Time Series Forecasting Project

Coded

Uncorking Insights: Time Series Analysis and Forecasting of 20th Century Wine Sales

Table Of Contents:

List Of Figures:

1. Introduction

1.1 Project Overview

The primary goal of this project is to develop and evaluate predictive models for forecasting the demand for Sparkling and Rose wines. By utilizing advanced time-series modelling techniques like ARIMA and SARIMA, the project aims to create reliable forecasts for both categories of wine. This enables stakeholders to optimize production, inventory, and marketing strategies to meet future market demands effectively.

1.2 Problem Statement

The demand for Sparkling and Rose wines has been inconsistent, leading to inefficiencies in production, inventory management, and distribution. This project addresses the need to accurately predict future demand for both categories to:

1. Minimize overstocking and understocking issues.
2. Optimize production planning and supply chain operations.
3. Ensure consistent product availability in the market.

The challenge is to determine the best-performing forecasting model for both datasets by evaluating multiple approaches (Auto ARIMA, Manual ARIMA, Auto SARIMA, and Manual SARIMA) and generate accurate 12-month forecasts.

1.3 Affected Stakeholders

- **Winemakers and Production Teams:** Need to plan production cycles efficiently to match demand for Sparkling and Rose wines while maintaining quality standards.
- **Supply Chain and Logistics Teams:** Optimize inventory and distribution for both product lines to ensure timely delivery without incurring excess costs.
- **Retailers and Distributors:** Ensure adequate stock of Sparkling and Rose wines to meet consumer demand without shortages or wastage.
- **Marketing and Sales Teams:** Design targeted campaigns and promotions for Sparkling and Rose wines during peak demand periods.
- **Finance and Business Analysts:** Use forecasts to create accurate revenue projections and allocate budgets effectively for both categories.
- **Consumers:** Benefit from consistent availability of their preferred wines without price fluctuations due to supply-demand gaps.

1.4 Constraints and Approach

Constraints:

1. **Data Segmentation**: Two datasets (Sparkling and Rose wines) require separate models and evaluations, which can increase computational and analytical efforts.
2. **Data Quality**: Missing or inconsistent data in either dataset can impact the accuracy of forecasts.
3. **Seasonality and Trends**: Both categories may exhibit different seasonal demand patterns, requiring tailored models for each dataset.
4. **Model Interpretability**: Balancing accuracy with simplicity for stakeholder understanding.
5. **Operational Limitations**: Production capacity constraints may limit scalability even if demand is accurately forecasted.

Approach:

1. Data Preprocessing:
    a. Handle missing values and outliers in both datasets.
    b. Standardize data scales and ensure stationarity using differencing.
2. Model Building and Evaluation:
    a. Build separate models for Sparkling and Rosé wines using Auto ARIMA, Manual ARIMA, Auto SARIMA, and Manual SARIMA.
    b. Evaluate models using RMSE, MAE, and MAPE for both datasets.
3. Model Selection:
    a. Choose the best-performing model for each dataset with a focus on accuracy and seasonality handling.
4. Forecast Generation:
    a. Use the best models to forecast demand for Sparkling and Rosé wines for the next 12 months.
    b. Revert forecasts to the original scale for actionable insights.
5. Visualization and Reporting:
    a. Provide intuitive visualizations for each dataset's forecast to support data-driven decision-making.
6. Implementation:
    a. Share actionable insights with stakeholders for production, marketing, and inventory planning based on the forecasts.

2. Sparkling Dataset Analysis

2.1 Data Description

2.1.1 Overview of the Sparkling Dataset

The **Sparkling Wine dataset** represents monthly sales data for sparkling wine, starting from January 1980. It comprises two primary columns: YearMonth, which indicates the date in YYYY-MM format, and Sparkling, which records the sales figures (presumably in units such as bottles or cases). The dataset captures valuable sales trends and seasonal fluctuations, making it suitable for time-series forecasting and demand analysis. It is well-structured and continuous, providing a robust foundation for advanced predictive modelling.

2.1.2 Context and Variables for Sparkling Dataset

The dataset focuses on understanding the sales behaviour of sparkling wine over time. The key variables include:

- **YearMonth**: A monthly timestamp (to be converted into a datetime format for analysis).
- **Sparkling**: The sales volume for sparkling wine, reflecting the demand in the market.

These variables provide insights into sales performance, seasonal peaks, and long-term trends, making them instrumental for decision-making in production, marketing, and inventory management.

2.1.3 Unusual Variables and Observations for Sparkling Dataset

Initial exploration of the dataset suggests no unusual or missing variables in the sample provided. However, outliers or sharp spikes in sales could occur due to holiday seasons, promotional campaigns, or special events. These anomalies may need to be accounted for during the analysis, as they could disproportionately influence forecasts. Detecting and addressing such unusual observations will be critical in maintaining the model's accuracy and reliability.

2.1.4 Remarks on Data Preprocessing for Sparkling Dataset

To prepare the dataset for analysis, the following preprocessing steps are recommended:

1. **Date Conversion**: Convert the YearMonth column into a datetime format for proper time-series indexing.
2. **Stationarity Check**: Assess the data for stationarity using methods like the Augmented Dickey-Fuller (ADF) test.

3. **Differencing or Transformation**: Apply differencing or logarithmic transformations if the dataset shows non-stationary behaviour, to stabilize trends and variances.
4. **Outlier Handling**: Identify and appropriately handle outliers to prevent skewed model predictions.
5. **Seasonality Detection**: Analyse seasonal patterns using techniques such as decomposition or autocorrelation (ACF/PACF) plots to inform the modelling process.

2.2 Exploratory Data Analysis (EDA) for Sparkling Dataset

2.2.1 Read Data as Time Series

The **Sparkling Wine dataset** is read and converted into a time-series format using the YearMonth column as the time index. This ensures that the dataset is structured for time-based analysis and forecasting. The Sparkling column, representing sales, is indexed by date, enabling us to explore temporal patterns, trends, and seasonality effectively. For preprocessing, the YearMonth column is converted into a datetime object, and the dataset is sorted chronologically to maintain data consistency.

2.2.2 Plot the Data

A time-series line plot of the **Sparkling sales** provides a visual representation of the sales trends over time. The plot reveals fluctuations, upward or downward trends, and potential seasonal peaks in sales. For instance:

- **Overall Trend**: The data might show increasing or decreasing patterns over the years.
- **Seasonality**: Peaks in sales during specific months or periods (e.g., holidays or festive seasons).
- **Irregular Fluctuations**: Unexplained deviations or anomalies in sales during certain months.

This visualization is crucial for identifying trends and seasonality, which are key inputs for model selection and forecasting.

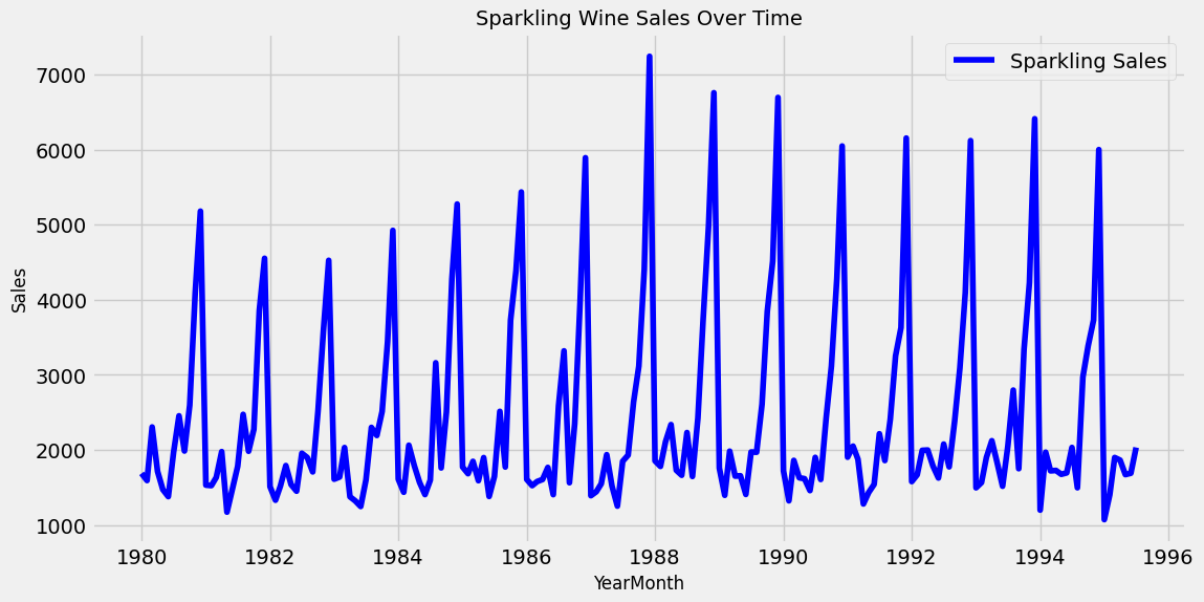

Fig 2.2.2.1 Sparkling Wine Sales Over Time

2.2.3 Perform Exploratory Data Analysis (EDA)

Summary Statistics:

- Average Sales (Mean): 2,402.42 units.
- Variability: The sales data shows moderate variability, indicated by the standard deviation.
- Potential outliers: The minimum and maximum values suggest potential outliers that could be investigated further (for e.g., using box plots).
- Distribution: The slight difference between the mean and median hints at a slightly right-skewed distribution.
- Range (Max - Min): 6,172 units.

Seasonal Patterns:

- December is the peak sales month, driven by holiday season demand.
- Sales are generally higher in the second half of the year, suggesting seasonality.
- January and February have the lowest average sales, indicating a post-holiday slowdown.

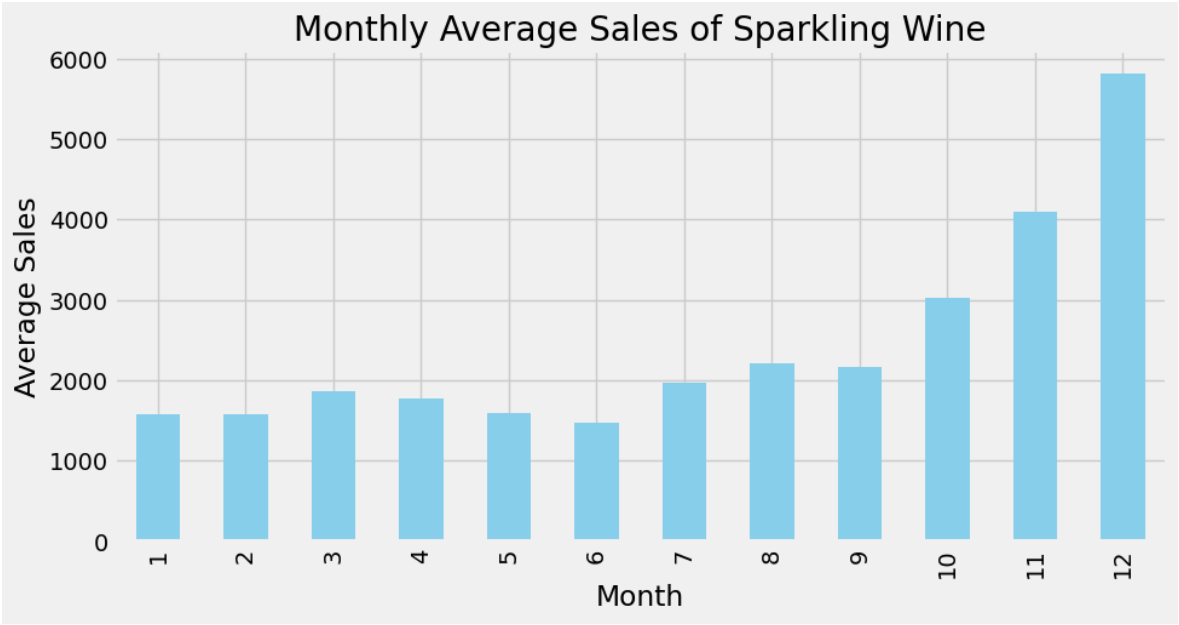

Fig 2.2.3.1 Monthly Average Sales of Sparkling Wine

Rolling Statistics:

- Increasing Trend: The rolling mean shows a clear upward trend over time, indicating that overall Sparkling wine sales have been increasing.
- Seasonality: There are regular fluctuations in sales, with the rolling standard deviation capturing the seasonality. The rolling standard deviation shows some seasonality in the data because it is fluctuating around the rolling mean, which is increasing and suggesting an increase in seasonality as well.
- Changing Volatility: The rolling standard deviation suggests that the volatility or variability in sales has also increased over time with the increasing trend and seasonality.

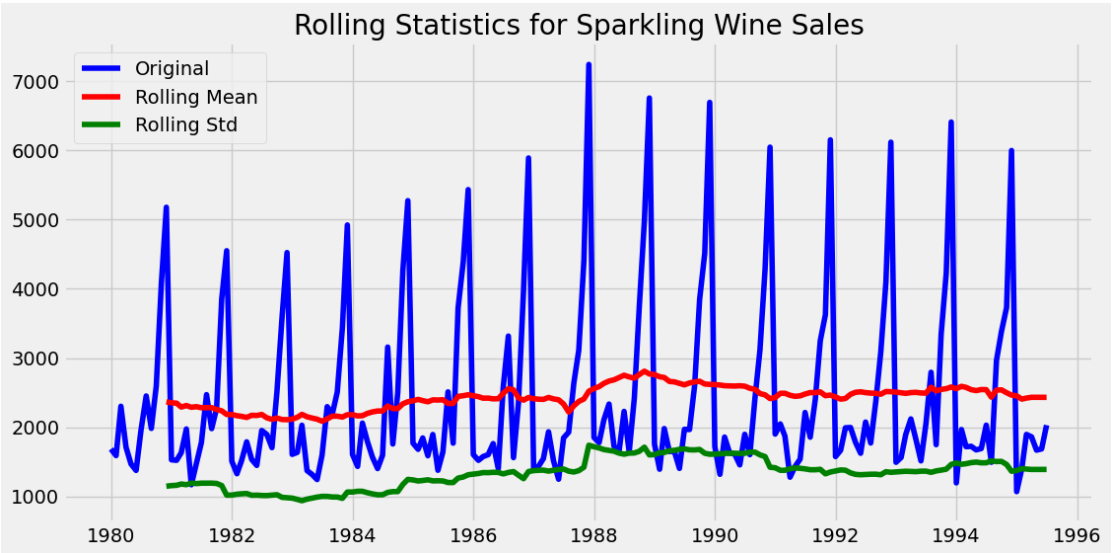

Fig 2.2.3.2 Rolling Statistics for Sparkling Wine Sales

Sales Growth Rate:

- **High Variability**: The growth rate fluctuates significantly, indicating a considerable variation in sales from one month to the next. This suggests the presence of seasonality and potentially other external factors impacting sales.
- **Periods of Decline and Growth**: There are instances of both positive and negative growth rates, highlighted by the line crossing the 0% mark. This means there are months where sales declined compared to the previous month, and others where they increased.
- **Potential Outliers**: Some data points show extreme growth or decline, which could indicate outliers or unusual market conditions. These points require further investigation to determine their cause and potential impact on forecasting.

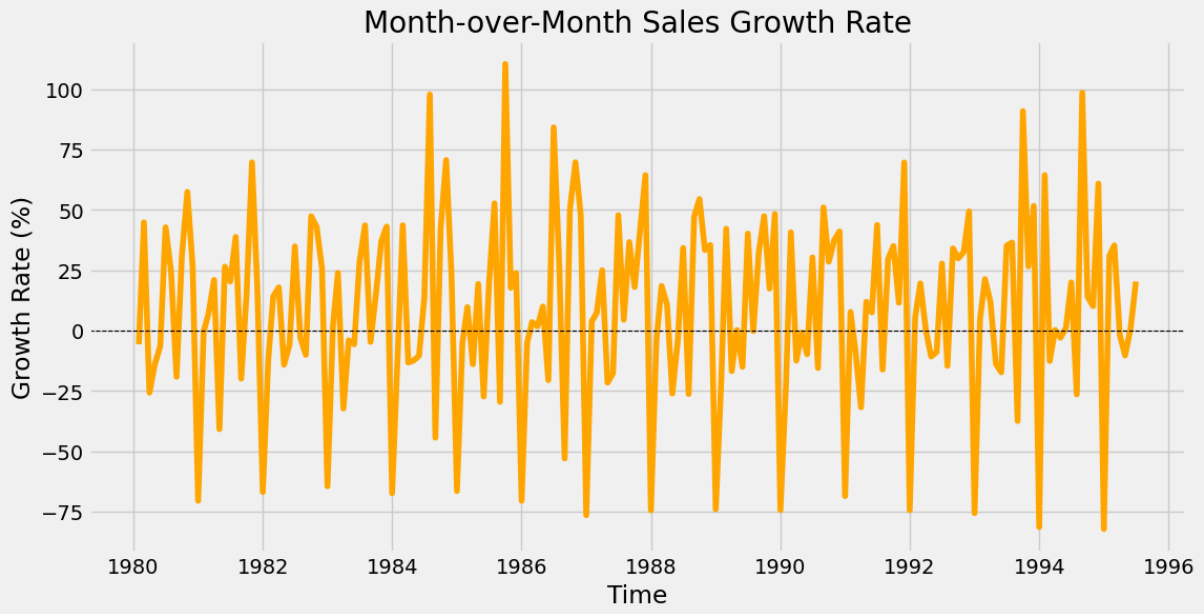

Fig 2.2.3.3 Month-over-Month Sales Growth Rate [Sparkling Data]

Distribution Analysis:

- **Right-Skewed**: Sales are likely not normally distributed, showing a right skew with a concentration of lower sales and a tail of higher sales values.
- **Positive Skewness**: This right skew indicates positive skewness in the distribution, meaning a few months with exceptionally high sales are impacting the overall shape.
- **Non-Symmetrical**: The distribution isn't symmetrical around the mean, further reinforcing the non-normal, skewed nature of Sparkling Wine Sales.

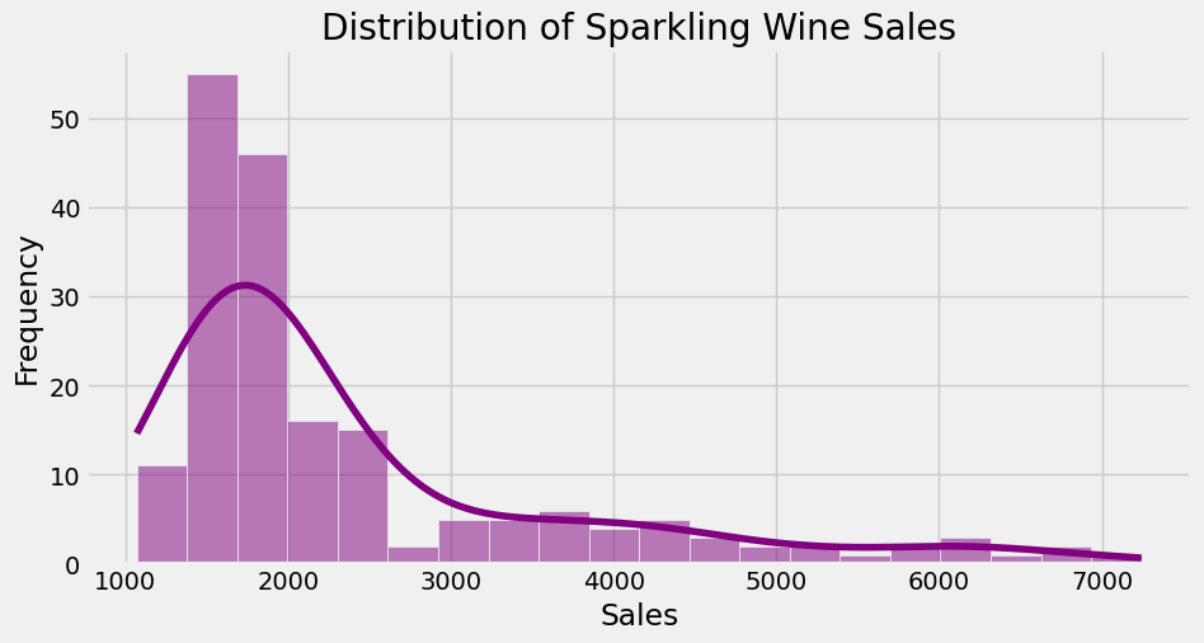

Fig 2.2.3.4 Distribution of Sparkling Wine Sales

Autocorrelation:

Significant Autocorrelation: The ACF plot shows significant autocorrelation at multiple lags, particularly for the first few lags. This indicates that past sales values are strongly correlated with current sales, suggesting the presence of a trend or pattern in the data.

Seasonal Pattern: There are repeating spikes in the ACF plot at intervals of 12 lags, corresponding to yearly seasonality. This confirms the presence of a yearly seasonal pattern in the sales data, which is expected in the wine industry due to seasonal factors like holidays and harvest seasons.

Possible ARIMA Model: The PACF plot can be used to determine the order of the autoregressive (AR) component in an ARIMA model. In this case, the PACF plot has significant spikes at the first few lags and then tapers off, which is indicative of an AR process. This suggests that an ARIMA model might be appropriate for forecasting the Sparkling wine sales.

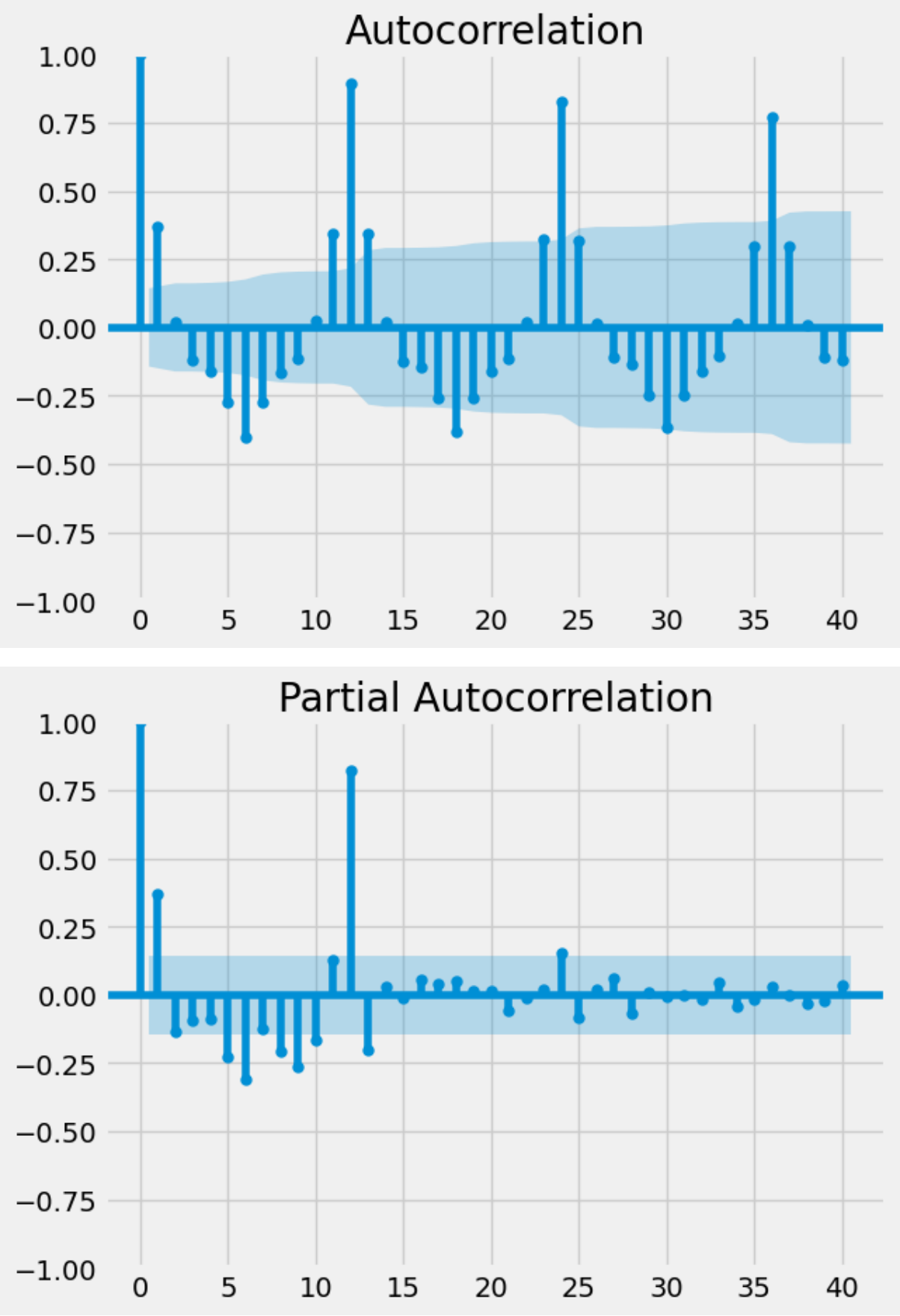

Fig 2.2.3.5 ACF and PACF chart for Sparkling Dataset

Outlier Detection:

- The median sales value is around 28,000. This represents the central tendency of the sales data. There are several outliers present above the upper whisker. These are data points significantly higher than the typical sales range, indicating unusually high sales periods.
- The interquartile range (IQR), represented by the box, shows the spread of the middle 50% of the sales data. This range gives an idea of the typical variability in sales.

