Triple Exponential Smoothing (Additive Season)	13.877335	304.247029

Figure 57 Triple Exponential Smoothing - Additive for Rose and Sparkling wine dataset

Observations:

- **Double Exponential Smoothing for Rose**:
    - The **RMSE of 14.62** for Rose indicates good performance for the Double Exponential Smoothing model, as it is lower than the **Simple Exponential Smoothing** (20.31) and **RegressionOnTime** (17.51) models. This suggests that the model is effectively capturing both the trend and level of Rose sales.
- **Double Exponential Smoothing for Sparkling**:
    - The **RMSE of 1340.45** for Sparkling is slightly better than the **RegressionOnTime** model (1349.04) but still relatively high. This suggests that Double Exponential Smoothing has limited success in predicting Sparkling sales, possibly due to more complex factors such as volatility or seasonality that the model does not account for.

- **Performance Comparison**:
    - For **Rose**, Double Exponential Smoothing improves forecast accuracy significantly by incorporating both trend and level adjustments.
    - For **Sparkling**, the model does not show much improvement over simpler models, indicating that more advanced techniques (like **Holt-Winters** or **ARIMA**) may be necessary to capture seasonal variations or other dynamics in the data.

- **Trend Sensitivity**:
    - Double Exponential Smoothing works well on datasets with a linear trend (as it appears to be the case with Rose) but may struggle with datasets that have more volatility, seasonality, or non-linear patterns (as seen with Sparkling).

4.5.8 Taking Multiplicative Seasonality- Rose

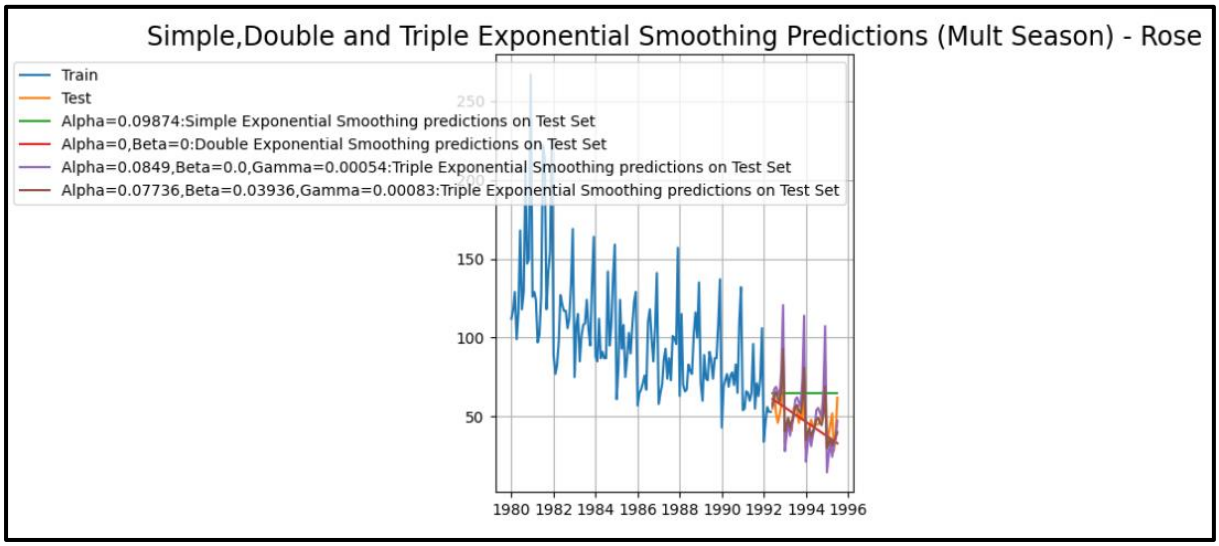


Figure 58 Simple, Double and Taking MULTIPLICATIVE SEASONALITY- ROSE

Rose Wine Insights

- **Sales Decline**:
    - The rose wine sales data reveals a general decline over time, with peaks reaching around 250 units. This indicates a weakening interest in rosé.
- **Prediction Models**:
    - **Simple Exponential Smoothing (green line)**:
        - Alpha = 0.09874: Captures the downward trend but shows limited responsiveness to fluctuations.

- **Double Exponential Smoothing (red line)**:
    - Alpha = 0.0, Beta = 0.0: This basic model implies no adjustments for trend, reflecting a relatively stagnant market.
- **Triple Exponential Smoothing (purple line)**:
    - Alpha = 0.07736, Beta = 0.03936, Gamma = 0.00083: Suggests minor seasonal adjustments but overall indicates an ongoing decline in interest and sales.

- **Stagnation and Challenges**:
    - The flat predictions at the end of the test period indicate potential stagnation, highlighting the need for innovative strategies to stimulate interest in rosé wine.
- **Need for Strategic Revitalization**:
    - Given the declining trend, exploring new marketing strategies or product innovations will be crucial for reversing sales trajectories.

4.5.9 Taking Multiplicative Seasonality- Sparkling

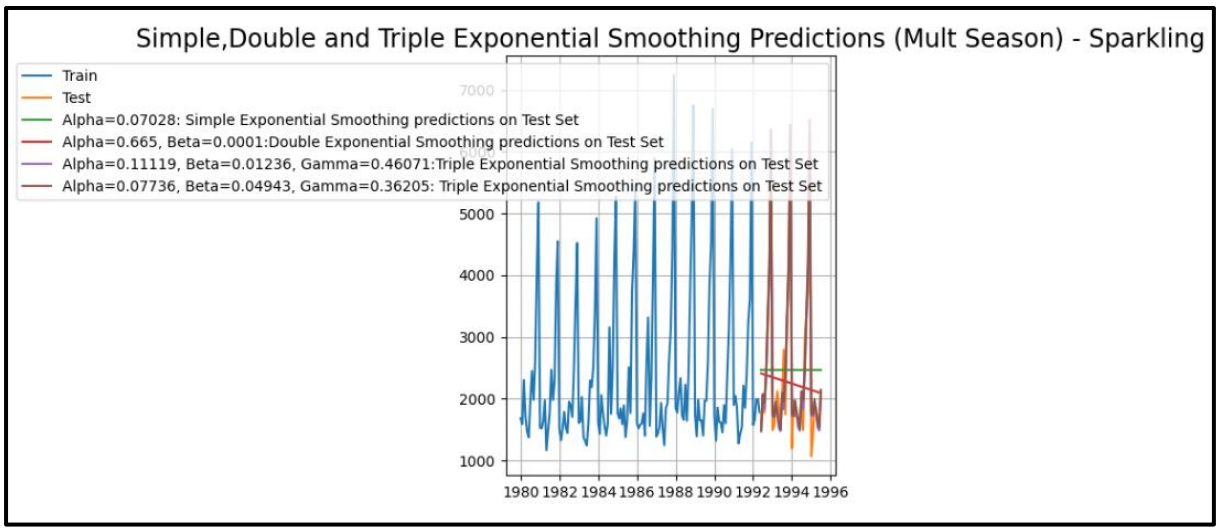


Figure 59 Simple, Double and Taking MULTIPLICATIVE SEASONALITY- Sparkling

Sparkling Wine Insights

- **Sales Fluctuations**:
    - The training data displays significant fluctuations, with notable peaks (exceeding 6000 units), indicating strong seasonal demand.
- **Prediction Models**:
    - **Simple Exponential Smoothing (green line)**:

        - Alpha = 0.07028: Provides a basic smoothing of the data but does not fully capture evolving trends.
    - **Double Exponential Smoothing (red line)**:
        - Alpha = 0.665, Beta = 0.0001: Accounts for trends, suggesting a more responsive model that better fits the training data. This model predicts smoother, more stable sales.
    - **Triple Exponential Smoothing (purple line)**:
        - Alpha = 0.11119, Beta = 0.01236, Gamma = 0.46071: This model captures seasonal variations effectively, providing the most detailed outlook for future sales and indicating a clear seasonal pattern.
- **Future Sales Trends**:
    - The predictive lines suggest potential stabilization in sales around 2500-3000 units, especially as the data progresses towards the end of the test period.
- **Effective Marketing Opportunities**:
    - The clear seasonal patterns indicate that targeted marketing efforts during peak seasons can significantly influence sales.

Triple Exponential Smoothing - Multiplicative for Rose and Sparkling wine dataset

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

Figure 60 Triple Exponential Smoothing - Multiplicative for Rose and Sparkling wine dataset

Observations:

- **Triple Exponential Smoothing (Multiplicative Seasonality) for Rose:**
    - The RMSE of 8.41 for Rose is the lowest among all the models tested so far, including Simple Exponential Smoothing (20.31), Double Exponential Smoothing (14.62), and Triple Exponential Smoothing (Additive Seasonality) (13.88). This suggests that Multiplicative Seasonality is the most effective in capturing the patterns in Rose sales, likely because sales fluctuations are proportional rather than constant.
- **Triple Exponential Smoothing (Multiplicative Seasonality) for Sparkling:**
    - The RMSE of 318.70 for Sparkling is a slight increase compared to Additive Seasonality (304.25), but it is still much lower than RegressionOnTime (1349.04) and Double Exponential Smoothing (1340.45). The multiplicative model performs well for Sparkling as well, suggesting that the seasonal

fluctuations in Sparkling sales are better captured by the multiplicative model, where seasonal variations scale with the level of the series.

- **Performance Comparison:**
  - For Rose, Multiplicative Seasonality in Triple Exponential Smoothing is the most accurate, significantly outperforming all other models, likely due to the proportional nature of the seasonal fluctuations.
  - For Sparkling, the multiplicative seasonality improves upon the additive seasonality, but still lags behind the Rose dataset in terms of accuracy, indicating that Sparkling has more complexity in its seasonal behavior.
- **Multiplicative vs. Additive Seasonality:**
  - Multiplicative seasonality works well for datasets where seasonal changes increase with the level of the data. This is likely the case for both Rose and Sparkling but is more pronounced in Rose, where the seasonal variations are proportionally significant.
  - Additive seasonality would be more suitable if the seasonal fluctuations are of similar magnitude, regardless of the level of the data, which seems less applicable in this case.

4.6 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 61 Checking the performance of the models built

Observations:

- **Best Model for Rose**:
  - **Triple Exponential Smoothing (Multiplicative Seasonality)** provides the **lowest RMSE (8.41)** for **Rose**, suggesting it is the best model for forecasting this data. This model handles both trend and seasonality effectively, capturing the seasonal fluctuations that are proportional to the level of sales.
- **Best Model for Sparkling**:
  - **Triple Exponential Smoothing (Additive Seasonality)** offers the **lowest RMSE (304.25)** for **Sparkling**, significantly improving accuracy over other models like **RegressionOnTime** (1349.04) and **Simple Exponential Smoothing** (1329.40). This model captures the seasonality and trend but still leaves room for improvement, as it doesn't fully account for all the complexities in the Sparkling data.
- **Models with High RMSE**:
  - The **Simple Average Model** (52.24 for Rose, 1331.04 for Sparkling) produces the worst results, indicating that it is not well-suited for forecasting either dataset. It fails to account for trends, seasonality, or fluctuations in the data.
- **Trailing Moving Averages**:
  - The **2-Point Trailing Moving Average** produces a relatively low RMSE for **Sparkling** (813.40), but it performs significantly worse for **Rose** (11.53), suggesting that this method is more responsive and accurate for datasets with more volatile trends like Sparkling.
- **Double Exponential Smoothing**:
  - **Double Exponential Smoothing** performs reasonably well for both datasets, with RMSE values of **14.62 for Rose** and **1340.45 for Sparkling**, suggesting it's useful for capturing the trend but may not be as effective at modeling seasonality compared to the Triple Exponential models.
- **Comparing the Trailing Moving Averages**:
  - As the window size increases (from **2-Point to 9-Point**), the **RMSE for Sparkling** generally increases, indicating that larger windows may smooth out too much of the data and fail to capture short-term fluctuations. The **4-Point Trailing Moving Average** (14.46 for Rose, 1156.59 for Sparkling) provides a good balance for **Rose**, while **2-Point** is better for **Sparkling**.

Best model:

- **For Rose**, **Triple Exponential Smoothing (Multiplicative Seasonality)** is the most accurate model, followed by **2-Point Trailing Moving Average** for shorter-term forecasting.
- **For Sparkling**, **Triple Exponential Smoothing (Additive Seasonality)** is the best model, with **2-Point Trailing Moving Average** providing a competitive alternative.

5. Steps to Build and Evaluate ARIMA Models:

5.1.A Check Stationarity of Rose Data

Check Stationarity:

- Use the Augmented Dickey-Fuller (ADF) test to check if the time series is stationary.
- Apply differencing if needed (d > 0) to make the series stationary.

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

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

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

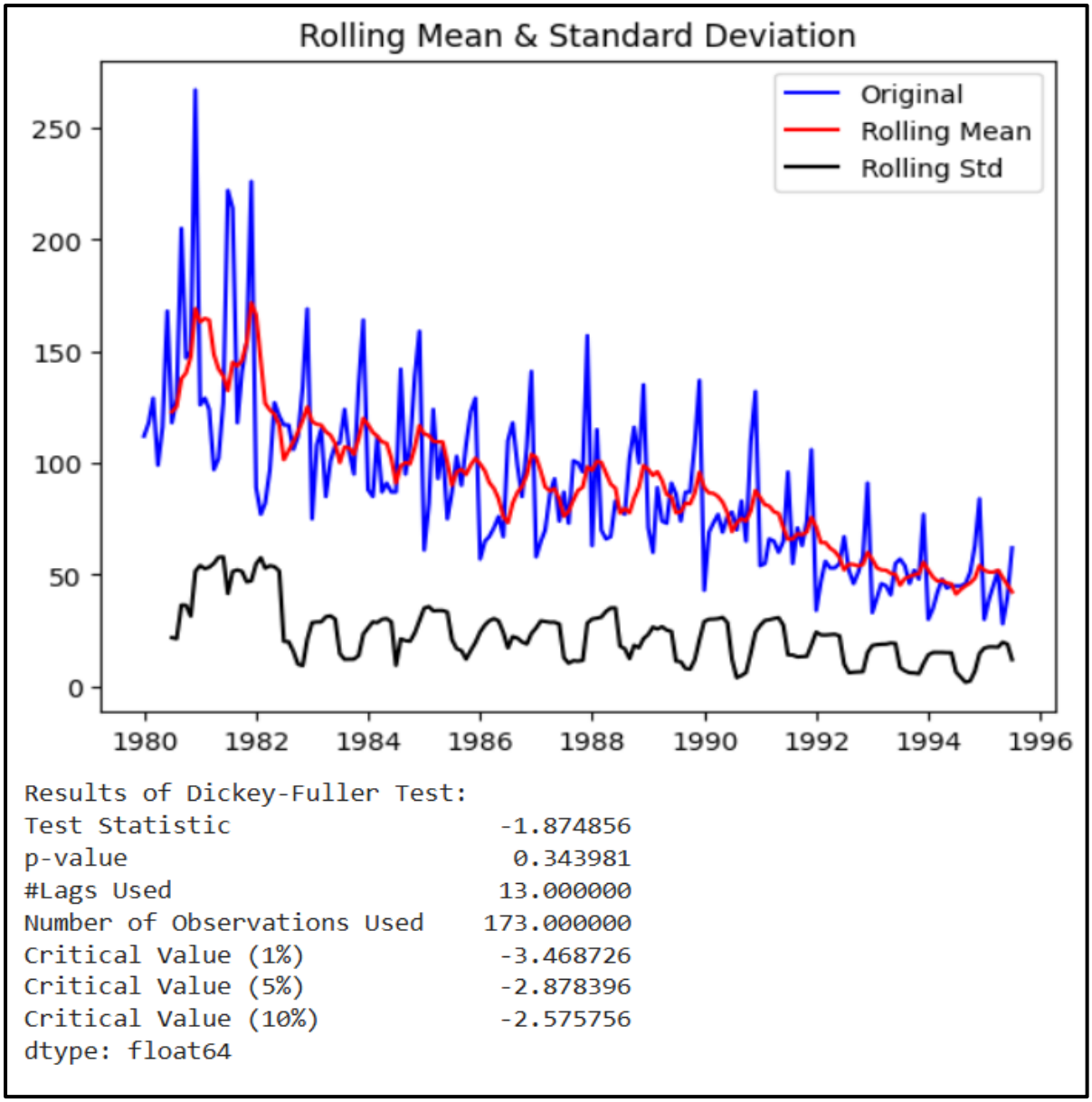


Figure 62 Result of Dickey - fuller test

Interpretation of the ADF Test Results:

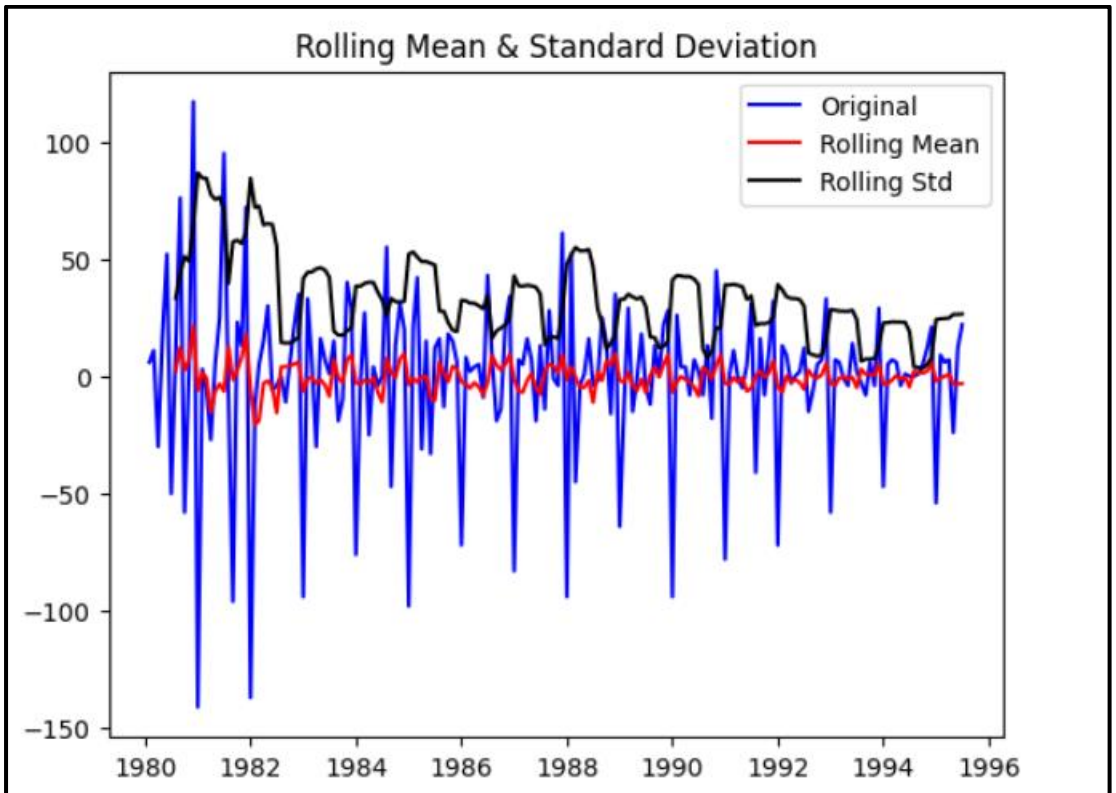
- **Test Statistic**:
  - The test statistic is **-1.875**.
  - This is compared to the critical values at different significance levels (1%, 5%, and 10%) to determine whether we can reject the null hypothesis.
- **p-value**:
  - The **p-value is 0.344**, which is **greater than the common significance levels** (e.g., 0.05 or 0.01).
  - This means we **fail to reject the null hypothesis** ($H0$), indicating that there is **insufficient evidence** to conclude the series is stationary.

