TIME SERIES FORECASTING

PROJECT REPORT

Pavithra Devi
PGPDSBA
Great Learning

INDEX

Problem:

To analyse and forecast ABC Estate Wine Sales in the 20th century.
Data set for the Problem: Sparkling.csv and Rose.csv

[For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.]

The analysis of wine sales for ABC Estate Wines holds significant importance for the company for several reasons:

1. Strategic Planning:
   - Understanding the historical trends and forecasting future sales enables ABC Estate Wines to engage in strategic planning. This includes inventory management, production planning, and resource allocation based on anticipated demand.

2. Optimizing Resources:
   - Accurate sales forecasts help in optimizing resources such as raw materials, production capacities, and manpower. This ensures that the company operates efficiently and avoids both overproduction and stockouts.

3. Marketing and Promotions:
   - Insights gained from the analysis can guide marketing and promotional activities. Understanding which types of wines are more popular during specific periods allows the company to tailor marketing efforts to boost sales during those times.

4. Financial Planning:
   - Reliable sales forecasts are crucial for financial planning. ABC Estate Wines can use the forecasts to estimate revenues, allocate budgets, and make informed financial decisions.

5. Customer Satisfaction:
   - Anticipating demand helps in maintaining sufficient stock to meet customer needs. This, in turn, contributes to customer satisfaction by ensuring that popular wines are consistently available.

6. Competitive Advantage:
   - The ability to accurately forecast wine sales provides ABC Estate Wines with a competitive advantage. By staying ahead of market trends and customer preferences, the company can position itself effectively against competitors.

7. Risk Management:
   - Understanding historical sales patterns allows ABC Estate Wines to identify potential risks and challenges. This proactive approach enables the company to implement risk mitigation strategies and respond effectively to market fluctuations.

8. Long-Term Sustainability:
   - Building a robust forecasting model contributes to the long-term sustainability of the business. ABC Estate Wines can adapt to changing market conditions, reduce waste, and maintain a strong market presence over time.

In summary, the analysis and forecasting of wine sales provide ABC Estate Wines with actionable insights that extend beyond mere sales numbers. It empowers the company to make informed decisions, enhance operational efficiency, and maintain a competitive edge in the dynamic wine market.

Data Overview:

Data sets used for analysis: Rose.csv and Sparkling.csv
Types of wines :Rose and Sparkling wine
Time period covered: Monthly Sales data from 1980 to 1995.

Sparkling and Rose datasets have 187 rows and 1 column, indicating a single time series of sales data. The shape (187, 1) suggests that you have 187 observations and 1 feature (in this case, the sales data of the respective wines).

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

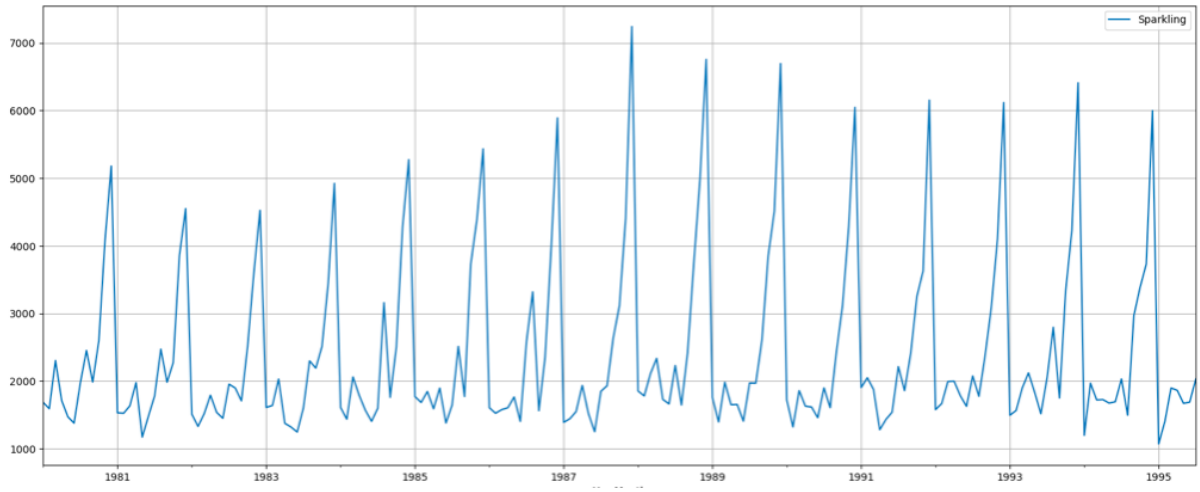
187 rows × 1 columns

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0
...	...
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

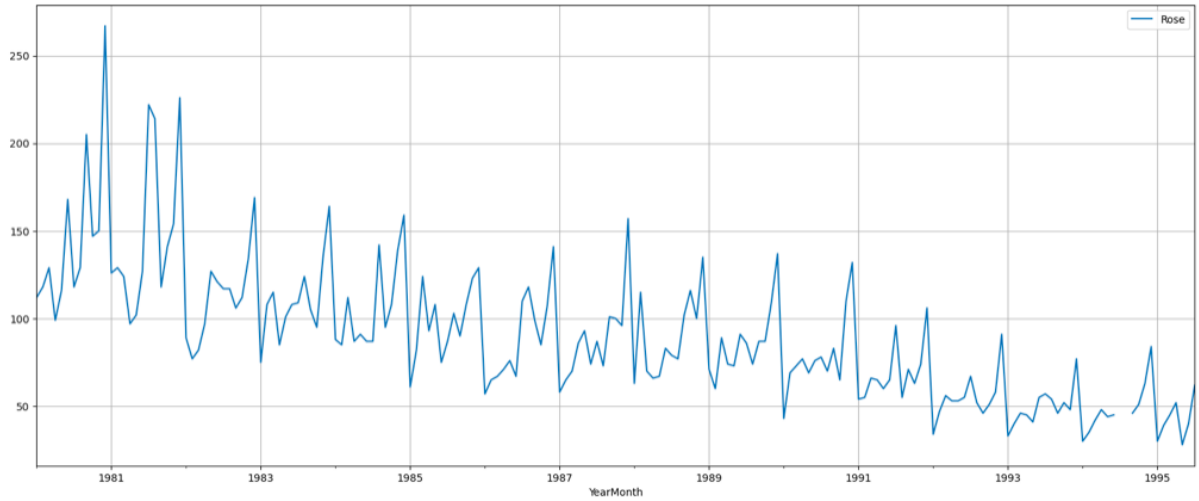
187 rows × 1 columns

Visualizing the data is a crucial step in understanding the sales patterns.

Sparkling dataset :



Rose dataset:



Exploratory Data Analysis (EDA):
EDA is a crucial step in understanding the characteristics of time series data.
Let's perform EDA and decomposition for both Sparkling and Rose wine sales:

The dataset has been enhanced by extracting separate columns for month and year from the 'YearMonth' column. Additionally, the 'Sparkling'. ‘Rose’ column has been renamed to 'Sales' to facilitate a more comprehensive analysis of the dataset. This restructuring allows for a clearer examination of sales trends over time, broken down by both month and year.

The Rose dataset contains two null values specifically observed for the months of July and August in the year 1994(Fig 1&2). This absence of data for these particular months may impact the overall analysis, and it's important to consider how to handle these missing values.

Rose Dataset

```
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Sales   185 non-null    float64
 1   Year    187 non-null    int32
 2   Month   187 non-null    int32
```

Figure 1

YearMonth	Sales	Year	Month
1994-07-01	NaN	1994	7
1994-08-01	NaN	1994	8

Figure 2

- **DatetimeIndex:** There are 187 entries ranging from January 1980 to July 1995.
- **Sales:** There are 185 non-null entries in the "Sales" column, indicating that there are two missing values.
- **Year:** The "Year" column contains integer values.
- **Month:** The "Month" column also contains integer values.

Sparkling Dataset

```
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Sales   187 non-null    int64
 1   Year    187 non-null    int32
 2   Month   187 non-null    int32
```

Figure 1

- **Sales Column:** There are 187 non-null entries with data type int64, indicating that it consists of integer values representing sales.
- **Year Column:** There are 187 non-null entries with data type int32, indicating that it consists of 32-bit integer values representing years.
- **Month Column:** There are 187 non-null entries with data type int32, indicating that it consists of 32-bit integer values representing months.

These statistics provide a snapshot of the central tendency, dispersion, and distribution of both Rose and sparkling dataset. (Fig 4&5)

	count	mean	std	min	25%	50%	75%	max
Sales	187.0	2402.417112	1295.111540	1070.0	1605.0	1874.0	2549.0	7242.0
Year	187.0	1987.299465	4.514749	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.406417	3.450972	1.0	3.0	6.0	9.0	12.0

Figure 2(Sparkling Dataset)

	count	mean	std	min	25%	50%	75%	max
Sales	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0
Year	187.0	1987.299465	4.514749	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.406417	3.450972	1.0	3.0	6.0	9.0	12.0

Figure 3(Rose Dataset)

Imputation using mean:

The missing values in the column for July and August 1994 are filled by computing the mean of sales values for the corresponding months from the surrounding years, specifically ranging from July 1993 to July 1995 for July and from August 1993 to August 1995 for August. This imputation approach utilizes the historical sales data from neighbouring years to estimate and complete the missing values.

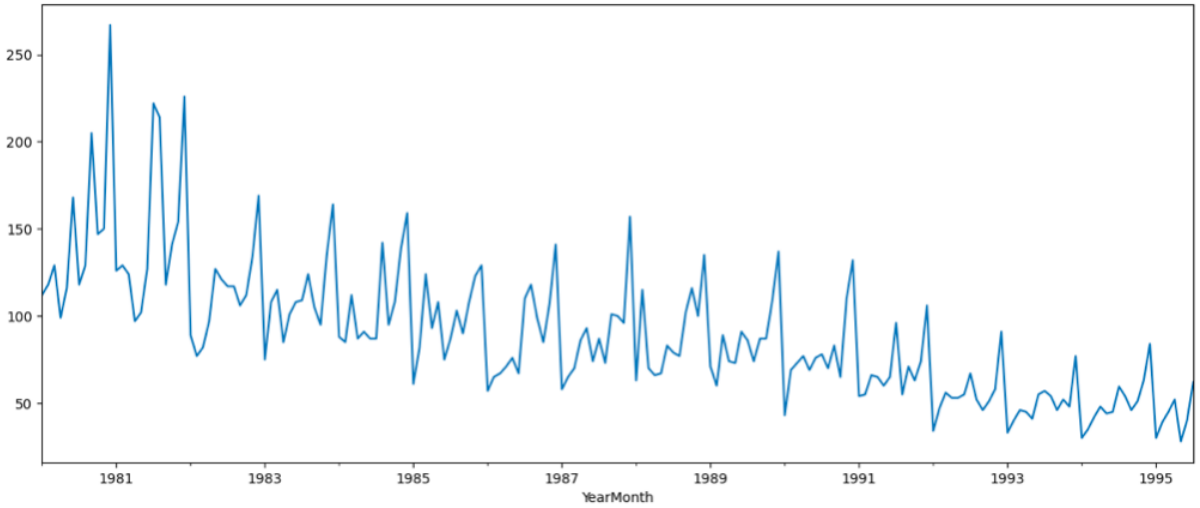


Figure 4

With missing values imputed, the dataset is now prepared for advanced analysis and exploration.

Rose Dataset:

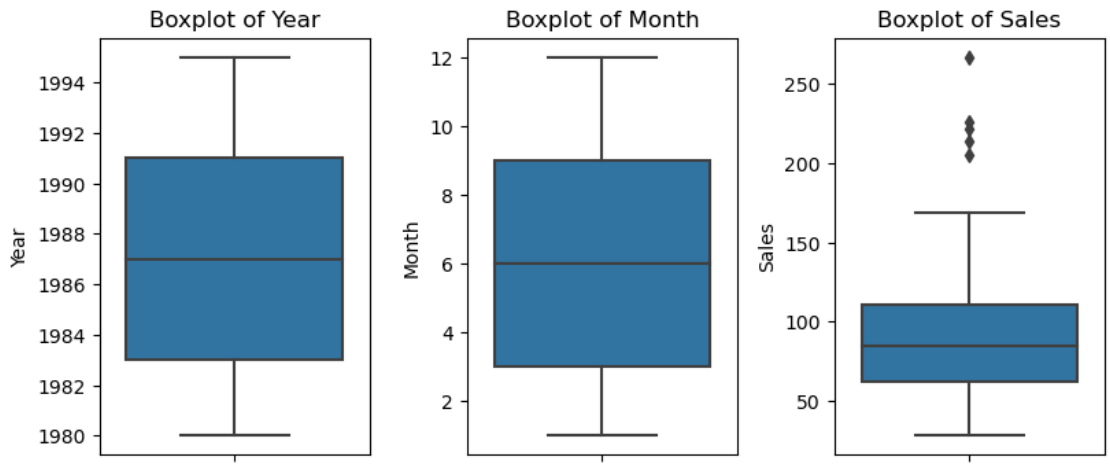


Figure 5

Outliers Detection and treatment:
We observe that there are outliers/extreme values in the sales column(Fig 5).
In this analysis, the decision not to treat outliers is deliberate and grounded in several considerations:

1. Preservation of Originality:
   - Opting not to manipulate outliers ensures the authenticity and integrity of the dataset, allowing the true variability and extreme values within the sales distribution to be faithfully represented.

2. Contextual Significance:
   - Recognizing that outliers can **carry meaningful insights**, especially during peak seasons or special events, refraining from their alteration allows us to capture the genuine dynamics and nuances of business operations.

3. Avoidance of Potential Bias:
   - Choosing not to impute or remove outliers is a precaution against introducing unintended biases or distorting the inherent characteristics of the sales distribution. This approach aims to maintain the accuracy and reliability of the original data.

By **refraining from outlier treatment**, we aim to conduct an analysis that reflects the raw and unaltered nature of the sales data, acknowledging the potential significance and context-dependent nature of extreme values within the business context.

Data Visualization:

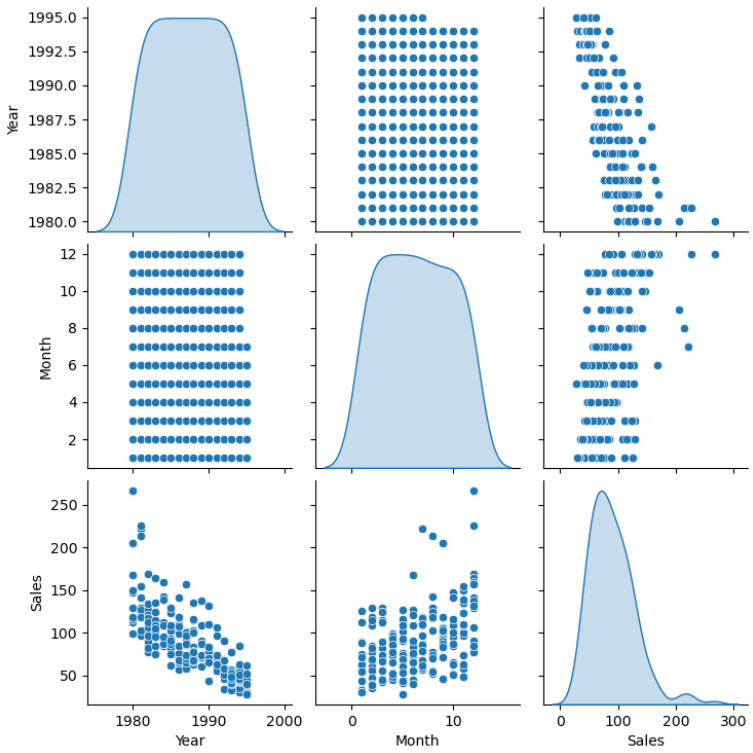


Figure 6

(Fig 6)The pair plot analysis reveals a discernible trend in the dataset, indicating a decrease in sales as the years progress. Additionally, it highlights a consistent pattern of higher sales

during the concluding months of each year. This observation suggests a potential seasonality effect, where wine sales tend to peak towards the end of the year, possibly influenced by factors such as holiday seasons or special events.

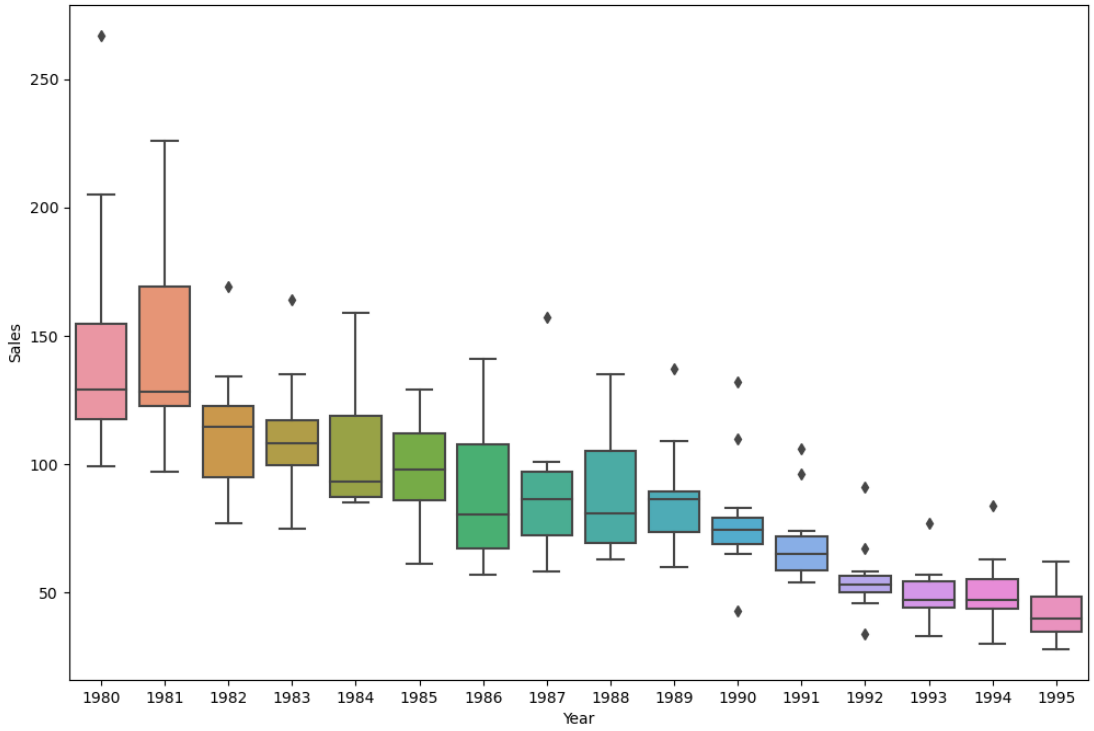


Figure 7

The above boxplot(Fig7) clearly depicts the gradual decrease in sales as the years progress. Let's deep dive into the patterns.

