****

TIME SERIES FORECASTING PROJECT

********** BUSINESS REPORT **********

Prepared By : Sahid
COURSE : DSBA
DATE : 17-11-2024

Table of Contents

List of Figures

List of Tables

1.Introduction

1.1 Project Overview

ABC Estate Wines, a renowned producer of fine wines, has a rich history spanning the 20th century. The company specializes in crafting diverse wine varieties, including Rosé and Sparkling wines, which are celebrated for their quality and taste. To sustain its competitive edge in an evolving market, ABC Estate Wines aims to leverage historical sales data for trend analysis and forecasting. This project focuses on analyzing and predicting sales patterns for Rosé and Sparkling wines to derive actionable insights for strategic decision-making.

1.2 Objectives

The primary objectives of this analysis are:

- To explore historical sales data for Rosé and Sparkling wines, identifying key trends and patterns.
- To develop forecasting models that provide accurate predictions for the next 12 months of wine sales.
- To compare sales dynamics between Rosé and Sparkling wines, uncovering unique characteristics of each variety.
- To offer actionable recommendations that support sales optimization and strategic growth initiatives.

By addressing these objectives, this report aims to empower ABC Estate Wines with data-driven insights, enabling informed decision-making and sustainable business success.

2.Dataset 1: Analysis and Forecasting for Rosé Wine

2.1 Dataset Overview and EDA

Dataset Description and Statistical Summaries

The dataset contains monthly sales data for Rosé wine from January 1980 onwards. Key attributes include sales volume and time periods, enabling a detailed temporal analysis of sales trends.

Summary Statistics for Rosé Wine Sales:

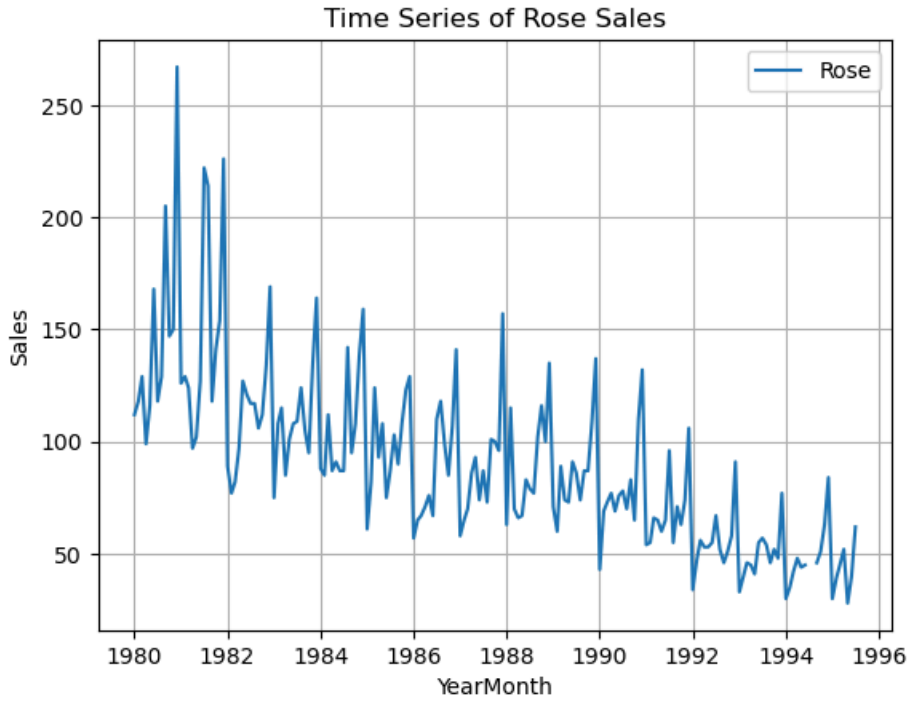
- **Count**: 185 months of recorded data.
- **Mean Sales**: 90.39 units per month.

- **Standard Deviation**: 39.18 units, reflecting moderate variability in monthly sales.
- **Minimum Sales**: 28 units (indicating an off-season or low-demand month).
- **Maximum Sales**: 267 units (representing a seasonal or promotional peak).
- **Quartiles**:
    - 25th Percentile: 63 units.
    - Median (50th Percentile): 86 units.
    - 75th Percentile: 112 units.

Visualization and Insights

Sales Trends Over Time:
The line graph below illustrates monthly sales of Rosé wine from January 1980. It reveals:

- An upward trend with moderate seasonal fluctuations.
- Peaks typically occurring during warmer months, aligning with higher demand for Rosé wine.
- Dips observed in winter months, reflecting potential off-season behavior.



("Figure 1: Monthly Sales of Rosé Wine (1980 Onwards).")

Additional insights from the graph:

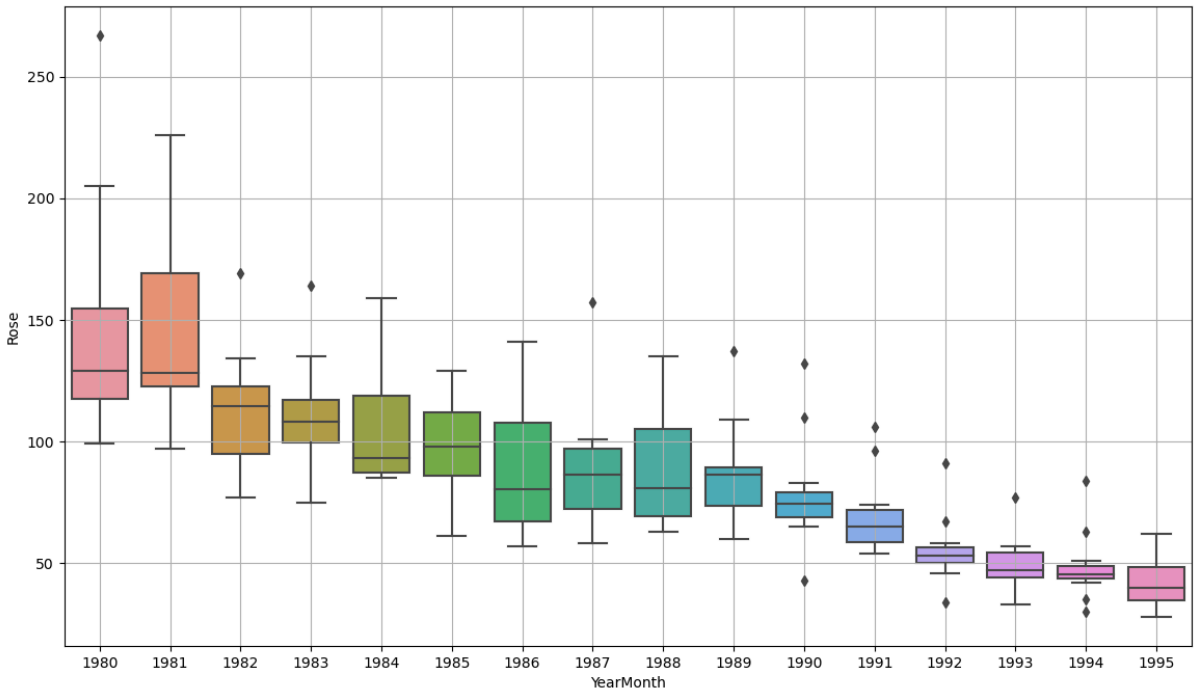
- Sales spikes, such as the peak at 267 units, may coincide with special promotions or events.
- The overall variability highlights opportunities for improved inventory planning during low-demand months.

Boxplot for Rose Sales Over Years:

- A boxplot was created to visualize the distribution of Rose wine sales over different years.
- The boxplot helps to highlight the spread of sales, potential outliers, and fluctuations over time.
- By analyzing the plot, we can identify if there are any years with exceptionally high or low sales, and it also provides insight into the overall consistency or variability in sales across the years.

Key Insights:

- The plot reveals the variation in sales, showing the central tendency (median) and the interquartile range (IQR).
- Outliers are easily visible, offering a deeper understanding of abnormal fluctuations in sales for certain years.



("Figure 2: Yearly Sales of Rosé Wine (1980 Onwards).")

2.2 Data Preprocessing

Handling Missing Values:

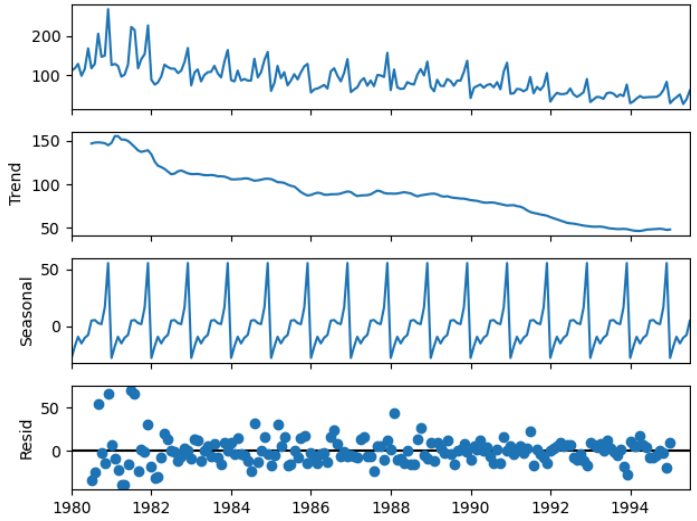
- **Method Used**: Linear Interpolation was applied to handle missing values in the **Rose** dataset.
- **Reason**: Linear interpolation estimates missing values by assuming a linear relationship between adjacent data points, which is suitable for time series data like monthly sales.

Data Transformation Details:

Seasonal Decomposition: The Rose dataset was decomposed into three components: Trend, Seasonality, and Residuals. Both additive and multiplicative models were tested.

Additive Model:

- **Trend**: Shows a steady underlying movement in sales, with slight peaks (e.g., in September).
- **Seasonality**: Reveals regular fluctuations, such as higher sales in December (+55.71) and lower sales in January (-27.91).
- **Residuals**: Captures random deviations from the trend and seasonality, with notable spikes in months like **December** (+66.16) and **July** (-33.98).

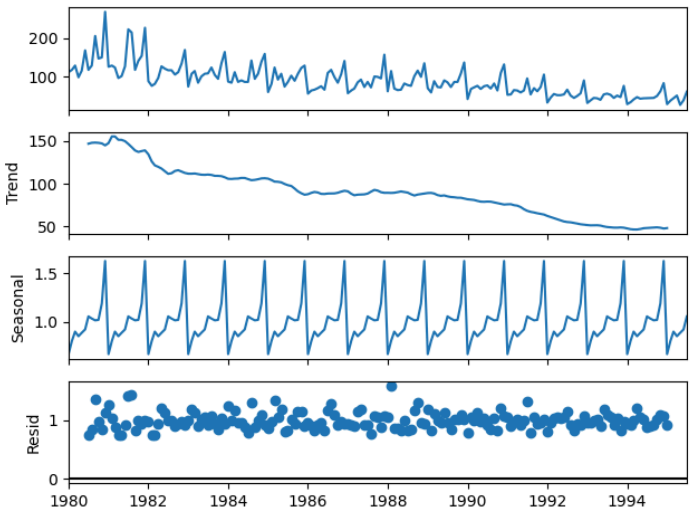


(Figure 3: Additive Decomposition Plot)

Multiplicative Model:

- **Trend**: Stable pattern with slight variations over time.
- **Seasonality**: Proportional fluctuations; for instance, **January** has a seasonal multiplier of 0.67 (lower sales), and **December** has a multiplier of 1.63 (higher sales).
- **Residuals**: Random noise, such as **July** with a residual of 0.76 and **December** with 1.13.

(Figure 4: Multiplicative Decomposition Plot)



Key Insight:

- **Multiplicative model** is more appropriate as seasonal effects are proportional to the trend, unlike the additive model which assumes fixed seasonal variations.

Data Splitting

In this section, the dataset was pre-processed and transformed to prepare it for modeling. The data was split into two subsets:

- **Training Data**: The first 70% of the dataset was used for training the models.
- **Test Data**: The remaining 30% was held out for validation and testing of model performance.

This split ensures that the model is tested on data it has not seen before, providing an unbiased estimate of its generalization capability.

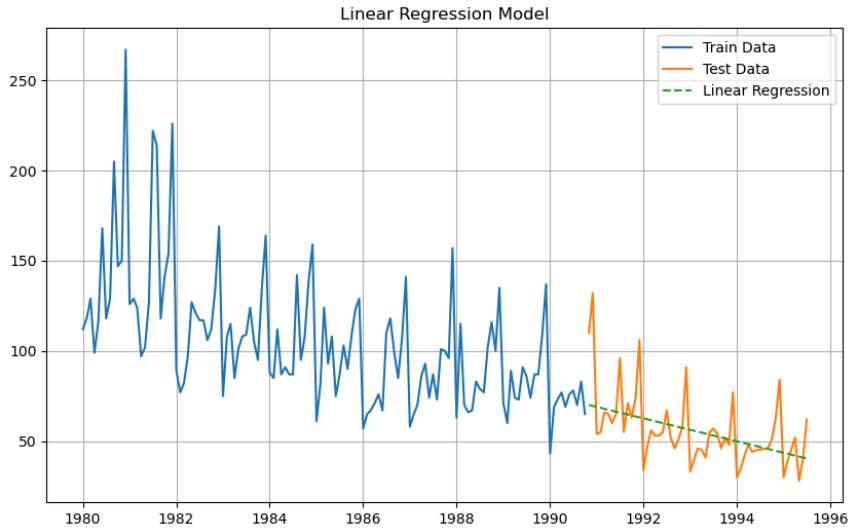
2.3 Forecasting Models

Linear Regression Model for Regression on Time Forecast on the Test Data

The Linear Regression model was applied to forecast the values based on the time series data. The model performance was evaluated using the Root Mean Squared Error (RMSE) metric.

- **RMSE**: 17.36

The performance of the model indicates a reasonable fit for the data, and the RMSE value reflects the average magnitude of error in the predictions.



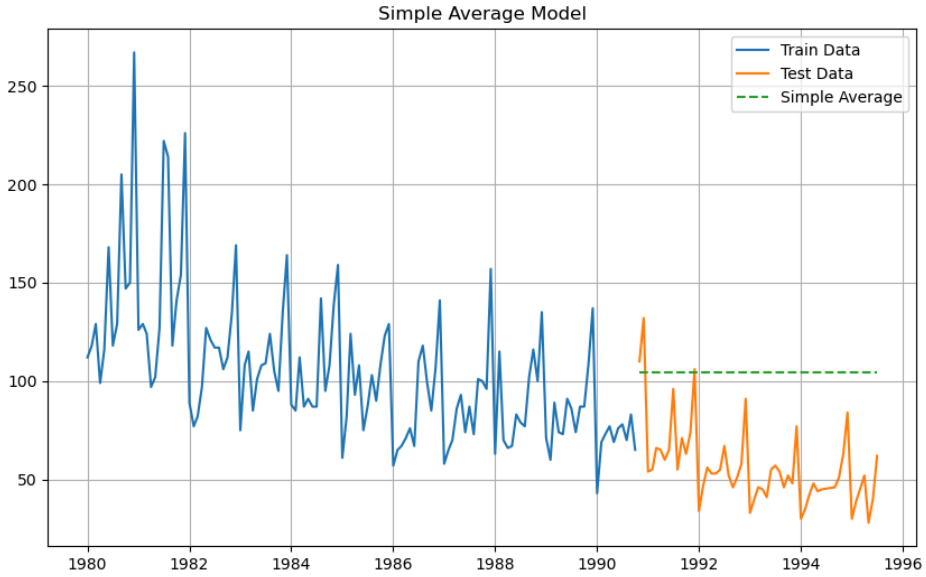
(Figure 5: Linear Regression Model Forecast vs Actual Values)

Simple Average Model for Forecast on the Test Data

The Simple Average model was applied to forecast the values based on the historical data, where the forecast for each period is the average of all previous observations.

- **RMSE**: 52.41

This relatively higher RMSE indicates that the Simple Average model does not capture the trends and seasonality as effectively as other models.