- **Critical Values**:
    - At **1%** significance level, the critical value is **-3.4687**.
    - At **5%** significance level, the critical value is **-2.8784**.
    - At **10%** significance level, the critical value is **-2.5758**.

Since the test statistic **-1.875** is **greater** than the critical value at **1%**, **5%**, and **10%** significance levels (i.e., closer to zero), we **fail to reject** the null hypothesis.

Conclusion:

- Since the **p-value (0.344)** is greater than the significance level (typically 0.05), and the **test statistic (-1.875)** is greater than the critical values, we **fail to reject the null hypothesis**.
- Therefore, **the time series is non-stationary**, and it likely has a **unit root**.




```
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 63 Result of Dickey - Fuller Test with diff(1)

Interpretation of the Updated ADF Test Results:

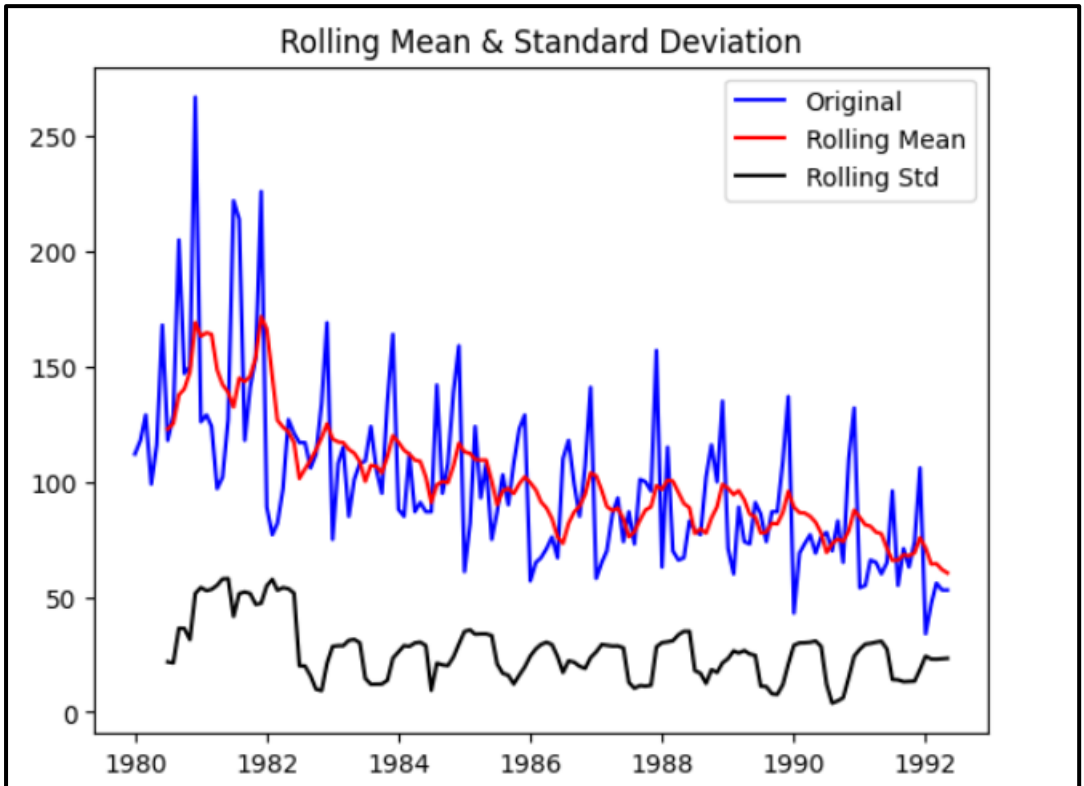
- **Test Statistic**:
    - The **test statistic** is **-8.0441**, which is a much more negative value compared to the previous result.
    - This suggests that the series is now likely **stationary**, as the test statistic is well below the critical values at all significance levels.
- **p-value**:
    - The **p-value is very small (1.81e-12)**, which is **much less than 0.05** (or even 0.01).
    - A small p-value indicates that we can **reject the null hypothesis**.
- **Critical Values**:
    - At **1% significance level**, the critical value is **-3.4687**.
    - At **5% significance level**, the critical value is **-2.8784**.
    - At **10% significance level**, the critical value is **-2.5758**.

The **test statistic (-8.0441)** is significantly more negative than all critical values, which further strengthens the conclusion.

Conclusion:

- Since the **p-value (1.81e-12)** is very small and the **test statistic (-8.0441)** is much lower than the critical values at all significance levels, we **reject the null hypothesis**.
- Therefore, we can conclude that **the time series is stationary**, meaning it does **not have a unit root** and can be used for modeling and forecasting without requiring further differencing or transformations.

Check for stationarity of the Training Data Time Series.




```
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 64 Results of Dickey - Fuller Test for train data

Interpretation of the ADF Test Results:

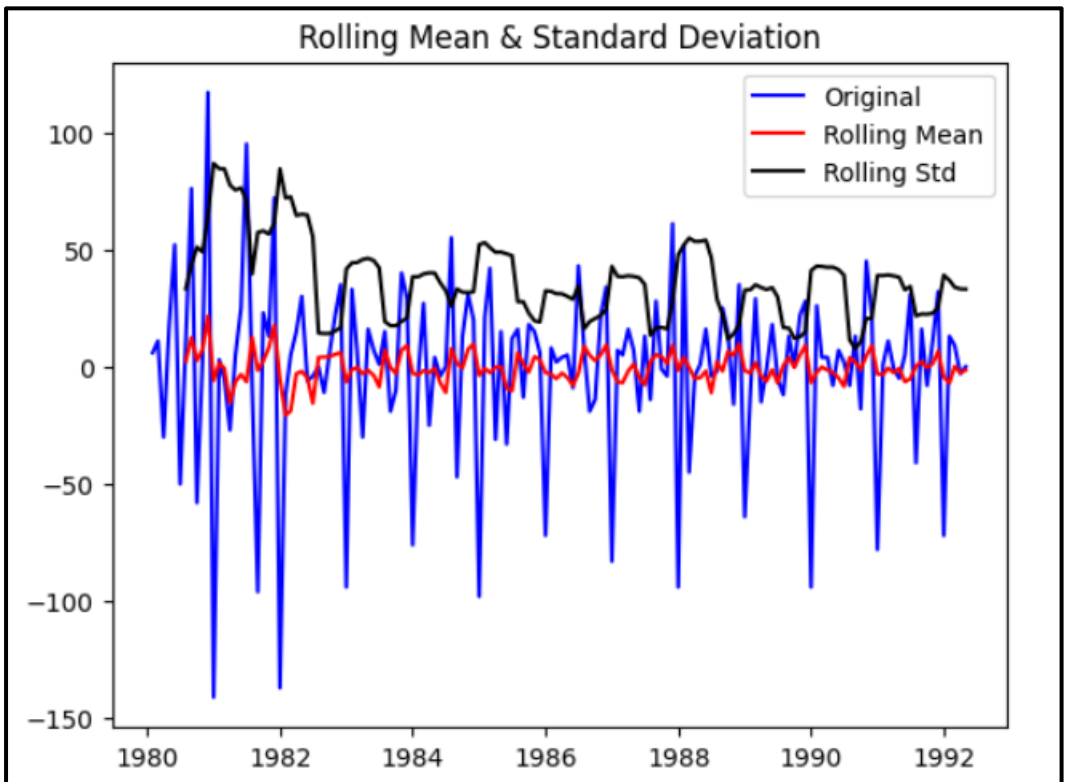
- **Test Statistic**:
  - The **test statistic** is **-1.6494**.
  - This value is **greater** (closer to zero) than the critical values at all significance levels (1%, 5%, and 10%).
- **p-value**:
  - The **p-value is 0.4574**, which is **greater** than the commonly used significance levels (e.g., 0.05 or 0.01).
  - This indicates that we **fail to reject the null hypothesis**.

- **Critical Values**:
    - At **1% significance level**, the critical value is **-3.4797**.
    - At **5% significance level**, the critical value is **-2.8832**.
    - At **10% significance level**, the critical value is **-2.5783**.

Since the **test statistic (-1.6494)** is **greater** than the critical values at all significance levels, we **fail to reject** the null hypothesis.

Conclusion:

- Given the **p-value (0.4574)** is greater than **0.05**, and the **test statistic (-1.6494)** is greater than the critical values, we **fail to reject the null hypothesis**.
- Therefore, the **time series is non-stationary**, and it 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 65 Results of Dickey - Fuller Test for train data with diff(1)

Interpretation of the Updated ADF Test Results:

- **Test Statistic**:
  - The **test statistic** is **-7.1321**, which is much more negative compared to the previous results.
  - This indicates that the time series is **likely stationary**, as it is well below the critical values at all significance levels.
- **p-value**:
  - The **p-value is 3.50e-10**, which is **extremely small** (far less than 0.05 or 0.01).
  - This suggests that we can **strongly reject the null hypothesis**.
- **Critical Values**:
  - At **1% significance level**, the critical value is **-3.4797**.
  - At **5% significance level**, the critical value is **-2.8832**.
  - At **10% significance level**, the critical value is **-2.5783**.

The **test statistic (-7.1321)** is significantly lower than all the critical values, which confirms that the time series is stationary.

Conclusion:

- Given the **p-value (3.50e-10)** is extremely small and the **test statistic (-7.1321)** is far below the critical values, we **reject the null hypothesis**.
- This means that **the time series is stationary** and does **not contain a unit root**.