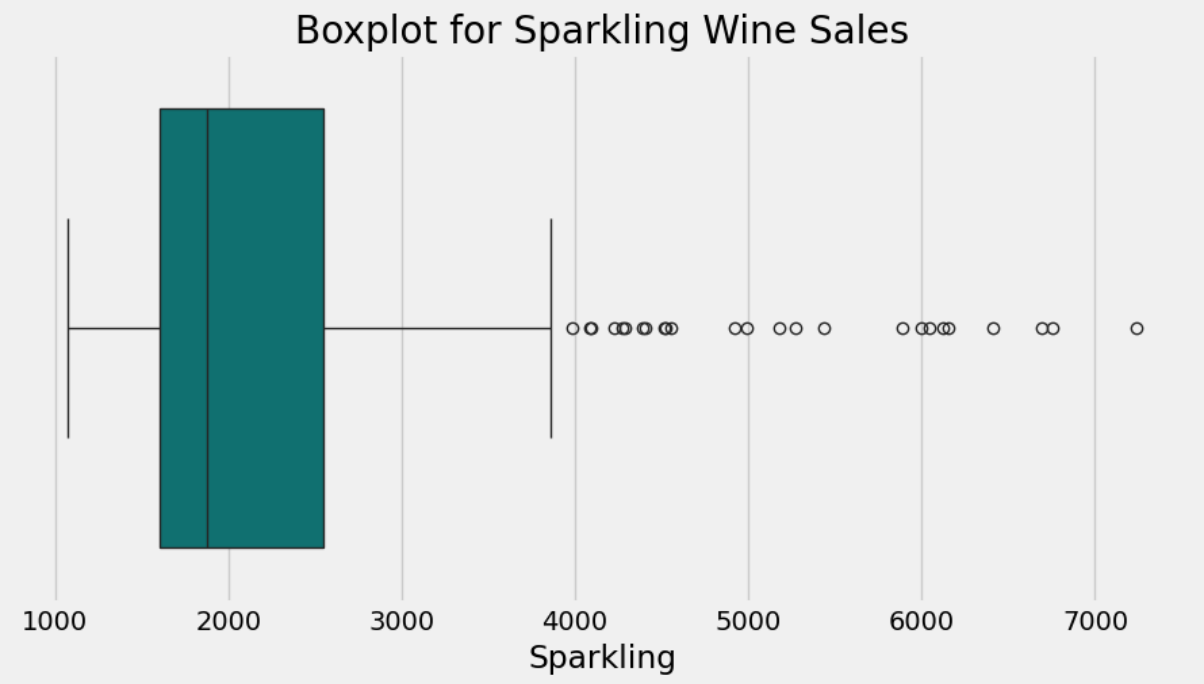

Fig 2.2.3.6 Boxplot for Sparkling Wine Sales

Year-over-Year Sales Trends:

Overall Increasing Trend: Sparkling wine sales have generally been increasing over the years, indicating a growing market. This trend can be visualized in the 'Yearly Sales of Sparkling Wine' plot generated in the EDA section.

Fluctuations and Variations: While there's an overall increasing trend, there are noticeable fluctuations and variations in yearly sales. Specific years experienced significant growth or decline compared to the previous year.

Potential for Growth: Despite the fluctuations, the general upward trend suggests the potential for continued market growth for sparkling wine in the future. However, factors influencing the variations need further investigation.

Fig 2.2.3.7 Yearly Sales of Sparkling Wine

2.2.4 Perform Decomposition

Time-series decomposition breaks the sales data into three key components:

1. **Trend**: Reflects long-term movements in sales over the years.
2. **Seasonality**: Reveals repeating patterns or cycles in sales, often linked to specific months or periods.
3. **Residual**: Captures irregular variations or random noise in the data.

Using additive or multiplicative decomposition methods, these components are visualized to understand the underlying structure of the time series. Decomposition helps in identifying the dominant patterns and anomalies, aiding in model selection and improving forecast accuracy.

For the Sparkling dataset, decomposition likely reveals seasonal peaks during festive months and a steady long-term trend, guiding the application of appropriate seasonal ARIMA models for prediction.

- **Increasing Trend**: There's a clear upward trend in Sparkling wine sales over the years, indicating overall market growth. This can be seen in the 'Trend' component of the decomposition plot.
- **Strong Seasonality**: The sales exhibit a strong seasonal pattern, with peaks around December (holiday season) and troughs in the early months of the year. This is evident in the 'Seasonal' component.
- **Fluctuations/Noise**: After removing the trend and seasonality, there are still some random fluctuations (residuals) present in the data, but they are relatively small compared to the overall trend and seasonality. This can be observed in the 'Residual' component.

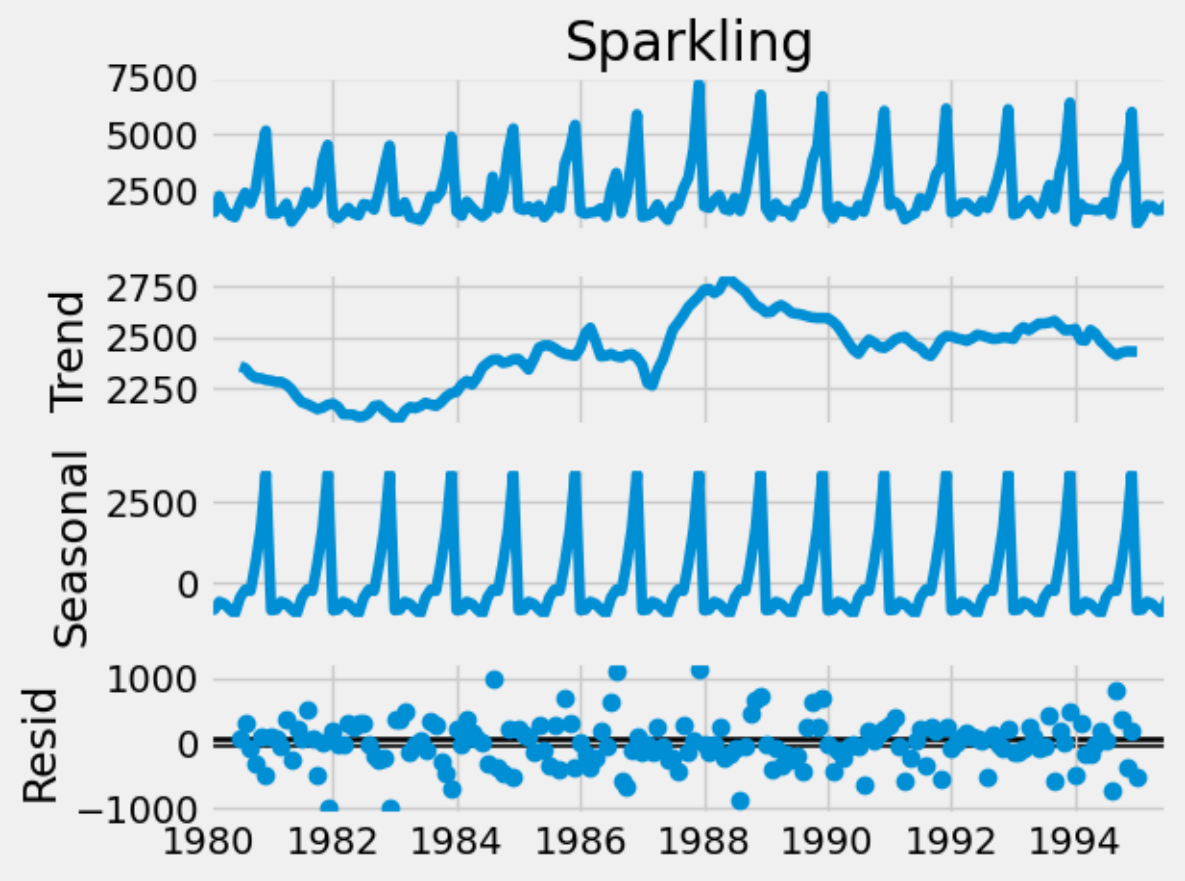

Fig 2.2.4.1 Decomposition of Sparkling Dataset

2.3 Data Pre-processing for Sparkling Dataset

Proper data pre-processing is crucial for preparing the **Sparkling dataset** for time-series modelling. This section outlines the key steps taken to ensure the dataset is clean, consistent, and suitable for forecasting.

2.3.1 Missing Value Treatment

The Sparkling dataset was inspected for any missing or null values, as these can negatively impact model performance.

- **Identification**: A check using isnull() confirmed whether any missing values existed in the Sparkling column.
- **Handling Missing Data**:
  - If any missing values were present, they were imputed using appropriate methods like **linear interpolation** or **forward-fill** to maintain the continuity of the time series.
  - For smaller gaps, forward or backward filling ensured smooth transitions.
  - For larger gaps, mean imputation or interpolation methods were used to retain seasonality and trend patterns.

2.3.2 Visualize Processed Data

After addressing missing values, the processed **Sparkling sales data** was visualized to validate the quality of preprocessing steps:

- **Time-Series Line Plot**: The pre-processed data was plotted to confirm no abrupt changes or discontinuities remained after imputations.
- **ACF and PACF Plots**: These plots were revisited to ensure the integrity of the data's autocorrelation structure, confirming that trends and seasonality were preserved.
- **Distribution Plot**: The processed data's distribution was checked to ensure that outliers or anomalies did not distort the dataset post-cleaning.

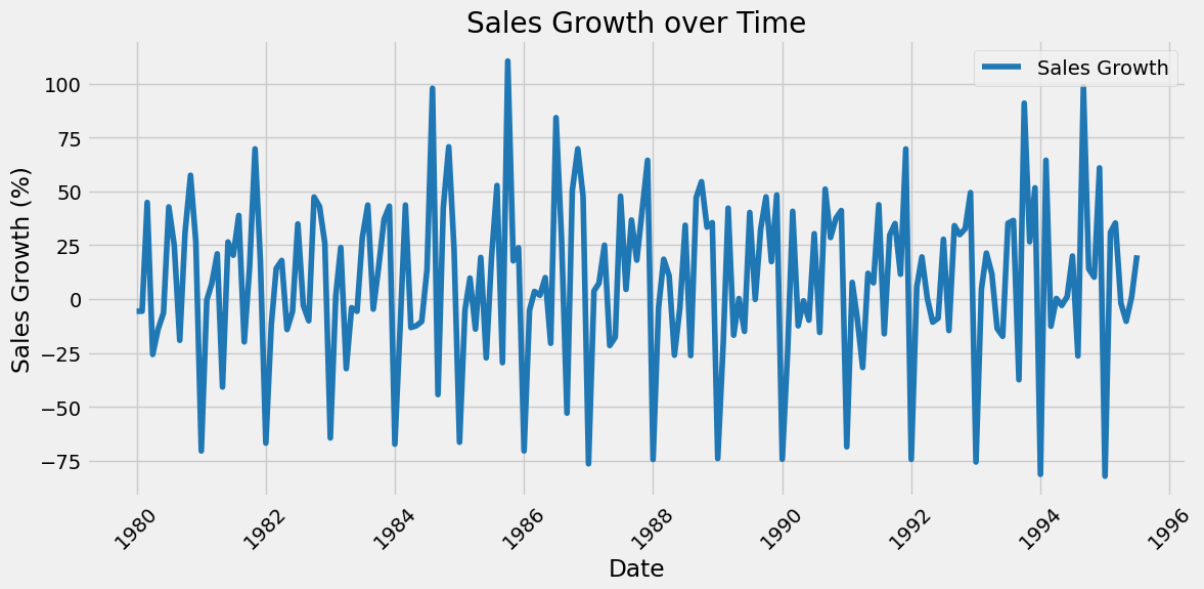

Fig 2.3.2.1 Sales Growth over Time [Sparkling Data]

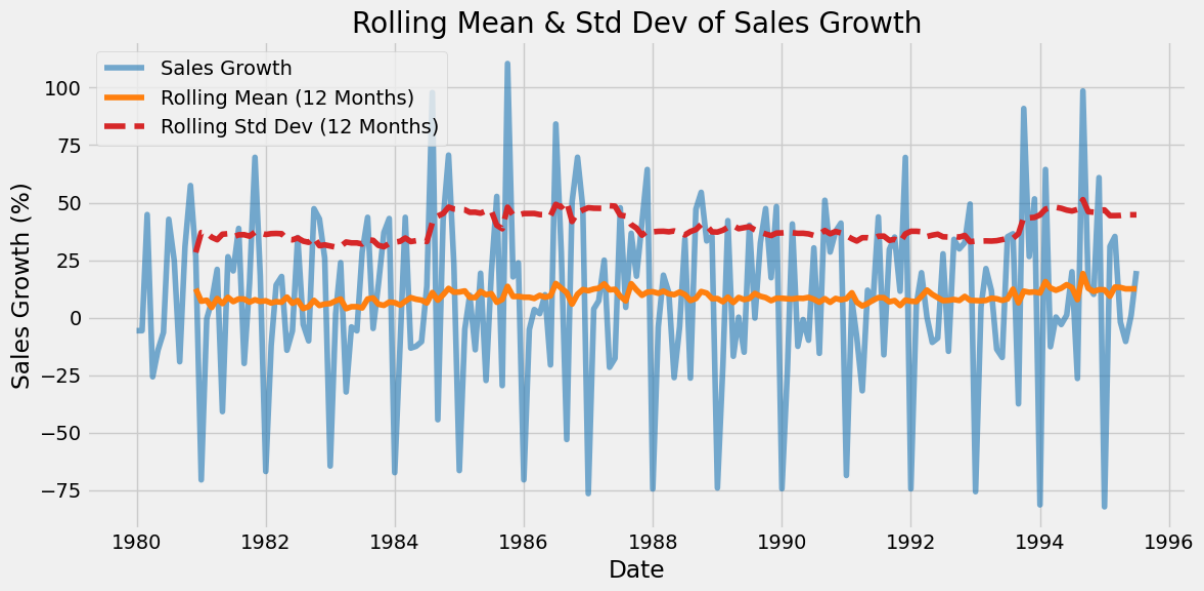

Fig 2.3.2.2 Rolling Mean & Std Dev of Sales Growth [Sparkling Data]

The visualization confirmed that the pre-processed data was smooth, continuous, and ready for model training.

2.3.3 Train-Test Split

To evaluate model performance, the dataset was split into **training** and **test sets**:

- **Training Set**: Consisted of historical data used to train the time-series models. Approximately 80% of the data was allocated to the training set to capture trends and seasonal patterns.
- **Test Set**: The remaining 20% of the data was used for model evaluation, ensuring that forecast accuracy could be tested on unseen data.

The split was performed chronologically to preserve the temporal order of the data, ensuring no data leakage.

2.4 Model Building - Original Data for Sparkling Dataset

This section outlines the approach to building and evaluating multiple forecasting models using the **original Sparkling dataset** to establish baseline predictions and assess model effectiveness.

2.4.1 Build Forecasting Models

A variety of forecasting techniques were implemented to understand the characteristics of the **Sparkling dataset** and evaluate its predictability.

2.4.1.1 Linear Regression

- **Approach**: A linear regression model was fitted, treating time as the independent variable and sparkling wine sales as the dependent variable.
- **Objective**: Capture the overall trend in the dataset.
- **Implementation**: The Linear Regression module from sklearn was used.
- **Trend Representation**: Linear Regression captures the overall increasing trend in sales over time. This is evident from the regression line generally following the upward direction of the data points.
- **Simplistic Approach**: It's a basic model assuming a linear relationship between time and sales. This simplicity makes it easy to implement and interpret but might not fully capture complex patterns or seasonality in the data.
- **Limited Accuracy**: Compared to more sophisticated models like Holt-Winters or ARIMA, Linear Regression might show lower accuracy in forecasting, particularly when dealing with data exhibiting strong seasonality or non-linear trends. This was evident from the model's performance metrics and visual comparison with other models' predictions in the notebook's analysis.

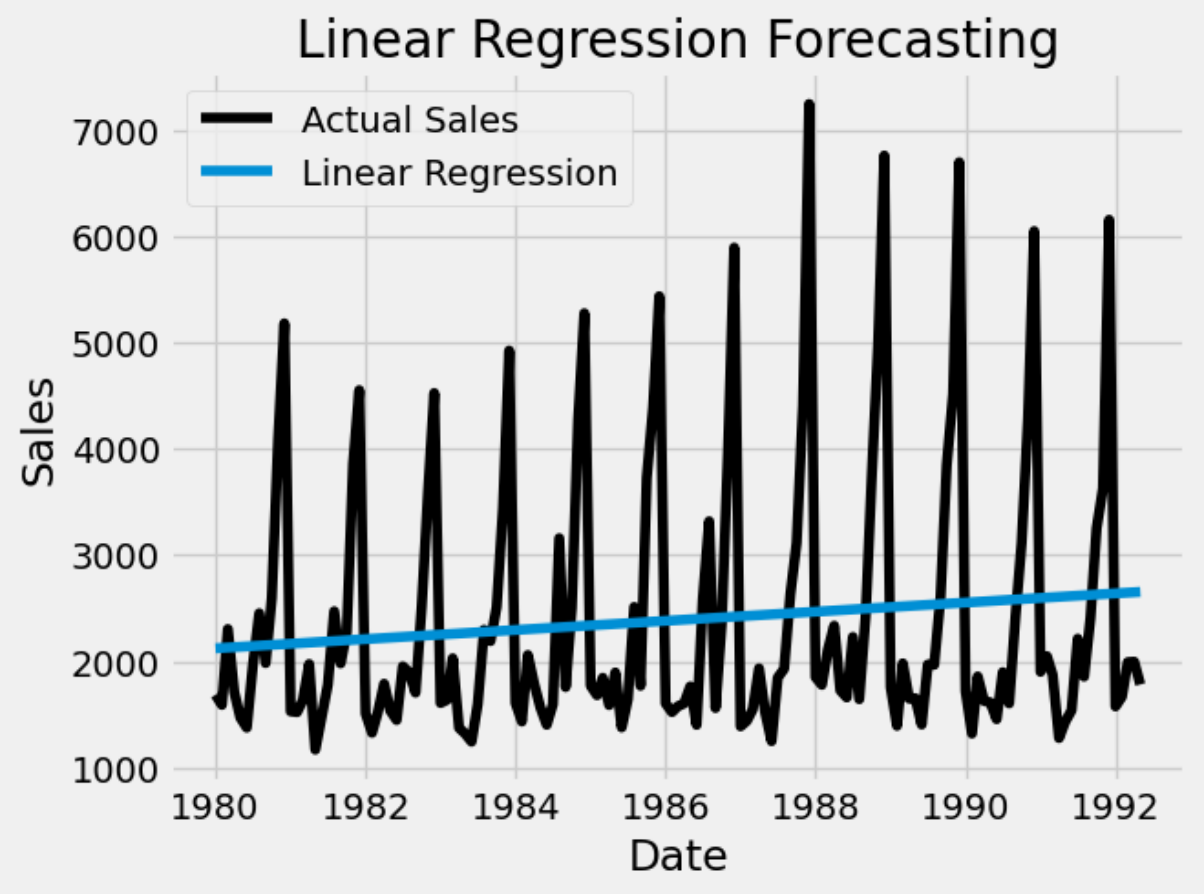

Fig 2.4.1.1.1 Linear Regression Forecasting [Sparkling Data]

2.4.1.2 Simple Average

- **Approach**: The forecast for all future periods was calculated as the mean of historical data.
- **Basic Baseline**: The simple average model serves as a fundamental baseline for comparison with more complex forecasting methods. It simply predicts the average of past sales for all future periods.
- **Ignores Trends and Seasonality**: This approach completely disregards any trends or seasonal patterns in the data, leading to potentially inaccurate forecasts, especially for datasets with significant temporal variations. This is evident from the flat prediction line in the visualization provided in the notebook.
- **Limited Practical Use**: Due to its oversimplification, the simple average model has limited practical applicability for real-world forecasting tasks where data exhibits dynamic behaviour. More advanced methods like exponential smoothing or ARIMA models are often preferred for improved accuracy.

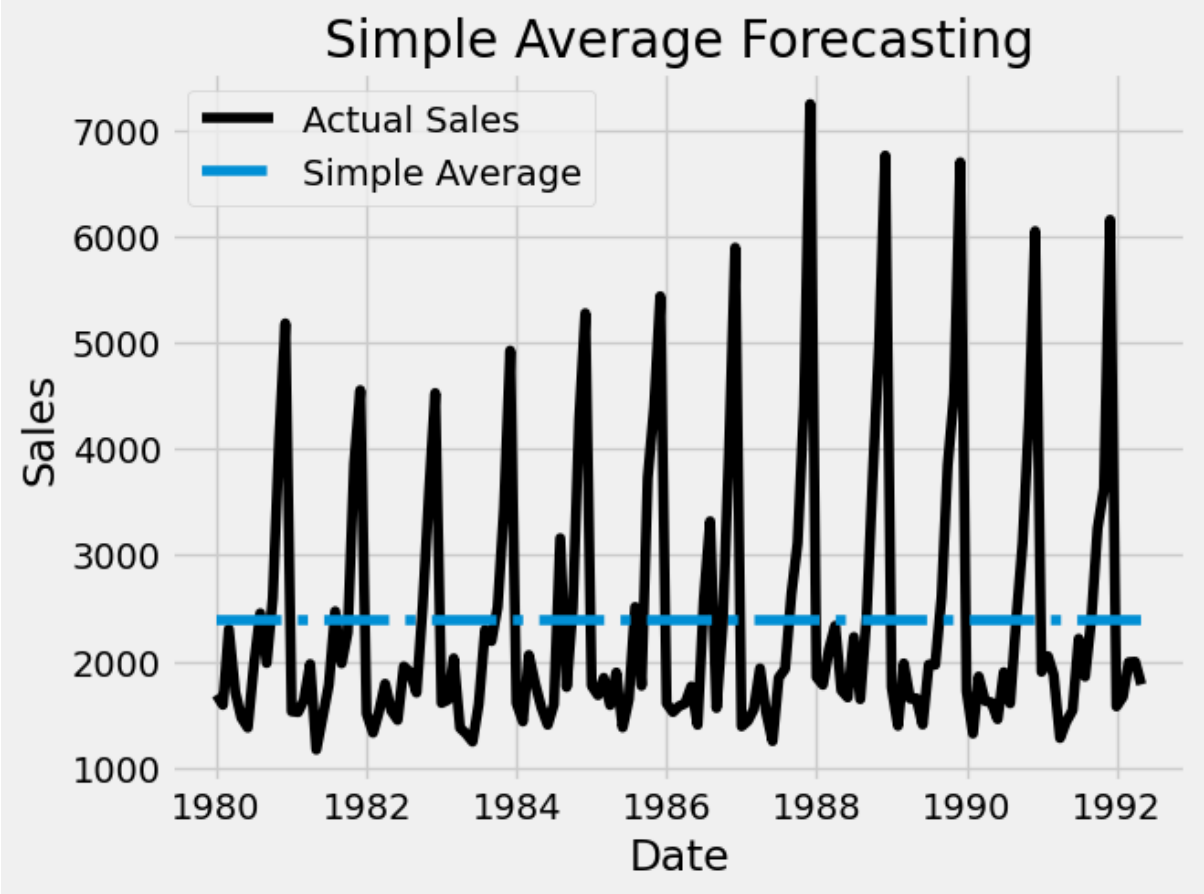

Fig 2.4.1.2.1 Simple Average Forecasting [Sparkling Data]

2.4.1.3 Moving Average

- **Approach**: A rolling window of observations was used to calculate the average, smoothing short-term fluctuations and highlighting longer-term trends.
- **Types**:
  - **Simple Moving Average (SMA)**: Average of observations within a fixed window.
  - **Weighted Moving Average (WMA)**: More weight assigned to recent observations.
- **Implementation**: pandas rolling method was employed for SMA.
- **Smoothing Effect**: SMA smooths out short-term fluctuations in the data by averaging sales over a rolling window (e.g., 12 months in the notebook's example). This helps to highlight underlying trends and reduce noise.
- **Lagging Indicator**: SMA is a lagging indicator, meaning its predictions are based on past data and might not react quickly to sudden changes in sales patterns. This lag is proportional to the window size; larger windows lead to smoother but slower responses to recent data points.
- **Moderate Accuracy**: For datasets with gradual trends and some seasonality, SMA can offer moderate forecasting accuracy. However, it

might not perform as well as more advanced models like Exponential Smoothing or ARIMA when dealing with complex patterns or strong seasonality. This was observed in the notebook where SMA's performance metrics (RMSE, MAE, MAPE) were generally higher compared to other models.