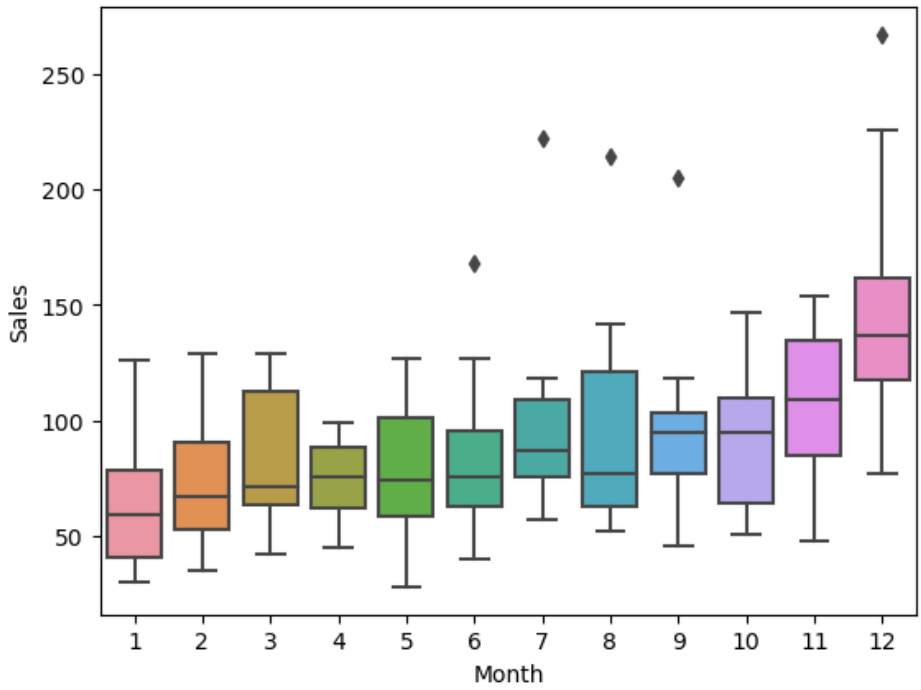


Figure 8

Overall, the monthly trend(Fig 8) in wine sales over the years exhibits several common patterns:

1. Consistency:
   - Across most months, there is a general consistency in sales, indicating a stable demand for wine over the years.

2. Seasonal Peaks:
   - Certain months consistently experience higher sales, with December being the most notable. This aligns with the holiday season and festivities.

3. Variability:
   - While most months show consistent patterns, some months exhibit variability in sales, with occasional peaks or dips.

In summary, the overall monthly trend reflects a stable demand for wine, with seasonality influencing sales, especially during festive months. Historical peaks in the early years suggest potential factors, such as marketing strategies or economic conditions, contributing to higher sales during those periods.

Month Year	1	2	3	4	5	6	7	8	9	10	11	12
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	59.5	54.0	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

Figure 9

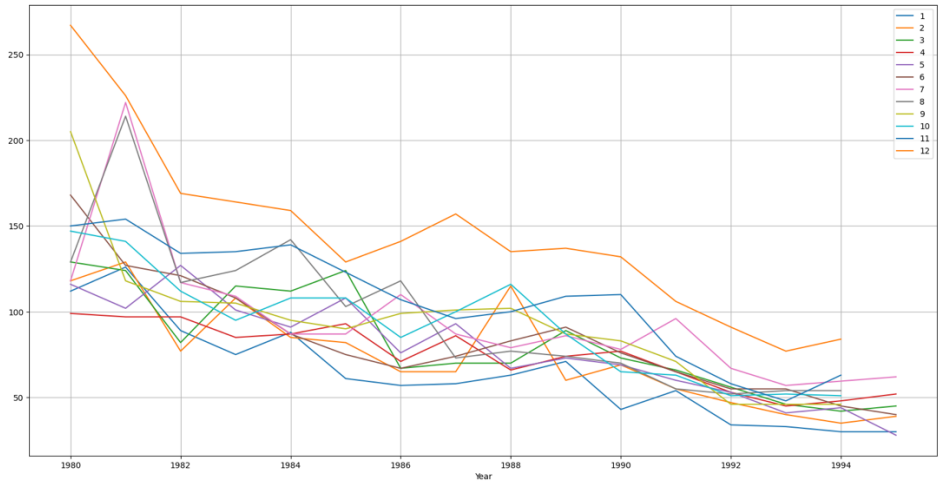


Figure 10

The provided table and (figure 10) summarizes the monthly sales figures for each year from 1980 to 1995. Here are some key findings:

1. **Yearly Sales Trends:**
    - The sales values vary across different years, indicating fluctuations in demand over time.
    - Some years, such as 1980, 1981, and 1987, exhibit higher sales peaks, while others, like 1982, 1986, and 1990, show relatively lower sales.
2. **Monthly Patterns:**
    - Certain months consistently demonstrate higher sales across multiple years. For example, November and December often have elevated sales, potentially due to holiday seasons.
    - July 1994 and August 1994 show imputed values (59.5 and 54.0, respectively) for sales, filling missing data points.
3. **Overall Variability:**
    - The sales data display variability both within and across years, suggesting the influence of various factors such as seasonality, promotional events, or market dynamics.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Month																
1	112.0	126.0	89.0	75.0	88.0	61.0	57.0	58.0	63.0	71.0	43.0	54.0	34.0	33.0	30.0	30.0
2	118.0	129.0	77.0	108.0	85.0	82.0	65.0	65.0	115.0	60.0	69.0	55.0	47.0	40.0	35.0	39.0
3	129.0	124.0	82.0	115.0	112.0	124.0	67.0	70.0	70.0	89.0	73.0	66.0	56.0	46.0	42.0	45.0
4	99.0	97.0	97.0	85.0	87.0	93.0	71.0	86.0	66.0	74.0	77.0	65.0	53.0	45.0	48.0	52.0
5	116.0	102.0	127.0	101.0	91.0	108.0	76.0	93.0	67.0	73.0	69.0	60.0	53.0	41.0	44.0	28.0
6	168.0	127.0	121.0	108.0	87.0	75.0	67.0	74.0	83.0	91.0	76.0	65.0	55.0	55.0	45.0	40.0
7	118.0	222.0	117.0	109.0	87.0	87.0	110.0	87.0	79.0	86.0	78.0	96.0	67.0	57.0	59.5	62.0
8	129.0	214.0	117.0	124.0	142.0	103.0	118.0	73.0	77.0	74.0	70.0	55.0	52.0	54.0	54.0	NaN
9	205.0	118.0	106.0	105.0	95.0	90.0	99.0	101.0	102.0	87.0	83.0	71.0	46.0	46.0	46.0	NaN
10	147.0	141.0	112.0	95.0	108.0	108.0	85.0	100.0	116.0	87.0	65.0	63.0	51.0	52.0	51.0	NaN
11	150.0	154.0	134.0	135.0	139.0	123.0	107.0	96.0	100.0	109.0	110.0	74.0	58.0	48.0	63.0	NaN
12	267.0	226.0	169.0	164.0	159.0	129.0	141.0	157.0	135.0	137.0	132.0	106.0	91.0	77.0	84.0	NaN

Figure 11

In specific years, certain months deviate from the overall trend, such as the dip in July 1986 or the lower sales in August 1994.
The early years, particularly 1980, 1981, and 1982, witnessed higher sales across multiple months, indicating historical peaks.

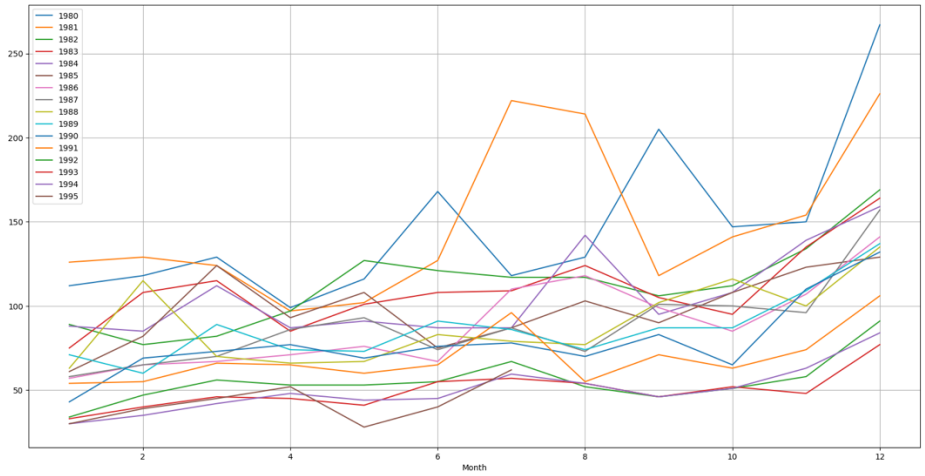


Figure 12

The presented table organizes monthly sales figures for each year from 1980 to 1995, providing a structured view of the sales distribution. Here are some key observations:

1. **Monthly Sales Trends:**
    - Variations in sales are evident across different months within each year, illustrating the seasonality or cyclicality of wine sales.
2. **Inter-Year Comparisons:**
    - Each row represents a specific month across the years, facilitating comparisons of sales performance during the same months across different years.
3. **Prominent Peaks:**
    - Certain months consistently demonstrate higher sales figures. Notably, November and December often exhibit elevated sales, potentially influenced by holiday seasons.

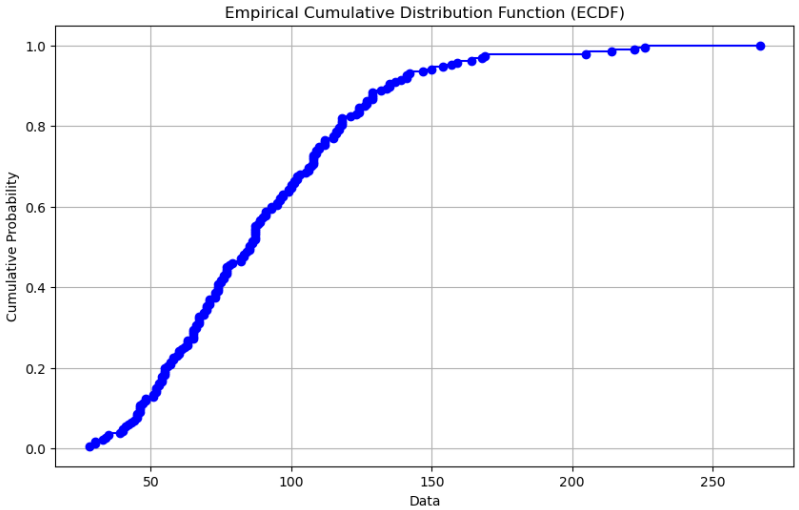


Figure 13

The Empirical Cumulative Distribution Function (ECDF) graph provides valuable insights into the distribution of sales data. Here are key observations:

1. 50% of Sales Less Than 100: The ECDF graph indicates that approximately 50% of the sales values fall below the 100 mark. This median value serves as a central point, suggesting that half of the observations are distributed below this threshold.

2. Highest Values Around 250:The plot illustrates that the highest sales values cluster around 250.

3. Approximately 90% of Sales Below 150: The ECDF curve also indicates that around 90% of the sales values are below the 150 mark.

In summary, the ECDF graph effectively communicates the distribution characteristics of the sales data, including central tendencies, upper limits, and percentile information.

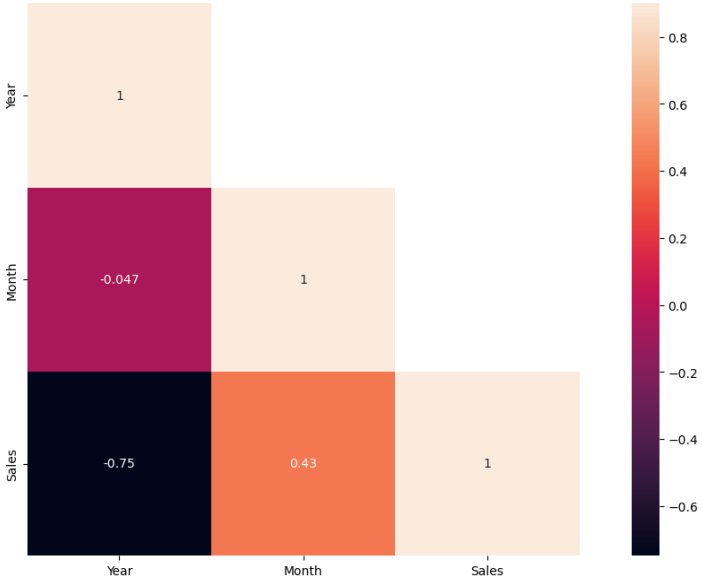


Figure 14

The heatmap analysis reveals interesting insights into the correlation patterns within the dataset. Contrary to a strong correlation with the years, the Sales column exhibits a relatively weak correlation. However, the correlation between the month and Sales columns is more pronounced, highlighting a clear seasonal pattern in the sales data. This suggests that specific months consistently witness higher or lower sales, emphasizing a recurring trend linked to the time of the year. The identification of this seasonal pattern is crucial for understanding the periodic variations in wine sales, potentially driven by factors such as holidays, events, or seasonal preferences.

Sparkling Dataset:

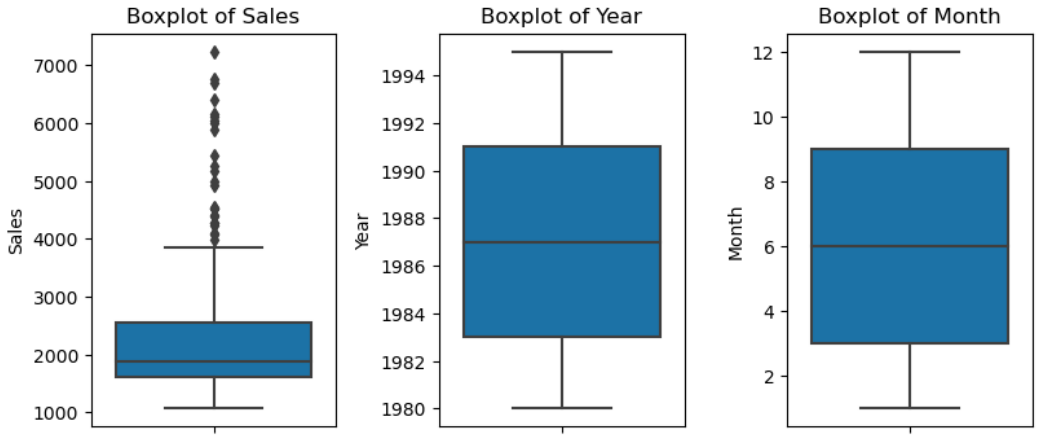


Figure 15

We observe that there are outliers/extreme values in the sales column.
We choose not to impute or remove outliers. This approach aims to maintain the accuracy and reliability of the original data.

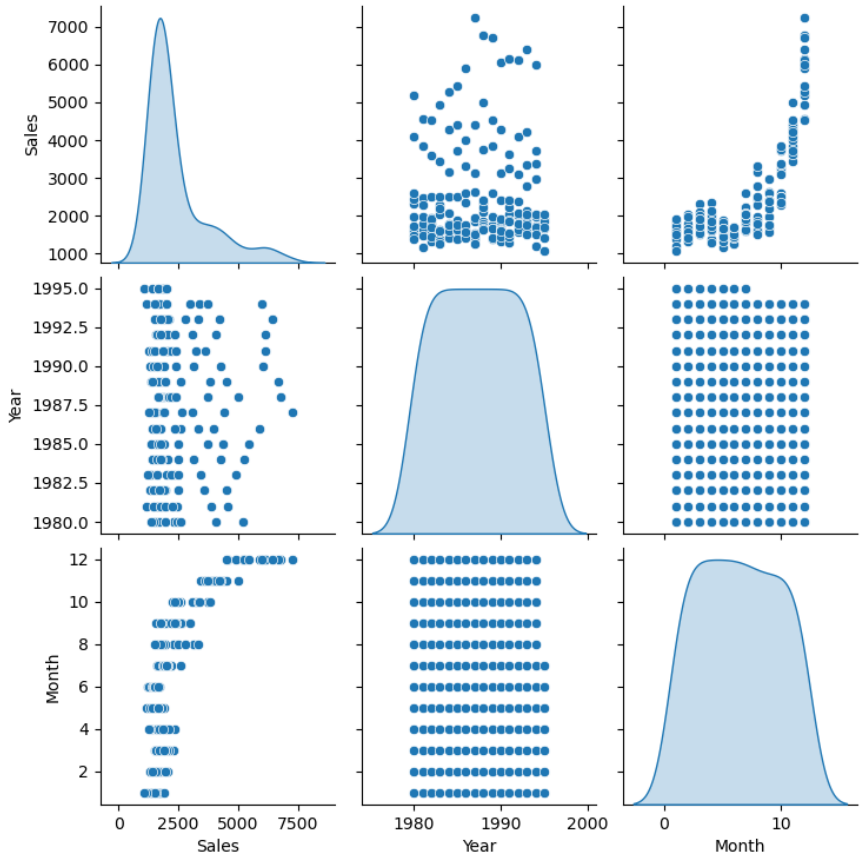
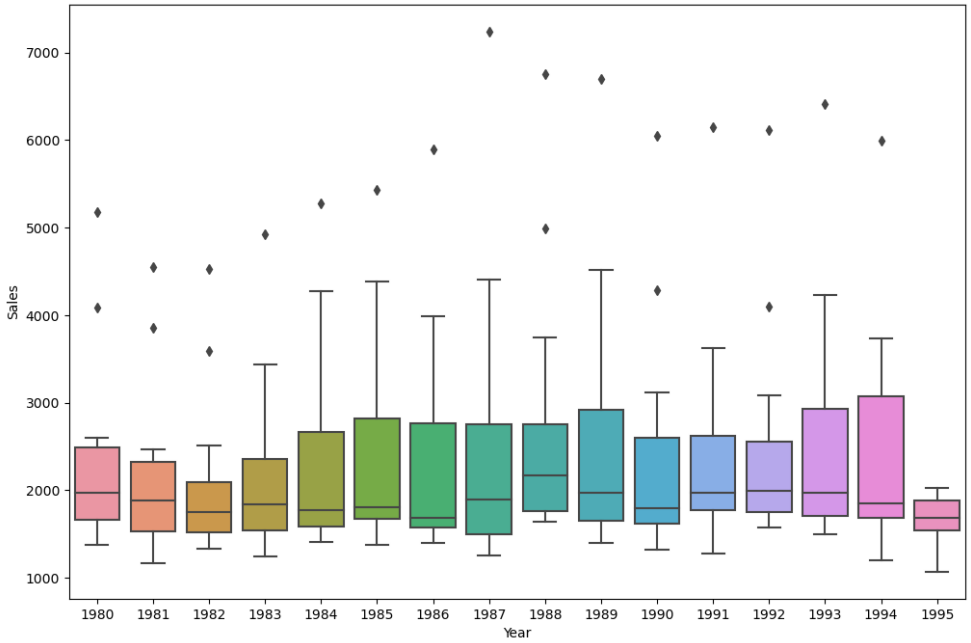


Figure 16

The pairplot(Fig 16) illustrates how sales values correlate with different months.



