Time Series Forecasting – Sparkling wine Coded Project

DSBA – Course

Business Report

Created by – Rishabh Gupta

Foreword

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.

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.

Contents

List of Tables

List of Figures

Objective

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

For this assignment, we will analyze data on different types of wine sales from the 20th century. Both datasets come from the same company but represent different wine varieties. As an analyst at ABC Estate Wines, our task is to analyze and forecast wine sales during this period.

We are going to individually perform the following tasks on each of the two datasets.

Data Analysis of problem statement –

The dataset contains data corresponding to polished and unpolished stones.

Sheet name –

1. *Sparkling.csv*
2. *Rose.csv*

1. Sparkling Wines Sales

Data Dictionary –

1. **YearMonth:** displays time for sales
2. **Sparkling:** No of wines sales of this type

Data Overview –

To start the data analysis, we need to import the necessary libraries, specify the working directory, and load the dataset. Then, we will view the first five rows using head () function to get a preliminary understanding of the data. The Dataset has 187 number of rows with 2 columns.

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

TABLE 1 - TOP 5 ROWS OF DATASET

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   YearMonth   187 non-null    object
 1   Sparkling   187 non-null    int64 
dtypes: int64(1), object(1)
memory usage: 3.1+ KB
```

TABLE 2 - BASIC INFO. OF DATASET

Regarding the datatype of the columns in the dataset, we can confirm that-

• We can observe that there around 1 numerical datatype and 1 object

• There are no missing values in the dataset

Data Pre-processing-

Missing value treatment and Analysis-

- On analysis, we can observe there are no null values in the dataset.
- Also, there are no duplicate entries.

Statistical Summary –

Using Describe () function, we can analyses the summary statistics of the dataset –

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

TABLE 3 - STATISTICAL SUMMARY OF DATASET

Observations-

- **Broad Sales Range**: The large gap between the minimum sales (1,070) and the maximum sales (7,242) highlights a significant variation in Sparkling wine sales. This fluctuation may be driven by changing demand, seasonal trends, or the introduction of new products and marketing strategies that influenced sales performance.

- **Positive Skewness:** With the mean (2,402.42) noticeably exceeding the median (1,874), the data exhibits a positive skew. This suggests that certain periods experienced exceptionally high sales, raising the overall average. Additionally, the maximum value is much farther from the 75th percentile compared to how the minimum value relates to the 25th percentile, further supporting this skewness.

- **Sales Volatility**: A high standard deviation (1,295.11) in relation to the mean indicates considerable fluctuations in Sparkling wine sales. This reinforces the presence of external factors or demand shifts affecting sales trends over time.

- **Sales Distribution**: Half of the sales figures fall within the range of 1,605 (25th percentile) to 2,549 (75th percentile), indicating that while there are extreme values, a significant portion of sales is concentrated around the median, reflecting typical sales patterns.

- **Potential Outliers**: The maximum sales figure (7,242) is substantially higher than the 75th percentile (2,549), suggesting the presence of potential outliers. These spikes could be attributed to special events, promotional efforts, or possible data anomalies that require further investigation to determine their validity.

Exploratory Data Analysis

A boxplot to understand the spread of sales across different years and within different months across years.

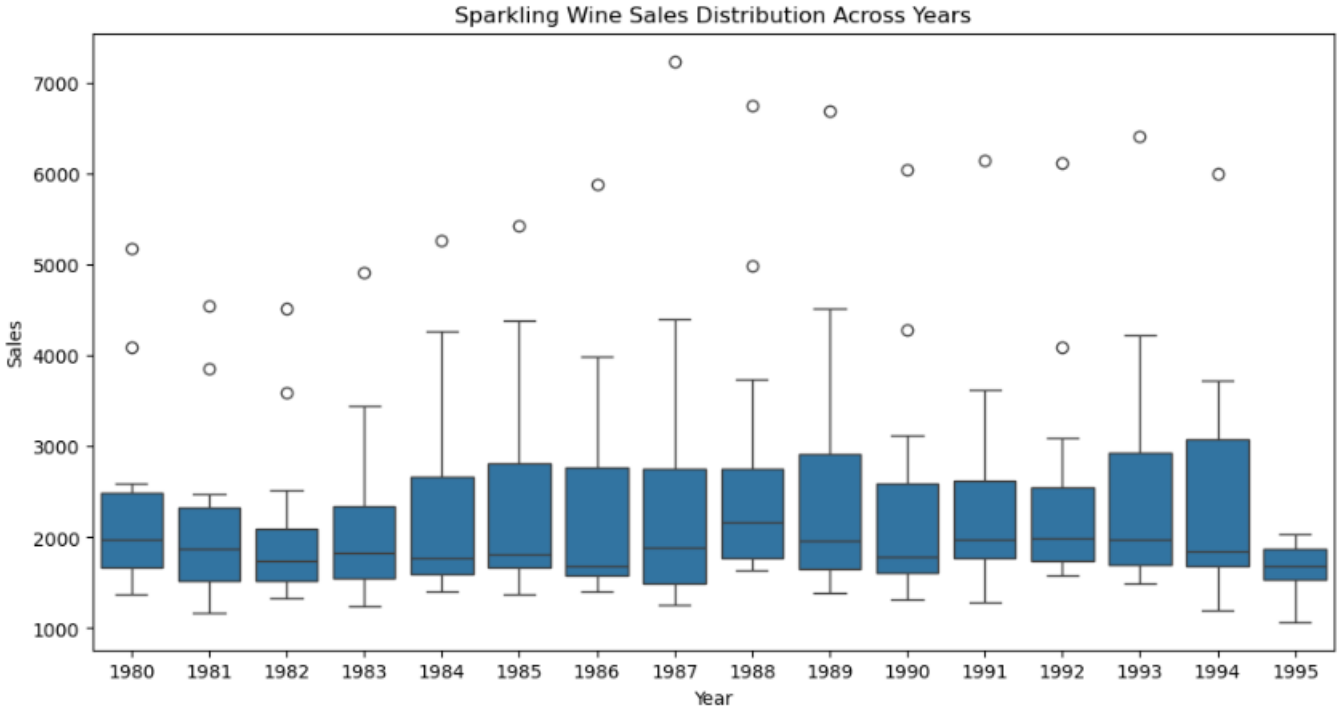

FIGURE 1-BOXPLOT FOR SALES VS YEARS

Observations-

- Stable Median Sales: Median sales remain relatively consistent across years, mostly within the 2000-3000 range, indicating steady typical sales.
- Varying Sales Spread: The Interquartile Range (IQR) fluctuates, showing that some years had more stable sales while others experienced higher variability.
- Outliers Present: Several high-sales outliers indicate occasional spikes in demand, reflecting exceptional sales periods.
- Increasing High Sales Outliers: From 1988 onwards, there is a subtle rise in both the frequency and magnitude of outliers, suggesting growing potential for peak sales.
- 1995 Anomaly: Sales in 1995 show a significantly lower median and compressed IQR, hinting at an unusually weak and consistent sales year, warranting further investigation.

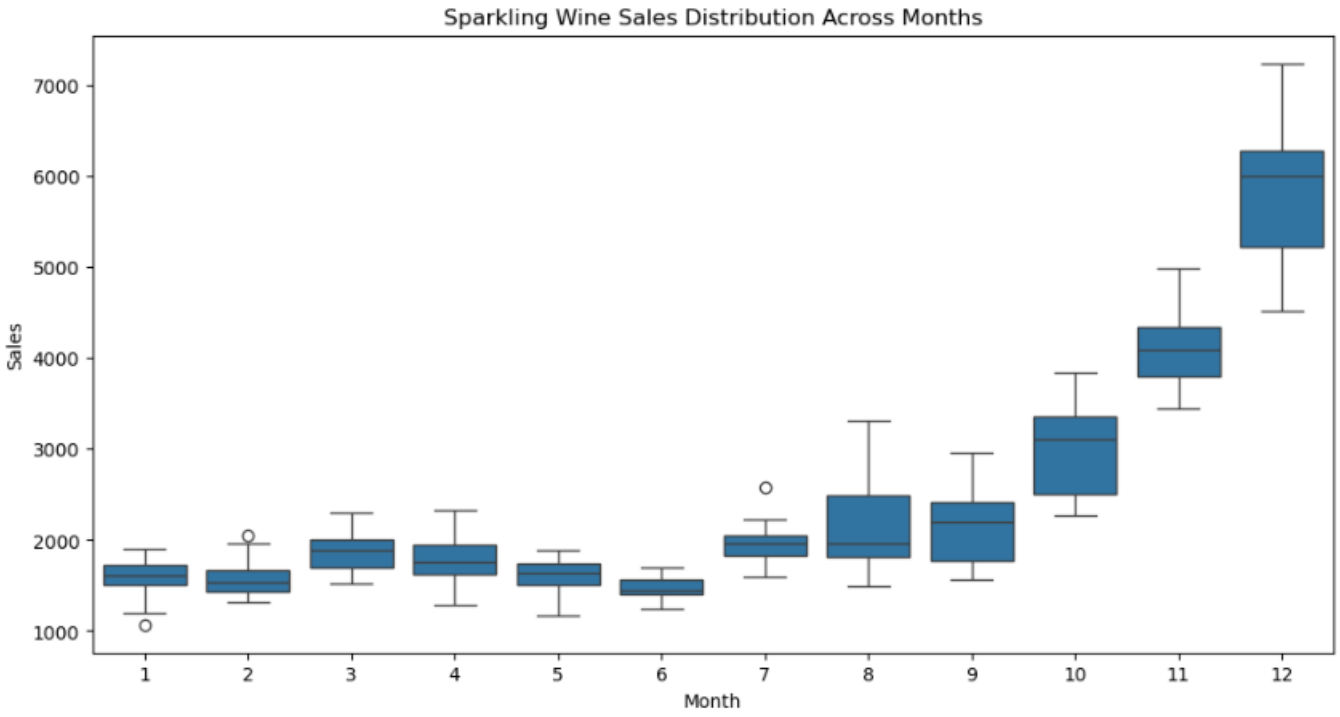

FIGURE 2-BOXPLOT FOR SALES VS YEARS

- Strong Seasonal Trend: Sales are low from January to July and rise sharply in the second half, peaking in December.
- December Peak: Highest median sales, largest spread, and most outliers, likely due to holiday demand.
- Gradual Increase in H2: Sales steadily climb from August to December, reflecting rising demand as the holiday season approaches.
- Stable First Half: Sales remain consistently low with minimal variability from January to June.
- July as a Turning Point: Shows a slight increase, possibly signaling early holiday promotions or shifting demand.
- Outliers Present: December has the most extreme outliers, but occasional spikes appear throughout the year, possibly linked to events or promotions.

Now let’s analyse overall Sparkling wines sales plot –

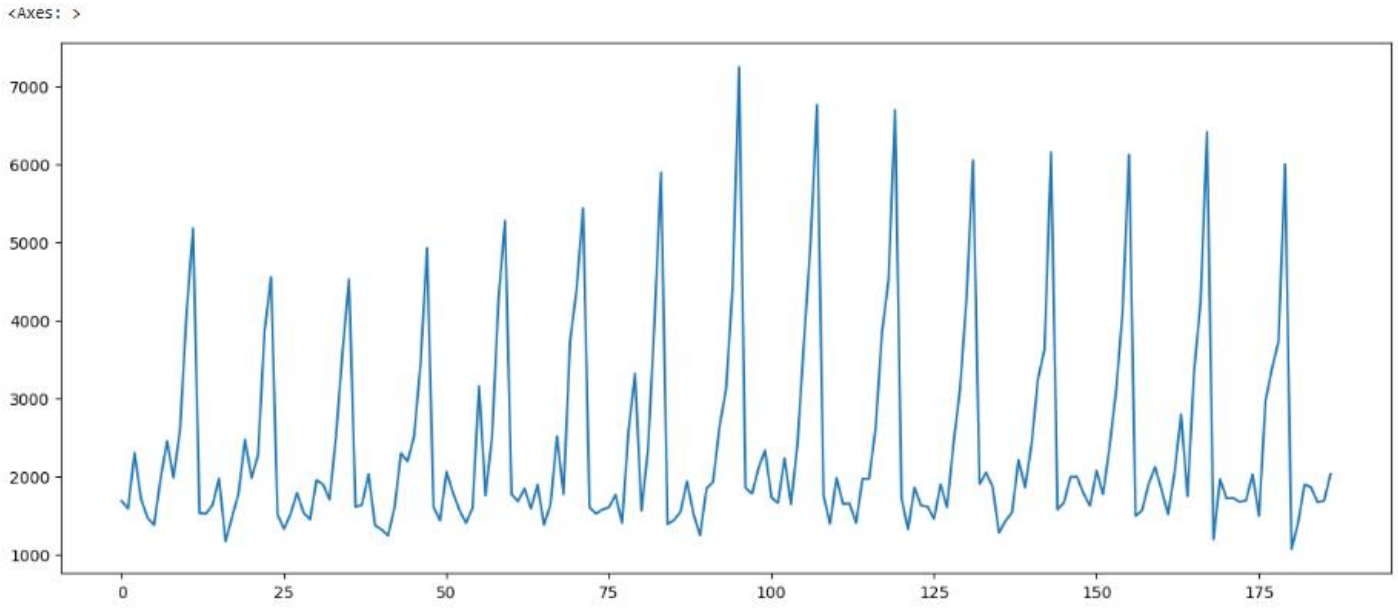

FIGURE 3-TIMES SERIES PLOT FOR SPARKLING SALES

Observations –

- Strong Seasonal Pattern: Clear, repetitive peaks and troughs indicate a strong cyclical influence on sales.
- Regular Peaks: Sales spikes occur at consistent intervals, likely driven by annual events or holidays.
- Consistent Troughs: Low points in sales remain steady, suggesting a stable baseline demand.
- Varying Peak Heights: While peaks are regular, their intensity fluctuates, possibly due to economic conditions or marketing efforts.
- Stable Baseline Sales: The underlying sales level between peaks remains relatively constant over time.
- Potential Outliers: Some peaks are significantly higher than others, warranting further investigation.

- No Clear Trend: Overall sales levels remain stable, with fluctuations primarily driven by seasonal factors.

Now let's analyses Empirical Cumulative Distribution Function (ECDF)

The Empirical Cumulative Distribution Function (ECDF) is a statistical tool that shows the proportion (or percentage) of data points that are less than or equal to a given value. It provides a stepwise, non-parametric estimate of the cumulative distribution of a dataset.

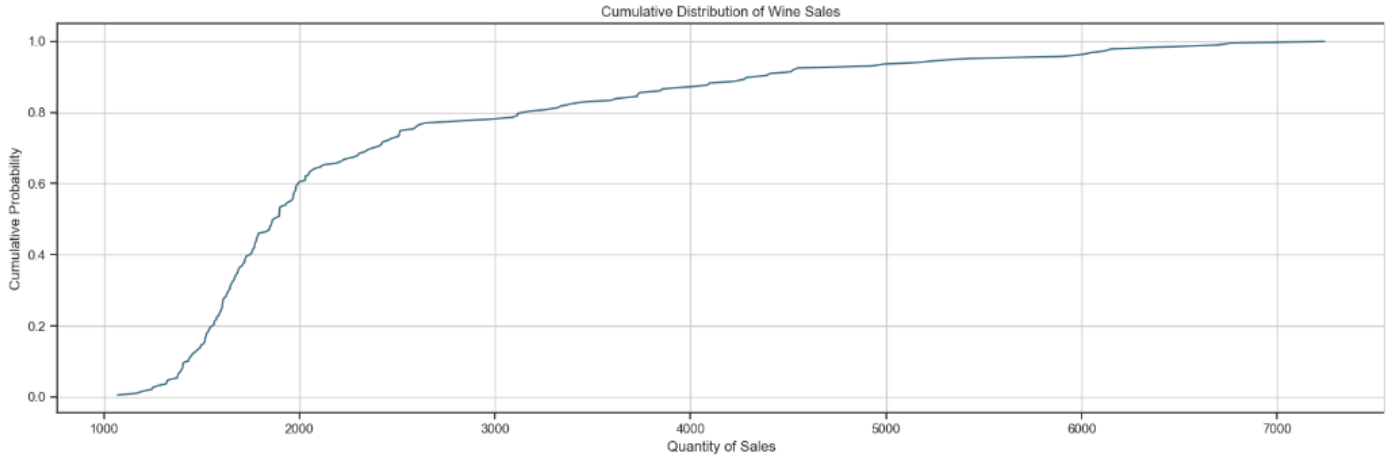

FIGURE 4-ECDF PLOT

- **Steep Initial Rise:** A large portion of sales falls within **1000–2000** units, indicating concentration at lower sales volumes.
- **Flattening at Higher Sales:** Sales above **4000 units** are less frequent, with the ECDF flattening, showing fewer extreme high-sales events.
- **Steps & Plateaus:** The stepwise nature reflects observed sales quantities, while plateaus indicate ranges with no new data points.
