Project 8

- Time Series Forecasting -

- (Time Series Sales Forecasting for Sparkling Wine using ARIMA, SARIMA & Holt-Winters) -

1.Introduction

1.1.Problem Definition

1.1.1 Introduction

Wine consumption and sales have been influenced by various factors over the years, including economic conditions, consumer preferences, and seasonal demand. ABC Estate Wines, a well-established player in the wine industry, has accumulated historical sales data spanning the 20th century. This dataset provides an opportunity to analyze trends and patterns that have shaped wine sales over time. By leveraging data analytics and forecasting techniques, we aim to extract valuable insights that can enhance strategic decision-making and optimize future sales strategies.

1.1.2.Business Problem

ABC Estate Wines operates in a highly competitive market where understanding sales trends is crucial for sustained growth. The company deals with multiple wine varieties, each exhibiting unique sales patterns influenced by seasonality, external economic factors, and shifts in consumer behavior. Without a clear grasp of these trends, ABC Estate Wines faces challenges in demand forecasting, inventory management, and production planning. Ineffective decision-making could lead to missed revenue opportunities, stock shortages, or excess inventory, ultimately affecting profitability.

1.1.3.Objective

The primary objective of this project is to analyze and forecast wine sales trends throughout the 20th century using historical data from ABC Estate Wines. Through this analysis, we aim to:

- Identify seasonal variations and long-term sales trends.
- Understand key factors influencing fluctuations in wine sales.
- Develop forecasting models to predict future sales performance.
- Provide actionable recommendations to optimize sales strategies.

By gaining deeper insights into historical sales trends, ABC Estate Wines can make informed decisions, improve operational efficiency, and maintain a competitive advantage in the wine industry.

1.2.Data Background and Contents

1.2.1.Dataset Overview:

The dataset consists of historical monthly sales data for sparkling wine from ABC Estate Wines. It contains 187 records with two columns:

- YearMonth: Represents the year and month of the sales data in a YYYY-MM format.
- Sparkling: Represents the number of sparkling wine units sold in that particular month.

There are no missing values in the dataset, ensuring completeness for analysis. The “YearMonth” column is currently in object format and may need to be converted to a datetime format for time series analysis. The dataset captures the long-term sales trends and seasonal patterns of sparkling wine over time.

1.2.2.Statistical Summary:

A descriptive analysis of the Sparkling sales column provides the following insights:

- Mean Sales: 2,402 units per month
- Standard Deviation: 1,295 units, indicating significant variation in monthly sales.
- Minimum Sales: 1,070 units in the lowest-selling month.
- Maximum Sales: 7,242 units in the highest-selling month.
- 25th Percentile (Q1): 1,605 units, meaning 25% of the months had sales below this value.
- Median (Q2/50th Percentile): 1,874 units, representing the middle of the sales distribution.
- 75th Percentile (Q3): 2,549 units, indicating that 75% of the months had sales below this value.

These statistics suggest a high variability in sales, with certain months experiencing significantly higher demand. The presence of extreme values, such as the maximum of 7,242 units, hints at potential seasonal trends or promotional effects that drive sales spikes. Further exploration of sales patterns over time will help uncover the underlying factors influencing these variations.

1.3.Reading the Data as a Time Series

The dataset consists of time-dependent observations, with a dedicated column representing dates or timestamps. Each entry corresponds to a recorded value at a specific point in time. To ensure accurate time series analysis, the dataset must be properly structured. This process begins with importing the data and examining its structure, followed by converting the date column into a proper DateTime format to enable time-based operations. The date column is then set as the index, ensuring that all subsequent analyses are conducted in a time-dependent manner. To maintain chronological order, the dataset is sorted based on the time index. An initial exploration is performed using summary statistics and visualizations to detect trends, patterns, or anomalies. These steps establish the dataset as a structured time series, forming the foundation for further exploratory analysis, decomposition, and forecasting.

1.4.Visualizing the Data and Exploratory Data Analysis

Before proceeding with time series modeling, it is essential to explore the dataset to understand its underlying patterns. Visualization helps identify trends, seasonality, and variability in sales over time. In this section, we analyze the sales data through different visualizations to gain insights into its structure.

1. Monthly Sparkling Sales Over the Years

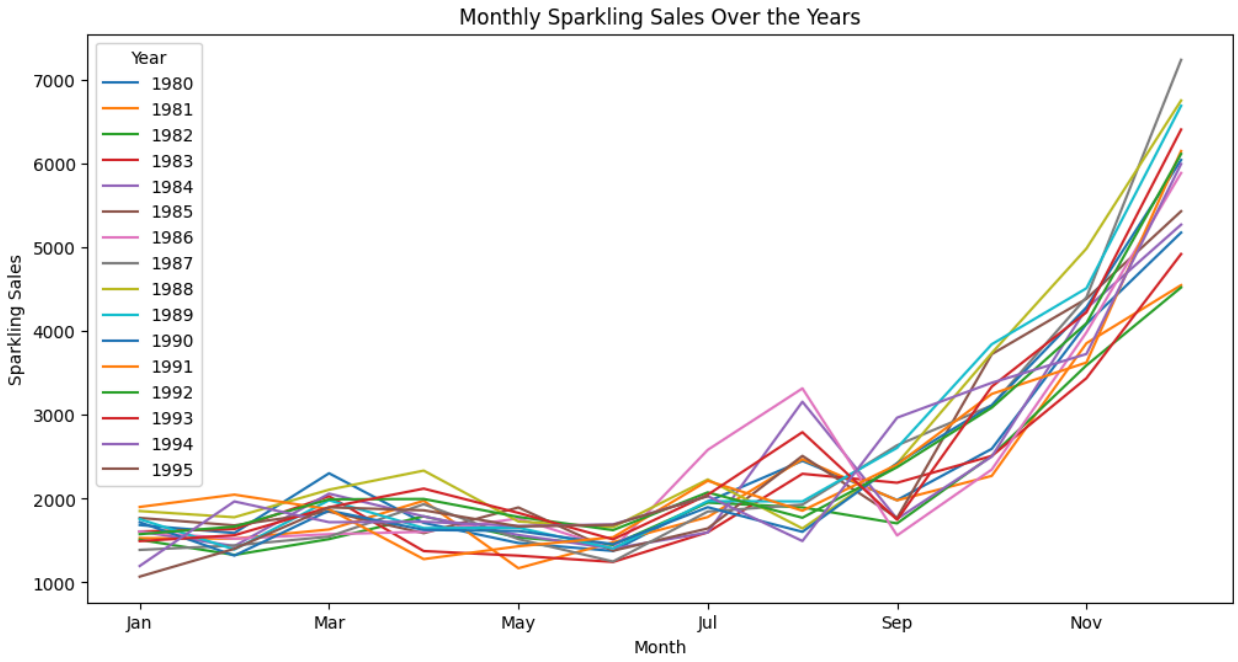

Figure 1. Monthly Sparkling Sales Over the Years

Interpretation :

The plot reveals a strong seasonal trend in sparkling sales, with relatively stable figures from January to July, followed by a noticeable increase starting in September and peaking in November and December. This consistent year-end surge across multiple years suggests a recurring seasonal pattern, likely driven by heightened demand during the holiday season. While sales in the early months exhibit some variability, the overall trend remains similar each year, reinforcing the seasonal nature of the data.