The yearly box plot (Fig17) for Sparkling wine reveals consistent sales patterns across different years, with a notable peak observed in 1988-1989. Throughout the years, there is a recurrent distribution of sales values, indicating stability in the overall performance.

Figure 17

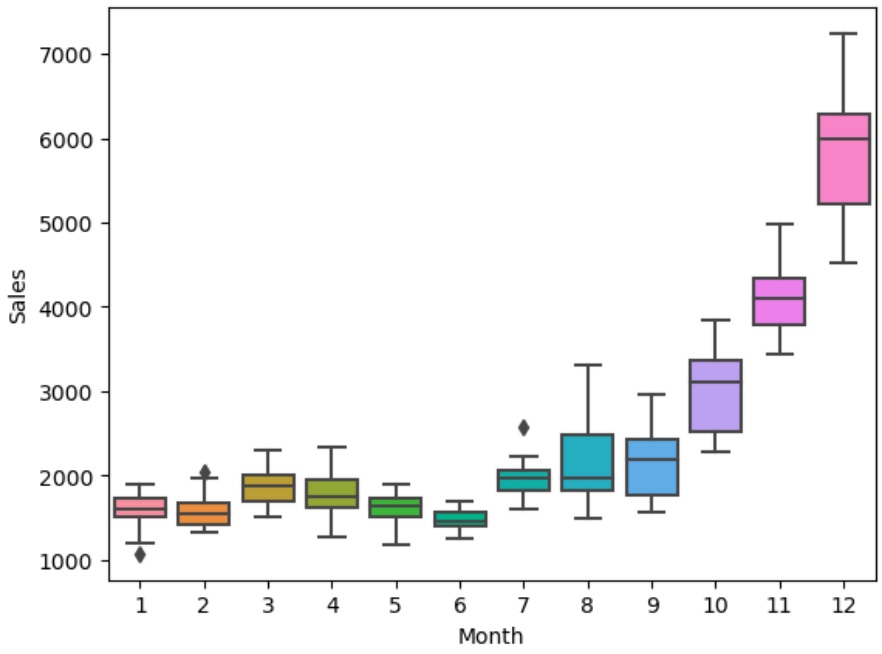


Figure 18

(Fig18)The plot illustrates a distinct monthly sales pattern, with December exhibiting the highest sales and January the lowest. From January to July, sales remain relatively consistent before experiencing an upward trend starting from August. Notably, outliers are observed in January, February, and July, suggesting occasional instances of significantly higher or lower sales during these months. The overall trend indicates seasonality and specific months contributing more significantly to the overall sales volume.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Year												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

Figure 19

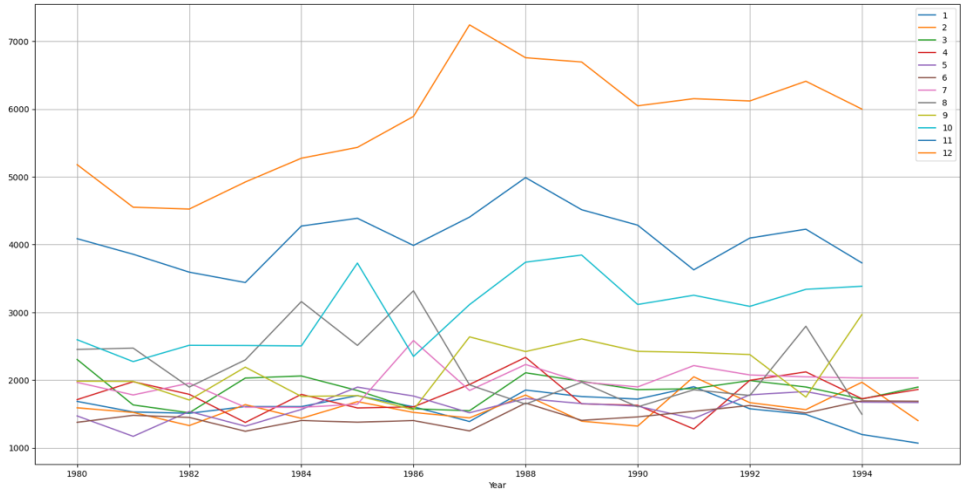


Figure 20

The yearly trend in wine sales data exhibits various patterns and trends over the decades. Here are some insights:

1. Overall Growth: The overall trend indicates growth in wine sales over the years, with generally increasing values from the 1980s to the 1990s.

2. Seasonal Fluctuations: Monthly sales show a recurring pattern, with peaks in certain months and lower sales in others. December consistently stands out as a month with the highest sales, while January tends to have lower sales.

3. Yearly Peaks: Certain years, such as 1980-1981, 1987-1988, and 1994-1995, experienced noticeable peaks in sales. These periods could be attributed to various factors, including marketing strategies, economic conditions, or specific events.

4. Consistency: Despite fluctuations, there is a level of consistency in the overall monthly and yearly patterns. The data reflects a degree of seasonality, with specific months contributing more significantly to the overall sales volume.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Month																
1	1686.0	1530.0	1510.0	1609.0	1609.0	1771.0	1606.0	1389.0	1853.0	1757.0	1720.0	1902.0	1577.0	1494.0	1197.0	1070.0
2	1591.0	1523.0	1329.0	1638.0	1435.0	1682.0	1523.0	1442.0	1779.0	1394.0	1321.0	2049.0	1667.0	1564.0	1968.0	1402.0
3	2304.0	1633.0	1518.0	2030.0	2061.0	1846.0	1577.0	1548.0	2108.0	1982.0	1859.0	1874.0	1993.0	1898.0	1720.0	1897.0
4	1712.0	1976.0	1790.0	1375.0	1789.0	1589.0	1605.0	1935.0	2336.0	1650.0	1628.0	1279.0	1997.0	2121.0	1725.0	1862.0
5	1471.0	1170.0	1537.0	1320.0	1567.0	1896.0	1765.0	1518.0	1728.0	1654.0	1615.0	1432.0	1783.0	1831.0	1674.0	1670.0
6	1377.0	1480.0	1449.0	1245.0	1404.0	1379.0	1403.0	1250.0	1661.0	1406.0	1457.0	1540.0	1625.0	1515.0	1693.0	1688.0
7	1966.0	1781.0	1954.0	1600.0	1597.0	1645.0	2584.0	1847.0	2230.0	1971.0	1899.0	2214.0	2076.0	2048.0	2031.0	2031.0
8	2453.0	2472.0	1897.0	2298.0	3159.0	2512.0	3318.0	1930.0	1645.0	1968.0	1605.0	1857.0	1773.0	2795.0	1495.0	NaN
9	1984.0	1981.0	1706.0	2191.0	1759.0	1771.0	1562.0	2638.0	2421.0	2608.0	2424.0	2408.0	2377.0	1749.0	2968.0	NaN
10	2596.0	2273.0	2514.0	2511.0	2504.0	3727.0	2349.0	3114.0	3740.0	3845.0	3116.0	3252.0	3088.0	3339.0	3385.0	NaN
11	4087.0	3857.0	3593.0	3440.0	4273.0	4388.0	3987.0	4405.0	4988.0	4514.0	4286.0	3627.0	4096.0	4227.0	3729.0	NaN
12	5179.0	4551.0	4524.0	4923.0	5274.0	5434.0	5891.0	7242.0	6757.0	6694.0	6047.0	6153.0	6119.0	6410.0	5999.0	NaN

Figure 21

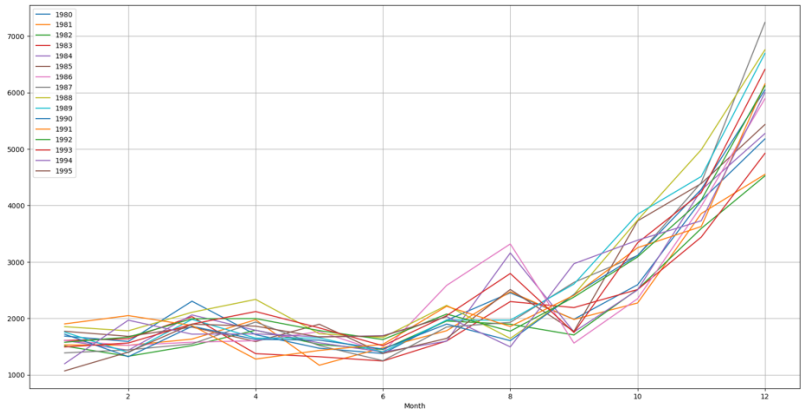


Figure 22

The plot illustrates that December consistently registers the highest sales over the years. Additionally, there is a noticeable peak in sales during the month of August.

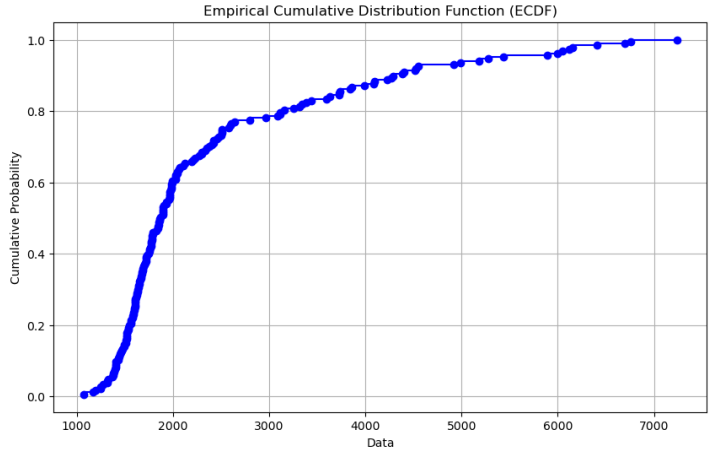


Figure 23

Fig 23, The ECDF (Empirical Cumulative Distribution Function) plot visually represents the distribution of the data.

- More than 50% of sales fall below the 2000 mark.
- The highest recorded values reach up to 7000 in the dataset.
- Approximately 80% of sales are below the 3000 threshold.

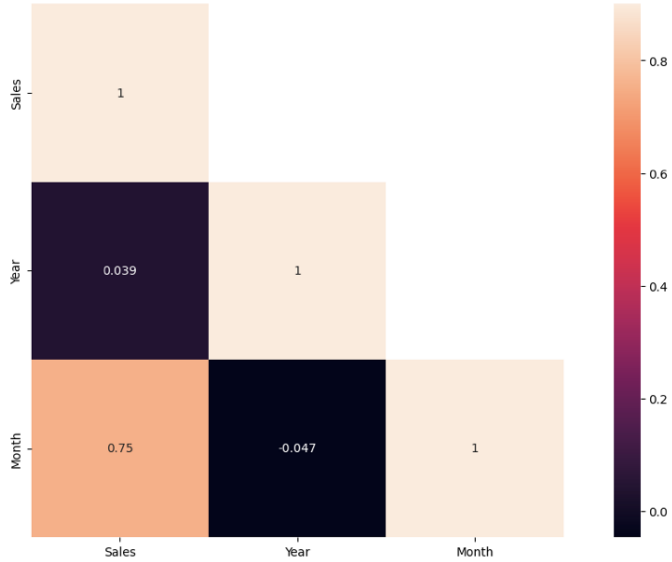
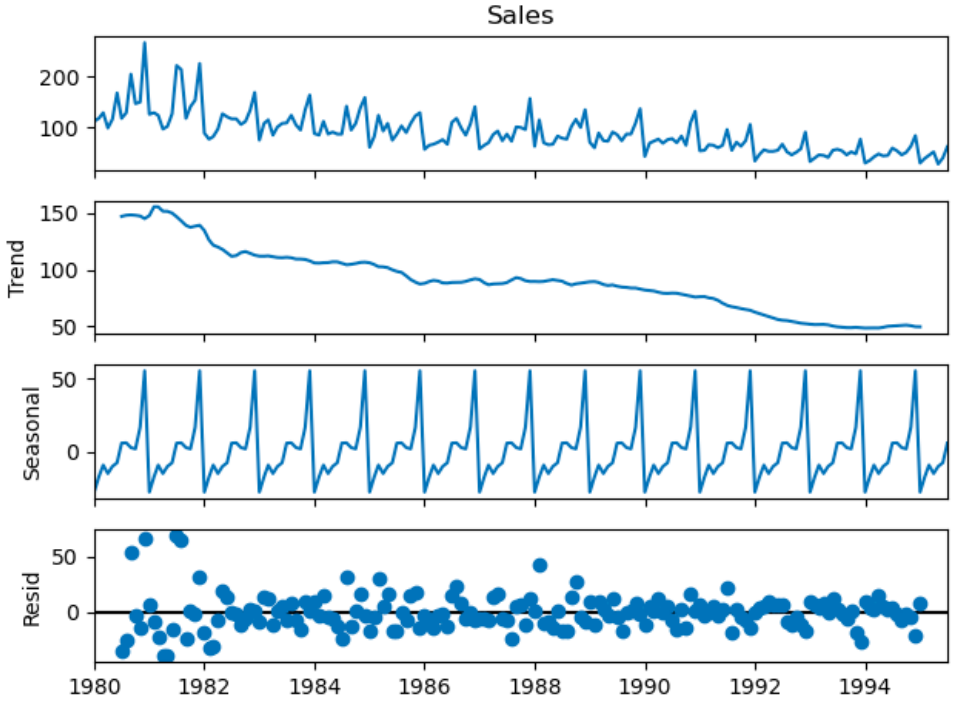


Figure 24

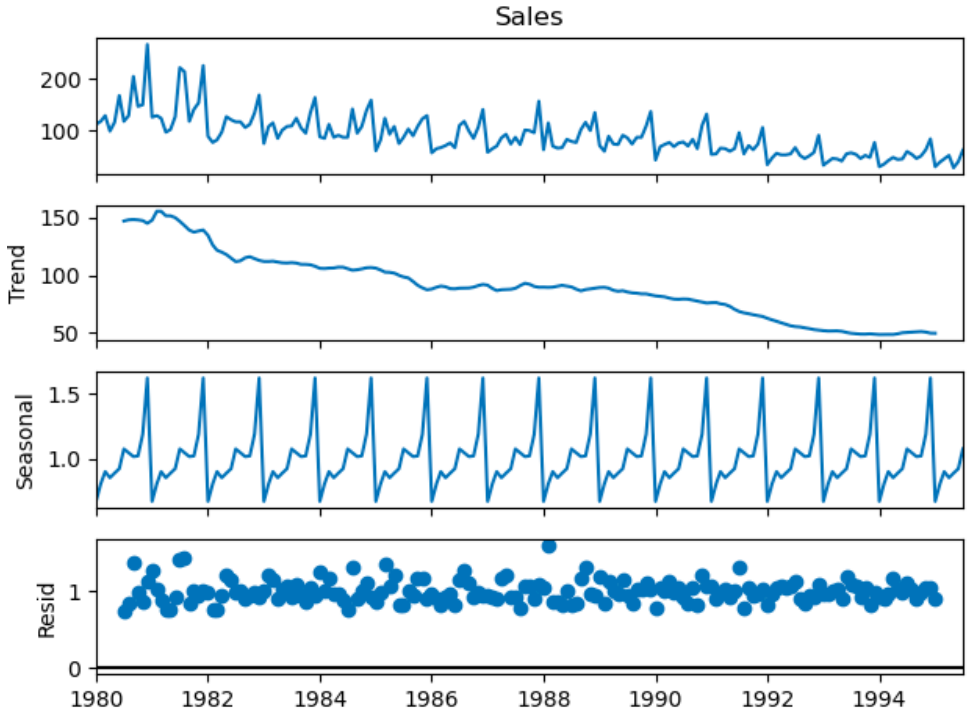
The heatmap highlights a modest correlation between sales and the year, emphasizing a more significant correlation between sales and the month. This points to distinct seasonal patterns in the sales data.