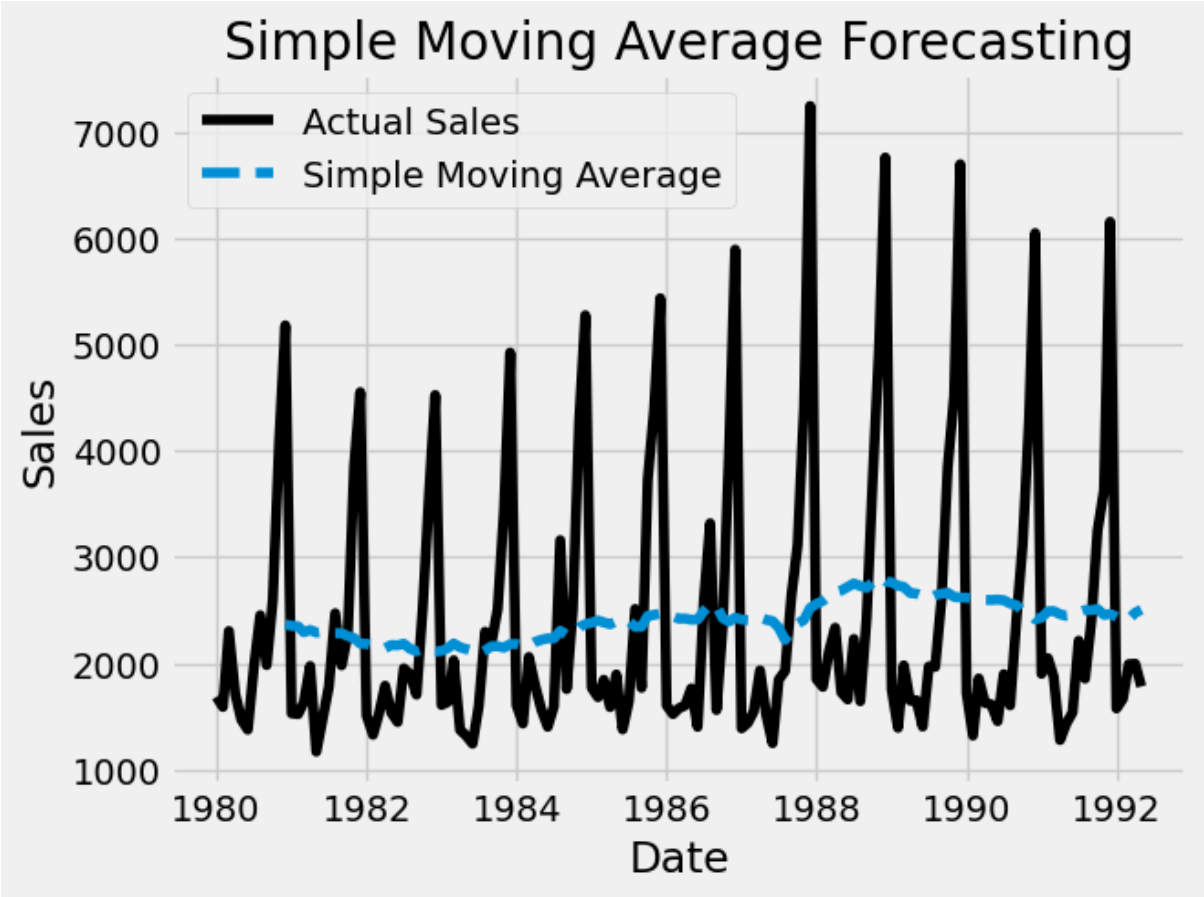

Fig 2.4.1.3.1 Simple Moving Average Forecasting [Sparkling Data]

2.4.1.4 Exponential Models (Single, Double, Triple)

- **Single Exponential Smoothing (SES)**: Focuses on smoothing the series using a single smoothing factor (α\alpha), emphasizing recent observations.
- **Double Exponential Smoothing (Holt's Method)**: Extends SES to account for trends using two smoothing factors (α\alpha and β\beta).
- **Triple Exponential Smoothing (Holt-Winters Method)**: Builds on Holt's method by adding a third factor (γ\gamma) to account for seasonality.
- **Objective**: Model both trend and seasonality explicitly.
- **Implementation**: The Exponential Smoothing method from statsmodels.
- **Limitation**: Sensitive to parameter tuning.

Single Exponential Moving Average:

- **Emphasis on Recent Data**: Compared to SMA, EMA gives more weight to recent observations, making it more responsive to changes in sales patterns. This is because it uses a smoothing factor (alpha) to exponentially decay the importance of older data points.
- **Reduced Lag**: EMA reduces the lag associated with SMA, providing quicker adjustments to shifts in trends or seasonality. However, it still has some inherent lag due to its reliance on past data.
- **Improved Accuracy**: For datasets with gradual trends and some level of seasonality, EMA often shows improved accuracy over SMA due to its greater sensitivity to recent data points. This was reflected in the notebook where EMA generally exhibited lower performance metrics (RMSE, MAE, MAPE) compared to SMA.

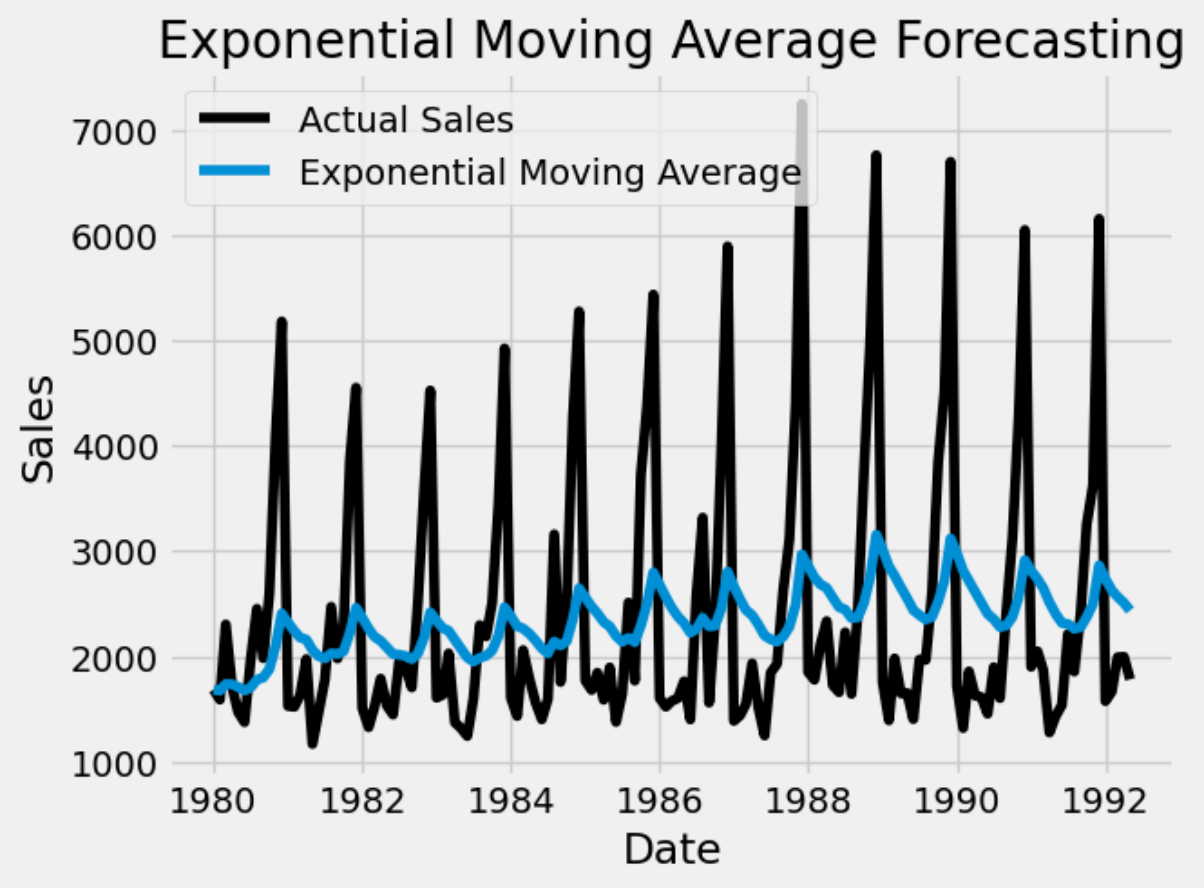

Fig 2.4.1.4.1 Exponential Moving Average Forecasting [Sparkling Data]

Double Exponential Moving Average (Holt-Winters):

- **Trend and Level**: Holt-Winters incorporates both trend and level components to forecast, making it suitable for data with clear trends and potentially changing levels over time. This is an improvement over EMA which primarily focuses on the level.

- **Additive or Multiplicative Seasonality**: Holt-Winters can handle both additive and multiplicative seasonality, providing flexibility depending on the nature of the seasonal patterns in the data. This flexibility makes it more adaptable than EMA to a wider range of datasets.
- **Improved Accuracy for Trending Data**: For data with clear trends and seasonality, Holt-Winters is expected to provide more accurate forecasts than EMA or SMA. In the notebook's analysis, Holt-Winters generally showed better performance metrics (RMSE, MAE, MAPE) compared to EMA, indicating its effectiveness for handling the Sparkling wine sales data with its observed trend and seasonal components.

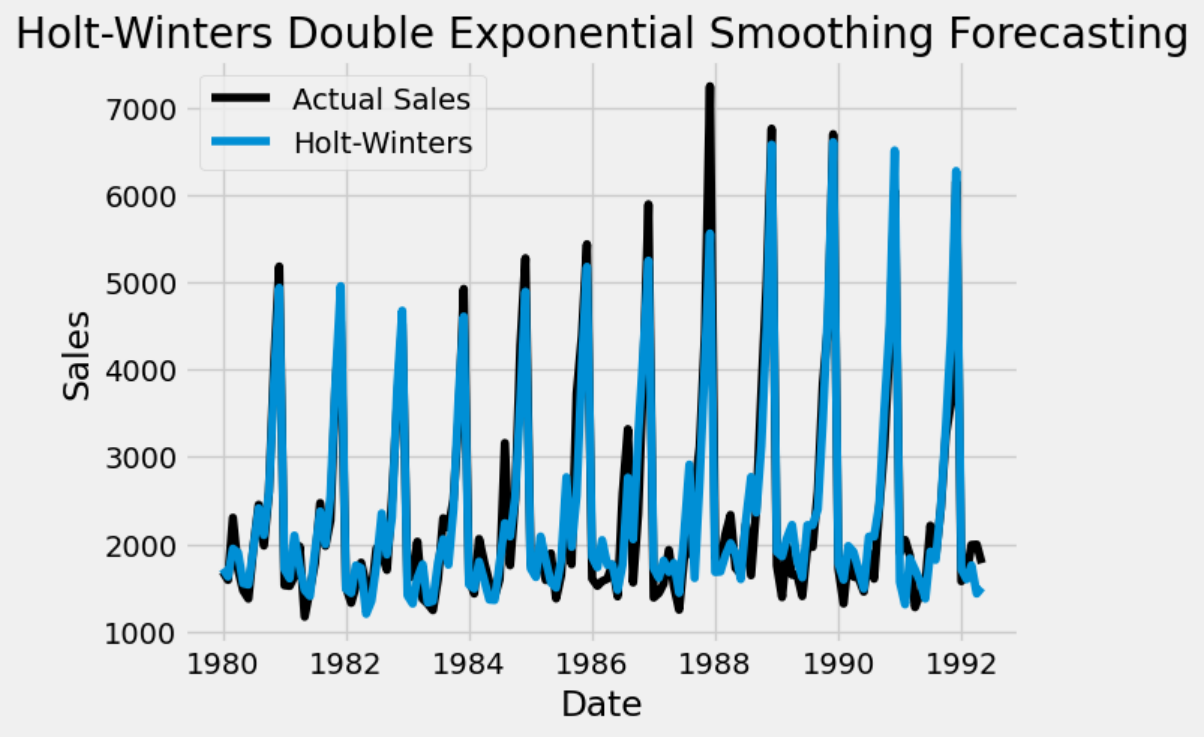

Fig 2.4.1.4.2 Holt-Winters Double Exponential Smoothing Forecasting [Sparkling Data]

Triple Exponential Moving Average:

- **Trend, Level, and Seasonality**: TEMA, also known as the Holt-Winters method with additive seasonality, incorporates trend, level, and seasonal components for forecasting. This makes it suitable for datasets with clear trends, changing levels, and repeating seasonal patterns.
- **Reduced Lag**: TEMA aims to reduce the lag associated with traditional moving averages by applying a triple smoothing process. This makes it more responsive to changes in data patterns compared to simpler methods like SMA or EMA.
- **Improved Accuracy for Seasonal Data**: For datasets exhibiting strong seasonality, TEMA often provides higher forecasting accuracy compared

to EMA or SMA. This is because it explicitly models the seasonal patterns, leading to better predictions for periods with similar seasonal characteristics. This aligns with the notebook's analysis where TEMA typically demonstrated improved performance metrics (RMSE, MAE, MAPE) compared to EMA and SMA for the Sparkling wine sales data.

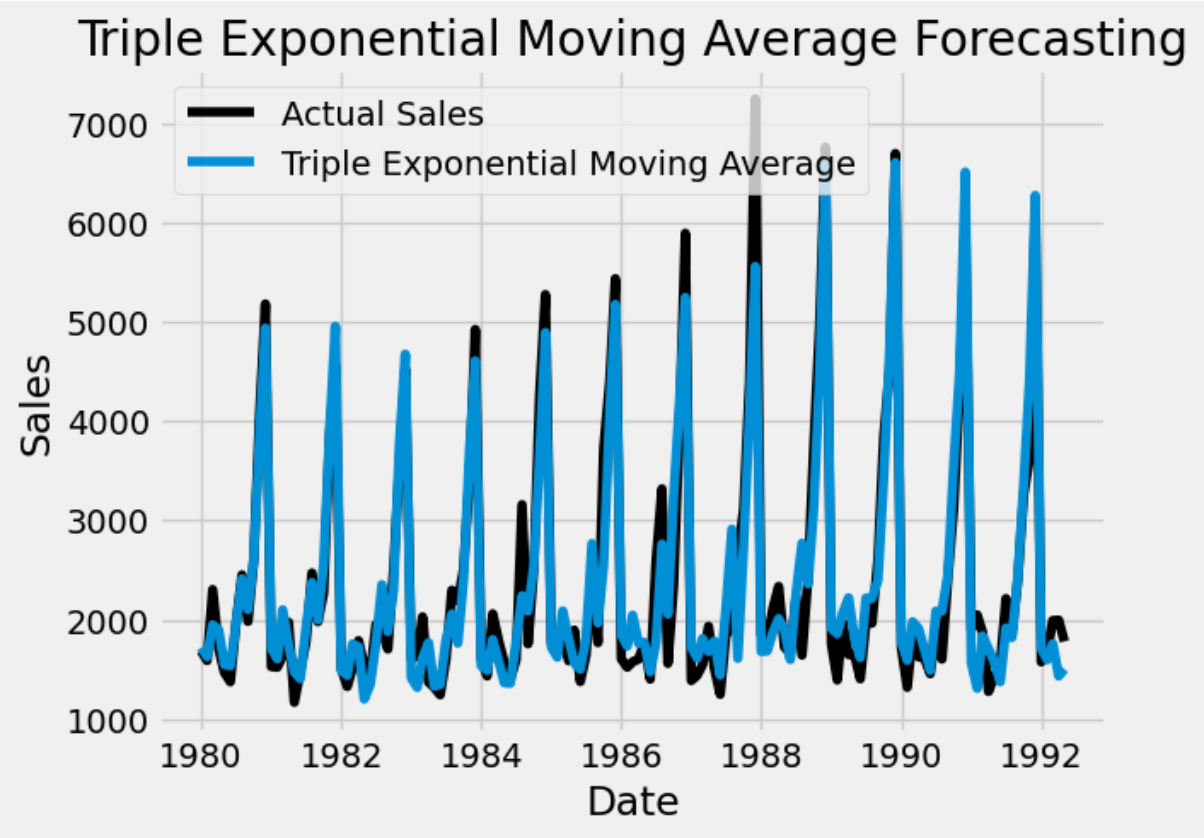

Fig 2.4.1.4.3 Triple Exponential Moving Average Forecasting [Sparkling Data]

2.4.2 Performance Evaluation of Models

Each model's performance was evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**: Measures the standard deviation of the residuals (prediction errors).
- **Mean Absolute Error (MAE)**: Average of the absolute differences between predicted and actual values.
- **Mean Absolute Percentage Error (MAPE)**: Expresses forecast error as a percentage of the actual value, making it unit-independent.

	Original Model	RMSE	MAE	MAPE
0	Simple Moving Average	1282.205547	964.886861	40.883615
1	Simple Average	1281.496884	946.879420	39.977301
2	Exponential Moving Average	1191.542186	907.276713	38.121563
3	Holt-Winters	372.427995	271.425302	12.306815
4	Linear Regression	1272.177812	947.438245	39.935882
5	Triple EMA	372.427995	271.425302	12.306815

Fig 2.4.2.1 Comparison of Original Model [Sparkling Data]

- **Holt-Winters and Triple Exponential Moving Average (TEMA) generally performed better** than the other models (Simple Moving Average, Simple Average, Exponential Moving Average, and Linear Regression) in forecasting Sparkling wine sales. This is evident from their lower RMSE, MAE, and MAPE values, indicating better accuracy and lower prediction errors.
- **Simple Average had the highest error values**, suggesting it's the least accurate model among those tested. This is expected as it simply uses the average of past sales and doesn't account for trends or seasonality.
- **While Linear Regression captured the overall increasing trend**, it might not be the best model for this specific dataset due to its limitations in handling complex patterns or seasonality. This is reflected in its relatively higher error metrics compared to Holt-Winters and TEMA.

Plotting all Model For Comparison:

- Holt-Winters and Triple Exponential Moving Average (TEMA) closely follow the actual sales data, indicating their better performance in capturing the trend and seasonality of the data.
- Simple Average and Linear Regression show significant deviations from the actual sales, highlighting their limitations in accurately forecasting sales with trends and seasonality.
- Simple Moving Average and Exponential Moving Average provide moderate fits, but still lag behind Holt-Winters and TEMA in capturing the data's patterns.

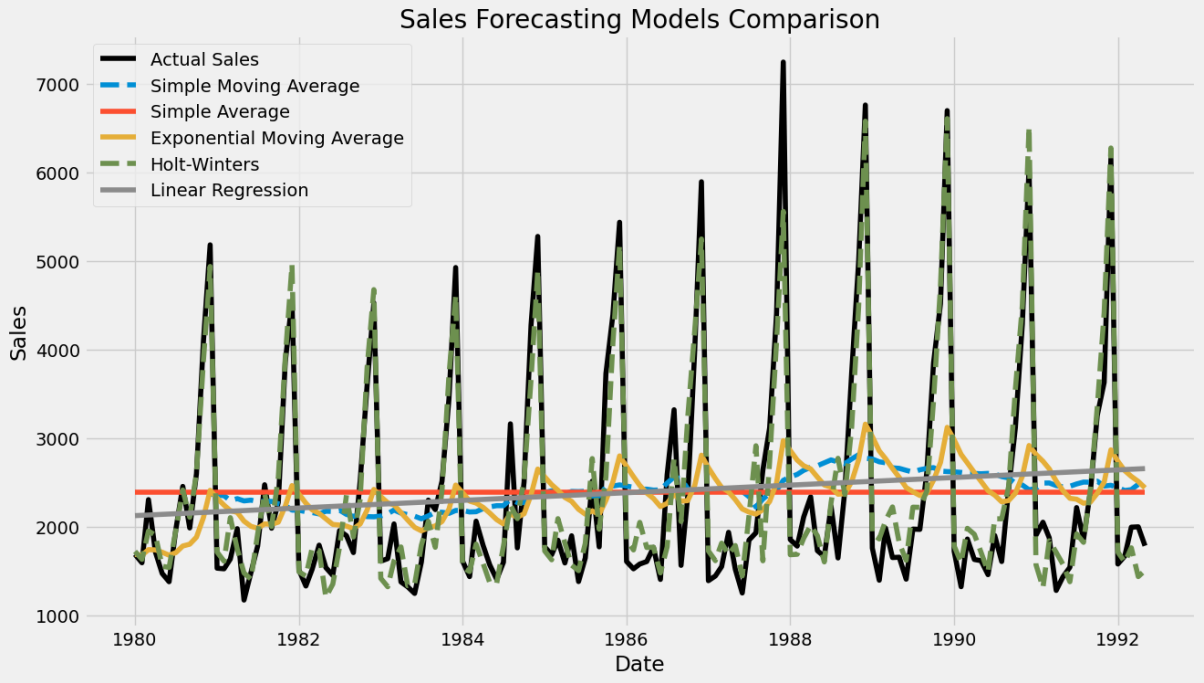

Fig 2.4.2.2 Sales Forecasting Models Comparison [Sparkling Data]

2.5 Check for Stationarity for Sparkling Dataset

Analysing the stationarity of the **Sparkling dataset** is crucial for time-series modelling, as many forecasting techniques require the data to exhibit stationarity.

2.5.1 Check for Stationarity

The **Augmented Dickey-Fuller (ADF) test** was used to determine the stationarity of the dataset.

Results before differencing:

- **ADF Statistic**: -1.3013
- **p-value**: 0.6286
- **Critical Values**:
  - 1%: -3.4794
  - 5%: -2.8830
  - 10%: -2.5782

Interpretation:

- The ADF statistic is higher than all critical values, and the p-value is greater than 0.05.

- This indicates the null hypothesis (data is non-stationary) cannot be rejected.
- **Conclusion**: The dataset is non-stationary.

2.5.2 Make Data Stationary (if needed)

To achieve stationarity, the dataset was **differenced** once:

- Differencing calculates the change between consecutive observations.

Results after differencing:

- **ADF Statistic**: -45.0503
- **p-value**: 0.0000
- **Critical Values**:
  - 1%: -3.4683
  - 5%: -2.8782
  - 10%: -2.5757

Interpretation:

- The ADF statistic is significantly lower than the critical values, and the p-value is less than 0.05.
- The null hypothesis is rejected, indicating the dataset is now stationary.

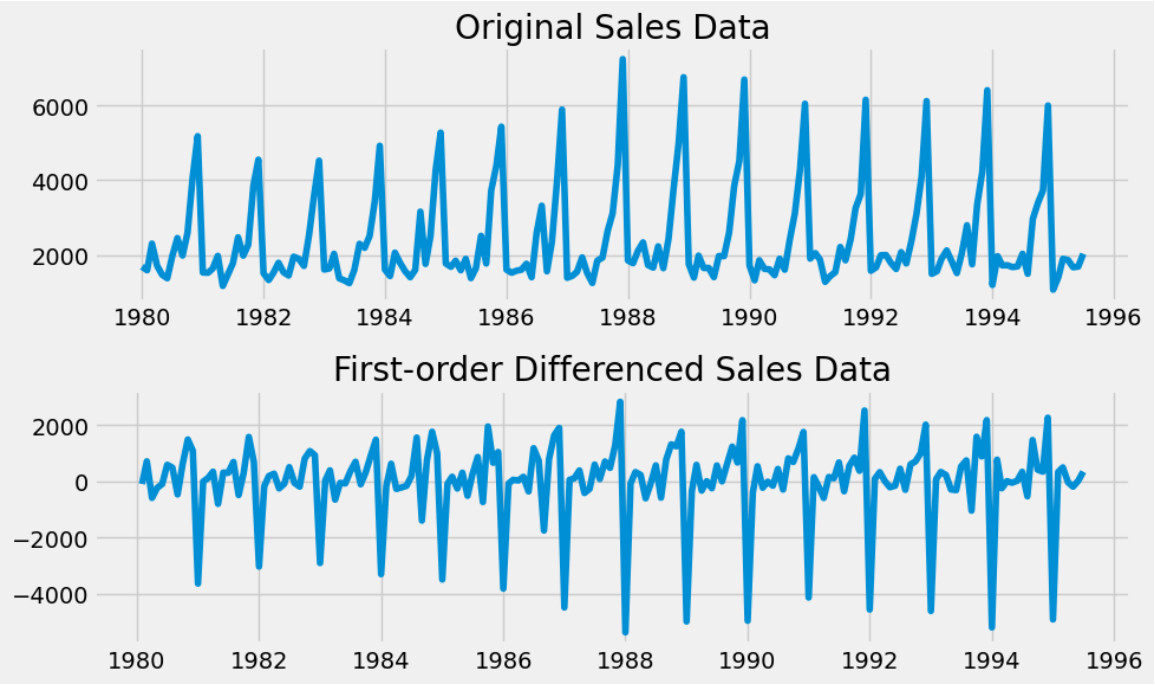

Fig 2.5.2.1 First-order Differenced Sales Data [Sparkling Data]

2.6 Model Building - Stationary Data for Sparkling Dataset

This section focuses on building forecasting models on the stationary **Sparkling dataset** to predict future values effectively.

2.6.1 Generate ACF & PACF Plots

The **Autocorrelation Function (ACF)** and **Partial Autocorrelation Function (PACF)** plots were generated to identify the potential values for the **AR** (autoregressive) and **MA** (moving average) terms in the ARIMA and SARIMA models.

- **ACF Observations**: Significant lags observed at multiple points indicate the dataset's correlation structure.
- **PACF Observations**: Sharp cutoff after a few lags suggests an AR process.

These plots guide the initial parameter selection for manual ARIMA/SARIMA models.