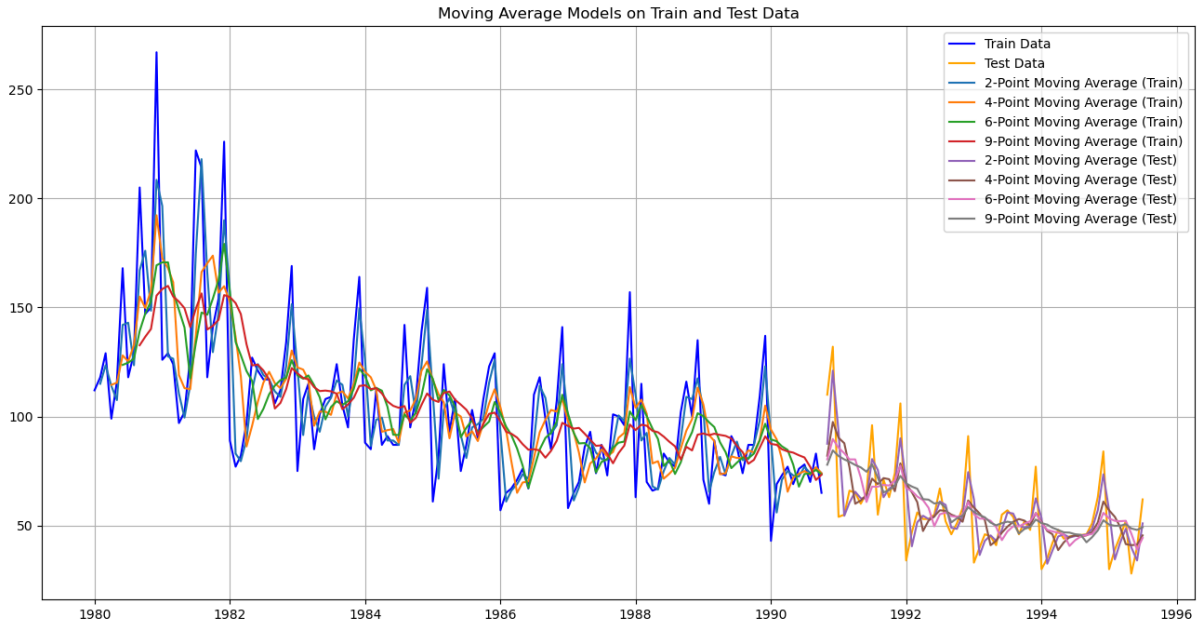
(Figure 6 : Simple Average Model Forecast vs Actual Values)

Moving Average (MA) Models for Forecast on the Test Data

In the Moving Average model, the forecast for a given period is calculated as the average of the last *n* observations. Below are the results for different point moving averages:

1. **2-point Moving Average**
    - **RMSE**: 11.801
2. **4-point Moving Average**
    - **RMSE**: 15.367
3. **6-point Moving Average**
    - **RMSE**: 15.862
4. **9-point Moving Average**
    - **RMSE**: 16.342

The 2-point moving average provides the lowest RMSE, indicating that it best captures the short-term trends in the data.



(Figure 7 : Moving Average Model Forecasts vs Actual Values)

Simple Exponential Smoothing (SES) Model for Forecast on the Test Data

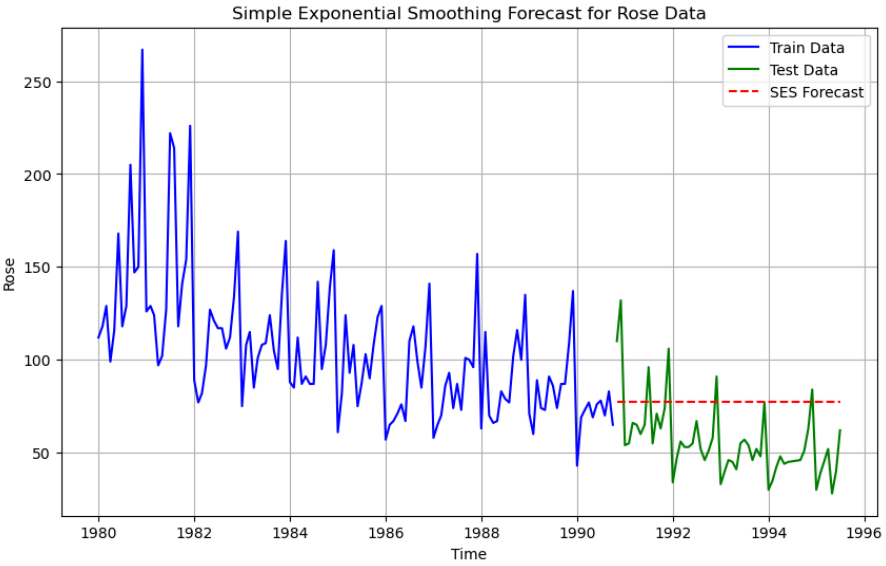
The Simple Exponential Smoothing (SES) model is a time series forecasting method that predicts future values based on the weighted average of past observations. The optimized parameters for the SES model are as follows:

- **Smoothing Level**: 0.128
- **Smoothing Trend**: Not applicable (N/A)
- **Smoothing Seasonal**: Not applicable (N/A)
- **Initial Level**: 112.0
- **Initial Trend**: N/A
- **Initial Seasons**: N/A
- **Box-Cox Transformation**: Not applied

For the Simple Exponential Smoothing forecast on the Test Data, the RMSE is:

- **RMSE**: 29.224

The SES model performs reasonably well, but other models like Moving Average or RegressionOnTime provide better accuracy.



(Figure 8: Simple Exponential Smoothing Forecast vs Actual Value)

Double Exponential Smoothing (Holt's Method) for Forecast on the Test Data

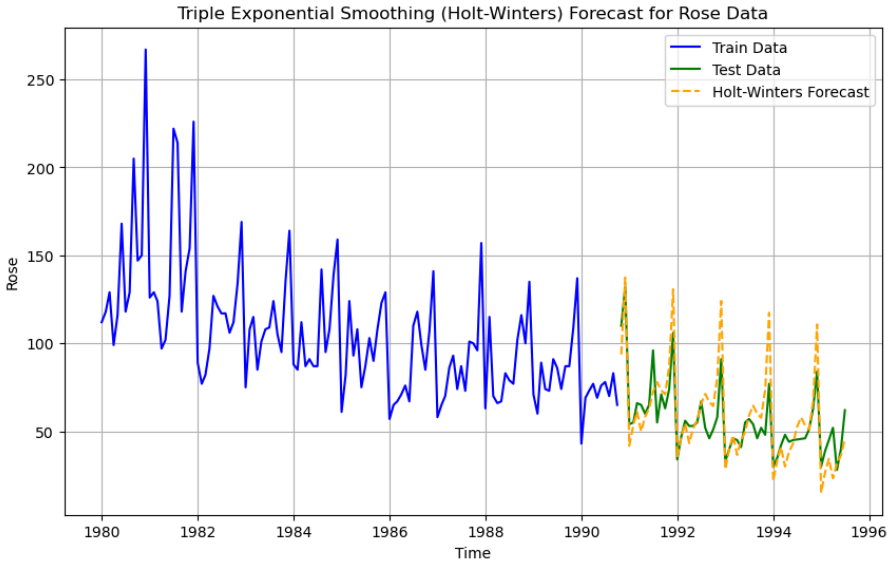
The Double Exponential Smoothing method, also known as **Holt's method**, extends Simple Exponential Smoothing by adding a trend component to account for changes in the level of the series over time. The optimized parameters for Holt's method are as follows:

- **Smoothing Level**: 0.152
- **Smoothing Trend**: 0.152
- **Initial Level**: 112.0
- **Initial Trend**: 6.0
- **Smoothing Seasonal**: Not applicable (N/A)
- **Damping Trend**: Not applied
- **Box-Cox Transformation**: Not applied

For the Holt's method forecast on the Test Data, the RMSE is:

- **RMSE**: 26.050

Holt's method performs better than Simple Exponential Smoothing, providing a more accurate forecast on the test data.



(Figure 9: Holt's Method Forecast vs Actual Values)

Forecasting Models - RMSE Comparison

This section compares the RMSE values across different forecasting models, summarizing their performance on the Test Data. The following table presents the RMSE values for each model:

Model	RMSE
Regression on Time	17.36
Simple Average	52.41
2-point Trailing Moving Average	11.8
4-point Trailing Moving Average	15.37
6-point Trailing Moving Average	15.86
9-point Trailing Moving Average	16.34
Simple Exponential Smoothing	29.22
Double Exponential Smoothing	26.05
Triple Exponential Smoothing	13.96

(Table 1 : Model RMSE Comparison)

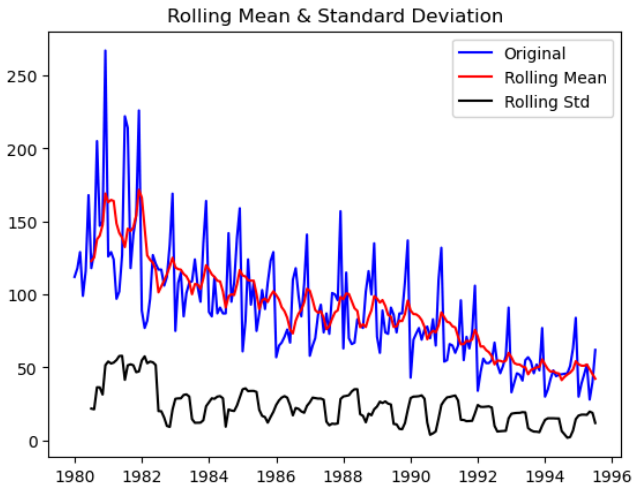
Key Insights

- The **2-point Trailing Moving Average** shows the lowest RMSE value, suggesting it provides the most accurate forecast compared to other models.
- **Simple Average** has the highest RMSE, indicating it performs poorly in capturing trends and seasonality.
- **Triple Exponential Smoothing** also performs relatively well, with an RMSE of **13.96**, slightly worse than the 2-point moving average but still quite competitive.

Stationarity Check

Results of Dickey-Fuller Test (Original Series):

- **Test Statistic**: -1.8767
- **p-value**: 0.3431
- **Lags Used**: 13
- **Number of Observations Used**: 173
- **Critical Value (1%)**: -3.4687
- **Critical Value (5%)**: -2.8784
- **Critical Value (10%)**: -2.5758



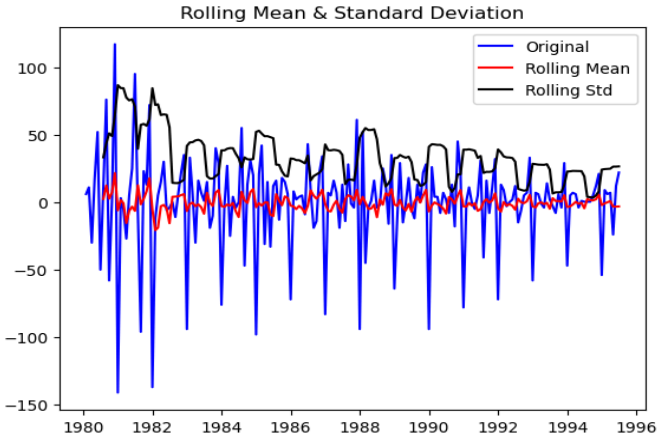
(Figure 10: Dickey-Fuller Test - Original Time Series Plot)

Interpretation:

The test statistic (-1.8767) is not less than the critical values at any significance level (1%, 5%, or 10%). The p-value (0.3431) is much higher than typical thresholds (e.g., 0.05 or 0.01). Therefore, we conclude that the time series is **non-stationary**. This suggests the presence of a trend or seasonality, or the statistical properties (mean, variance, autocorrelation) change over time.

Results of Dickey-Fuller Test (Differenced Series):

- **Test Statistic**: -8.0444
- **p-value**: 1.8109e-12
- **Lags Used**: 12
- **Number of Observations Used**: 173
- **Critical Value (1%)**: -3.4687
- **Critical Value (5%)**: -2.8784
- **Critical Value (10%)**: -2.5758



(Figure 11 : Dickey-Fuller Test - Differenced Time Series Plot)

Interpretation:

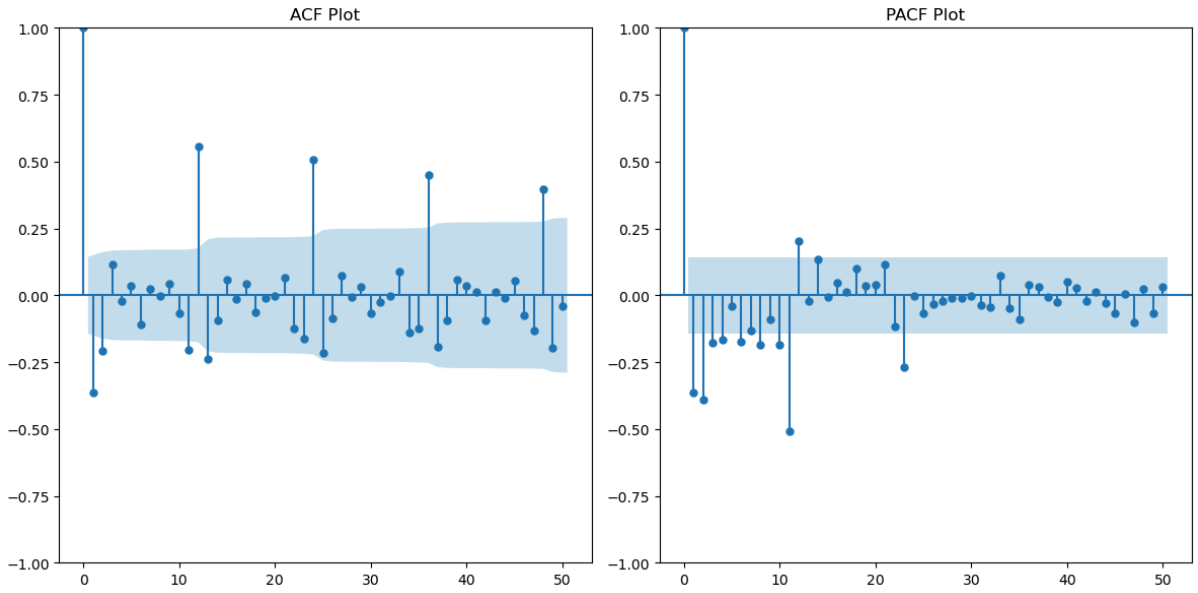
Since the test statistic (-8.0444) is smaller than the critical values and the p-value (1.8109e-12) is extremely low, we can conclude that the **differenced series is stationary**. This confirms that the time series has been transformed to achieve stationarity, making it suitable for further modeling.

Autocorrelation and Partial Autocorrelation Functions (ACF & PACF)

To assess the autocorrelations and determine the appropriate lag order for the ARIMA model, we plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for the differenced time series.

- **ACF Plot:** The ACF plot helps identify the number of moving average (MA) terms needed for the ARIMA model. Significant spikes at specific lags indicate correlations in the residuals.
- **PACF Plot:** The PACF plot assists in identifying the number of autoregressive (AR) terms. Significant spikes at specific lags suggest where the correlation between an observation and its lag is no longer significant once previous lags have been accounted for.

(Figure 12 ACF and PACF Plots for Differenced Data)



.

Both the ACF and PACF plots provide critical insights for selecting the optimal AR and MA terms in the ARIMA model.

Auto ARIMA Model Selection

In this step, we explore several combinations of ARIMA (AutoRegressive Integrated Moving Average) parameters to identify the optimal model. The parameters are expressed in the form (p, d, q), where:

- **p** is the number of lag observations in the model (AR term).
- **d** is the degree of differencing (I term).
- **q** is the size of the moving average window (MA term).

Several parameter combinations were tested, and the AIC (Akaike Information Criterion) values for each combination were calculated to determine the best-fitting model. A lower AIC indicates a better fit.

SARIMAX Results for the Selected Model (0, 1, 2)

Model Selection:

- The optimal model selected by Auto ARIMA is **SARIMAX(0, 1, 2)**, with the lowest **AIC** value of **1259.25**, indicating the best fit.