Information of Train data of Rose

```
<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 66 Information of Train data of Rose

Key Points from the Output:

- DatetimeIndex: The index is a DatetimeIndex, indicating the data points are time series data.
- 149 Entries: You have data from 149 time points.
- Non-null Count: There are no missing values in the dataset, as all entries have valid values for the "Rose" column.

- Column Type: The "Rose" column is of type float64, which means it contains numeric data.
- Memory Usage: The dataset is small (2.3 KB in memory), so it can be easily processed.

5.2.A Identify ARIMA Parameters:

Check for stationarity of the Training Data - Rose

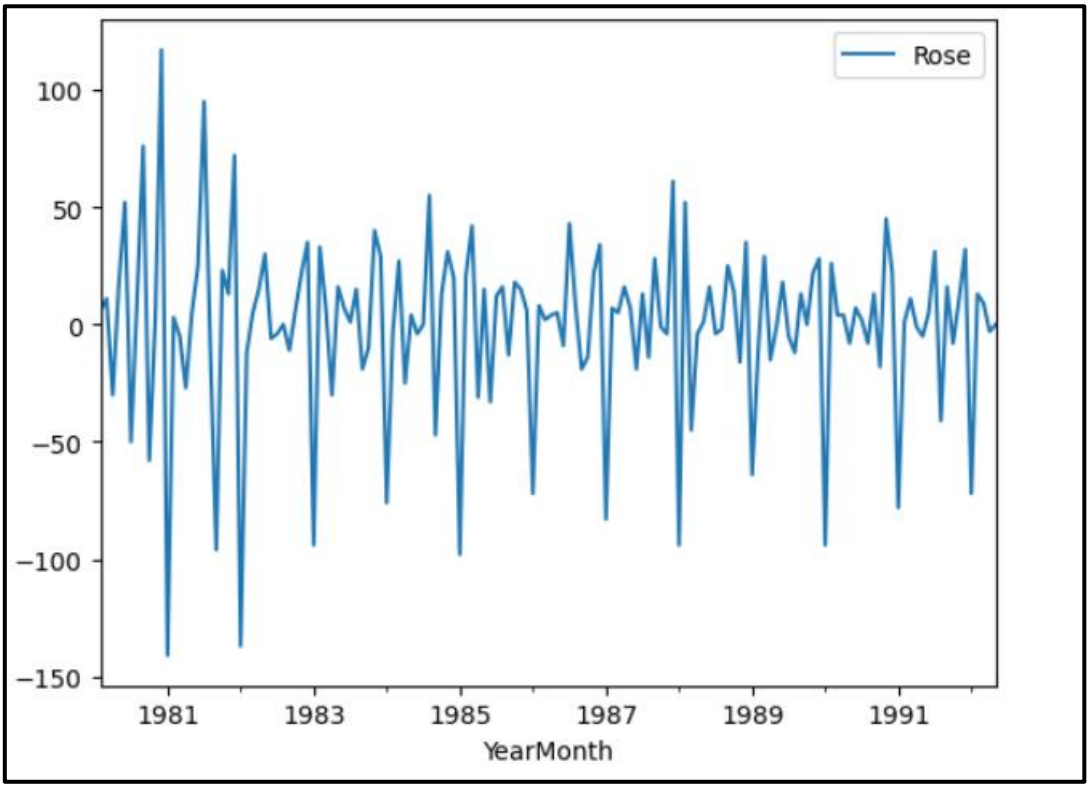


Figure 67 Check for stationarity of the Training Data - Rose

Insights from Rosé Wine Sales Data

- **Sales Variability**:
  - The graph shows substantial fluctuations in rosé wine sales over the period from 1981 to 1991. This suggests inconsistent consumer interest or sales performance.
- **Cycle Characteristics**:
  - The peaks and troughs in the data indicate cyclical behavior, which could be linked to seasonal factors or specific events that influence consumer buying patterns.

- **Negative Values**:
    - The presence of negative values suggests instances where sales may have dramatically dropped, possibly indicating stock issues, returns, or other market disruptions.
- **Lack of Clear Trend**:
    - Overall, there seems to be no clear upward or downward trend in sales, suggesting that the market for rosé might be stable but unexciting, without visible growth over time.
- **Potential for Improvement**:
    - The fluctuations present opportunities to explore factors that could stabilize sales, such as targeted marketing campaigns during peak seasons or enhancing product visibility.
- **Market Dynamics**:
    - The variability may reflect a response to changing consumer preferences or competition within the wine market, emphasizing the need for market analysis to understand these shifts.

Generate ACF & PACF Plot and find 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.

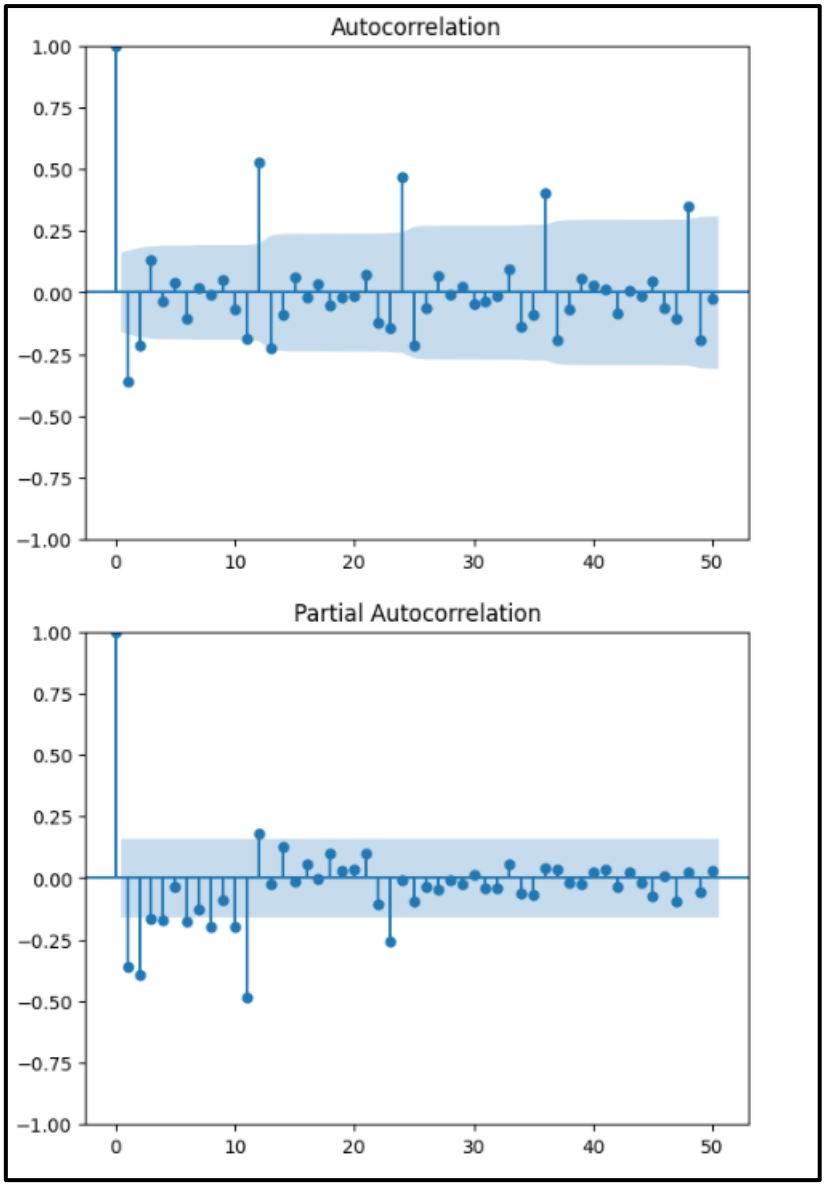


Figure 68 Generate ACF & PACF Plot

p(Auto Regressive Order):

- PACF shows a significant value at lag 1, with a rapid drop-off afterward. This indicates that an AR(1) or AR(2) model would be suitable. The first lag plays a key role in explaining the current value.
- p = 1 or p = 2 (since there is some decay but still significant structure at lag 1).

q (Moving Average Order):

- ACF shows a strong autocorrelation at lag 1 and some significant peaks at lags 3, 6, 12, and 24. These significant lags suggest that moving averages at these lags may capture important relationships.
- Given the significant autocorrelation at lag 1 and at higher lags (3, 6, 12, 24), q = 1 or q = 2 might be appropriate.

d (Differencing Order):

- Based on the Dickey-Fuller test (ADF test), if the series is already stationary, d = 0 (no differencing). If further differencing is needed, d = 1 could be an option to address any non-stationarity.
- d = 0 (if the series is already stationary).

P, Q, and D (Seasonal Parameters for SARIMA):

- ACF shows significant peaks at lag 12 and 24, suggesting seasonal patterns at these intervals. This could imply yearly seasonality, with the periodicity repeating every 12 lags (months, assuming monthly data).
    - P = 1 (for capturing the seasonal autoregressive component at lag 12).
    - Q = 2 (to capture the seasonal moving average effects at lags 12 and 24).
    - D = 1 (seasonal differencing might be necessary to address seasonality).

Final Model Parameters:

- p = 1 or p = 2
- q = 1 or q = 2
- d = 0 (or possibly 1 if further differencing is needed)
- P = 1
- Q = 2
- D = 1 (for seasonal differencing if seasonality is detected)

Suggested Models:

- ARIMA(1,0,1) or ARIMA(2,0,2) (if non-seasonal)
- 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).

5.2.A.1 Auto 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).

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.

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

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

Observations:

- The **model with the lowest AIC** is **(2, 0, 2)** with an AIC of **1448.97**, indicating it has the best fit among the tested models.
- The next best models are **(1, 0, 2)** with an AIC of **1449.51**, and **(2, 0, 1)** with an AIC of **1450.46**, which are still relatively good fits, but slightly worse than **(2, 0, 2)**.
- The models with higher AIC values (e.g., **(0, 0, 0)** with **1501.77**) represent worse fits, as indicated by the larger AIC numbers.

Conclusion:

- Based on the AIC values, the **(2, 0, 2)** model provides the best fit to the data among the ones tested. This model should be preferred for forecasting or further analysis.
- It's important to note that a model with a lower AIC is preferred, but other factors (such as interpretability, overfitting, and prediction accuracy) should also be considered when selecting the final model.

SARIMAX Results

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

Figure 70 SARIMAX Results for ARIMA (2,0,2)

Interpretation of Key Coefficients:

- const: The intercept term is 100.1395, meaning that the baseline value of the "Rose" series is around 100.14.
- ar.L1 (AR1): The coefficient is 0.5032, which suggests a positive relationship between the value at time t-1 and the value at time t. This indicates that past values have a positive impact on the current value of the series.
- ar.L2 (AR2): Similarly, the coefficient is 0.4858, suggesting a positive impact of the value at time t-2 on the current value.
- ma.L1 (MA1): The coefficient is -0.2144, which suggests a negative relationship between the previous forecast error and the current value.
- ma.L2 (MA2): The coefficient is -0.6135, indicating a stronger negative relationship between the forecast error at t-2 and the value at t.
- sigma2: This is the estimated variance of the residuals (894.4983), indicating the degree of variability in the model's predictions.

Conclusion:

- The ARIMA(2, 0, 2) model has a low AIC and appears to provide a good fit to the data based on the Ljung-Box test for autocorrelation.
- The moving average term (ma.L2) is statistically significant with a p-value of 0.001, indicating that past forecast errors are important for the model.
- The residuals do not show significant autocorrelation, but the normality test (Jarque-Bera) indicates some 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 71 Test RMSE and Test MAPE ARIMA (2,0,2)

Interpretation of the ARIMA(2, 0, 2) Model Performance:

Here are the performance metrics for the ARIMA(2, 0, 2) model based on the Test Data:

- Test RMSE (Root Mean Squared Error): 26.135542
    - RMSE is a measure of the average magnitude of the errors between the predicted and actual values.
    - A lower RMSE indicates better predictive accuracy. An RMSE of 26.14 suggests that, on average, the model's predictions deviate by around 26.14 units from the actual "Rose" values.
- Test MAPE (Mean Absolute Percentage Error): 56.709217
    - MAPE measures the percentage error between the predicted and actual values.
    - A lower MAPE indicates a better fit, with 56.71% suggesting that, on average, the predictions are off by around 56.71% from the actual values. While this is relatively high, it still indicates that the model provides useful forecasts, but there may be room for improvement.

Summary:

- The ARIMA(2, 0, 2) model appears to be a decent fit for the data with an RMSE of 26.14 and a MAPE of 56.71%.
- While the model provides reasonable accuracy, the MAPE value suggests that it may not be perfect, and further refinement or the use of alternative models could improve predictive performance.

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

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.

Sort the 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 72 Sort the AIC values in the ascending order to get the parameters for the minimum AIC value

Interpretation:

- The (1, 1, 2) model has the lowest AIC value of 1435.66, indicating it is the most optimal model among those considered in terms of fit and complexity.
- Models like (0, 1, 2) and (1, 1, 1) have slightly higher AIC values but may still be worth considering depending on the trade-off between fit and model simplicity.
- Models such as (2, 1, 0), (1, 1, 0), and (0, 1, 0) have relatively high AIC values, suggesting that they are less optimal.

Conclusion:

- Based on the AIC, the best-fitting model appears to be (1, 1, 2), followed closely by (0, 1, 2).
- You could consider using the (1, 1, 2) model for your forecasting task, but further diagnostic checks and validation on test data may be necessary to confirm its robustness.

SARIMAX Results

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	HQIC	1440.528
	- 05-01-1992		
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

Figure 73 SARIMAX Results for Rose ARIMA(1, 1, 2)

Parameter Estimates:

- AR.L1 (Autoregressive term lag 1): -0.4888 (p-value = 0.042)
    - This indicates a negative relationship with the previous value in the series. The p-value suggests statistical significance (since it is below 0.05).
- MA.L1 (Moving Average term lag 1): -0.2206 (p-value = 0.318)
    - This parameter is not statistically significant (p-value > 0.05). It suggests that the model might not need this lag to capture the series' behavior.
- MA.L2 (Moving Average term lag 2): -0.6175 (p-value = 0.001)
    - This is statistically significant, indicating that there is a strong negative influence from lag 2 of the error terms.
- Sigma2 (Variance of the error term): 895.98
    - The error term has a standard deviation of around 30 units (since the square root of 895.98 is approximately 30), indicating variability in the series not explained by the model.

Model Diagnostics:

- Ljung-Box (L1) Test: p-value = 0.79
    - The p-value is high, indicating that there is no significant autocorrelation in the residuals, which is a good sign for the model.

- Jarque-Bera Test: p-value = 0.00
    - The residuals are not normally distributed, indicating potential issues with model assumptions.
- Heteroskedasticity Test: p-value = 0.00
    - There is significant heteroskedasticity, meaning the variance of the residuals is not constant over time.
- Skewness: 0.83
    - The residuals have a positive skew, meaning the distribution has a longer tail on the right.
- Kurtosis: 5.15
    - The residuals have high kurtosis, indicating heavy tails (more extreme values than expected).

Interpretation:

- The model fits the data reasonably well based on the AIC (1435.66), but the diagnostic tests suggest that there are some issues:
    - Non-normal residuals and heteroskedasticity indicate that the model might not fully capture all the underlying patterns in the data.
    - The Ljung-Box test indicates that there is no significant autocorrelation, suggesting the model residuals are uncorrelated.

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
ARIMA(1,1,2)	20.915405	56.709217

Figure 74 Test RMSE and Test MAPE ARIMA (2,0,2) (1,1,2)

Interpretation:

- ARIMA(1,1,2) provides better performance than ARIMA(2,0,2), as evidenced by its lower RMSE value (20.92 compared to 26.14).
    - The RMSE measures the average magnitude of error, so a lower RMSE indicates better predictive accuracy.
- MAPE is identical for both models (56.71), which means both models have a similar relative percentage error when predicting the time series values.

Conclusion:

- ARIMA(1,1,2) appears to be the better model in terms of RMSE, making it more suitable for forecasting the "Rose" time series.
- Although both models show the same MAPE, the lower RMSE in ARIMA(1,1,2) suggests it provides more accurate point predictions.

5.2.A.2 Manual ARIMA

SARIMAX Results

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	HQIC	1456.784
	- 05-01-1992		
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 75 SARIMAX Results for rose ARIMA (1,0,1)

Insights:

- The AR(1) coefficient is close to 1 (0.9912), which suggests strong persistence in the time series, where previous values have a significant impact on future values.
- The MA(1) coefficient is negative (-0.8774), showing the moving average component's influence on the residuals.
- Despite the model's good fit based on the Ljung-Box test, the residuals display non-normality and some heteroskedasticity, suggesting potential improvements in the model might be needed (such as transforming the data or adding more seasonal components).

Diagnostics Plot

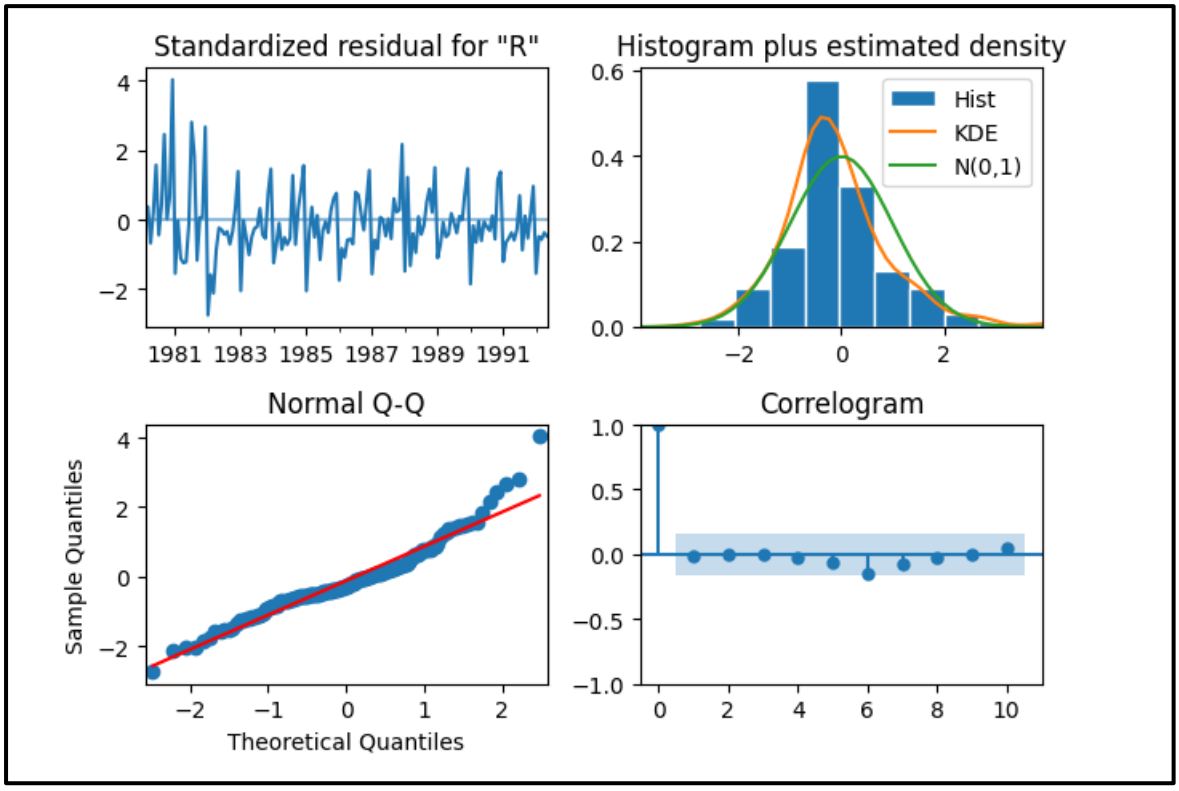


Figure 76 Diagnostics Plot

Standardized Residuals

- **Plot Interpretation:**