- **Median Sales Estimate:** Around **2000–2500 units**, meaning half of all sales are below this quantity.
- **Sales Skewness:** The steep early rise suggests a right-skewed distribution, common in sales data.

- **Percentile Estimation:** The ECDF allows quick estimation of any percentile (e.g., 90% of sales fall below a certain quantity).
- **Comparing Distributions:** ECDFs for different wine types or time periods can reveal variations in sales patterns.
- **Fit to Theoretical Distributions:** Comparing ECDF to normal/log-normal distributions or using statistical tests (e.g., Kolmogorov-Smirnov) can assess distribution fit.

Now let's analyses Line plot and Heatmap depicting monthly sales over years -

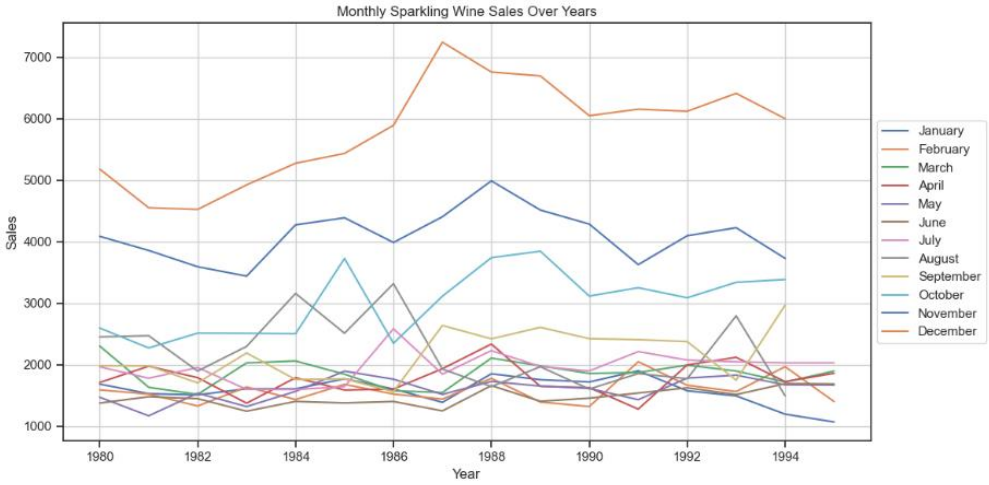

FIGURE 5 – MONTHLY SALES LINE PLOT

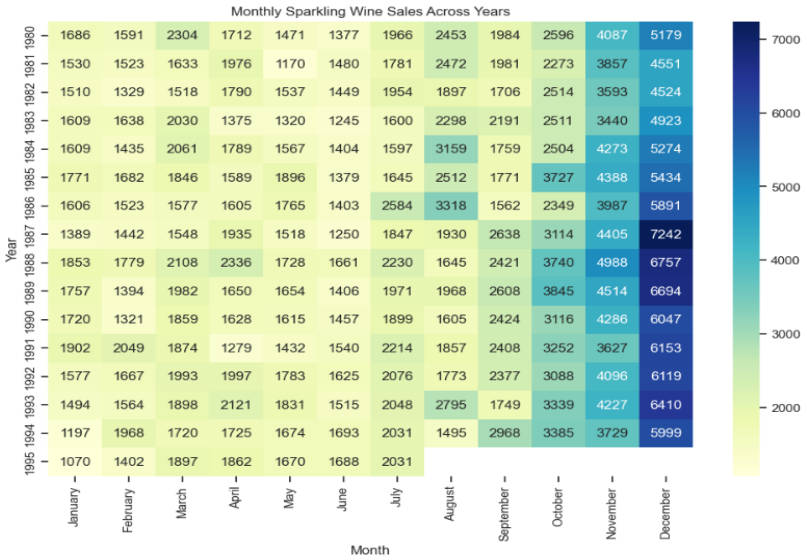

Figure 6 – Monthly sales Heat Map

Observations-

- **Strong Seasonal Trend:** Sales follow a clear seasonal cycle, with peaks and troughs repeating annually.
- **December Peak:** December consistently records the highest sales, driven by holiday celebrations.
- **November Surge:** Sales rise sharply in November, indicating pre-holiday stocking.
- **Lower Demand in Early Months:** Sales remain relatively low and stable from January to June.
- **July Transition:** A slight increase in July suggests early holiday preparations.
- **Varying Peak Heights:** December peaks differ across years, hinting at external influences like economic conditions and marketing.
- **Stable Overall Pattern:** No clear long-term trend, suggesting a mature and predictable market.
- **Holiday Influence:** December and November spikes align with festive buying behavior.
- **Promotional Effectiveness:** Analysing sales in relation to campaigns can optimize marketing strategies.
- **Inventory Planning:** Stocking strategies should align with seasonal demand patterns

Now let's analyses correlation matrix -

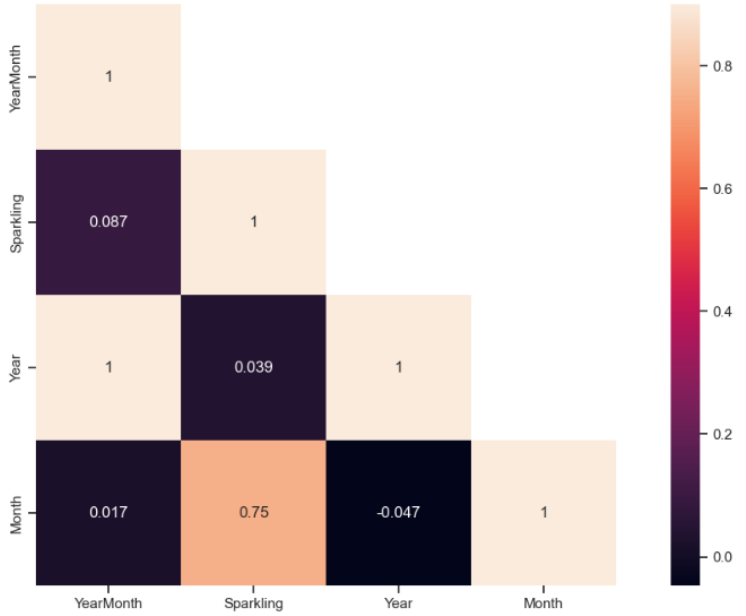

FIGURE 7 – CORRELATION MATRIX

- **Perfect Autocorrelation (1.0):** The diagonal shows a perfect correlation between variables and themselves, which is expected.
- **Strong Positive Correlation (0.75) – Month & Sparkling Sales:** Sales are highly influenced by the month, confirming strong seasonality, especially during the holiday season.
- **Weak Positive Correlation (0.087) – YearMonth & Sparkling Sales:** A slight upward trend over time, but not a strong predictor of sales.
- **Very Weak Positive Correlation (0.039) – Year & Sparkling Sales:** Almost no impact of the year itself on sales, reinforcing that other factors (e.g., seasonality) drive sales.
- **Very Weak Positive Correlation (0.017) – YearMonth & Month:** Likely due to the repeating cyclical nature of months in the dataset.
- **Weak Negative Correlation (-0.047) – Year & Month:** Likely an artifact of the data structure rather than a meaningful relationship.
- **Seasonality is the Primary Driver:** Month strongly influences sales, reinforcing the importance of seasonal demand.

- **Minimal Long-Term Growth Trend:** Year-over-year sales changes are weak, suggesting external factors (marketing, economy) play a bigger role.
- **Month-Based Analysis is Crucial:** Forecasting and planning should focus on monthly trends rather than yearly trends.

Decomposition of a Time Series

Time series decomposition helps break down a time series into its fundamental components, making it easier to analyze patterns, trends, and irregularities. It typically splits the data into three key components:

- **Trend Component**
  - Shows the long-term direction of the data (upward, downward, or stable).
  - Helps identify if sales, demand, or any metric is increasing or decreasing over time.
- **Seasonal Component**
  - Captures repeating patterns at regular intervals (e.g., monthly, quarterly, yearly).
  - Useful for identifying seasonal fluctuations, like higher sparkling wine sales in December.
- **Residual (Irregular) Component**
  - Represents random noise or unexplained variation in the data.
  - Helps detect anomalies, outliers, or events impacting sales.

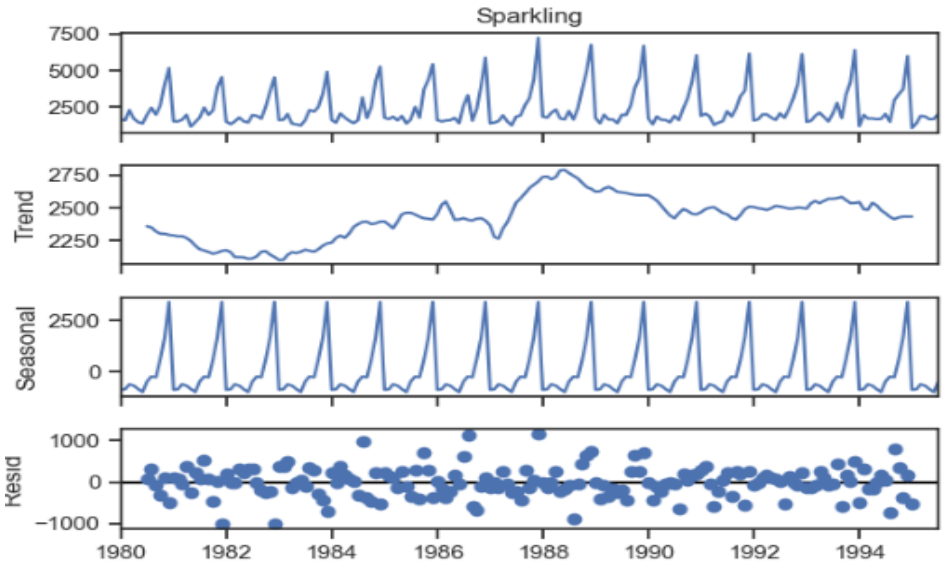

Figure 8 – Decomposition time series additive

Additive Seasonal Decomposition of Time Series (STL) – Observations: -

1. Observed (Top Panel):

- Displays the actual sales data, showing clear seasonal peaks and troughs.
- Yearly fluctuations are visible, but the overall pattern remains consistent.

2. Trend (Second Panel):

- Represents the long-term movement of sales, smoothing out seasonal effects.
- Relatively stable trend with a slight upward tendency in later years.
- A noticeable dip in sales appears around the late 1980s.

3. Seasonal (Third Panel):

- Clearly isolates the repeating seasonal pattern in sales.
- Consistent peaks and troughs confirm strong seasonality.
- The magnitude of seasonal fluctuations remains stable over time.

4. Residual (Bottom Panel):

- Captures the irregular (unexplained) variations after removing trend and seasonality.
- Residuals appear randomly distributed around zero, indicating no major unexplained patterns.
- Any systematic patterns in residuals may suggest additional influencing factors.

Also,

- **Strong Seasonality** – Sales follow a predictable annual pattern, peaking consistently
- **Stable Trend** – Long-term sales remain steady with minor fluctuations.
- **Residual Randomness** – No major patterns in residuals, indicating good decomposition.

Multiplicative Seasonal Decomposition of Time Series (STL) – Observations: -

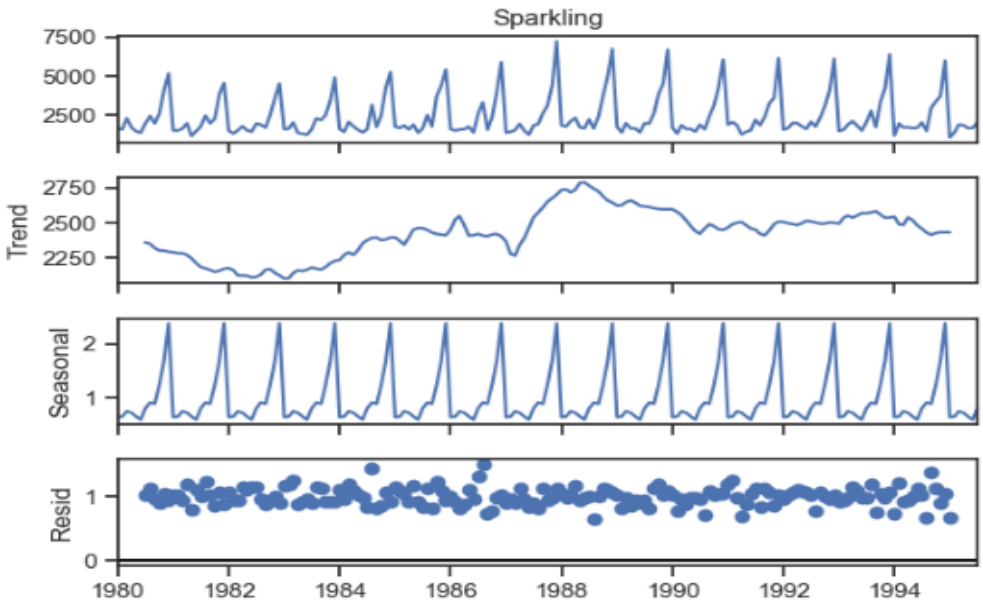

Figure 9 – Decomposition time series Multiplicative

1. Observed (Top Panel):

- Displays the original time series with visible seasonal peaks and trends.

2. Trend (Second Panel):

- Represents the overall level of sales, around which seasonal fluctuations occur.
- Shows the long-term direction of sales, acting as a multiplying factor for seasonality.

3. Seasonal (Third Panel):

- Expressed as a ratio:
    - **Above 1** → Sales are higher than the trend.
    - **Below 1** → Sales are lower than the trend.
    - **Close to 1** → Sales are near the expected level.
- Example: A value of **1.2 in December** means December sales are typically **20% above trend**.

4. Residual (Bottom Panel):

- Represents random fluctuations as ratios relative to expected values.

- Ideally, should vary randomly around **1**, indicating a well-fitted model.
- **Log Transformation** – Helps stabilize variance if sales fluctuations grow with time.
- **Easy Interpretation** – Seasonal components can be read as percentage changes from trend.

Test Train Split

We have split the data into training and testing sets.

Lets have a view to top 10 rows from training set and test set –

First few rows of Training Data

YearMonth	Sparkling	Year	Month
1980-01-01	1686	1980	1
1980-02-01	1591	1980	2
1980-03-01	2304	1980	3
1980-04-01	1712	1980	4
1980-05-01	1471	1980	5
1980-06-01	1377	1980	6
1980-07-01	1966	1980	7
1980-08-01	2453	1980	8
1980-09-01	1984	1980	9
1980-10-01	2596	1980	10

Last few rows of Training Data

TABLE 4 – TRAIN SET

First few rows of Test Data

YearMonth	Sparkling	Year	Month
1991-01-01	1902	1991	1
1991-02-01	2049	1991	2
1991-03-01	1874	1991	3
1991-04-01	1279	1991	4
1991-05-01	1432	1991	5
1991-06-01	1540	1991	6
1991-07-01	2214	1991	7
1991-08-01	1857	1991	8
1991-09-01	2408	1991	9
1991-10-01	3252	1991	10

Last few rows of Test Data

TABLE 5 – TEST SET