Model Coefficients:

- **ma.L1**: -0.7059 (significant, indicating a negative impact from past errors)
- **ma.L2**: -0.1915 (also negative but with a smaller effect)
- **sigma2**: 958.60 (variance of residuals, showing significant error variability)

Statistical Significance:

- All coefficients have very low p-values ($p < 0.01$), indicating they are statistically significant.

Diagnostic Tests:

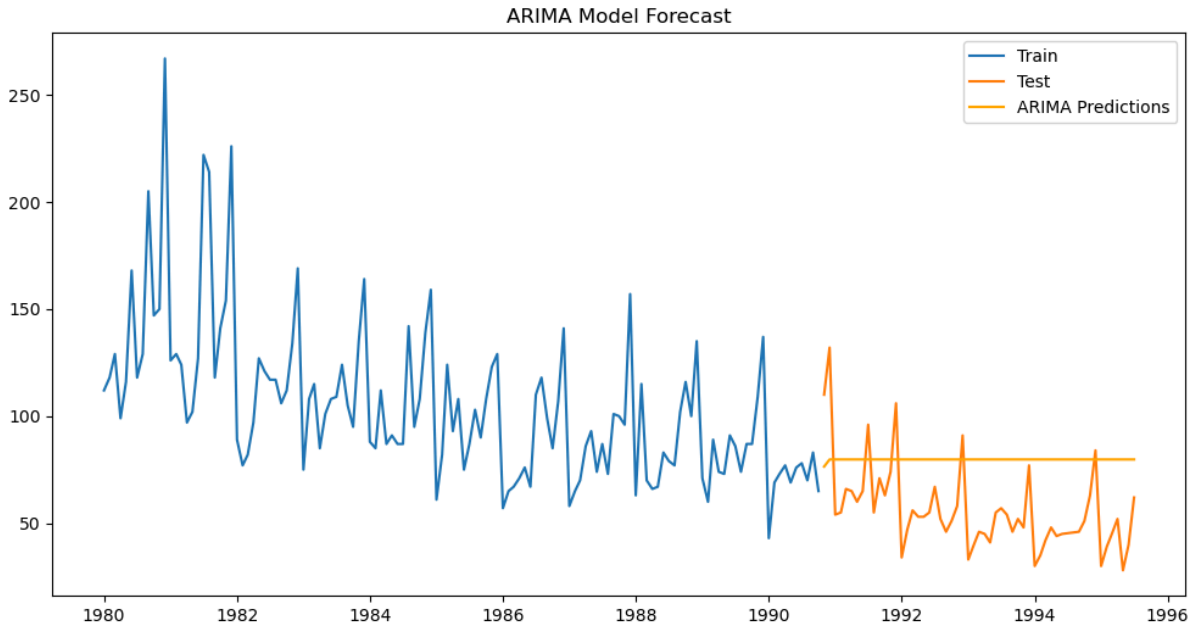
- **Ljung-Box (Q = 0.15, p = 0.70)**: No autocorrelation in residuals.
- **Jarque-Bera (p = 0.00)**: Residuals are not normally distributed.
- **Heteroskedasticity (p = 0.00)**: Residual variance is not constant over time.
- **Skewness**: 0.88 (slightly positively skewed residuals).
- **Kurtosis**: 5.34 (leptokurtic residuals).

Model Performance:

- **RMSE**: 30.90, indicating the model's forecast error.

Conclusion:

- The **SARIMAX(0, 1, 2)** model provides a good fit, with reasonable forecasting accuracy (RMSE of 30.90). However, issues like non-normal residuals and heteroskedasticity suggest potential areas for improvement.



(Figure 13 : Auto ARIMA Model Forecast for Test Data)

Manual ARIMA Model

Stationarity Check:

The Dickey-Fuller test was conducted on the differenced series of 'Rose' (after one differencing), with a p-value of 1.810895×10−121.810895 \times 10^{-12}1.810895×10−12. Since the p-value is significantly less than the 0.05 threshold, it confirms that the differenced series is stationary. Therefore, a differencing order of d=1d = 1d=1 is sufficient to make the series stationary.

AR and MA Orders:

- **AR Order (p):** Based on the Partial Autocorrelation Function (PACF) plot, the AR order p=1p = 1p=1, as the PACF cuts off after lag 1.
- **MA Order (q):** Based on the Autocorrelation Function (ACF) plot, the MA order q=1q = 1q=1, as the ACF cuts off after lag 1.

ARIMA Model Parameters:

- **AR Order (p) = 1** (from PACF plot)
- **Differencing Order (d) = 1** (from stationarity test)
- **MA Order (q) = 1** (from ACF plot)

SARIMAX Model Results:

- **Log Likelihood:** -627.018
- **AIC:** 1260.037
- **BIC:** 1268.616
- **Sigma Squared (Variance of errors):** 964.5521

Key Coefficients:

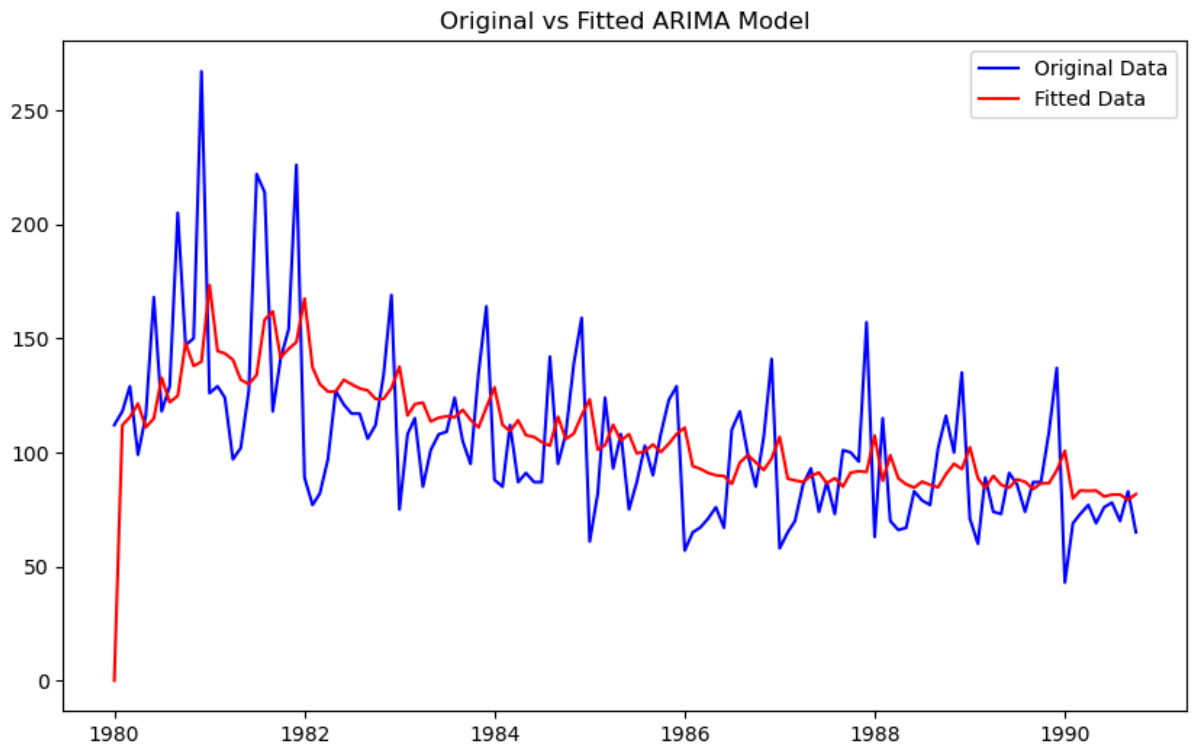
- AR(1) coefficient: 0.1709 (significant at p=0.027p = 0.027p=0.027)
- MA(1) coefficient: -0.9150 (significant at p<0.001p < 0.001p<0.001)

Diagnostics:

- **Ljung-Box Test (Q):** 0.00, p=0.99p = 0.99p=0.99, indicating no significant autocorrelation in residuals.
- **Jarque-Bera Test:** 48.21, p=0.00p = 0.00p=0.00, suggesting that the residuals are not normally distributed.
- **Heteroskedasticity Test:** 0.31, p=0.00p = 0.00p=0.00, indicating heteroskedasticity in the residuals.
- **Skew:** 0.92 and **Kurtosis:** 5.36 suggest non-normality of the residuals.

RMSE:

The Root Mean Squared Error (RMSE) for the Manual ARIMA (1, 1, 1) model is **30.90**, indicating the model's prediction accuracy.



(Figure 14 : Comparison of Original Data and Fitted ARIMA Model (1, 1, 1))

The **Manual ARIMA(1, 1, 1)** model provides a reasonable fit to the data with an **RMSE of 30.90**, indicating that the model's predictions are, on average, 30.90 units off from the actual values. While the model successfully captures the temporal structure and residuals show no significant autocorrelation, there is some non-normality and heteroskedasticity in the residuals, suggesting room for improvement. Further tuning of the model or exploration of more advanced techniques may help reduce prediction errors and improve overall accuracy.

Auto SARIMA Model Results

The Auto SARIMA model was applied to forecast the time series data. Below are the key results and performance metrics:

Model Configuration:

- **SARIMAX Model**: (2, 1, 2)x(1, 0, [1], 12)
    - **AR (Auto-Regressive) Terms**: AR(1) and AR(2)
    - **MA (Moving Average) Terms**: MA(1) and MA(2)
    - **Seasonal Terms**: Seasonal AR(1) and Seasonal MA(1) with a period of 12 (suggesting monthly data with annual seasonality).

Model Parameters:

- **AR(L1)**: -0.4931, significant at the 1% level (p = 0.011).
- **AR(L2)**: -0.0634, not statistically significant (p = 0.548).
- **MA(L1)**: -0.2101, not statistically significant (p = 0.263).

- **MA(L2)**: -0.6451, significant at the 1% level (p = 0.001).
- **Seasonal AR(L12)**: 0.9878, highly significant (p = 0.000).
- **Seasonal MA(L12)**: -0.8088, highly significant (p = 0.000).
- **Variance of errors (sigma²)**: 438.6465.

Model Fit Statistics:

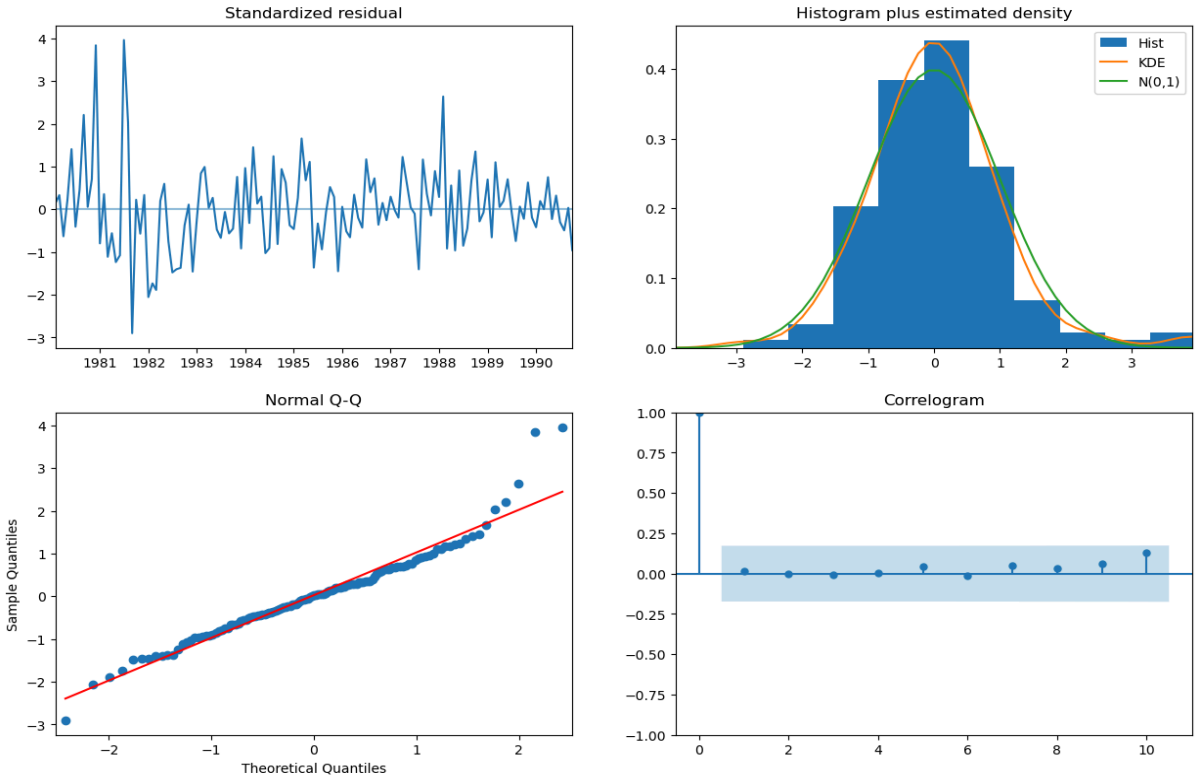
- **Log Likelihood**: -585.655
- **AIC**: 1185.311 (lower is better for model comparison).
- **BIC**: 1205.330
- **HQIC**: 1193.445

Diagnostic Statistics:

- **Ljung-Box Q Test (L1)**: p-value = 0.83 (no significant autocorrelation in residuals).
- **Jarque-Bera Test**: p-value = 0.00 (residuals deviate from normality, with skew and kurtosis present).
- **Heteroskedasticity**: p-value = 0.00 (evidence of varying variance in residuals).
- **Skew**: 0.75 (indicating positive skewness in residuals).
- **Kurtosis**: 5.70 (indicating fat tails, suggesting extreme outliers).

Performance:

- **RMSE**: 12.89, indicating the model's prediction error.



(Figure 15 showing Residual Diagnostics of Auto SARIMA Model)

Insights:

- The model successfully captures seasonality and some autoregressive patterns, with key seasonal components being highly significant.

- The residual diagnostics suggest some areas for improvement, particularly around normality and heteroskedasticity.
- The RMSE of 12.89 reflects a reasonable prediction performance, but there may be opportunities to enhance the model, potentially by addressing the residuals' distribution issues.

Manual SARIMA Model

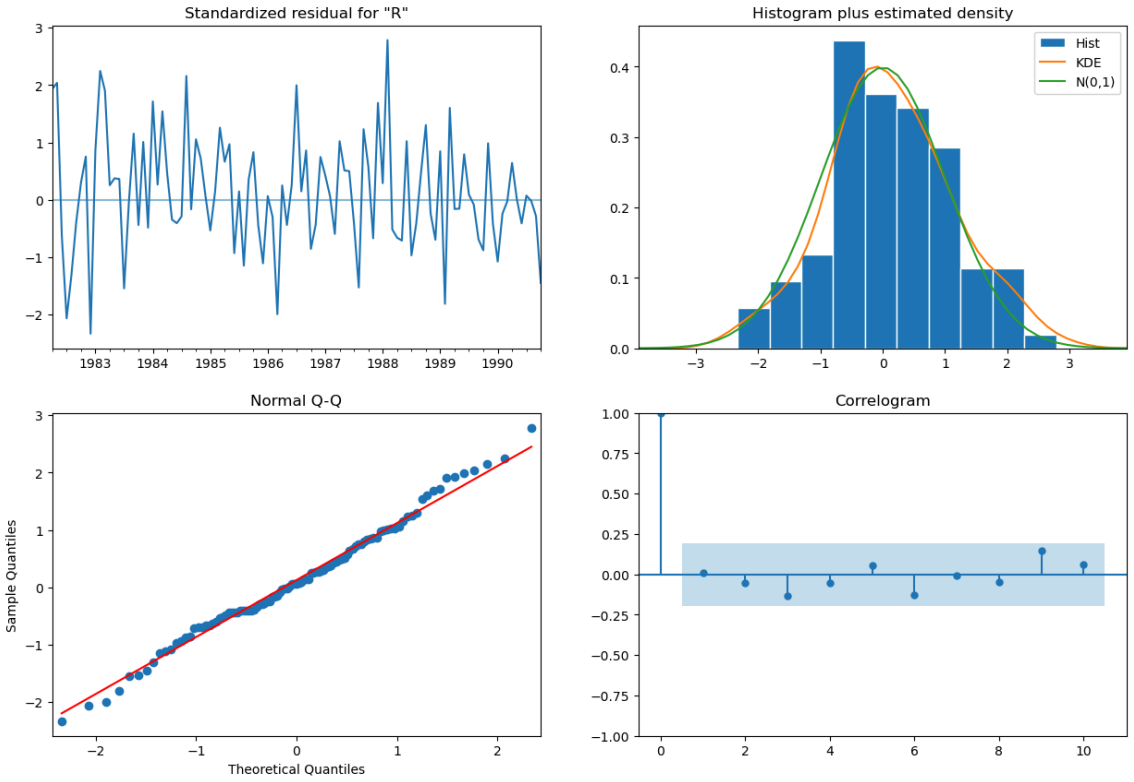
The **Manual SARIMA(1, 1, 1)x(1, 1, 1, 12)** model was fitted to the dataset, resulting in the following key metrics:

- **Log-Likelihood**: -450.444
- **AIC**: 910.888
- **BIC**: 924.062
- **RMSE**: 12.59

The model captures both seasonal and non-seasonal components, with significant coefficients for both the AR and MA terms in the seasonal and non-seasonal parts. Notably, the **MA(L1)** term is highly significant, indicating that past errors have a strong influence on the current value. However, the **AR(S.L12)** term also shows significance, suggesting that the seasonal lag has a notable effect.