Rose dataset:

Decomposition- Additive



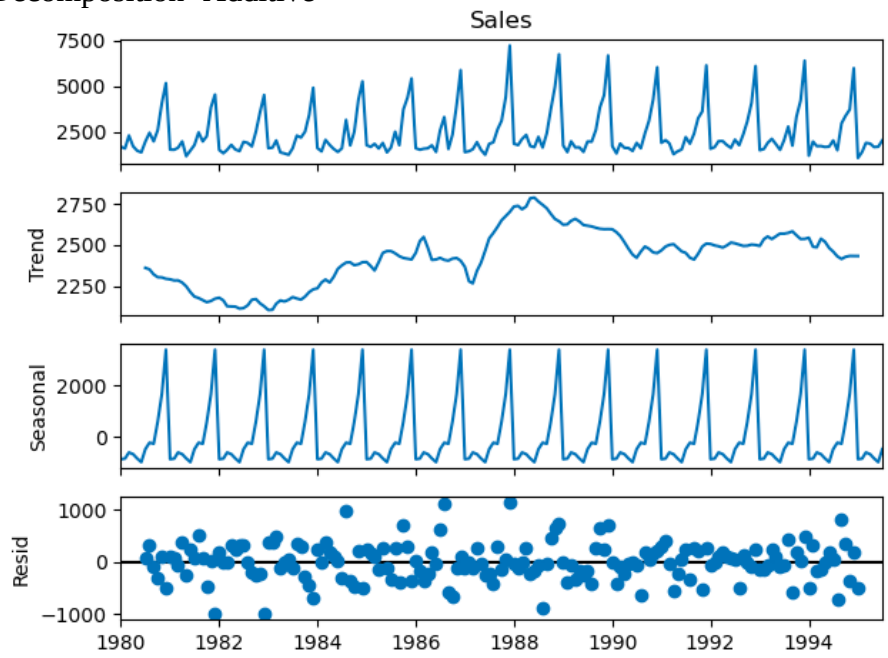
Decomposition- Multiplicative



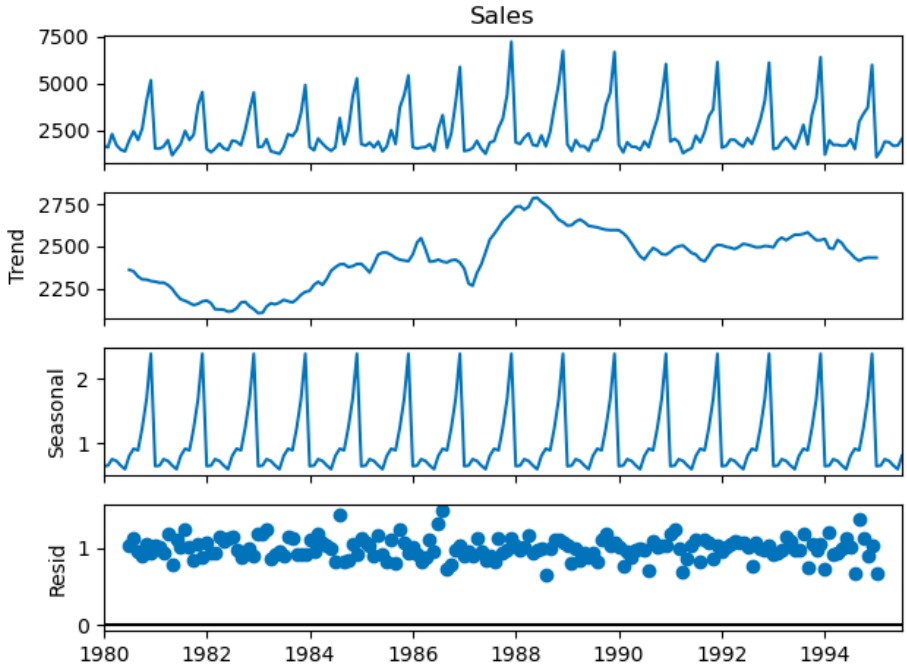
In both cases, the decompositions suggest a clear trend and seasonality in the data. [Decreasing trend and seasonality within months][Yt= Tt*St*It]
And the data exhibit multiplicative decomposition, with the residual component when plotted, appears to be scattered with no discernible pattern, resembling an approximately straight line.

Sparkling dataset:

Decomposition- Additive



Decomposition- Multiplicative



The choice between additive and multiplicative decomposition depends on whether the seasonality and trend exhibit constant or proportional fluctuations.
The data exhibit multiplicative decomposition, with the residual component when plotted, appears to be scattered with no discernible pattern, resembling an approximately straight line.

Splitting the dataset:

The training dataset comprises 132 rows and 3 columns, covering the period from 1980 to 1990. In contrast, the testing dataset consists of 55 rows and 3 columns, encompassing the years from 1991 to 1995.

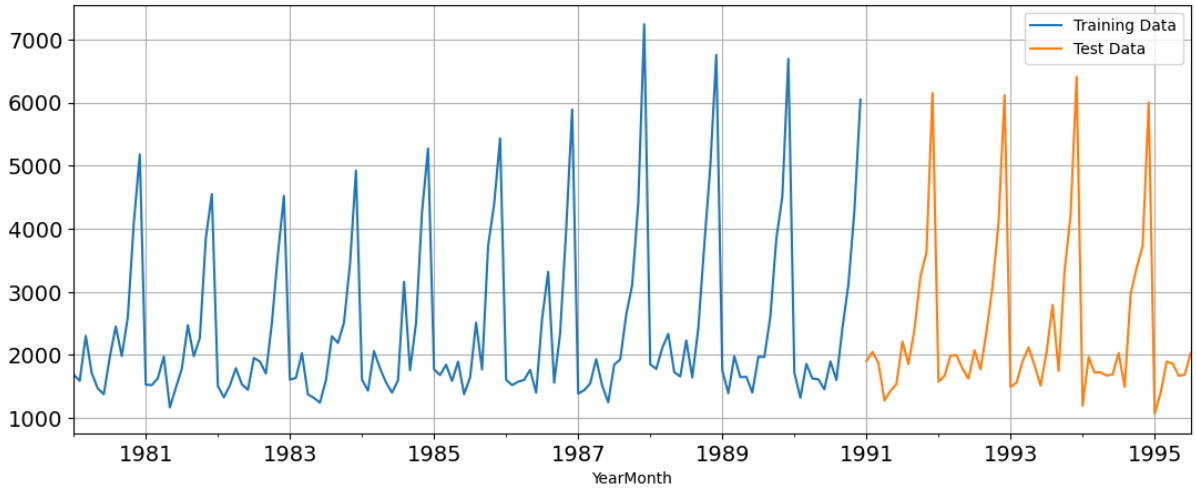


Figure 25 Sparkling dataset

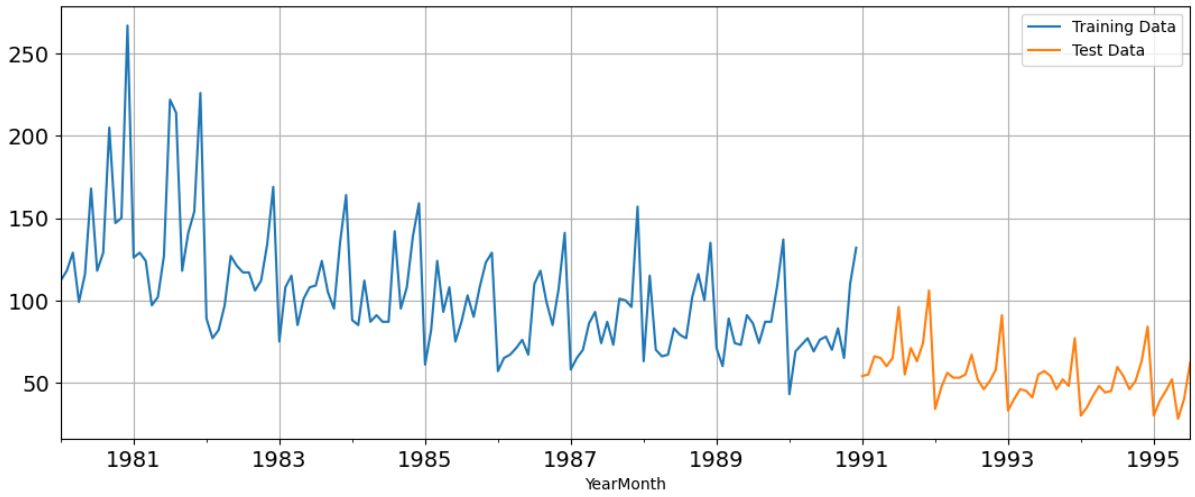


Figure 26 Rose Dataset

Model building:

In this analysis, various forecasting models will be explored and evaluated on the provided wine sales dataset. The primary goal is to develop robust models that effectively predict wine sales, particularly focusing on the years 1991 to 1995 as the test period. The training data spans from 1980 to 1990. A range of exponential smoothing models will be considered, encompassing different iterations and parameters to optimize their predictive capabilities.

In addition to exponential smoothing models, other baseline models such as regression, naïve forecast models, and simple average models will also be constructed using the training data. These diverse models will be assessed based on their Root Mean Squared Error (RMSE) when applied to the test data. By employing various forecasting techniques, this analysis aims to identify the most suitable model or combination of models for accurately predicting wine sales, providing valuable insights for future sales forecasting at ABC Estate Wines.

Root Mean Square Error (RMSE)

What is Root Mean Square Error (RMSE)?

Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance. **The lower the RMSE, the better the model and its predictions.**

Regression Model:

A time series regression forecasts a time series as a linear relationship with the independent variables. $y\ t = X\ t\ \beta + \epsilon\ t$. The linear regression model assumes there is a linear relationship between the forecast variable and the predictor variables.

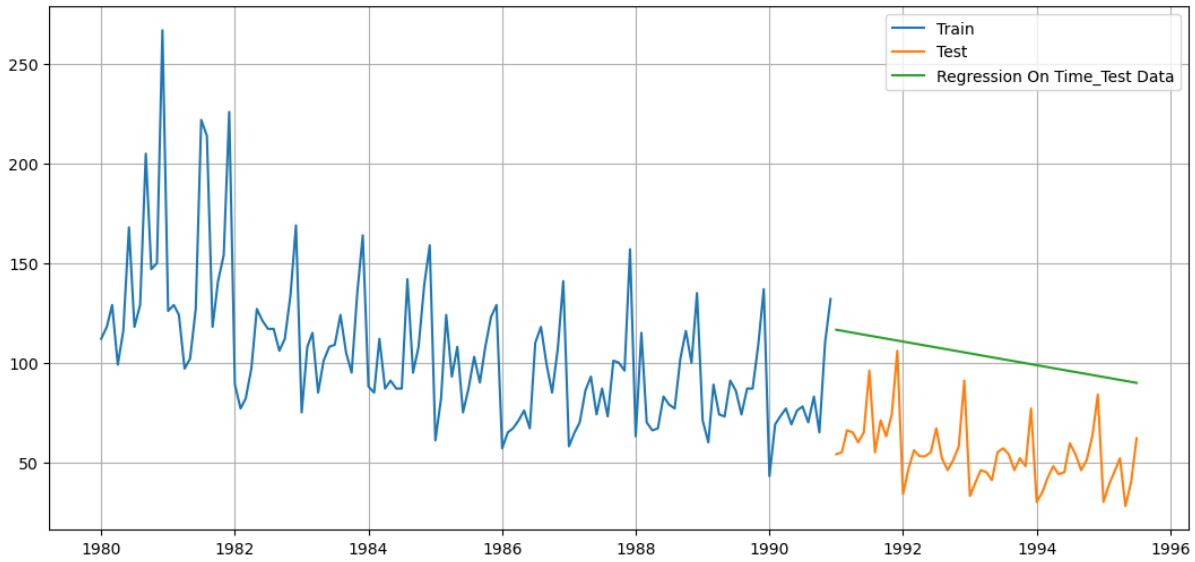


Figure 27 Rose Dataset

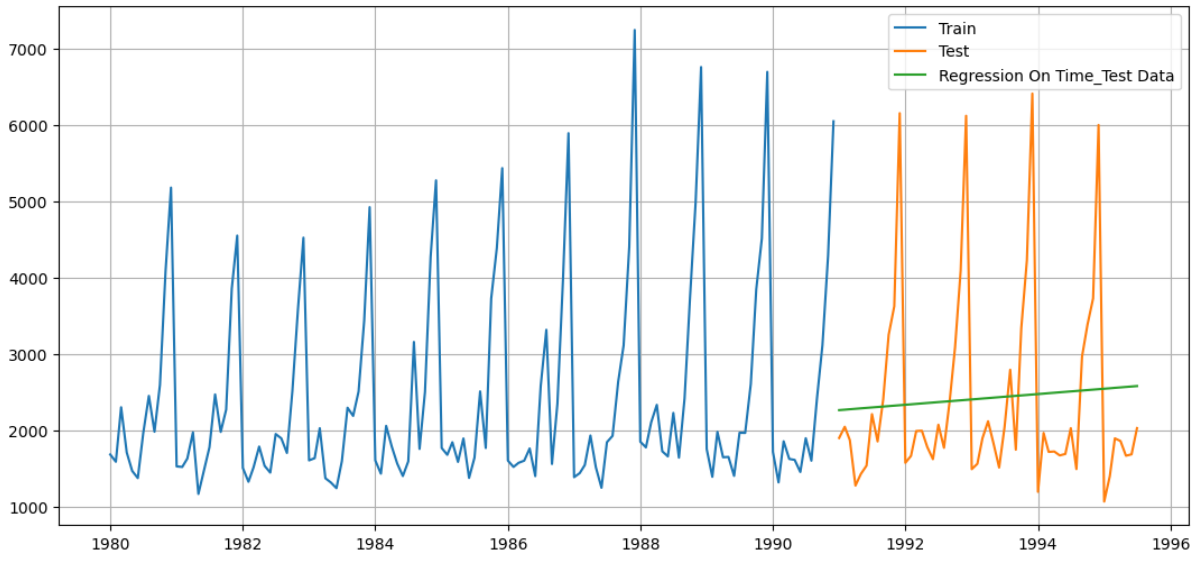


Figure 28 Sparkling Dataset

The model's predictions, represented by the green line, exhibit a considerable deviation from the actual test values depicted in orange. This discrepancy highlights a significant variance between the predicted and observed outcomes. The evaluation of the model, utilizing the Root Mean Squared Error (RMSE) metric, further emphasizes this disparity. The calculated RMSE for the Linear Regression model of Rose Dataset is 51.080941 and for sparkling dataset is 1275.86. In this case, a higher RMSE suggests a notable level of inaccuracy in the predictions made by the Linear Regression model.

Naive Approach:

The naive approach in time series forecasting involves making predictions based on the assumption that future values will be the same as the most recent observed value. In other words, the forecast for the next period is simply the value of the last observed data point. This approach is straightforward and easy to implement but may not capture more complex patterns or trends present in the data.

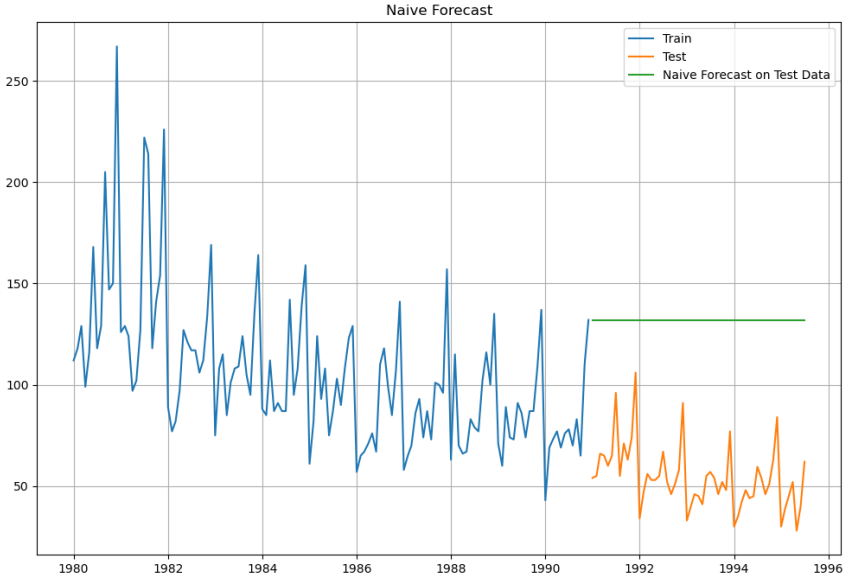


Figure 29 Rose Dataset

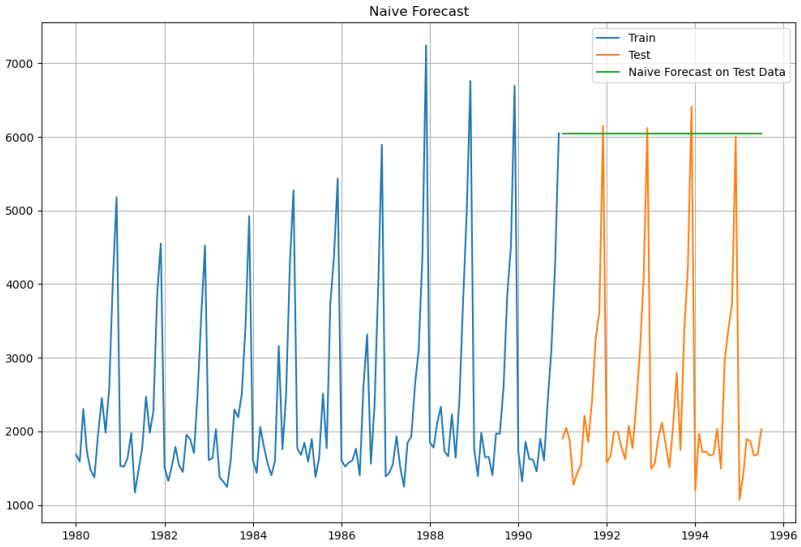


Figure 30 Sparkling Dataset

The model's predictions, depicted by the green line, diverge significantly from the actual test values represented by the orange line. The disparity is evident, indicating that the naive model, which relies on predicting future values based on the most recent observation, does not perform well in capturing the underlying patterns of the time series.

The Root Mean Squared Error (RMSE) was employed to evaluate the model's accuracy. The calculated RMSE for the Naive Model for Rose dataset is 79.30 and for Sparkling Dataset is 3864.27, reflecting the magnitude of the errors between predicted and actual values.

Simple Average Forecast:

The simple average time series forecast involves predicting future values by taking the average of historical observations. This straightforward method assumes that the future pattern of the time series will be similar to the average pattern observed in the past.

This method is easy to implement and provides a baseline for comparison with more sophisticated forecasting models. However, it may not capture underlying trends, seasonality, or other complex patterns present in the time series data. The simple average forecast is particularly suitable for stable and stationary time series with minimal variation.