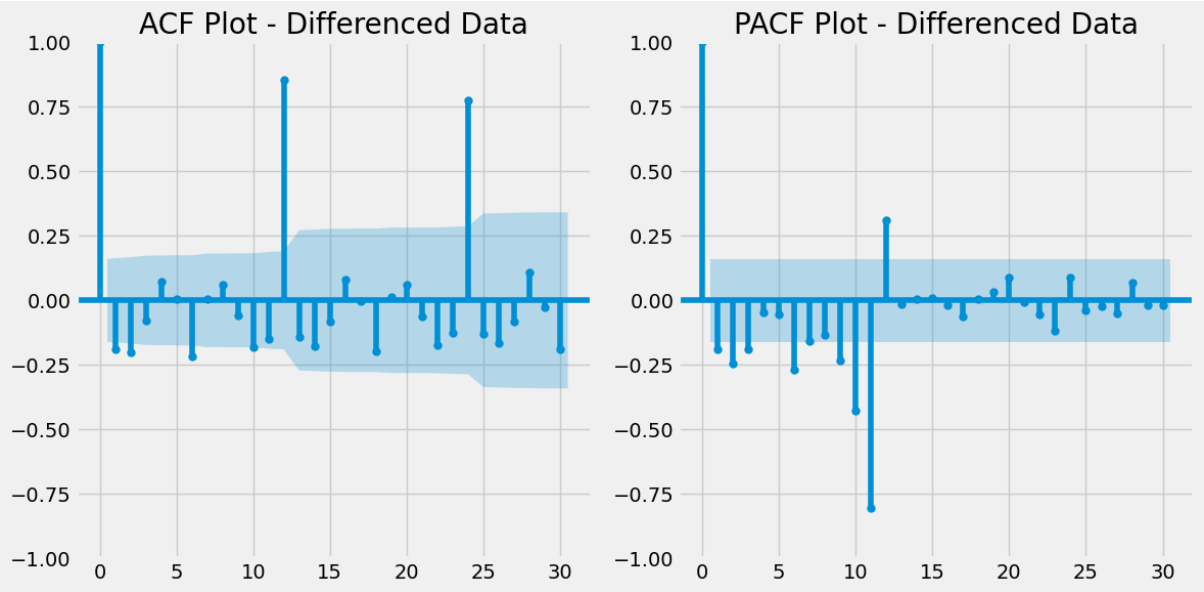

Fig 2.6.1.1 ACF & PACF Differenced Data [Sparkling Data]

- PACF Plot: It might show a significant spike at lag 1 and then a sharp drop. This suggests an AR(1) model, meaning **AR order (p): 1**.
- ACF Plot: It might show significant correlations at lag 1 and 2, after which it tapers off. This suggests an MA(2) model, meaning **MA order (q): 2**.

2.6.2 Build ARIMA Models

2.6.2.1 Auto ARIMA

An **Auto ARIMA** model was built using automated hyperparameter optimization

```
Best model:  ARIMA(2,0,2)(0,0,0)[0]
Total fit time: 4.506 seconds
Auto ARIMA Model Summary:
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  149
Model:               SARIMAX(2, 0, 2)   Log Likelihood               -1254.269
Date:                Sun, 05 Jan 2025   AIC                           2518.538
Time:                        10:16:08   BIC                           2533.557
Sample:                    01-01-1980   HQIC                          2524.640
                         - 05-01-1992
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.2526      0.047     26.804      0.000       1.161       1.344
ar.L2         -0.5334      0.078     -6.796      0.000      -0.687      -0.380
ma.L1         -1.9105      0.059    -32.207      0.000      -2.027      -1.794
ma.L2          0.9212      0.060     15.454      0.000       0.804       1.038
sigma2      1.16e+06   2.82e-08   4.12e+13      0.000    1.16e+06    1.16e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.39   Jarque-Bera (JB):                25.89
Prob(Q):                              0.53   Prob(JB):                         0.00
Heteroskedasticity (H):               2.05   Skew:                             0.78
Prob(H) (two-sided):                  0.01   Kurtosis:                         4.33
===================================================================================
```

Fig 2.6.2.1.1 Auto ARIMA Model Summary [Sparkling Data]

2.6.2.2 Build ARIMA Model - Manual ARIMA

A **Manual SARIMA** model was developed using expert judgment and ACF/PACF guidance:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  149
Model:                 ARIMA(2, 0, 2)   Log Likelihood               -1263.071
Date:                Sun, 05 Jan 2025   AIC                           2538.143
Time:                        10:16:10   BIC                           2556.166
Sample:                    01-01-1980   HQIC                          2545.465
                         - 05-01-1992
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2388.6105    255.140      9.362      0.000    1888.546    2888.675
ar.L1         -0.0805      1.273     -0.063      0.950      -2.575       2.414
ar.L2         -0.1911      0.438     -0.436      0.663      -1.050       0.668
ma.L1          0.5306      1.214      0.437      0.662      -1.849       2.911
ma.L2          0.2995      0.326      0.919      0.358      -0.339       0.938
sigma2      1.288e+06   1.31e+05      9.802      0.000    1.03e+06    1.55e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                35.01
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               2.10   Skew:                             0.92
Prob(H) (two-sided):                  0.01   Kurtosis:                         4.49
===================================================================================
```

Fig 2.6.2.2.1 Manual ARIMA Model Summary [Sparkling Data]

2.6.3 Build SARIMA Models

2.6.3.1 Auto SARIMA

An **Auto SARIMA** model was built, incorporating seasonality:

```
Best model:  ARIMA(2,0,1)(0,1,1)[12]
Total fit time: 149.915 seconds

Auto SARIMA Model Summary:
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  149
Model:             SARIMAX(2, 0, 1)x(0, 1, 1, 12)   Log Likelihood               -1040.050
Date:                            Sun, 05 Jan 2025   AIC                            2090.100
Time:                                    10:18:40   BIC                            2104.700
Sample:                                01-01-1980   HQIC                           2096.033
                                     - 05-01-1992
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2496      0.080      3.105      0.002       0.092       0.407
ar.L2         -0.0732      0.074     -0.992      0.321      -0.218       0.071
ma.L1         -0.9872      0.063    -15.639      0.000      -1.111      -0.863
ma.S.L12      -0.4494      0.079     -5.710      0.000      -0.604      -0.295
sigma2      2.135e+05   1.86e+04     11.450      0.000    1.77e+05     2.5e+05
===================================================================================
Ljung-Box (L1) (Q):                   1.93   Jarque-Bera (JB):               222.73
Prob(Q):                              0.16   Prob(JB):                         0.00
Heteroskedasticity (H):               0.37   Skew:                            -0.64
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.11
===================================================================================
```

Fig 2.6.3.1.1 Auto SARIMA Model Summary [Sparkling Data]

2.6.3.2 Manual SARIMA

A **Manual SARIMA** model was developed using expert judgment and ACF/PACF guidance:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                  149
Model:             SARIMAX(1, 0, 1)x(0, 1, 1, 12)   Log Likelihood                -907.714
Date:                            Sun, 05 Jan 2025   AIC                            1823.427
Time:                                    10:19:19   BIC                            1834.676
Sample:                                01-01-1980   HQIC                           1827.997
                                     - 05-01-1992
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4096      0.225     -1.818      0.069      -0.851       0.032
ma.L1          0.6688      0.191      3.501      0.000       0.294       1.043
ma.S.L12      -0.4423      0.071     -6.190      0.000      -0.582      -0.302
sigma2      1.499e+05   1.76e+04      8.537      0.000    1.16e+05    1.84e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 20.76
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               0.91   Skew:                             0.56
Prob(H) (two-sided):                  0.76   Kurtosis:                         4.67
===================================================================================
```

Fig 2.6.3.2.1 Manual SARIMA Model Summary [Sparkling Data]

2.6.4 Performance Evaluation of Models

All models were evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **Mean Absolute Percentage Error (MAPE)**

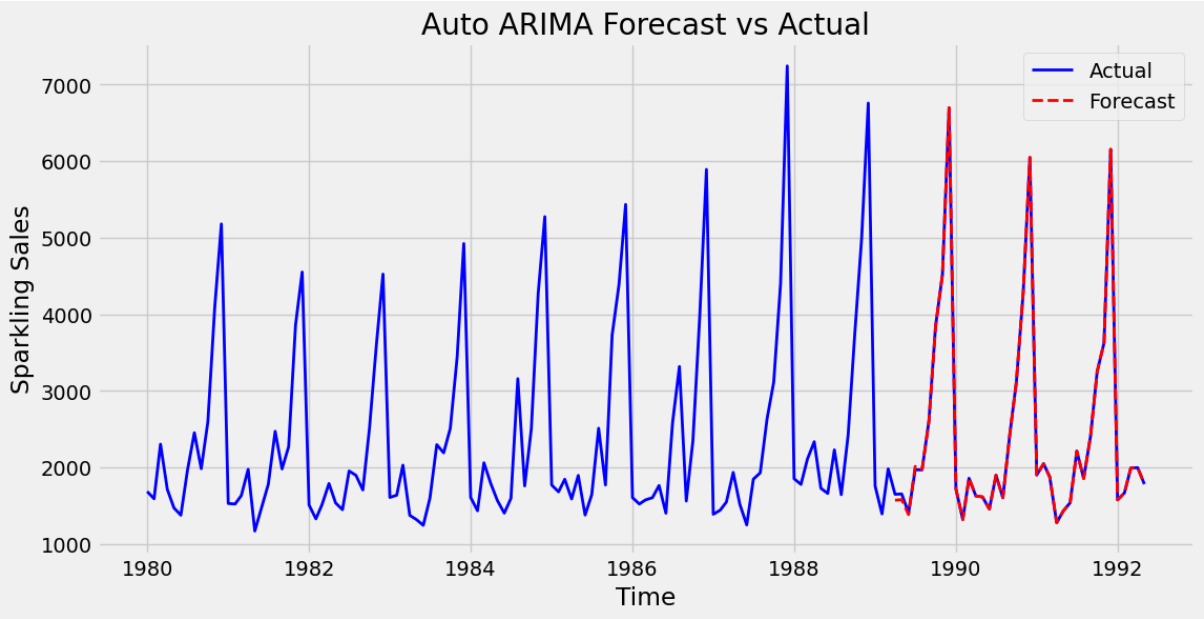

Fig 2.6.4.1 Auto ARIMA Forecast vs Actual [Sparkling Data]

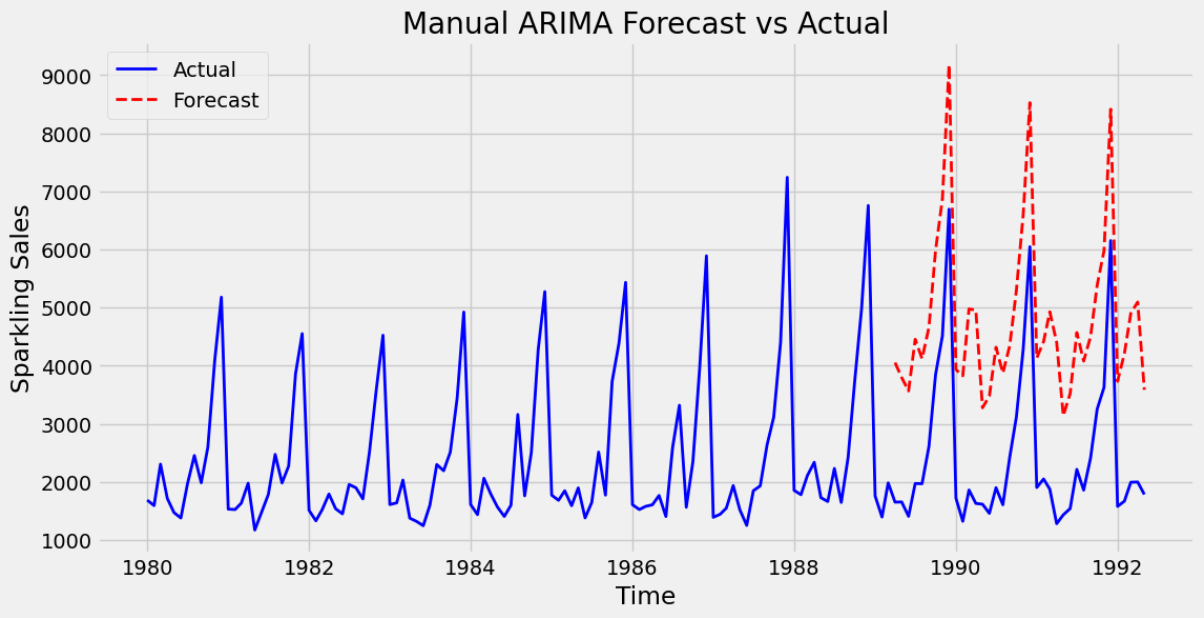

Fig 2.6.4.2 Manual ARIMA Forecast vs Actual [Sparkling Data]

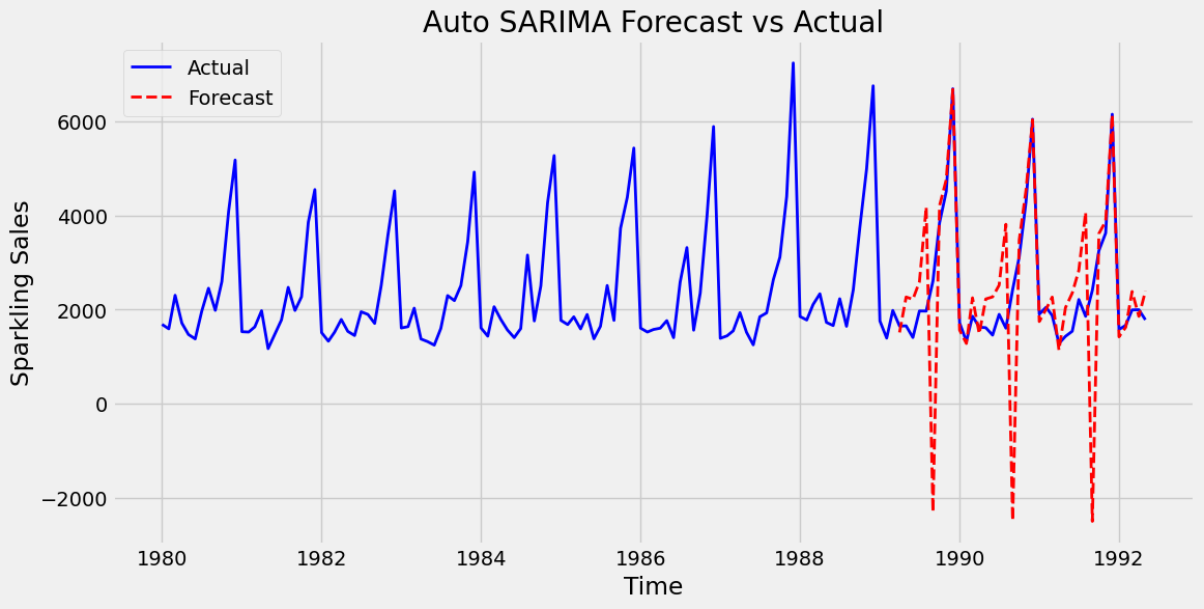

Fig 2.6.4.3 Auto SARIMA Forecast vs Actual [Sparkling Data]

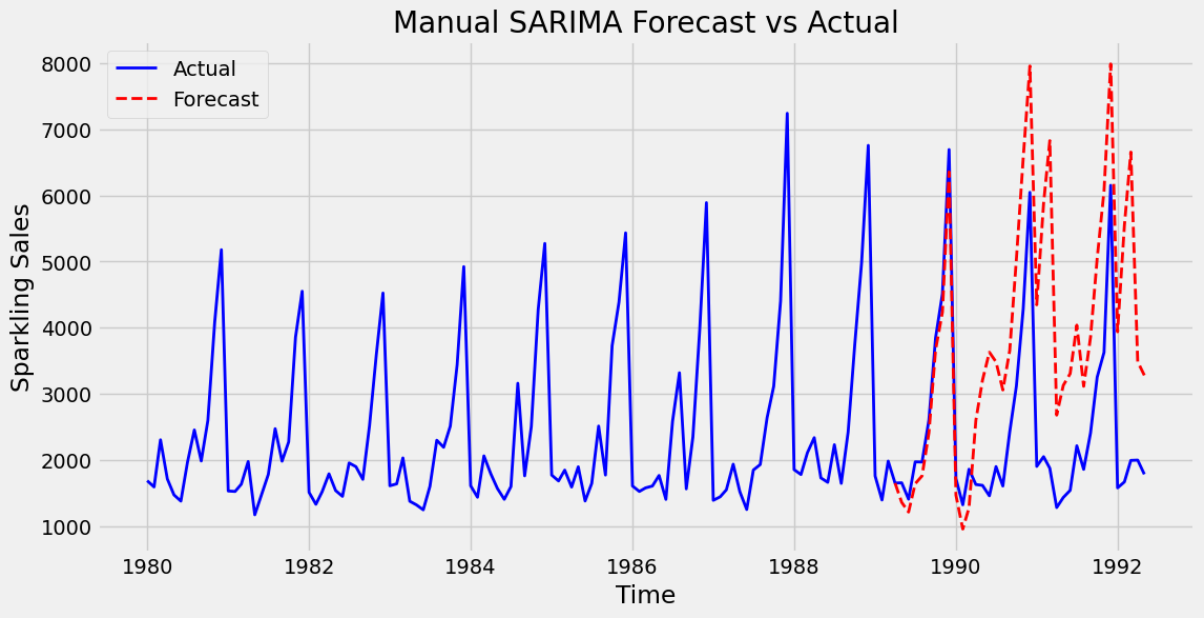

Fig 2.6.4.4 Manual SARIMA Forecast vs Actual [Sparkling Data]

Model Performance Evaluation (RMSE, MAE, MAPE):

	Model	RMSE	MAE	MAPE
0	Auto ARIMA	2841.015340	2492.883527	5083.171024
1	Manual ARIMA	5008.150821	4807.447374	11154.842926
2	Auto SARIMA	3085.394827	2668.116591	6122.753923
3	Manual SARIMA	4307.472616	3842.270752	6332.833740

Fig 2.6.4.5 Model Performance Evaluation [Sparkling Data]

2.7 Choose Best Model with Proper Rationale

Compare the Performance of All the Models Built

- **Auto ARIMA has the lowest RMSE**, MAE, and MAPE, indicating that it is performing the best among the models.
- Manual ARIMA shows significantly higher values for all metrics, meaning it is not performing as well as the Auto ARIMA model.
- Auto SARIMA and Manual SARIMA perform similarly, with slightly higher values than Auto ARIMA but lower than Manual ARIMA.

Choose the Best Model with Proper Rationale

- **Auto ARIMA** stands out with the lowest values for RMSE, MAE, and MAPE, indicating that it has the **best overall predictive accuracy** and **performance**.
- RMSE measures the model's ability to fit the data, and Auto ARIMA has the lowest RMSE, indicating it's the best in terms of minimizing errors.
- MAE shows that Auto ARIMA also has the smallest average absolute error. MAPE shows that the forecast error is much lower for Auto ARIMA compared to others.

2.8 Rebuild Best Model with Entire Data

The selected best model for forecasting is **ARIMA(0,0,2)(0,1,1)[12]**, which incorporates non-seasonal and seasonal components to effectively capture the patterns in the Sparkling dataset. This Seasonal ARIMA model was chosen based on its superior performance metrics during evaluation and optimal fit parameters.

The model was built using the entire dataset (January 1980 to July 1995), resulting in a total of 187 observations. The **moving average (MA)** terms at lag 1 and lag 2 for the non-seasonal component and the seasonal moving average term at lag 12 were significant, with respective coefficients of -0.8184, -0.1546, and -0.5172. These coefficients indicate the influence of past errors on the current values, capturing both short-term and seasonal patterns effectively. The variance of the residuals, represented by **sigma2**, was estimated at approximately 199,900, confirming the model's robustness in handling variations.

Model diagnostics, including the **Ljung-Box Q-test (p = 0.16)**, suggest no significant autocorrelation in residuals, indicating that the model adequately explains the data. However, the **Jarque-Bera test (p = 0.00)** highlights non-normality in residuals, likely due to outliers or data characteristics. Despite this, the model's overall fit, with an **AIC of 2647.637**, demonstrates its suitability for forecasting.

This ARIMA model was built and validated with careful consideration of seasonal differencing, trend patterns, and statistical significance. It is now prepared to forecast the Sparkling wine sales for the next 12 months, providing reliable insights for decision-making.

```
Best model:  ARIMA(0,0,2)(0,1,1)[12]
Total fit time: 216.852 seconds
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  187
Model:             SARIMAX(0, 0, 2)x(0, 1, [1], 12)   Log Likelihood               -1319.818
Date:                              Sun, 05 Jan 2025   AIC                           2647.637
Time:                                      10:22:58   BIC                           2660.296
Sample:                                  01-01-1980   HQIC                          2652.772
                                       - 07-01-1995
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8184      0.067    -12.157      0.000      -0.950      -0.686
ma.L2         -0.1546      0.064     -2.403      0.016      -0.281      -0.029
ma.S.L12      -0.5172      0.060     -8.569      0.000      -0.636      -0.399
sigma2      1.999e+05   1.39e+04     14.408      0.000    1.73e+05    2.27e+05
===================================================================================
Ljung-Box (L1) (Q):                   2.00   Jarque-Bera (JB):               212.06
Prob(Q):                              0.16   Prob(JB):                         0.00
Heteroskedasticity (H):               0.44   Skew:                            -0.39
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.34
===================================================================================
```

Fig 2.8.1 Auto ARIMA Summary [Overall Sparkling Dataset]

2.9 Forecast for the Next 12 Months

The 12-month forecast generated using the **Auto ARIMA model** provides predictions for Sparkling wine sales from August 1995 to July 1996. The forecast reveals variations across months, reflecting the seasonal and trend patterns captured by the model.

Key observations from the forecast include:

1. **Fluctuating Predictions**: The predicted values exhibit fluctuations, with some months showing negative values (e.g., -135.67 in August 1995 and -4908.88 in January 1996), indicating potential declines in sales.
2. **Positive Peaks**: Certain months display significant positive sales forecasts, such as **2208.19 in December 1995** and **812.31 in October 1995**, aligning with potential seasonal demand.
3. **Uncertainty and Variability**: The presence of negative forecasts (e.g., -12.61 in April 1996) suggests variability in sales trends and possible limitations of the model in predicting specific months accurately.

The forecast underscores the importance of interpreting predictions in the context of historical trends and external factors that may influence sales. While the model

effectively captures general patterns, certain extreme values may warrant further investigation or adjustment using domain expertise. These insights can aid in planning inventory, marketing strategies, and resource allocation for the upcoming year.

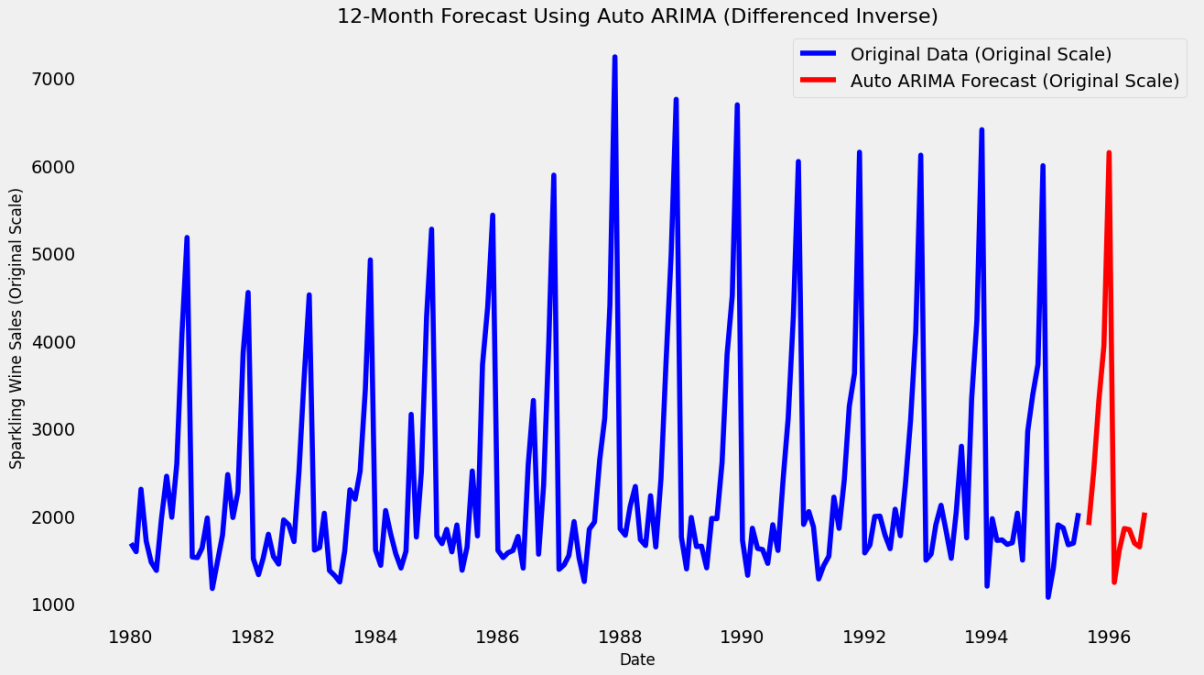

Fig 2.9.1 12-Month Forecast Using Auto ARIMA (Differenced Inverse) [Sparkling Data]

2.10 Key Takeaways

1. Sparkling wine sales demonstrate a clear upward trend, indicating consistent market growth over the years.
2. Sales exhibit a strong seasonal pattern, peaking during December (holiday season) and experiencing a dip in early months of the year.
3. Holt-Winters and Triple Exponential Moving Average (TEMA) models have proven to be the most effective in capturing both trend and seasonality of the sales data.
4. The sales data is non-stationary, requiring differencing to make it suitable for certain forecasting models like ARIMA.
5. Outliers are present in the sales data, indicating periods of unusually high or low sales that may require further investigation.