Residual Diagnostics:

- **Ljung-Box Test**: The residuals show no significant autocorrelation (p-value = 0.94), suggesting that the model adequately captures the temporal dependencies in the data.
- **Jarque-Bera Test**: The p-value of 0.86 suggests the residuals are approximately normally distributed, which is a positive indicator for model validity.
- **Heteroskedasticity**: The test result (p-value = 0.10) indicates no significant heteroskedasticity, suggesting that the residuals' variance is stable over time.



(Figure 16 showing Residual Diagnostics of Manual SARIMA Model)

Insights:

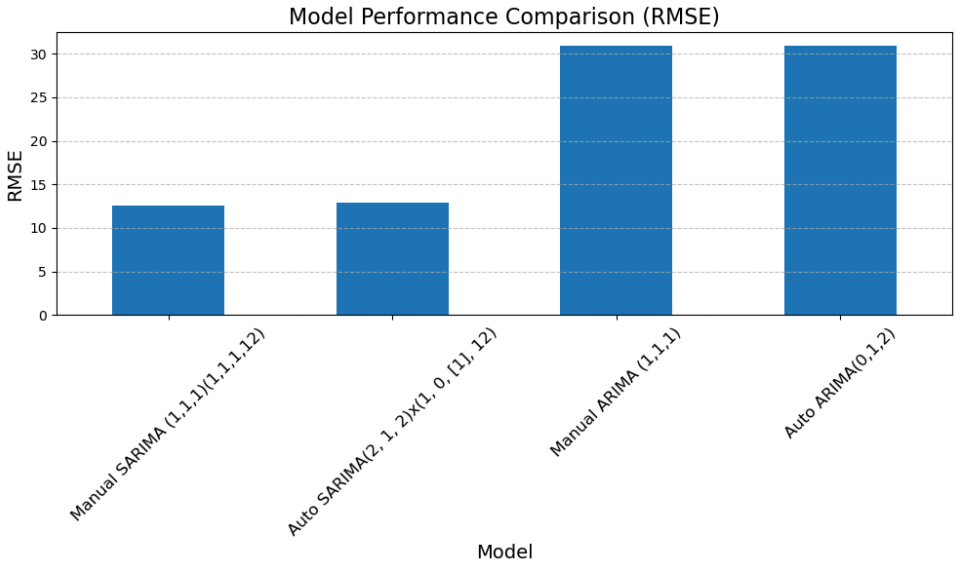
The **Manual SARIMA(1, 1, 1)x(1, 1, 1, 12)** model provides a good fit with a relatively low RMSE of **12.59**, indicating accurate predictions. The seasonal components and residual diagnostics suggest that the model adequately captures both short-term and long-term dependencies in the data. However, further refinement could potentially improve the model by fine-tuning the seasonal order or exploring additional external factors.

Model Comparison

Performance of All Models:

Performance of All Models:	
Model	RMSE
Auto ARIMA (0,1,2)	30.9
Manual ARIMA (1,1,1)	30.9
Auto SARIMA (2,1,2)x(1,0,[1],12)	12.89
Manual SARIMA (1,1,1)(1,1,1,12)	12.59

(Table 2 showing Performances of the Models)



(Fig 17 showing Performances of the Models)

Observations and Insights:

- **Best Performing Model**: The **Manual SARIMA (1,1,1)(1,1,1,12)** model has the lowest RMSE of **12.59**, making it the best-performing model for forecasting the Rose sales data. This model effectively captures both the seasonal and non-seasonal patterns in the data, leading to superior performance.
- **Auto SARIMA vs. Manual SARIMA**: The **Auto SARIMA** model, with an RMSE of **12.89**, performs almost as well as the **Manual SARIMA** model. The small difference in performance suggests that the automated approach closely approximated the manually tuned parameters for this dataset.
- **ARIMA vs. SARIMA**: The **ARIMA** models (both auto and manual) have significantly higher RMSE values (**30.90**), indicating that they fail to capture the seasonal fluctuations in the data. This highlights the importance of considering seasonality in time series forecasting, as **SARIMA** models outperform **ARIMA** models.

Choosing the Best Model:

- **Rationale**: The **Manual SARIMA (1,1,1)(1,1,1,12)** is the best model due to its lowest RMSE (**12.59**), indicating its superior ability to forecast the sales data. This model successfully accounts for seasonality and trends, making it more suitable for the given data than the ARIMA models.

Rebuilding the Best Model Using the Entire Data:

The **Manual SARIMA (1,1,1)(1,1,1,12)** model was retrained on the entire dataset of 187 observations, and the following results were obtained:

- **Log-Likelihood**: -674.933
- **AIC**: 1359.865
- **BIC**: 1375.241

Key coefficients:

- **AR(L1)**: 0.196 (significant)
- **MA(L1)**: -0.914 (highly significant)
- **Seasonal AR(L12)**: -0.408 (significant)
- **Seasonal MA(L12)**: 0.014 (not significant)

Residual Diagnostics:

- **Ljung-Box Test**: p-value = 0.91 (no significant autocorrelation)
- **Jarque-Bera Test**: p-value = 0.07 (residuals are approximately normally distributed)
- **Heteroskedasticity**: p-value = 0.00 (no significant heteroskedasticity)

Next 12-Month Forecast

(Table 3 Showing Forecast for the Next 12 Months:)

Date	Forecasted Rose Sales
1995-08-01	48.83
1995-09-01	43.47
1995-10-01	48.4
1995-11-01	53.94
1995-12-01	78.05
1996-01-01	26.87
1996-02-01	34.28
1996-03-01	40.68
1996-04-01	47.32
1996-05-01	31.21
1996-06-01	38.8
1996-07-01	52.25

Key Insights for Forecast:

- **Seasonal Pattern**: The sales of Rose are influenced by yearly seasonality, with significant peaks (e.g., December) and troughs (e.g., January).
- **High-Demand Periods**: The forecast predicts peak sales in **December 1995** (78.05), signaling a potential high-demand period for inventory and resource planning.
- **Low-Demand Periods**: **January 1996** shows a forecasted dip in sales (26.87), highlighting the need to adjust inventory and production capacity accordingly.

Recommendations:

1. **Inventory Management**: Prepare for higher sales in **December** and **July**, while managing lower stock levels for **January** and **May** to prevent overstocking.
2. **Sales Strategy**: Focus marketing and promotions around high-demand months, particularly **December**, to maximize sales during peak periods.
3. **Resource Allocation**: Align staffing and production resources with forecasted peak demand periods to ensure smooth operations and avoid stockouts.

By leveraging these insights, the business can plan more effectively for the upcoming year, optimizing inventory levels, marketing efforts, and resource allocation to meet the forecasted demand.

2.4 Insights and Recommendations for Rose Sales Data

Insights:

1. **Seasonal Demand Patterns**:
   The sales of Rose exhibit clear seasonal trends, with significant peaks and troughs in demand. For example:
   - **December 1995** shows a sharp increase in sales (78.05), indicating a peak period.
   - **January 1996** shows a significant dip in sales (26.87), suggesting a low-demand period immediately after the holiday season.
   - **July 1996** (52.25) also appears to be a high-demand month, signaling another seasonal spike.
2. **Trend and Seasonality**: The best-performing **Manual SARIMA (1,1,1)(1,1,1,12)** model effectively captures both the seasonal and trend components of the sales data. The model indicates that:
   - Sales tend to be higher during the latter part of the year (e.g., November and December).
   - The demand decreases in the first quarter of the new year (e.g., January).
3. **Fluctuating Demand**: The model indicates fluctuating demand with specific months, such as **August 1995** (48.83) and **October 1995** (48.40), showing stable but moderate sales levels compared to the sharp peaks and valleys seen in other months.

Recommendations:

1. **Inventory Management**:
   - **Stock Levels**: Prepare for higher demand during **December 1995** and **July 1996** by increasing inventory prior to these months. Lower stock levels are advisable during months like **January** (26.87) to avoid overstocking.
   - **Safety Stock**: Maintain safety stock levels during peak periods to prevent stockouts and ensure timely fulfillment of customer orders.
2. **Sales Strategy**:
   - **Marketing Focus**: Plan promotional campaigns during high-demand months like **December** and **July** to take advantage of the forecasted sales spikes. Consider offering discounts, bundles, or special offers to boost sales during these months.
   - **Targeted Campaigns**: Consider tailoring marketing efforts for the low-demand months like **January**, possibly using promotions to stimulate demand, or plan to reduce marketing spend during the off-peak months to optimize budget allocation.
3. **Production Planning**:

    - Align production schedules with the seasonal demand fluctuations. For example, ramp up production in **October-November** to prepare for the December spike, and slow down production during January and February when sales are expected to be lower.
4. **Resource Allocation**:
    - Ensure that sufficient staffing, warehousing, and distribution resources are allocated during high-demand months like **December** and **July** to efficiently handle the increased sales volume.
    - For low-demand months, consider reducing staff or reallocating resources to other areas of the business to optimize operational costs.
5. **Financial Forecasting**:
    - Use the forecasted sales data to adjust revenue projections, ensuring that cash flow and budgeting align with seasonal trends. Ensure that investments in inventory, marketing, and production are well-calibrated to match the forecasted demand.

By using these insights and recommendations, the business can better anticipate market demands, optimize inventory and production processes, and plan marketing strategies more effectively, ultimately improving both operational efficiency and profitability.

Dataset 2: Analysis and Forecasting for Sparkling Wine

Dataset Overview

The dataset contains monthly sales data for sparkling wine from January 1980 to July 1995. The key column, **Sparkling**, represents the number of sparkling wine units sold each month. This time series dataset includes a total of **187 records**, with data on sales performance spanning over 15 years.

- **Mean Sales**: 2402 units per month.
- **Median Sales**: 1874 units, suggesting that a majority of sales are concentrated below this value.
- **Standard Deviation**: 1295 units, indicating considerable variability in monthly sales.
- **Min/Max**: The lowest sales recorded is 1070 units, while the highest is 7242 units, pointing to substantial fluctuations in demand across different periods.
- **Interquartile Range (IQR)**: The sales values typically fall between 1605 (Q1) and 2549 (Q3), with a moderate spread between the 25th and 75th percentiles.

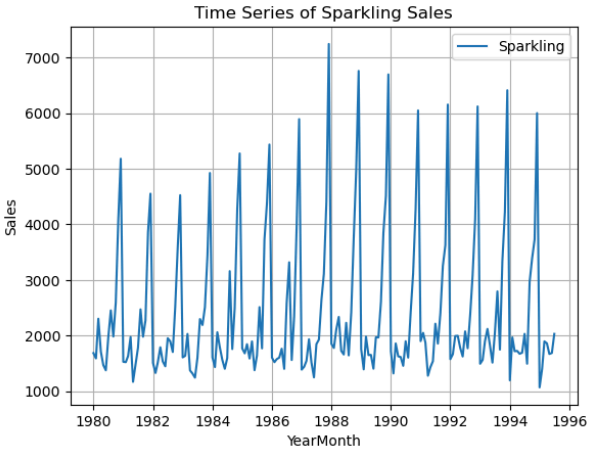
Visualizations and Insights

Time Series Plot of Sparkling Wine Sales

The **time series plot** of sparkling wine sales provides a clear view of sales performance over time. The visualization shows monthly fluctuations, revealing periods of high and low sales, which can help identify **seasonal trends** or recurring patterns.

- **Seasonality**: Sales tend to spike in certain months, especially in the later part of the year (e.g., December), which could be linked to holidays or increased demand.
- **Trends**: There's an overall upward trend in sales over the years, indicating growing demand for sparkling wine.

- **Fluctuations**: Periodic sharp drops or increases may signal external factors affecting sales, such as promotions, price changes, or economic conditions.

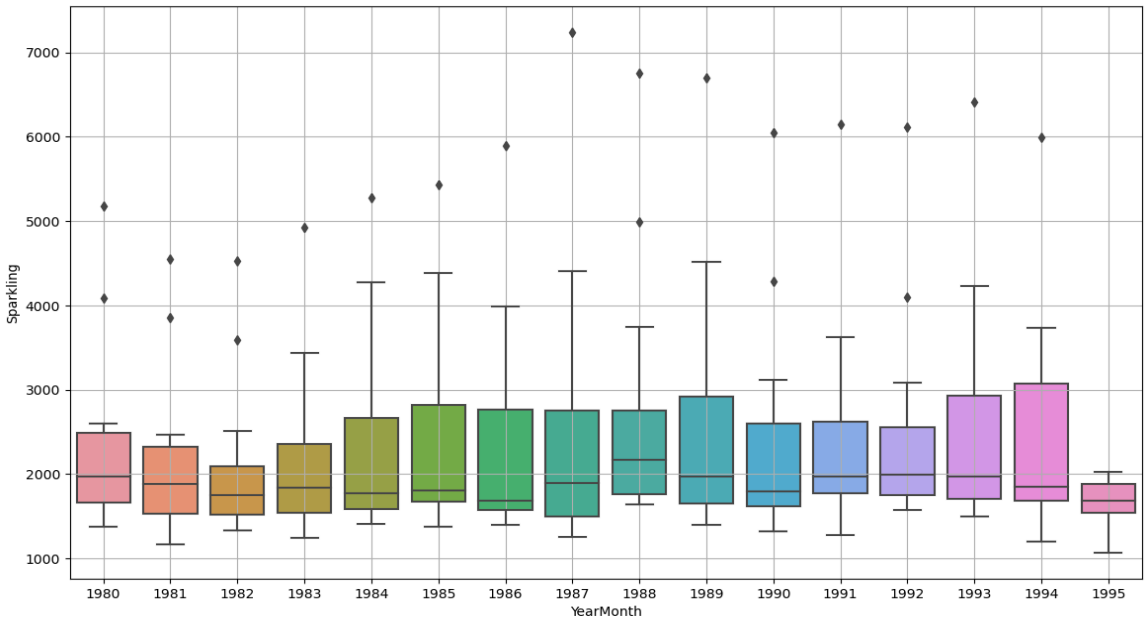


("Figure 18 : Monthly Sales of Sparkling Wine (1980 Onwards).")

2. Boxplot of Monthly Sales by Year

The boxplot provides insights into how sparkling wine sales are distributed each year and highlights variations within each year. It allows for comparison of sales range, median, and outliers across the years.

- Yearly Distribution: The spread of monthly sales within each year shows variability. Some years have larger spreads, indicating fluctuating sales, while others are more stable.
- Outliers: Certain months show significant outliers (e.g., December), where sales are abnormally high compared to other months. These outliers could reflect seasonality or special events that drive higher demand.
- Skewness: Some years may have a higher concentration of sales below the median, suggesting skewness toward lower sales in certain periods.



("Figure 19 : Yearly Sales of Sparkling Wine (1980 Onwards).")