2. Yearly Sparkling sales Trend by Month

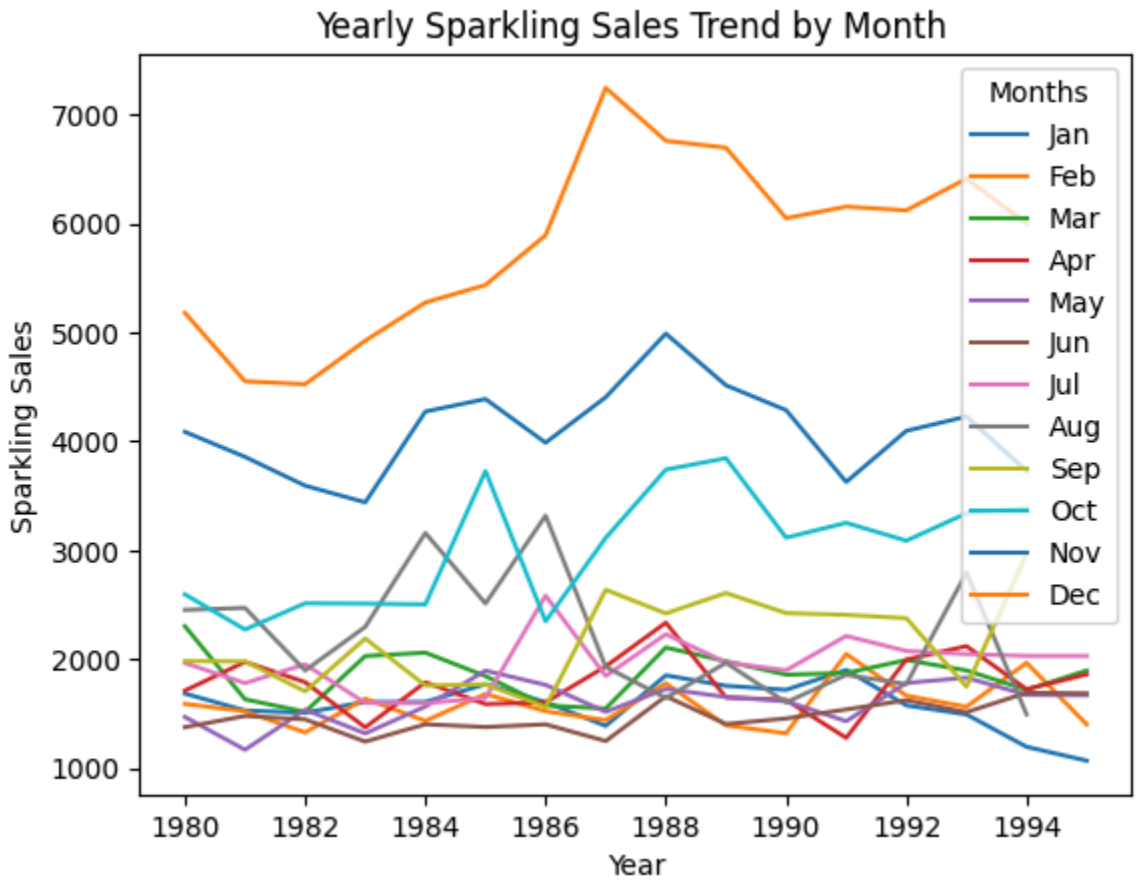

Figure 2. Yearly Sparkling sales Trend by Month

Interpretation :

The line plot illustrates the yearly trends in sparkling sales for each month. It shows that November and December consistently exhibit the highest sales volumes compared to other months, reinforcing the strong seasonal pattern observed earlier. The sales peak significantly in December, suggesting a surge in demand, possibly due to holiday or festive season sales. February and January have moderate sales, while the other months maintain relatively stable sales with minor fluctuations. This visualization highlights the strong seasonality in sales, with clear distinctions between high and low sales months across different years.

3. Box Plot of Yearly Sales Distribution

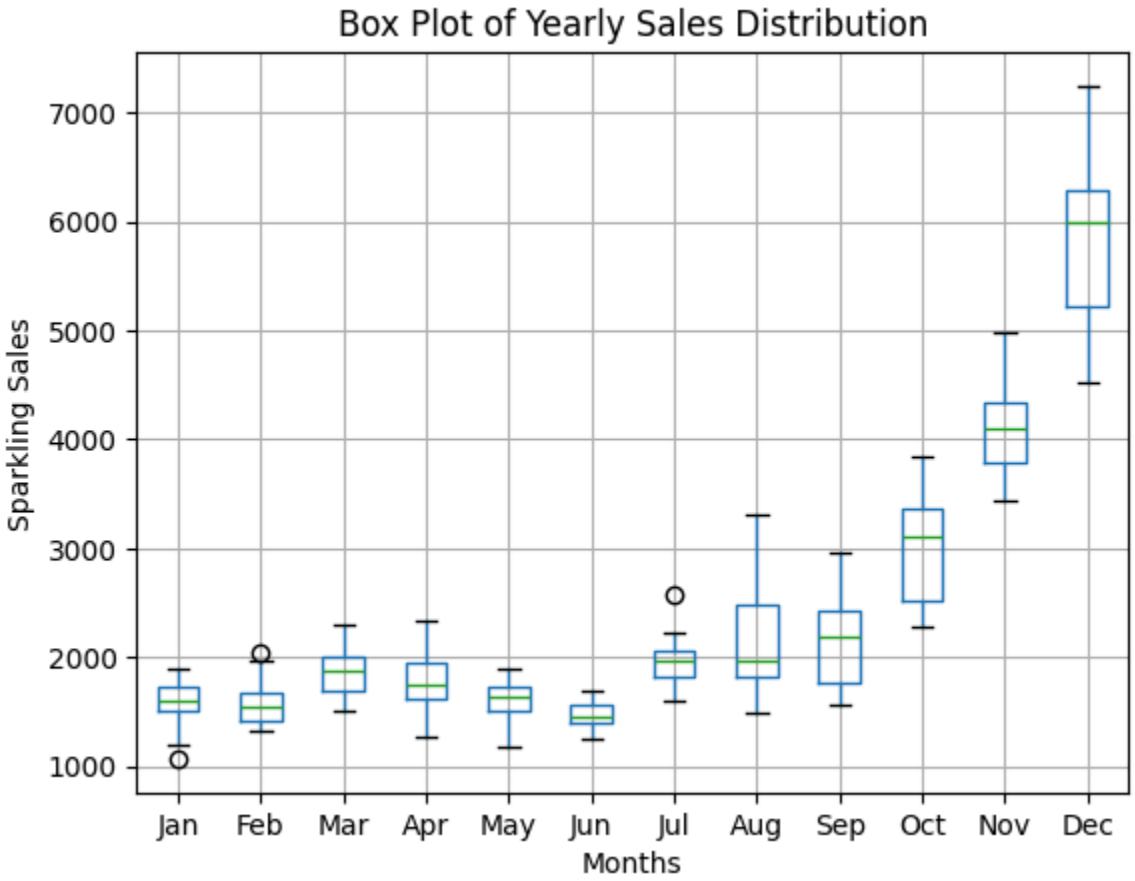

Figure 3. Box Plot of Yearly Sales Distribution

Interpretation :

The box plot illustrates the distribution of sparkling sales across different months over the years. It reveals that sales tend to be significantly higher in the later months, particularly in November and December, compared to the earlier months. The median sales values show a gradual increase as the months progress, with December having the highest median and variability. The interquartile range (IQR) expands towards the end of the year, indicating greater fluctuations in sales. Additionally, the presence of outliers in certain months, such as February and July, suggests occasional extreme sales variations. This trend highlights a strong seasonal effect, where sales peak in the final quarter,.

Insights from Exploratory Data Analysis:

- Sales data exhibit a strong seasonal pattern, with peaks occurring consistently toward the end of each year.
- Yearly sales trends show periods of rapid growth and fluctuations, reflecting market dynamics.
- The box plot highlights increasing variability in sales, suggesting growing demand with occasional outliers.

1.5. Perform Time Series Decomposition

Time series decomposition is a technique used to break down a time series into its key components: trend, seasonality, and residuals. This helps in understanding the underlying patterns and variations in the data, making it easier to select an appropriate forecasting model. The decomposition can be additive (when components are independent of each other) or multiplicative (when components depend on the level of the time series).

We opt for a multiplicative decomposition because the seasonal variations in sparkling sales appear to increase as the overall sales trend rises. This suggests that the magnitude of seasonal fluctuations is proportional to the level of the trend, making a multiplicative model more suitable than an additive one.