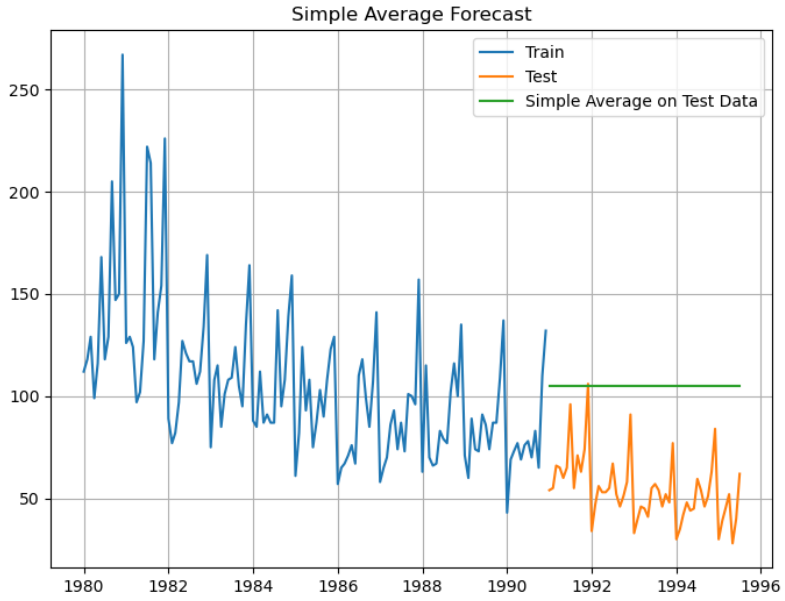


Figure 31 Rose Dataset

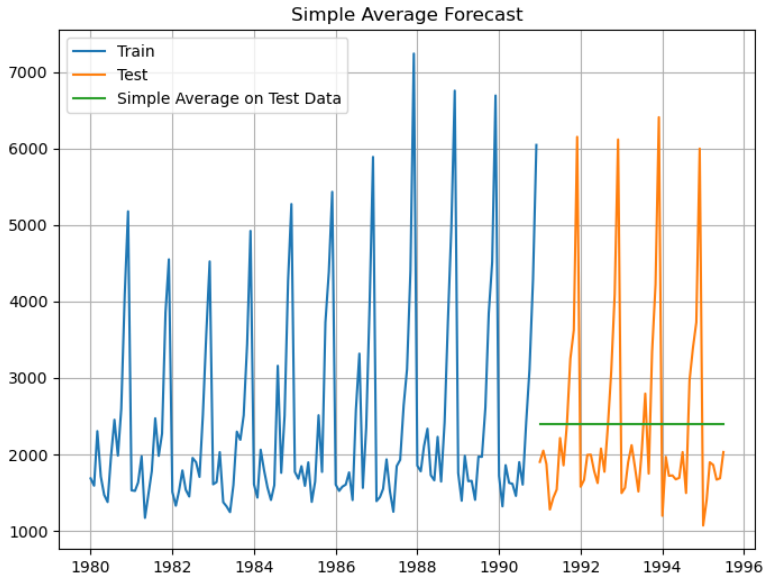


Figure 32 Sparkling Dataset

The green line in the plot represents the predictions generated by the model, while the orange line corresponds to the actual test values. Clearly, the predicted values deviate significantly from the actual values, indicating a limited predictive capability of the Simple Average

Model. The model's performance was assessed using the Root Mean Squared Error (RMSE) metric, yielding a value of 53.049755 for Rose dataset and 1275.08 for Sparkling dataset. This metric quantifies the average magnitude of the prediction errors, highlighting the disparities between the predicted and actual values.

Moving Average:

Moving Average is a statistical calculation used to analyze data over a certain period of time by creating a series of averages of different subsets of the full dataset. This method is commonly employed in time series analysis to smooth out short-term fluctuations and highlight longer-term trends or cycles.

The process involves taking the average of a specific number of consecutive data points (referred to as the "window" or "lag") and shifting the window through the dataset. The result is a new series of averages, which provides a clearer representation of the underlying patterns in the data.

The figure 33 shows the creation of trailing averages for the given dataset. Four types of trailing averages, with window sizes 2, 4, 6, and 9, have been calculated for each data point in the "YearMonth" column. The trailing averages represent the average values over the specified number of previous observations.

YearMonth	Sales	Year	Month	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	1686	1980	1	NaN	NaN	NaN	NaN
1980-02-01	1591	1980	2	1638.5	NaN	NaN	NaN
1980-03-01	2304	1980	3	1947.5	NaN	NaN	NaN
1980-04-01	1712	1980	4	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1980	5	1591.5	1769.50	NaN	NaN

Figure 33 Sparkling Dataset

YearMonth	Year	Month	Sales	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	1980	1	112.0	NaN	NaN	NaN	NaN
1980-02-01	1980	2	118.0	115.0	NaN	NaN	NaN
1980-03-01	1980	3	129.0	123.5	NaN	NaN	NaN
1980-04-01	1980	4	99.0	114.0	114.5	NaN	NaN
1980-05-01	1980	5	116.0	107.5	115.5	NaN	NaN

Figure 34 Rose Dataset

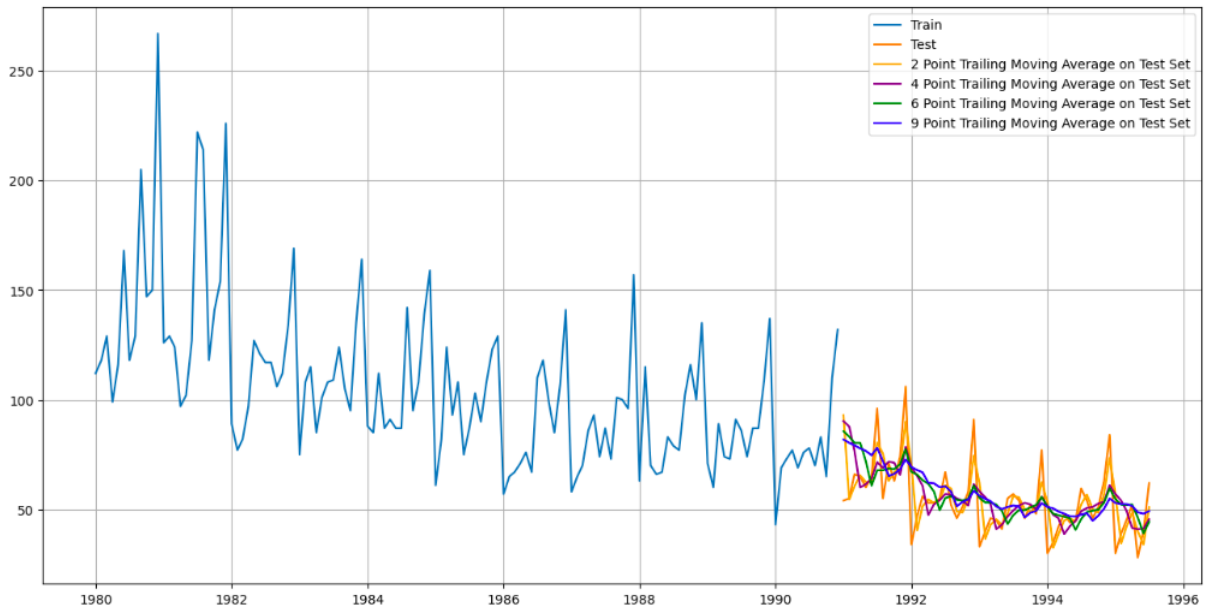


Figure 35 Rose Dataset

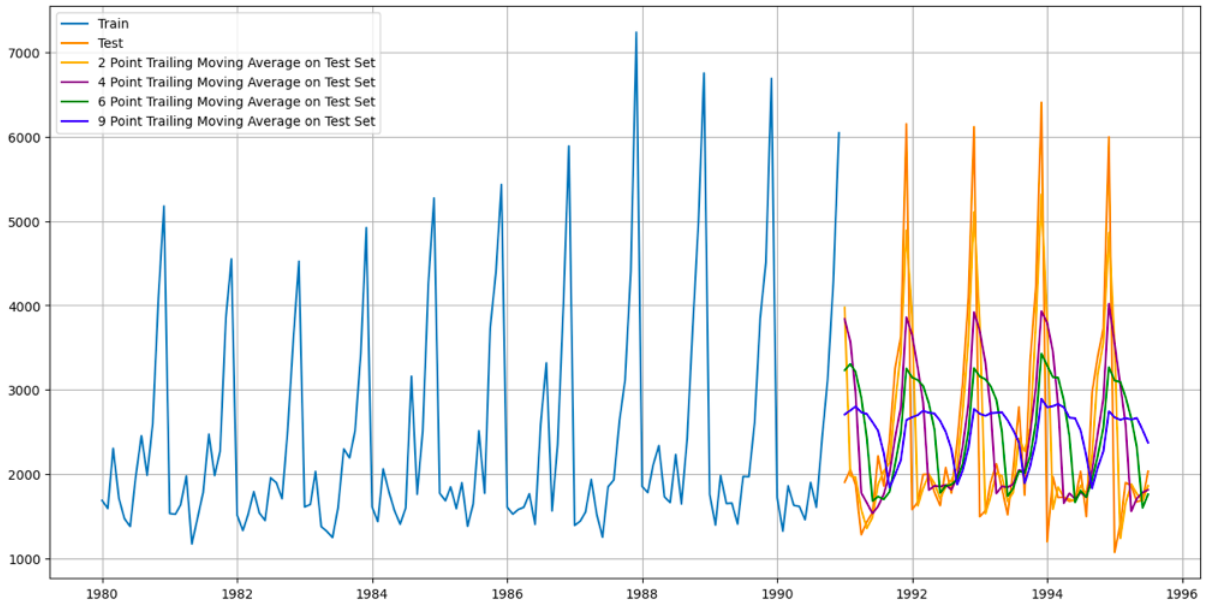


Figure 36 Sparkling Dataset

Utilizing a moving average is an improvement over a simple average as it considers only the preceding n values for making predictions. By focusing on recent trends, moving averages provide a more accurate representation. To assess the effectiveness of different moving windows, we will evaluate the Root Mean Square Error (RMSE) and determine which window size yields optimal results.

RMSE

2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

Figure 37 Sparkling Dataset

2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139

Figure 38 Rose Dataset

For both datasets, the Moving Average with a window size of 2 exhibits the lowest Root Mean Square Error (RMSE). This suggests that, among the tested window sizes (2, 4, 6, and 9), the 2-point Trailing Moving Average provides the most accurate predictions for the given time series data.

Simple Exponential Smoothing:

Simple Exponential Smoothing is a time series forecasting method that assigns exponentially decreasing weights to past observations. It is particularly useful for forecasting data with a constant level and no clear trend or seasonality. The method uses a smoothing parameter (alpha) that determines the weight given to the most recent observation.

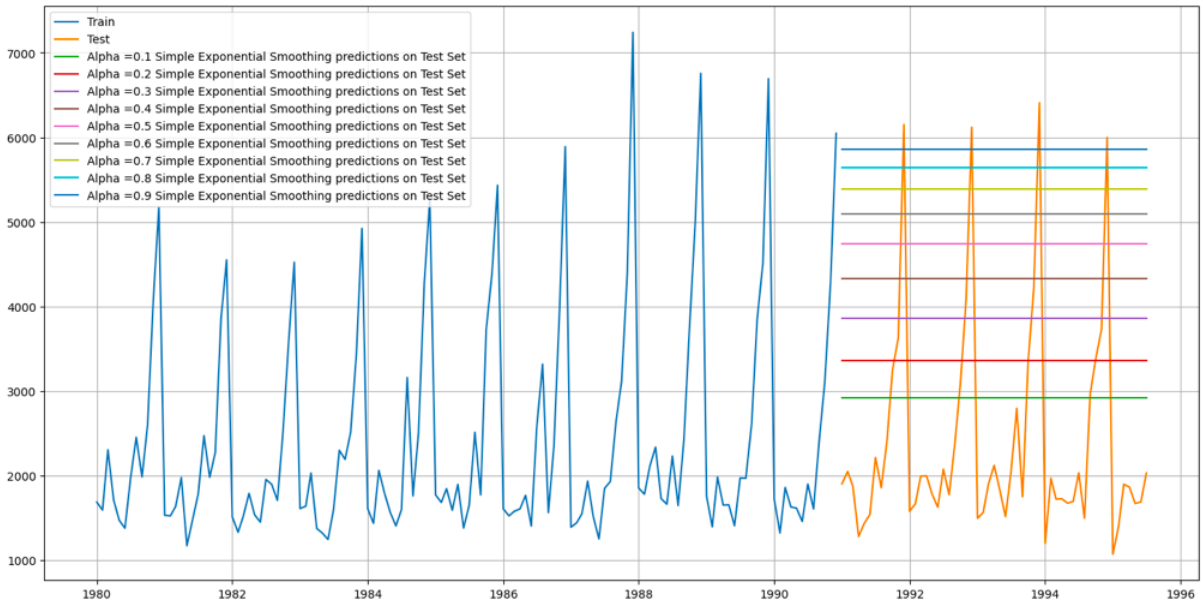


Figure 39 Sparkling Dataset

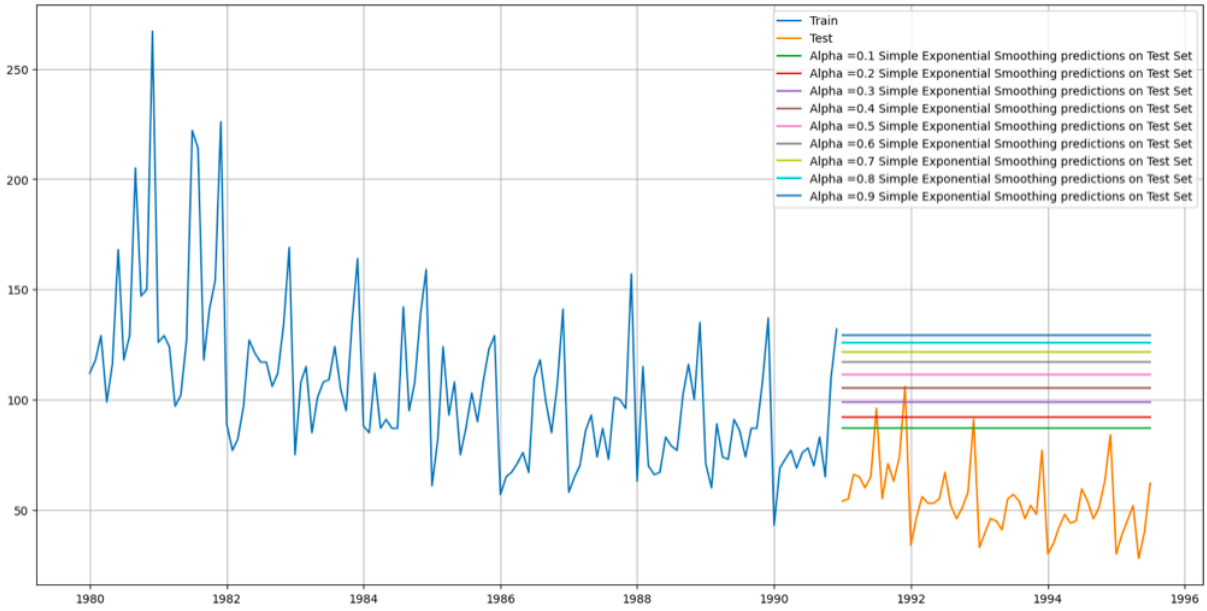


Figure 40 Rose Dataset

	Alpha Values	Train RMSE	Test RMSE
0	0.1	31.815610	36.429535
1	0.2	31.979391	40.957988
2	0.3	32.470164	47.096522
3	0.4	33.035130	53.356493
4	0.5	33.682839	59.229384
5	0.6	34.441171	64.558022
6	0.7	35.323261	69.284383
7	0.8	36.334596	73.359904
8	0.9	37.482782	76.725002

Alpha=0.1,SimpleExponentialSmoothing	36.429535

Rose Dataset

	Alpha Values	Train RMSE	Test RMSE
0	0.1	1333.873836	1375.393398
1	0.2	1356.042987	1595.206839
2	0.3	1359.511747	1935.507132
3	0.4	1352.588879	2311.919615
4	0.5	1344.004369	2666.351413
5	0.6	1338.805381	2979.204388
6	0.7	1338.844308	3249.944092
7	0.8	1344.462091	3483.801006
8	0.9	1355.723518	3686.794285

Alpha=0.1,SimpleExponentialSmoothing	1375.393398

Sparkling Dataset

Simple Exponential Smoothing (SES) is performed with various alpha values to find the one that minimizes the Root Mean Squared Error (RMSE). The table shows the results for different alpha values:

The results indicate that the SES model with an alpha value of 0.1 has the least RMSE, making it the preferred choice for forecasting in this case comparatively.

Double Exponential Smoothing (Holt's Model):

Double Exponential Smoothing, also known as Holt's method, is a time series forecasting technique that extends simple exponential smoothing by adding a trend component to the forecast model. It is used to handle time series data with a trend.

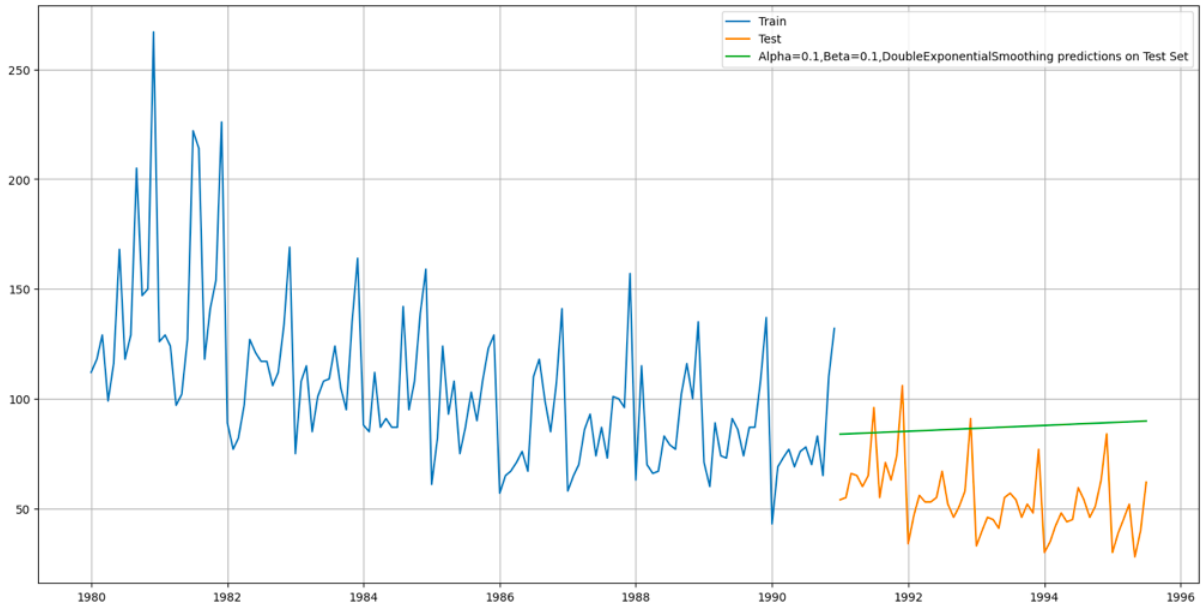


Figure 41 Rose Dataset

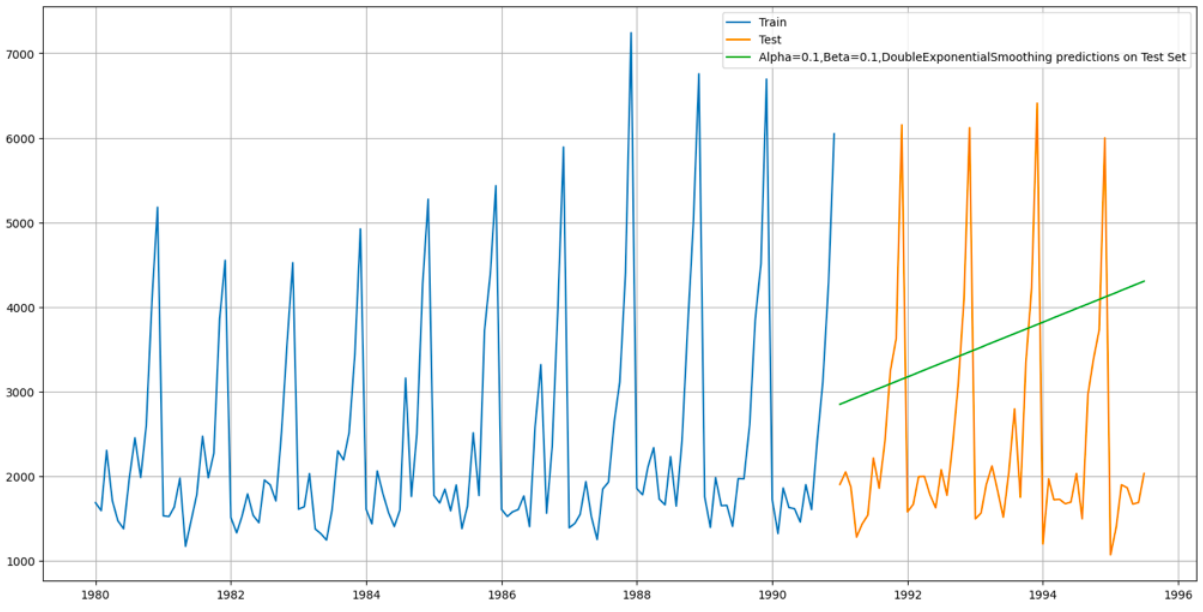


Figure 42 Sparkling Dataset

The forecast generated by the Double Exponential Smoothing (Holt's Model) with the selected parameters shows a significant deviation from the actual test values. The model seems to capture the trend component but falls short in accounting for the seasonality, resulting in a noticeable mismatch between the predicted and actual values.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	1382.520870	1778.564670
1	0.1	0.2	1413.598835	2599.439986
2	0.1	0.3	1445.762015	4293.084674
3	0.1	0.4	1480.897776	6039.537339
4	0.1	0.5	1521.108657	7390.522201
...	...	...	...	...
95	1.0	0.6	1753.402326	49327.087977
96	1.0	0.7	1825.187155	52655.765663
97	1.0	0.8	1902.013709	55442.273880
98	1.0	0.9	1985.368445	57823.177011
99	1.0	1.0	2077.672157	59877.076519