  - The standardized residual plot shows variations in residuals over time.
  - While the residuals appear centered around zero, there is noticeable volatility, suggesting potential non-constant variance (heteroscedasticity).
- **Implications:**
  - Variability in residuals may indicate that the model could be improved by incorporating more explanatory variables or applying different modeling techniques to better capture underlying patterns.

2. Histogram and Density Estimation

- **Histogram Analysis:**
  - The histogram, along with the Kernel Density Estimate (KDE) and normal distribution overlay, indicates that the residuals are somewhat normally distributed but exhibit some asymmetric behavior.

- **Key Observations:**
    - The peak is slightly skewed to the left, with some heavier tails. This underlines the importance of checking for outliers or extreme values impacting model performance.
    - Comparison with the normal distribution curve shows deviations, which suggests that the normality assumption of the errors may not hold perfectly.

3. Normal Q-Q Plot

- **Plot Comments:**
    - The normal Q-Q plot compares the quantiles of the residuals to the quantiles of a normal distribution.
    - Most points fall along the red line, but there are noticeable deviations in the tails.
- **Conclusion:**
    - Deviations from the line at both ends suggest slight departures from perfect normality, which could affect hypothesis tests and confidence intervals derived from the model.

4. Correlogram

- **Autocorrelation Insights:**
    - The correlogram shows that autocorrelations for lags are generally close to zero after lag 0, with no significant spikes beyond the confidence interval.
- **Indication:**
    - This implies that there is no significant autocorrelation in the residuals, which is a good sign. It suggests the model's errors are independent, aligning with model assumptions.

Summary Conclusion

- **Model Assessment:** The residual diagnostics indicate that while the model is performing adequately overall, there are areas for improvement:
    - **Variance and Normality:** Addressing potential heteroscedasticity and adjusting for non-normality could enhance model accuracy.
    - **Outliers:** Investigating outliers or extreme residuals may help refine predictions.

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

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

Figure 77 Test RMSE and Test MAPE

The evaluation metrics you have provided indicate the following:

- **RMSE (Root Mean Square Error): 25.63**

  This is the square root of the average of the squared differences between predicted and actual values. In this case, an RMSE of 25.63 means that the model's predictions deviate, on average, by about 25.63 units from the actual values.

- **MAPE (Mean Absolute Percentage Error): 55.45%**

  MAPE expresses the average absolute percentage difference between the predicted and actual values. A MAPE of 55.45% indicates that, on average, the model's predictions are off by about 55.45% in terms of percentage. This is quite high and suggests that the model may not be performing optimally, especially if the goal is high-accuracy predictions.

In summary:

- The RMSE shows a moderate prediction error in absolute terms.
- The MAPE indicates that the model's relative prediction error is quite high (over 50%).

SARIMAX Results

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 78 SARIMAX Results for ARIMA (2,0,2)

- AR(2) and MA(2) terms seem important for this model, while the AR(1) and MA(1) terms are less impactful.
- The model seems to perform well in terms of autocorrelation and residuals, although some signs of non-normality and heteroskedasticity are present in the residuals.
- AIC and BIC suggest the model fits the data well, with lower values indicating a better model fit.

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

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

Figure 79 Test RMSE and Test MAPE

Interpretation:

- The RMSE suggests that the model's predictions deviate from the actual values by approximately 26.14 units on average. This is a relatively high error, indicating that the model may not be fully capturing the underlying patterns in the data.
- The MAPE indicates that the model's predictions are off by about 56.7% on average, which is quite large. This suggests that the model is not very accurate in terms of

relative errors, especially for time series forecasting where lower MAPE values are desired.

Conclusion:

These metrics suggest that while the model may provide a general sense of the data, it could likely benefit from further refinement or adjustment. Exploring alternative model configurations, incorporating additional features, or even considering seasonal variations (if applicable) might improve the accuracy.

5.3.A Fit Multiple ARIMA Models:

5.3.A.1 Fit models with different combinations of (p, d, q)

	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 Fit models with different combinations of (p, d, q)

Observations:

- **ARIMA(1,1,2)** provides a significantly lower RMSE (20.9154), suggesting that this model has a better fit in terms of absolute errors.
- The **MAPE** for both **ARIMA(2,0,2)** and **ARIMA(1,1,2)** is the same (56.7092%), indicating that while **ARIMA(1,1,2)** does better in terms of RMSE, the relative percentage error is identical between the two models.
- The final entry seems to repeat the results of **ARIMA(2,0,2)** but with a different MAPE value of **0.5671**. This may represent a different scenario or model evaluation, possibly from another subset of data or a different calculation method.

Conclusion:

- **ARIMA(1,1,2)** appears to be the better model based on RMSE, as it yields lower values compared to **ARIMA(2,0,2)**.
- If you aim for minimizing absolute error, **ARIMA(1,1,2)** would be more favorable.
- If the goal is to reduce percentage errors, the models show equal performance, although the **MAPE** of 56.7092% still indicates a need for improvement in both cases.

5.3.A.2 Include Seasonal ARIMA (SARIMA) if seasonality is present (with parameters P, D, Q, and m)

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

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).

5.3.A.2.1 Auto SARIMA

Automated version of a SARIMA model for that the best parameters are selected with the lowest Akaike Information Criteria (AIC) - ROSE DATA

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

	param	seasonal_param	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1181.946217
80	(2, 1, 2)	(2, 0, 2, 6)	1183.927088
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 81 Sort the AIC values in the ascending order to get the parameters for the minimum AIC value

SARIMAX Results

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:	02:26:02	BIC	1205.069
Sample:	0	HQIC	1191.342
	- 149		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5985	0.130	-4.618	0.000	-0.853	-0.345
ma.L1	-0.1832	211.633	-0.001	0.999	-414.975	414.609
ma.L2	-0.8169	172.902	-0.005	0.996	-339.698	338.065
ar.S.L6	-0.0458	0.029	-1.561	0.118	-0.103	0.012
ar.S.L12	0.8760	0.029	30.337	0.000	0.819	0.933
ma.S.L6	0.1859	211.701	0.001	0.999	-414.740	415.112
ma.S.L12	-0.8141	172.297	-0.005	0.996	-338.511	336.882
sigma2	313.8603	1.123	279.376	0.000	311.658	316.062

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

Figure 82 SARIMAX Results SARIMA (1, 1,2) (2,0,2,6)

Coefficients:

- AR(1): -0.5985 (p-value < 0.01) - Significant
- MA(1): -0.1832 (p-value = 0.999) - Not Significant
- MA(2): -0.8169 (p-value = 0.996) - Not Significant
- Seasonal AR(6): -0.0458 (p-value = 0.118) - Not Significant
- Seasonal AR(12): 0.8760 (p-value < 0.01) - Significant
- Seasonal MA(6): 0.1859 (p-value = 0.999) - Not Significant
- Seasonal MA(12): -0.8141 (p-value = 0.996) - Not Significant
- Sigma^2: 313.8603 (p-value < 0.01) - Significant

Diagnostic Tests:

- Ljung-Box (Q) (L1): 0.02 (p-value = 0.88) - No significant autocorrelation
- Jarque-Bera (JB): 70.87 (p-value < 0.01) - Indicates non-normality in residuals
- Heteroskedasticity (H): 0.30 (p-value < 0.01) - Indicates heteroskedasticity in the model
- Skew: 0.55 - Slightly positive skew in residuals
- Kurtosis: 6.40 - High kurtosis, indicating fat tails

Interpretation:

- **Significant Parameters:**
  - AR(1) and Seasonal AR(12) are significant, meaning they play a role in predicting the time series.
  - Sigma^2 (variance of residuals) is also significant, indicating a well-defined model fit for error terms.
- **Non-Significant Parameters:**
  - MA(1), MA(2), and the seasonal moving average terms (SMA(6), SMA(12)) are not significant. These terms do not contribute substantially to the model's performance and could potentially be removed for a more parsimonious model.
- **Model Diagnostics:**
  - Ljung-Box Test shows no significant autocorrelation in the residuals, which is a good sign for the model.
  - Jarque-Bera Test suggests non-normality in the residuals, which may affect some statistical tests and inference.
  - Heteroskedasticity indicates the presence of changing variance over time, which may require further adjustments to the model (e.g., GARCH models).
  - The residuals exhibit skewness and kurtosis, implying that the errors are not perfectly normally distributed.

Conclusion:

- The model seems to capture the main temporal and seasonal effects with AR(1) and Seasonal AR(12) being the most important parameters.
- However, there may be a need for further refinement due to non-normality, heteroskedasticity, and the presence of insignificant terms.

Diagnostics Plot

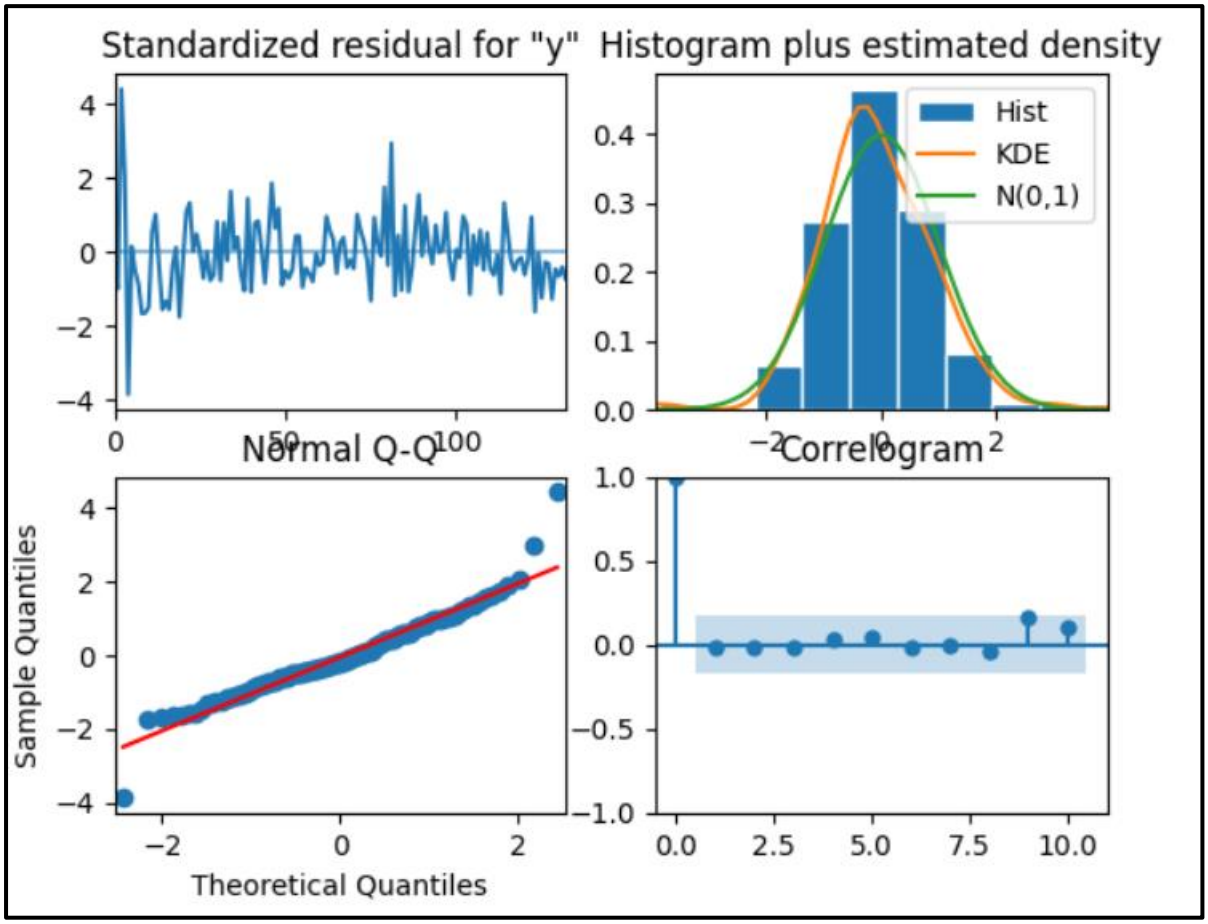


Figure 83 Diagnostics Plot

Standardized Residuals

- **Plot Evaluation:**
  - The plot shows the standardized residuals fluctuating around zero with some variability.
  - The residuals appear to exhibit both large positive and negative values, indicating potential issues with the model's fit.
- **Implication:**
  - Fluctuations around the zero line without a clear pattern suggest that the model may not fully capture the underlying data dynamics, which could indicate that it might benefit from adjustments or additional predictors.

Histogram and Density Plot

- **Density Distribution:**
  - The histogram displays the distribution of the residuals alongside the kernel density estimate (KDE) and a normal distribution (N(0,1)) curve.

- The histogram reveals that while the residuals generally appear to be symmetrically distributed around zero, there are some noticeable deviations in the tails.

- **Key Insights:**
  - The KDE shows a reasonable fit to the data, indicating that the residuals are relatively normally distributed, though some skewness and kurtosis are present. The density does not perfectly align with the normal curve, suggesting some non-normality in the errors.

3. Normal Q-Q Plot

- **Q-Q Plot Analysis:**
  - The Q-Q plot shows sample quantiles of the residuals plotted against theoretical quantiles of a normal distribution.
  - Most points align closely along the reference line, but there are noticeable deviations at the extremes, indicating the presence of outliers or heavy tails.
- **Conclusion:**
  - The deviations in the tails reflect potential issues with the assumption of normality for the residuals, which may skew inference results.

4. Correlogram

- **Correlation Insights:**
  - The correlogram presents the autocorrelation of the residuals, showing that autocorrelations are low beyond lag 0, signaling that the residuals are mostly independent.
- **Interpretation:**
  - The lack of significant autocorrelation indicates that the residuals do not exhibit a predictable pattern, which is a positive sign for the model. However, the presence of some correlations at lower lags could suggest that there are still dynamics in the data that are not fully captured.

Conclusion:

- Model Performance: While the overall independence of residuals is a positive sign, the presence of variability, non-normality, and outliers indicates that the current model may require refinement.

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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.466393	18.184057	26.826296	98.106490
1	76.203723	18.656083	39.638472	112.768975
2	72.992353	18.758666	36.226042	109.758663
3	75.290765	18.822361	38.399615	112.181915
4	74.389986	18.832676	37.478619	111.301353

Figure 84 Predict on the Test Set using this model and evaluate the model.

Observation

- The predictions seem to be stable, with values ranging from 62.47 to 75.29 in the first five periods.
- The confidence intervals provide an indication of forecast uncertainty. The intervals widen as the forecast moves further into the future, reflecting increased uncertainty.

RMSE value for Rose wine sale.