The below plot shows the Decomposition of multiplicative time series:

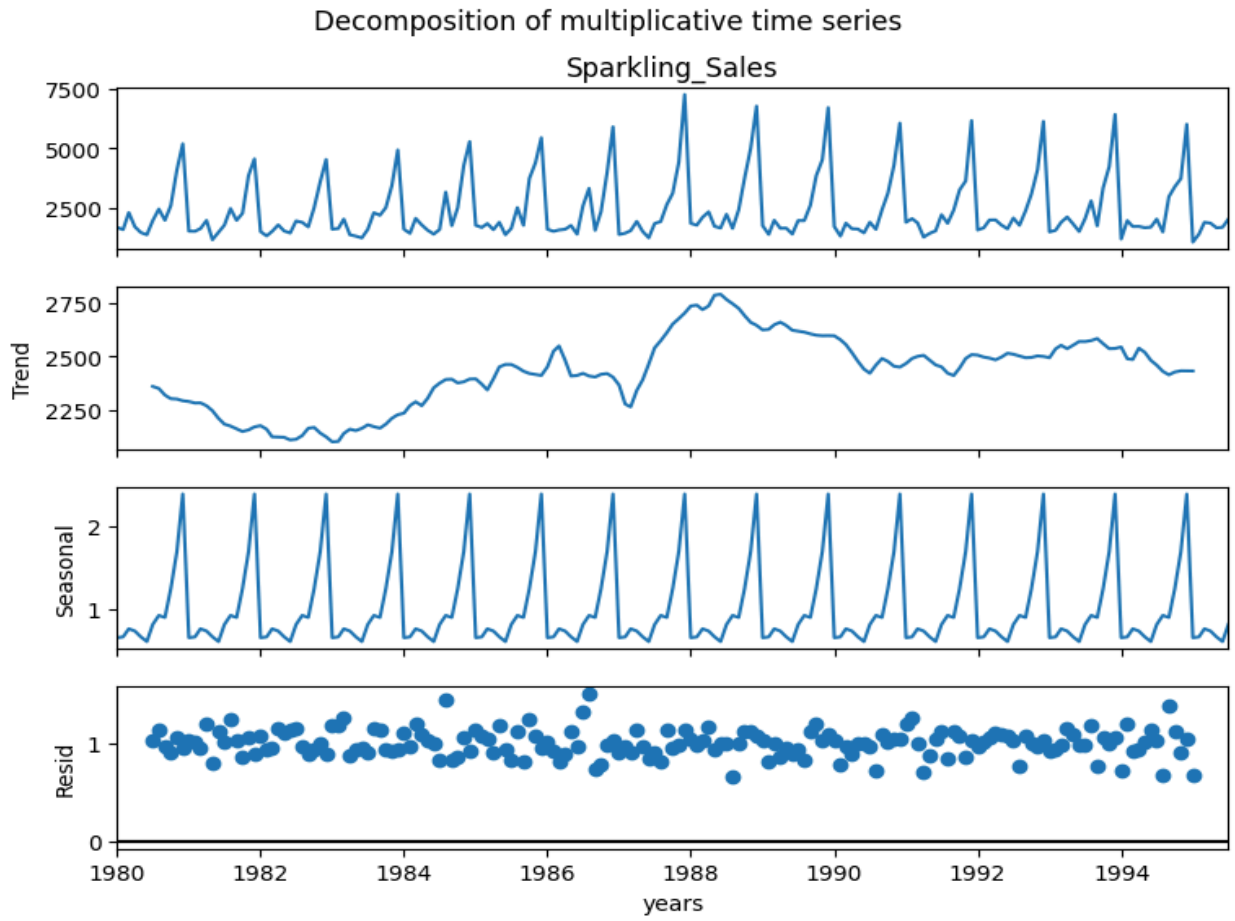

Figure 4. Decomposition of multiplicative time series

Interpretation :

The decomposition of the sparkling sales time series reveals the following insights:

- Trend Component: The trend plot shows an overall increasing pattern in sales, with a sharp rise around 1987, followed by some stabilization.
- Seasonality Component: The seasonal component exhibits strong periodic patterns, indicating that sales tend to peak at certain times of the year.
- Residual Component: The residuals appear to be randomly scattered, suggesting that most of the systematic variations in the data have been captured by the trend and seasonal components.

2.Data Pre-processing

2.1 Handling Missing Values

In time series forecasting, missing values must be carefully handled to ensure the accuracy of the model. Upon examining the dataset, no missing values were detected in the original data. However, after restructuring the dataset into a pivot table format, missing values were observed in the last five months (August to December 1995).

The original dataset contains data from January 1980 to July 1995, meaning that sales records for August–December 1995 were never provided. Since these missing values correspond to future months, they cannot be imputed using traditional statistical methods (such as mean or median imputation) without introducing bias.

Instead of removing or imputing these missing values manually, they will be left as NaN and later predicted using time series forecasting models. The forecasting models will use historical sales patterns to estimate the sales values for the missing months.

2.2.Visualizing the Processed Data

After handling missing values, it is essential to visualize the processed dataset to confirm that the data is correctly structured and ready for modeling. This step helps verify that missing values have been appropriately addressed and allows us to observe the trends in sales over time.

The below plot shows the yearly Sparkling Sales Trend (Processed Data)

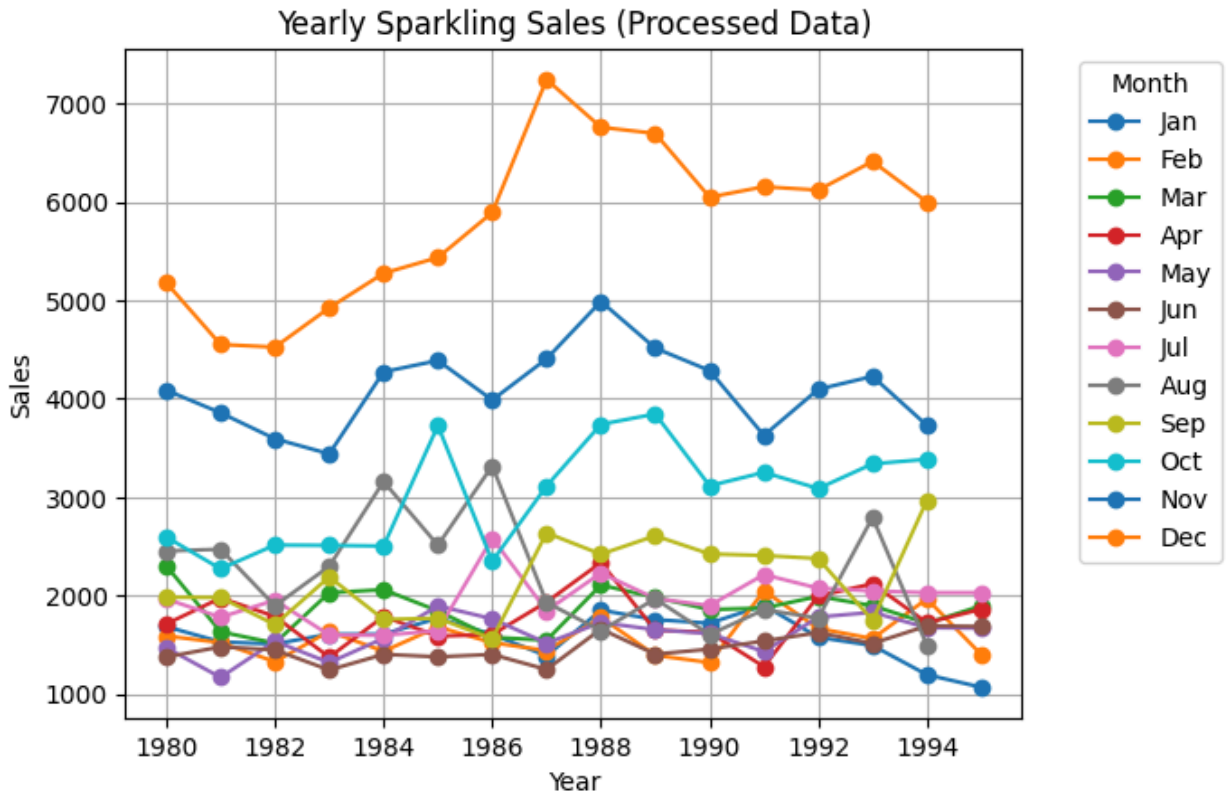

Figure 5. The yearly Sparkling Sales Trend (Processed Data)

Interpretation :

This line plot illustrates the yearly sparkling sales trend after handling missing values. The overall pattern remains consistent, showing a steady increase in sales over time. The seasonal trend is still evident, with sales peaking in the latter months of each year, particularly in November and December. The processed data ensures continuity, making it suitable for further analysis and forecasting.