```
---------------------------------------
Number of observations in Train data   :   (132, 3)
Number of observations in Test data    :   (55, 3)
Total Observations                     :   187
---------------------------------------
```

Time Series Plot for Train vs Test set

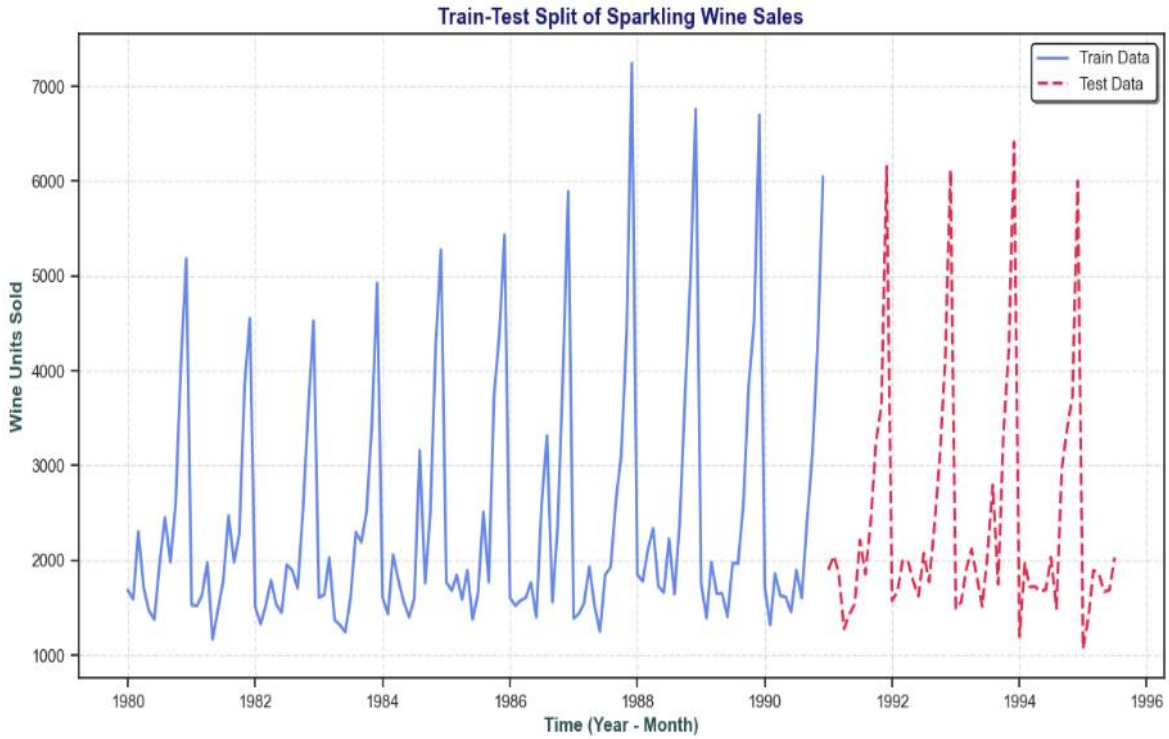

Figure 10 – Train – Test split Plot

- **Chronological Train-Test Split:** The training set (blue) covers the earlier period, while the test set (red, dashed) represents the later period, ensuring a realistic forecasting approach.
- **Clear Separation:** A distinct transition between train and test sets makes it easy to visualize the split and understand model evaluation timing.
- **Test Set Duration:** The test set spans multiple seasonal cycles, allowing for a meaningful evaluation of how well the model captures seasonality.
- **Consistent Seasonal Patterns:** Both training and test sets exhibit recurring peaks and troughs, confirming that the seasonal nature of Sparkling Wine sales persists across time.
- **Balanced Evaluation Opportunity:** The test set includes both high and low sales periods, providing a robust way to assess forecast accuracy across different demand levels.

- **Train-Test Ratio:** The exact split (e.g., 80/20, 70/30) should be chosen carefully to balance model learning and evaluation.
- **Model Selection:** Depending on trend and seasonality, ARIMA, Exponential Smoothing, or ML models with time-series features can be explored.
- **Performance Metrics:** Use RMSE, MAE, or MAPE to evaluate forecast accuracy effectively.
- **Time-Series Cross-Validation:** Techniques like rolling or expanding windows can provide a more robust assessment with limited data.

Model Building

Model 1: Linear Regression

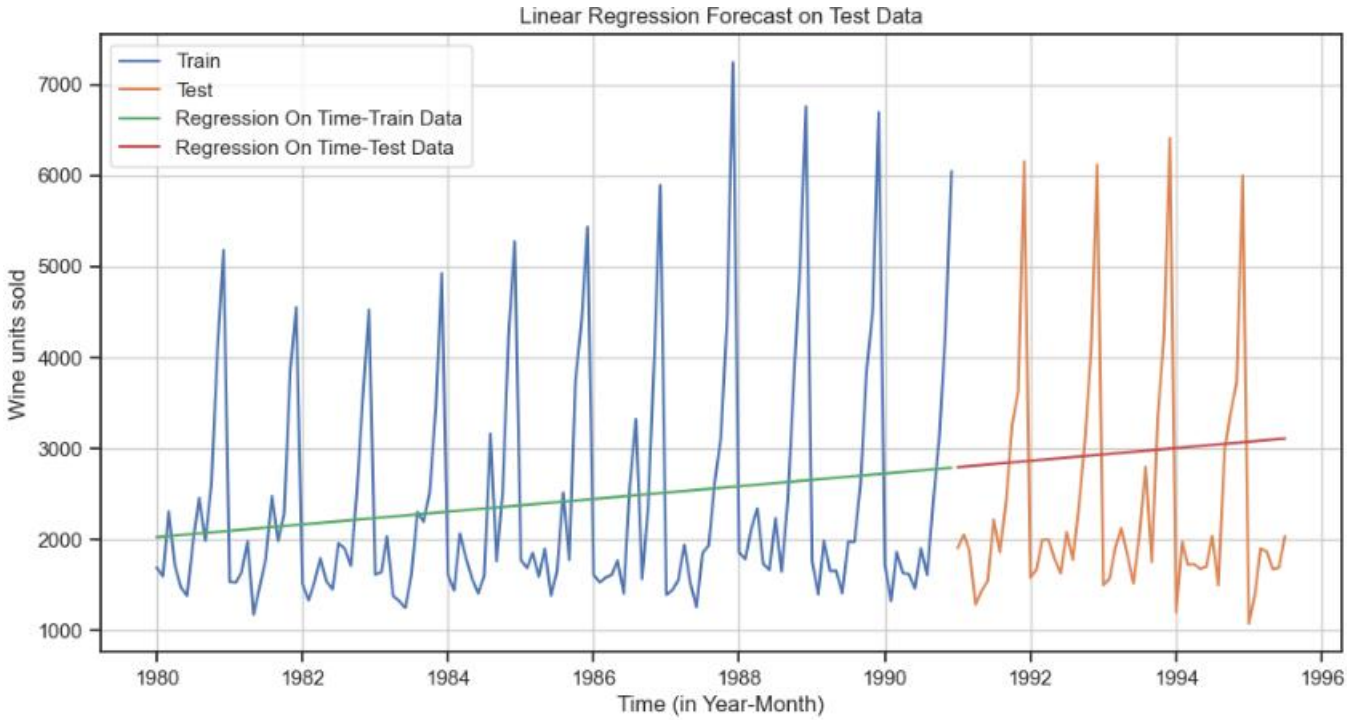

Figure 11 – Linear regression model – Test data

Observations-

- Fails to Capture Seasonality:
    - Sales data has clear peaks and troughs, but the model predicts a straight-line trend.
    - Ignores the cyclical nature of Sparkling Wine sales.
- Limited Predictive Power:
    - Forecasts (red line) miss the real variations in sales.
    - Predictions do not align with actual demand fluctuations.
- Trend Capture but Overshadowed:
    - The model detects a slight upward trend.
    - However, the seasonal effect is much stronger, making the linear trend alone insufficient.
- Oversimplification of Data:
    - Linear regression assumes a constant relationship over time.
    - Not a suitable approach when strong seasonality is present.

Model evaluation –

As per code, For RegressionOnTime forecast on the Test Data, RMSE is 1389.135

	Test RMSE
Linear Regression	1389.135175

TABLE 6 – LINEAR MODEL RMSE

Model 2: Simple Average

For this particular simple average method, we will forecast by using the average of the training values.

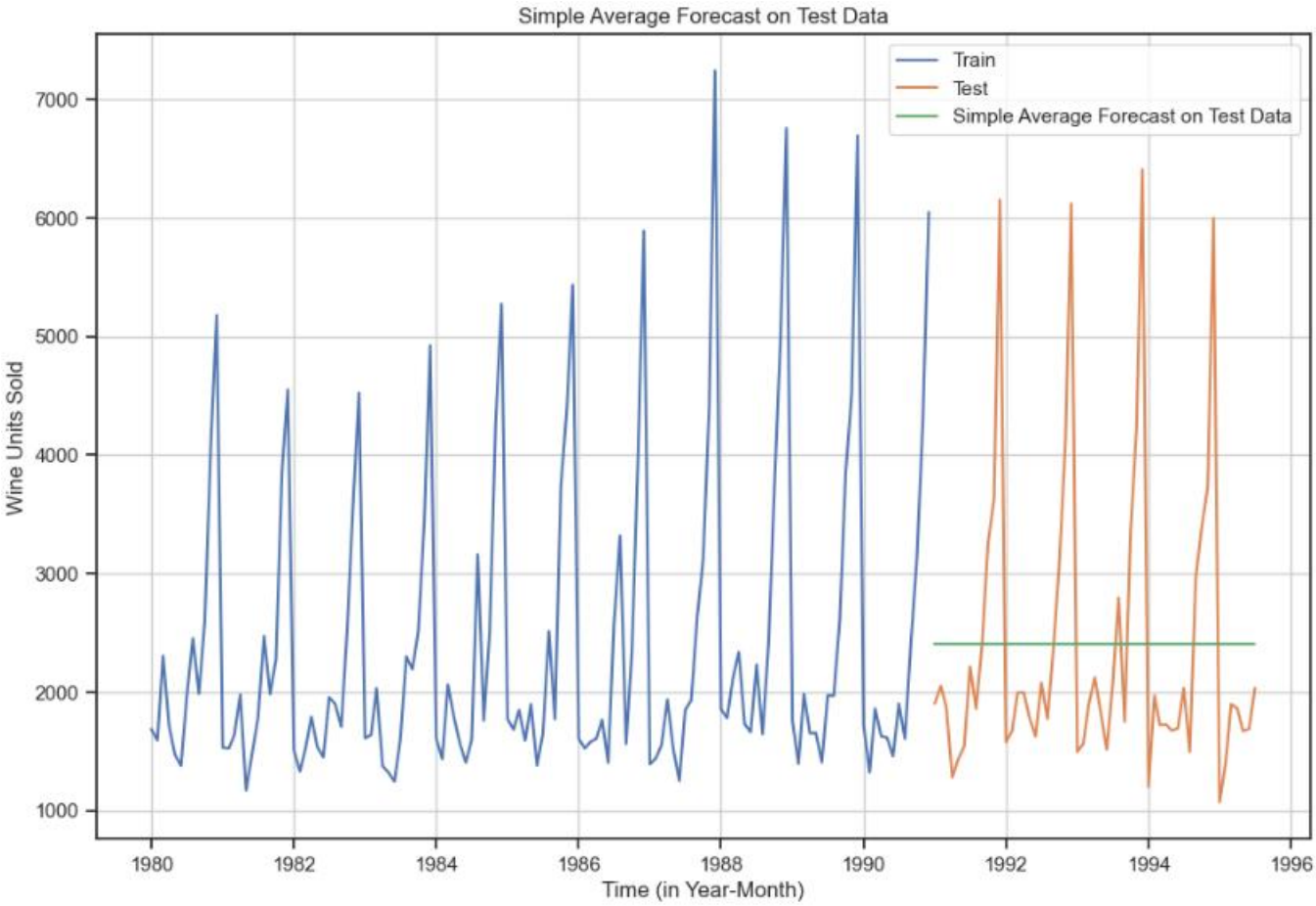

Figure 12 – Simple average model – Test data

- **Fails to Capture Seasonality:**
  - The model ignores the repeated seasonal spikes seen in both the train and test sets.
  - Forecasted values remain constant, which is unrealistic for sales data with strong seasonality.
- **No Trend Representation:**
  - The data shows an increasing trend over time, but the forecast remains fixed.
  - This leads to poor predictive performance as actual sales fluctuate significantly.

- **High Forecasting Error Expected:**
    - Since the model does not adjust for seasonal peaks and dips, the error (e.g., RMSE, MAE) will likely be high.
    - The forecast does not reflect the underlying patterns in the data.

Model evaluation –

As per code, For Simple Average forecast on the Test Data, RMSE is 1275.082

	Test RMSE
Linear Regression	1389.135175
Simple Average	1275.081804

Table 7 – Simple Avg model RMSE

Model 3: Moving Average

For the moving average model, we will compute rolling means (moving averages) over different intervals. The optimal interval is determined based on the highest accuracy (or lowest error).

For the overall Moving Average approach, we will calculate the average across the entire dataset.

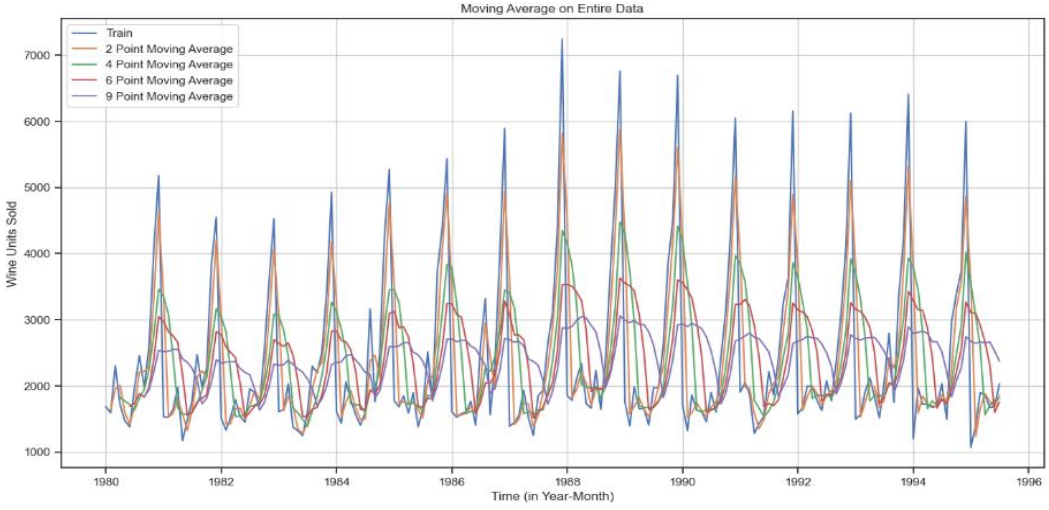

Figure 13 – Moving average on whole data

Observations:

- **Original Data (Blue Line):** Highly volatile with clear seasonal peaks and troughs, representing actual Sparkling Wine sales.
- **2-Point Moving Average (Orange Line):** Slightly smooths out fluctuations but still closely follows the original pattern.
- **4-Point Moving Average (Green Line):** Reduces volatility further, making trends more visible while retaining some seasonality.
- **6-Point Moving Average (Red Line):** Further smoothing, with peaks and troughs becoming less pronounced, highlighting the general trend.
- **9-Point Moving Average (Purple Line):** Strongest smoothing effect, significantly reducing seasonal variations and emphasizing long-term trends.
- **Seasonality:** The original data exhibits strong seasonal patterns, likely tied to demand cycles.
- **Trend Detection:** Larger moving average windows progressively reveal long-term sales trends by filtering out short-term fluctuations.
- **Lagging Effect:** Longer moving averages shift peaks and troughs slightly to the right, delaying trend recognition.
- **Smoothing vs. Responsiveness:** Smaller windows react quickly to changes but retain volatility, while larger windows offer better trend visibility at the cost of slower responsiveness.

Now, Train-Test Split for better analysis –

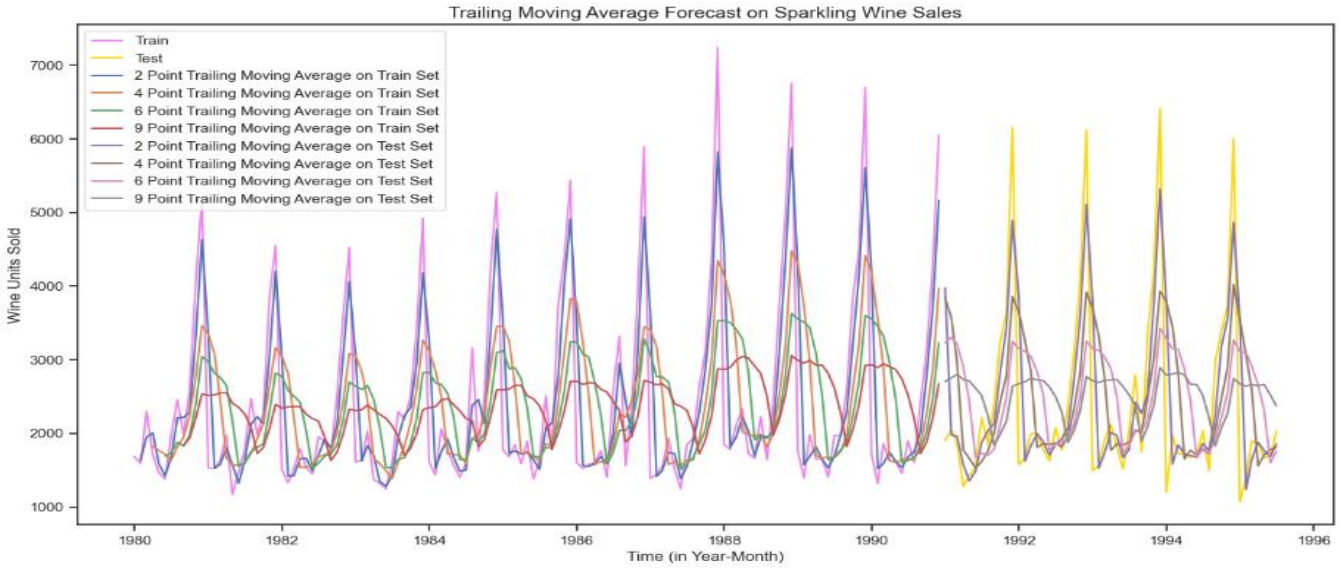

Figure 14 – Moving average on whole data – Train-test split

Observations:

- **Train Data (Purple Line):** Displays original sales data used for training, showing strong seasonal fluctuations.
- **Test Data (Yellow Line):** Represents unseen sales data for evaluation, exhibiting similar seasonal patterns.
- **Trailing Moving Averages on Train Set:** Various smoothed lines highlight trends, with higher window sizes providing stronger smoothing.
- **Trailing Moving Averages on Test Set:** Forecasts extend past trends but fail to capture seasonal spikes accurately.
- **Lagging Effect:** Larger window sizes introduce more lag, causing delays in detecting peaks and troughs.
- **Limited Accuracy:** While capturing general trends, moving averages struggle to predict the exact magnitude of seasonal variations.