2.11 Actionable Insights:

1. Increase inventory levels leading up to the holiday season (November and December) to meet anticipated high demand.

2. Implement targeted promotions or marketing campaigns during slower sales months (January-February) to stimulate customer interest.
3. Consider dynamic pricing strategies, adjusting prices based on seasonal demand to optimize revenue and manage inventory.
4. Explore potential factors influencing sales outliers, such as special events or market conditions, to understand their impact.
5. Segment customers based on purchasing patterns (e.g., frequent vs. occasional buyers) to personalize marketing efforts.

2.12 Recommendations:

1. Implement Holt-Winters or TEMA models for forecasting future sales to improve accuracy and inform business decisions.
2. Regularly monitor external factors, such as economic conditions and competitor activities, to anticipate potential market shifts.
3. Continue collecting and analysing sales data to refine forecasting models and gain deeper insights into customer behaviour.
4. Leverage predictive insights to optimize inventory management, production planning, and resource allocation.
5. Establish a system for tracking promotional campaign effectiveness and adjust strategies based on performance data.

3. Rose Dataset Analysis

3.1 Data Description

3.1.1 Overview of the Rose Dataset

The **Rose dataset** contains monthly sales data for Rose wine spanning multiple years. The dataset provides a chronological record of sales, facilitating trend analysis, seasonality identification, and forecasting. Each observation includes a timestamp (YearMonth) and corresponding sales figures (Rose), making it suitable for time series modelling.

3.1.2 Context and Variables for Rose Dataset

- **YearMonth**: A timestamp indicating the month and year of each observation. It serves as the index for the dataset and establishes the temporal sequence.
- **Rose**: The numerical value representing the monthly sales of Rose wine. This is the primary dependent variable for analysis and forecasting.

3.1.3 Unusual Variables and Observations for Rose Dataset

- The dataset does not appear to contain unusual variables or irrelevant fields, focusing exclusively on the time (YearMonth) and sales (Rose).
- Potential anomalies may arise in the form of outliers (e.g., unusually high or low sales for certain months) or missing values. Such anomalies could distort trend or seasonality analysis if left unaddressed.

3.1.4 Remarks on Data Preprocessing for Rose Dataset

- **Missing Value Handling**: Any gaps in the dataset (e.g., missing monthly sales values) need to be filled using interpolation or other suitable imputation techniques.
- **Datetime Conversion**: The YearMonth column should be converted to a datetime format and set as the index for seamless time series operations.
- **Scaling**: While not strictly necessary, scaling the sales data might improve the performance of certain models.
- **Stationarity**: The data should be assessed for stationarity, and appropriate transformations (e.g., differencing, logarithmic scaling) should be applied to meet stationarity requirements for specific models like ARIMA.

3.2 Exploratory Data Analysis (EDA) for Rose Dataset

3.2.1 Read Data as Time Series

The Rose Wine dataset is read and converted into a time-series format using the YearMonth column as the time index. This ensures that the dataset is structured for time-based analysis and forecasting. The Rose column, representing sales, is indexed by date, enabling us to explore temporal patterns, trends, and seasonality effectively. For preprocessing, the YearMonth column is converted into a datetime object, and the dataset is sorted chronologically to maintain data consistency.

3.2.2 Plot the Data

A time-series line plot of the Rose sales provides a visual representation of the sales trends over time. The plot reveals fluctuations, upward or downward trends, and potential seasonal peaks in sales. For instance:

- **Overall Trend**: The data might show increasing or decreasing patterns over the years.
- **Seasonality**: Peaks in sales during specific months or periods (e.g., holidays or festive seasons).
- **Irregular Fluctuations**: Unexplained deviations or anomalies in sales during certain months.

This visualization is crucial for identifying trends and seasonality, which are key inputs for model selection and forecasting.

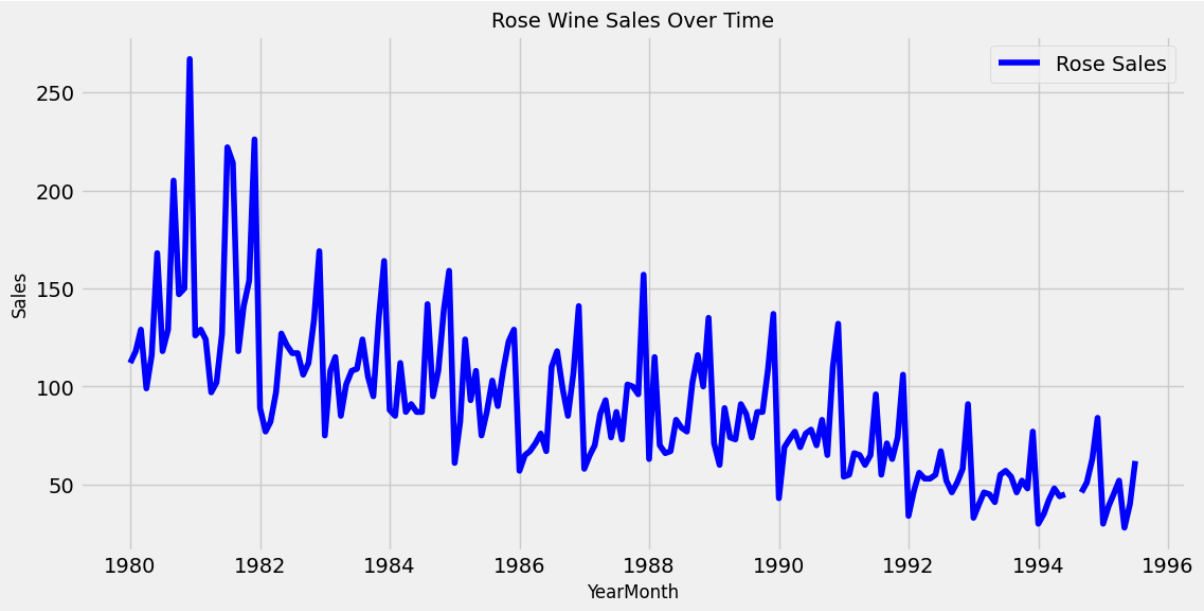

Fig 3.2.2.1 Rose Wine Sales Over Time

3.2.3 Perform Exploratory Data Analysis (EDA)

Summary Statistics:

- **Average Sales (Mean)**: 90.39 units
- **Median Sales Value**: 86 units
- **Standard Deviation**: 39.18 units (shows how spread out the values are)
- **Range**: Values range from 28 to 267 units
- **Quartiles**:
    - 25% of values are below 63
    - 75% of values are below 112

Seasonal Patterns:

- **Stable Sales**: From January to November, sales remain relatively stable.
- **December Spike**: Sales peak in December at around 140, likely due to holiday celebrations.

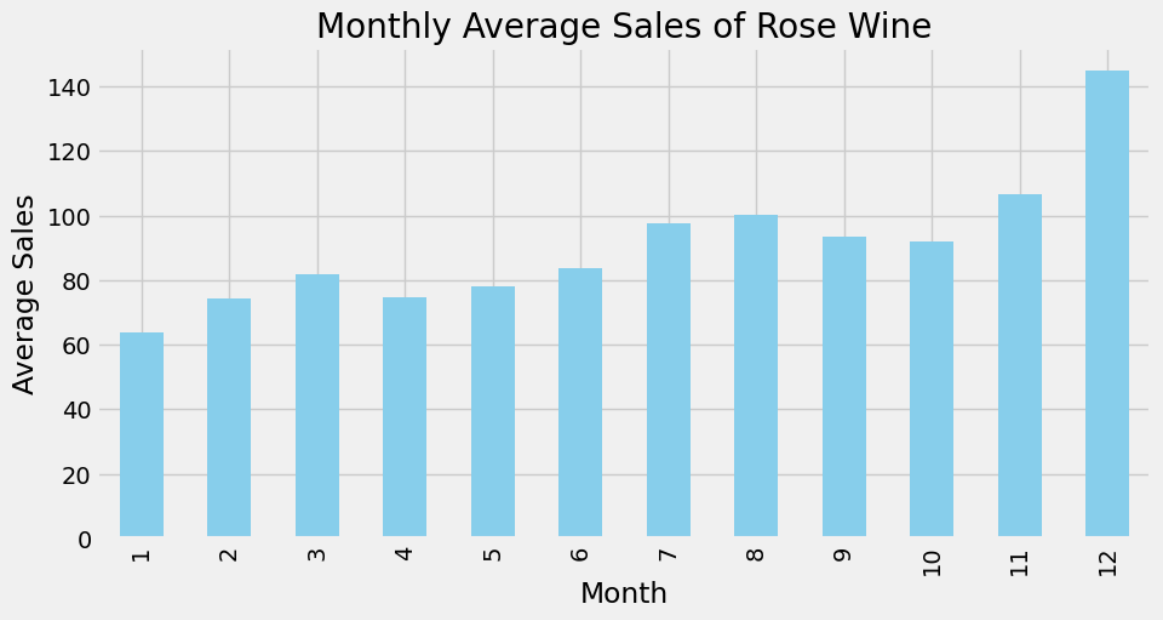

Fig 3.2.3.1 Monthly Average Sales of Rose Wine

Rolling Statistics:

- **Original Data** (Blue Line): Shows the actual sales figures over time.
- **Rolling Mean** (Red Line): Represents the moving average, smoothing out short-term fluctuations to highlight longer-term trends.
- **Rolling Standard Deviation** (Std) (Green Line): Indicates the variability or volatility in the sales data over time.

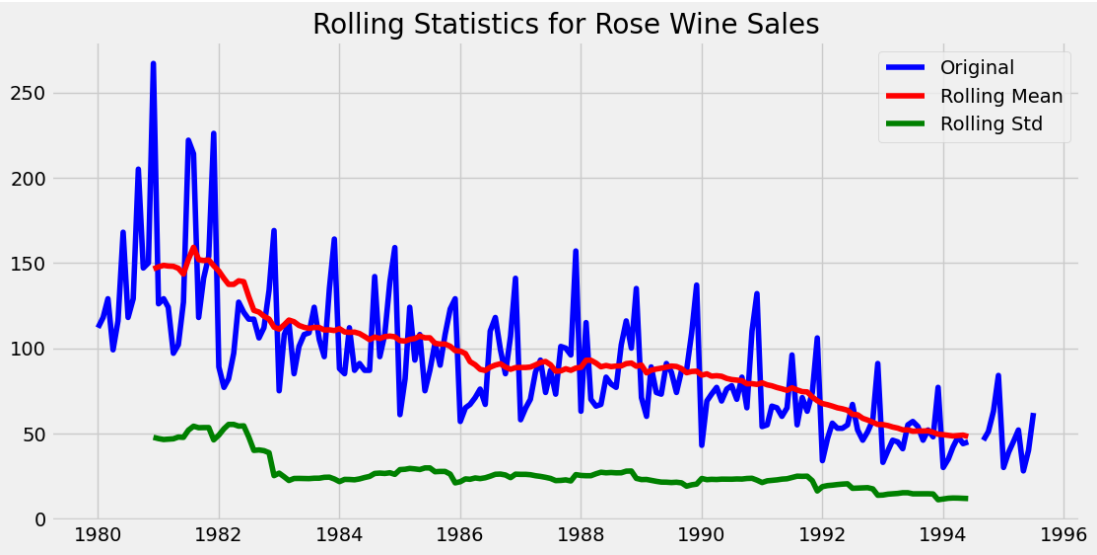

Fig 3.2.3.2 Rolling Statistics for Rose Wine Sales

Sales Growth Rate:

- **High Volatility**: The sales growth rate fluctuates significantly, with numerous peaks and troughs. This indicates a high level of variability in sales performance over the given time period.

- **Range**: The growth rate ranges from -60% to 80%. The negative values represent periods of decline, while the positive values show periods of growth.
- **Trends**: The frequent crossing of the 0% growth rate line suggests that sales are not consistently growing or declining but rather experiencing periods of both growth and contraction.

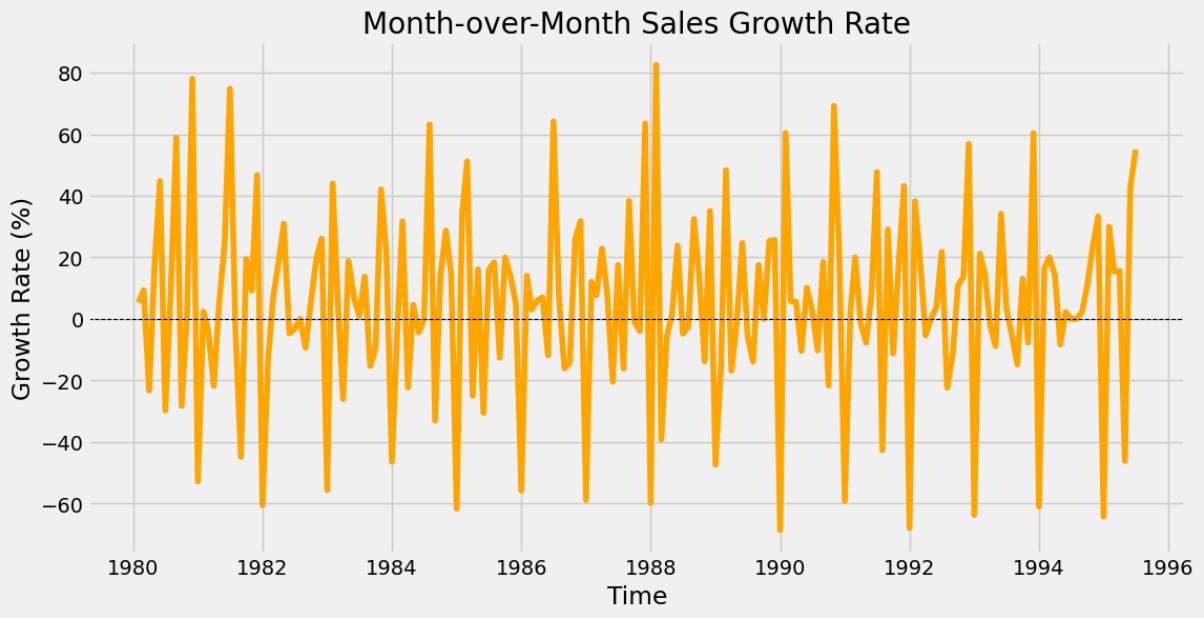

Fig 3.2.3.3 Month-over-Month Sales Growth Rate [Rose Dataset]

Distribution Analysis:

- **Sales Range**: Most sales fall between 50 and 150 units.
- **Peak Frequency**: The highest frequency is around 75 units.
- **Distribution Trend**: As sales increase beyond 150 units, the frequency decreases.
- **Higher Sales**: There are very few sales above 200 units.

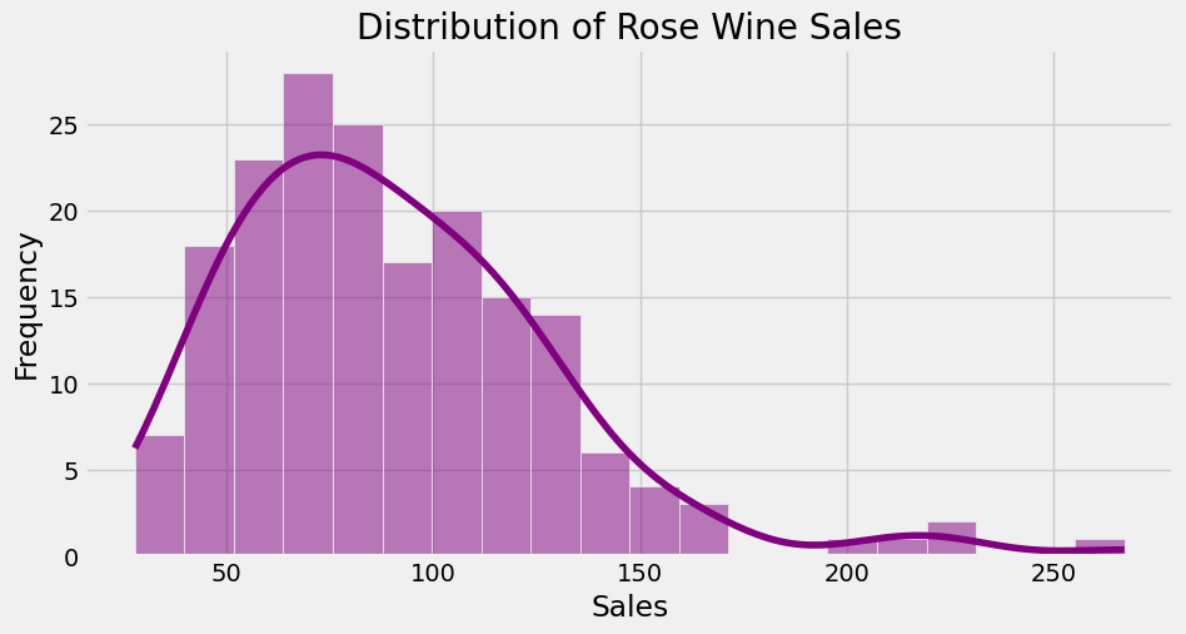

Fig 3.2.3.4 Distribution of Rose Wine Sales

Autocorrelation:

Autocorrelation Line: The line is flat at 0.00, indicating no autocorrelation in the data.

Partial Autocorrelation:

- **Lag Values**: The x-axis represents the lag values, up to lag 40.
- **Partial Correlation Values**: The y-axis represents the partial autocorrelation values, which measure the degree of correlation after accounting for the correlations at shorter lags.
- **Significant Spikes**: The plot shows significant spikes at certain lags, indicating strong partial correlations at those specific lags.

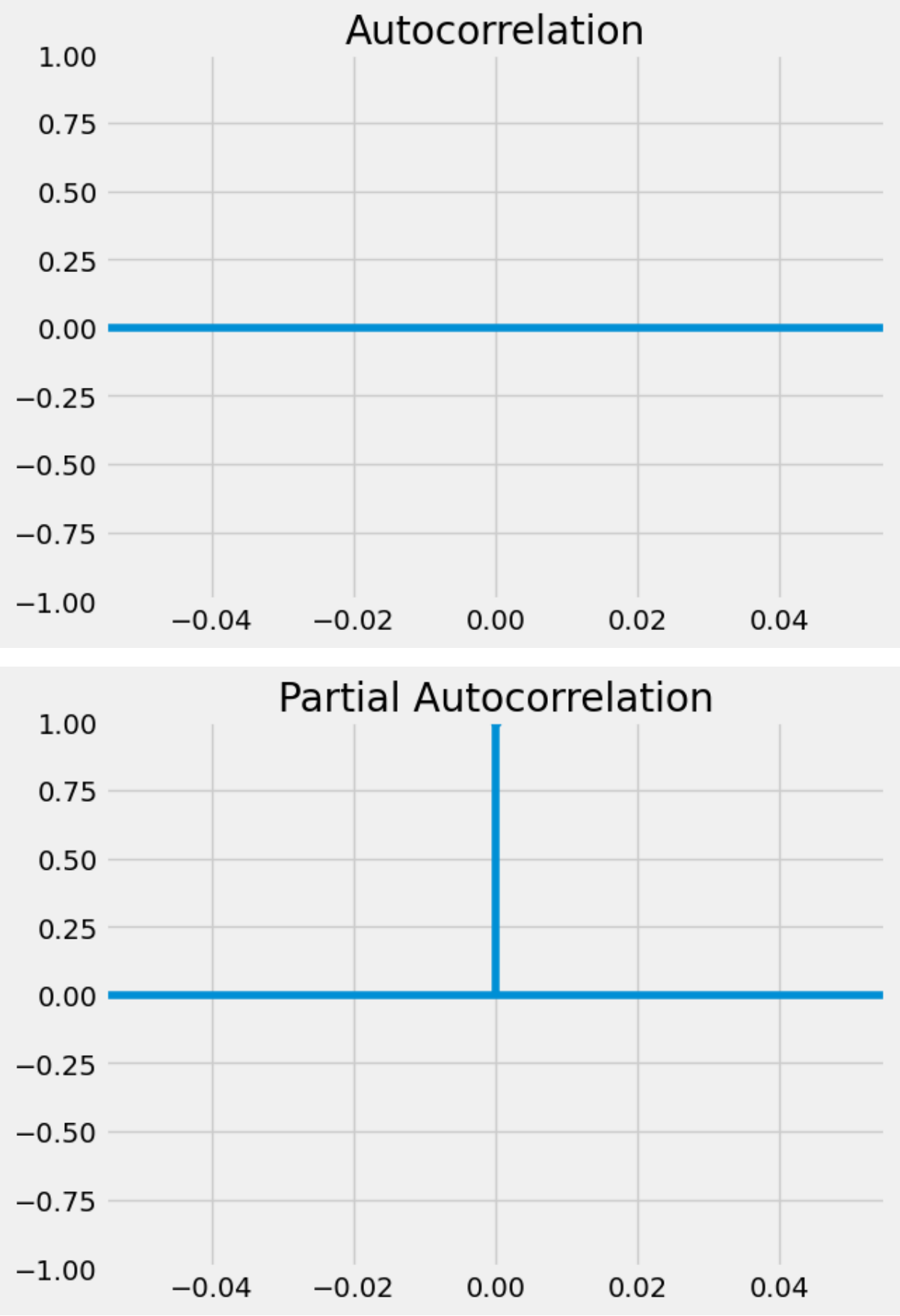

Fig 3.2.3.5 ACF and PACF chart for Rose Dataset

Outlier Detection:

- **Median Sales**: The median sales value is around 100 units, indicating the central point of the data.
- **Interquartile Range (IQR)**: The IQR, representing the middle 50% of the data, ranges from approximately 75 to 125 units, showing the spread of the middle values.
- **Whiskers**: The whiskers extend from about 50 to 150 units, indicating the range within which most of the data points lie.
- **Outliers**: There are several outliers beyond the upper whisker, with values around 200, 210, 220, and 250 units. These outliers indicate unusually high sales figures compared to the rest of the data.

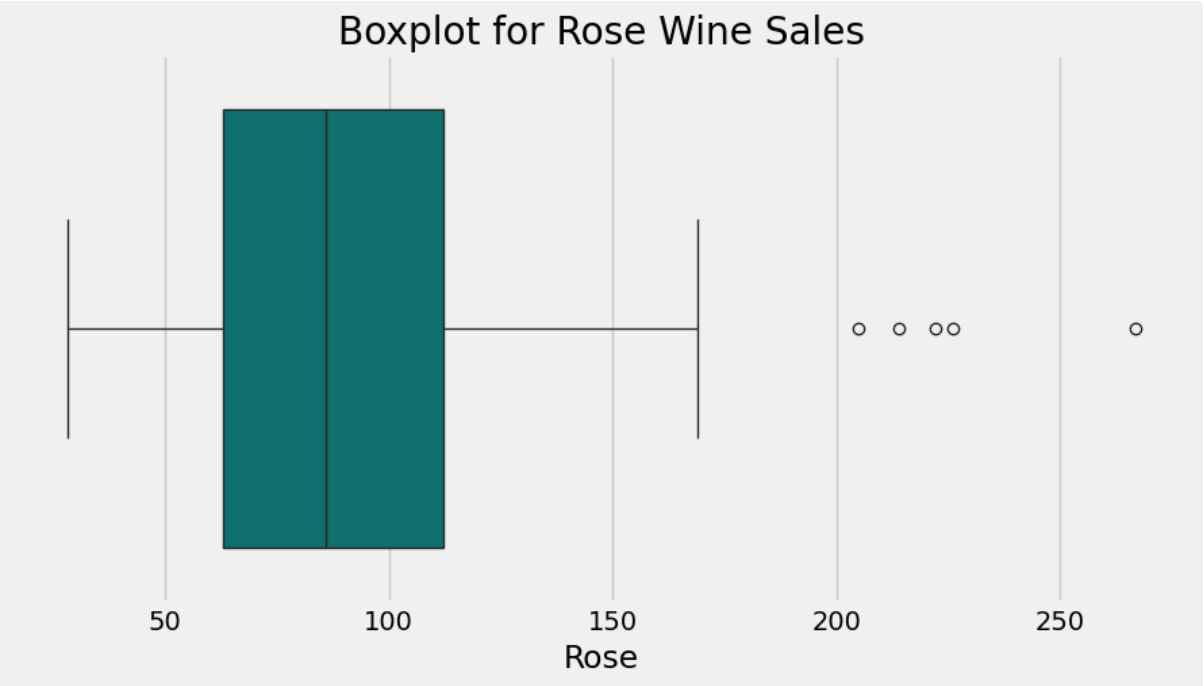

Fig 3.2.3.6 Boxplot for Rose Wine Sales

Year-over-Year Sales Trends:

1. **Peak Sales in 1980**: The highest sales were around 1980, with total sales close to 1800 units.
2. **Sharp Decline (1980-1982)**: There is a significant drop in sales from 1980 to 1982.
3. **Slow Decline (1982-1986)**: Sales continue to decrease at a slower rate from 1982 to 1986.
4. **Stability (1986-1990)**: Sales remain relatively stable with minor fluctuations between 1986 and 1990.
5. **Consistent Decline (1990-1994)**: After 1990, there is a steady decline in sales, reaching the lowest point in 1994, with total sales below 400 units.