2.3 Train-Test Split

To build an effective forecasting model, we need to split the dataset into training and testing sets. The training set is used to develop the model, while the test set is reserved for evaluating its predictive performance. In time series forecasting, it is crucial to split data based on time rather than randomly, ensuring that past data is used to predict future values.

We use data before January 1994 as the training set and data from January 1994 onward as the test set. This approach closely resembles real-world forecasting, where future values remain unknown during model training. By assessing model performance on the test set, we can determine its ability to generalize to new, unseen data.

3. Build Forecasting Models

Forecasting is a crucial aspect of time series analysis, allowing businesses to anticipate future trends based on historical data. In this section, we implement multiple forecasting models to predict future sparkling sales. Each model follows a unique approach to capture trends, seasonality, and overall patterns in the data.

The forecasting models chosen for this project range from traditional statistical methods to more advanced smoothing techniques:

1. Linear Regression – A fundamental predictive model that identifies linear relationships between time and sales.
2. Simple Average – A naive method that forecasts future values based on the average of past observations.
3. Moving Average – A smoothing technique that reduces short-term fluctuations to highlight underlying trends.
4. Exponential Smoothing – A family of models (Single, Double, Triple) that assigns exponentially decreasing weights to past observations, making them more effective for time series data with trends and seasonality.

Each model's effectiveness will be evaluated based on its accuracy and ability to capture the sales patterns.

3.1 Linear Regression

Linear Regression is one of the simplest forecasting techniques used in time series analysis. It assumes a linear relationship between the dependent variable (Sparkling Sales) and time. In this approach, time is transformed into a numerical feature, and a regression model is trained to identify the best-fitting trend line. This method is useful when sales exhibit a consistent increasing or decreasing pattern over time.

For this analysis, we trained a Linear Regression model using historical sales data up to 1993 and predicted sales for the test period from 1994 onward. The model estimates future sales based on the established trend, which helps in understanding long-term growth patterns.

The plot below visualizes the actual vs. predicted sales using Linear Regression.

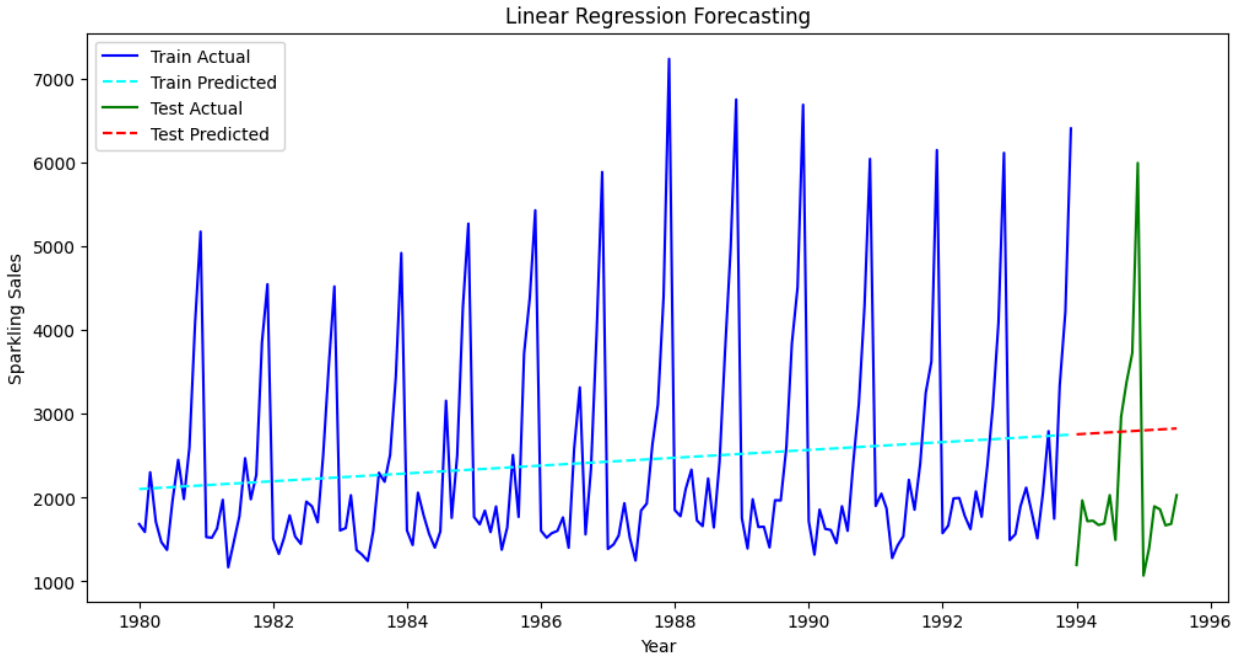

Figure 6. Linear Regression Forecasting Plot

Interpretation :

- The blue and green lines represent the actual sales for the training and test datasets, respectively.
- The cyan and red dashed lines represent the predicted sales by the Linear Regression model.
- The model captures the overall trend in sales, but since time series data often exhibits seasonality and irregular variations, a simple linear approach may not always be sufficient.

3.2 Simple Average Method

The Simple Average method is one of the most straightforward forecasting techniques. It assumes that future values can be estimated by taking the historical average of past observations. This approach is effective when the time series data does not exhibit strong trends or seasonality, as it provides a stable baseline for comparison with more complex models.

To implement this method, we calculated the average of all sales values in the training dataset and used this constant value as the forecast for the test period. While this method does not account for patterns or fluctuations in the data, it serves as a useful benchmark for evaluating the performance of other forecasting models.

The below plot shows the Simple Average Method Forecasting

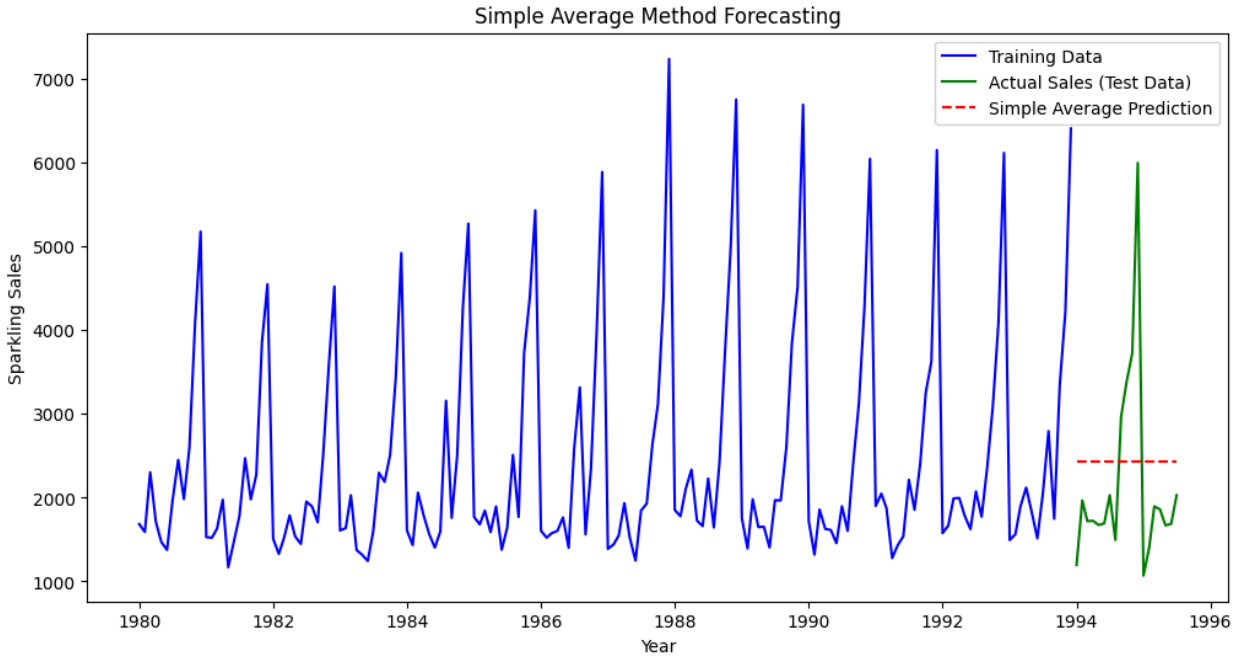

Figure 7. Simple Average Method Forecasting

Interpretation:

The plot above illustrates the actual sales data from the training and test sets, alongside the predictions generated using the Simple Average method. The red dashed line represents the forecasted values, which remain constant throughout the test period. We observe that this method does not adapt to the increasing trend or seasonal variations in sales. As a result, the predicted values may deviate significantly from actual sales, especially in periods with noticeable growth or fluctuations. This highlights the need for more sophisticated forecasting techniques that consider trends and seasonality.