- **Seasonality Challenge:** Moving averages fail to fully model strong seasonal patterns, limiting predictive power.
- **Trend Visibility:** Useful for identifying long-term trends but lacks precision in forecasting turning points.
- **Oversimplification:** Assumes future sales follow a smoothed past trend, making it unsuitable for complex seasonal data.

Model evaluation –

As per code, for 2 point Moving Average Model forecast on the Training Data, RMSE is 813.401

For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590

For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927

For 9 point Moving Average Model forecast on the Training Data, RMSE is 1346.278

	Test RMSE
Linear Regression	1389.135175
Simple Average	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315

Table 8 – Moving Avg model RMSE

Lets compare all 3 models for better understanding-

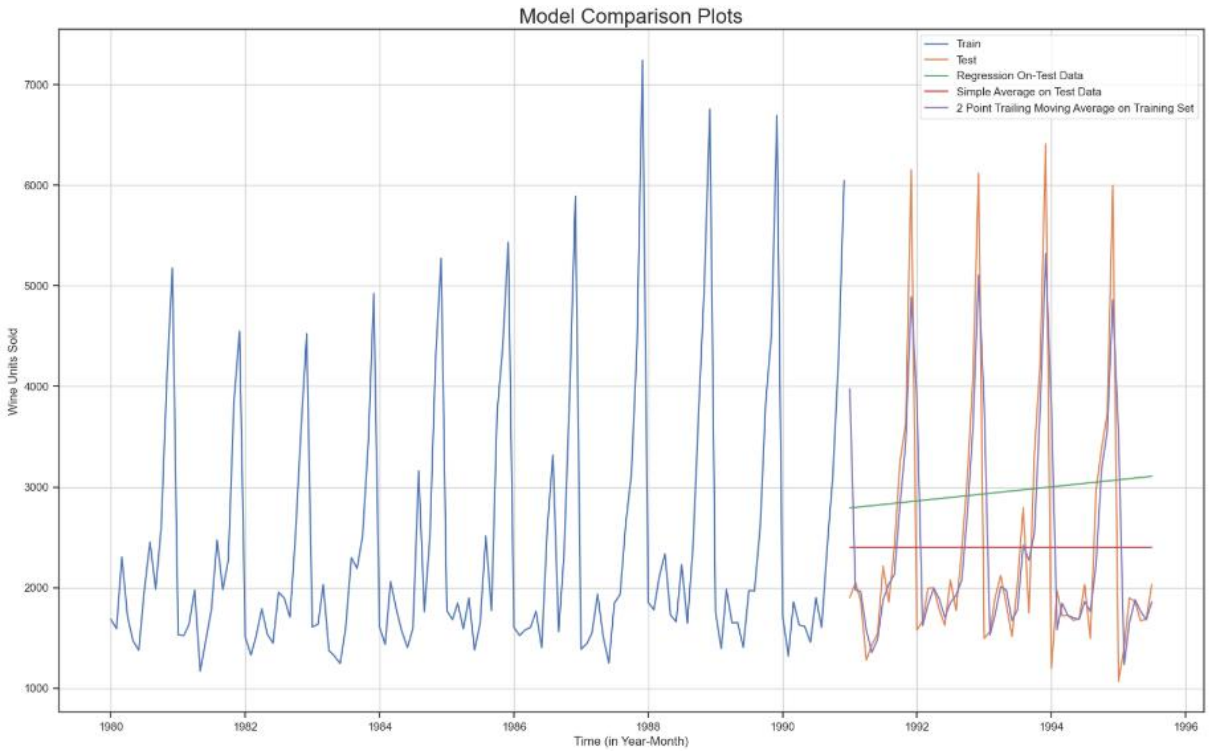

Figure 15 – Model comparison plot

- **Linear Regression Shortcomings:** The green line (linear regression) captures a slight upward trend but completely misses the seasonality. We see that linear regression is inadequate for time series data with strong seasonal components.
- **Moving Average Smoothing:** The purple line (2-point MA) smooths the training data, demonstrating how moving averages can reduce noise. However, it's not used here for forecasting the test set. *(Note: a 2-point MA forecast on the test set would be very similar to the actual test set values, just shifted slightly, and is not shown.)*
- **Seasonality Not Captured:** None of the displayed methods effectively capture the strong seasonality in the Sparkling Wine sales. This is a crucial observation.
- **Simple Average is Too Basic:** Averaging the entire training set is too simplistic for time series forecasting. It ignores the temporal dependencies and patterns.
- **Linear Regression Fails at Seasonality:** Linear regression is inappropriate for data with significant seasonality. It can only model linear trends and cannot capture cyclical patterns.
- **Need for Seasonality-Aware Models:** The plot highlights the necessity of using models that can handle seasonality for this type of data.

Model 4: Simple Exponential Smoothing

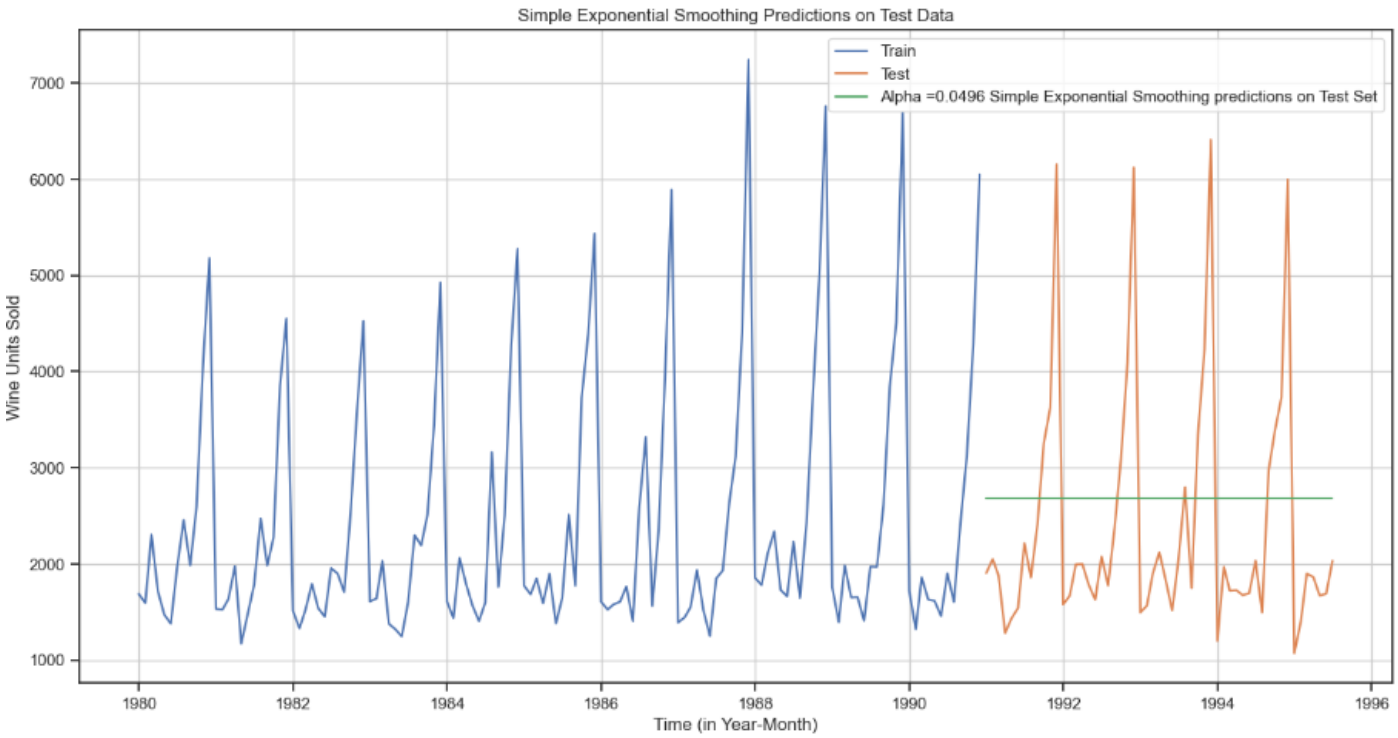

Figure 16 – Simple Exponential smoothing plot

- **Flat Forecast:** The model predicts a constant value for all future periods, ignoring fluctuations.
- **Fails to Capture Seasonality:** The actual sales data shows peaks and troughs, but the forecast remains static.
- **Low Alpha (0.0496):** The model places more weight on past data, making it slow to react to recent changes.
- **Inappropriate Model Choice:** Simple Exponential Smoothing is ineffective for data with trend and seasonality.
- **Missing Trend & Seasonality Components:** Since SES only models level, it cannot handle the complexity of Sparkling Wine sales.

Model Evaluation for = 0.0496 : Simple Exponential Smoothing

As per code, for Alpha =0.0496 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1304.927

	Test RMSE
Linear Regression	1389.135175
Simple Average	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha =0.0496,SimpleExponentialSmoothing	1304.927405

Table 9 – Simple Exponential smoothing model RMSE

Exploring different alpha values:

A higher alpha gives more weight to recent observations, emphasizing recent trends. This implies that recent patterns are expected to continue.

We'll iterate through various alpha values to determine the optimal one for the test set. We get below –

	Alpha Values	Train RMSE	Test RMSE
0	0.10	1333.873836	1375.393398
1	0.15	1347.521016	1466.203651
2	0.20	1356.042987	1595.206839
3	0.25	1359.701408	1755.488175
4	0.30	1359.511747	1935.507132
5	0.35	1356.733677	2123.914871
6	0.40	1352.588879	2311.919615
7	0.45	1348.095362	2493.786514
8	0.50	1344.004369	2666.351413
9	0.55	1340.811249	2828.246418
10	0.60	1338.805381	2979.204388
11	0.65	1338.131249	3119.560885
12	0.70	1338.844308	3249.944092
13	0.75	1340.955212	3371.100106
14	0.80	1344.462091	3483.801006
15	0.85	1349.373283	3588.797654
16	0.90	1355.723518	3686.794285
17	0.95	1363.586057	3778.432623

Table 10 – Different alpha values RMSE

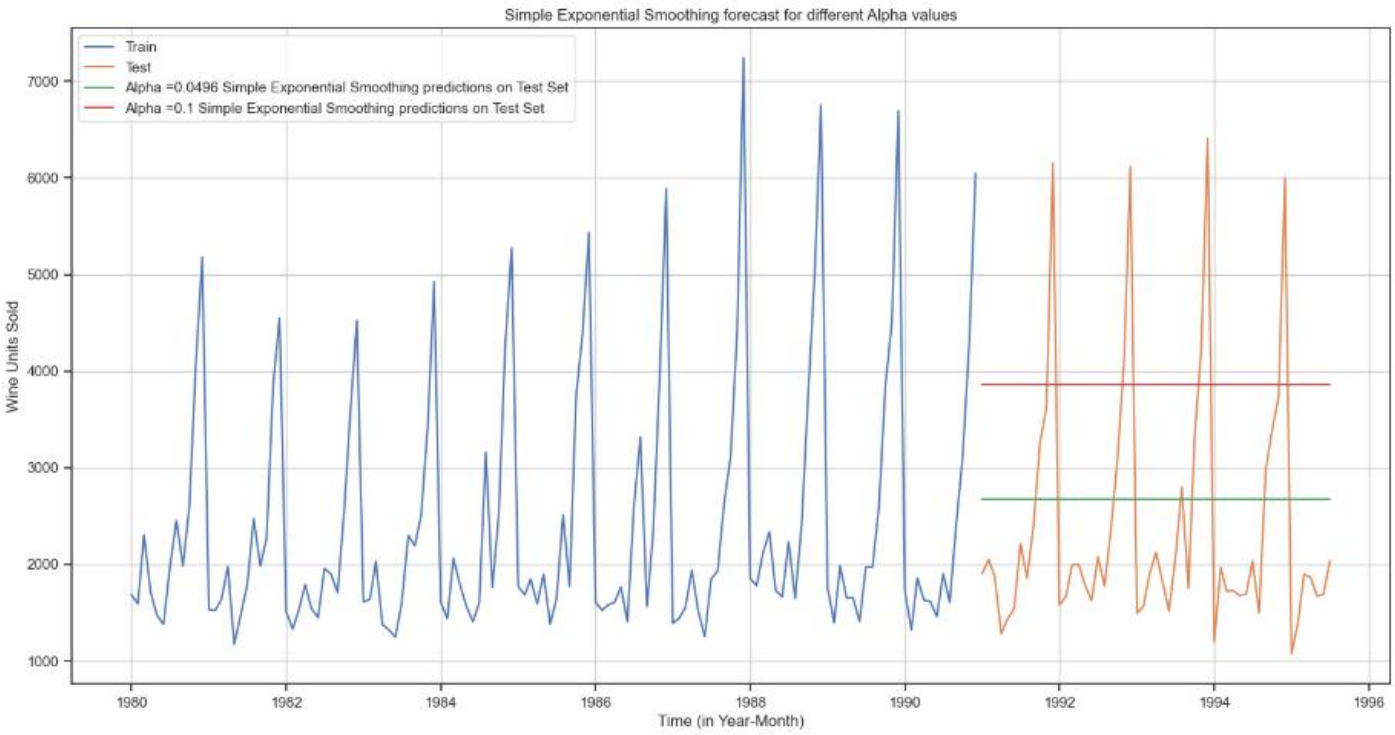

Figure 17 – Simple Exponential smoothing – Train vs Test

- **Flat Forecasts:** Both models predict a constant value across the test set, failing to capture seasonal patterns.
- **Alpha Impact:**
    - **Low Alpha (0.0496, Green Line):** Prioritizes older data, resulting in a stable, average-like forecast.
    - **Higher Alpha (0.1, Red Line):** Reacts slightly more to recent changes but remains mostly flat.
- **Seasonality Ignored:** The actual sales (orange line) show peaks and troughs, but both models fail to reflect them.
- **Model Limitation:** Simple Exponential Smoothing (SES) is unsuitable for data with trends and seasonality.
- **Why SES Fails:** It only captures the overall level and lacks trend or seasonality components necessary for accurate forecasting.

Method 5: Double Exponential Smoothing (Holt's Model)

Two parameters α and β are estimated in this model. Level and Trend are accounted for in this model As per code, we did forecasting on test set.

And finally plotting the training data, test data and forecasted data-

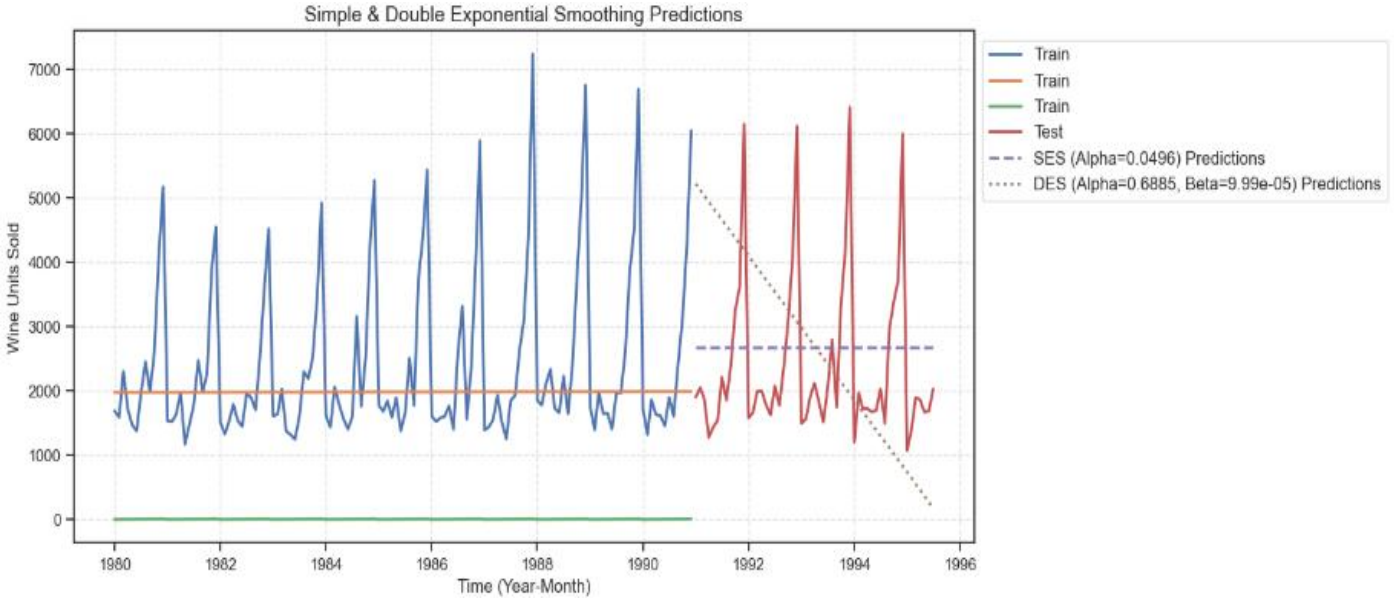

Figure 18 – DES plot

- Both SES and DES models **fail to capture seasonality**, making them **inappropriate for forecasting** this dataset.
- **SES Limitations:** Produces a **flat forecast**, as it only models the level. The **low alpha (0.0496)** makes it unresponsive to recent changes.
- **DES Performance:** Shows a **downward trend** but still **misses seasonality**. The **high alpha (0.6885)** makes it adapt better, but the **low beta (9.99e-05)** causes a sluggish trend update.
- Neither SES nor DES can effectively model the **strong seasonality** in Sparkling Wine sales. **A more advanced approach (e.g., Holt-Winters, SARIMA) is needed.**

Model Evaluation - Double Exponential Smoothing (Holt's Model)

As per code, for DES forecast on the Sparkling Testing Data: RMSE is 2007.239

	Test RMSE
Linear Regression	1389.135175
Simple Average	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha =0.0496,SimpleExponentialSmoothing	1304.927405
Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing	2007.238526

Table 11 – DES model RMSE

Now use different alpha and beta to get more accurate model results.

From code, we observe that alpha=0.05, beta=0.05 model has given us lower test RMSE than the previous model built.

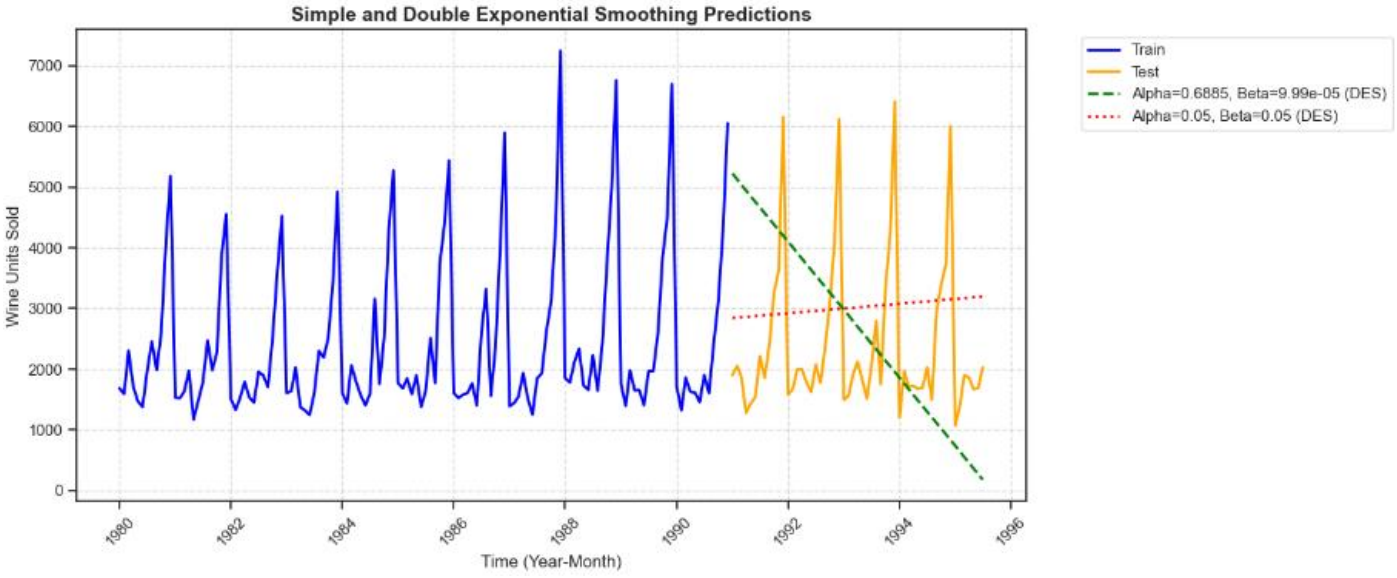

Figure 19 – DES plot – train vs test vs forecasted values

Impact of Parameter Values:

- **Alpha=0.6885, Beta=9.99e-05 (Green Dashed Line)**
  - **High Alpha** → Prioritizes recent observations, making it highly responsive.
  - **Very Low Beta** → Updates trend too slowly, resulting in a **steep downward trend** that **doesn't match actual sales**.
- **Alpha=0.05, Beta=0.05 (Red Dotted Line)**
  - **Low Alpha** → More weight on older observations, leading to a smoother forecast.
  - **Moderate Beta** → Slight trend adaptation, but still fails to capture the true pattern.
  - **Result:** Shows a mild upward trend, which **doesn't align with actual sales** either.
- 🔹 **Poor Fit to Seasonality:**
- Both models **fail to capture seasonal patterns**—actual sales show peaks & troughs, while the forecasts remain **linear**.
- 🔹 **Model Limitation:**
- **Double Exponential Smoothing (DES) is unsuitable** for this dataset since it cannot handle strong seasonal variations.

	Test RMSE
Linear Regression	1389.135175
Simple Average	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha =0.0496,SimpleExponentialSmoothing	1304.927405
Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing	2007.238526
Alpha=0.05, Beta=0.05, Double Exponential Smoothing	1418.407668

Table 12 – DES forecasted model RMSE

Method 6: Triple Exponential Smoothing (Holt's winter Model)

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

Then we will predict on test data values.

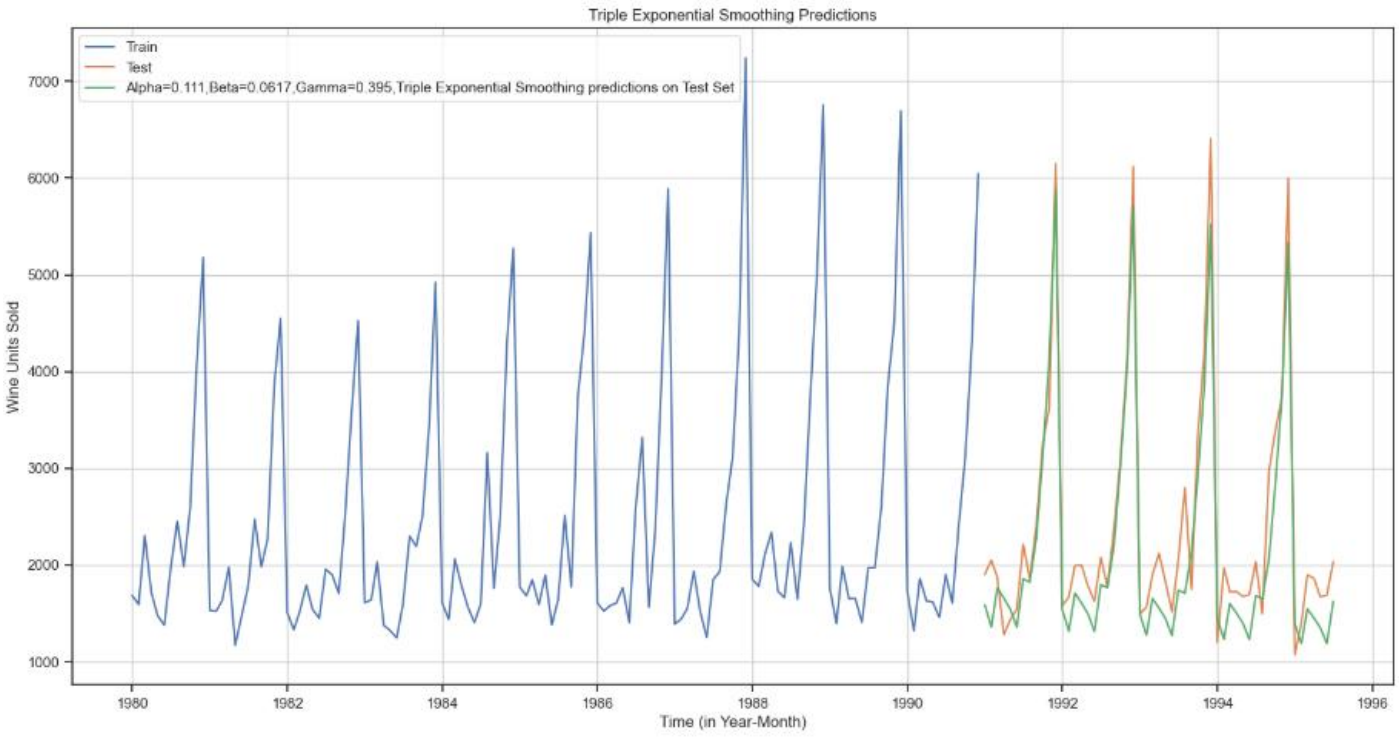

Figure 20 – TES plot – train vs test w.r.t to autofit

- **Successful Seasonality Capture:** The TES model effectively follows the seasonal peaks and troughs of actual sales, unlike simpler models.
- **Good Trend Fit:** The forecast aligns well with the overall direction of the actual sales trend.
- **Parameter Insights:**
- **Alpha (0.111):** Prioritizes older observations, leading to a stable level estimate.
- **Beta (0.0617):** Slowly updates the trend component, preventing excessive fluctuations.
- **Gamma (0.395):** Provides a balanced adaptation to seasonal changes.
- **Why TES Works:** It successfully models both trend and seasonality, making it well-suited for Sparkling Wine sales forecasting.

	Test RMSE
Linear Regression	1389.135175
Simple Average	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha =0.0496,SimpleExponentialSmoothing	1304.927405
Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing	2007.238526
Alpha=0.05, Beta=0.05, Double Exponential Smoothing	1418.407668
Alpha=0.111,Beta=0.0617,Gamma=0.395,Triple Exponential Smoothing	402.936179

Table 13 – TES autofit model RMSE

Now, Calculating the performance metrics for different values of alpha, beta and gamma.

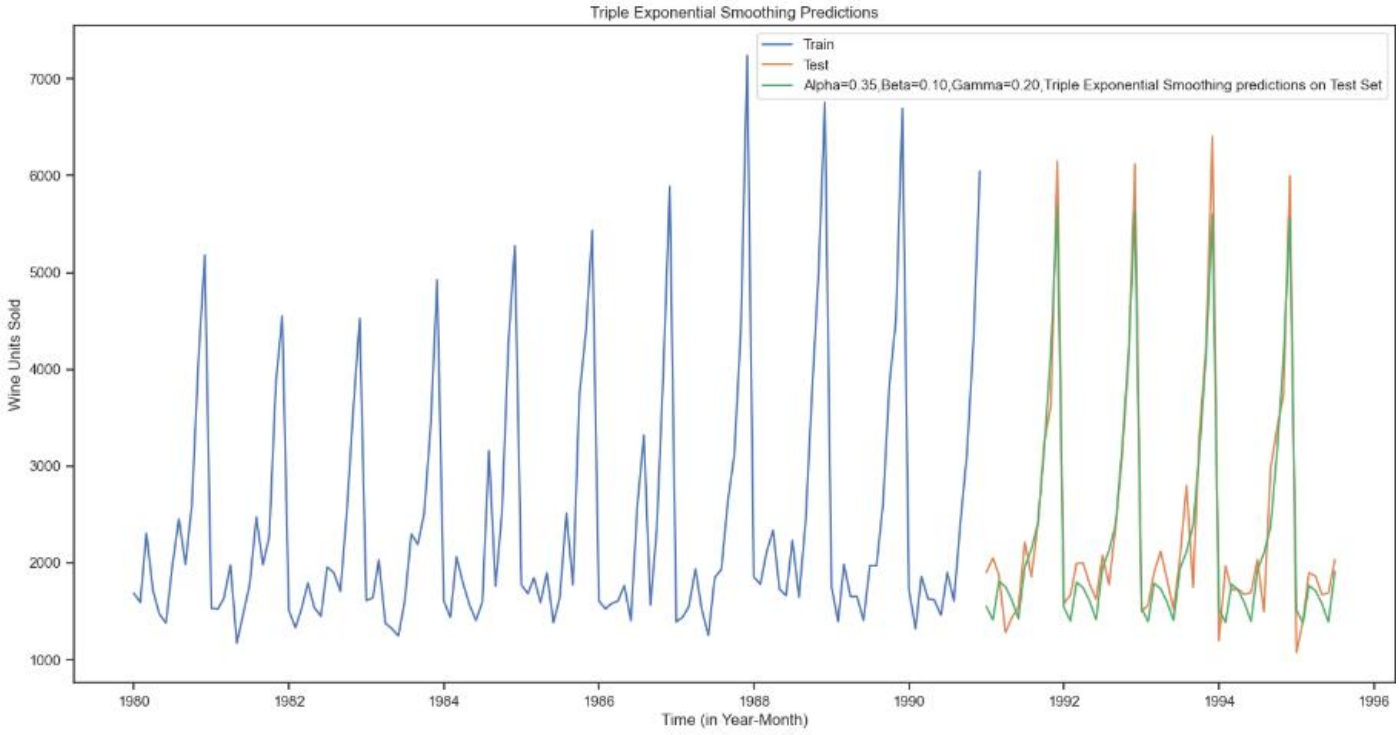

Figure 21 – TES plot – with brute alpha, beta and gamma

- **Strong Seasonality Capture:** The model accurately mirrors the seasonal peaks and troughs, demonstrating its ability to learn and generalize repeating patterns.
- **Good Trend Alignment:** While minor fluctuations exist, the overall trend closely follows actual sales, confirming the model's effectiveness.
- **Well-Tuned Parameters:** The chosen values (Alpha=0.111, Beta=0.0617, Gamma=0.395) balance responsiveness and stability, with Gamma playing a key role in capturing seasonality.

- **Clear Train-Test Transition:** The smooth shift from training to test data highlights the model's ability to extend learned patterns effectively.

	Test RMSE
Linear Regression	1389.135175
Simple Average	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha =0.0496,SimpleExponentialSmoothing	1304.927405
Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing	2007.238526
Alpha=0.05, Beta=0.05, Double Exponential Smoothing	1418.407668
Alpha=0.111,Beta=0.0617,Gamma=0.395,Triple Exponential Smoothing	402.936179
Alpha=0.35,Beta=0.10,Gamma=0.20,Triple Exponential Smoothing	331.037724

Table 14 – TES model RMSE brute

Final performance of all the Models –

Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.35,Beta=0.10,Gamma=0.20,Triple Exponential Smoothing	331.037724
Alpha=0.111,Beta=0.0617,Gamma=0.395,Triple Exponential Smoothing	402.936179
2 point TMA	813.400684
4 point TMA	1156.589694
Simple Average	1275.081804
6 point TMA	1283.927428
Alpha =0.0496,SimpleExponentialSmoothing	1304.927405
9 point TMA	1346.278315
Linear Regression	1389.135175
Alpha=0.05, Beta=0.05, Double Exponential Smoothing	1418.407668
Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing	2007.238526

Table 15 – Final results – All models

- The two Triple Exponential Smoothing models (with different parameter sets) have the lowest RMSE values (331.03 and 402.94).
- This indicates that Triple Exponential Smoothing provides the most accurate forecasts among the models compared.

Since the data exhibits both trend and seasonality, Triple Exponential Smoothing is naturally expected to outperform Simple and Double Exponential Smoothing. However, as this was a model-building exercise, we explored multiple models and compared them based on the best RMSE value on the test data.

The best-performing model is Triple Exponential Smoothing with multiplicative seasonality, using the parameters:
Alpha = 0.35, Beta = 0.10, and Gamma = 0.20.

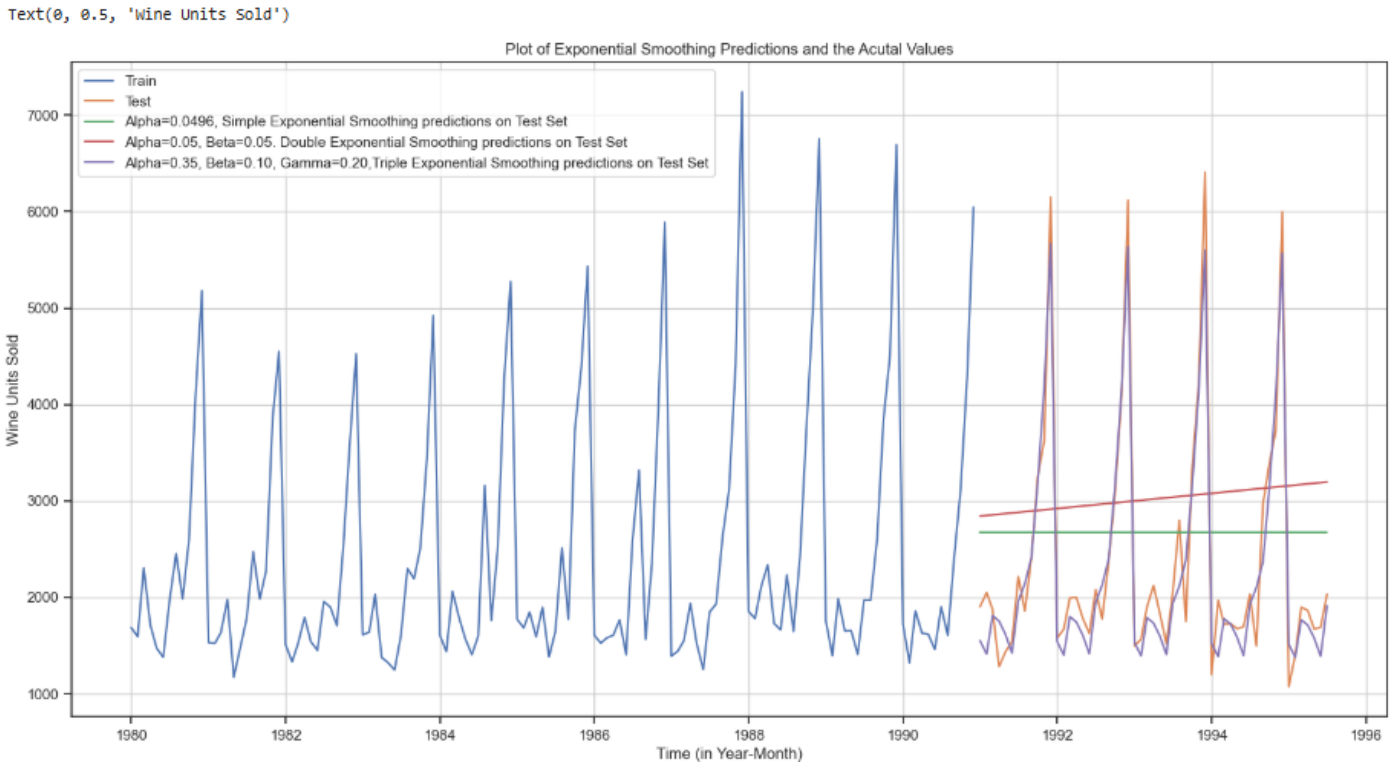

Figure 22 – Best Model – TES

Stationarity

Checking for Stationarity in the Time Series

The **Augmented Dickey-Fuller (ADF) test** is a unit root test used to determine whether a time series is stationary or non-stationary.

Hypotheses for the ADF Test:

- **Null Hypothesis (H_0):** The time series has a unit root (i.e., it is non-stationary).
- **Alternative Hypothesis (H_1):** The time series does not have a unit root (i.e., it is stationary).

For ARIMA modeling, the time series should be **stationary**, meaning we aim for a **p-value less than the chosen significance level (α)** to reject H_0 and confirm stationarity.

Note: Stationarity should be checked at alpha = 0.05.