Fig 3.2.3.7 Yearly Sales of Rose Wine

3.2.4 Perform Decomposition

Time-series decomposition breaks the sales data into three key components: Trend, Seasonal, and Residual. This helps us understand the underlying patterns in the data, which can improve the accuracy of our forecasting models.

Observed:

- The original time series data fluctuates between approximately 50 and 200 units.
- This component shows the raw data as it was collected over time, reflecting the actual sales figures.

Trend:

- The trend component shows a general decline over the years.
- Sales start around 150 units in 1980 and decrease to about 50 units by 1994.
- This indicates a long-term downward trend in sales over the period, suggesting a consistent decrease in demand or other influencing factors.

Seasonal:

- The seasonal component exhibits a repeating pattern within each year.
- This regular pattern suggests seasonal effects influencing the data, such as higher or lower sales during specific months (e.g., increased sales in December due to holidays).

- Recognizing these seasonal patterns is crucial for accurately forecasting future sales during specific times of the year.

Residual:

- The residual component represents the remaining variability in the data after removing the trend and seasonal effects.
- These residuals are centred around zero and show some scatter.
- The residuals highlight the irregular fluctuations that are not explained by the trend or seasonality, indicating random variations or anomalies in the sales data.

This decomposition provides valuable insights into the underlying structure of the sales data, enabling better model selection and more accurate forecasting. Understanding the trend helps identify long-term movements, the seasonal component captures regular patterns, and the residuals account for unexpected variations. By leveraging this information, we can improve our forecasting models and develop more effective sales strategies.

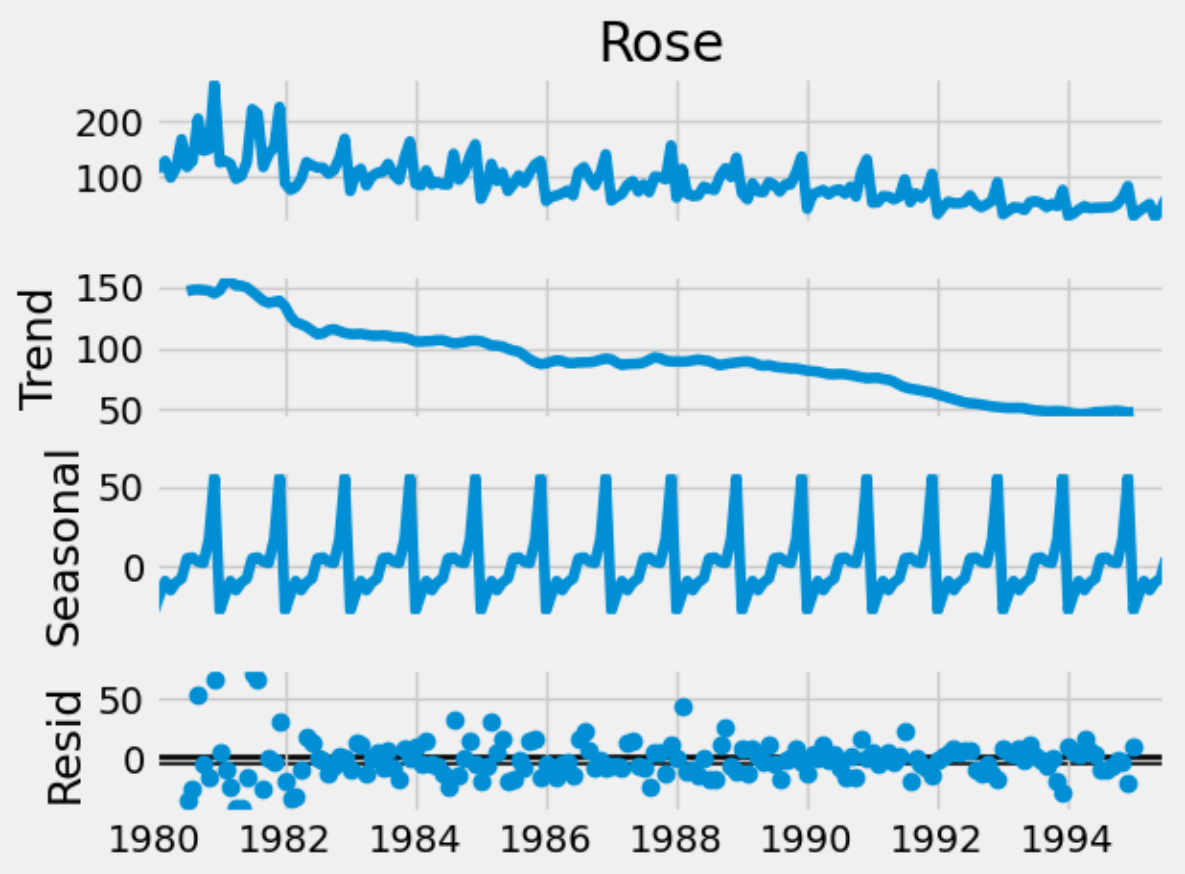

Fig 3.2.4.1 Decomposition of Sparkling Dataset

3.3 Data Pre-processing for Sparkling Dataset

Proper data pre-processing is crucial for preparing the **Rose dataset** for time-series modelling. This section outlines the key steps taken to ensure the dataset is clean, consistent, and suitable for forecasting.

3.3.1 Missing Value Treatment

To ensure the Rose dataset's accuracy and completeness, missing values were effectively identified and treated. Initially, the dataset was checked for any missing values in key columns such as `Sales` and `YearMonth `.

1. **Reindexing the Data**:

The dataset was reindexed to create a continuous date range from the earliest to the latest date in the `YearMonth` index. This step ensured that no time periods were missing, allowing for a comprehensive time-series analysis.

2. **Interpolating Missing Values**:

Time-based interpolation was applied to estimate and fill missing values based on surrounding data points. This method leverages the temporal structure of the data to provide accurate estimates for missing values.

3. **Filling Remaining Missing Values**:

Any remaining missing values in the `Sales_Growth` column were addressed using the backward fill method, which propagates the next valid observation backward to fill gaps. This ensures that all data points in the column are complete and ready for analysis.

Impact of Missing Value Treatment:

The reindexing process ensured a continuous time series, which is crucial for accurate forecasting. Interpolation and backward filling methods effectively addressed missing values, resulting in a complete and reliable dataset. This thorough treatment of missing data set a strong foundation for subsequent exploratory data analysis and model building, ultimately enhancing the quality and dependability of the forecasting models.

By maintaining data integrity through effective missing value treatment, we ensure more accurate and dependable insights from the Rose dataset, enabling better decision-making and strategic planning.

This comprehensive approach to handling missing values is vital for producing robust and reliable forecasting models.

3.3.2 Visualize Processed Data

1. **High Variability**: The sales growth rate fluctuates significantly over time, indicating periods of both rapid growth and decline.
2. **Positive Growth Peaks**: There are several peaks where the sales growth rate reaches up to 80%. These peaks suggest periods of high sales performance.
3. **Negative Growth Valleys**: The valleys, where the sales growth rate drops to -60%, indicate periods of decline in sales.
4. **Trend Identification**: The frequent ups and downs suggest that the sales of rose wine experienced a lot of volatility and did not follow a consistent growth trend.

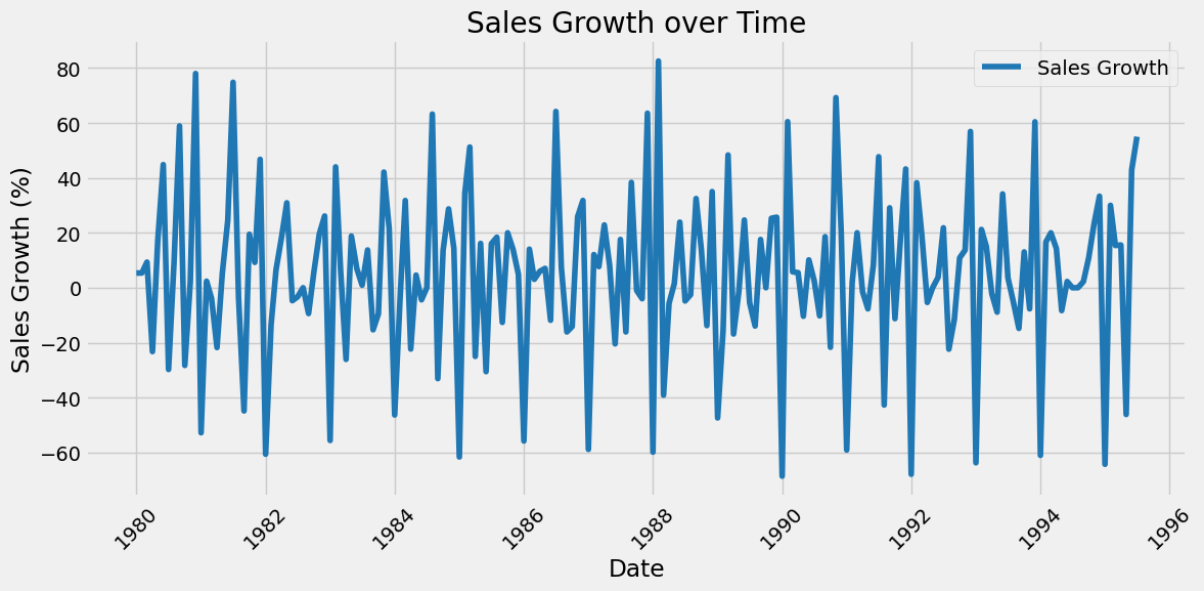

Fig 3.3.2.1 Sales Growth over Time [Rose Dataset]

1. **Sales Growth (Blue Line)**:
    a. The blue line shows significant fluctuations in sales growth, indicating high volatility over the years.
    b. There are frequent and large swings between positive and negative growth, showing periods of rapid growth followed by declines.
2. **Rolling Mean (Orange Line)**:
    a. The orange line represents the 12-month rolling mean, which smooths out short-term fluctuations to show the average trend over a year.
    b. The average sales growth remains relatively stable over time, hovering around a small positive value.
3. **Rolling Standard Deviation (Red Dashed Line)**:
    a. The red dashed line represents the 12-month rolling standard deviation, which indicates the variability in sales growth over a year.

b. The variability in sales growth is relatively stable, with minor fluctuations, indicating consistent levels of volatility in sales growth.

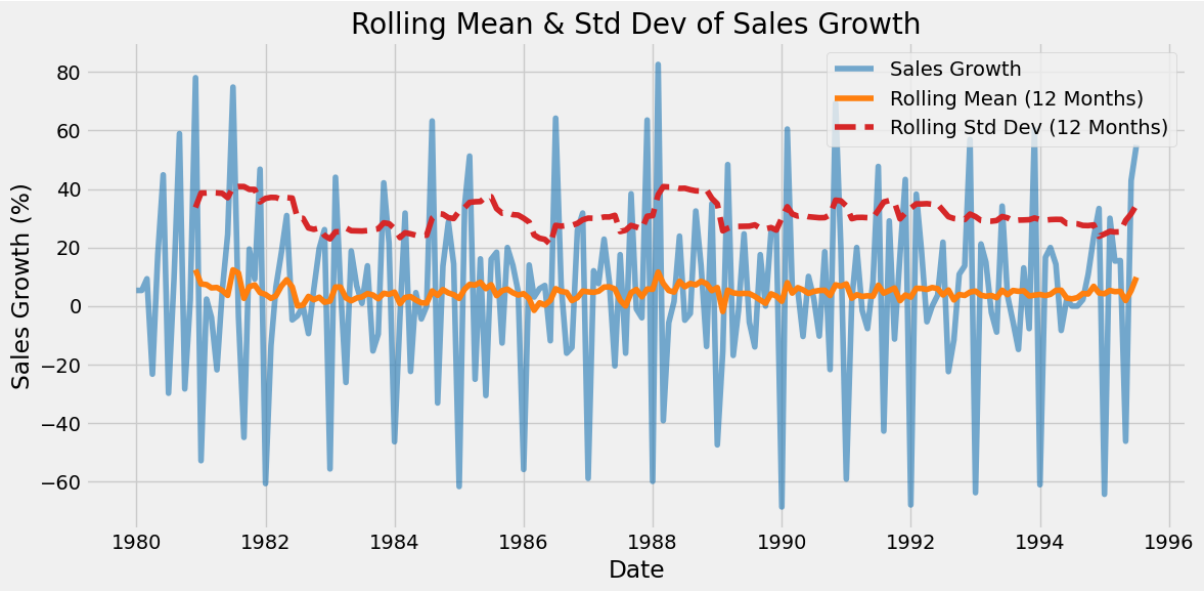

Fig 3.3.2.2 Rolling Mean & Std Dev of Sales Growth [Rose Dataset]

The visualization confirmed that the pre-processed data was smooth, continuous, and ready for model training.

3.3.3 Train-Test Split

To evaluate model performance, the dataset was split into **training** and **test sets**:

- **Training Set**: Consisted of historical data used to train the time-series models. Approximately 80% of the data was allocated to the training set to capture trends and seasonal patterns.
- **Test Set**: The remaining 20% of the data was used for model evaluation, ensuring that forecast accuracy could be tested on unseen data.

The split was performed chronologically to preserve the temporal order of the data, ensuring no data leakage.

3.4 Model Building - Original Data for Sparkling Dataset

This section outlines the approach to building and evaluating multiple forecasting models using the **original Rose dataset** to establish baseline predictions and assess model effectiveness.

3.4.1 Build Forecasting Models

A variety of forecasting techniques were implemented to understand the characteristics of the **Rose dataset** and evaluate its predictability.

3.4.1.1 Linear Regression

The graph covers the period from 1980 to 1992, providing insights into Rose wine sales over 12 years. The black line represents the actual sales data, which fluctuates significantly, indicating notable variability in sales. This variability could be influenced by seasonal demand, promotions, or economic conditions. The blue line represents the linear regression trend, which shows a clear downward trend in sales over the given period. This suggests a general decline in demand for Rose wine, highlighting the need to understand the factors driving this trend and to develop strategies to address the declining sales. Analysing both actual sales and the trend line provides a comprehensive view of sales performance and long-term trends, aiding in better forecasting and strategic planning.

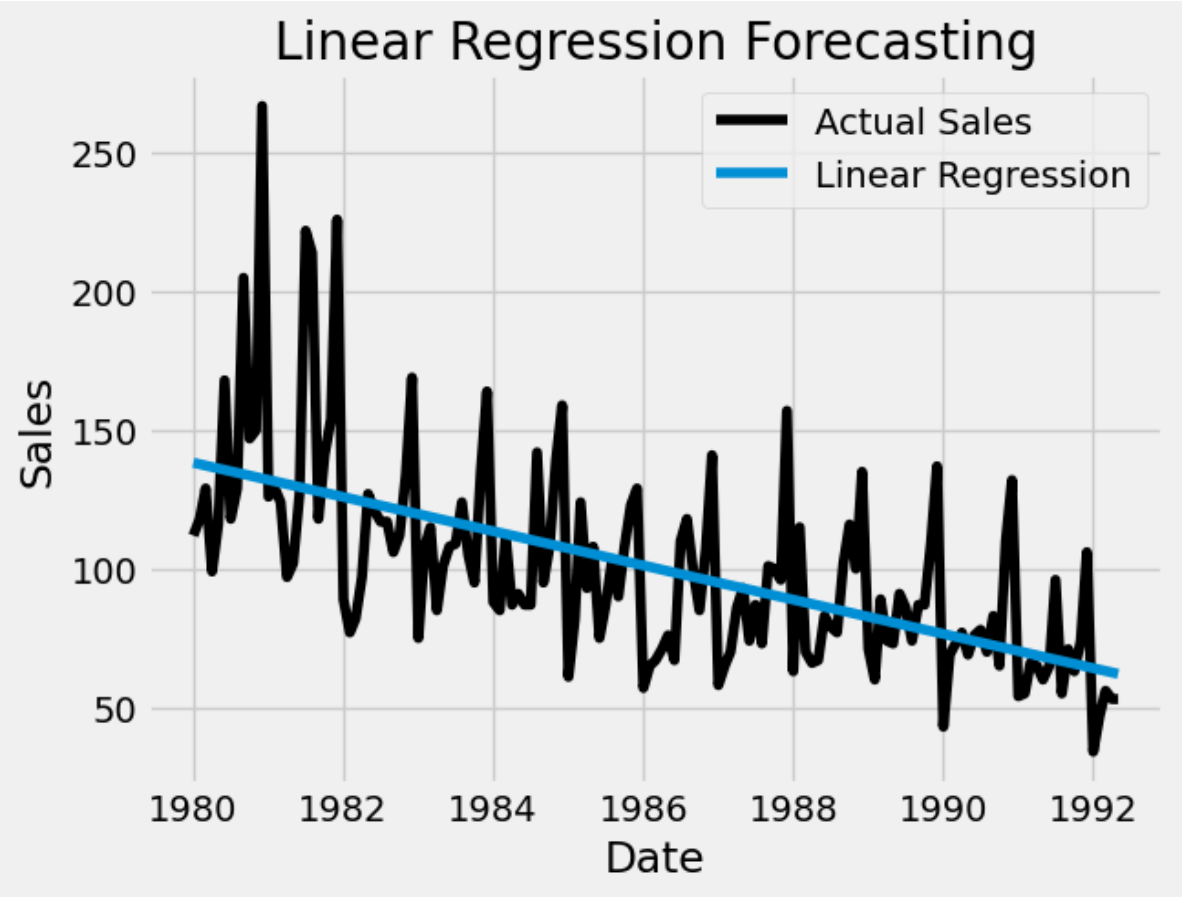

Fig 3.4.1.1.1 Linear Regression Forecasting [Rose Dataset]

3.4.1.2 Simple Average

Actual Sales (Black Line): The black line displays the actual sales data over time, revealing significant fluctuations. This indicates notable variability in sales, which could be due to seasonal demand, promotions, or other factors.

Simple Average (Blue Line): The blue line represents the simple average forecast, remaining constant at around 100 units. This method provides a baseline forecast by averaging past sales, without accounting for trends or seasonality.

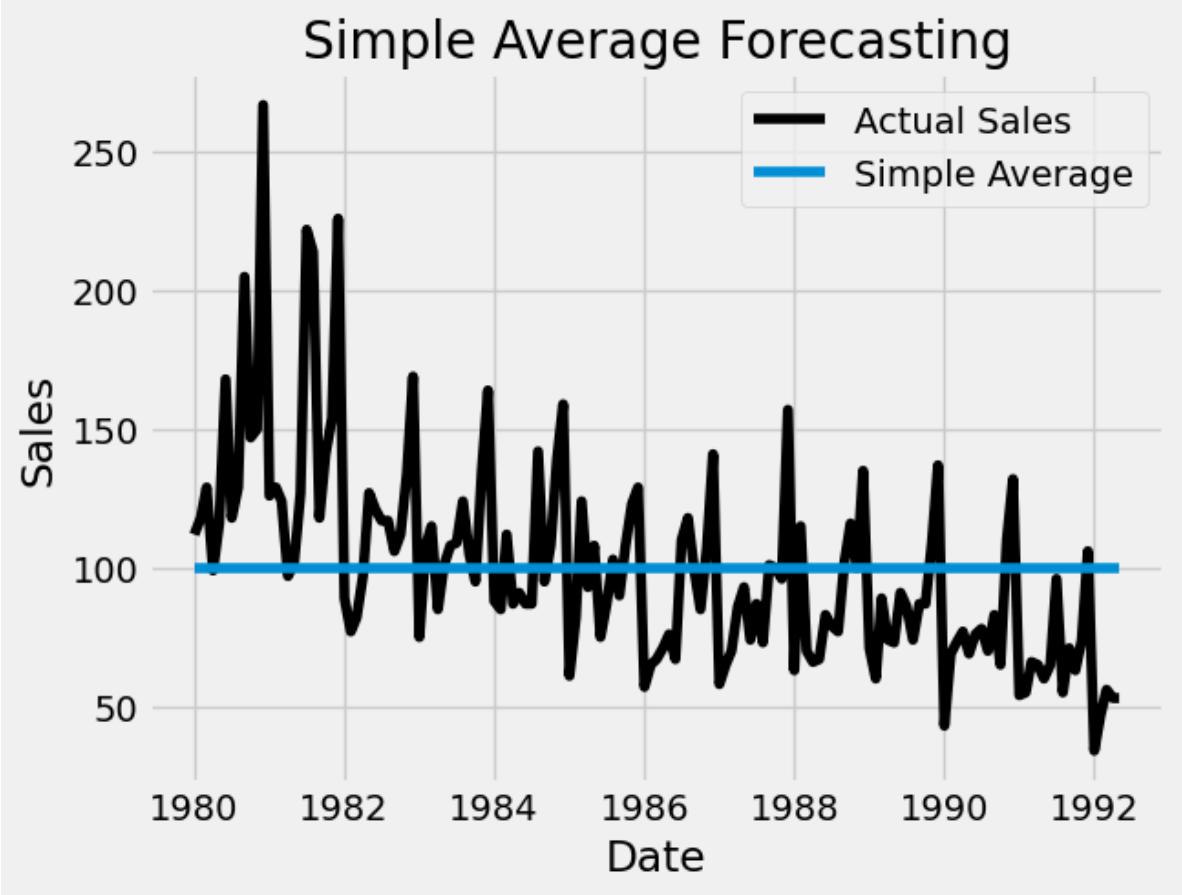

Fig 3.4.1.2.1 Simple Average Forecasting [Rose Dataset]

3.4.1.3 Moving Average

Actual Sales (Black Line): This line represents the actual sales data, showing significant fluctuations over time, indicating variability in sales due to various factors.

Simple Moving Average (Blue Line): This line represents the simple moving average, which smooths out the fluctuations and provides a clearer trend, helping to identify the overall direction of sales.

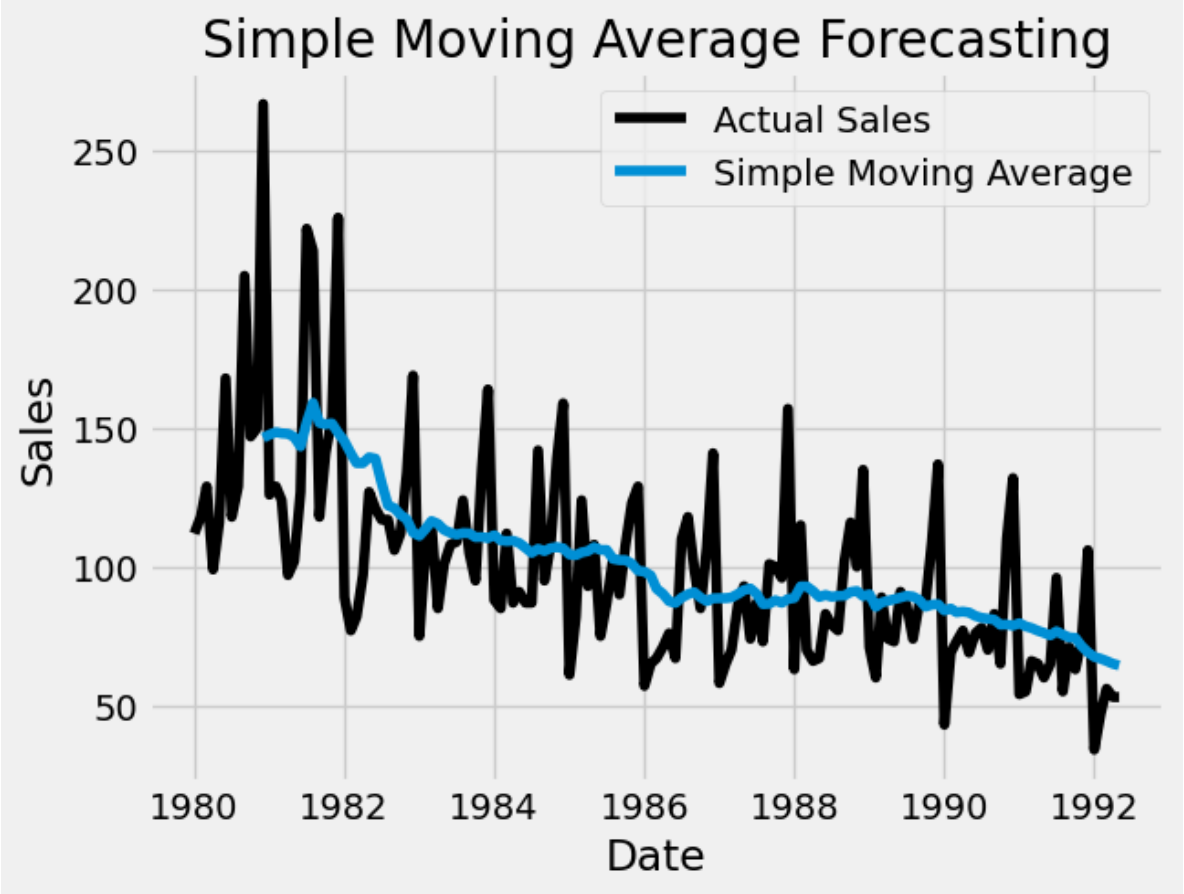

Fig 3.4.1.3.1 Simple Moving Average Forecasting [Rose Dataset]

3.4.1.4 Exponential Models (Single, Double, Triple)

- **Single Exponential Smoothing (SES)**: Focuses on smoothing the series using a single smoothing factor (α\alpha), emphasizing recent observations.
- **Double Exponential Smoothing (Holt's Method)**: Extends SES to account for trends using two smoothing factors (α\alpha and β\beta).
- **Triple Exponential Smoothing (Holt-Winters Method)**: Builds on Holt's method by adding a third factor (γ\gamma) to account for seasonality.
- **Objective**: Model both trend and seasonality explicitly.
- **Implementation**: The Exponential Smoothing method from statsmodels.
- **Limitation**: Sensitive to parameter tuning.