3.3 Moving Average Method

The Moving Average method is a simple yet effective forecasting technique that helps smooth out short-term fluctuations and highlight long-term trends in the data. This method calculates the average of a fixed number of previous observations (known as the window size) and uses this value to predict future points.

For this analysis, a 3-month moving average has been applied. The forecast for the test period is generated using the last available moving average from the training data. This approach ensures that our predictions reflect historical patterns while minimizing the impact of random variations.

The below plot shows the Moving Average Forecasting:

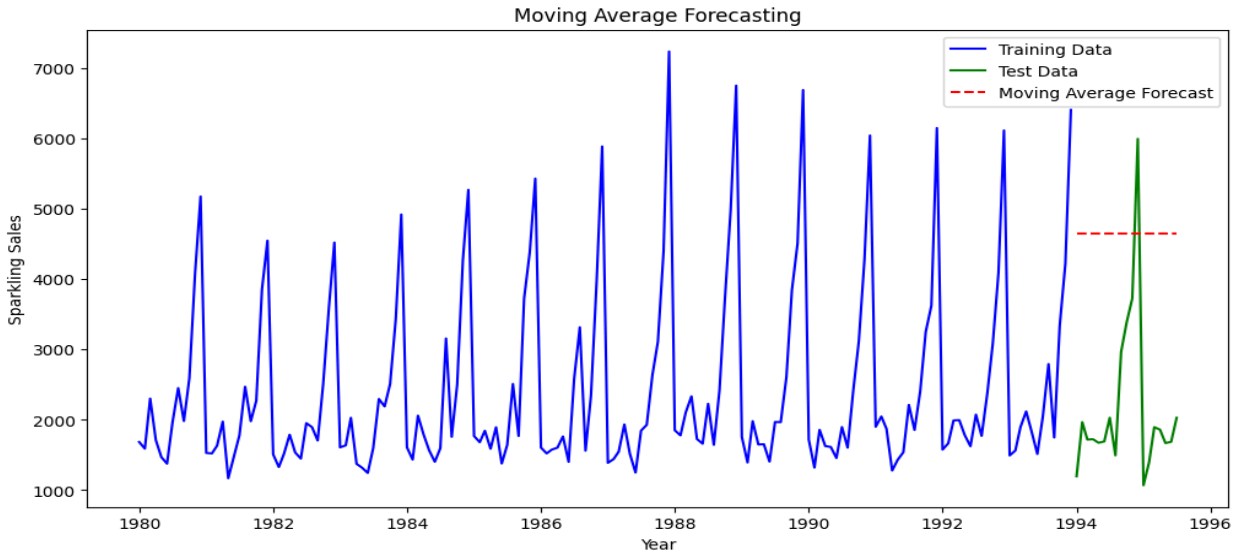

Figure 8. Moving Average Forecasting

Interpretation:

The blue line represents the actual sales values in the training dataset, while the green line represents the test data. The red dashed line shows the predicted values using the moving average method. Since moving averages rely on past observations, the forecasts appear as a smooth continuation of past trends. However, this method may not be effective in capturing seasonal patterns or sudden shifts in sales trends.

3.4 Exponential Smoothing

Exponential Smoothing is a widely used forecasting technique that applies exponentially decreasing weights to past observations. Unlike simple moving averages, it gives more importance to recent data points, making it suitable for capturing short-term patterns. This method is particularly useful when data exhibits trends and seasonality.

Exponential Smoothing has three variations:

- Single Exponential Smoothing (for data with no trend or seasonality)
- Double Exponential Smoothing (for data with trends)
- Triple Exponential Smoothing (Holt-Winters Method) (for data with both trend and seasonality).

3.4.1 Single Exponential Smoothing (SES)

Single Exponential Smoothing (SES) is a simple yet effective forecasting technique used for time series data with no strong trend or seasonality. It applies exponentially decreasing weights to past observations, giving more importance to recent values. The model uses a smoothing parameter (alpha) to determine the rate at which past data points lose influence.

SES has been applied to forecast future sales based on historical values. The model captures short-term fluctuations effectively but may not fully account for long-term trends.

The below plot shows the Single Exponential Smoothing Forecast:

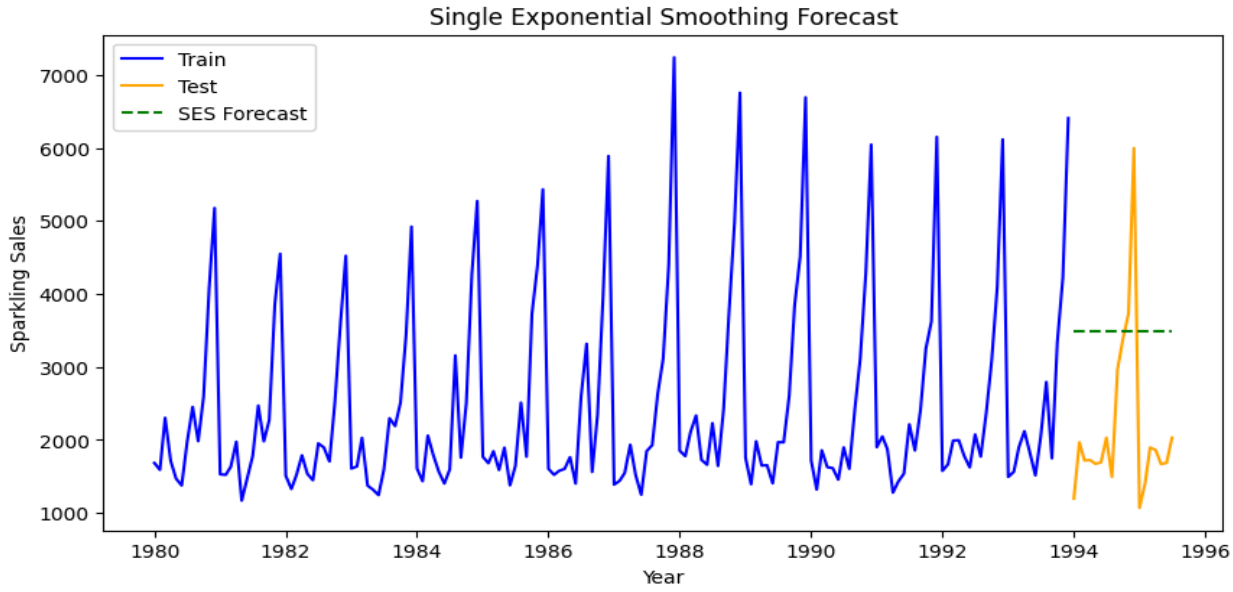

Figure 9. Single Exponential Smoothing Forecast

Interpretation:

The SES forecast (green dashed line) follows the general pattern of the test data but does not fully capture significant trends or seasonal effects. Since SES primarily smooths the data, it is most effective when there is no strong trend. However, given the increasing trend in sales over the years, more advanced methods may provide better accuracy.

3.4.2 Double Exponential Smoothing (Holt's Method)

Holt's Method, also known as Double Exponential Smoothing, extends Single Exponential Smoothing by adding a trend component. This method is useful when the time series data exhibits a trend but no seasonality. It smooths both the level (base value) and the trend separately using two different smoothing parameters:

- Alpha (Level Smoothing Parameter): Controls how much weight is given to recent observations.
- Beta (Trend Smoothing Parameter): Adjusts the trend component to adapt to data fluctuations.