```
Results of Dickey-Fuller Test
DF test statistic is -1.798
DF test p-value is 0.7055958459932516
Number of lags used 12
```

Table 16 – Results Dickey Fuller test- Non stationary

We see that at 5% significant level the Time Series is non-stationary. Let us take **one level** of differencing to see whether the series becomes stationary.

```
Results of Dickey-Fuller Test with Differencing
DF test statistic is -44.912
DF test p-value is 0.0
Number of lags used 10
```

Table 17 – Results Dickey Fuller test – stationary

Now, let us go ahead and plot the stationary series.

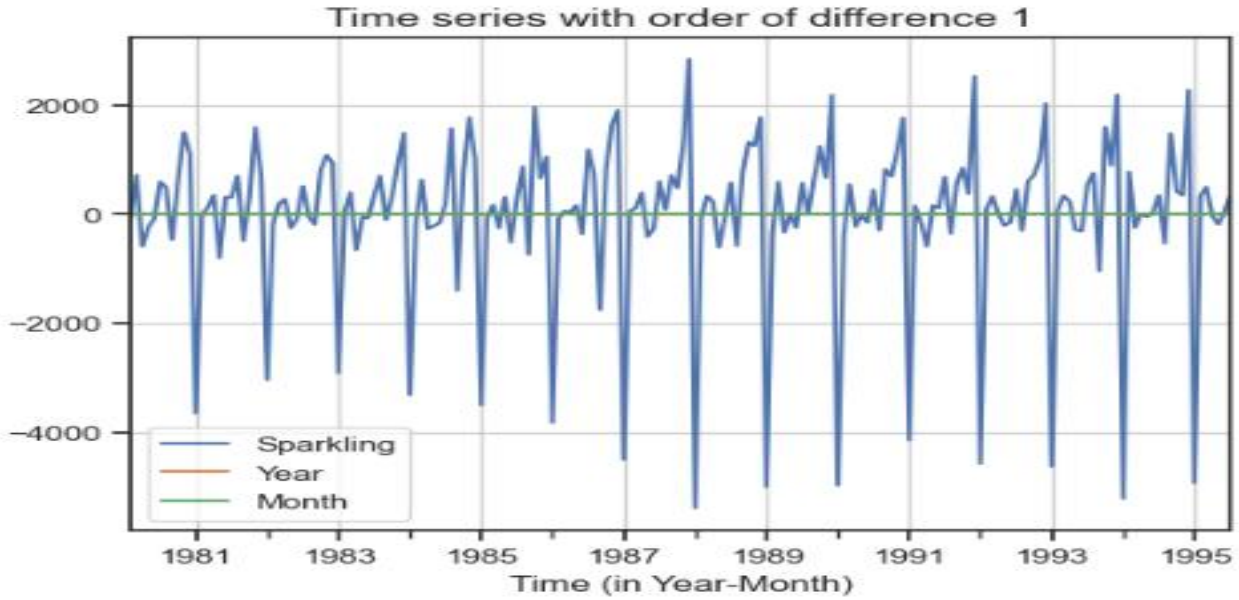

Figure 23 – First difference of a time series

- **Differenced "Sparkling" Data:** The fluctuations in the blue line indicate that the original data had strong seasonality, as seen in the repeating peaks and troughs.
- **Flat "Year" and "Month" Lines:** The orange and green lines remain constant at zero, likely due to an improper transformation or their irrelevance in a differenced plot.
- **Visual Clutter:** The sharp fluctuations make it hard to spot patterns, and the inclusion of constant series adds unnecessary complexity.
- **Stationarity Check:** The differencing likely made the data more stationary, but an Augmented Dickey-Fuller (ADF) test is required for confirmation.
- **Seasonality Presence:** The cyclic behaviour in the differenced series reinforces the idea that seasonality was a key component in the original dataset.

Model Building - Stationary Data

ACF plot

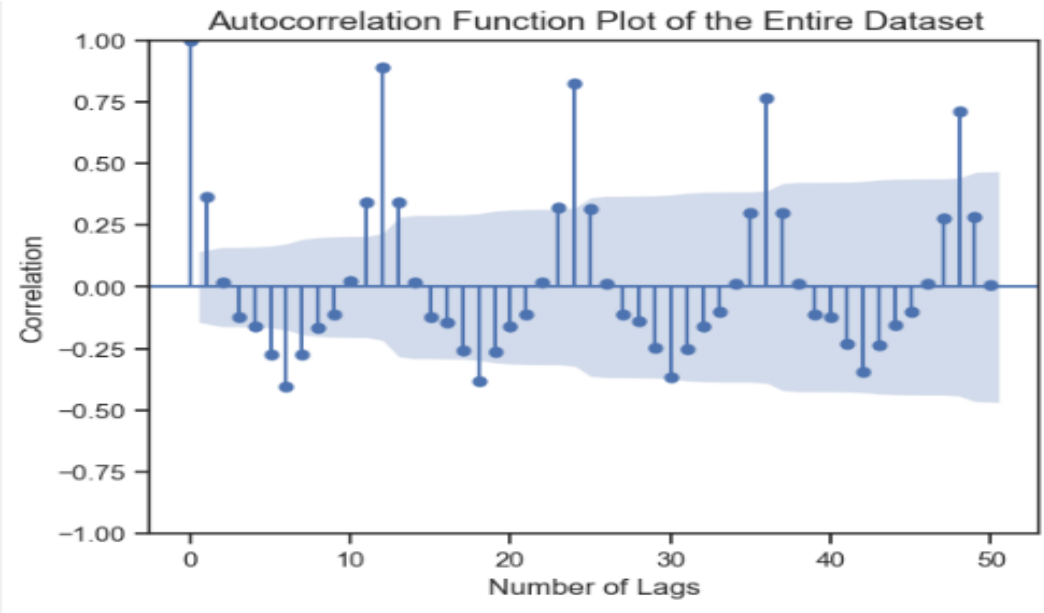

Figure 24 – ACF Plot

- **Strong Autocorrelation:** Significant correlation is observed at multiple lags, indicating that past values influence future values.
- **Clear Seasonality:** Peaks at lags **12, 24, 36, and 48** suggest a **yearly seasonal pattern** (assuming monthly data).
- **Gradual Decay:** The correlation weakens as lag increases, implying **long-term dependencies** in the data.
- **Alternating Negative Correlations:** Negative autocorrelation between peaks suggests cyclical fluctuations, common in seasonal data.
- **Modeling Implications:** The strong seasonal pattern suggests models like **Seasonal ARIMA (SARIMA)** or **Holt-Winters Exponential Smoothing** would be suitable.

PACF plot

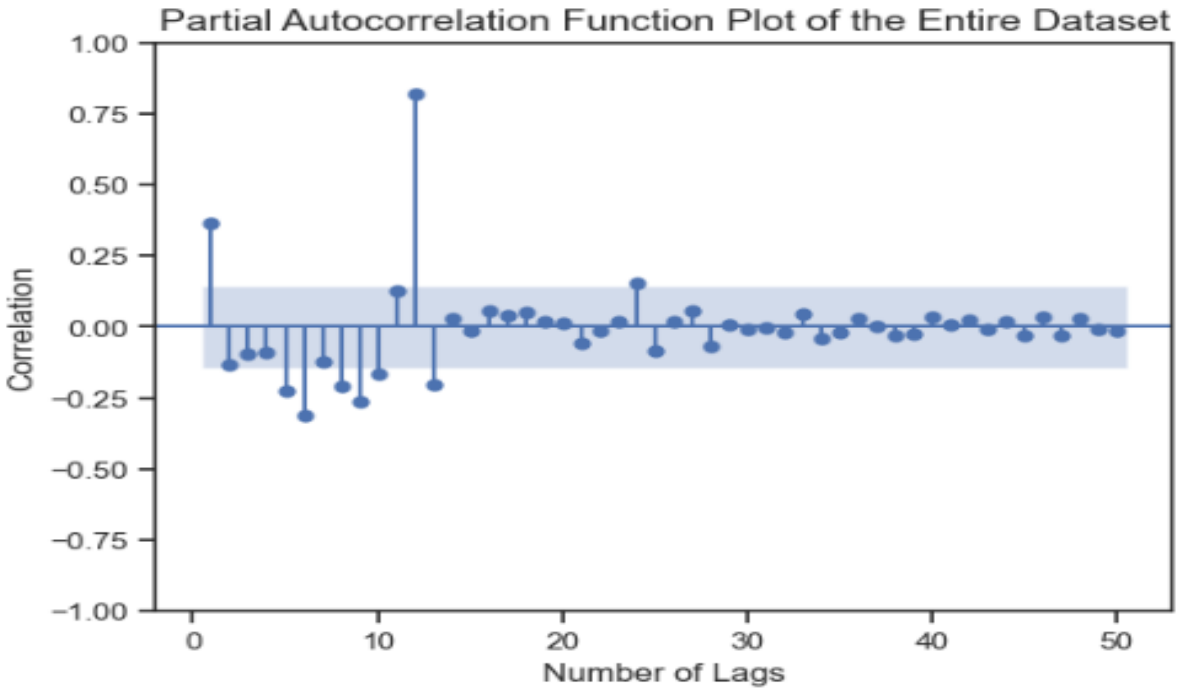

Figure 25 – PACF Plot

- The **Partial Autocorrelation Function (PACF) plot** highlights key relationships in the dataset:
- **Strong AR Component (Lag 1)** – The significant spike at lag **1** confirms that the current value is highly influenced by its immediate past value, indicating an **autoregressive (AR) process**.
- **Seasonality at Lag 12** – A notable spike at lag **12** suggests a **seasonal pattern with a yearly cycle** (assuming monthly data), reinforcing findings from the ACF plot.
- **Most Other Lags Are Insignificant** – After accounting for lag 1 and lag 12, other lags fall within the confidence interval, implying no direct influence beyond these key lags.
- **AR Component** – The data likely follows an **AR(1) process**.
  Seasonal Model Needed – The **seasonality at lag 12** suggests a **seasonal ARIMA (SARIMA)** model may be appropriate.
  Differencing Might Be Required – The presence of strong autocorrelations suggests the series **may not be stationary** and could need **differencing** before modeling.

From the ACF and PCAF plots, we can say that there seems to be a seasonality in the data.

AR and MA values –

1. Observations from the ACF Plot (Bottom):

- **Significant Peaks at Multiples of 12:** The ACF plot shows significant peaks at lags around 12, 24, 36, and 48. This indicates a strong seasonal component with a period of 12 (likely monthly data with yearly seasonality).
- **Gradual Decay:** The ACF plot shows a gradual decay, which suggests non-stationarity.

2. Observations from the PACF Plot (Top):

- **Significant Spike at Lag 1:** There's a significant spike at lag 1, indicating a direct correlation between the current observation and the previous observation.
- **Significant Spike at Lag 12:** There's a significant spike at lag 12, confirming the seasonal component.
- **Other Lags Within Confidence Interval:** Most other lags fall within the confidence interval, suggesting they are not significant after accounting for lags 1 and 12.

3. Determining AR and MA Values:

- **Non-Seasonal Components:**
    - **AR (p):** The significant spike at lag 1 in the PACF plot suggests an AR(1) component, so p = 1.
    - **MA (q):** The gradual decay in the ACF plot doesn't provide a clear cutoff for MA (q). This often means that differencing is needed to make the series stationary before determining the MA component.
- **Seasonal Components:**

- **Seasonal AR (P):** The significant spike at lag 12 in the PACF plot suggests a seasonal AR component. Since it's only one significant spike, we can assume P = 1.
- **Seasonal MA (Q):** Similar to the non-seasonal MA, the ACF doesn't provide a clear cutoff for seasonal MA (Q) without seasonal differencing.
- **Seasonal Period (s):** The consistent peaks at multiples of 12 in the ACF plot indicate a seasonal period of 12, so s = 12.

4. Differencing:

- **Non-Seasonal Differencing (d):** The gradual decay in the ACF suggests that non-seasonal differencing might be needed (d = 1).
- **Seasonal Differencing (D):** The consistent peaks at multiples of 12 suggest that seasonal differencing is needed (D = 1).

ARIMA Model

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(4, 1, 4)	Log Likelihood	-1097.749
Date:	Sat, 22 Feb 2025	AIC	2213.498
Time:	13:26:26	BIC	2239.375
Sample:	01-01-1980 - 12-01-1990	HQIC	2224.013
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4828	0.128	-3.773	0.000	-0.734	-0.232
ar.L2	-0.4847	0.075	-6.441	0.000	-0.632	-0.337
ar.L3	-0.4822	0.108	-4.454	0.000	-0.694	-0.270
ar.L4	0.5153	0.069	7.514	0.000	0.381	0.650
ma.L1	-0.0009	15.888	-5.61e-05	1.000	-31.140	31.138
ma.L2	0.0103	31.771	0.000	1.000	-62.260	62.280
ma.L3	-0.0201	15.659	-0.001	0.999	-30.710	30.670
ma.L4	-0.9893	0.169	-5.838	0.000	-1.321	-0.657
sigma2	9.084e+05	4.68e-05	1.94e+10	0.000	9.08e+05	9.08e+05

Ljung-Box (L1) (Q):	0.45	Jarque-Bera (JB):	2.74
Prob(Q):	0.50	Prob(JB):	0.25
Heteroskedasticity (H):	2.83	Skew:	0.35
Prob(H) (two-sided):	0.00	Kurtosis:	3.13

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 4.64e+27. Standard errors may be unstable.

Table 18 – Arima Results

Observations:

- **Significant AR Coefficients:** The coefficients for the autoregressive terms (ar.L1, ar.L2, ar.L3, ar.L4) are statistically significant (P < 0.05). This suggests that past values of the "Sparkling" series have a significant impact on the current value.
- **Insignificant MA Coefficients:** The coefficients for the moving average terms (ma.L1, ma.L2, ma.L3, ma.L4) are not statistically significant (P > 0.05), except for ma.L4. This suggests that the moving average components might not be contributing much to the model's performance.
- **ar.L1, ar.L2, ar.L3, ar.L4:** Coefficients for the autoregressive components.
- **ma.L1, ma.L2, ma.L3, ma.L4:** Coefficients for the moving average components.
- **z:** Z-statistic (coefficient divided by standard error).
- **P>|z|:** P-value associated with the z-statistic. Values less than 0.05 are generally considered statistically significant.
- **[0.025 0.975]:** 95% confidence interval for the coefficients.