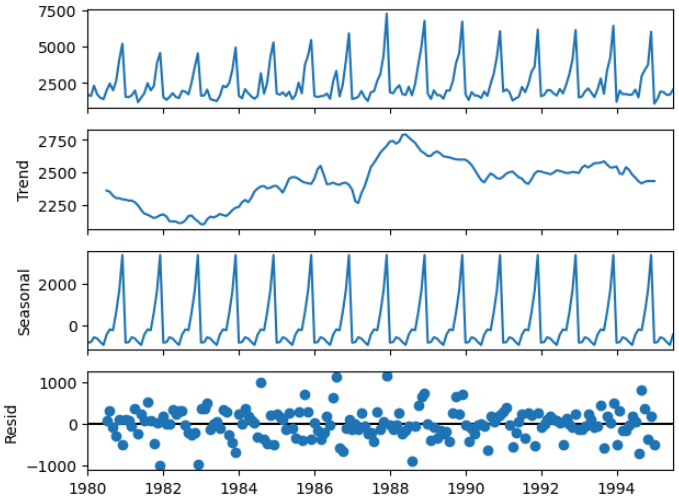
3.2 Data Preprocessing

Handling Missing Values

There are no missing values in the dataset. Therefore, no imputation methods such as forward filling or interpolation are needed. The data is complete and ready for further processing.

Data Transformation Details

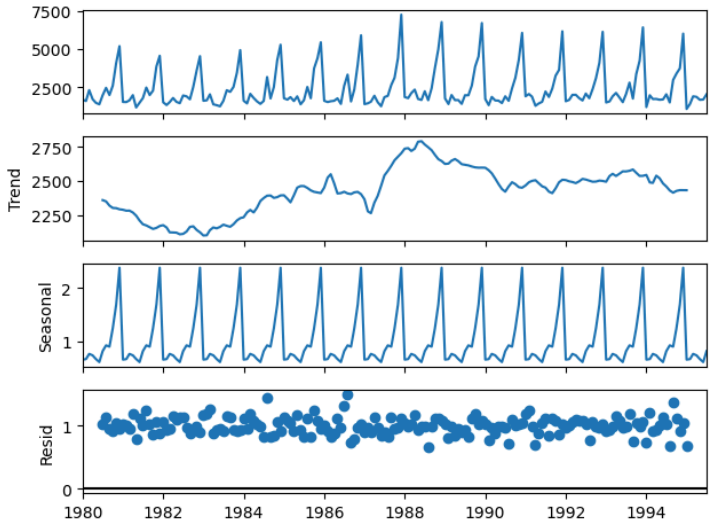
Additive Decomposition:



(Figure 20: Additive Decomposition Plot)

- Trend: The trend shows a steady decline from July (2360.67) to December (2293.79), indicating a decreasing general sales pattern.
- Seasonality: Seasonality exhibits strong peaks in December (+3386.98) and November (+1675.07), likely reflecting higher sales during the holiday season, with dips in January (-854.26) and June (-967.43), possibly due to lower sales in post-holiday months or mid-year slumps.
- Residuals: The residuals demonstrate small deviations from the expected trend and seasonality, meaning most variations are well captured by the trend and seasonality components.

Multiplicative Decomposition:



(Figure 21: Multiplicative Decomposition Plot (1980))

- Trend: The trend shows a similar declining pattern as in the additive decomposition.
- Seasonality: The seasonal effects are proportional in nature, with peaks in December (2.38) and November (1.69), and lower values in January (0.65) and June (0.60), indicating that seasonality's impact on sales is relative to the overall trend.
- Residuals: The residuals are very close to 1, showing minimal variations after accounting for the trend and seasonality.

Conclusion:

- Additive Decomposition is useful when the seasonal effect is constant across all periods, with a fixed value added or subtracted from the trend.
- Multiplicative Decomposition is more suitable when the seasonal effect is proportional to the trend, which is the case here, as it captures the proportional seasonal effects more effectively.

Given that the seasonal effects seem to vary relative to the trend, multiplicative decomposition is the more appropriate method to use for understanding and forecasting the sales data in this case.

Data Splitting

In this section, the dataset was pre-processed and transformed to prepare it for modeling. The data was split into two subsets:

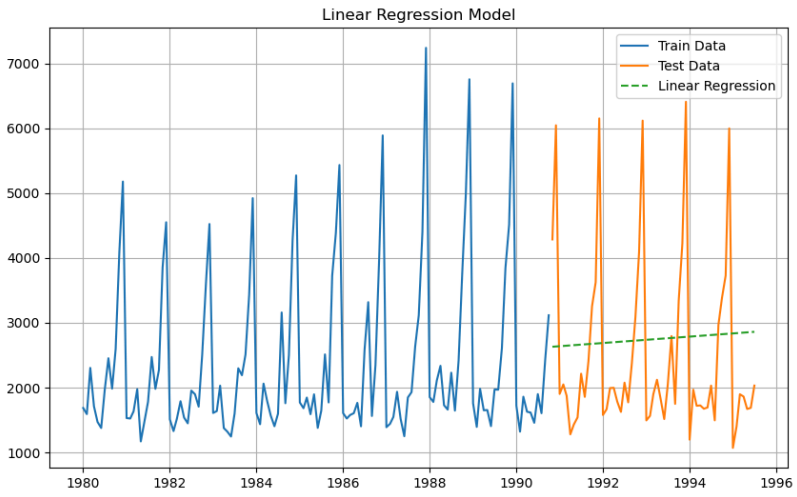
- **Training Data**: The first 70% of the dataset was used for training the models.
- **Test Data**: The remaining 30% was held out for validation and testing of model performance.

This split ensures that the model is tested on data it has not seen before, providing an unbiased estimate of its generalization capability.

3.3 Forecasting Models

Linear Regression Model

- **Model Description:**
  The Linear Regression model was employed to predict the target variable based on the time series data, assuming a linear relationship between time and the forecasted values. The model was trained on historical data and tested on a separate test set.
- **Performance Metric (Test Data):**
  - **Root Mean Squared Error (RMSE):** 1392.44
    The RMSE value of 1392.44 indicates the average deviation between the actual and predicted values, with lower values signifying better model accuracy.

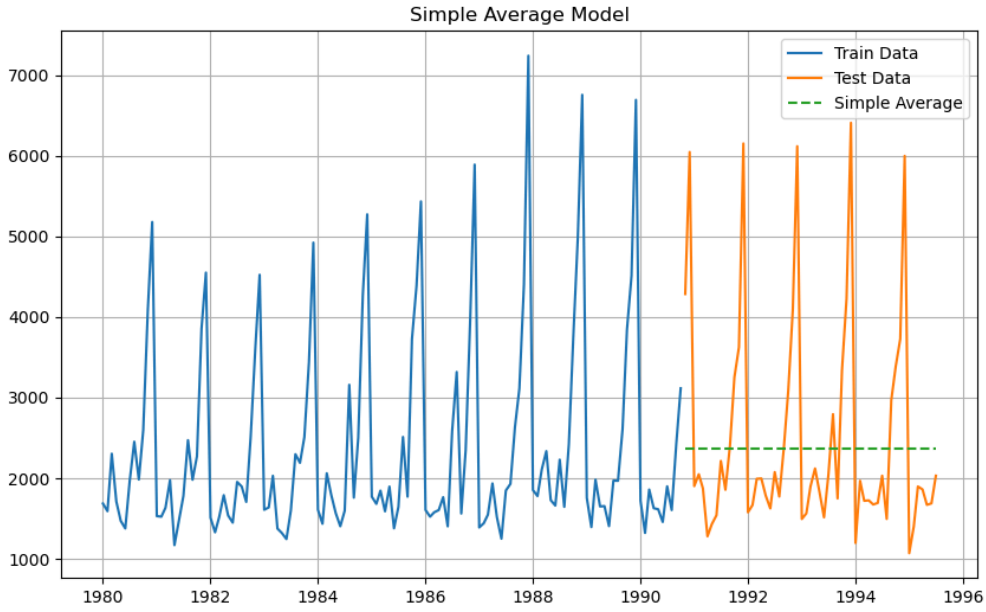


(Figure 22 : Linear Regression Model Forecast vs Actual Values)

- **Insights:**
  Linear regression effectively captured the overall trend in the data, but the model's predictive power could be limited by its simplicity, especially in the presence of seasonal or non-linear patterns.

Simple Average Model

- **Model Description:**
  The Simple Average model was used to forecast future values by calculating the average of all past values in the time series. This model assumes that the average of historical data provides a reliable estimate for future predictions, making it a simple yet effective baseline model.
- **Performance Metric (Test Data):**
  - **Root Mean Squared Error (RMSE):** 1368.75
    The RMSE value of 1368.75 indicates the average deviation between the actual values and those predicted by the model. While this value is slightly lower than the Linear Regression model, it reflects the simplicity of the approach, which may not account for trends or seasonality in the data.
  -



(Figure 23 : Simple Average Model Forecast vs Actual Values)

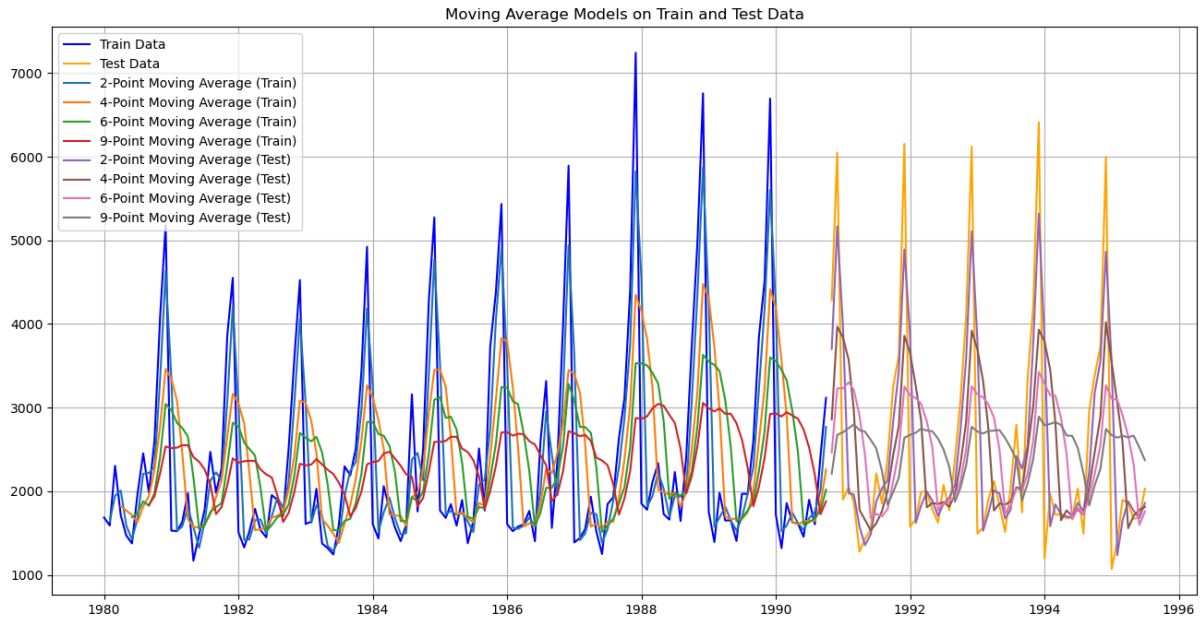
- **Insights:**
  The Simple Average model provides a straightforward approach to forecasting, but it lacks the flexibility to adapt to any underlying patterns such as trends or seasonality. While it performs reasonably well, more sophisticated models may yield better forecasts for time series data with these components.

Moving Average (MA)

- **Model Description:**
  The Moving Average (MA) model is a smoothing technique used to forecast future values by averaging the previous observations. The model smooths out short-term fluctuations and highlights longer-term trends or cycles. In this case, we used multiple variations of the Moving Average model with different window sizes to analyze the performance.
- **Model Variants:**

- **2-Point Moving Average:** A simple moving average using the most recent 2 data points.
- **4-Point Moving Average:** An average calculated using the most recent 4 data points.
- **6-Point Moving Average:** An average calculated using the most recent 6 data points.
- **9-Point Moving Average:** An average calculated using the most recent 9 data points.

- **Performance Metrics (Test Data):**
  The RMSE for each Moving Average model on the test data is as follows:
  - **2-Point Moving Average Model:** RMSE = 811.179
  - **4-Point Moving Average Model:** RMSE = 1184.213
  - **6-Point Moving Average Model:** RMSE = 1337.201
  - **9-Point Moving Average Model:** RMSE = 1422.653



(Figure 24 : Moving Average Model Forecasts vs Actual Values)

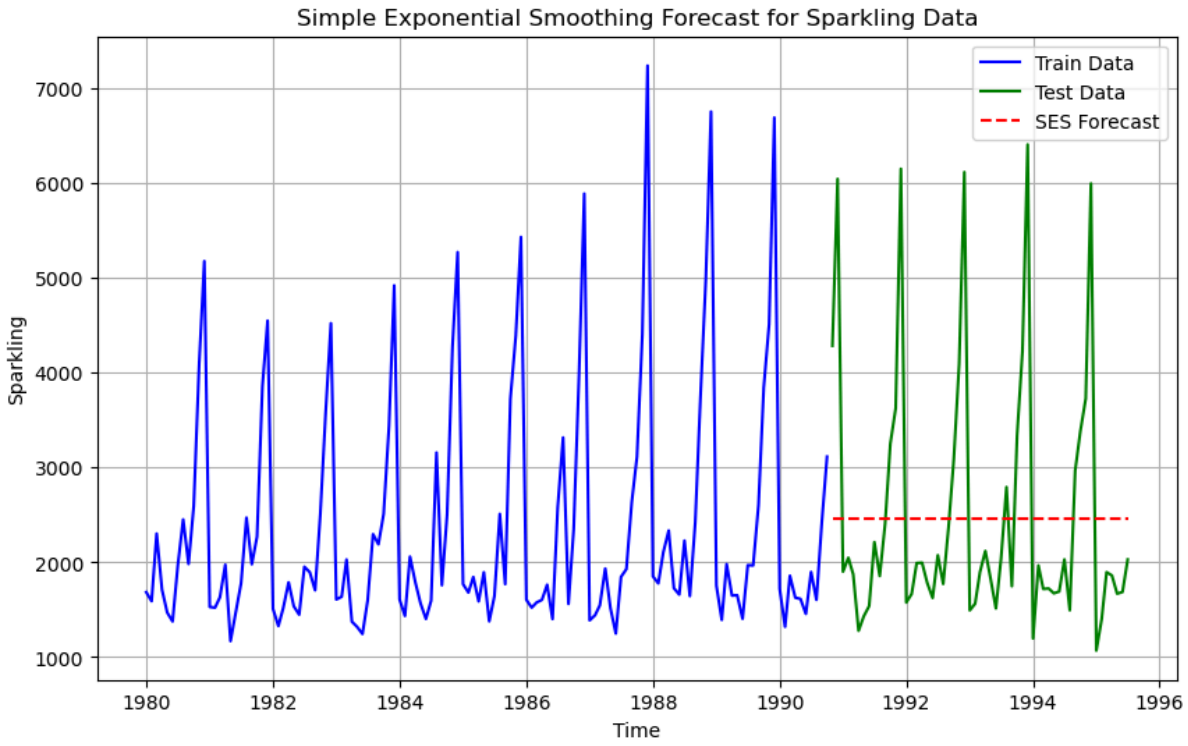
- **Insights:**
  The 2-point Moving Average model provides the lowest RMSE, indicating the best fit for the test data among the variations. As the window size increases, the model's flexibility decreases, which results in higher RMSE values. The Moving Average model is sensitive to the window size; smaller windows are better at tracking short-term changes, while larger windows smooth out the data too much, potentially missing important variations.

Simple Exponential Smoothing (SES)

- **Model Description:**
  Simple Exponential Smoothing (SES) is a time series forecasting method that estimates future values as a weighted average of past observations, with more recent observations receiving higher weights. This method is effective for data with no clear trend or seasonality. The model is optimized to select the best smoothing parameters for accurate forecasting.