The below plot shows the Double Exponential Smoothing Forecast:

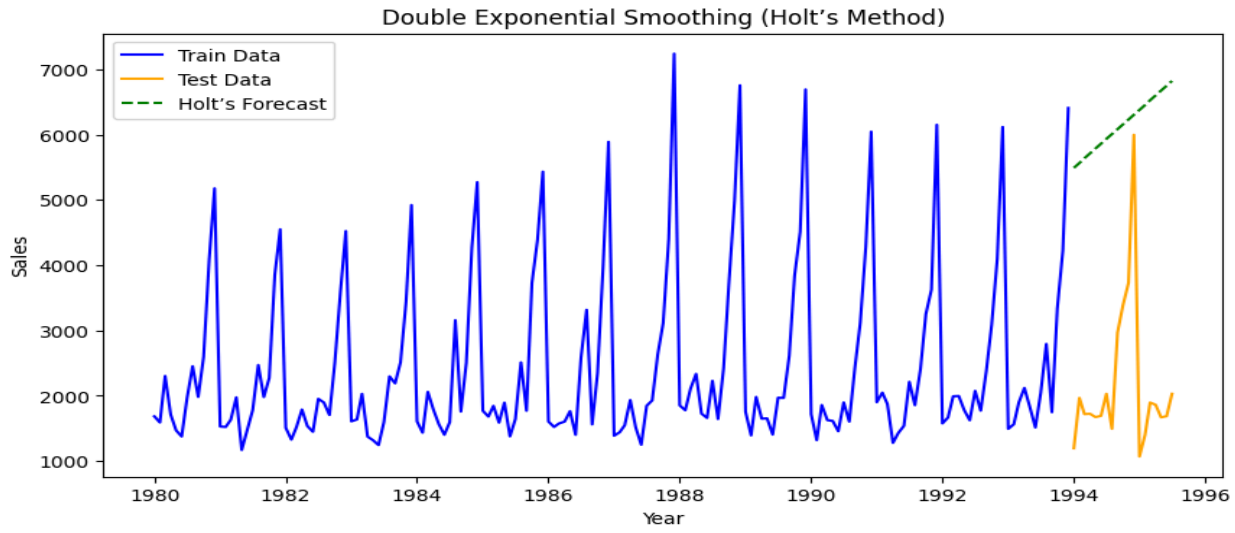

Figure 10. Double Exponential Smoothing Forecast

Interpretation:

The forecasted values (green dashed line) capture the overall trend in sales better than SES. The model effectively follows the upward sales pattern, making it more suitable for time series data with a clear trend. However, since it does not consider seasonality, it may not fully capture recurring seasonal fluctuations.

3.4.3 Triple Exponential Smoothing (Holt-Winters Method)

Holt-Winters Method, also known as Triple Exponential Smoothing, extends Holt's Method by incorporating a seasonal component. This makes it an ideal choice for time series data that exhibits both trend and seasonality.

It uses three smoothing parameters:

- Alpha (Level Smoothing): Adjusts the base value.
- Beta (Trend Smoothing): Controls how much weight is given to recent trends.
- Gamma (Seasonality Smoothing): Captures repeating seasonal patterns.

We use the multiplicative seasonal model because the seasonal variations change proportionally with the level of the data.

The below plot shows the Triple Exponential Smoothing Forecast:

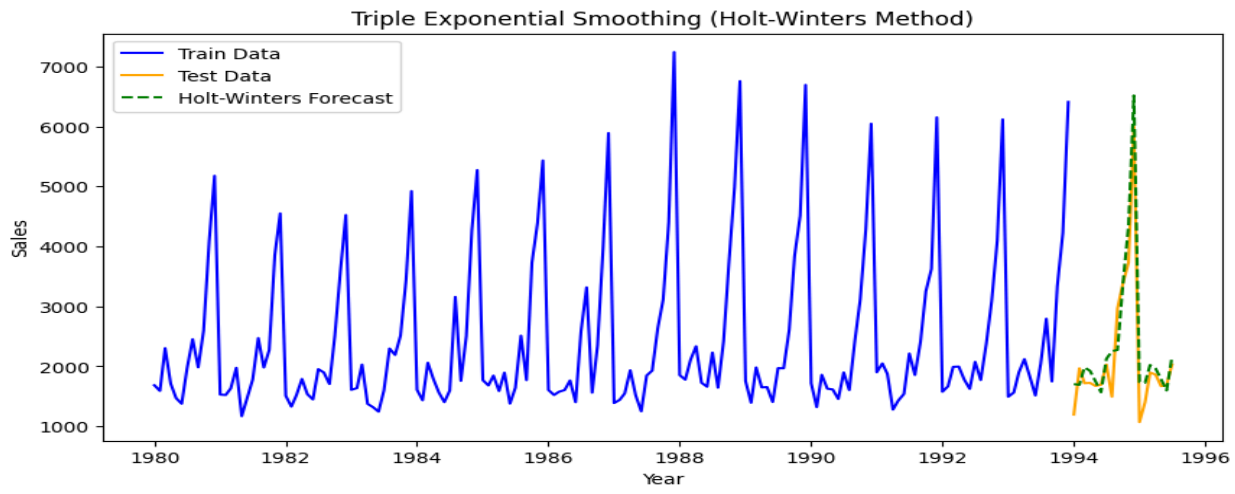

Figure 11. Triple Exponential Smoothing Forecast

Interpretation:

The forecasted values (green dashed line) effectively follow both the trend and seasonal fluctuations in the sales data. Unlike SES and Holt's Method, Holt-Winters captures recurring seasonal peaks and dips, making it more suitable for our dataset.

3.5 Model Evaluation using Mean Squared Error (MSE)

To assess the performance of the forecasting models, we utilized Mean Squared Error (MSE), which quantifies the average squared difference between actual and predicted values. A lower MSE indicates a more accurate model.

The MSE results for each model are as follows:

- Holt-Winters Method achieved the lowest MSE of 147,279.4, making it the best-performing model.
- Simple Average performed reasonably well, with an MSE of 1,331,513.0, ranking second.
- Linear Regression followed closely behind, with an MSE of 1,648,571.0, indicating moderate accuracy.
- Simple Exponential Smoothing had a higher error of 3,028,854.0, suggesting limitations in capturing trends.
- Moving Average resulted in an even higher MSE of 7,464,071.0, showing poor predictive capability.
- Holt's Method performed the worst, with an MSE of 17,247,000.0, indicating significant forecasting errors.

From this evaluation, we conclude that the Holt-Winters Method is the most effective forecasting model, as it produced the lowest MSE. Given its superior performance, we will proceed with this model for further analysis and forecasting.

4.Stationarity Assessment and Transformation

4.1 Assessing Stationarity in Time Series Data

For any time series forecasting model to provide reliable and meaningful results, it is crucial to ensure that the data is stationary. A time series is considered stationary when its statistical properties, such as mean, variance, and autocovariance, remain constant over time. If a time series is non-stationary, it can lead to inaccurate model estimations and poor forecasting performance.

To assess the stationarity of the given dataset, we employ both visual inspection and a statistical test.

Visual Inspection using Rolling Statistics

One of the simplest ways to check for stationarity is by plotting the rolling mean and rolling standard deviation over time. This approach helps in identifying trends and variations in the dataset.

- The rolling mean represents a smoothed version of the data over a specified window (12 months in this case), allowing us to observe any potential upward or downward trends.
- The rolling standard deviation provides insights into whether the variability of the dataset remains constant over time.

If the rolling mean and rolling standard deviation fluctuate significantly over time rather than remaining steady, the data is likely to be non-stationary.