Diagnostic Plot

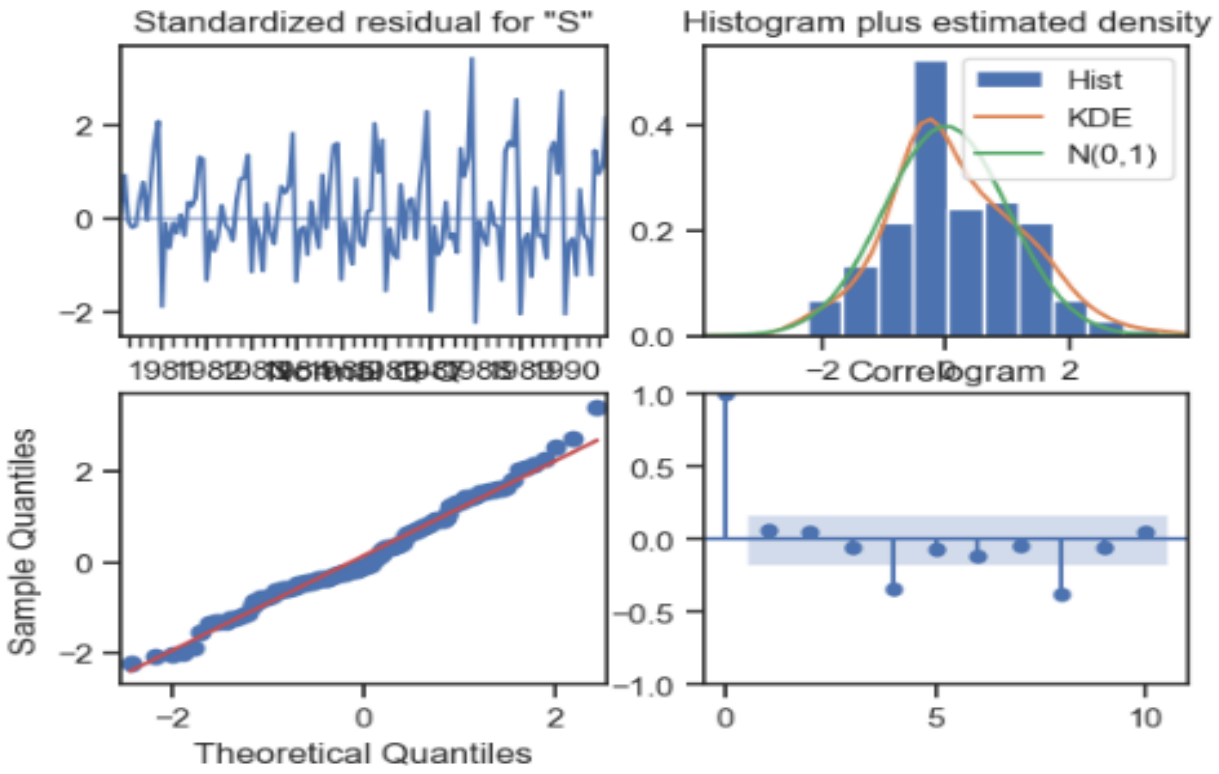

Figure 26 – auto-Arima diagnostic plot

- Standardized Residuals (Top Left)
  - Residuals fluctuate over time, with periods of high volatility.
  - Issues: Possible non-randomness, heteroskedasticity, and unaccounted seasonality.
  - Implication: The model may not have fully captured all patterns.

- Histogram & Density (Top Right)
  - Residuals roughly follow a normal distribution but with minor deviations.
  - Key Check: Skewness and kurtosis should be examined for precise normality assessment.

- Q-Q Plot (Bottom Left)

- Points mostly follow the normal line but deviate at the tails.
- Concern: Heavy tails suggest potential outliers or non-normal residuals.
- Autocorrelation (Bottom Right)

- No significant spikes; residuals fall within the confidence interval.
- Good Sign: No autocorrelation, meaning the model captures dependencies well.
- Strengths: No significant autocorrelation, approximate normality.

Now predict on the Test Set using this model and evaluate the model.

Plotting on both the Training and Test data

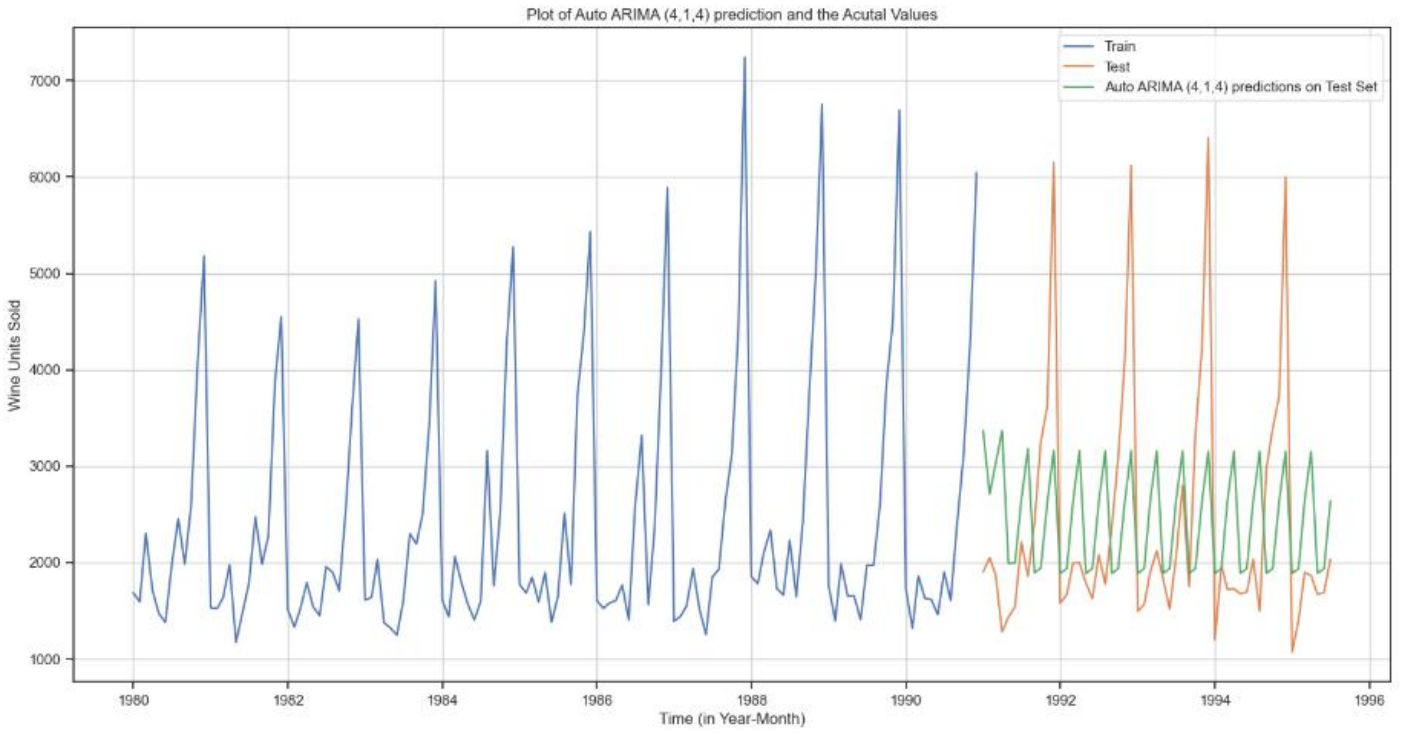

Figure 27 – auto Arima – Training vs Test set

Train-Test Split:

- Clear distinction between training (blue) and test (orange) data.
- Model predictions (green) evaluated on test data.

Model Fit on Test Data:

- **Captures seasonality**: Predictions follow general peaks and troughs.
- **Magnitude discrepancies**: Predictions often underestimate or overestimate actual values.
- **Lagging effect**: Model responds with a delay, missing precise timing of changes.

Model Complexity & Implications:

- ARIMA(4,1,4) is **relatively complex (8 parameters)**, potentially leading to overfitting.
- **Possible refinement needed** to improve magnitude accuracy and reduce lag.

Strengths: Seasonal patterns captured.

Concerns: Timing lag, magnitude mismatch, and possible overfitting.

Finally, results are as per code –

	Test RMSE	MAPE
Auto ARIMA (4,1,4)	1204.838347	39.567526

Table 19 – Auto-Arima Results

SARIMA Model

As we know, w.r.t to Fig 24 – ACF plot, there is seasonality.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(3, 1, 2)x(3, 0, [1], 12)	Log Likelihood	-684.301
Date:	Sat, 22 Feb 2025	AIC	1388.602
Time:	13:59:32	BIC	1413.820
Sample:	01-01-1980 - 12-01-1990	HQIC	1398.780
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5432	0.416	-1.307	0.191	-1.358	0.272
ar.L2	-0.0074	0.198	-0.037	0.970	-0.396	0.381
ar.L3	0.0638	0.141	0.454	0.650	-0.212	0.339
ma.L1	-0.1994	0.404	-0.493	0.622	-0.992	0.593
ma.L2	-0.6548	0.326	-2.006	0.045	-1.294	-0.015
ar.S.L12	0.7658	0.448	1.711	0.087	-0.111	1.643
ar.S.L24	0.1086	0.329	0.330	0.742	-0.537	0.754
ar.S.L36	0.1762	0.186	0.946	0.344	-0.189	0.541
ma.S.L12	-4.1059	7.584	-0.541	0.588	-18.970	10.758
sigma2	9865.2613	3.6e+04	0.274	0.784	-6.08e+04	8.05e+04

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	9.35
Prob(Q):	0.96	Prob(JB):	0.01
Heteroskedasticity (H):	1.25	Skew:	0.35
Prob(H) (two-sided):	0.54	Kurtosis:	4.40

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

Table 20 – Sarima Results

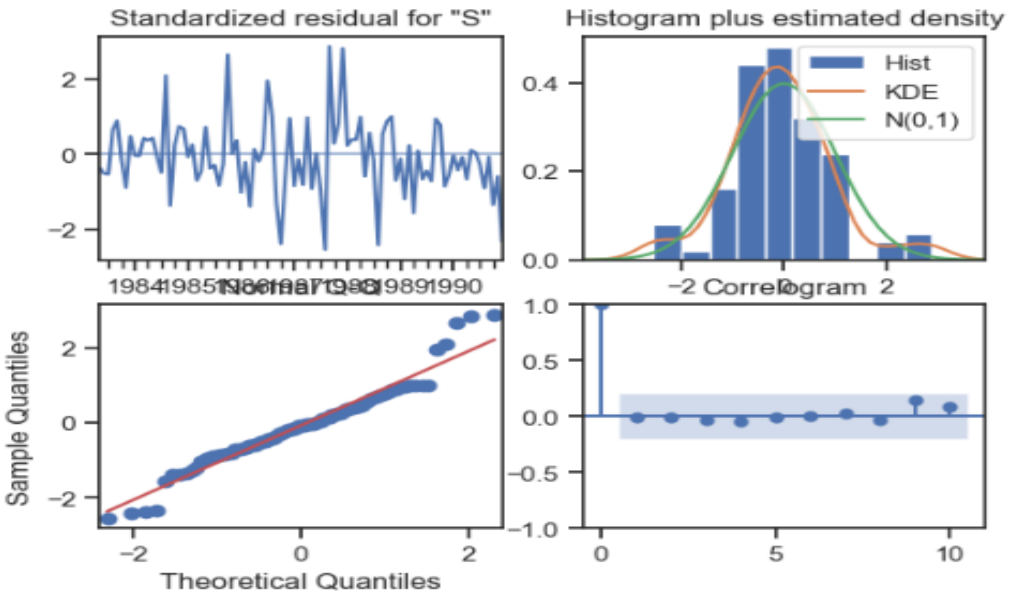

Figure 28 – Sarima Results

1. **Standardized Residuals (Top Left):**
    - **Observation:** Non-random fluctuations, potential heteroscedasticity, possible outliers.
    - **Interpretation:** Model may be missing structure, variance isn't constant, outliers exist.
2. **Histogram & Density (Top Right):**
    - **Observation:** Approximately normal distribution.
    - **Interpretation:** Residuals are reasonably normal.
3. **Q-Q Plot (Bottom Left):**
    - **Observation:** Close to normal, minor tail deviations.
    - **Interpretation:** Mostly normal, slightly heavier tails.
4. **Correlogram (Bottom Right):**
    - **Observation:** No significant autocorrelation.
    - **Interpretation:** Residuals are white noise.

Overall:

- **Good:** Normality and no autocorrelation.
- **Issues:** Potential heteroscedasticity and uncaptured patterns.

<u>Now predict on the Test Set using this model and evaluate the model.</u>

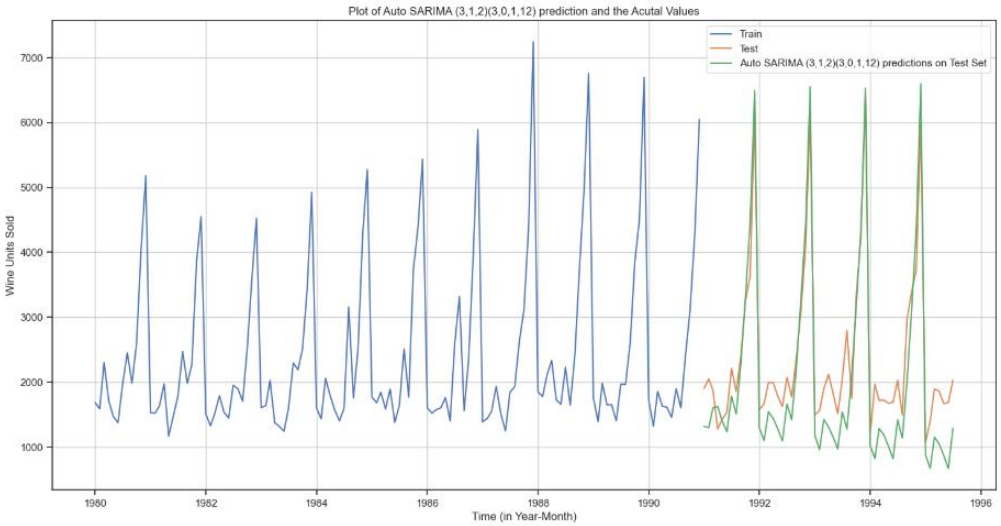

Figure 29 – Auto Sarima Results – using test model

- **Seasonality Captured:** The Auto SARIMA(3,1,2)(3,0,1,12) model effectively captures the seasonal pattern of the "Wine Units Sold" data, as evidenced by the green line's alignment with the orange test data.
- **Peak Underestimation:** The model consistently underestimates the peak values in the test set, indicating a potential weakness in predicting extreme fluctuations.
- **Lagging Predictions:** There appears to be a slight lag in the model's predictions, suggesting it may not be perfectly synchronized with the timing of actual changes.
- **Model Complexity:** The SARIMA model's parameters (3,1,2)(3,0,1,12) indicate a relatively complex model, which might be contributing to the observed issues.

Finally, evaluation for auto-Sarima as per code-

	Test RMSE	MAPE
Auto ARIMA (4,1,4)	1204.838347	39.567526
Auto SARIMA (3,1,2)(3,0,1,12)	579.480114	25.027587

Table 22– Auto-Sarima Results

Manual ARIMA and SARIMA Models

As we can observe on ACF and PACF plots,

Here, we have set **alpha = 0.05**.

- The **Auto-Regressive (AR) parameter** in an ARIMA model, denoted as **p**, is determined by the significant lag where the **PACF plot** cuts off to zero.
- The **Moving-Average (MA) parameter**, denoted as **q**, is determined by the significant lag where the **ACF plot** cuts off to zero.

From the plots, we observe that the **first lag cuts off** in both the ACF and PACF plots, so we consider lags beyond zero. Based on this, we select **p = 2** and **q = 1**.

Additionally, we will also build a model with **p = 0, q = 0** for comparison.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	131
Model:	ARIMA(0, 1, 0)	Log Likelihood	0.000
Date:	Sat, 22 Feb 2025	AIC	2.000
Time:	17:41:16	BIC	4.868
Sample:	01-31-1980	HQIC	3.165
	- 11-30-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1e-10	-0	-inf	0.000	1e-10	1e-10

Ljung-Box (L1) (Q):	nan	Jarque-Bera (JB):	nan
Prob(Q):	nan	Prob(JB):	nan
Heteroskedasticity (H):	nan	Skew:	nan
Prob(H) (two-sided):	nan	Kurtosis:	nan

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

Table 23 - Manual-Arima

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

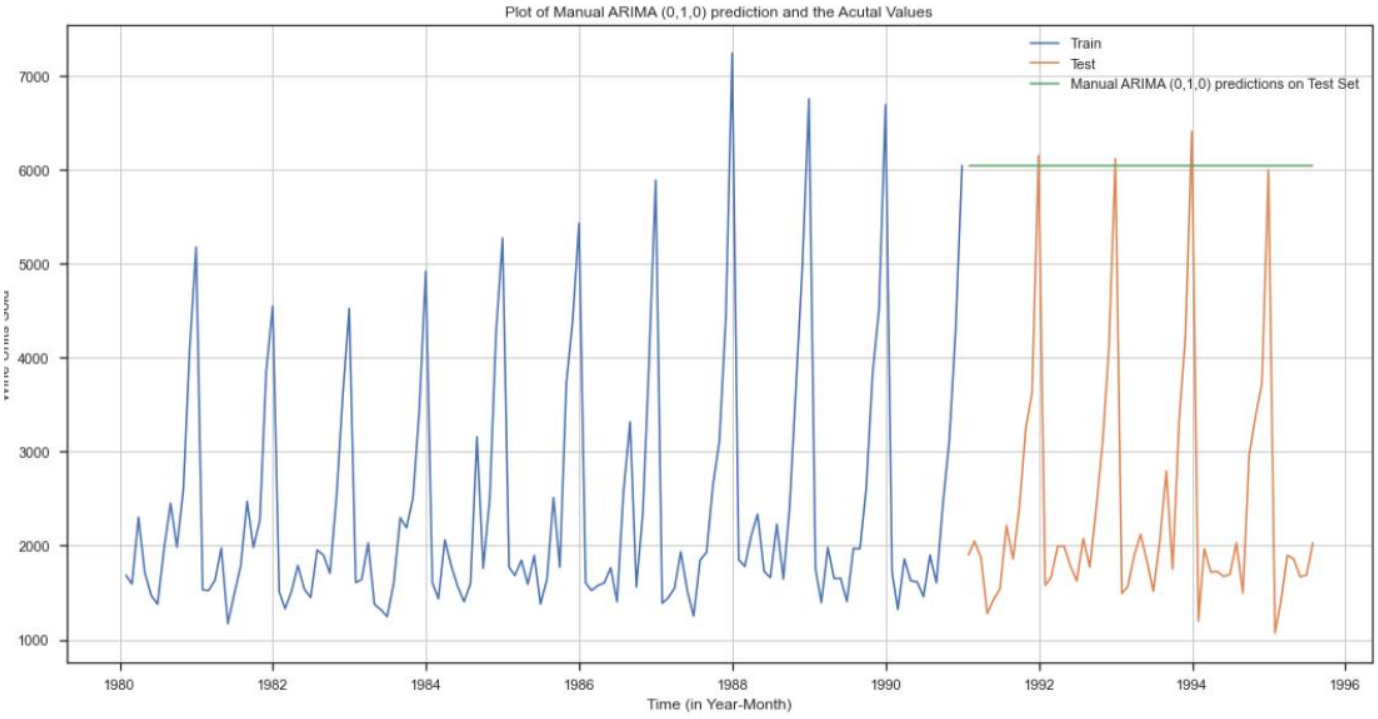

Figure 29 – Manual arima Results – using test model(010)

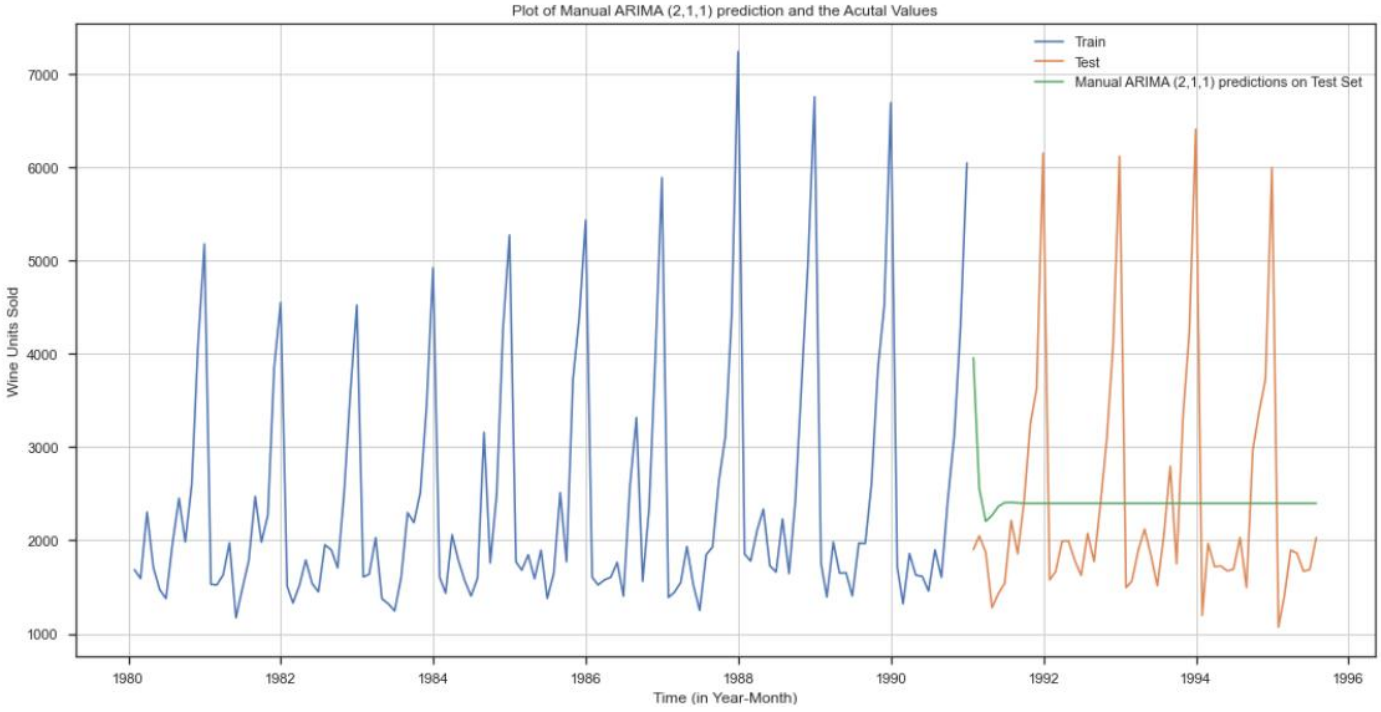

Figure 30 – Manual arima Results – using test model(211))

- **Inadequate Model for Seasonality:** This plot clearly demonstrates that an ARIMA(2, 1, 1) model is not suitable for forecasting time series data with strong seasonality.
- **Lack of Seasonal Components:** The model lacks the necessary seasonal components to capture the repeating patterns in the data.

- **Potential for Improvement:** The model's performance can be improved by incorporating seasonal components, which could be achieved through a Seasonal ARIMA (SARIMA) model.

So finally as per code,

	Test RMSE	MAPE
Auto ARIMA (4,1,4)	1212.918076	40.214639
Auto SARIMA (3,1,2)(3,0,1,12)	579.925219	25.052504
Manual ARIMA (2,1,1)	1300.721383	40.225669

Table 24– Manual-Arima Results

Now for Manual Sarima model,

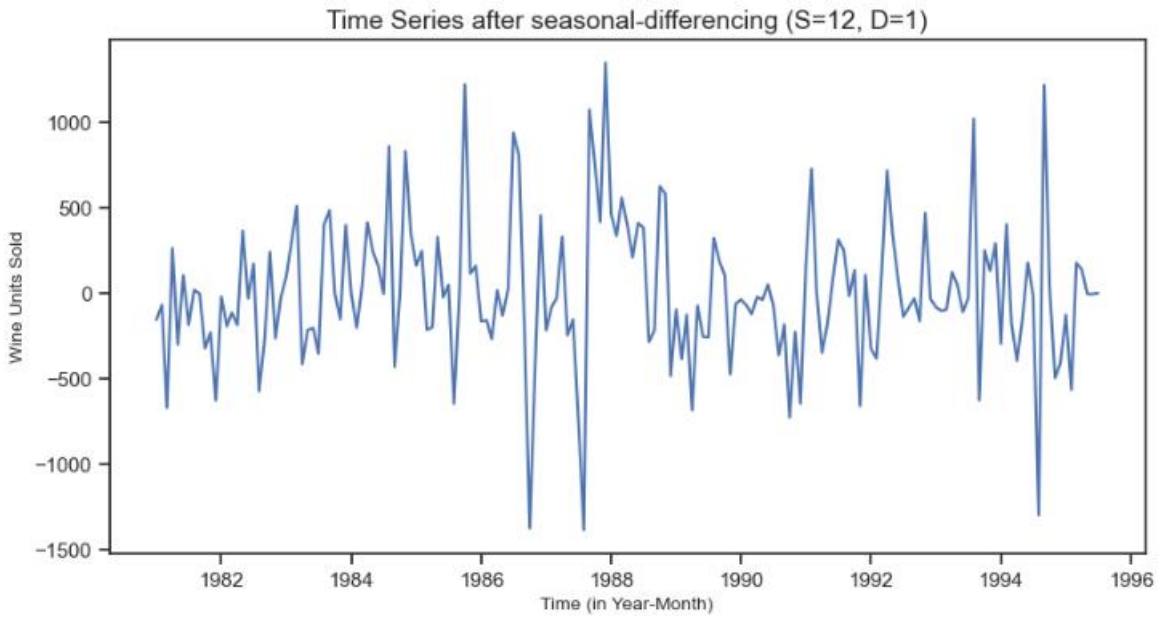

Figure 31 – Time series plot with deiffrencing S=12, D=1)