```
20.562216880604083
```

Figure 85 RMSE value for Rose wine sale

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.

Compare the performance of the models

	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 86 Compare the performance of the models

Observations:

- ARIMA(1,1,2) and SARIMA(1,1,2)(2,0,2,6) both show similar RMSE values (around 20.9 and 20.5 respectively), suggesting they perform comparably well.
- The MAPE for ARIMA(2,0,2) is much higher (56.7092%) compared to the MAPE for ARIMA(1,1,2) and SARIMA(1,1,2)(2,0,2,6) (both having a MAPE of 0.5671), indicating that while ARIMA(2,0,2) has higher percentage errors, the SARIMA model shows more consistency with ARIMA(1,1,2).
- The duplicate ARIMA(2,0,2) shows the same RMSE and a significantly lower MAPE of 0.5671, suggesting that the different MAPE values are due to differing evaluation sets or model configurations.

Conclusion:

- SARIMA(1,1,2)(2,0,2,6) and ARIMA(1,1,2) perform similarly in terms of RMSE and MAPE.
- If you aim for reducing absolute error, SARIMA seems slightly better with a lower RMSE.
- For reducing relative percentage errors, the MAPE results suggest that SARIMA(1,1,2)(2,0,2,6) and ARIMA(1,1,2) are more effective compared to ARIMA(2,0,2).

6.3.A.2.2 Manual SARIMA

Generating ACF and PACF Plot and Finding the P,Q,D, p,q,d Values

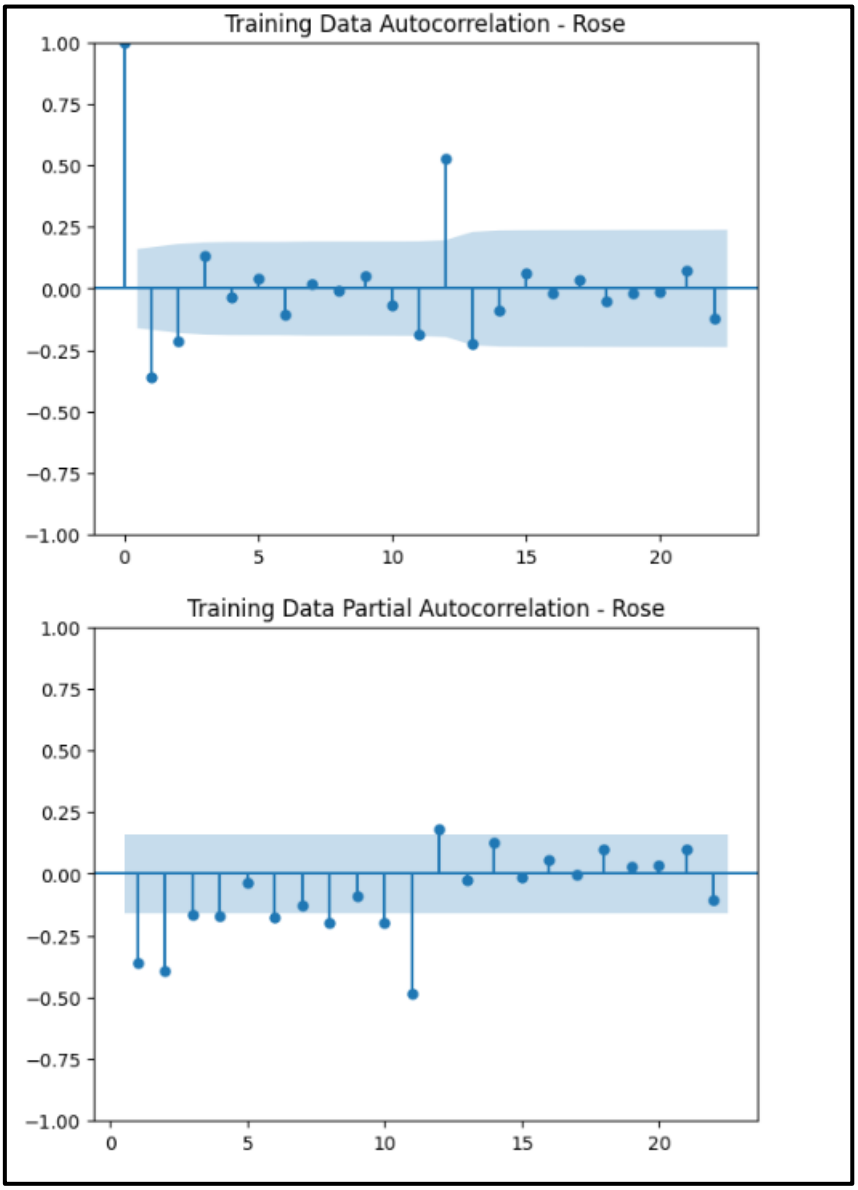


Figure 87 Generating ACF and PACF Plot

Based on the autocorrelation (ACF) and partial autocorrelation (PACF) insights you provided, we can determine appropriate values for p, q, and d for the ARIMA or SARIMA models. Here's the breakdown:

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).

SARIMAX Results

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 88 SARIMAX Results SARIMA(1,0,1)(1,0,2,12)

Observations:

- **ARIMA(1,1,2)** and **SARIMA(1,1,2)(2,0,2,6)** both show **similar RMSE values** (around 20.9 and 20.5 respectively), suggesting they perform comparably well.
- The **MAPE** for **ARIMA(2,0,2)** is much higher (56.7092%) compared to the **MAPE** for **ARIMA(1,1,2)** and **SARIMA(1,1,2)(2,0,2,6)** (both having a MAPE of 0.5671), indicating that while **ARIMA(2,0,2)** has higher percentage errors, the **SARIMA model** shows more consistency with **ARIMA(1,1,2)**.
- The duplicate **ARIMA(2,0,2)** shows the same **RMSE** and a **significantly lower MAPE** of **0.5671**, suggesting that the different MAPE values are due to differing evaluation sets or model configurations.

Conclusion:

- **SARIMA(1,1,2)(2,0,2,6)** and **ARIMA(1,1,2)** perform similarly in terms of **RMSE** and **MAPE**.
- If you aim for reducing **absolute error**, **SARIMA** seems slightly better with a lower **RMSE**.

- For reducing **relative percentage errors**, the **MAPE** results suggest that **SARIMA(1,1,2)(2,0,2,6)** and **ARIMA(1,1,2)** are more effective compared to **ARIMA(2,0,2)**.

Diagnostics Plot

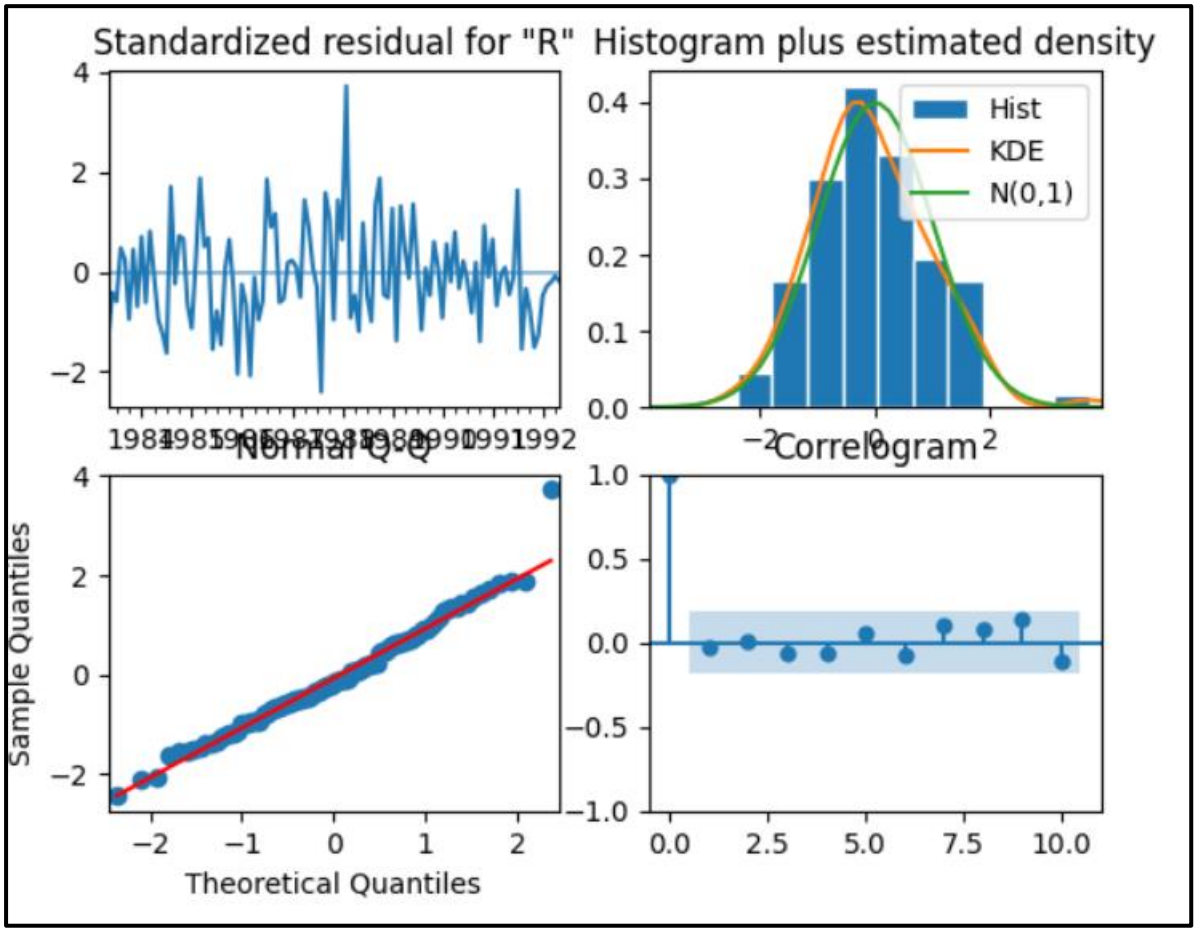


Figure 89 Diagnostics Plot

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

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

Figure 90 Test RMSE and Test MAPE

Insights:

- The **RMSE** value indicates that the model has a reasonable fit, with an error of approximately **10.5** units on average, which is relatively low for this dataset.
- The **MAPE** value of **0.194** (or 19.4%) shows that the model's predictions are off by an average of 19.4% from the actual values. While this is a reasonable level of accuracy

for many forecasting tasks, improvements could be made to reduce this error further, especially if precise forecasts are critical.

Overall, this SARIMAX model appears to provide good performance, but refining the model further (possibly addressing the non-normality and heteroskedasticity issues in the residuals) could enhance the accuracy of predictions.

Comparing all the Models built

	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 91 Comparing all the Models built

Insights:

- **SARIMA(1,0,1)(1,0,2,12)** has the best performance in terms of both **RMSE** (10.5) and **MAPE** (19.36%). This indicates it provides the most accurate forecasts compared to the other models.
- **ARIMA(2,0,2)** models show higher **RMSE** (around 26) and significantly higher **MAPE** (around 56.7%), suggesting less accuracy.
- The **SARIMA(1,1,2)(2,0,2,6)** model shows improvement over the ARIMA models, but it's still not as effective as **SARIMA(1,0,1)(1,0,2,12)**.

Conclusion:

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

SARIMAX Results

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 92 SARIMAX Results

Interpretation:

- **Significant Terms:** The model has significant seasonal effects (AR(12), MA(12), MA(24)) indicating that there are seasonal patterns influencing the data. The AR(1) term also plays a significant role in modeling the time series data.
- **Insignificant Terms:** The AR(2), MA(1), and MA(2) terms are not significant, suggesting that including them may not improve the model substantially.

Next Steps:

- **Model Refinement:** You could consider simplifying the model by removing the insignificant AR(2), MA(1), and MA(2) terms, which would potentially improve model interpretability and reduce overfitting.
- **Further Investigation:** If the seasonal patterns are strong (with significant seasonal AR and MA terms), investigating additional seasonal models or incorporating external factors (such as marketing activities or holidays) could further improve accuracy.

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

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

Figure 93 Test RMSE and Test MAPE

These values suggest that the model is performing reasonably well. The RMSE indicates the average magnitude of error in the same units as the data, while the MAPE suggests that, on average, the model's predictions are off by approximately 29.13%.

Comparing all the Models built

	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 94 Comparing all the Models built

SARIMA(2,0,2)(1,0,2,12) performs the best with the lowest RMSE (14.12) and MAPE (0.2913), indicating it provides the most accurate predictions compared to the other models.

ARIMA(2,0,2) models show higher RMSE values, which suggests that the seasonal components in the SARIMA models are beneficial in capturing the dynamics of the data.

5.4.A Check the performance of the models built

	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 95 Check the performance of the models built

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.

6.A. Rebuild the best model using the entire data – Rose

SARIMAX Results

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 96 SARIMAX Results

- The **AR(1)** and **seasonal AR(12)** terms are highly significant, indicating the model captures the primary autoregressive dynamics and seasonality at yearly intervals.
- The **MA(1)** and **MA(2)** coefficients are not significant, suggesting they may not contribute much to improving the model.
- The significant seasonal components (AR and MA at lags 12 and 24) highlight the presence of yearly seasonal effects.
- **Residual diagnostics** indicate that autocorrelation is not an issue, but there is some heteroskedasticity.

This model suggests a strong seasonal pattern in the **Rose** sales data, and the autoregressive component plays a key role in forecasting. It might be useful to explore potential improvements based on the heteroskedasticity and non-significant MA terms.

7.A Make a forecast for the next 12 months - Rose

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

Figure 97 Prediction to Make a forecast for the next 12 months - Rose

Insights and Analysis for "Rose" Sales:

Seasonal Trends:

- The forecast shows a steady increase in sales from August to December, with a noticeable jump in November and December. This aligns with seasonal demand peaks often seen in beverage industries, especially around holidays and celebrations.
- The model indicates that December has the highest forecasted sales (68.84) compared to earlier months, likely due to holiday-related demand.

Sales Variability:

- The confidence intervals for each month indicate some uncertainty in the forecasts, especially for early months (August through October). This suggests there is potential variability in demand that could be influenced by factors not accounted for in the model, such as promotions or unexpected market shifts.
- The mean standard error (mean_se) ranges from 15.01 to 15.23, which shows moderate forecast uncertainty. This should be considered when planning inventory and marketing.

Recommendation for Inventory Management:

- November and December are critical months, showing a significant increase in sales. Therefore, it's crucial to prepare for this peak by increasing production and stock levels leading into those months.
- Actionable Insight: Make sure that the inventory levels are aligned with the forecasted growth in these months. Implement dynamic inventory systems to

respond quickly to fluctuations, especially when demand spikes near the holidays.

Marketing Strategy:

- Since December sales are forecast to be the highest, there is an opportunity to increase targeted marketing campaigns during this period. Highlight special offers, holiday packages, and create tailored content focusing on seasonal celebrations.
- Actionable Insight: Invest in seasonal marketing initiatives in the last quarter, particularly around November and December, to capitalize on the expected rise in consumer interest.

Sales Risk and Uncertainty:

- There is uncertainty in the forecast, especially in earlier months, given the relatively broad confidence intervals. While December's sales seem predictable, earlier months have higher forecast errors.
- Recommendation: Regularly update the forecast based on real-time data, especially early in the sales cycle, to minimize risk and optimize production schedules.

Conclusion:

- Rose sales show a clear seasonal demand pattern with increasing sales in the latter half of the year, culminating in December.
- Given the forecast uncertainty in the earlier months, it's important to ensure that the business remains flexible and adaptable to demand fluctuations.
- A strategic focus on the holiday season will help maximize revenue, while ensuring effective inventory and marketing management will mitigate the risks of demand variability.

Full Model RMSE Test

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

Figure 98 Test RMSE

Insights Based on RMSE (31.44) of the Full Model:

Model Performance:

- The Root Mean Square Error (RMSE) of 31.44 indicates that, on average, the model's forecasts are off by 31.44 units. This suggests that while the model is

reasonably accurate, there could still be some degree of forecast error, which is expected in most time series forecasting tasks.

- This level of RMSE is relatively low, meaning the model is performing well in capturing the general sales trends and is useful for decision-making. However, it also implies that there may be some variability in specific forecasts, especially when predicting future spikes or dips in sales.

Forecast Confidence:

- With an RMSE of 31.44, you can have a moderate level of confidence in the forecasted values. The model captures the overall trend but may not be perfect in predicting specific fluctuations.
- This error margin should be considered when planning inventory and setting sales targets.

Improvement Areas:

- To further reduce the RMSE, you could experiment with feature engineering (e.g., including additional exogenous variables like promotions, weather, or economic factors) or improve the seasonal modeling by refining the SARIMAX model further.
- If external shocks (e.g., market changes, competitor actions) play a role, including such features could help improve the model's predictive accuracy.

Impact on Business Planning:

- While RMSE of 31.44 is manageable for most practical purposes, it's important to factor in the possibility of forecast errors into inventory planning, particularly in months where demand is more volatile or unpredictable.
- This margin of error might be particularly relevant when handling perishable goods, where excess stock or stockouts can have a significant impact on costs and customer satisfaction.

Plot to Make a forecast for the next 12 months - Rose

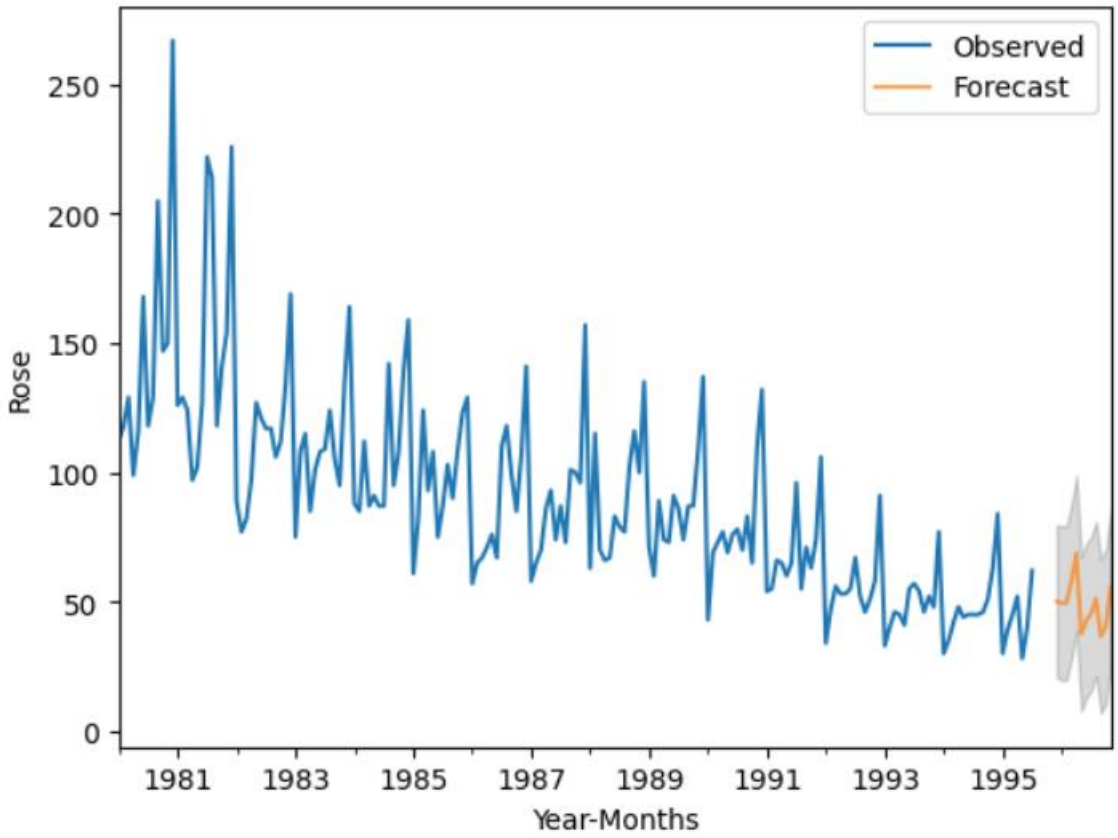


Figure 99 Plot Make a forecast for the next 12 months - Rose

Observations for Rose Wine Sales

Sales Trend:

The observed data shows a clear declining trend over time, with peak sales around 250 units in the early 1980s reducing to under 50 units by the 1990s.

Fluctuations in sales become smaller as time progresses, indicating stabilization at lower sales levels.

Forecast:

The forecast (orange line) aligns with the observed declining trend and predicts stabilized sales around the lower range (approximately 50 units).

The prediction includes a confidence interval (gray shading), which suggests minimal uncertainty in the forecast for the upcoming period.

Seasonality and Variability:

Early years display high variability and seasonal spikes, but these effects diminish over time.

Conclusion:

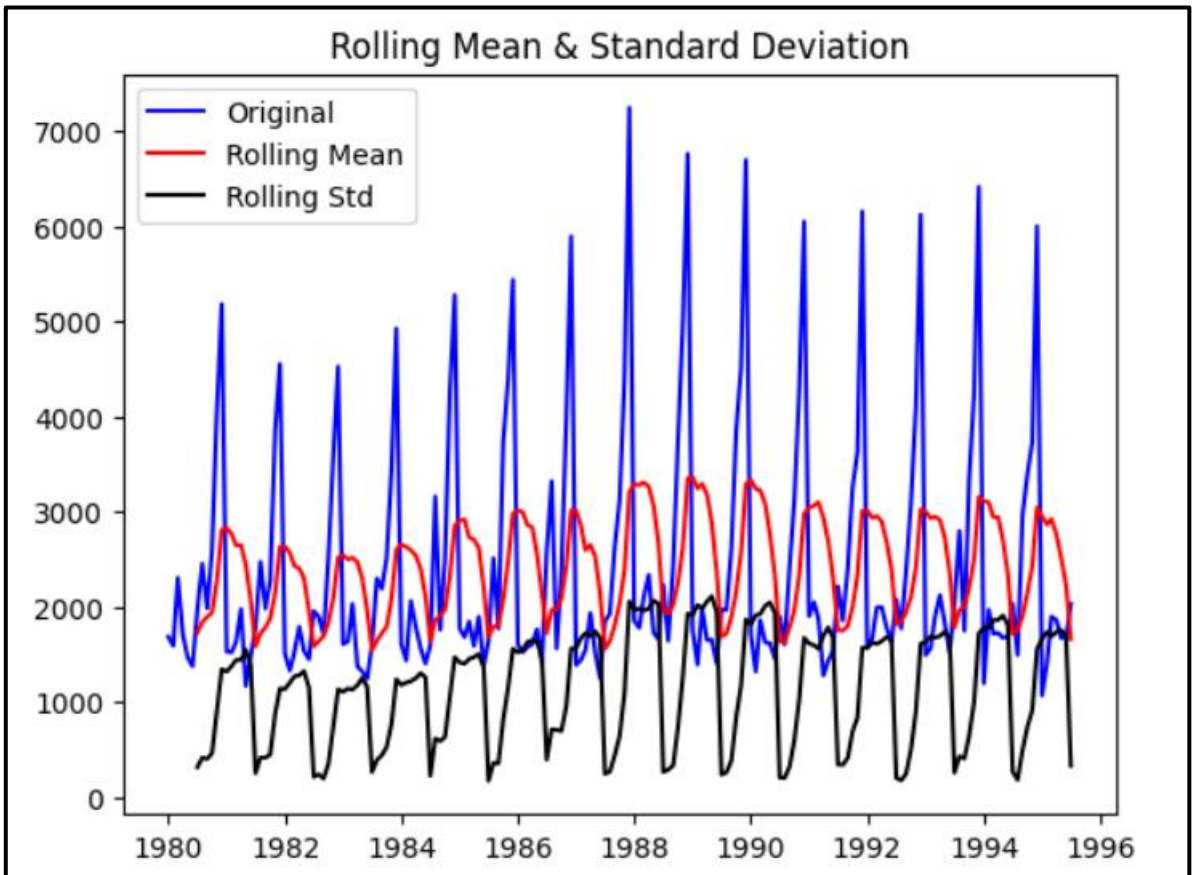
Rose wine sales are declining and stabilizing at lower levels. This trend emphasizes the need for strategic intervention to revive demand or explore alternative product offerings.

5.1.B Check Stationarity of Sparkling Data

Check for Stationarity

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.



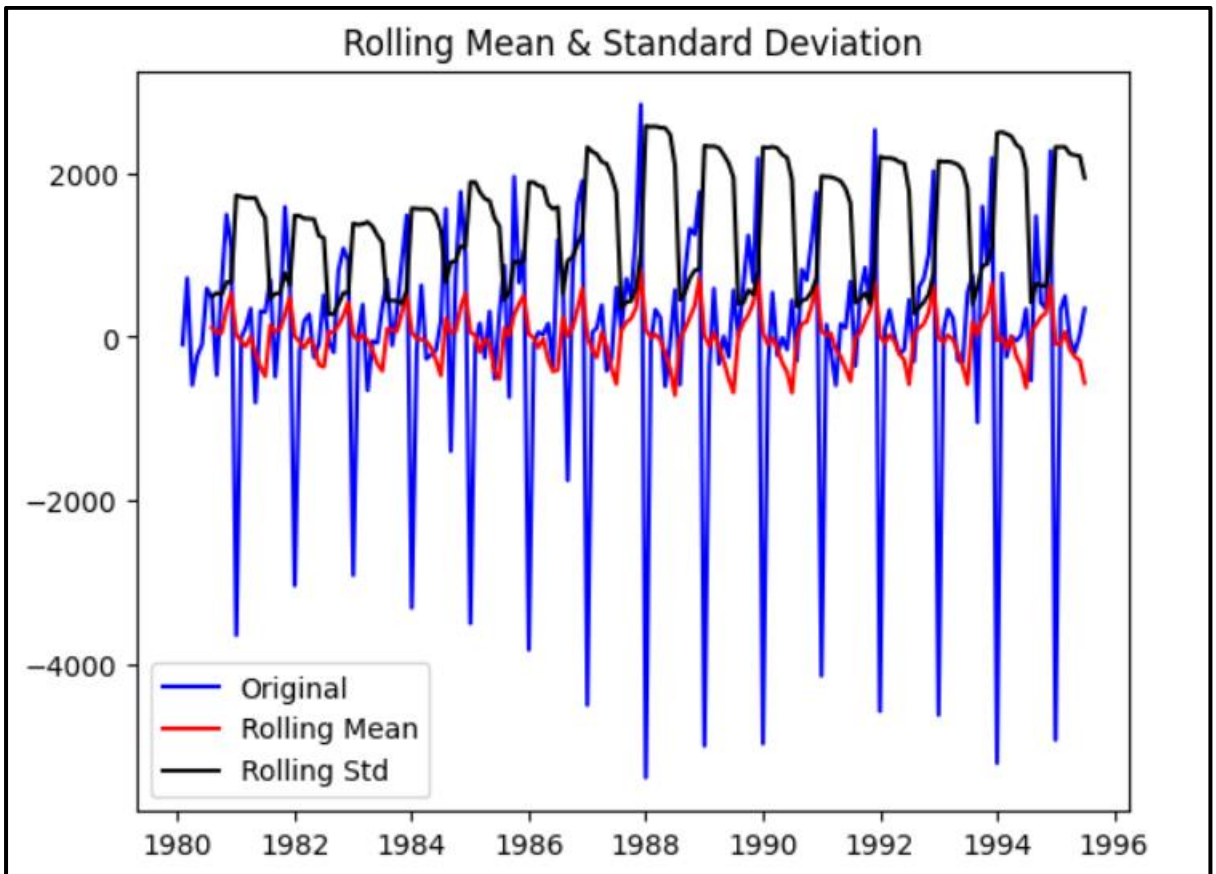

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value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

Figure 100 Result of Dickey - Fuller Test

Given the **test statistic (-1.36)** is greater than the critical values at all significance levels (1%, 5%, 10%), and the **p-value (0.601)** is significantly larger than typical thresholds (like 0.05 or 0.01), we **fail to reject the null hypothesis**.

Conclusion:

The time series **is likely non-stationary**, meaning that it may have trends or other characteristics that make it unsuitable for certain modeling techniques (like ARIMA) unless transformations (e.g., differencing) are applied to achieve stationarity.