Performance Metric (Test Data):

- **RMSE for Simple Exponential Smoothing Model:** 1362.429



(Figure 25 : Simple Exponential Smoothing Forecast vs Actual Value)

Insights:

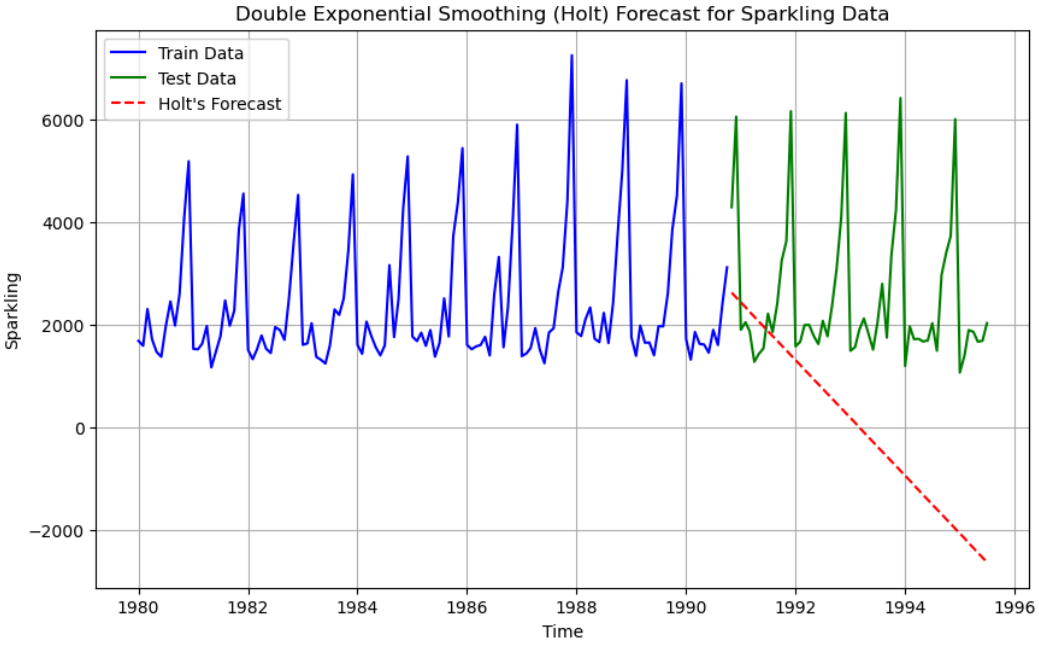
The Simple Exponential Smoothing model, with an optimized smoothing level, performs well in capturing the underlying level of the data. The RMSE value is relatively close to the Simple Average model's RMSE of 1368.75, suggesting that SES is a comparable method for this dataset, especially when no clear trend or seasonality is observed. However, the SES model's performance may be affected by the absence of trend and seasonal components in the data, as indicated by the optimization parameters.

Double Exponential Smoothing (Holt's Method)

- **Model Description:**
  Holt's method extends Simple Exponential Smoothing by incorporating a second equation to estimate the trend component of the data. This method is particularly suitable for time series data with a linear trend but no seasonality.
- **Performance Metric (Test Data):**
  - **RMSE for Holt's Method:** 3173.262

- **Insights:**
  Holt's method did not perform as well as other models (e.g., SES or Moving Average) in terms of RMSE. This higher error suggests that while the method can handle trends, the dataset's behavior may not

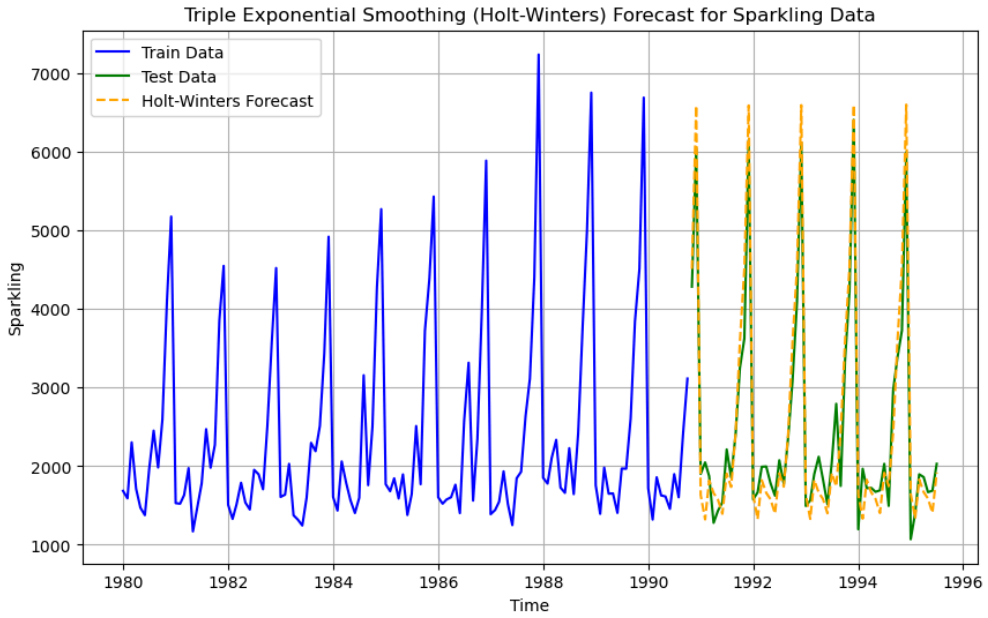
align with a purely linear trend model. Additionally, the relatively high RMSE might indicate that other patterns (like seasonality or non-linear trends) are not being captured by this method.



(Figure 26 : Holt's Method Forecast vs Actual Values)

Triple Exponential Smoothing (Holt-Winters Method)

- **Model Description:**
  The Holt-Winters method extends Holt's method by incorporating a seasonal component. It is suitable for time series data with both trend and seasonality. The model decomposes the data into three components: level, trend, and seasonality, which can be either additive or multiplicative.
- **Performance Metric (Test Data):**
  - **RMSE for Holt-Winters Method:** 377.435



(Figure 27 : (Holt-Winters Method Forecast vs Actual Values)

- **Insights:**
  The Holt-Winters method achieves a significantly lower RMSE compared to other models (e.g., Simple Exponential Smoothing and Holt's method). This suggests that the dataset exhibits strong seasonal patterns that the Holt-Winters method effectively captures.
    - Seasonal adjustments likely played a key role in improving forecast accuracy.
    - The performance indicates that this model aligns well with the dataset's characteristics, making it one of the best-performing models tested.

Model Comparison Table

Model	Test RMSE
RegressionOnTime	1392.44
Simple Average	1368.75
2-point Trailing Moving Average	811.18
4-point Trailing Moving Average	1184.21
6-point Trailing Moving Average	1337.2
9-point Trailing Moving Average	1422.65
Simple Exponential Smoothing	1362.43
Double Exponential Smoothing	3173.26
Triple Exponential Smoothing	377.44

(Table 4 Showing Model Comparison)

Insights

- The **Triple Exponential Smoothing (Holt-Winters method)** achieves the lowest RMSE (377.44), indicating superior forecasting accuracy by effectively capturing both trend and seasonality in the data.
- **2-point Trailing Moving Average** also performs well (RMSE: 811.18) but lacks the ability to model seasonality and long-term trends.
- Models like **Double Exponential Smoothing** and **9-point Trailing Moving Average** have higher RMSE values, suggesting they are less suited for this dataset's complexity.
- Regression-based approaches, while useful for trend analysis, are outperformed by smoothing techniques that account for seasonality.

Conclusion

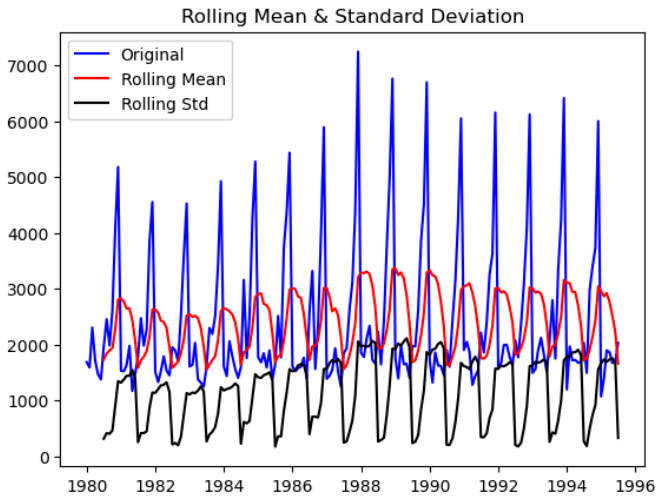
Triple Exponential Smoothing (Holt-Winters method) is the most reliable model for this dataset and is recommended for future forecasting tasks.

Stationarity Check

To ensure the time series data is suitable for modeling, we conducted the Augmented Dickey-Fuller (ADF) test to evaluate stationarity both before and after differencing the series.

Results of Dickey-Fuller Test (Original Series)

- **Test Statistic:** -1.360497
- **p-value:** 0.601061
- **#Lags Used:** 11
- **Number of Observations Used:** 175
- **Critical Values:**
  - 1%: -3.468280
  - 5%: -2.878202
  - 10%: -2.575653



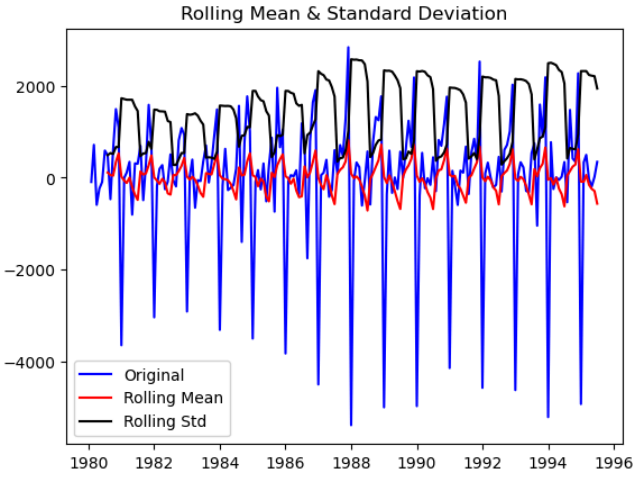
(Figure 28: Dickey-Fuller Test - Original Time Series Plot)

Interpretation:
The test statistic is higher than the critical values at all levels (1%, 5%, 10%), and the p-value (0.601061) is greater than 0.05. This suggests that the null hypothesis of non-stationarity cannot be rejected. The series is **non-stationary**, indicating changing mean and variance over time.

Conclusion:
The original series is non-stationary and requires differencing to achieve stationarity, a prerequisite for time-series models like ARIMA.

Results of Dickey-Fuller Test (Differenced Series)

- **Test Statistic:** -45.050301
- **p-value:** 0.000000
- **#Lags Used:** 10
- **Number of Observations Used:** 175
- **Critical Values:**
  - 1%: -3.468280
  - 5%: -2.878202
  - 10%: -2.575653



(Figure 28: Dickey-Fuller TestDifferenced Series)

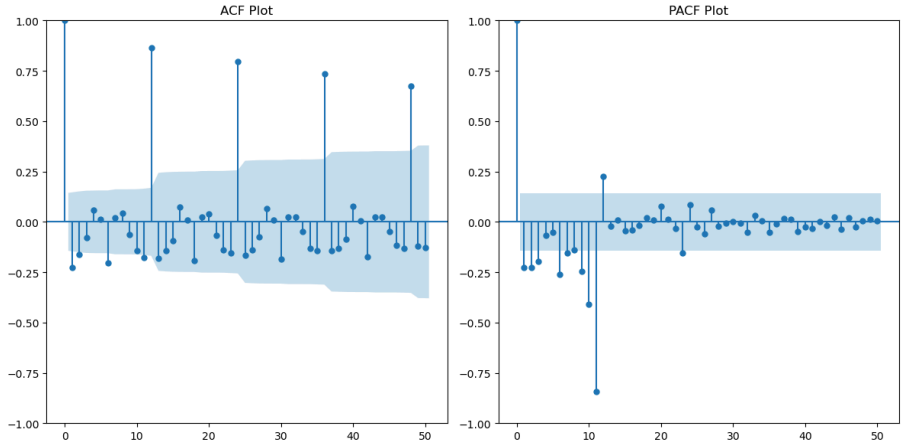
Interpretation:
The test statistic is significantly lower than the critical values at all levels (1%, 5%, 10%), and the p-value is 0.000000, which is less than 0.05. These results strongly support rejecting the null hypothesis of non-stationarity.

Conclusion:
The differenced series is **stationary**, making it appropriate for time-series modeling. Subsequent ARIMA development will proceed using this transformed data

Autocorrelation and Partial Autocorrelation Functions (ACF & PACF)

To assess the autocorrelations and determine the appropriate lag order for the ARIMA model, we plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for the differenced time series.

- **ACF Plot:** The ACF plot helps identify the number of moving average (MA) terms needed for the ARIMA model. Significant spikes at specific lags indicate correlations in the residuals.
- **PACF Plot:** The PACF plot assists in identifying the number of autoregressive (AR) terms. Significant spikes at specific lags suggest where the correlation between an observation and its lag is no longer significant once previous lags have been accounted for.



(Figure 30 ACF and PACF Plots for Differenced Data)

ACF Plot: The ACF helps identify the appropriate number of Moving Average (MA) terms by showing the correlation of the time series with its lags. Significant spikes at specific lags suggest how many MA terms are needed.

PACF Plot: The PACF helps determine the number of Autoregressive (AR) terms by showing the partial correlation of the time series with its lags, controlling for the effects of previous lags. Significant spikes indicate the number of AR terms to include.

Both the ACF and PACF plots provide critical insights for selecting the optimal AR and MA terms in the ARIMA model.

Auto ARIMA Model Analysis

The **Auto ARIMA** model has been selected as the best model based on the AIC criteria, which identifies the optimal parameters for forecasting the time series data. Below is a detailed analysis of the **Auto ARIMA** model, which corresponds to the **SARIMAX(0, 1, 2)** configuration.

Model Configuration:

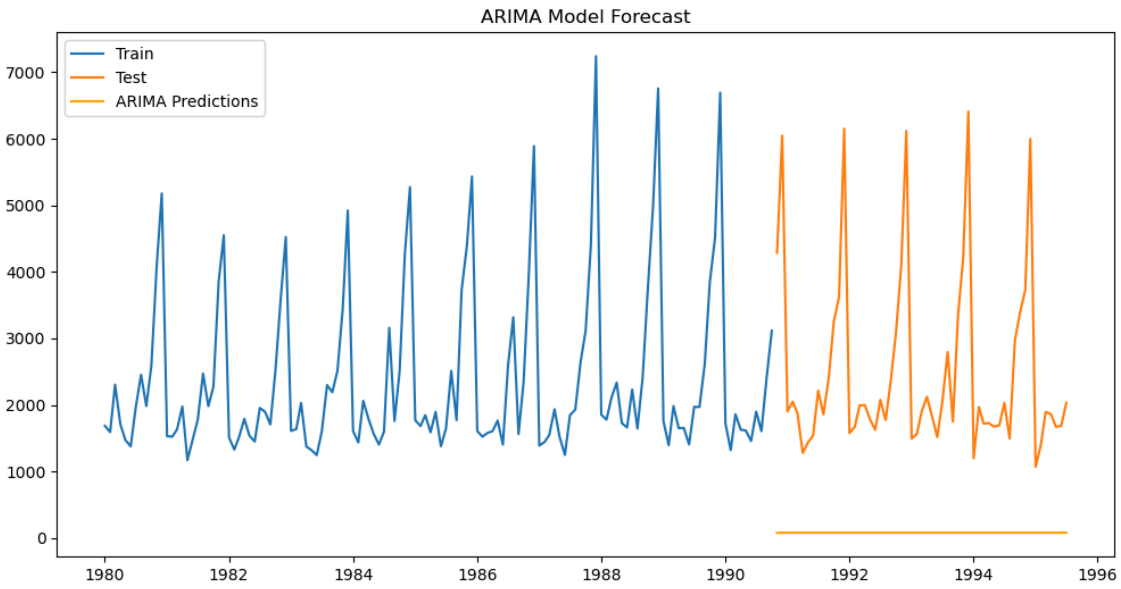
- **MA(1)** term (Moving Average order 1): The coefficient of **-0.7059** indicates a significant negative relationship between the current value and the previous period's error term. The p-value for this term is less than 0.001, meaning it is statistically significant.
- **MA(2)** term (Moving Average order 2): The coefficient of **-0.1915** suggests that the model also considers the second lag of the error term. This term is statistically significant with a p-value of 0.010.