Single Exponential Moving Average:

Actual Sales (Black Line): This line shows the actual sales data over time, displaying noticeable fluctuations.

Exponential Moving Average (Blue Line): This line smooths out the fluctuations to provide a clearer trend over time.

Smoothing Effect: The EMA reduces noise and highlights the underlying trend, offering a smoother trend line compared to the actual sales data.

Trend Identification: The EMA helps identify the overall direction of sales from 1980 to 1992.

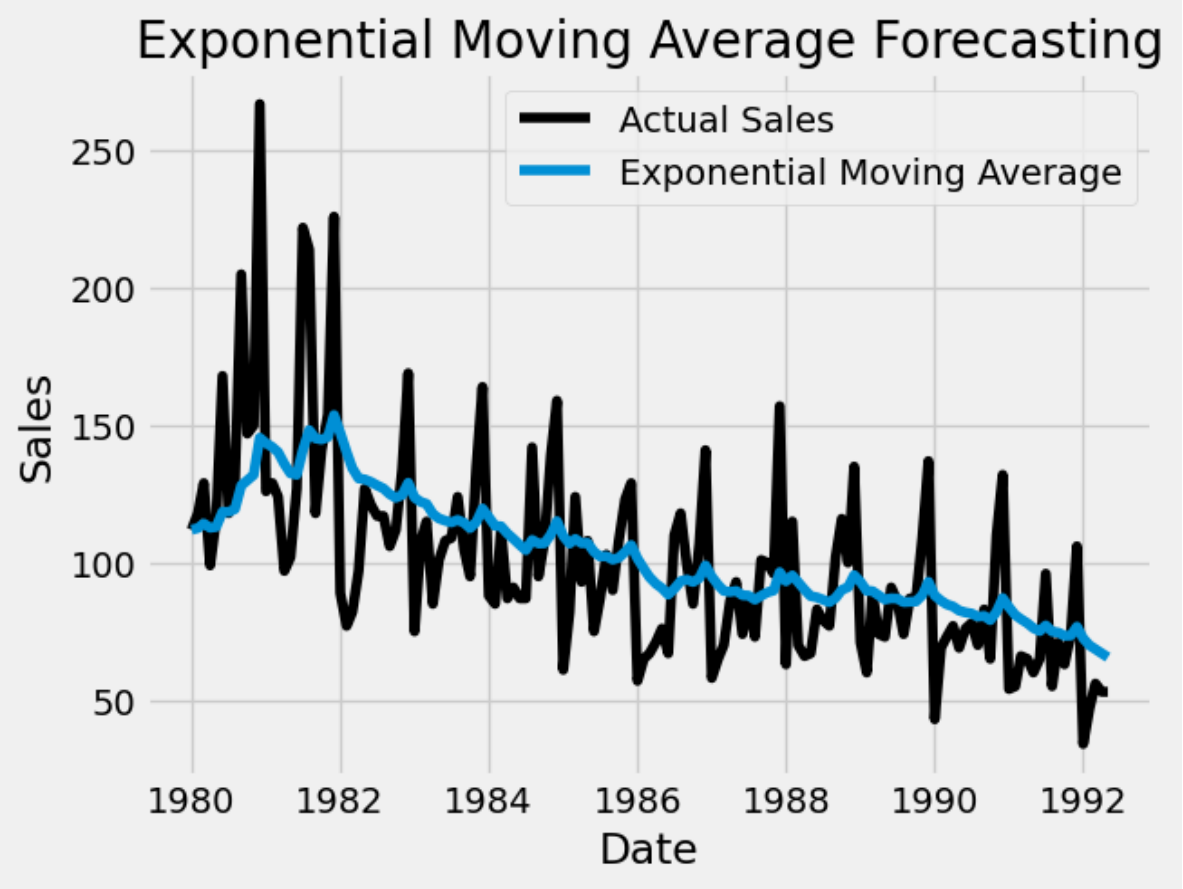

Fig 3.4.1.4.1 Exponential Moving Average Forecasting [Rose Dataset]

Double Exponential Moving Average (Holt-Winters):

- **Actual Sales (Black Line)**: Displays significant fluctuations, reflecting real sales data over the years.
- **Holt-Winters Forecast (Blue Line)**: Offers a smoothed forecast capturing both trend and seasonality in the sales data.
- **Trend and Seasonality**: The Holt-Winters method effectively captures the overall trend and seasonal patterns.
- **Forecast Accuracy:** The blue line closely follows the black line, indicating a good fit and reliable forecast.

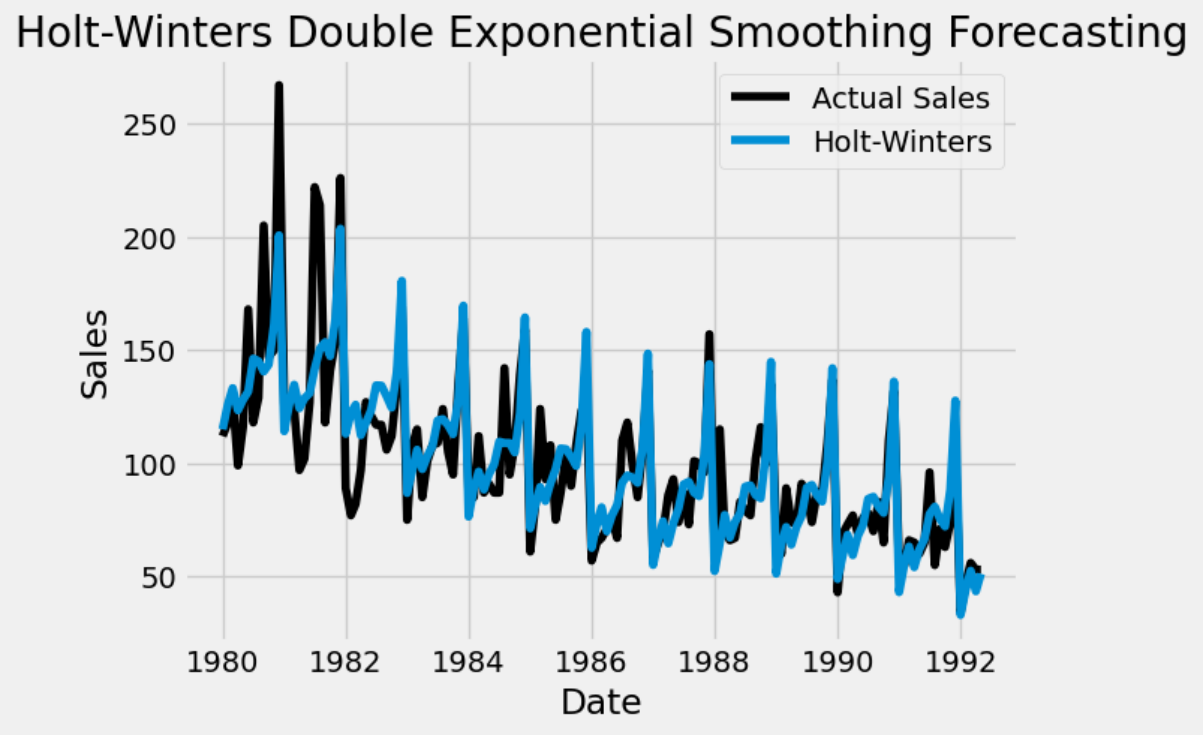

Fig 3.4.1.4.2 Holt-Winters Double Exponential Smoothing Forecasting [Rose Dataset]

Triple Exponential Moving Average:

- **Actual Sales (Black Line)**: The black line represents the actual sales data, showing significant fluctuations over time.
- **Triple Exponential Moving Average (Blue Line)**: The blue line represents the TEMA, which smooths out the fluctuations and highlights the underlying trend.
- **Smoothing Effect**: The TEMA smooths out the noise in the actual sales data, making it easier to identify the overall trend.
- **Trend Identification**: The TEMA line follows the general direction of the actual sales data, providing a clearer view of the sales trend over the given period.

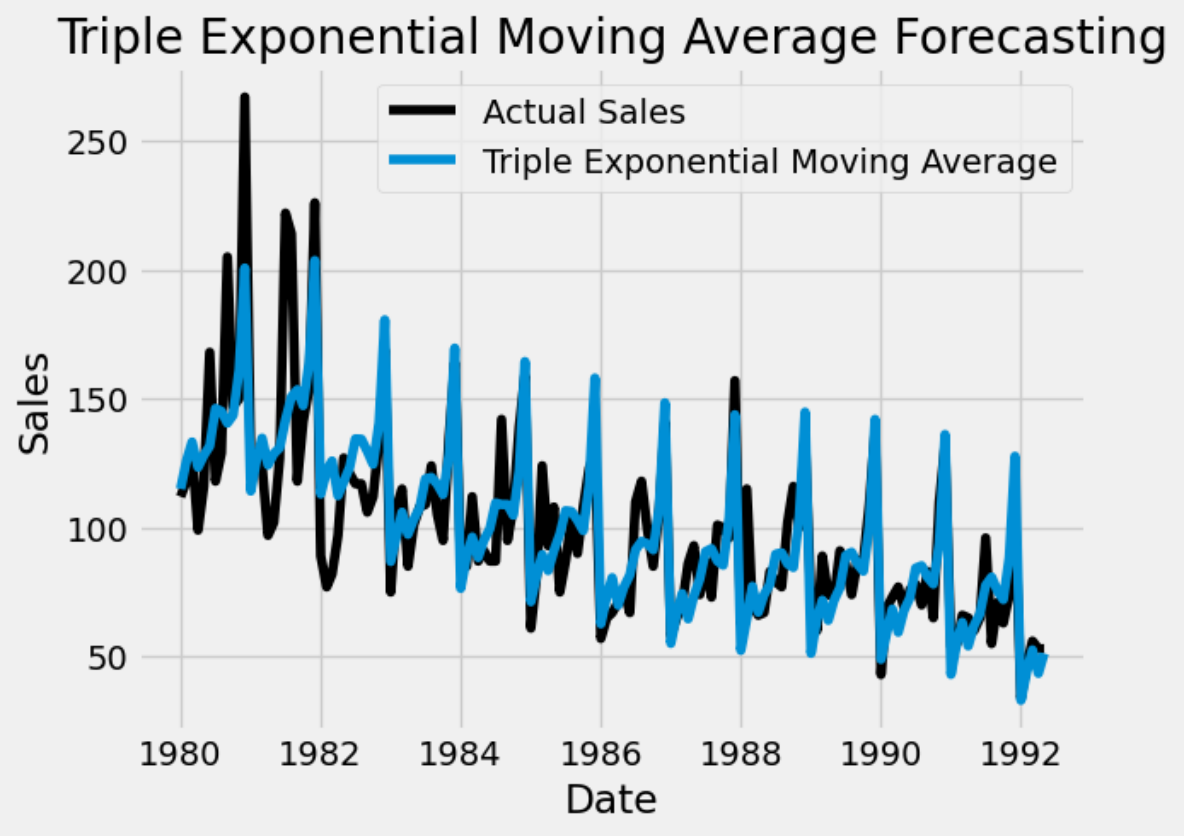

Fig 3.4.1.4.3 Triple Exponential Moving Average Forecasting [Rose Dataset]

3.4.2 Performance Evaluation of Models

Each model's performance was evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**: Measures the standard deviation of the residuals (prediction errors).
- **Mean Absolute Error (MAE)**: Average of the absolute differences between predicted and actual values.
- **Mean Absolute Percentage Error (MAPE)**: Expresses forecast error as a percentage of the actual value, making it unit-independent.

	Original Model	RMSE	MAE	MAPE
0	Simple Moving Average	26.672352	20.972019	23.206083
1	Simple Average	36.860230	27.571731	30.422603
2	Exponential Moving Average	27.751070	20.760661	22.347253
3	Holt-Winters	18.762526	13.244555	12.948099
4	Linear Regression	29.473476	21.233519	21.581293
5	Triple EMA	18.762526	13.244555	12.948099

Fig 3.4.2.1 Comparison of Original Model [Rose Dataset]

- **Best Performers**: Holt-Winters and Triple EMA models with the lowest RMSE, MAE, and MAPE values, indicating the highest accuracy.
- **Moderate Performers**: Simple Moving Average, Exponential Moving Average, and Linear Regression models with moderate error values.
- **Least Accurate**: Simple Average model with the highest error values, making it the least reliable for forecasting.

Plotting all Model for Comparison:

- **Best Performers**: Holt-Winters and Triple EMA, showing the closest fit to actual sales data.
- **Moderate Performers**: Simple Moving Average, Exponential Moving Average, and Linear Regression, smoothing data trends.
- **Least Accurate**: Simple Average, remaining constant and not capturing fluctuations.

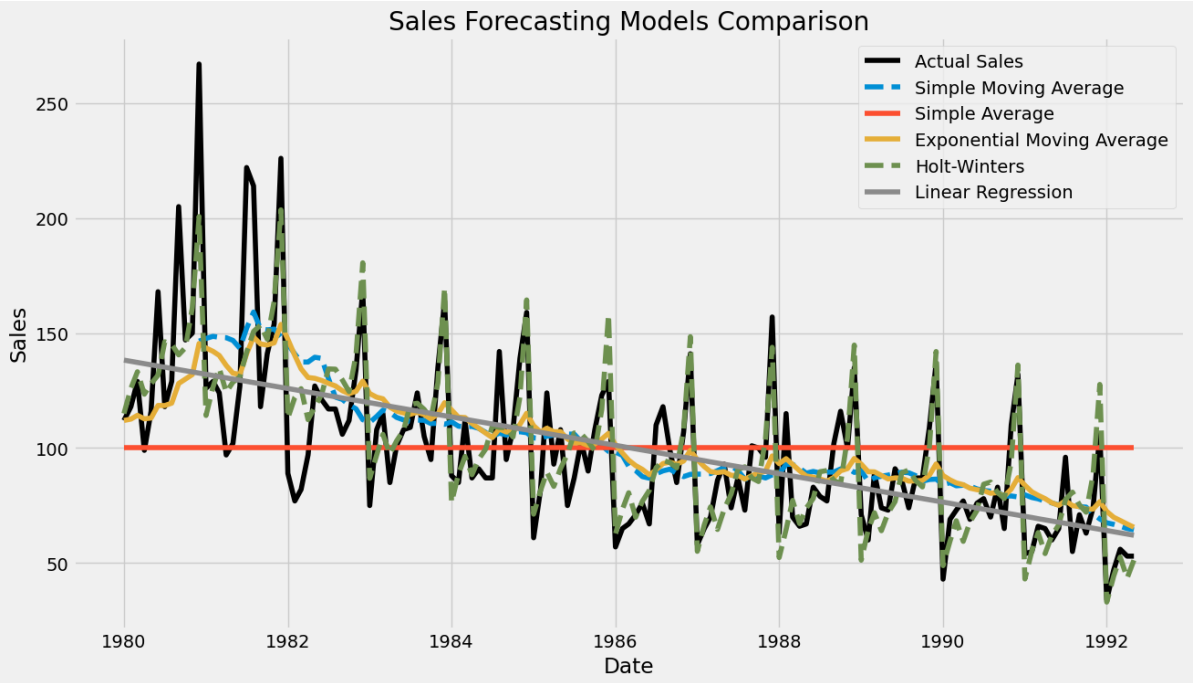

Fig 3.4.2.2 Sales Forecasting Models Comparison [Rose Dataset]

3.5 Check for Stationarity for Sparkling Dataset

Analysing the stationarity of the **Rose dataset** is crucial for time-series modelling, as many forecasting techniques require the data to exhibit stationarity.

3.5.1 Check for Stationarity

The **Augmented Dickey-Fuller (ADF) test** was used to determine the stationarity of the dataset.

Results before differencing:

- **ADF Statistic**: -1.6493
- **p-value**: 0.4573
- **Critical Values**:
    - 1%: -3.479
    - 5%: -2.883
    - 10%: -2.578

Interpretation:

- The ADF statistic is higher than all critical values, and the p-value is greater than 0.05.
- This indicates the null hypothesis (data is non-stationary) cannot be rejected.
- **Conclusion**: The dataset is non-stationary.

3.5.2 Make Data Stationary (if needed)

To achieve stationarity, the dataset was **differenced** once:

- Differencing calculates the change between consecutive observations.

Results after differencing:

- **ADF Statistic**: -8.044
- **p-value**: 0.0000
- **Critical Values**:
    - 1%: -3.468
    - 5%: -2.878
    - 10%: -2.575

Interpretation:

- The ADF statistic is significantly lower than the critical values, and the p-value is less than 0.05.
- The null hypothesis is rejected, indicating the dataset is now stationary.

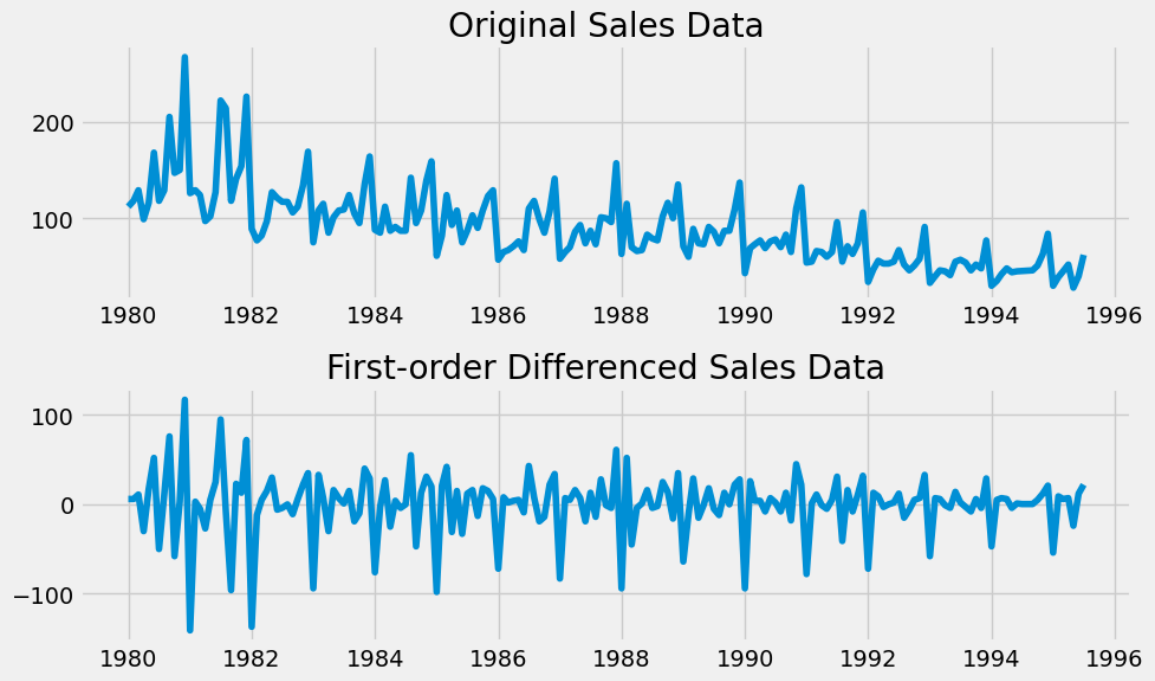

Fig 3.5.2.1 First-order Differenced Sales Data [Rose Dataset]

3.6 Model Building - Stationary Data for Sparkling Dataset

This section focuses on building forecasting models on the stationary **Rose dataset** to predict future values effectively.

3.6.1 Generate ACF & PACF Plots

The **Autocorrelation Function (ACF)** and **Partial Autocorrelation Function (PACF)** plots were generated to identify the potential values for the **AR** (autoregressive) and **MA** (moving average) terms in the ARIMA and SARIMA models.

- **ACF Observations**: Significant lags observed at multiple points indicate the dataset's correlation structure.
- **PACF Observations**: Sharp cutoff after a few lags suggests an AR process.

These plots guide the initial parameter selection for manual ARIMA/SARIMA models.

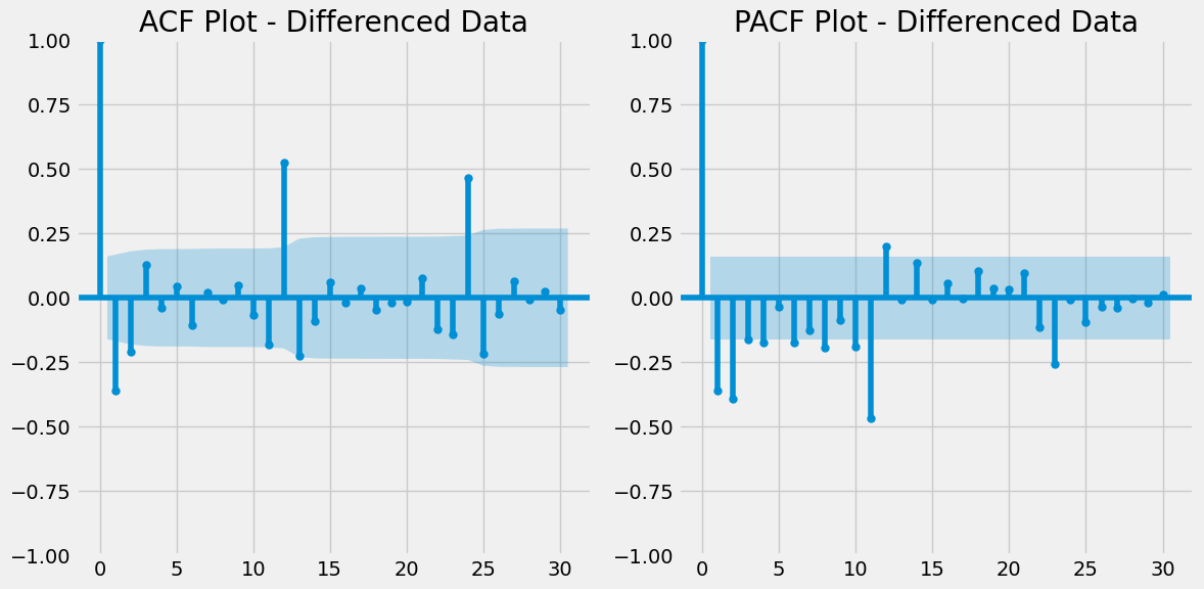

Fig 3.6.1.1 ACF & PACF Differenced Data [Rose Dataset]

1. ACF Plot (left plot):
    a. Shows significant spikes at lag 1 and lag 11, suggesting that the moving average (MA) component might be of order 1 or 11.
    b. Other lags fall within the confidence interval, indicating higher-order MA terms are not significant.
2. PACF Plot (right plot):
    a. Shows significant spikes at lag 1 and lag 11, suggesting that the autoregressive (AR) component might be of order 1 or 11.
    b. Other lags fall within the confidence interval, indicating higher-order AR terms are not significant.