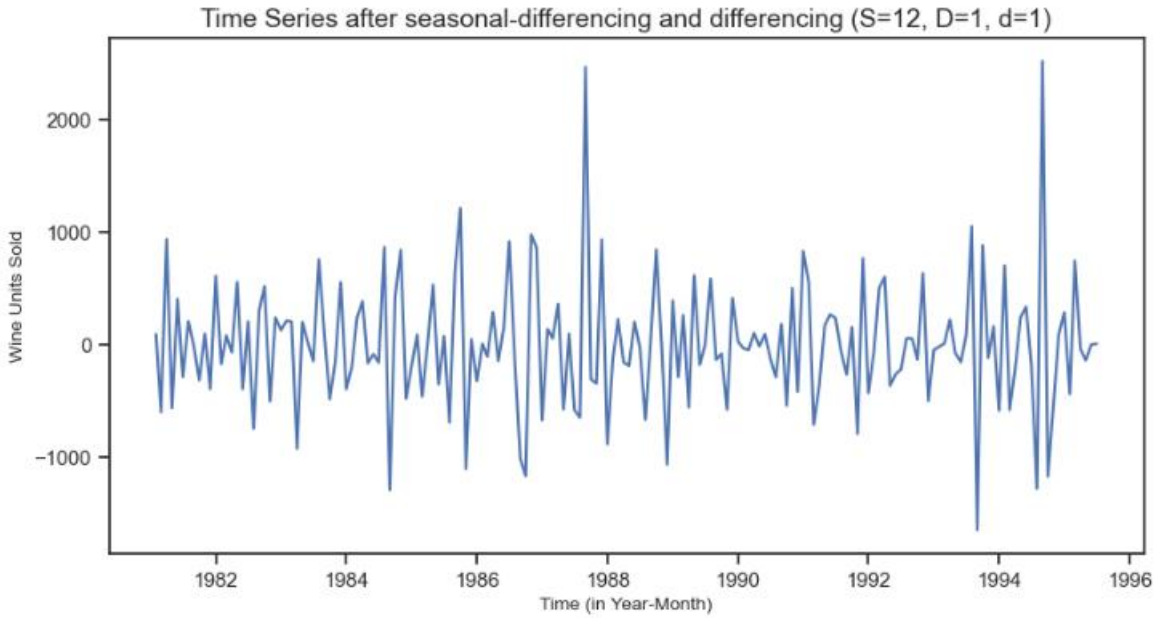

Figure 32 – Time series plot with deiffrencing S=12, D=1, d=1)

- **Removal of Seasonality:**
    - The most noticeable feature is the absence of the strong, repeating seasonal patterns that were likely present in the original data. The seasonal differencing has effectively removed the yearly seasonality.
- **Stationarity:**
    - The time series appears to be more stationary after seasonal differencing. The fluctuations seem to be more random around a mean of approximately zero.
- **Spikes and Outliers:** Several sharp spikes and potential outliers are visible, suggesting significant short-term variability in the data.
- **Fluctuations Around Zero:** The series fluctuates around zero, as expected after differencing

We have chosen **alpha = 0.05** and a **seasonal period of 12**.

- **PACF Analysis**: Significant lags are observed up to lag 4 before cutting off, so we set **AR term (p) = 4**. At the seasonal lag of 12, there is no significance, so **seasonal AR (P) = 0**.
- **ACF Analysis**: Significant lags appear at **lags 1 and 2**, so we set **MA term (q) = 2**. At the seasonal lag of 12, a significant lag is present, but no clear seasonal lags are observed at 24, 36, or beyond, so we set **seasonal MA (Q) = 1**.

Thus, the **final SARIMA model** is **(4,1,2) x (0,1,1,12)** based on the ACF and PACF plots.

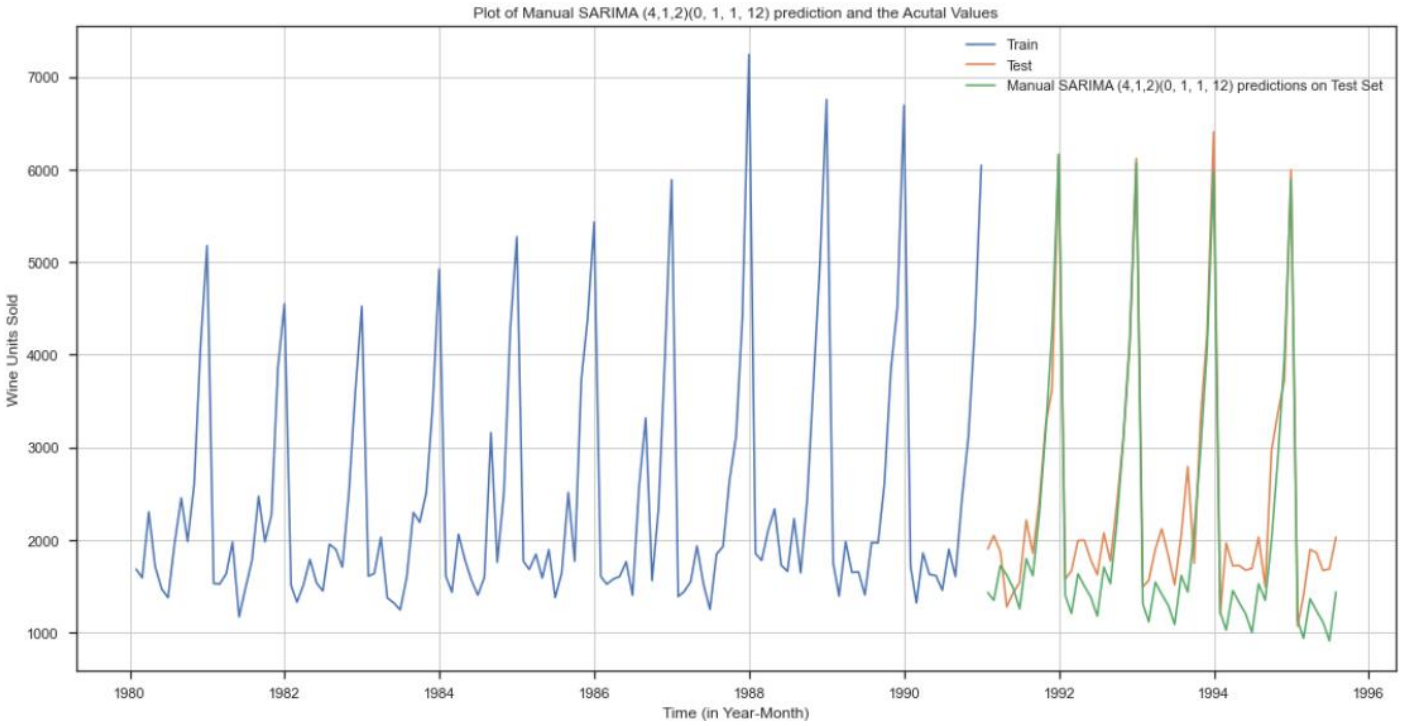

Figure 33 – Final Manual Sarima plot

- **Seasonality Captured:** The green line (SARIMA predictions) closely follows the seasonal pattern of the orange line (actual test data), indicating that the model successfully captured the seasonality.
- **Magnitude Discrepancies:** While the seasonality is captured, there are noticeable discrepancies in the magnitude of the predictions. The model sometimes underestimates or overestimates the actual values.
- **Lagging Predictions:** There's a slight lag in the model's predictions, particularly around the peaks and troughs. The green line often follows the orange line with a delay.
- **Model Complexity:** SARIMA(4, 1, 2)(0, 1, 1, 12) is a moderately complex model, with several parameters.

Finally, as per code,

	Test RMSE	MAPE
Auto ARIMA (4,1,4)	1212.918076	40.214639
Auto SARIMA (3,1,2)(3,0,1,12)	579.925219	25.052504
Manual ARIMA (2,1,1)	1300.721383	40.225669
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	468.677589	19.324914

Table 25– Manual-Sarima Results

Final Model building on analysing all models –

From the above results we can see that Triple exponential model is the optimum model followed by Trailing moving average models.

From below plot, we can confirm that TES is the best model.

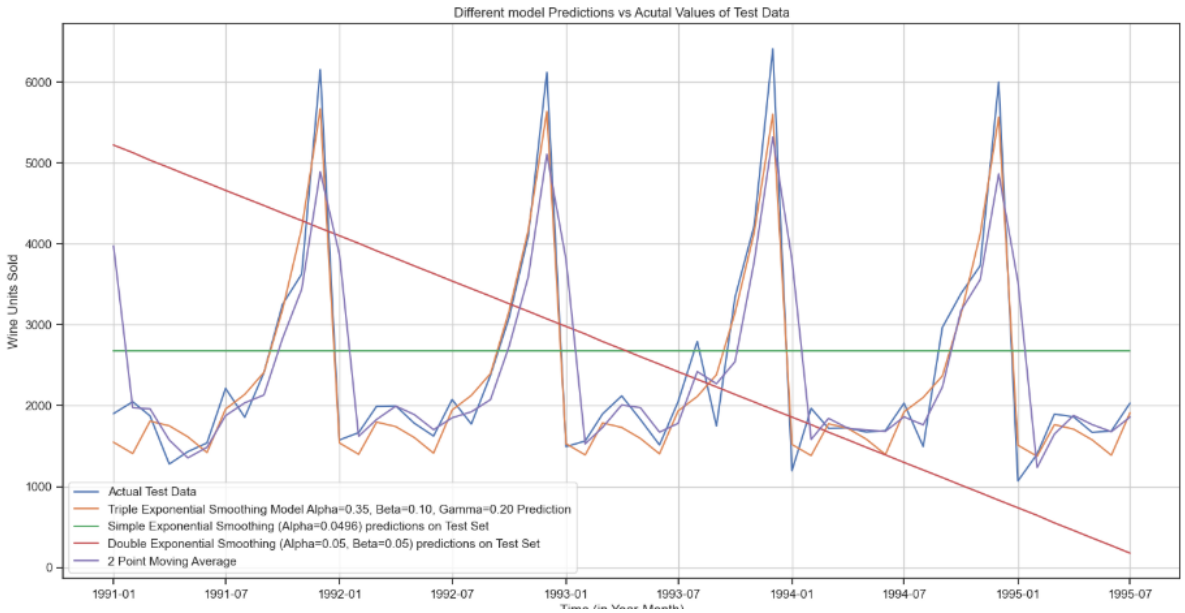

Figure 34 – Final comparison Plot

- **Actual Test Data (Blue Line):** This represents the true values of "Wine Units Sold" during the test period. It clearly shows a strong seasonal pattern with peaks and troughs.
- **Triple Exponential Smoothing (Orange Line):**
    - This model (with specific alpha, beta, and gamma parameters) closely follows the actual test data, capturing both the trend and seasonality.
    - It appears to be the most accurate model among those shown.
- **Simple Exponential Smoothing (Green Line):**
    - This model produces a flat, constant prediction.
    - It fails to capture any trend or seasonality, indicating it's a poor fit for this data.
- **Double Exponential Smoothing (Red Line):**
    - This model captures a downward trend but fails to capture the seasonality.

- It deviates significantly from the actual data, especially during peak periods.

- **2 Point Moving Average (Gray Line):**
  - This model follows the general trend of the data, but is very close to the actual test data.
  - It does not capture the seasonality well.

Out of TES and SARIMA models, lets deep dive to select our choice-

Optimum Model - Triple Exponential Smoothing Model (Alpha=0.35,Beta=0.10,Gamma=0.20)

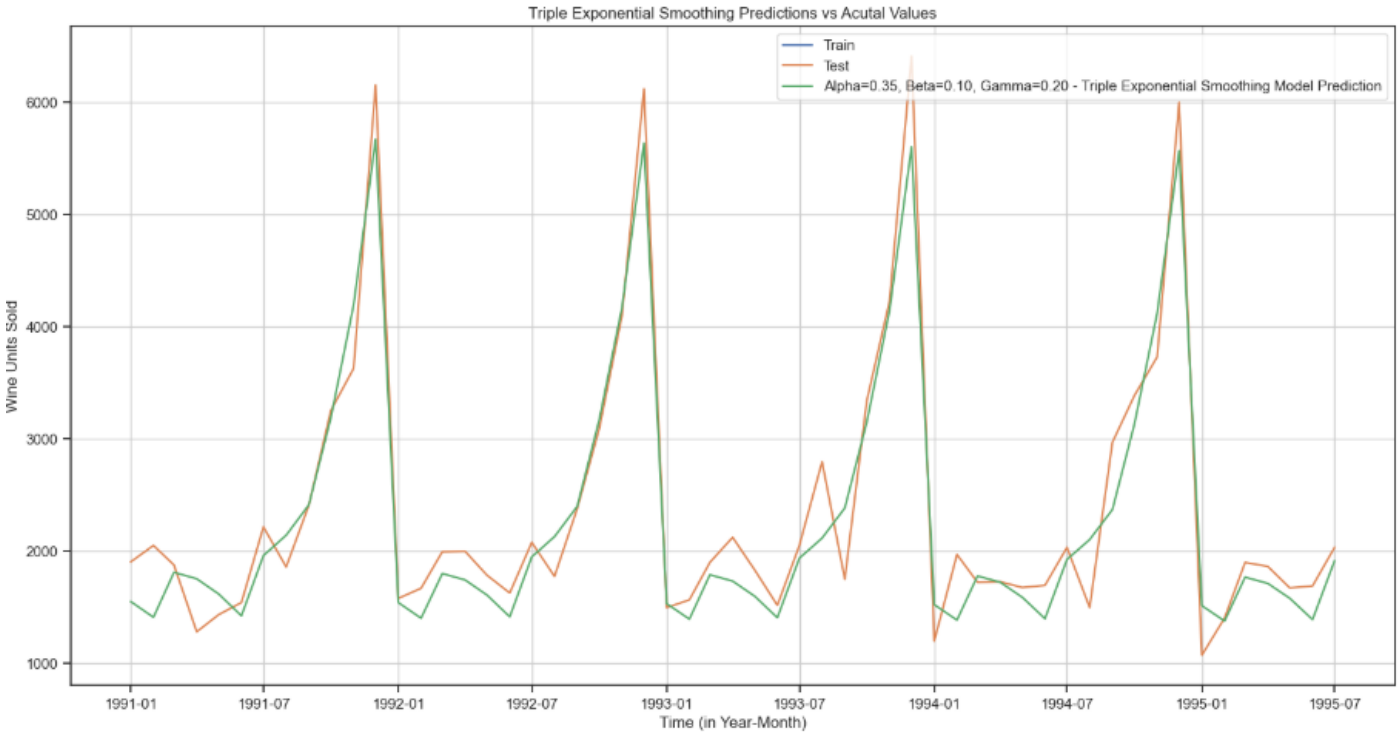

Figure 35 – TES predictions vs Actual values on both Test and Train data

- Excellent Seasonality Capture: The green line (model predictions) closely mirrors the orange line (actual test data), demonstrating the model's ability to accurately capture the seasonal patterns in the data.
- Good Fit to Trend: The model also captures the overall trend of the data, with the green line following the general direction of the orange line.
- Parameter Effectiveness: The chosen alpha, beta, and gamma values appear to be well-suited for the data, allowing the model to effectively capture both trend and seasonality.

Also on model checkup, we are getting RMSE as - 373.26457572243464

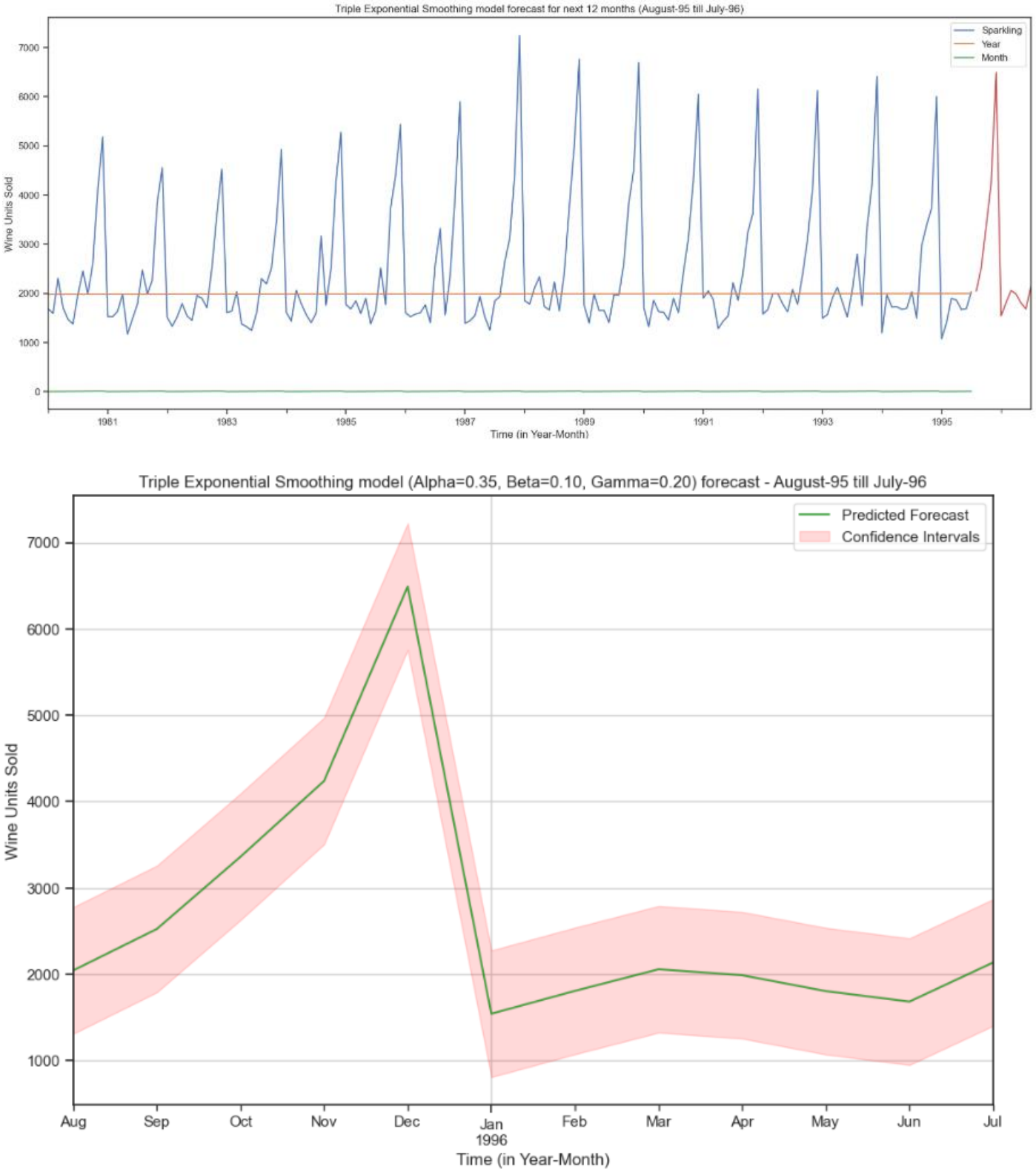

Figure 36 – TES predictions for next 12 months

On other hand, in Sarima model –

As er code, we get RMSE - 577.5088

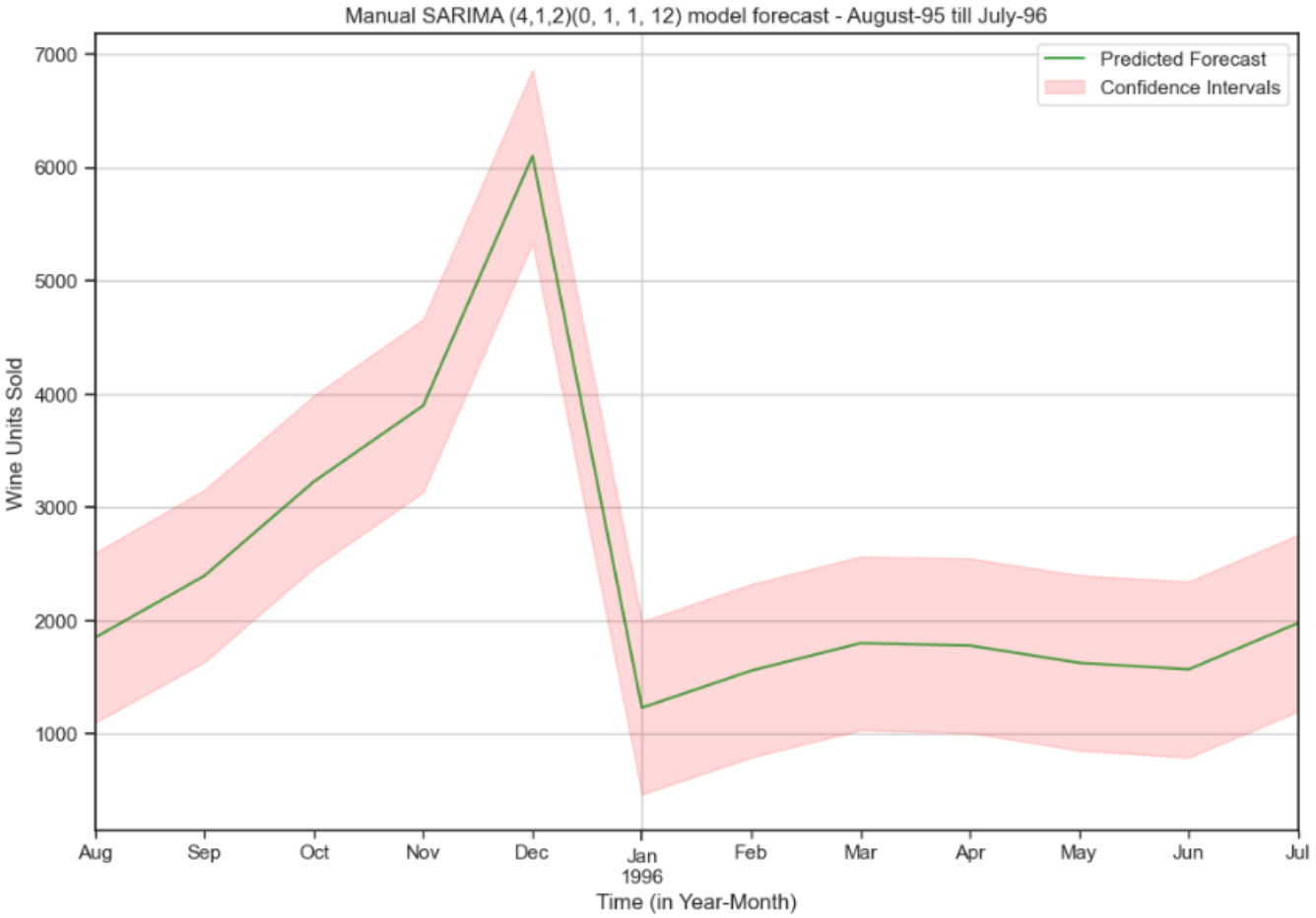

Figure 37 – Sarima predictions for next 12 months

Hence, TES is the optimum choice on our 12 month prediction on better RMSE values.

Actionable Insights & Recommendations

Model Insights:

- **Model Selection:** Triple Exponential Smoothing and SARIMA models were identified as most effective due to their ability to handle both trend and stable seasonality present in the data.
- **Performance Evaluation:** Model performance was assessed using Root Mean Squared Error (RMSE), with lower RMSE indicating better fit. Both Triple Exponential Smoothing and SARIMA demonstrated the lowest RMSE and closest alignment with test data.

Historical Insights:

- **Sales Trend:** Sparkling wine sales exhibited a slight upward trend with consistent seasonality. Peak sales occurred in 1988, with a low in 1995 (data to July).
- **Monthly Seasonality:** Sales increase progressively towards year-end, peaking in December and reaching a low in January. Sales gradually rise from January to August, then sharply increase.
- **Sales Distribution:** Average monthly sales were 2402 units, with 50% falling between 1605 and 2549 units. Sales ranged from 1070 to 7242 units, with 75% below 2549 units. The majority of sales (60-70%) were below 2500 units, and 80% were below 4000 units.

Forecast Insights (Triple Exponential Smoothing):

- **Increased Average Sales:** Forecasted average sales are 2639 units, a 10% increase from the historical average.
- **Higher Minimum Sales:** Forecasted minimum sales are 1540 units, a 43% increase from the historical minimum.
- **Lower Maximum Sales:** Forecasted maximum sales are 6487 units, a 10% decrease from the historical maximum.
- **Increased Volatility:** Forecasted standard deviation is 1439 units, an 11% increase from the historical standard deviation.
- **Seasonal Pattern:** Forecasted sales show increased activity in October, November, and December, with a sharp decline in January and gradual growth until October.

Recommendations:

- **Leverage Festive Season:** Focus on September to December, which account for 50% of forecasted sales, by implementing promotional offers, targeted marketing, and bulk order incentives.
- **Address Low Sales Period:** Conduct market research to understand factors impacting sales from January to June and consider introducing market-friendly product variations.

- **Enhance Forecast Accuracy:** Perform in-depth market research to identify additional factors influencing sales and incorporate them into the forecasting model.