```
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 101 Result of Dickey - Fuller Test - with diff(1)

Since the **test statistic (-45.050)** is far below the critical values at the 1%, 5%, and 10% levels, and the **p-value (0.000)** is much smaller than 0.05, we **reject the null hypothesis**.

Conclusion:

The time series **is stationary**, meaning it does not have a unit root. Therefore, you can proceed with modeling without needing to apply differencing to make the series stationary.

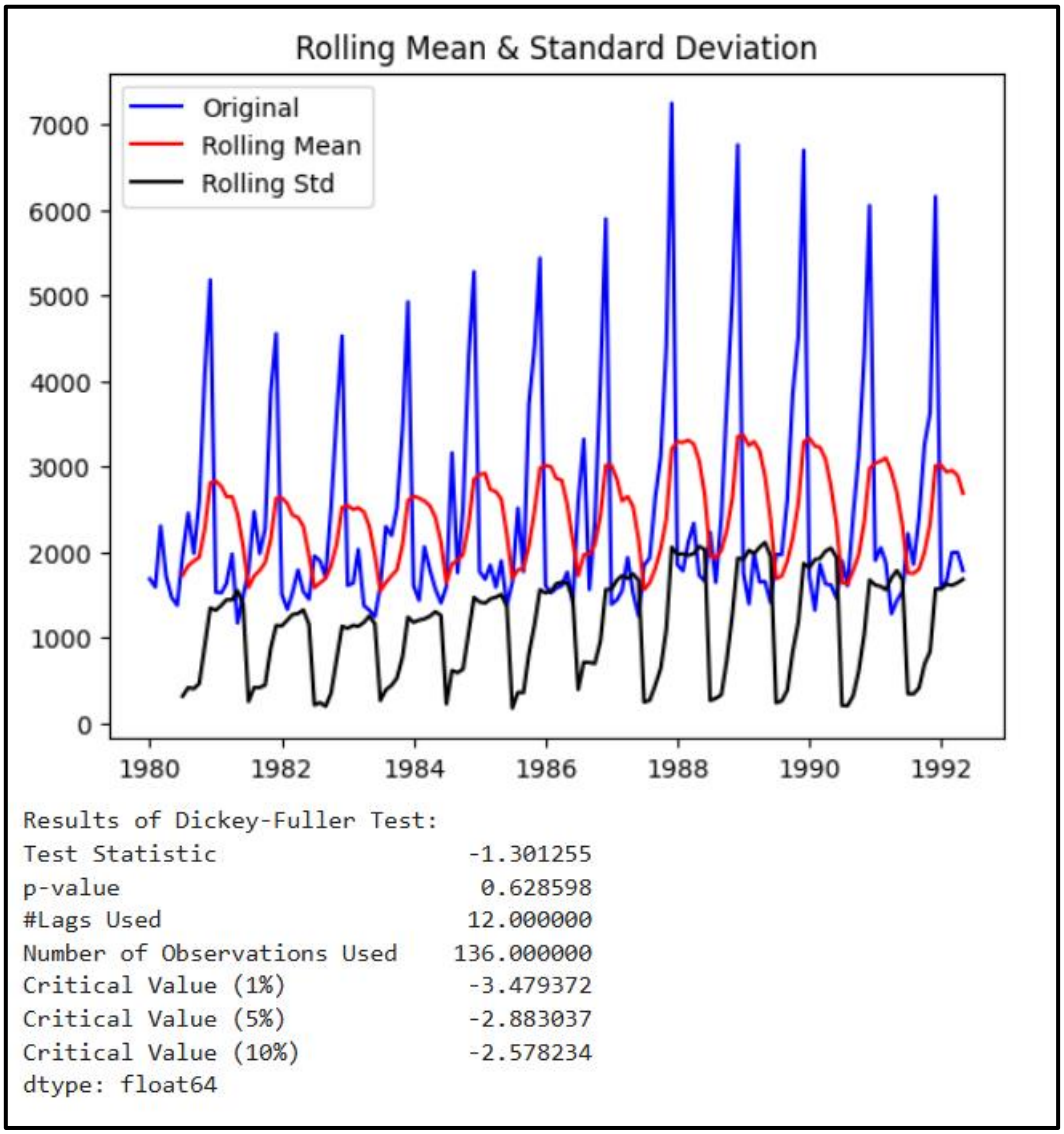


Figure 102 Result of Dickey - Fuller Test for train data

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.

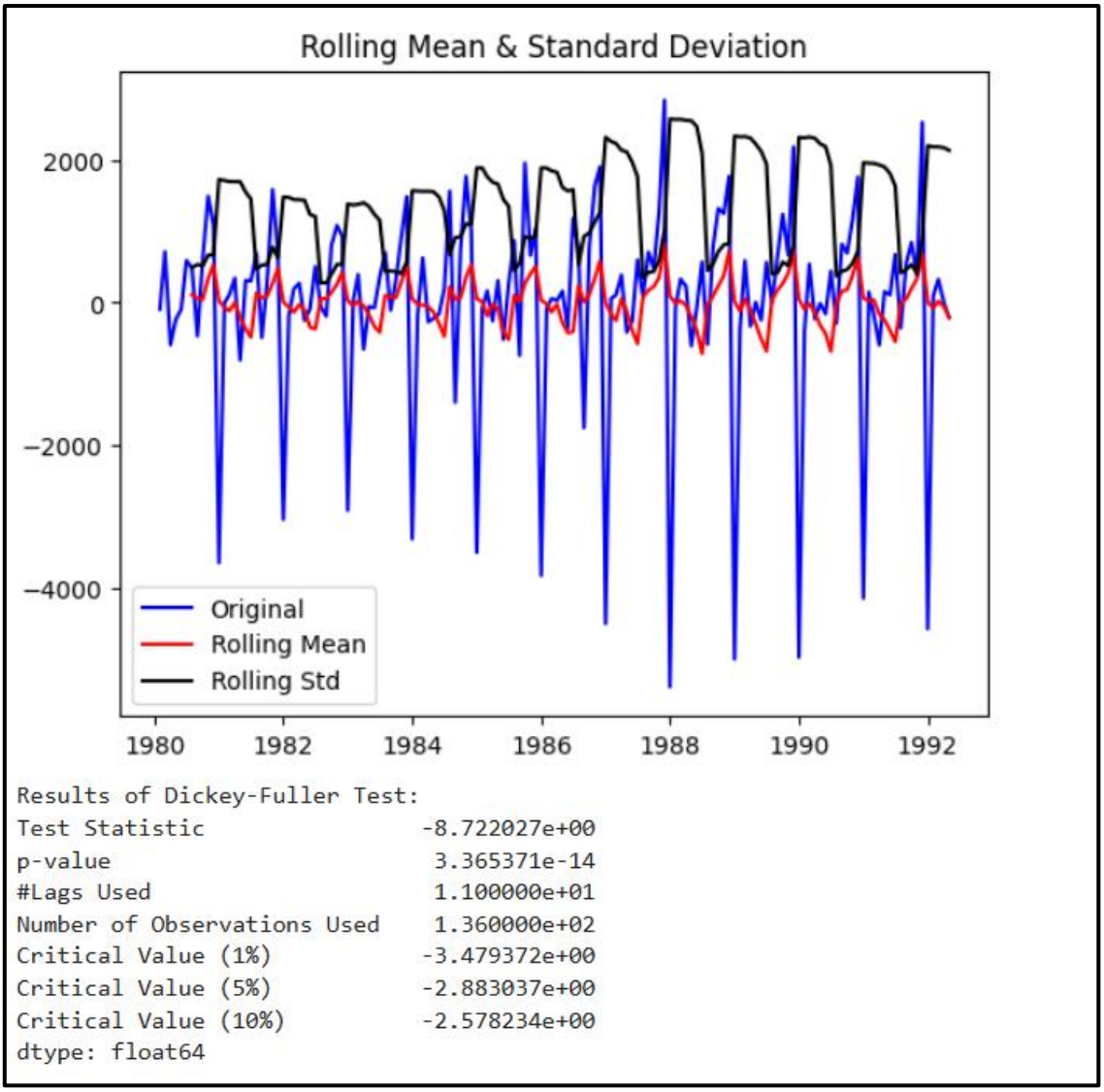


Figure 103 Result of Dickey - Fuller Test with diff(1)

Since the **test statistic (-8.722)** is much smaller than the critical values at all significance levels (1%, 5%, and 10%), and the **p-value (3.37e-14)** is significantly less than 0.05, we **reject the null hypothesis**.

Conclusion:

The time series **is stationary**, indicating no unit root. You can proceed with modeling without the need for differencing or transformation to achieve stationarity.