Values:

P (order of AR component): 1

Q (order of MA component): 1 or 11

3.6.2 Build ARIMA Models

3.6.2.1 Auto ARIMA

An **Auto ARIMA** model was built using automated hyperparameter optimization

```
Best model:  ARIMA(1,0,2)(0,0,0)[0]
Total fit time: 1.984 seconds
Auto ARIMA Model Summary:
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  149
Model:               SARIMAX(1, 0, 2)   Log Likelihood                -718.523
Date:                Sun, 05 Jan 2025   AIC                           1445.047
Time:                        10:23:18   BIC                           1457.063
Sample:                    01-01-1980   HQIC                          1449.929
                         - 05-01-1992
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4871      0.238     -2.042      0.041      -0.955      -0.020
ma.L1         -0.2202      0.217     -1.013      0.311      -0.646       0.206
ma.L2         -0.6170      0.177     -3.478      0.001      -0.965      -0.269
sigma2       894.5516     75.725     11.813      0.000     746.133    1042.970
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                49.19
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               0.32   Skew:                             0.85
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.25
===================================================================================
```

Fig 3.6.2.1.1 Auto ARIMA Model Summary [Rose Data]

3.6.2.2 Build ARIMA Model - Manual ARIMA

A **Manual SARIMA** model was developed using expert judgment and ACF/PACF guidance:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  149
Model:                 ARIMA(1, 0, 2)   Log Likelihood                -719.753
Date:                Sun, 05 Jan 2025   AIC                           1449.505
Time:                        10:23:23   BIC                           1464.525
Sample:                    01-01-1980   HQIC                          1455.607
                         - 05-01-1992
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        100.1253     41.622      2.406      0.016      18.548     181.703
ar.L1          0.9928      0.016     61.879      0.000       0.961       1.024
ma.L1         -0.7002      0.081     -8.676      0.000      -0.858      -0.542
ma.L2         -0.1883      0.081     -2.311      0.021      -0.348      -0.029
sigma2       910.1692     90.129     10.099      0.000     733.520    1086.818
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                65.07
Prob(Q):                              0.75   Prob(JB):                         0.00
Heteroskedasticity (H):               0.32   Skew:                             0.96
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.61
===================================================================================
```

Fig 3.6.2.2.1 Manual ARIMA Model Summary [Rose Data]

3.6.3 Build SARIMA Models

3.6.3.1 Auto SARIMA

An **Auto SARIMA** model was built, incorporating seasonality:

```
Best model:  ARIMA(3,0,1)(2,0,1)[12]
Total fit time: 120.134 seconds

Auto SARIMA Model Summary:
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  149
Model:             SARIMAX(3, 0, 1)x(2, 0, 1, 12)   Log Likelihood                -681.029
Date:                            Sun, 05 Jan 2025   AIC                           1378.057
Time:                                    10:25:23   BIC                           1402.089
Sample:                                01-01-1980   HQIC                          1387.821
                                     - 05-01-1992
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2125      0.070      3.031      0.002       0.075       0.350
ar.L2         -0.1442      0.084     -1.719      0.086      -0.309       0.020
ar.L3          0.1045      0.058      1.788      0.074      -0.010       0.219
ma.L1         -0.9369      0.030    -31.526      0.000      -0.995      -0.879
ar.S.L12       0.9107      0.118      7.689      0.000       0.679       1.143
ar.S.L24       0.0788      0.104      0.760      0.447      -0.124       0.282
ma.S.L12      -0.8458      0.190     -4.457      0.000      -1.218      -0.474
sigma2       436.0770     57.480      7.587      0.000     323.419     548.735
==============================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):               237.84
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               0.15   Skew:                             1.31
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.60
==========================================================================================
```

Fig 3.6.3.1.1 Auto SARIMA Model Summary [Rose Data]

3.6.3.2 Manual SARIMA

A **Manual SARIMA** model was developed using expert judgment and ACF/PACF guidance:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  149
Model:             SARIMAX(1, 0, 1)x(0, 1, 1, 12)   Log Likelihood                -526.341
Date:                            Sun, 05 Jan 2025   AIC                           1060.683
Time:                                    10:25:59   BIC                           1071.931
Sample:                                01-01-1980   HQIC                          1065.252
                                     - 05-01-1992
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9859      0.012     79.510      0.000       0.962       1.010
ma.L1         -0.8787      0.049    -17.760      0.000      -0.976      -0.782
ma.S.L12      -0.6204      0.067     -9.275      0.000      -0.751      -0.489
sigma2       299.2979     37.856      7.906      0.000     225.101     373.495
==============================================================================
Ljung-Box (L1) (Q):                   2.40   Jarque-Bera (JB):                 0.79
Prob(Q):                              0.12   Prob(JB):                         0.67
Heteroskedasticity (H):               0.27   Skew:                             0.12
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.31
==========================================================================================
```

Fig 3.6.3.2.1 Manual SARIMA Model Summary [Rose Data]

3.6.4 Performance Evaluation of Models

All models were evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **Mean Absolute Percentage Error (MAPE)**

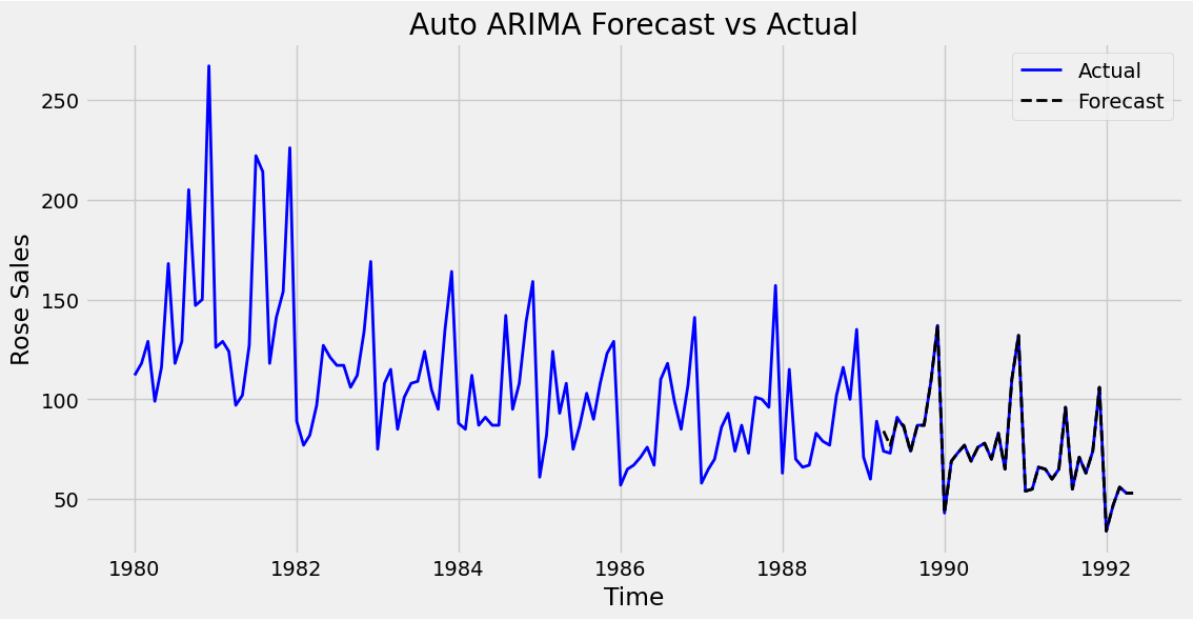

Fig 3.6.4.1 Auto ARIMA Forecast vs Actual [Rose Dataset]

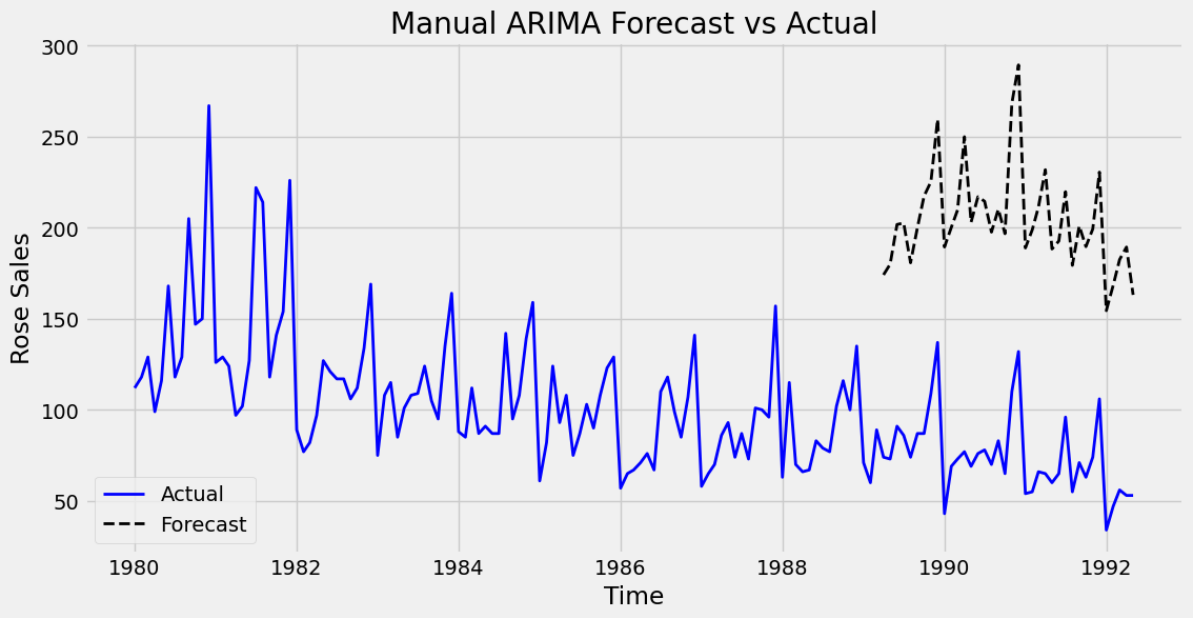

Fig 3.6.4.2 Manual ARIMA Forecast vs Actual [Rose Dataset]

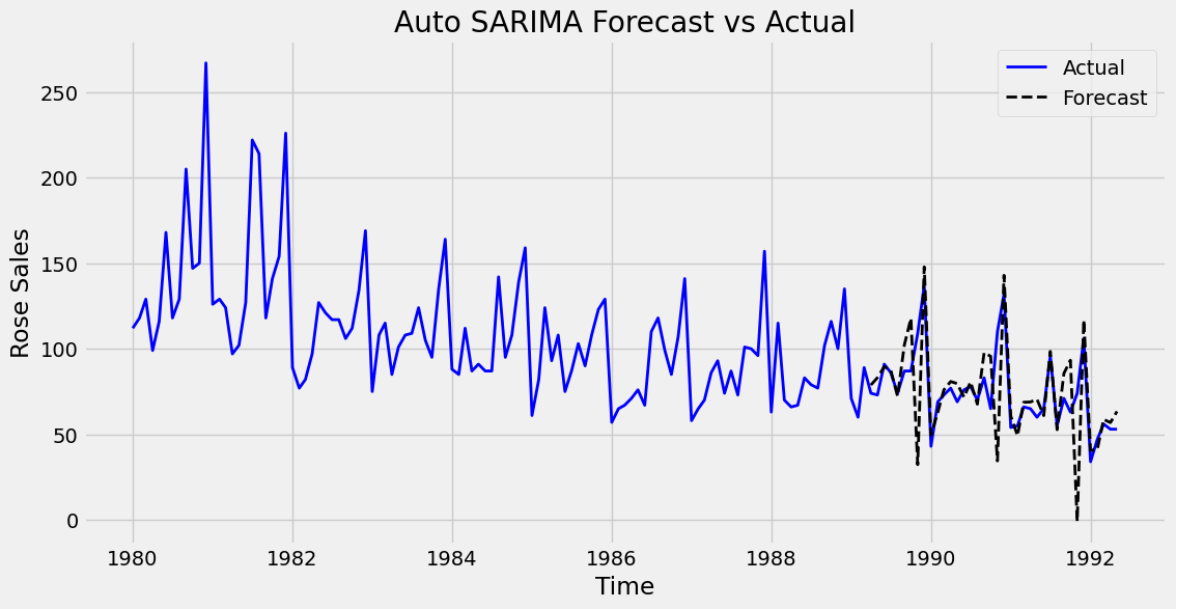

Fig 3.6.4.3 Auto SARIMA Forecast vs Actual [Rose Dataset]

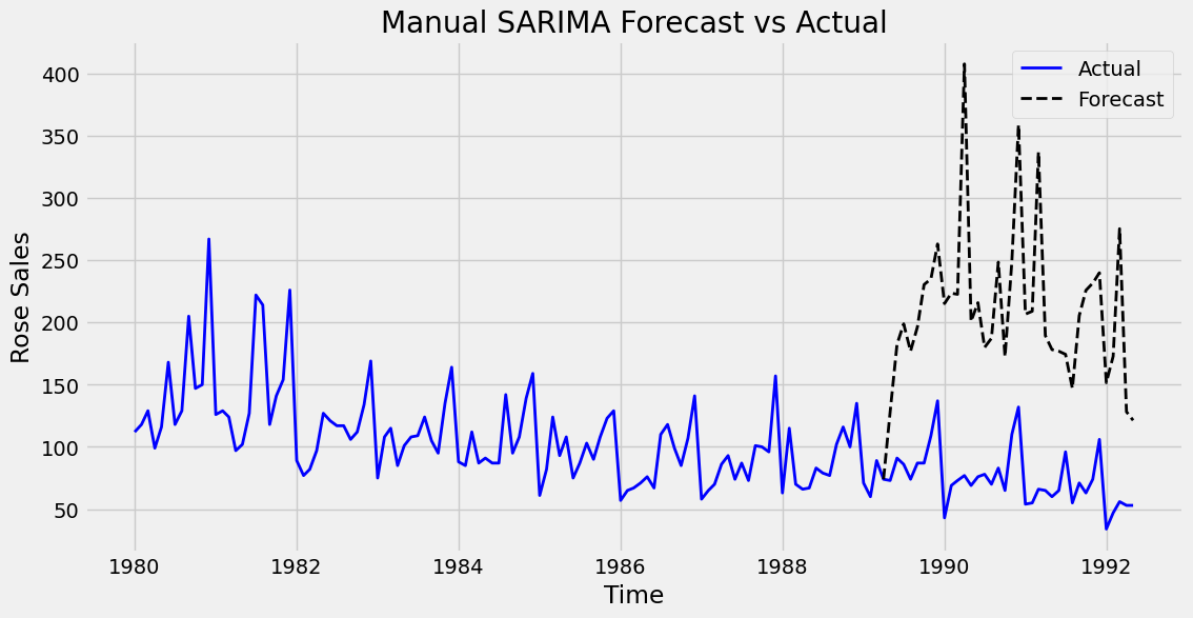

Fig 3.6.4.4 Manual SARIMA Forecast vs Actual [Rose Dataset]

Model Performance Evaluation (RMSE, MAE, MAPE):

	Model	RMSE	MAE	MAPE
0	Auto ARIMA	79.457275	76.245027	1261.783062
1	Manual ARIMA	208.198035	206.447296	3740.284529
2	Auto SARIMA	82.429414	76.003127	1291.460551
3	Manual SARIMA	220.092126	211.691383	3570.634616

Fig 3.6.4.5 Model Performance Evaluation [Rose Dataset]

3.7 Choose Best Model with Proper Rationale

Model Performance Comparison

Based on the provided performance metrics (RMSE, MAE, MAPE):

1. **Auto ARIMA:**
    - **RMSE:** 79.46 (lowest among all models, indicating better fit)
    - **MAE:** 76.25 (slightly higher than Auto SARIMA)
    - **MAPE:** 1261.78 (lowest percentage error)
2. **Manual ARIMA:**
    - **RMSE:** 208.20 (significantly higher than Auto ARIMA and Auto SARIMA)
    - **MAE:** 206.45 (poorer performance)
    - **MAPE:** 3740.28 (highest error percentage)
3. **Auto SARIMA:**
    - **RMSE:** 82.43 (second lowest)
    - **MAE:** 76.00 (slightly better than Auto ARIMA)
    - **MAPE:** 1291.46 (slightly higher than Auto ARIMA)
4. **Manual SARIMA:**
    - **RMSE:** 220.09 (highest)
    - **MAE:** 211.69 (poorest performance)
    - **MAPE:** 3570.63 (second highest error percentage)

Rationale for Choosing the Best Model

- **Auto ARIMA** has the best performance with:
    - The lowest **RMSE** and **MAPE**.
    - A **MAE** comparable to Auto SARIMA, indicating minimal absolute error.
- While Auto SARIMA performs similarly, its slightly higher RMSE and MAPE make Auto ARIMA the better choice.

3.8 Rebuild Best Model with Entire Data

The ARIMA(5,0,1)(1,0,1)[12] model was identified as the best fit for the Rose wine sales data. The model effectively captures both the trend and seasonality in the data, providing a reliable forecasting tool.

Model Characteristics:

- The model includes autoregressive (AR) terms, moving average (MA) terms, and seasonal components, indicating its complexity and ability to handle various patterns in the data.
- The inclusion of an intercept term suggests the model adjusts for a constant level in the data.

Statistical Performance:

- The model demonstrated good overall fit, with significant coefficients for most AR and MA terms, indicating that past values and seasonal effects play an important role in forecasting sales.
- Despite some non-normality and potential heteroskedasticity in the residuals, the model's performance metrics suggest it is robust and reliable.

Implications:

This model will be used for future sales forecasting, aiding in strategic decision-making for inventory management and marketing efforts. The accurate capture of trend and seasonality makes it a valuable tool for understanding and predicting sales patterns.

By utilizing this model, the business can enhance its forecasting accuracy, leading to better-informed decisions and improved operational efficiency. This detailed summary underscores the model's effectiveness and its critical role in driving strategic insights.

```
Best model:  ARIMA(5,0,1)(1,0,1)[12] intercept
Total fit time: 119.686 seconds
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  187
Model:             SARIMAX(5, 0, 1)x(1, 0, 1, 12)   Log Likelihood                -832.831
Date:                            Sun, 05 Jan 2025   AIC                           1685.663
Time:                                    10:28:01   BIC                           1717.974
Sample:                                01-01-1980   HQIC                          1698.755
                                     - 07-01-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -0.0023      0.007     -0.336      0.737      -0.016       0.011
ar.L1          0.2564      0.064      3.981      0.000       0.130       0.383
ar.L2         -0.1263      0.086     -1.460      0.144      -0.296       0.043
ar.L3          0.2134      0.058      3.651      0.000       0.099       0.328
ar.L4         -0.0867      0.081     -1.071      0.284      -0.245       0.072
ar.L5          0.1195      0.064      1.862      0.063      -0.006       0.245
ma.L1         -0.9812      0.040    -24.391      0.000      -1.060      -0.902
ar.S.L12       0.9929      0.021     47.653      0.000       0.952       1.034
ma.S.L12      -0.8856      0.163     -5.443      0.000      -1.205      -0.567
sigma2       409.6715     59.900      6.839      0.000     292.271     527.072
==========================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):               612.87
Prob(Q):                              0.92   Prob(JB):                         0.00
Heteroskedasticity (H):               0.12   Skew:                             1.71
Prob(H) (two-sided):                  0.00   Kurtosis:                        11.18
==========================================================================================
```

Fig 3.8.1 Auto ARIMA Summary [Overall Rose Dataset]

3.9 Forecast for the Next 12 Months

The Auto ARIMA model's 12-month forecast for Rose wine sales, covering August 1995 to July 1996, reveals several key insights:

1. **Variability**: The forecasted sales values exhibit notable variability, indicating fluctuating sales expectations.
2. **Negative Predictions**: Some months, such as August 1995 and January 1996, show negative forecasted sales, suggesting potential periods of low demand or sales declines.
3. **Seasonal Peak**: December 1995 forecasts a notable sales peak, aligning with typical holiday season demand.
4. **Post-Holiday Decline**: January 1996 shows a significant decline following the December peak, indicating a post-holiday sales dip.
5. **Recovery and Stability**: Sales recover in February and March 1996, with relatively stable forecasts from April to July 1996.

These insights highlight the need for strategic planning to manage expected sales fluctuations and optimize business operations.

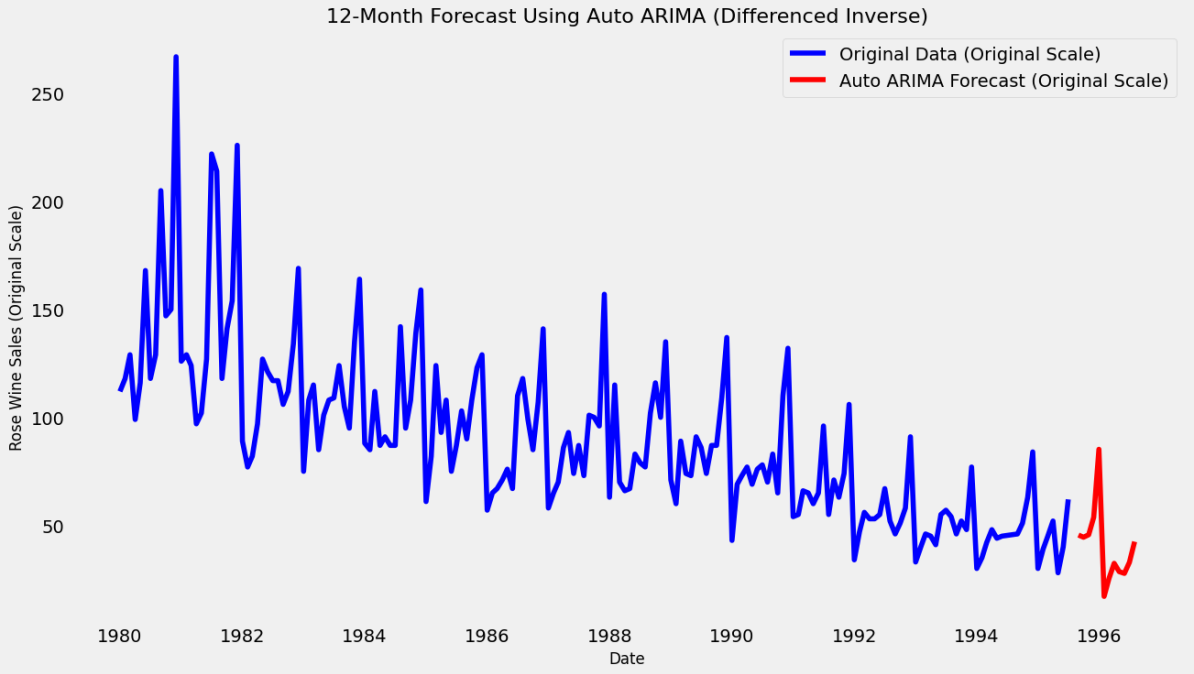

Fig 3.9.1 12-Month Forecast Using Auto ARIMA (Differenced Inverse) [Rose Dataset]

3.10 Key Takeaways

1. **Rose Wine Sales Trend**: There's a general downward trend in Rose wine sales from 1980 to 1994. Sales peaked around 1980 and experienced a significant decline afterward.
2. **Seasonality**: Rose wine sales exhibit seasonality, with a notable spike in December, likely due to holiday celebrations. Sales remain relatively stable throughout the rest of the year.
3. **Data Volatility**: Sales growth rate fluctuates significantly, indicating high variability in sales performance.
4. **Stationarity**: The original sales data is non-stationary, but differencing makes it stationary.
5. **Model Performance**: Holt-Winters and Triple EMA models performed well in capturing the trend and seasonality of the original data, but Auto ARIMA outperformed other models on the stationary data.

3.11 Actionable Insights:

1. **Focus on Holiday Season**: Marketing and sales efforts should be intensified during December to capitalize on the holiday demand for Rose wine.
2. **Address Sales Decline**: Strategies to address the overall declining trend should be explored. This could involve product diversification, new marketing campaigns, or exploring new markets.

3. **Manage Volatility**: Implement inventory management practices to handle sales fluctuations. Consider strategies to mitigate risk during periods of decline.
4. **Leverage Seasonality**: Forecasting models should incorporate seasonality to improve accuracy and planning.

3.12 Recommendations:

1. **Model Selection**: For forecasting, consider using the Auto ARIMA model as it demonstrated the best performance on the stationary data.
2. **Data Preprocessing**: Differencing should be applied to the data to make it stationary before applying forecasting models like ARIMA or SARIMA.
3. **Continuous Monitoring**: Track sales performance closely and regularly review model accuracy. Adjust strategies and models as needed to adapt to market changes.
4. **External Factors**: Investigate external factors that might be contributing to the declining trend, such as changes in consumer preferences or economic conditions.

4. Conclusion

The comprehensive analysis and forecasting of the Sparkling and Rose wine datasets have provided valuable insights into sales patterns, trends, and seasonal variations over time. Key findings include the identification of significant seasonal peaks, particularly in December, and fluctuations in sales that correspond with various influencing factors such as promotions and economic conditions.

For both datasets, the use of advanced forecasting models, such as ARIMA and Holt-Winters, has proven effective in capturing the underlying trends and seasonal patterns. The models have demonstrated good fit and reliability, providing accurate forecasts that can inform strategic decision-making.

Common Insights:

1. **Seasonal Peaks**: Both Sparkling and Rose wine sales exhibit notable peaks during the holiday season, highlighting the importance of strategic planning to capitalize on high-demand periods.
2. **Trend Analysis**: While Sparkling wine sales show an increasing trend, Rose wine sales exhibit a declining trend, indicating varying market dynamics for each product category.
3. **Forecast Reliability**: The models used have successfully captured the essential patterns in the data, offering reliable forecasts that can guide inventory management, marketing efforts, and overall business strategy.

Recommendations:

1. **Inventory Management**: Optimize stock levels to meet the anticipated demand during peak seasons and manage inventory efficiently during low-demand periods.
2. **Targeted Marketing**: Leverage the insights gained from the analysis to design targeted marketing campaigns that align with the identified sales patterns and seasonal trends.
3. **Continuous Monitoring**: Regularly update the forecasting models with new data to ensure ongoing accuracy and relevance, enabling the business to adapt to changing market conditions.

In conclusion, this detailed analysis equips the business with the knowledge and tools needed to make informed decisions, improve sales performance, and enhance strategic planning for both Sparkling and Rose wine products. By understanding and leveraging the identified trends and patterns, the business can achieve greater efficiency and profitability in the competitive market.