Figure 43 Sparkling Dataset

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	34.439111	36.510010
1	0.1	0.2	33.450729	48.221436
2	0.1	0.3	33.145789	77.649847
3	0.1	0.4	33.262191	99.064536
4	0.1	0.5	33.688415	123.742433
...	...	...	...	...
95	1.0	0.6	51.831610	801.137173
96	1.0	0.7	54.497039	841.349112
97	1.0	0.8	57.365879	853.421959
98	1.0	0.9	60.474309	834.167545
99	1.0	1.0	63.873454	779.536777

Figure 44 Rose Dataset

The optimal parameters for the Double Exponential Smoothing (Holt's Model) were determined through experimentation, and the configuration with an alpha value of 0.1 and a beta value of 0.1 yielded the least Root Mean Squared Error (RMSE). For the Rose dataset, the resulting RMSE was 36.50, while for the Sparkling dataset, it was 1778.56.

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

The three aspects of the time series behavior—value, trend, and seasonality—are expressed as three types of exponential smoothing, so Holt-Winters is called triple exponential smoothing. The model predicts a current or future value by computing the combined effects of these three influences.

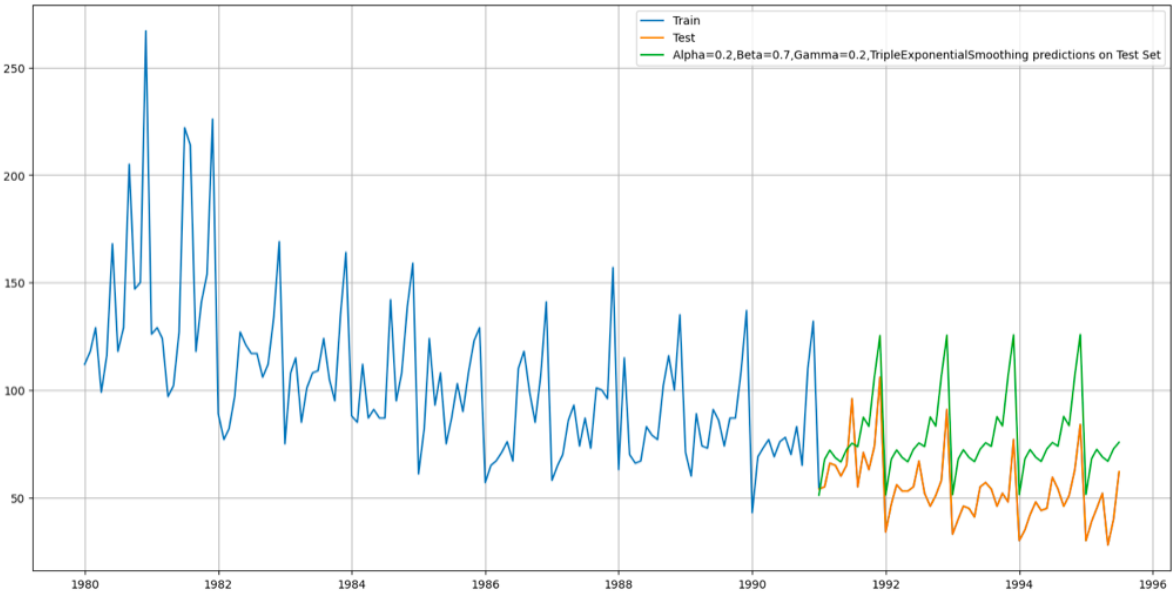


Figure 45 Rose Dataset

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE	Method
2136	0.2	0.7	0.2	24.042290	8.992350	tm_sm
1010	0.1	0.2	0.1	19.770392	9.221020	ta_sm
1011	0.1	0.2	0.2	20.253487	9.543696	ta_sm
1151	0.2	0.6	0.2	23.129850	9.922552	ta_sm
1012	0.1	0.2	0.3	20.871304	9.952909	ta_sm

Alpha=0.2,Beta=0.7,Gamma=0.2,TripleExponentialSmoothing 8.992350

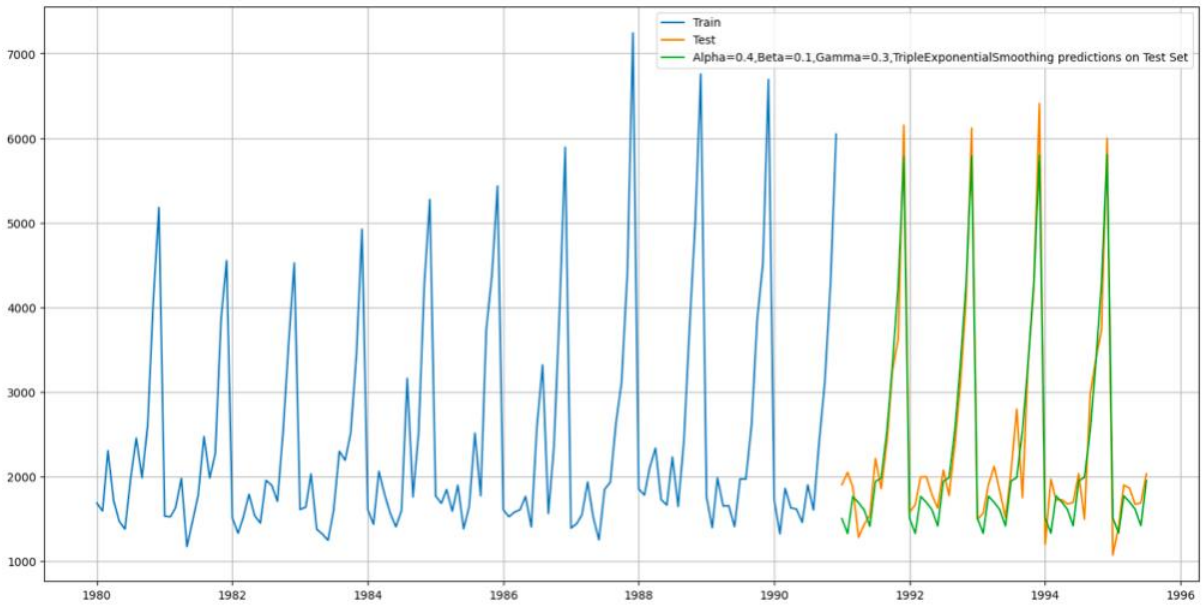


Figure 46 Sparkling Dataset

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE	Method
1301	0.4	0.1	0.2	384.467709	317.434302	ta_sm
2245	0.4	0.1	0.3	381.106645	326.579641	tm_sm
1211	0.3	0.2	0.2	388.544148	329.037543	ta_sm
1200	0.3	0.1	0.1	388.220071	337.080969	ta_sm
1110	0.2	0.2	0.1	398.482510	340.186457	ta_sm

Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing 317.434302

The Double Exponential Smoothing (Holt's Model) with respective alpha, beta and gamma values stands out as the most effective model among the ones tested, providing the lowest RMSE values of 8.99 and 317.43. This model successfully captures both the seasonality and trend in the dataset, and the predictions closely resemble the patterns observed in the test data graph. The superior performance of this model makes it a favorable choice for forecasting in this scenario.

Stationarity:

A time series is said to have Stationarity if the joint distribution of its values at any set of time points is the same as the joint probability distribution of its values at any other set of time points. In simpler terms, the statistical properties of the data should not change at different time intervals(a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality).

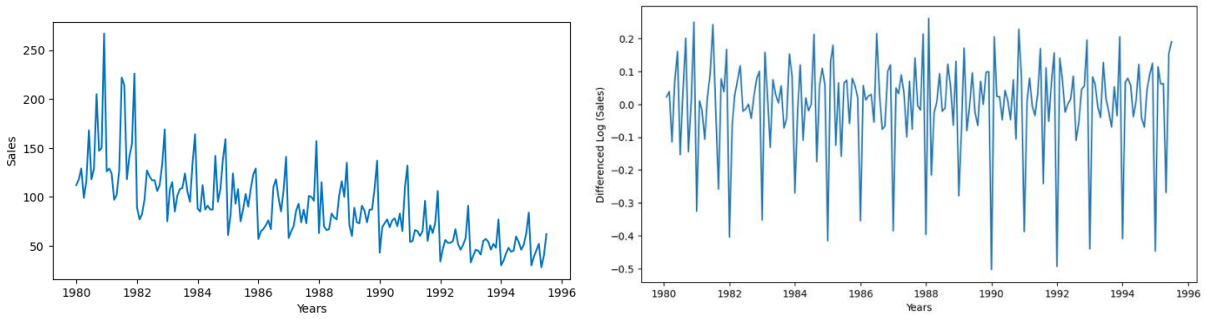
A non-stationary time series can exhibit trends or seasonality, making it challenging for models that assume stationarity. To assess stationarity, appropriate statistical tests are employed, typically at a significance level of 0.05 (alpha = 0.05).

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not . It is one of the most commonly used statistical test when it comes to analyzing the stationary of a series.

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.

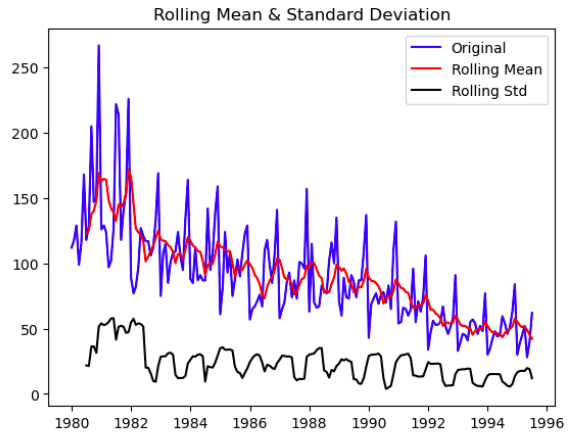
The objective in building ARIMA models is to work with stationary time series data. To assess the stationarity of our series, we conducted the Augmented Dickey-Fuller (ADF) test. At a 5% significance level.

Rose Dataset: The initial test yielded a p-value of 0.335674, indicating non-stationarity as we failed to reject the null hypothesis.

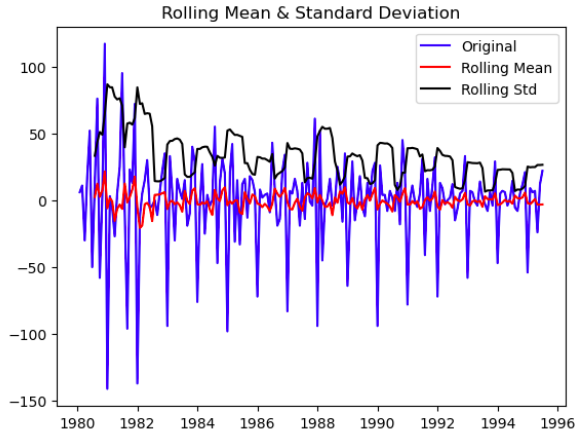


Dataset before and after differencing (d=1) on the log transformed time series

To transform the series into a stationary one, we employed the differencing approach using the .diff() function with a default difference value of 1(Refer he above graph). Subsequent ADF testing at this point resulted in a significantly lower p-value of 1.938803e-12, leading us to reject the null hypothesis. This signifies that the differenced series has become stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.892338
p-value                          0.335674
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```



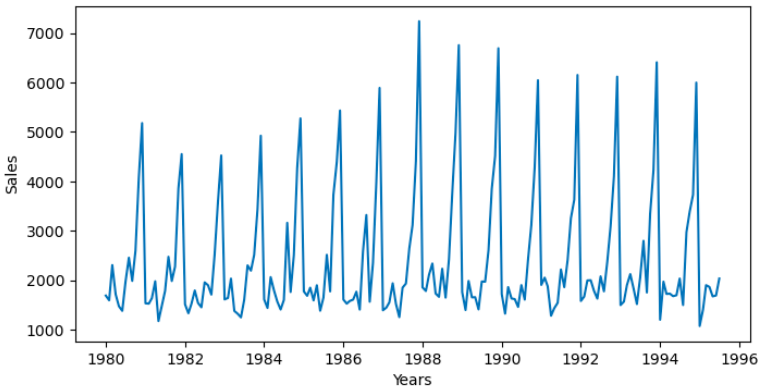

```
Results of Dickey-Fuller Test:
Test Statistic                  -8.032729e+00
p-value                          1.938803e-12
#Lags Used                       1.200000e+01
Number of Observations Used      1.730000e+02
Critical Value (1%)             -3.468726e+00
Critical Value (5%)             -2.878396e+00
Critical Value (10%)            -2.575756e+00
dtype: float64
```

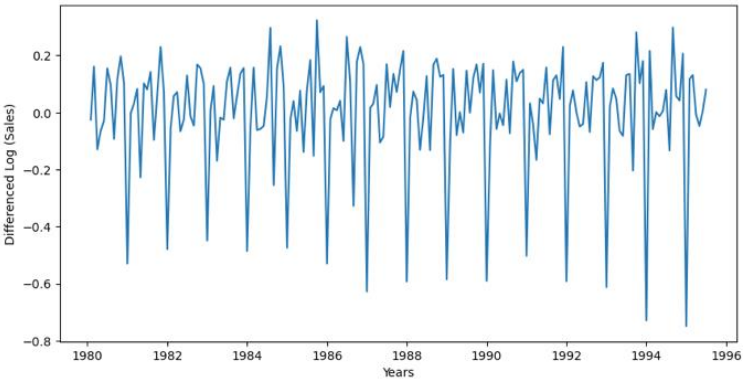
Before and after performing stationarity

With the series now meeting the stationarity requirement, we can confidently proceed with the construction of ARIMA/SARIMA models for our analysis.

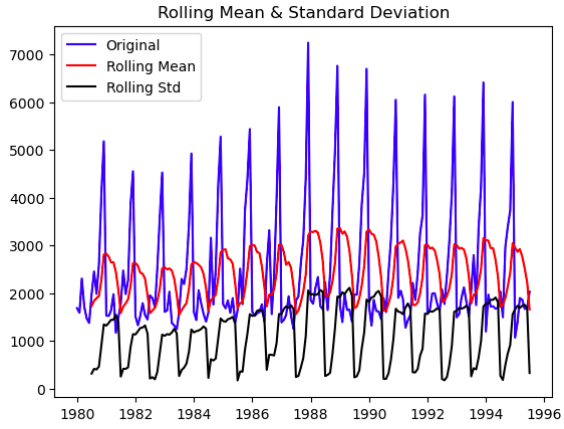
Sparkling Dataset:

The p-value of 0.601061 obtained from the Dickey-Fuller test indicates that, at the 5% significance level (alpha = 0.05), we fail to reject the null hypothesis. Differencing is done just like the Rose dataset.

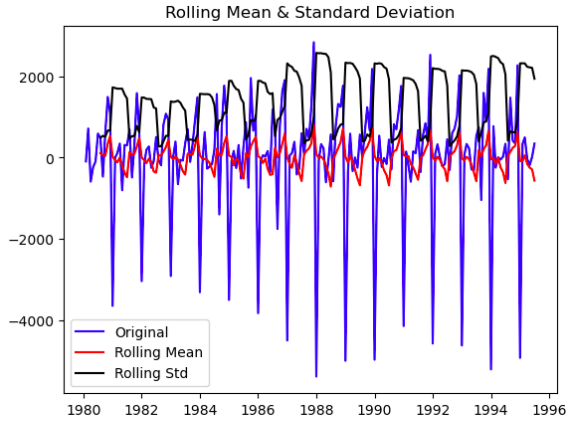




Dataset before and after differencing (d=1) on the log transformed time series




```
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
```




```
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
```

Subsequent ADF testing at this point resulted in a significantly lower p-value of 0.000000 leading us to reject the null hypothesis. This signifies that the differenced series has become stationary.

With the series now meeting the stationarity requirement, we can confidently proceed with the construction of ARIMA/SARIMA models for our analysis.