The below plot shows the Rolling Statistics - Stationarity check:

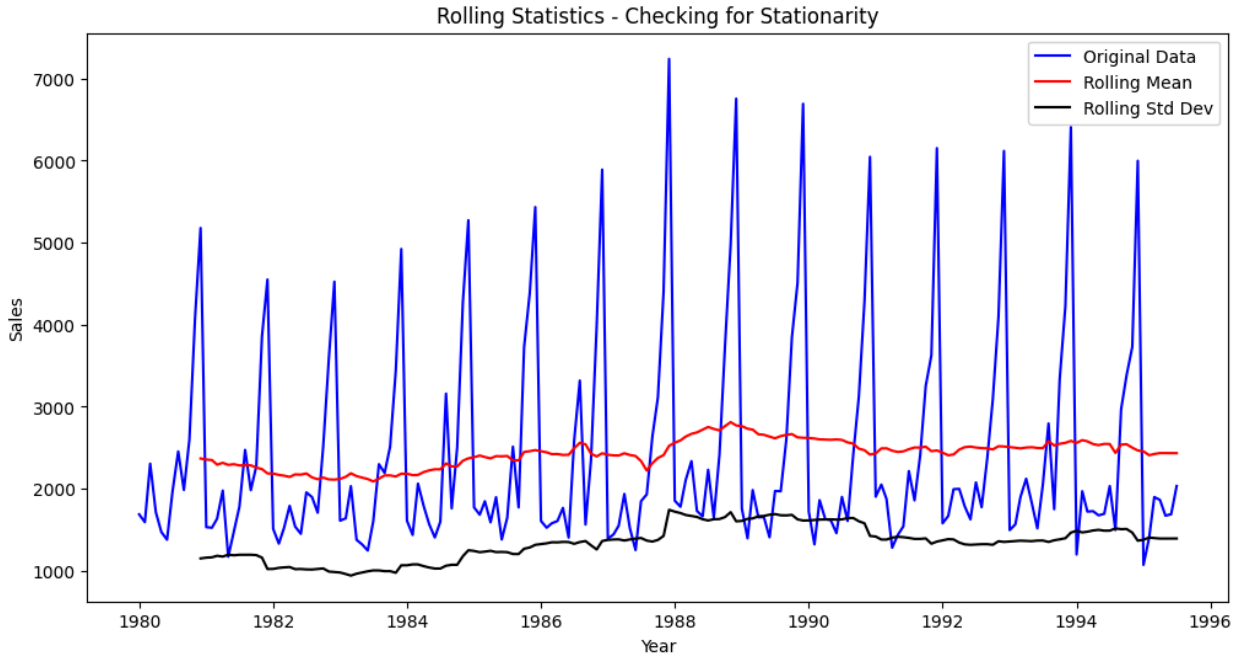

Figure 12. Rolling Statistics - Stationarity check

Interpretation :

The rolling statistics plot provides a visual representation of how the mean and standard deviation of the dataset change over time. In an ideal stationary series, these values should remain relatively constant.

- The rolling mean (red line) shows an upward trend, indicating the presence of a long-term increase in sales.
- The rolling standard deviation (Black line) also fluctuates over time, suggesting varying levels of dispersion in the dataset.
- The original time series (blue line) exhibits seasonal patterns with periodic spikes, reinforcing the presence of trends and seasonality.

These observations confirm that the dataset is non-stationary and must be transformed before applying forecasting models.

Augmented Dickey-Fuller (ADF) Test for Stationarity:

While visual inspection provides an initial understanding, a more formal way to confirm stationarity is through the Augmented Dickey-Fuller (ADF) test. This is a widely used statistical test that determines whether a given time series has a unit root, which indicates non-stationarity.

The ADF test works by testing the following hypotheses:

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

If the p-value obtained from the ADF test is less than 0.05, we reject the null hypothesis and conclude that the series is stationary. Conversely, if the p-value is greater than 0.05, we fail to reject the null hypothesis, implying that the data is non-stationary.

The results of the ADF test for our dataset are as follows:

- Test Statistic: -1.360
- P-Value: 0.601
- Critical Values: {'1%': -3.468, '5%': -2.878, '10%': -2.576}

Since the p-value is significantly higher than 0.05, we fail to reject the null hypothesis, confirming that the time series is non-stationary. This means that further transformations will be required to make the data suitable for forecasting models. The next step involves techniques such as differencing and log transformations to remove trends and make the series stationary.

4.2 Transforming Data to Achieve Stationarity

After identifying that the original time series data was non-stationary, transformations were applied to stabilize the mean and eliminate trends. First-order differencing was used to remove the trend component and make the data more suitable for time series modeling. This transformation ensures that fluctuations in the data remain consistent over time, reducing any systematic patterns.

The below plot shows Differenced Data Plot

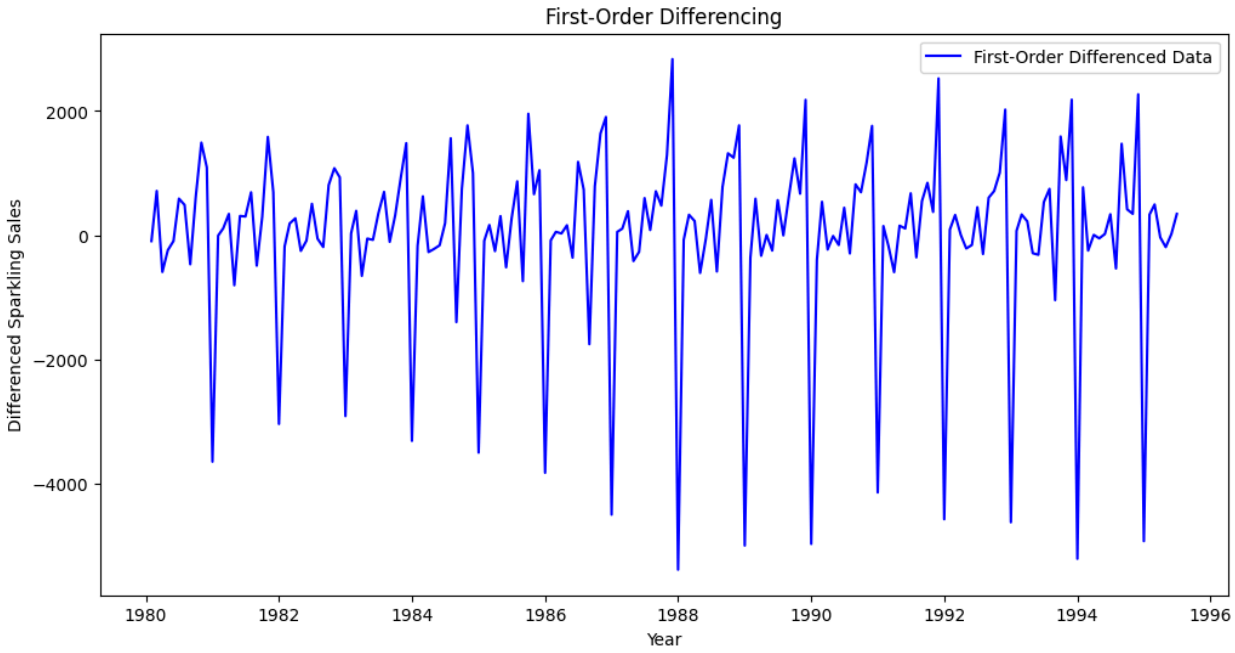

Figure 13. Differenced Data Plot

Interpretation :

The plot of the first-order differenced data exhibits a more stable pattern compared to the original time series. The noticeable trend present in the raw data has been eliminated, and the values now fluctuate around a constant mean. This suggests that the series is moving towards stationarity, which is a key requirement for effective time series forecasting.

ADF Test Results After First-Order Differencing

- Test Statistic: -45.05
- P-Value: 0.0
- Critical Values: {'1%': -3.468, '5%': -2.878, '10%': -2.576}

The results of the Augmented Dickey-Fuller (ADF) test confirm that the transformed data is now stationary. The test statistic is significantly lower than the critical values at all significance levels, and the p-value is 0.0, indicating strong evidence to reject the null hypothesis of non-stationarity. With the data now stationary, it is suitable for time series forecasting techniques that require stationarity, such as ARIMA and SARIMA models.