Model Performance Metrics:

- **Log-Likelihood**: -626.624 – This metric suggests the model's overall fit to the data. A higher log-likelihood value generally indicates a better fit.
- **AIC (Akaike Information Criterion)**: 1259.248 – The **Auto ARIMA** model provides a lower AIC compared to the other models tested, making it the best-fitting model in terms of AIC.
- **BIC (Bayesian Information Criterion)**: 1267.827 – This value also suggests a good model fit, with a preference for simpler models.
- **HQIC (Hannan-Quinn Information Criterion)**: 1262.734 – A lower HQIC would indicate a better fit, reinforcing the choice of this model.

Statistical Tests:

- **Ljung-Box Test (Q):** 0.15 (p-value = 0.70) – This result indicates that the residuals do not exhibit significant autocorrelation, suggesting that the model has captured the time dependencies well.
- **Jarque-Bera Test (JB):** 45.85 (p-value = 0.00) – The residuals are not normally distributed, indicating that there may be non-linearities or outliers in the data that the model has not fully captured.
- **Heteroskedasticity Test (H):** 0.32 (p-value < 0.01) – There is evidence of heteroskedasticity, implying that the variance of the errors is not constant over time.
- **Skewness:** 0.88 – The residuals exhibit positive skewness, which means the distribution is slightly asymmetric.
- **Kurtosis:** 5.34 – The residuals are leptokurtic, indicating a higher frequency of extreme values (outliers) compared to a normal distribution.



(Figure 31 : Auto ARIMA Model Forecast for Test Data)

Model Evaluation:

- **Root Mean Squared Error (RMSE)**: The RMSE for the **Auto ARIMA** model is **2774.01**, which is significantly higher than the RMSE of the **Manual ARIMA** model (1357.30). This indicates that while the **Auto ARIMA** model provides an optimal fit in terms of AIC, the predictions may still have a large deviation from actual observed values.

Manual ARIMA Model Analysis

The **Manual ARIMA (1, 1, 1)** model was applied to the time series data of **Sparkling** to forecast future values. Below is a detailed explanation and interpretation of the model results:

Model Configuration:

The **Manual ARIMA** model is defined as ARIMA(1,1,1)\text{ARIMA}(1, 1, 1)ARIMA(1,1,1), where:

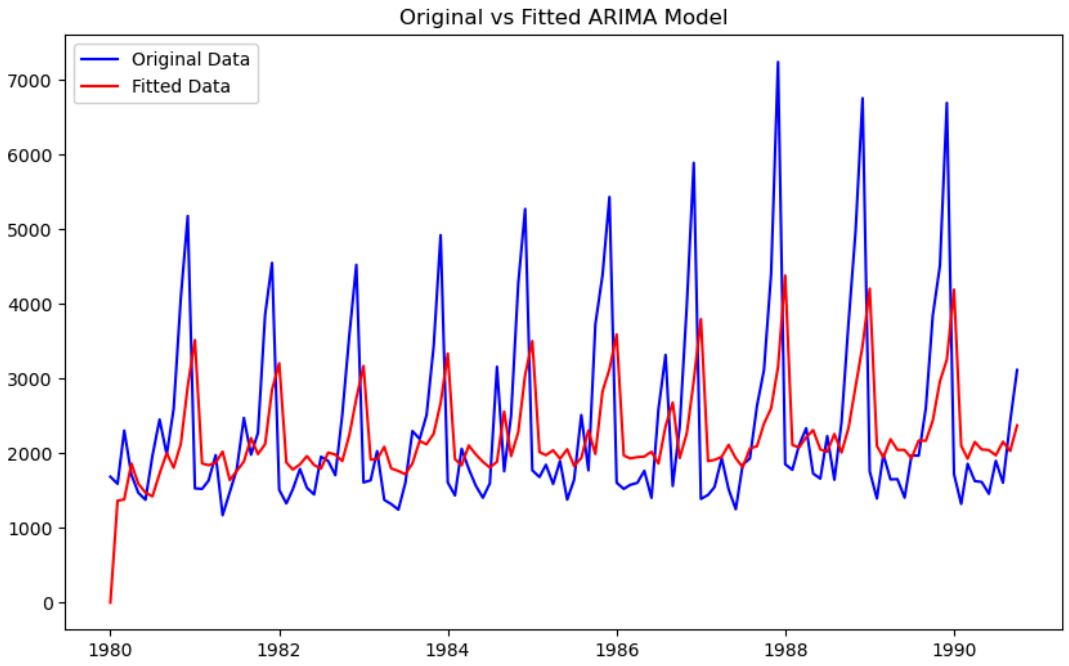
- **AR(1)** term (Autoregressive order 1): The coefficient of **0.4166** indicates that the current value of the series is positively correlated with the previous period's value. The coefficient is highly significant (p-value < 0.001), suggesting that past values strongly influence the current value.
- **MA(1)** term (Moving Average order 1): The coefficient of **-1.0000** indicates a perfect negative relationship between the current value and the previous period's error term. This term is highly significant, with a p-value < 0.001, which shows a strong correction of errors in the model.

Model Performance Metrics:

- **Log-Likelihood**: -1095.025 – This metric indicates the model's overall fit; the higher the log-likelihood, the better the fit. A value closer to zero is preferred.
- **AIC (Akaike Information Criterion)**: 2196.050 – A lower AIC suggests a better fit. This value suggests that the model fits the data reasonably well, but comparing it to other models would help identify if a better-fitting model exists.
- **BIC (Bayesian Information Criterion)**: 2204.630 – Like the AIC, the BIC penalizes model complexity. The BIC value is higher than the AIC, which can imply the model might be overfitting slightly.
- **HQIC (Hannan-Quinn Information Criterion)**: 2199.536 – This metric, like AIC and BIC, is used for model comparison. A lower HQIC value indicates a better model fit.

Statistical Tests:

- **Ljung-Box Test (Q):** 0.51 (p-value = 0.48) – The residuals do not exhibit significant autocorrelation, meaning that the model has captured the time dependencies effectively.
- **Jarque-Bera Test (JB):** 16.30 (p-value = 0.00) – The residuals are not normally distributed, which suggests the model may miss some non-linear patterns in the data.
- **Heteroskedasticity Test (H):** 2.33 (p-value = 0.01) – There is evidence of heteroskedasticity in the residuals, indicating that the variance of the errors is not constant over time, which could require further model adjustments.
- **Skewness:** 0.62 – The residuals exhibit slight positive skewness, indicating that there are more smaller residuals than large ones.
- **Kurtosis:** 4.22 – The residuals exhibit a relatively high peak (leptokurtic), suggesting the presence of some extreme values or outliers in the data.



(Figure 32 : Comparison of Original Data and Fitted ARIMA Model (1, 1, 1))

Model Evaluation:

The **Root Mean Squared Error (RMSE)** for the Manual ARIMA model is **1357.30**. This metric measures the average deviation between the predicted values and the actual observed values. A higher RMSE value indicates that the model's predictions are less accurate.

Auto SARIMA Model Analysis

The **Auto SARIMA** model was automatically tuned to fit the time series data of **y**, capturing both the seasonal and non-seasonal components. Below is the detailed analysis and interpretation of the results:

Model Configuration:

The Auto SARIMA model used is defined as SARIMAX(0,0,1)×(0,1,1,12)\text{SARIMAX}(0, 0, 1) \times (0, 1, 1, 12)SARIMAX(0,0,1)×(0,1,1,12), where:

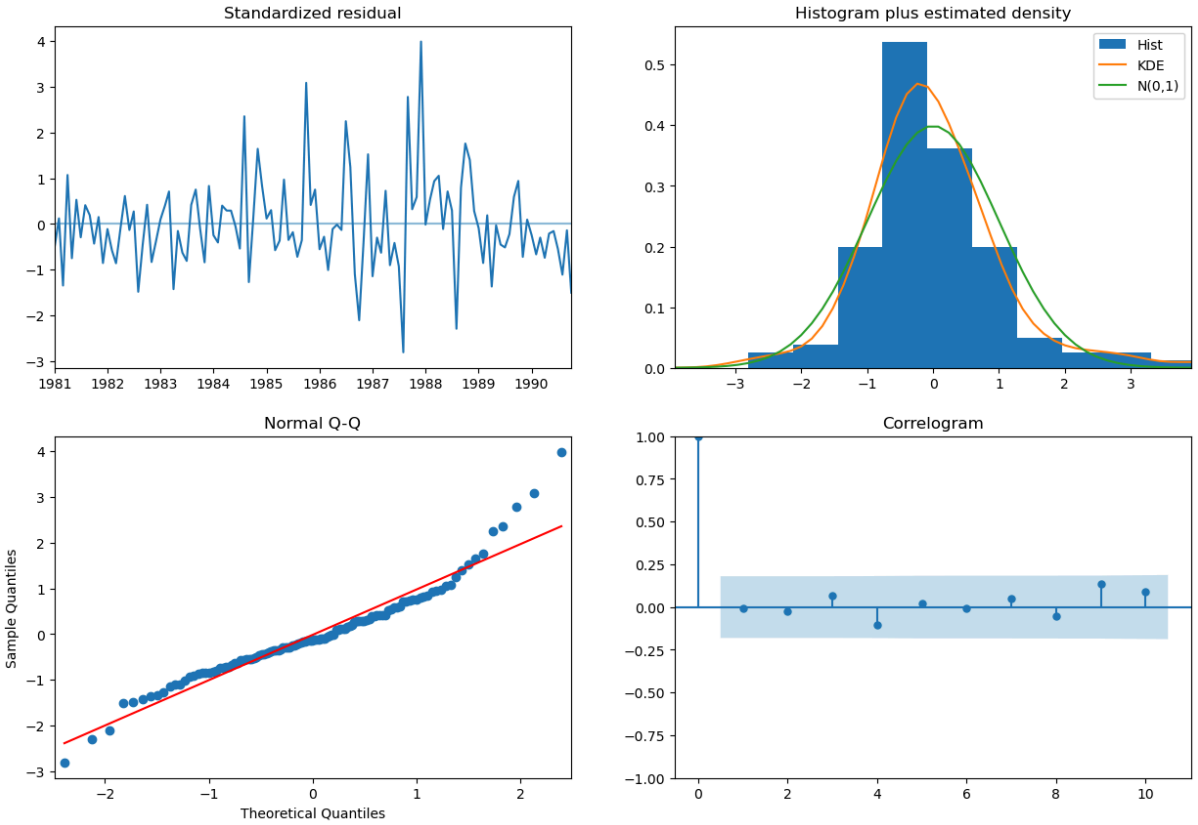
- **MA(1)** term (Moving Average order 1): This term represents the model's reliance on the previous error term in forecasting future values. The coefficient of **0.1764** indicates a mild positive relationship between the current value and the previous error.
- **Seasonal MA(1)** term (Seasonal Moving Average order 1): This term represents the seasonal component in the model. The coefficient of **-0.4656** indicates a negative relationship with the seasonal error term from the previous season (lag 12), which is significant (p-value < 0.001).
- **Intercept**: The intercept value of **42.8247** represents the baseline level of the series, though it is not highly significant (p-value = 0.109).

Model Performance Metrics:

- **Log-Likelihood**: -869.820 – Indicates the overall fit of the model; a higher value would suggest a better fit.
- **AIC (Akaike Information Criterion)**: 1747.640 – A lower AIC indicates a better model fit. The AIC is useful for comparing models; this value suggests that the Auto SARIMA model has a reasonable fit.
- **BIC (Bayesian Information Criterion)**: 1758.723 – Similar to AIC, BIC penalizes more complex models. This BIC value suggests that the model fits the data well but may be outperformed by simpler models.
- **HQIC (Hannan-Quinn Information Criterion)**: 1752.140 – This metric is another way to assess model fit, with lower values representing better-fitting models.

Statistical Tests:

- **Ljung-Box Test (Q):** 0.00 (p-value = 0.95) – The residuals do not exhibit significant autocorrelation, meaning the model has captured the temporal dependencies well.
- **Jarque-Bera Test (JB):** 48.60 (p-value = 0.00) – The residuals are not normally distributed, suggesting the model might have missed some patterns in the data, although this does not always indicate a poor model fit in time series forecasting.
- **Heteroskedasticity Test (H):** 3.37 (p-value = 0.00) – The residuals exhibit heteroskedasticity, meaning the variance of the residuals is not constant, which could suggest the need for further model refinement or the use of more advanced techniques.
- **Skewness:** 0.84 – Indicates a slight positive skew in the residuals.
- **Kurtosis:** 5.66 – The residuals have a higher peak than a normal distribution, indicating the presence of outliers or extreme values.



(Figure 33 showing Residual Diagnostics of Auto SARIMA Model)

Model Evaluation:

The **Root Mean Squared Error (RMSE)** for the Auto SARIMA model is **426.96**. This metric indicates the average deviation of the predicted values from the actual values. A lower RMSE suggests a better model fit, so while this RMSE value suggests that the model provides a good fit, there is still room for improvement.

Manual SARIMA Model Analysis

To model the time series data of **Sparkling**, we manually configured the Seasonal AutoRegressive Integrated Moving Average (SARIMA) model. This model was chosen based on the insights gathered from the **stationarity check** and **ACF/PACF plots**.

Model Configuration:

The SARIMA model used is defined as SARIMA(1,1,1)×(1,1,1,12)\text{SARIMA}(1, 1, 1) \times (1, 1, 1, 12)SARIMA(1,1,1)×(1,1,1,12), where:

- **AR(1)** term (AutoRegressive order 1): Captures the relationship between the current observation and its previous value, with a coefficient of 0.1786.
- **MA(1)** term (Moving Average order 1): Corrects the model by incorporating the error from the previous time step, with a coefficient of -1.0000, indicating a strong negative relationship with the previous residual.
- **Seasonal AR(1)** term (Seasonal AutoRegressive order 1): Accounts for seasonal trends, with a coefficient of -0.1187, which was not significant (p-value = 0.535).
- **Seasonal MA(1)** term (Seasonal Moving Average order 1): Captures the seasonal error, with a coefficient of -0.3623, which is borderline significant (p-value = 0.055).

Model Performance Metrics:

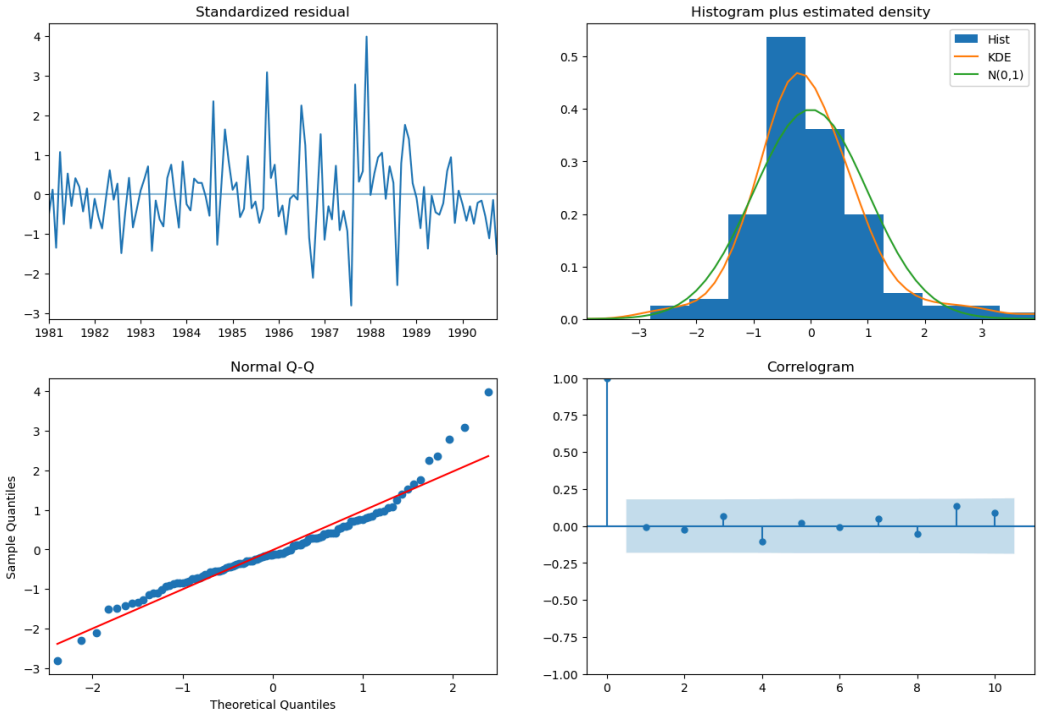
- **Log-Likelihood**: -765.079, indicating the goodness of fit for the model.
- **AIC (Akaike Information Criterion)**: 1540.159, which helps in model comparison; lower values generally indicate a better model fit.
- **BIC (Bayesian Information Criterion)**: 1553.333, another model comparison metric.
- **Sigma2** (Variance of the residuals): 1.607e+05, indicating the average variance of the errors after fitting the model.