Akaike Information Criteria (AIC):

In the process of automating the ARIMA/SARIMA modeling, we aim to develop a systematic approach for selecting the most suitable parameters. This is achieved by leveraging the Akaike Information Criteria (AIC) on the training data. The AIC is a statistical measure that balances the goodness of fit against the complexity of the model. In essence, lower AIC values suggest a better trade-off between accuracy and simplicity.

ARIMA:

ARIMA stands for auto-regressive integrated moving average. It's a way of modelling time series data for forecasting (i.e., for predicting future points in the series), in such a way that:

- a pattern of growth/decline in the data is accounted for (hence the "auto-regressive" part)
- the rate of change of the growth/decline in the data is accounted for (hence the "integrated" part)
- noise between consecutive time points is accounted for (hence the "moving average" part).

To identify the optimal values for the ARIMA model parameters (p, d, q), we implemented a for loop. Here, p represents the order of the Auto-Regressive (AR) component, q signifies the order of the Moving Average (MA) component, and d denotes the differencing required to achieve stationarity in the time series data.

Within the for loop, we explored a range of values for p and q, specifically testing values from 0 to 4. This choice aimed to comprehensively search for the most suitable parameters within this specified range. Meanwhile, a fixed value of 1 was assigned to d, as we had previously determined this value through stationarity testing using the Augmented Dickey-Fuller (ADF) test.

This systematic exploration of parameter combinations allows us to fine-tune the ARIMA model for optimal performance on the training data.

Rose Dataset:

The Akaike Information Criterion (AIC) was computed for various ARIMA models, and the model with the lowest AIC value was chosen for optimal performance. For the rose dataset, the selected ARIMA model is (2, 1, 3), where the parameters represent the order of the Auto-Regressive (AR), differencing (d), and Moving Average (MA) components, respectively. The corresponding AIC value for this model is 1274.70, indicating its relative goodness of fit.

Additionally, the Root Mean Squared Error (RMSE) for this ARIMA(2, 1, 3) model on the test data was calculated to be 36.41, providing a quantitative measure of the model's accuracy in predicting the observed values. Refer Figure 47.

	param	AIC
11	(2, 1, 3)	1274.695116
15	(3, 1, 3)	1278.663811
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.574230
9	(2, 1, 1)	1281.507862
10	(2, 1, 2)	1281.870722
7	(1, 1, 3)	1281.870722
1	(0, 1, 1)	1282.309832
13	(3, 1, 1)	1282.419278
14	(3, 1, 2)	1283.720741
12	(3, 1, 0)	1297.481092
8	(2, 1, 0)	1298.611034
4	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	−631.348
Date:	Wed, 17 Jan 2024	AIC	1274.695
Time:	12:22:30	BIC	1291.946
Sample:	01-01-1980	HQIC	1281.705
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	−1.6779	0.084	−20.050	0.000	−1.842	−1.514
ar.L2	−0.7288	0.084	−8.710	0.000	−0.893	−0.565
ma.L1	1.0448	0.664	1.573	0.116	−0.257	2.346
ma.L2	−0.7719	0.135	−5.713	0.000	−1.037	−0.507
ma.L3	−0.9047	0.603	−1.501	0.133	−2.086	0.277
sigma2	858.1925	559.019	1.535	0.125	−237.465	1953.850

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.45
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	4.57

Auto_ARIMA	36.418927

Figure 47

Sparkling Dataset:

The Akaike Information Criterion (AIC) was computed for various ARIMA models, and the model with the lowest AIC value was chosen for optimal performance. For the sparkling dataset, the selected ARIMA model is (2, 1, 2), where the parameters represent the order of the Auto-Regressive (AR), differencing (d), and Moving Average (MA) components, respectively. The corresponding AIC value for this model is 2213.50, indicating its relative goodness of fit.

Additionally, the Root Mean Squared Error (RMSE) for this ARIMA(2, 1, 2) model on the test data was calculated to be 1299.98, providing a quantitative measure of the model's accuracy in predicting the observed values. Refer Figure 48.

	param	AIC
10	(2, 1, 2)	2213.509213
15	(3, 1, 3)	2221.454407
14	(3, 1, 2)	2230.778713
11	(2, 1, 3)	2232.944705
9	(2, 1, 1)	2233.777626
3	(0, 1, 3)	2233.994858
2	(0, 1, 2)	2234.408323
6	(1, 1, 2)	2234.527200
13	(3, 1, 1)	2235.498829
7	(1, 1, 3)	2235.607816
5	(1, 1, 1)	2235.755095
12	(3, 1, 0)	2257.723379
8	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
4	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Fri, 19 Jan 2024	AIC	2213.509
Time:	13:30:24	BIC	2227.885
Sample:	01-01-1980 - 12-01-1990	HQIC	2219.351
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.784	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.741	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.218	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.109	0.000	0.785	1.215
sigma2	1.099e+06	1.99e-07	5.51e+12	0.000	1.1e+06	1.1e+06

Ljung-Box (L1) (Q):	0.19	Jarque-Bera (JB):	14.46
Prob(Q):	0.67	Prob(JB):	0.00
Heteroskedasticity (H):	2.43	Skew:	0.61
Prob(H) (two-sided):	0.00	Kurtosis:	4.08

Auto_ARIMA	1299.982793

Figure 48

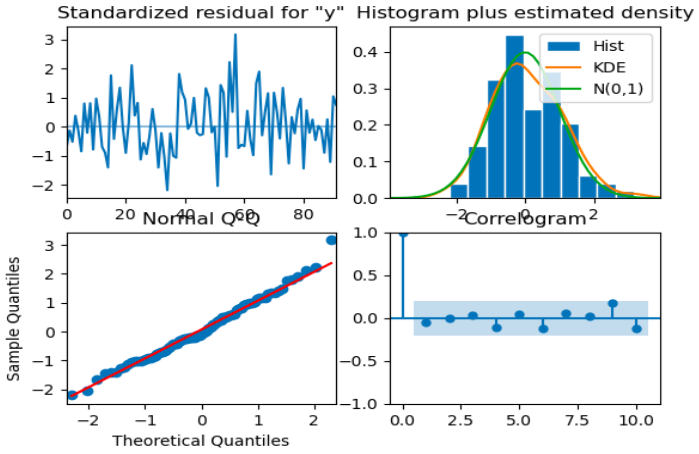
SARIMA:

SARIMA (Seasonal Auto-Regressive Integrated Moving Average) is an extension of the **ARIMA (Autoregressive Integrated Moving Average)** model that incorporates *seasonality* in addition to the non-seasonal components.

Rose Dataset:

The Akaike Information Criterion (AIC) values were computed for various models, and the model with the lowest AIC value was chosen. The optimal SARIMA model identified was (3,1,1)(3,0,2,12), and the summary report for this model indicates a root mean square error (RSME) of 18.5358.

	param	seasonal	AIC
222	(3, 1, 1)	(3, 0, 2, 12)	774.400287
238	(3, 1, 2)	(3, 0, 2, 12)	774.880934
220	(3, 1, 1)	(3, 0, 0, 12)	775.426699
221	(3, 1, 1)	(3, 0, 1, 12)	775.49533
252	(3, 1, 3)	(3, 0, 0, 12)	775.561018



SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)	Log Likelihood	-377.200
Date:	Wed, 17 Jan 2024	AIC	774.400
Time:	12:27:45	BIC	799.618
Sample:	0	HQIC	784.578
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0464	0.126	0.367	0.714	-0.201	0.294
ar.L2	-0.0060	0.120	-0.050	0.960	-0.241	0.229
ar.L3	-0.1808	0.098	-1.838	0.066	-0.374	0.012
ma.L1	-0.9370	0.067	-13.905	0.000	-1.069	-0.805
ar.S.L12	0.7639	0.165	4.640	0.000	0.441	1.087
ar.S.L24	0.0841	0.159	0.527	0.598	-0.228	0.397
ar.S.L36	0.0727	0.095	0.764	0.445	-0.114	0.259
ma.S.L12	-0.4969	0.250	-1.988	0.047	-0.987	-0.007
ma.S.L24	-0.2191	0.210	-1.044	0.296	-0.630	0.192
sigma2	192.1402	39.625	4.849	0.000	114.477	269.803

Ljung-Box (L1) (Q):	0.30	Jarque-Bera (JB):	1.64
Prob(Q):	0.58	Prob(JB):	0.44
Heteroskedasticity (H):	1.11	Skew:	0.33
Prob(H) (two-sided):	0.77	Kurtosis:	3.03

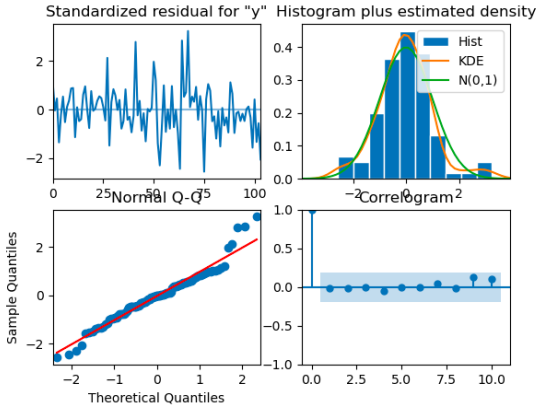
(3,1,1),(3,0,2,12),Auto_SARIMA 18.535865

The following graphs depict the residuals for the best auto-tuned SARIMA model. The diagnostic plots, especially residual plots, are crucial for assessing whether a time series model, such as SARIMA, meets certain assumptions.

Sparkling Dataset:

The Akaike Information Criterion (AIC) values were computed for various models, and the model with the lowest AIC value was chosen. The optimal SARIMA model identified was (1,1,2)(1,0,2,12), and the summary report for this model indicates a root mean square error (RSME) of 52860.

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584248
53	(1, 1, 2)	(2, 0, 2, 12)	1555.934563
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121564
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340403



SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Fri, 19 Jan 2024	AIC	1555.584
Time:	13:41:15	BIC	1574.095
Sample:	0	HQIC	1563.083
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6281	0.255	-2.463	0.014	-1.128	-0.128
ma.L1	-0.1041	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7276	0.154	-4.734	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.843	0.000	1.016	1.072
ma.S.L12	-0.5551	0.098	-5.663	0.000	-0.747	-0.363
ma.S.L24	-0.1355	0.120	-1.133	0.257	-0.370	0.099
sigma2	1.506e+05	2.03e+04	7.400	0.000	1.11e+05	1.9e+05

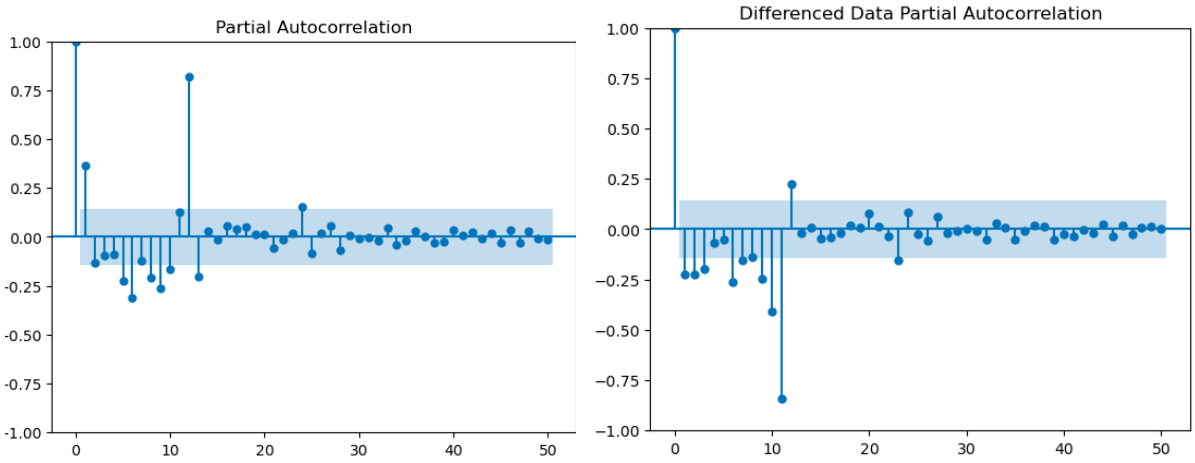
Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	11.72
Prob(Q):	0.84	Prob(JB):	0.00
Heteroskedasticity (H):	1.47	Skew:	0.36
Prob(H) (two-sided):	0.26	Kurtosis:	4.48

(1,1,2),(1,0,2,12),Auto_SARIMA 528.603103

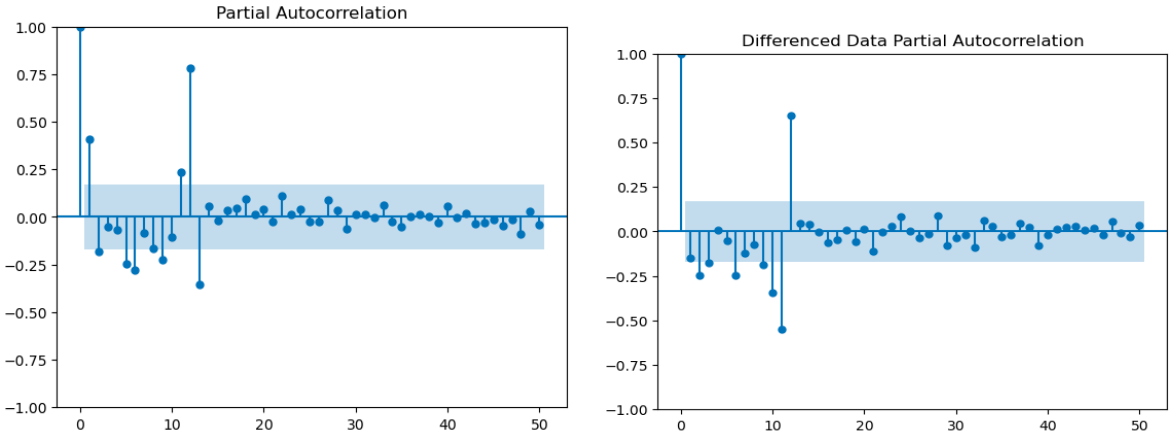
ARIMA and SARIMA models based on the cut-off points of ACF and PACF on the training data:

Sparkling dataset:

PACF the ACF plot on data :



PACF the ACF plot on training data :



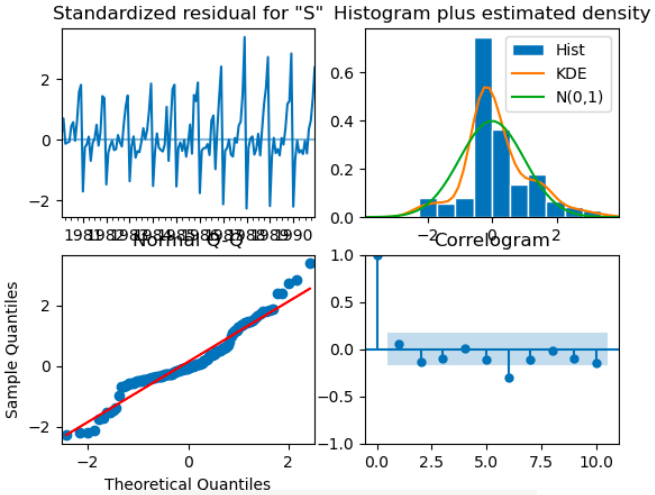
Looking at PACF plot we can again see significant bars till lag 1 for differenced series which is stationary in nature, post 1 the decay is large enough. Hence we choose p value to be 1. i.e. p=1. d values will be 1, since we had seen earlier that the series is stationary with lag1. Hence the values selected for manual ARIMA:- p=1, d=1, q=1 summary from this manual ARIMA model.

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	-1114.878
Date:	Fri, 19 Jan 2024	AIC	2235.755
Time:	15:29:53	BIC	2244.381
Sample:	01-01-1980	HQIC	2239.260
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4494	0.043	10.366	0.000	0.364	0.534
ma.L1	-0.9996	0.102	-9.811	0.000	-1.199	-0.800
sigma2	1.401e+06	7.57e-08	1.85e+13	0.000	1.4e+06	1.4e+06

Ljung-Box (L1) (Q):	0.50	Jarque-Bera (JB):	10.42
Prob(Q):	0.48	Prob(JB):	0.01
Heteroskedasticity (H):	2.64	Skew:	0.46
Prob(H) (two-sided):	0.00	Kurtosis:	4.03



ARIMA(3,1,3) 1319.936734

Model Evaluation: RSME - 1319.9367298218867

SARIMA:

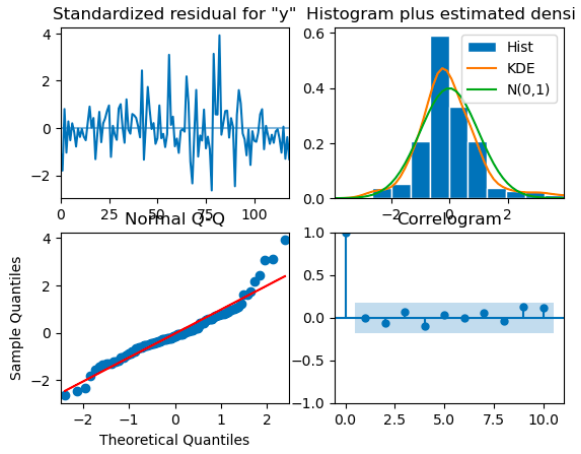
SARIMAX(1, 1, 1)x(1, 1, 1, 12)
Below is the summary of the manual SARIMA model

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	Log Likelihood	-882.088
Date:	Fri, 19 Jan 2024	AIC	1774.175
Time:	15:29:55	BIC	1788.071
Sample:	0	HQIC	1779.818
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1957	0.104	1.878	0.060	-0.009	0.400
ma.L1	-0.9404	0.053	-17.897	0.000	-1.043	-0.837
ar.S.L12	0.0711	0.242	0.294	0.769	-0.404	0.546
ma.S.L12	-0.5035	0.221	-2.277	0.023	-0.937	-0.070
sigma2	1.51e+05	1.33e+04	11.371	0.000	1.25e+05	1.77e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	45.66
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	2.61	Skew:	0.82
Prob(H) (two-sided):	0.00	Kurtosis:	5.56