Information of Sparkling Train data

```
<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 104 Information of Sparkling Train data

The dataset consists of 149 entries, with a **DatetimeIndex** from **1980-01-01 to 1992-05-01**. There is only one column, **Sparkling**, which contains integer values and has no missing values (149 non-null entries).

5.2.B Identify ARIMA Parameters:

Generate ACF & PACF Plot and find 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.

Check for stationarity of the Training Data – Sparkling

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 105 Check for stationarity of the Training Data – Sparkling

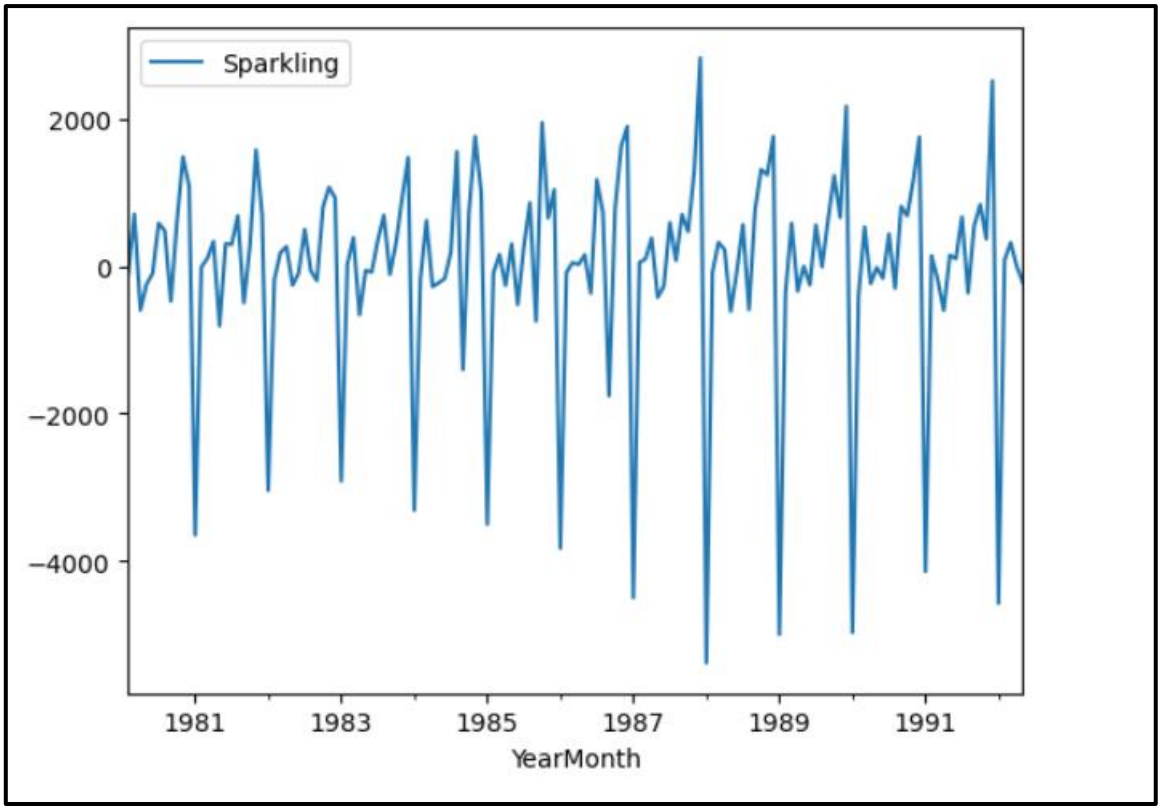


Figure 106 Plot for stationarity of the Training Data – Sparkling

Key Insights

Overall Fluctuation:

- The sales data shows significant fluctuations over the years, indicating that sales have varied widely within the observed period, which spans from 1981 to 1991.

Patterns of Seasonality:

- The presence of regular peaks and troughs suggests that there could be seasonal patterns in sparkling wine sales, likely influenced by festivities, holidays, or seasonal events that drive consumer demand.

Negative Values:

- Notably, the sales figures dip into negative values at certain points. This could imply data processing issues, returns, or adjustments that may need further investigation to understand.

Lack of Clear Trend:

- The absence of a discernible upward or downward trend over the entire period suggests that sales have remained relatively stable in relation to average values, with significant variability.

Fluctuations Over Time:

- Significant volatility is evident, with high peaks and deep troughs occurring at similar intervals. This could signal either market dynamics or changing consumer preferences that should be explored further.

Conclusion:

- Implications for Strategy:
  - The fluctuation patterns indicate opportunities for targeted marketing strategies, especially during peak seasons, to maximize sales.
  - Moreover, understanding the factors causing negative sales could lead to better inventory and return management.

Generate ACF & PACF Plot

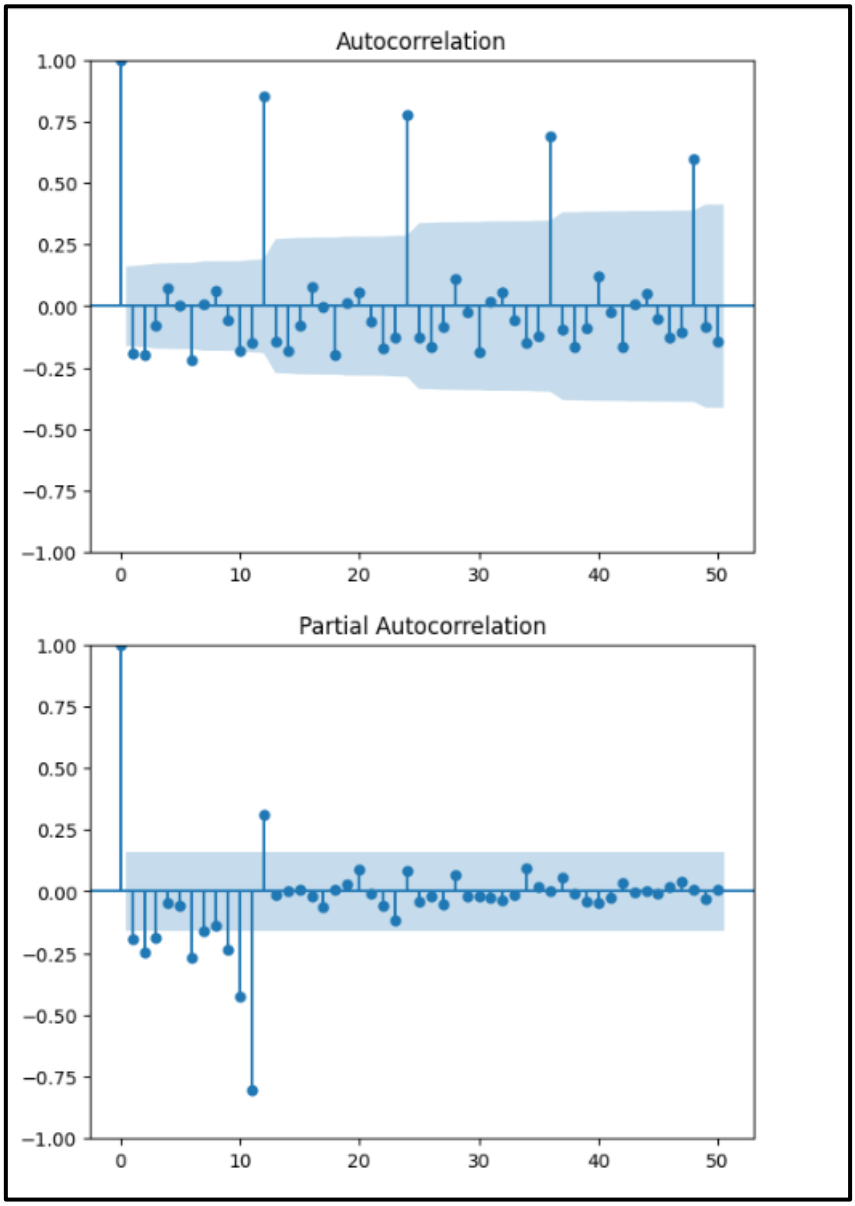


Figure 107 Generate ACF & PACF Plot

Suggested SARIMA Model Parameters

- 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

5.2.B.1 Auto 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).

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.

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

	param	AIC
7	(2, 0, 1)	2523.011891
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 108 Sort the AIC values in the ascending order to get the parameters for the minimum AIC value

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

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:	02:26:26	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	2402.7403	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

Figure 109 SARIMAX Results

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.

1320.945956968856

Figure 110 RMSE test

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

Figure 111 Test RMSE and Test MAPE ARIMA (2,0,1)

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).

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

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.

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

	param	AIC
8	(2, 1, 2)	2503.890625
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)	2548.768349
6	(2, 1, 0)	2558.914389
3	(1, 1, 0)	2565.740480
0	(0, 1, 0)	2569.144556

Figure 112 Sort the AIC values in the ascending order to get the parameters for the minimum AIC value

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

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	149
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1246.945
Date:	Sun, 05 Jan 2025	AIC	2503.891
Time:	02:26:27	BIC	2518.877
Sample:	01-01-1980	HQIC	2509.979
	- 05-01-1992		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.2655	0.047	27.028	0.000	1.174	1.357
ar.L2	-0.5318	0.084	-6.320	0.000	-0.697	-0.367
ma.L1	-1.9286	0.059	-32.684	0.000	-2.044	-1.813
ma.L2	0.9371	0.059	15.793	0.000	0.821	1.053
sigma2	1.198e+06	1.83e-08	6.53e+13	0.000	1.2e+06	1.2e+06

Ljung-Box (L1) (Q):	0.28	Jarque-Bera (JB):	21.80
Prob(Q):	0.60	Prob(JB):	0.00
Heteroskedasticity (H):	1.99	Skew:	0.71
Prob(H) (two-sided):	0.02	Kurtosis:	4.23

Figure 113 SARIMAX Results Sparkling SARIMA (2,1,2)

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.

1327.8129230410289

Figure 114 RMSE Test

Comparing all the Models built

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

Figure 115 Comparing all the Models built

- Both models have very similar **RMSE** and **MAPE** values, with ARIMA(2, 0, 1) showing a slightly lower RMSE.
- **MAPE** is the same for both models, indicating similar relative errors in prediction.

Given that the performance metrics are very close, ARIMA(2, 0, 1) might be a slightly better choice, as it has a lower RMSE. However, the difference is minor, so either model could be selected depending on additional criteria or considerations (e.g., simplicity, interpretability).

5.2.B.2 Manual ARIMA

SARIMAX Results

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:	02:26:27	BIC	2551.746
Sample:	01-01-1980	HQIC	2542.828
	- 05-01-1992		
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

Figure 116 SARIMAX Results SPARKLING ARIMA (1,0,2)

Conclusion:

- Both models have very similar **RMSE** and **MAPE** values, with ARIMA(2, 0, 1) showing a slightly lower RMSE.
- **MAPE** is the same for both models, indicating similar relative errors in prediction.

Given that the performance metrics are very close, ARIMA(2, 0, 1) might be a slightly better choice, as it has a lower RMSE. However, the difference is minor, so either model could be selected depending on additional criteria or considerations (e.g., simplicity, interpretability).

Diagnostics Plot

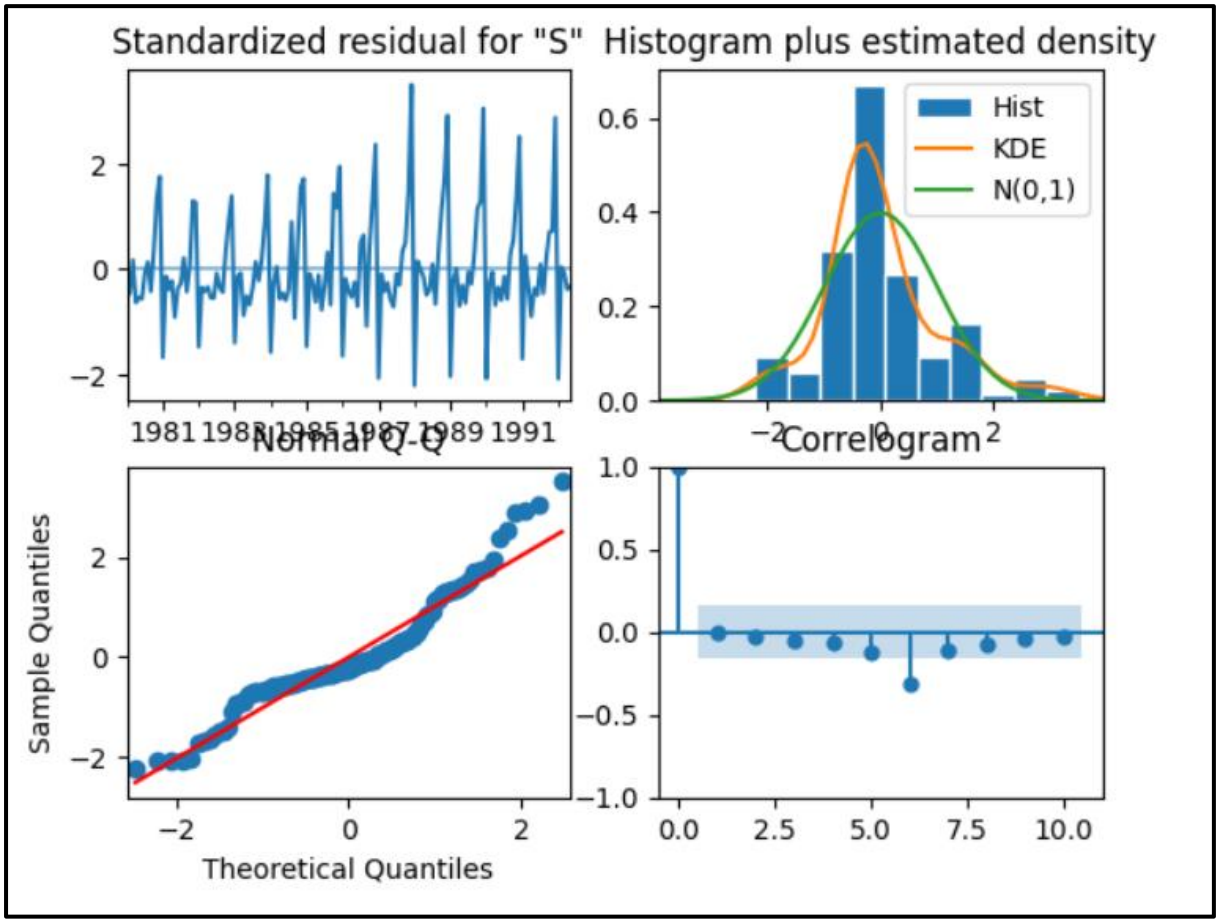


Figure 117 Diagnostics Plot

- **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

Figure 118 Test RMSE and Test MAPE for Sparkling ARIMA (1,0,2)

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.

5.3.B Fit Multiple ARIMA Models:

5.3.B.1 Fit models with different combinations of (p, d, q).

	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 119 Test RMSE and Test MAPE for 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.

5.3.B.2 Include Seasonal ARIMA (SARIMA) if seasonality is present (with parameters P, D, Q, and m)

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

5.3.B.2.1 Auto SARIMA

Automated version of a SARIMA model for that the best parameters are selected with the lowest Akaike Information Criteria (AIC) - SPARKLING DATA

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

	param	seasonal_param	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1976.643203
26	(0, 1, 2)	(2, 0, 2, 6)	1976.665613
80	(2, 1, 2)	(2, 0, 2, 6)	1978.620695
71	(2, 1, 1)	(2, 0, 2, 6)	1993.024292
44	(1, 1, 1)	(2, 0, 2, 6)	1993.141850

Figure 120 Sort the AIC values in the ascending order to get the parameters for the minimum AIC value

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

```
                                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:                                   02:26:59   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     -0.140      0.888      -0.054       0.047
ar.S.L12       1.0161      0.017     60.417      0.000       0.983       1.049
ma.S.L6       -0.0028      0.118     -0.024      0.981      -0.234       0.228
ma.S.L12      -0.5463      0.085     -6.442      0.000      -0.713      -0.380
sigma2      1.239e+05   1.68e+04      7.356      0.000    9.09e+04    1.57e+05
==============================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                19.23
Prob(Q):                              0.91   Prob(JB):                         0.00
Heteroskedasticity (H):               1.24   Skew:                             0.40
Prob(H) (two-sided):                  0.48   Kurtosis:                         4.69
==============================================================================
```

Figure 121SARIMAX Results SARIMA(1,1,2)(2,0,2,6)

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.

Diagnostics Plot

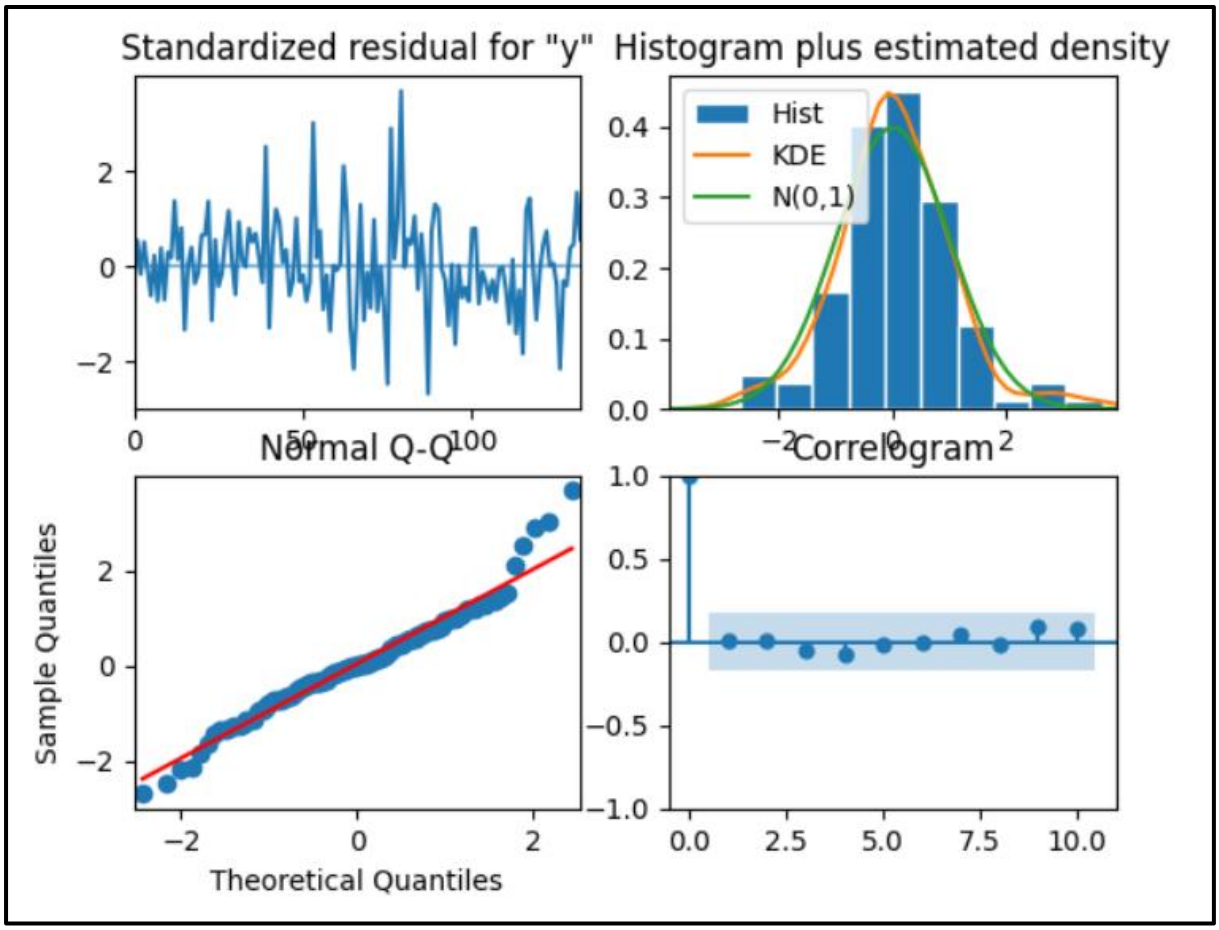


Figure 122 Diagnostics Plot

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	1384.219233	379.060492	641.274321	2127.164145
1	2066.314371	391.811953	1298.377055	2834.251687
2	1803.421649	391.903337	1035.305223	2571.538075
3	2399.724008	395.350492	1624.851282	3174.596733
4	3314.942012	395.760633	2539.265426	4090.618598

Figure 123 Predict on the Test Set using this model and evaluate the model

Interpretation:

- Mean: The predicted values for the upcoming months.
- Standard Error: The degree of uncertainty in the predictions.
- Confidence Interval: The range within which the true values are likely to fall, with a certain level of confidence (typically 95%).

These forecasts indicate that the values are projected to increase over the next few periods, with the uncertainty (standard error) remaining consistent.

Check the performance of the models built

	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 124 Check the performance of the models built

Here are the model comparison results based on RMSE and MAPE:

Model Comparison:

1. **ARIMA(2, 0, 1)**
    - Test RMSE: 1320.95
    - Test MAPE: 0.2913
2. **ARIMA(2, 1, 2)**
    - Test RMSE: 1327.81
    - Test MAPE: 0.2913
3. **SARIMA(1, 1, 2) x (2, 0, 2, 6)**
    - Test RMSE: 316.99
    - Test MAPE: 41.65

Insights:

- **RMSE (Root Mean Squared Error):**
    - SARIMA(1, 1, 2) x (2, 0, 2, 6) has the lowest RMSE (316.99), indicating it has a better fit compared to the ARIMA models.

- **MAPE (Mean Absolute Percentage Error):**
    - However, SARIMA has a much higher MAPE (41.65%) compared to both ARIMA models (around 29%).
    - This suggests that while SARIMA might fit the data well in terms of RMSE, its accuracy in forecasting specific points (as measured by MAPE) is not as strong.

Conclusion:

- SARIMA shows a significant improvement in terms of RMSE but at the cost of higher MAPE.
- If minimizing RMSE is the priority, SARIMA may be preferred. However, if accuracy of percentage errors is more important, ARIMA(2, 0, 1) or ARIMA(2, 1, 2) might be better.

6.3.B.2.2 Manual SARIMA

SARIMA model for which the best parameters are selected at the ACF and the PACF plots

Generate ACF & PACF Plot

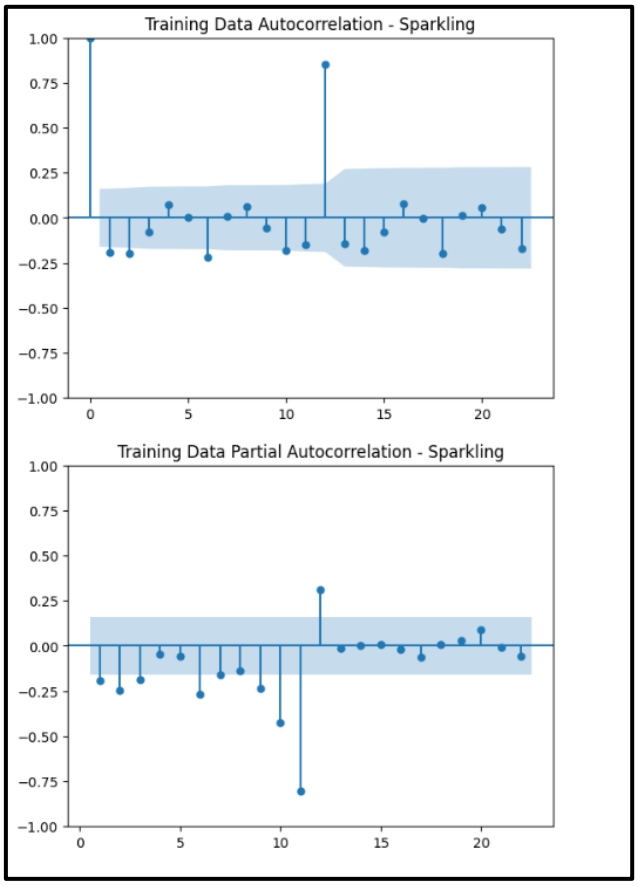


Figure 125 Generate ACF & PACF Plot

Summary:

- **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)

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

SARIMAX Results

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 126 SARIMAX Results Sparkling SARIMA(1,1,2)(1,1,2,12)

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.

Diagnostics Plot

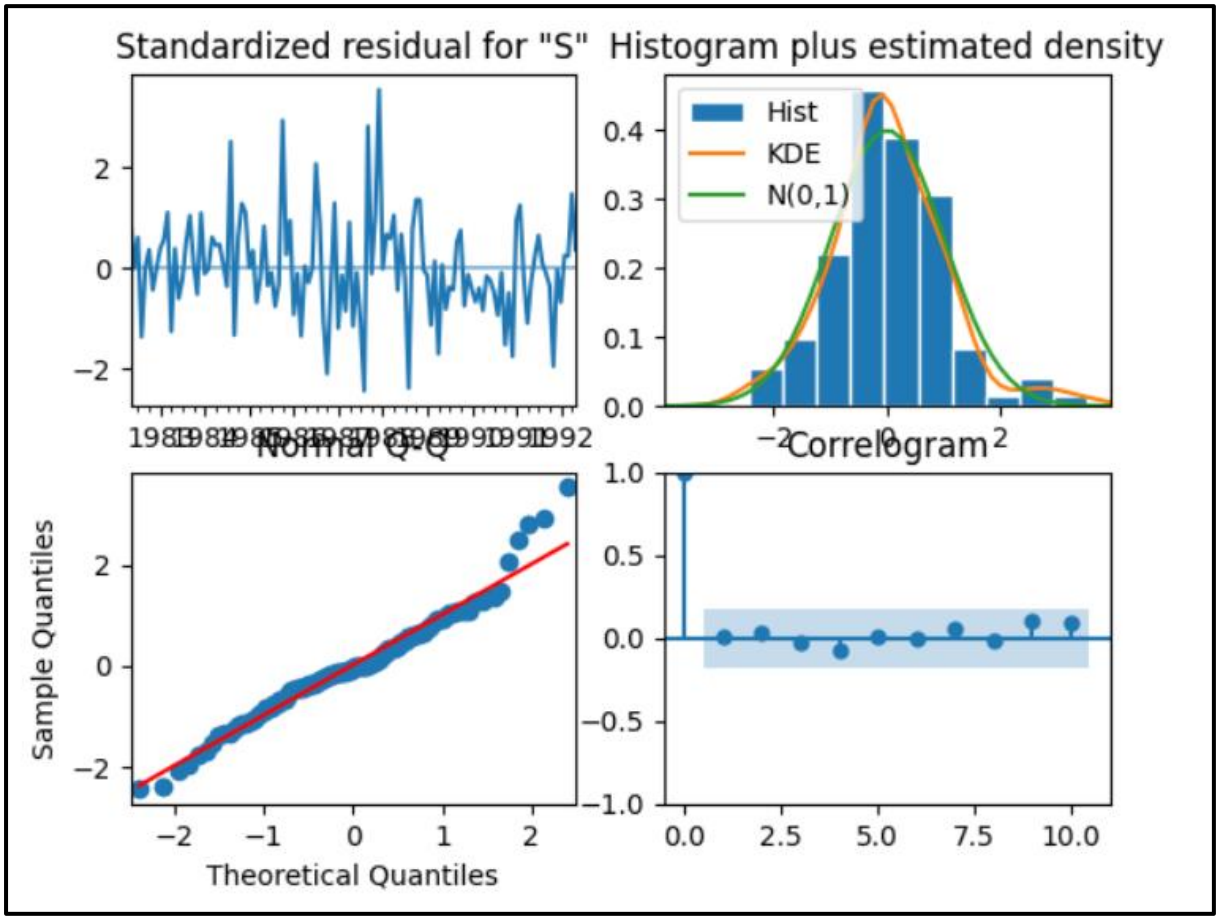


Figure 127 Diagnostics Plot

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