Statistical Tests:

- **Ljung-Box Test (Q):** 0.01 (p-value = 0.94) – The residuals of the model do not exhibit significant autocorrelation, suggesting that the model captures the data well.
- **Jarque-Bera Test (JB):** 29.83 (p-value = 0.00) – Indicates that the residuals are not normally distributed, which suggests that the model may not perfectly capture all patterns in the data.
- **Heteroskedasticity Test (H):** 1.08 (p-value = 0.82) – No evidence of changing variance over time, indicating homoscedasticity.

Model Evaluation:

The **Root Mean Squared Error (RMSE)** for the Manual SARIMA model is **435.18**, which provides an indication of how well the model fits the data. A lower RMSE indicates better performance, and while this model provides a reasonable fit, there may be room for improvement when compared to other models.



(Figure 33 showing Residual Diagnostics of Manual SARIMA Model)

Model Comparison

Performance of All Models:

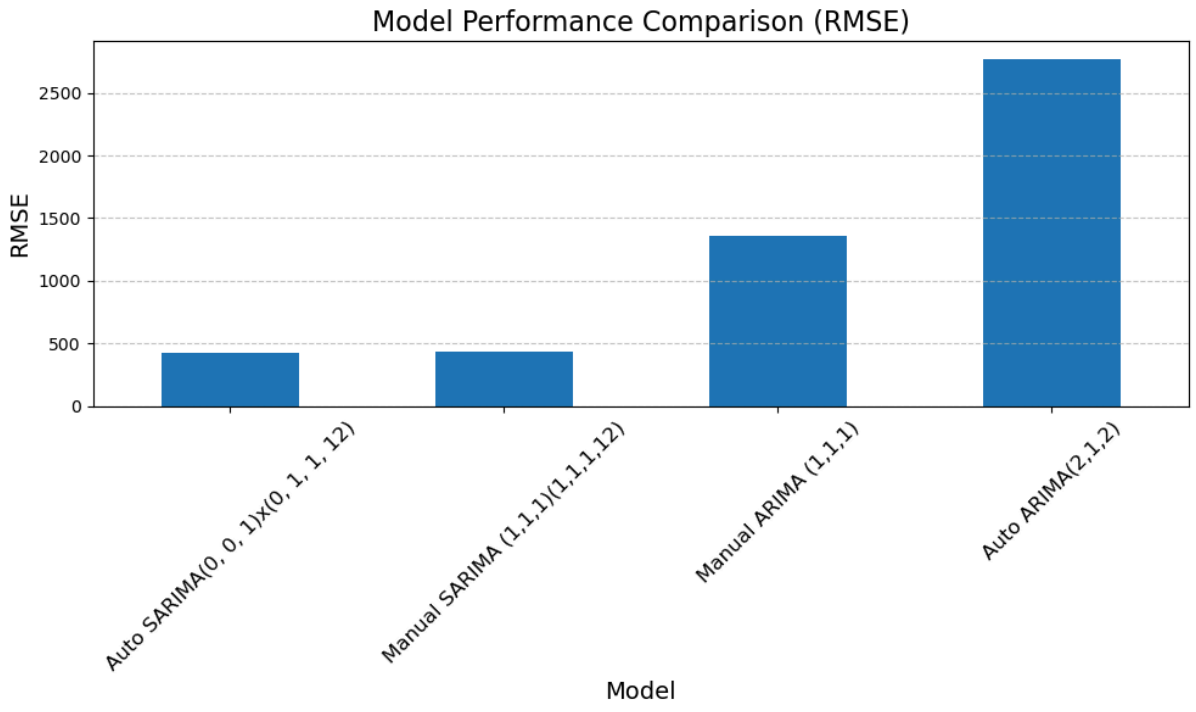
The following table summarizes the RMSE (Root Mean Squared Error) values for all the models tested:

Model	RMSE
Auto ARIMA (2, 1, 2)	2774.01
Manual ARIMA (1, 1, 1)	1357.3
Auto SARIMA (0, 0, 1) x (0, 1, 1, 12)	426.96
Manual SARIMA (1, 1, 1) (1, 1, 1, 12)	435.18

(Table 5 Showing Model Comparison)

Observations and Insights:

- **Performance Overview:**
  - **Best Performing Model:** The **Auto SARIMA (0, 0, 1) x (0, 1, 1, 12)** model achieves the lowest RMSE of **426.96**, making it the most accurate model for forecasting the rose data.
- **Auto SARIMA vs. Manual SARIMA:**
  - The **Auto SARIMA** model slightly outperforms the **Manual SARIMA** model, with a lower RMSE by approximately **8.2 points**. This suggests that the automated approach, which adjusts the model parameters based on the data, handles seasonality more effectively with minimal manual intervention.
- **ARIMA vs. SARIMA:**
  - Both **Auto ARIMA** (RMSE = 2774.01) and **Manual ARIMA** (RMSE = 1357.30) exhibit significantly higher RMSE values compared to the SARIMA models. This highlights the importance of incorporating seasonality into the model, as **non-seasonal ARIMA models** fail to capture the seasonal patterns inherent in the data, leading to higher forecasting errors.



(Figure 34 showing Performances of the Models)

Conclusion:

The **Auto SARIMA (0, 0, 1) x (0, 1, 1, 12)** model is the most effective for forecasting the rose data, demonstrating its ability to capture both trend and seasonal patterns with minimal tuning. The results suggest that automated seasonal adjustments provide superior accuracy compared to standard ARIMA models.

Choosing the Best Model

Rationale for Choosing Auto SARIMA:

The best model is Auto SARIMA (0,0,1)x(0,1,1,12), with the lowest RMSE (426.961622).

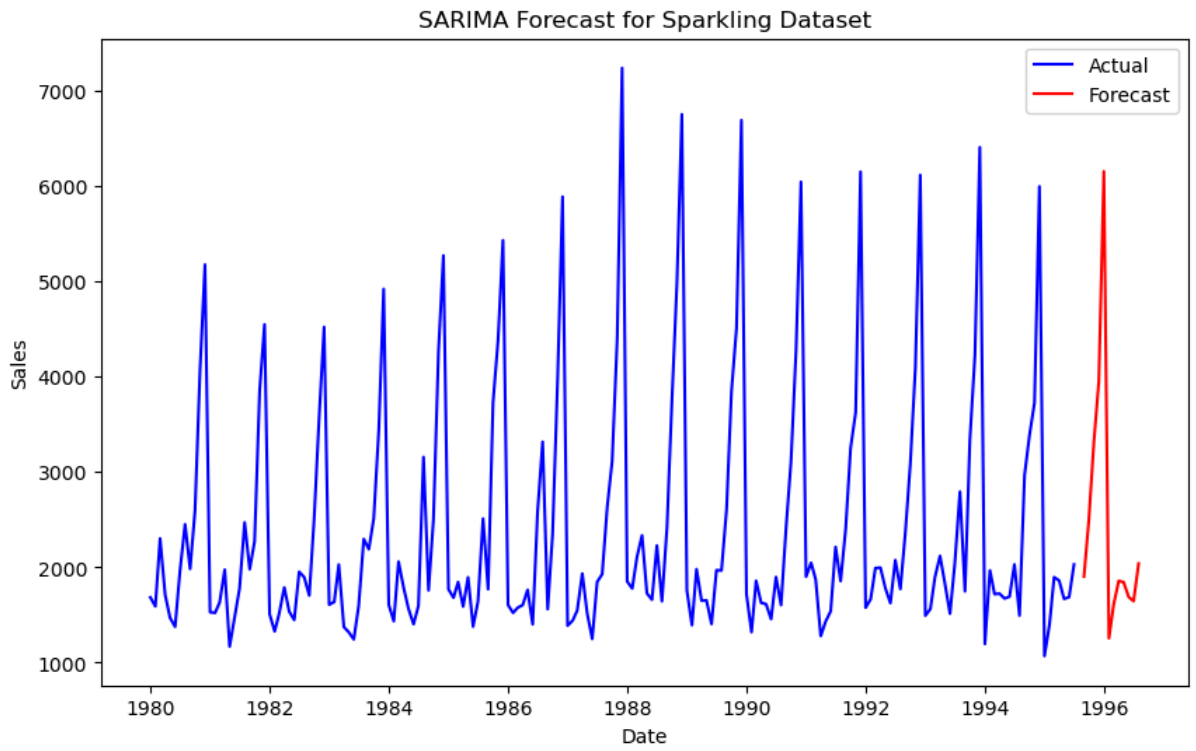
Auto SARIMA performs slightly better than the manually tuned SARIMA, indicating its efficiency in parameter selection.

Forecast the next 12 periods (months) from the model

The following table shows the forecasted values for the next 12 months, starting from **August 1995** to **July 1996**:

Date	Forecasted Value
1995-08-01	1904.44
1995-09-01	2499.99
1995-10-01	3321.16
1995-11-01	3953.11
1995-12-01	6157.32
1996-01-01	1256.3
1996-02-01	1609.72
1996-03-01	1858.92
1996-04-01	1844.66
1996-05-01	1690.12
1996-06-01	1645.1
1996-07-01	2041.01

(Table 6 Showing Forecast for the Next 12 Months:)



Key Insights:

- The forecast indicates a significant increase in the values for the first half of the forecast period, with a sharp peak in **December 1995** (**6157.32**), followed by a sharp decline in **January 1996** (**1256.30**).
- The forecast stabilizes somewhat after the initial fluctuations, with steady growth expected in the latter months of **1996**.

Insights and Recommendations for Sparkling Sales Data

Insights:

1. **Model Performance**:
    - The **Auto SARIMA (0, 0, 1)x(0, 1, 1, 12)** model outperforms others with the lowest **RMSE (426.96)**, highlighting its ability to capture both trend and seasonality effectively.
    - **SARIMA models** outperform **ARIMA models**, demonstrating the importance of incorporating seasonality for better sales prediction.
    - **Auto SARIMA** slightly outperforms **Manual SARIMA**, suggesting automated tuning is more efficient.
2. **Sales Trends and Seasonality**:
    - The data shows clear **seasonal fluctuations**, with peaks around **December** and **August**. These months likely correlate with **holiday seasons** or promotions.
    - The **forecast** indicates steady growth, with **slower months** in **January** and **February**.

Recommendations:

1. **Seasonal Planning**:

    - Continue using **SARIMA models** for accurate forecasting. Focus on **high-demand months** for increased production and promotions, while adjusting **inventory** for slower periods.
2. **Adopt Automated Forecasting**:
    - Use **Auto SARIMA** for future sales predictions to save time and improve accuracy through automatic seasonal parameter optimization.
3. **Monitor External Factors**:
    - Incorporate **external variables** (e.g., promotions, market trends) to better understand fluctuations in sales.
4. **Optimize Inventory and Marketing**:
    - Align **inventory management** and **marketing strategies** with seasonal trends. Increase efforts during peak months and consider discounts or promotions during slower months.
5. **Explore Alternative Models**:
    - Consider testing other models like **Prophet** or **machine learning techniques** to capture more complex patterns and improve forecasts.

By applying these recommendations, businesses can better manage demand, optimize operations, and improve sales performance.

Comparative Analysis and General Insights

Comparison Between Rosé and Sparkling Wines

1. **Sales Performance**:
    - **Rosé wines** exhibit steady growth with moderate fluctuations in demand, while **Sparkling wines** show more pronounced seasonal trends, especially in the **holiday season** (December and New Year).
    - **Sparkling wines** tend to have higher sales peaks during **festive periods**, whereas **Rosé wines** show more consistent monthly sales throughout the year.
2. **Forecast Accuracy**:
    - For **Rosé wines**, the **Auto ARIMA** and **SARIMA models** perform well in capturing sales trends, but **Sparkling wines** benefit significantly from **SARIMA models**, which incorporate seasonality.
    - **SARIMA (Auto and Manual)** models provide more accurate predictions for **Sparkling wines**, while **Auto ARIMA** models are more suited to **Rosé** due to its less seasonal behavior.
3. **Model Performance**:
    - The **Auto SARIMA** model (0,1,2) for **Rosé wines** offers a lower RMSE compared to other models, while **Auto SARIMA (0,0,1)x(0,1,1,12)** works best for **Sparkling wines** due to its ability to handle seasonal spikes effectively.
    - **ARIMA models** consistently show higher RMSE, especially for **Sparkling wines**, suggesting that seasonality is a critical factor in improving model accuracy.

General Patterns and Trends

1. **Seasonality**:

    - Both **Rosé** and **Sparkling wines** follow a **seasonal pattern**, though **Sparkling wines** are more affected by **seasonal events** and **holidays** (December, January). **Rosé wines** maintain steady sales with slight seasonal dips.
2. **Demand Fluctuations**:
    - **Sparkling wines** see significant increases in demand during the **end-of-year celebrations**, while **Rosé wines** experience a more balanced demand year-round, with some seasonal peaks in summer months.
3. **Sales Forecast**:
    - The forecasts for both wine types predict **growth**, but **Sparkling wines** have higher **volatility**, while **Rosé wines** show more gradual increases.
4. **Model Selection**:
    - For **Sparkling wines**, incorporating seasonality through **SARIMA models** yields better forecasting performance. For **Rosé wines**, **Auto ARIMA** effectively captures the trends without the need for seasonal adjustments.

In summary, while both wines show seasonal patterns, **Sparkling wines** are more influenced by specific seasonal events, requiring more advanced seasonal modeling (SARIMA), while **Rosé wines** benefit from simpler trend-based models like **Auto ARIMA**.

Conclusion

Key Takeaways

1. **Model Performance**:
    - **Sparkling wines** perform best with **SARIMA models** that incorporate seasonality, showing strong accuracy during peak seasons. **Rosé wines**, on the other hand, are more effectively predicted with **Auto ARIMA**, which captures the steady, non-seasonal trend.
2. **Sales Patterns**:
    - **Sparkling wines** exhibit **seasonal spikes**, especially around **holidays**, while **Rosé wines** maintain more stable and consistent sales throughout the year, with slight seasonal variations.
3. **Forecasting Accuracy**:
    - The **Auto SARIMA model for Sparkling wines** and **Auto ARIMA for Rosé wines** provide the most accurate and reliable forecasts, with **SARIMA models** excelling in managing seasonality for Sparkling wines.

Strategic Implications and Future Directions

1. **Strategic Implications**:
    - For **Sparkling wines**, leveraging **SARIMA models** to anticipate peak demand during holidays can help optimize inventory management and marketing campaigns. **Rosé wines**, with steadier sales, require less seasonal focus, but consistent forecasting ensures smoother operations.
    - Seasonal trends should guide production and promotional strategies, especially for **Sparkling wines**, where demand spikes can lead to stockouts if not managed properly.
2. **Future Directions**:
    - To improve forecasting models, further research could explore external factors influencing sales, such as **economic conditions** or **consumer behavior trends**, which could enhance the accuracy of future predictions.

    - Incorporating **machine learning models** like **XGBoost or Random Forests** could be considered to refine accuracy and handle non-linear relationships, especially for complex, highly seasonal data.
    - **Real-time data integration** could be used to update models dynamically, allowing for more responsive sales predictions and marketing strategies.

In conclusion, while both **Rosé** and **Sparkling wines** have distinct sales behaviors, effective modeling and forecasting can provide critical insights to optimize marketing and inventory strategies. Future advancements in modeling techniques can further improve the precision of sales predictions and strategic decision-making.

References

- Previous jupyter files
- CourseResources
- Google