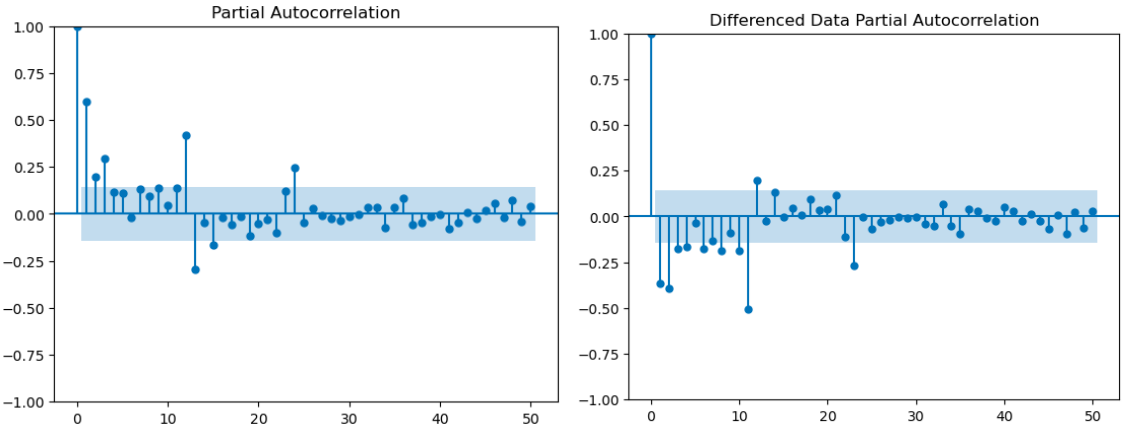
(2,1,2)(2,1,2,12),Manual_SARIMA	359.612453

Model Evaluation: RSME-359.612454

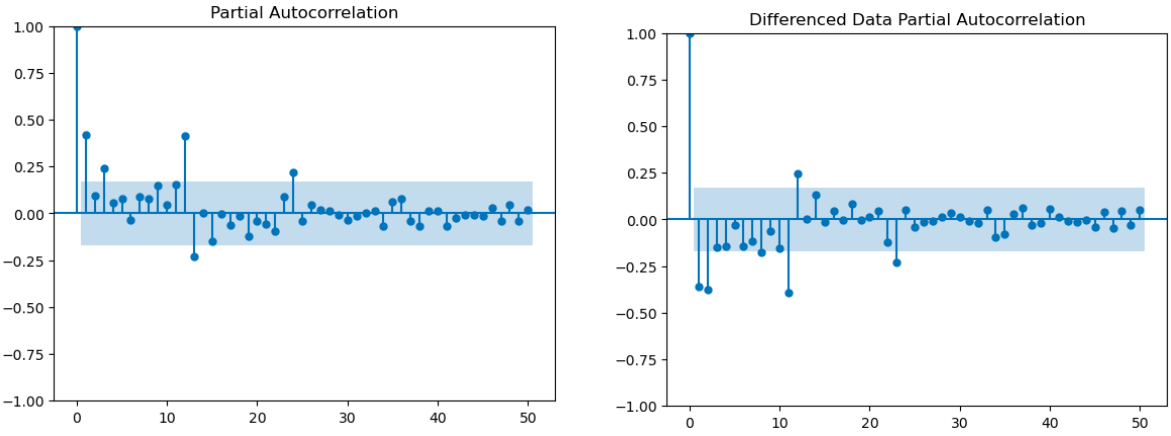
ARIMA and SARIMA models based on the cut-off points of ACF and PACF on the training data:

Rose Dataset:

PACF and ACF plot on data :



PACF and ACF plot on training data :



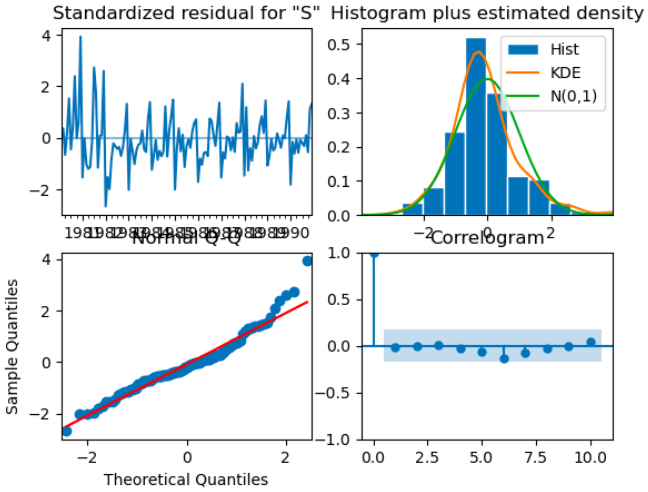
Therefore, the chosen values for the manual ARIMA model are p=2, d=1, q=2. Here is a summary of the manual ARIMA model with these parameter values.

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Wed, 17 Jan 2024	AIC	1281.871
Time:	12:27:48	BIC	1296.247
Sample:	01-01-1980 - 12-01-1990	HQIC	1287.712
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4540	0.469	-0.969	0.333	-1.372	0.464
ar.L2	0.0001	0.170	0.001	0.999	-0.334	0.334
ma.L1	-0.2541	0.459	-0.554	0.580	-1.154	0.646
ma.L2	-0.5984	0.430	-1.390	0.164	-1.442	0.245
sigma2	952.1601	91.424	10.415	0.000	772.973	1131.347

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	34.16
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.79
Prob(H) (two-sided):	0.00	Kurtosis:	4.94



ARIMA(3,1,3) 36.473225

The model evaluation using the Root Mean Square Error (RMSE) yields a value of 36.47.

SARIMA:

Upon examining the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots for the training data, prominent spikes are observed at lags 12, 24, 36, 48, etc., indicating a seasonality of 12. Consequently, the parameters selected for the manual Seasonal Autoregressive Integrated Moving Average (SARIMA) model are SARIMAX(2, 1, 2)x(2, 1, 2, 12).

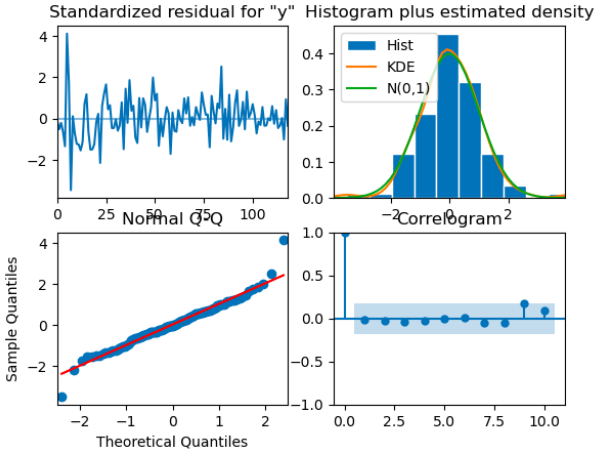
Here is a summary of the manual SARIMA model:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-538.016
Date:	Wed, 17 Jan 2024	AIC	1094.031
Time:	12:27:53	BIC	1119.044
Sample:	0	HQIC	1104.188
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5491	0.228	-2.408	0.016	-0.996	-0.102
ar.L2	-0.0744	0.099	-0.753	0.452	-0.268	0.119
ma.L1	-0.1703	0.216	-0.787	0.431	-0.595	0.254
ma.L2	-0.6693	0.228	-2.936	0.003	-1.116	-0.222
ar.S.L12	-1.0135	0.524	-1.936	0.053	-2.040	0.013
ar.S.L24	-0.1002	0.175	-0.572	0.567	-0.444	0.243
ma.S.L12	0.2914	85.793	0.003	0.997	-167.859	168.442
ma.S.L24	-0.7081	60.859	-0.012	0.991	-119.990	118.574
sigma2	430.1842	3.67e+04	0.012	0.991	-7.15e+04	7.24e+04

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	27.16
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	0.33	Skew:	0.26
Prob(H) (two-sided):	0.00	Kurtosis:	5.28



(2,1,2)(2,1,2,12),Manual_SARIMA 14.975316

The RMSE is 14.975041301618377.

Visualization of the RMSE:

To systematically organize and present the information regarding the models built, their corresponding parameters, and the associated Root Mean Square Error (RMSE) values on the test data, we can create a table or a data frame.
This table serves as a comprehensive summary, allowing for a clear comparison of different model configurations and their performance metrics.

Rose Dataset:

	Test RMSE
Alpha=0.2,Beta=0.7,Gamma=0.2,TripleExponentialSmoothing	8.992350
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
(2,1,2)(2,1,2,12),Manual_SARIMA	14.975316
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535865
Auto_ARIMA	36.418927
Alpha=0.1,SimpleExponentialSmoothing	36.429535
ARIMA(3,1,3)	36.473225
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Linear Regression	51.080941
Simple Average Model	53.049755
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	62.560681
Naive Model	79.304391

Sparkling Dataset:

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing	317.434302
(2,1,2)(2,1,2,12),Manual_SARIMA	359.612453
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612453
(1,1,1),(2,0,3,12),Auto_SARIMA	528.603103
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
Simple Average Model	1275.081804
Linear Regression	1275.867052
6pointTrailingMovingAverage	1283.927428
Auto_ARIMA	1299.982793
ARIMA(3,1,3)	1319.936734
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	2007.238526
Naive Model	3864.279352

Optimum Model:

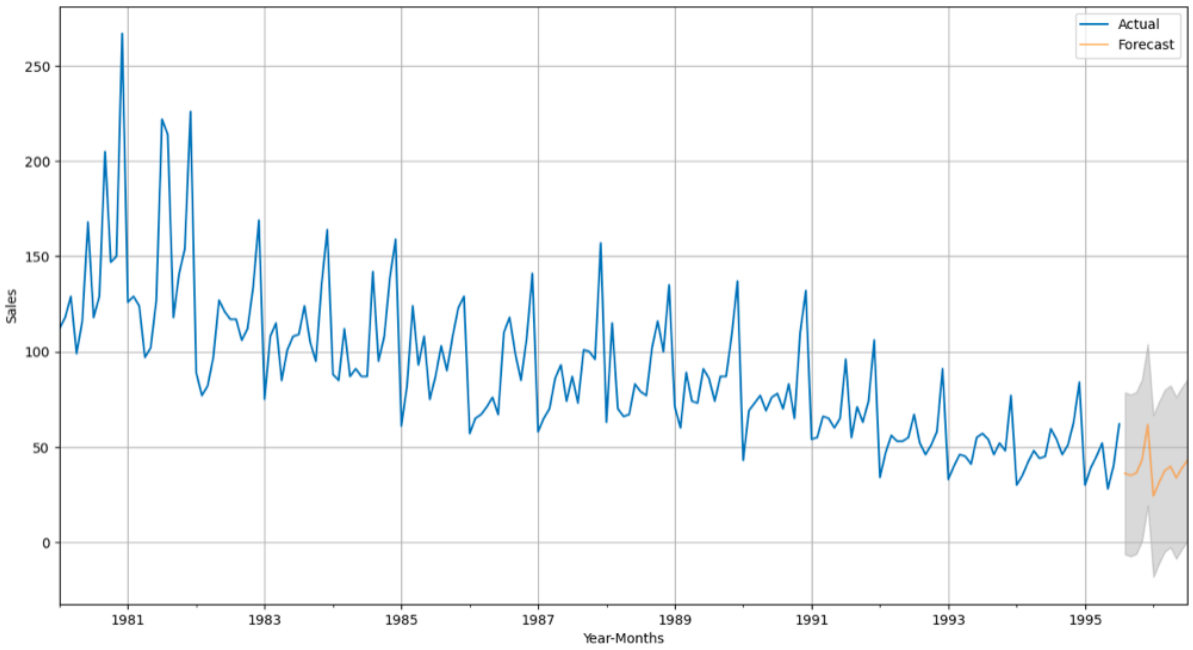
Rose Dataset:

After comparing the performance of all the models constructed, it is evident that the triple exponential smoothing, also known as the Holt-Winters model, consistently yields the lowest Root Mean Square Error (RMSE). Therefore, based on this evaluation, it can be concluded that the Holt-Winters model is the most optimal choice among the considered models for predicting the observed time series data.

Let us predict 12 months into the future with appropriate confidence intervals/bands.

	Sales_Predictions
1995-08-01	36.096841
1995-09-01	34.999961
1995-10-01	36.289937
1995-11-01	43.126839
1995-12-01	61.593978
1996-01-01	24.293852
1996-02-01	31.406019
1996-03-01	37.545514
1996-04-01	39.735393
1996-05-01	33.753457
1996-06-01	38.868148
1996-07-01	43.093112

Now, let's visualize the trend and seasonality of the predicted values….



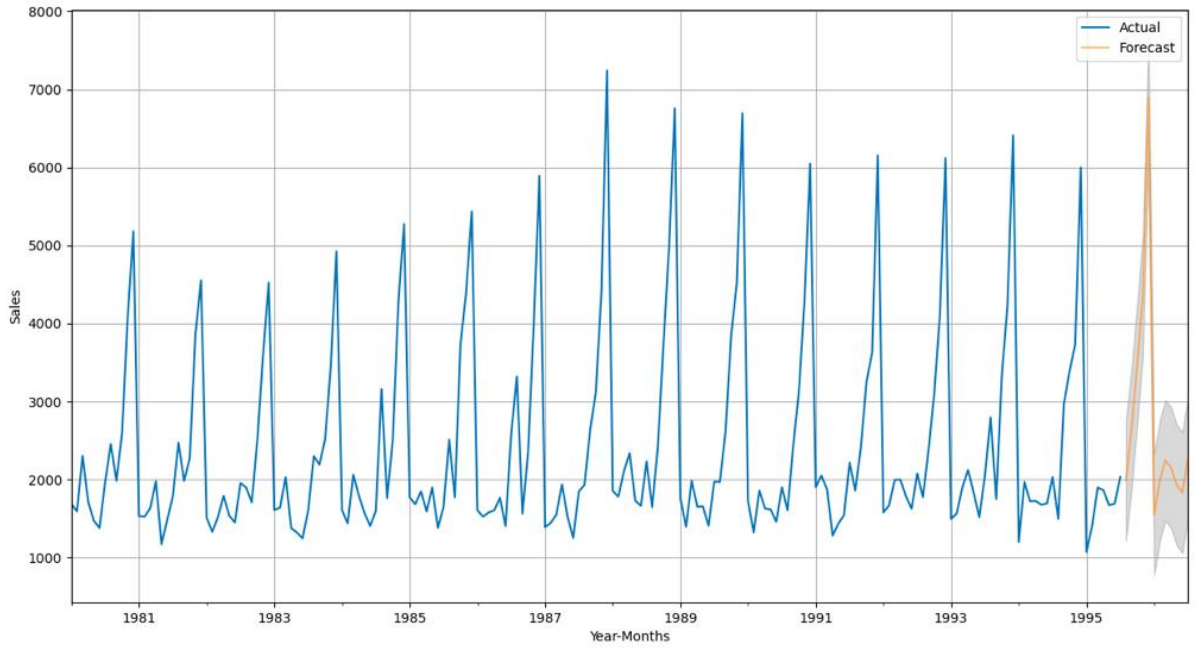
Sparkling Dataset:

After analyzing and comparing the performance of the various models developed, it is evident that the triple exponential smoothing, specifically the Holt-Winters model, consistently exhibits the lowest Root Mean Square Error (RMSE). Consequently, it can be concluded that the Holt-Winters model is the most optimal choice among the considered models for accurately predicting the observed time series data.

Let us predict 12 months into the future with appropriate confidence intervals/bands.

	Sales_Predictions
1995-08-01	1988.782193
1995-09-01	2652.762887
1995-10-01	3483.872246
1995-11-01	4354.989747
1995-12-01	6900.103171
1996-01-01	1546.800546
1996-02-01	1981.361768
1996-03-01	2245.459724
1996-04-01	2151.066942
1996-05-01	1929.355815
1996-06-01	1830.619260
1996-07-01	2272.156151

Now, let's visualize the trend and seasonality of the predicted values….



Conclusion:

The strengths of the Holt-Winters model lie in its ability to capture trends, seasonality, and variations in the data, making it well-suited for time series forecasting. Additionally, the triple exponential smoothing approach enables the model to adapt to changes in the underlying patterns over time.

Rose Dataset:

The analysis of the wine sales data reveals several key insights that should guide the company's strategic decisions:

1. Clear Downward Trend for Rose Wine:
   - The data indicates a distinct downward trend for the Rose wine variety, persisting for more than a decade. This decline in popularity is expected to continue in the future, as projected by the most optimal forecasting model.

2. Seasonal Influence on Wine Sales:
   - Wine sales exhibit a strong correlation with seasonal changes, with an increase during festival seasons and a decline in peak winter months, particularly in January.

3. Campaign Strategy Recommendations:
   - To address the declining popularity of Rose wine, the company should consider running targeted campaigns to boost consumption during the lean period (April to June), when sales are typically subdued. Focusing efforts during this time could have a maximum impact on overall annual performance.

   - Campaigns during peak periods, such as festivals, may not generate significant additional impact, as sales are already high during these times. Moreover, running campaigns during peak winter (January) is not recommended due to lower consumer propensity to purchase wine during this period.

4. Investigate Decline in Popularity:
   - The company should conduct a thorough investigation into the reasons behind the decline in the popularity of the Rose wine variety. This may involve revisiting production and marketing strategies to revamp and regain market share.

In summary, the company should strategize its marketing campaigns based on the identified trends and seasonality in wine sales. Focusing on boosting sales during the lean period, understanding the causes of the decline in Rose wine popularity, and refining strategies accordingly will be essential for maintaining a competitive edge in the market.

Sparkling Dataset:

In summary, the analysis of Sparkling wine sales for the company provides the following insights:

1. Positive Sales Outlook:
   - The predictive model suggests that Sparkling wine sales are expected to at least match, if not surpass, the figures from the previous year. There is potential for peak sales in the upcoming year to be higher than the current year.

2. Consistent Popularity:
   - Sparkling wine has maintained a consistent level of popularity among customers. Despite a marginal decline, it has remained a preferred choice, especially considering its peak popularity in the late 1980s.

3. Seasonal Impact on Sales:
   - Seasonality significantly influences Sparkling wine sales, with a slow period in the first half of the year followed by a pickup from August to December.

4. Campaign Recommendations:
   - Given the slow sales in the first half of the year, it is recommended for the company to run targeted campaigns during this period, particularly in the months of March to July. These campaigns could focus on promoting Sparkling wine and increasing awareness during the slower months.

   - To enhance the impact of promotions, the company may consider creating special offers that pair Sparkling wine with a less popular variety, such as "Rose wine." This strategy aims to encourage customers to try the underperforming wine, potentially leading to increased sales.

Overall, leveraging the consistent popularity of Sparkling wine and strategically planning campaigns during slower periods can contribute to sustained or increased sales. Additionally, exploring innovative promotional pairings may positively impact the overall performance of the company's wine varieties.