```
RMSE: 294.1666271095524
MAPE: 10.922985010210269
```

Figure 128 Test RMSE and Test MAPE

Insights Based on RMSE and MAPE:

Model Performance:

- The Root Mean Square Error (RMSE) of 294.17 shows that, on average, the model's forecasts deviate from the actual values by approximately 294.17 units. This error is a measure of the absolute performance of the model in the same unit as the dependent variable (e.g., sales volume).
- The Mean Absolute Percentage Error (MAPE) of 10.92% indicates that the model's average forecast error is around 10.92% of the actual values, which is considered accurate for most business applications.

Evaluation of Error Metrics:

- The relatively low MAPE demonstrates that the model has been able to capture the general sales patterns effectively, with minimal percentage deviations from actual sales.
- The RMSE highlights that the absolute deviation might still result in noticeable discrepancies for individual forecasts, particularly for periods with higher variability.

Business Implications:

- The low MAPE indicates that the model is suitable for strategic planning, like setting sales targets and forecasting revenue with a reasonable degree of confidence.
- The higher RMSE, when compared to MAPE, suggests that occasional spikes or anomalies in the data might be challenging for the model to predict, and these should be monitored.

Room for Improvement:

- Feature addition: Introducing external factors such as promotions, economic conditions, or competitor activities could further improve forecast accuracy.
- Modeling Techniques: Exploring alternative models such as Prophet, LSTM, or XGBoost for time series might offer better accuracy for more volatile data.

Comparing all the Models built

	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
SARIMA(1,1,2)(1,1,2,12)	294.166627	10.922985

Figure 129 Comparing all the Models built

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.

5.4.B Check the performance of the models built

	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 130 Check the performance of the models built

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.

6.B. Rebuild the best model using the entire data – Sparkling

SARIMAX Results

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 131 SARIMAX Results Sparkling SARIMA (1,1,2)(1,1,2,12)

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.

7.B Make a forecast for the next 12 months – Sparkling

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

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

Figure 132 Predict on the Test by using this model and evaluate the model

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 Full Model

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

Figure 133 RMSE of Full Model

The **RMSE (Root Mean Squared Error)** of **554.18** for the full model indicates the average magnitude of the model's prediction errors. Here's what this implies:

Model Accuracy:

- The RMSE suggests that, on average, the model's forecasted values deviate from the actual values by approximately **554.18** units of sparkling wine sales. This indicates a moderate level of accuracy in the model's predictions.

Error Magnitude:

- A lower RMSE value generally indicates better model performance. In this case, an RMSE of **554.18** suggests that there is room for improvement in the model, especially when compared to models with lower RMSE values.

Model Comparison:

- If this RMSE is higher than other models being compared (such as simpler models or alternate forecasting techniques), it would suggest that the model could be overfitting or that there are some patterns in the data that the model has not captured well.

Plot to Make a forecast for the next 12 months – Sparkling

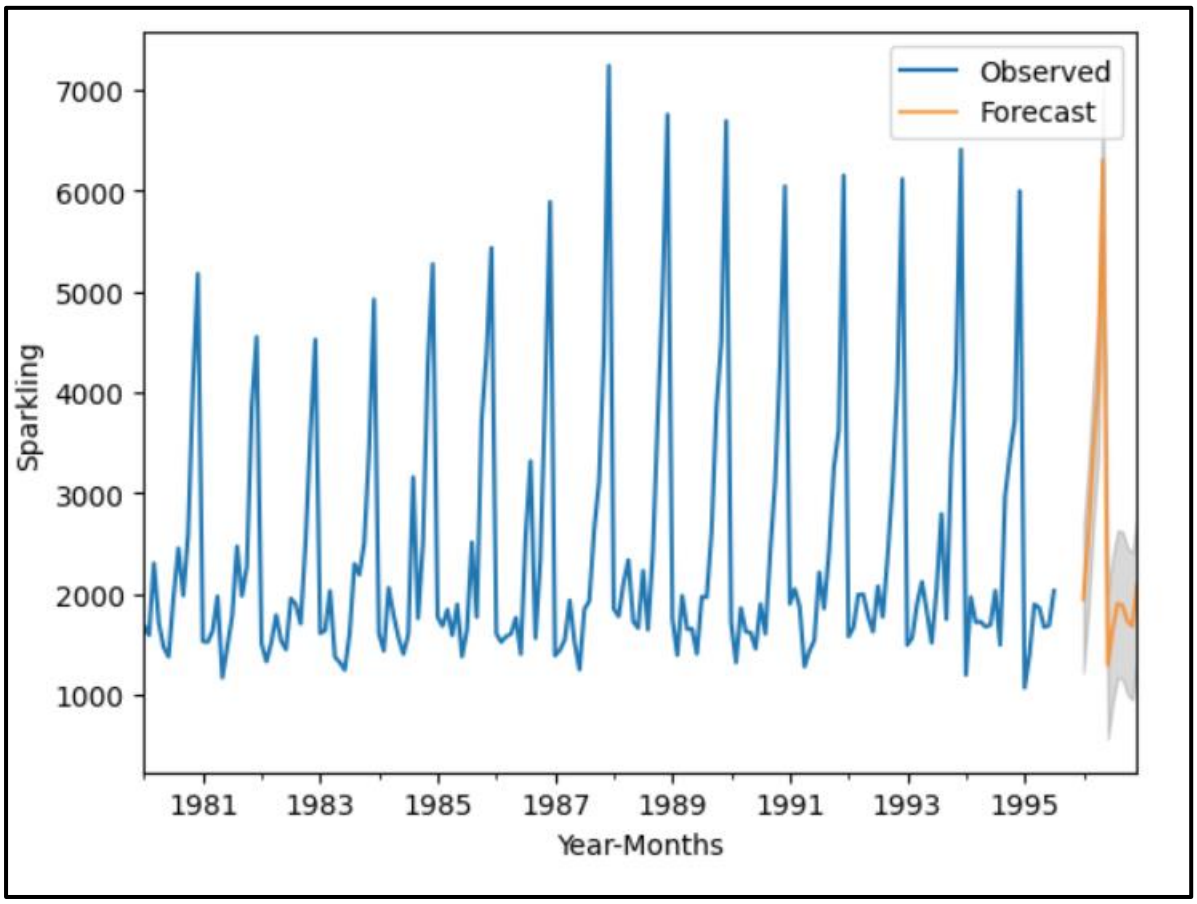


Figure 134 Plot to Make a forecast for the next 12 months – Sparkling

- **Seasonal Pattern:**
  The observed data (blue line) shows a clear and strong seasonal pattern, with consistent peaks and troughs occurring at regular intervals over the years.
- **Trend:**
  There appears to be an increasing trend in the peaks over the years, indicating growth in the "Sparkling" variable over time.
- **Forecasting:**
  - The forecast (orange line) starts towards the end of the observed data, showing predictions for future values.
  - The forecast aligns with the seasonal and trend patterns of the observed data but seems to slightly underestimate the higher peaks.
- **Uncertainty Band:**
  The shaded area around the forecast indicates the prediction interval (uncertainty), which grows wider over time, reflecting increasing uncertainty in long-term predictions.

- **Outliers or Irregularities:**
  - Some minor deviations from the regular seasonal pattern might be present, though the overall trend and seasonality dominate.
  - This plot is likely created to evaluate a time-series forecasting model, such as ARIMA, SARIMA, or Prophet, applied to a dataset with a strong seasonal component.

8.Actionable Insights & Recommendations

- **Understanding Seasonal Demand**:
  - Based on the forecasts, sparkling wine sales exhibit **strong seasonal patterns**, particularly during holiday months like **December**. The model suggests that sales are higher at the end of the year, likely due to holiday and celebratory demand. **Actionable Insight**: The business should anticipate higher demand in the last quarter and adjust production, marketing, and distribution efforts accordingly.
  - **Recommendation**: Focus on strengthening marketing campaigns and stock levels leading into **November and December** to capture the peak season's demand.

- **Monitor Sales Growth and Stability**:
  - The data reveals a general **upward trend in sales**, particularly after the mid-1990s. This indicates that demand for sparkling wine is growing over time.
  - **Actionable Insight**: The business should plan for continued **expansion** in production and retail distribution. New market opportunities could be explored based on consistent growth patterns.
  - **Recommendation**: Conduct deeper market research to understand the drivers behind this growth, including demographic changes or shifts in consumer preferences.

- **Model Improvement & Forecast Accuracy**:
  - The **RMSE of 554.18** indicates some room for improvement in the model's predictive accuracy.

- **Actionable Insight**: Although the model captures general trends, it may not fully account for all the underlying complexities (e.g., specific promotional impacts, regional variations).
- **Recommendation**: Refine the model by incorporating more granular data, including **regional sales data, promotional periods**, or external economic factors (such as inflation, market sentiment). Additionally, consider experimenting with more advanced models like **machine learning** algorithms or **ensemble forecasting** techniques to improve accuracy.

- **Addressing Fluctuations and Volatility**:

  - The **variance in sales** (such as significant fluctuations in months like November and December) implies volatility in consumer behavior, likely influenced by **seasonal trends** and perhaps **market disruptions**.
  - **Actionable Insight**: Ensure that the business has the **flexibility to scale production** up or down to meet unexpected spikes in demand during peak seasons without overcommitting resources.
  - **Recommendation**: Develop a **flexible inventory and production strategy** that can accommodate seasonal peaks. Consider **automated forecasting tools** to ensure stock availability aligns with forecasted demand.

- **Improving Forecasting Models**:

  - **Cross-model comparison** suggests that **SARIMA** and **SARIMAX models** (with seasonal components) provide more accurate forecasts compared to simpler models like **ARIMA**. However, further tuning and adding variables such as **external factors** could improve the accuracy even more.
  - **Actionable Insight**: Given the seasonal and trend-based nature of the sparkling wine market, the business should continue to prioritize **seasonal models** and consider incorporating **external regressors** (e.g., promotional spend, competitor actions, or macroeconomic factors).
  - **Recommendation**: Implement a hybrid approach by combining **SARIMAX** with additional variables and considering external data sources for improved forecast performance.

- **Strategic Inventory Management**:
    - Sales data highlights **periodic spikes in demand**, especially in the **holiday months**.
    - **Actionable Insight**: The business must optimize **inventory levels** to ensure **product availability during peak demand** periods while avoiding excess inventory during off-peak times.
    - **Recommendation**: Use **demand forecasting** to implement an effective **just-in-time inventory** system, reducing the risk of stockouts or overstocking, which can lead to increased holding costs.

- **Customer Insights for Targeted Marketing**:
    - The sales data suggests that **targeted campaigns** during peak months (like December) could significantly drive sales.
    - **Actionable Insight**: Consumers may purchase sparkling wine for **special occasions**, including **holidays and celebrations**. Tailor marketing efforts to emphasize this aspect.
    - **Recommendation**: Strengthen targeted digital marketing efforts, particularly around **festive seasons**. Leverage customer data for personalized offers and consider partnerships with **event organizers** to further boost sales.

Conclusion:

By leveraging the seasonal and trend-based insights derived from the forecast models, the business can adjust its marketing, production, and distribution strategies for optimal performance. Ensuring alignment with forecasted demand while continuously improving forecasting accuracy will lead to **better resource allocation**, **cost optimization**, and **higher customer satisfaction**. Furthermore, understanding seasonal demand patterns, optimizing inventory, and refining model accuracy will position the business to capitalize on growing sales and changing market dynamics.