5. Stationary Data Model Building & Evaluation

5.1.ACF & PACF Analysis for AR, MA Identification:

To build an appropriate ARIMA model, it is crucial to determine the values of Auto-Regressive (AR) term (p) and Moving Average (MA) term (q). This can be done using the Autocorrelation Function (ACF) plot and the Partial Autocorrelation Function (PACF) plot.

ACF Plot Analysis

The Autocorrelation Function (ACF) plot helps identify the order of the Moving Average (MA) term (q) by showing how past values influence the present value over different lag periods.

The below plot shows ACF Plot:

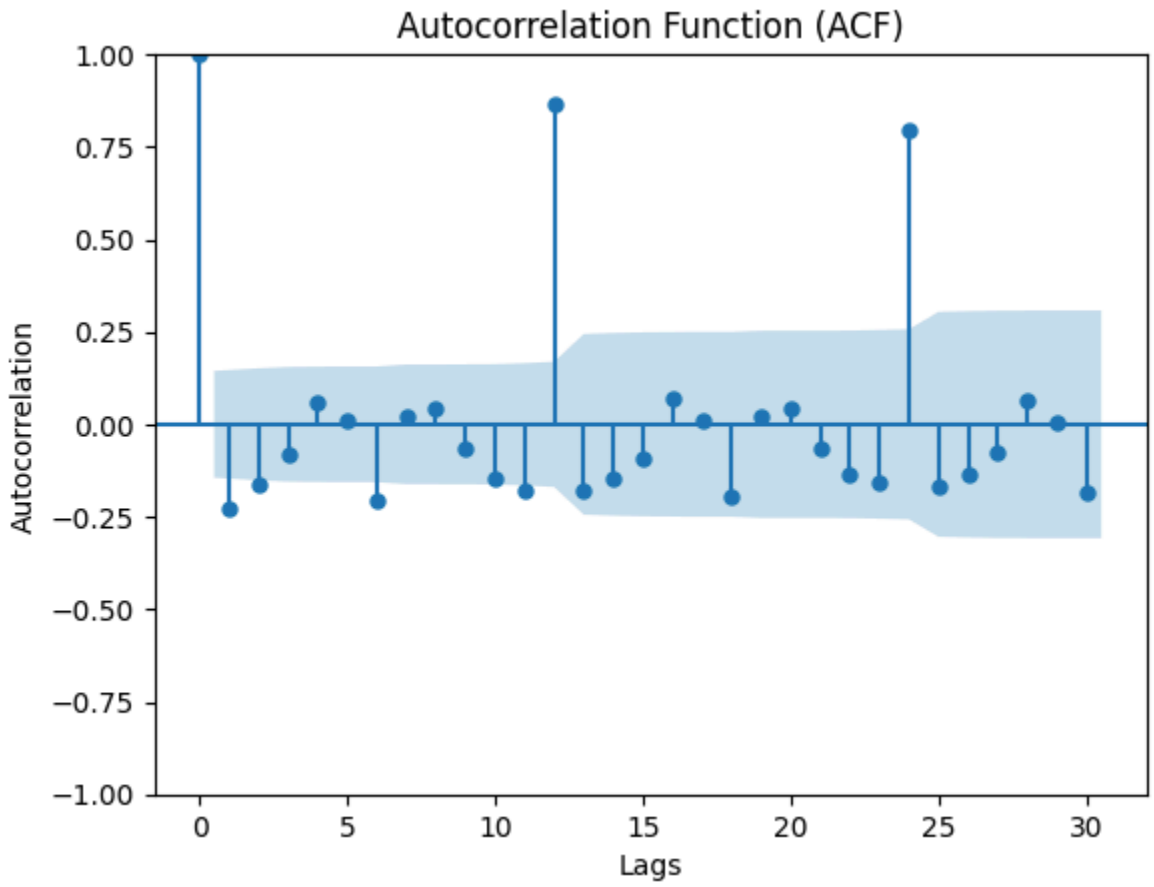

Figure 14. ACF Plot

Interpretation:

- The first spike at lag 0 is always at 1, representing the correlation of the series with itself.
- The blue-shaded region represents the confidence interval (CI), where fluctuations within this range are considered statistically insignificant.
- The significant spikes outside the CI indicate meaningful correlations at specific lags.
- In this case, we observe that the first two spikes beyond lag 0 (at lag 1 and 2) exceed the CI, suggesting that past errors up to lag 2 impact the current value.
- Based on this observation, we select q = 2 for the Moving Average term.

PACF Plot Analysis:

The Partial Autocorrelation Function (PACF) plot helps determine the order of the Auto-Regressive (AR) term (p) by measuring the direct influence of past observations while removing intermediate effects.

The below plot shows PACF Plot:

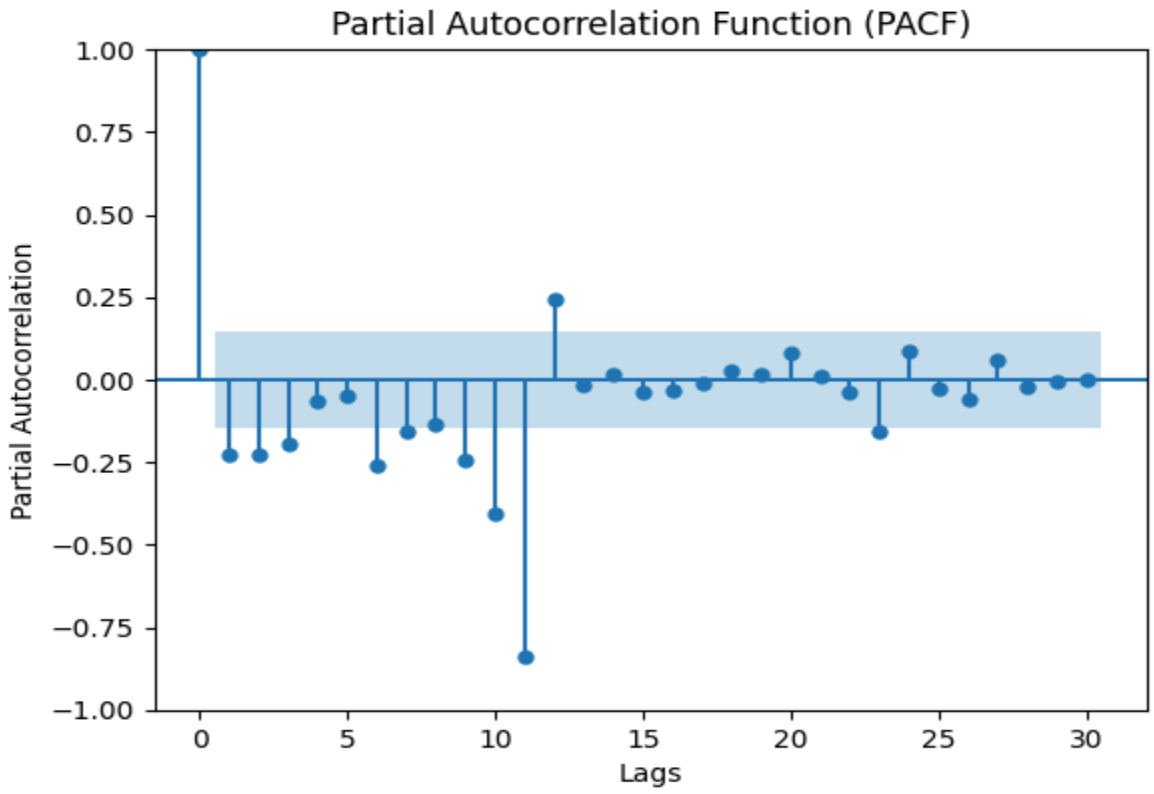

Figure 15. PACF Plot

Interpretation:

- As in the ACF plot, the first spike at lag 0 is always at 1.
- The blue-shaded region represents the confidence interval, where values within this range are statistically insignificant.
- The first three spikes beyond lag 0 (at lag 1, 2, and 3) exceed the CI, indicating a significant direct correlation of these past values with the present value.
- As a result, we determine p = 3 for the Auto-Regressive term.

5.2 Build Different ARIMA Models

5.2.1 Auto ARIMA Model

Auto ARIMA is an automated approach to selecting the optimal ARIMA parameters (p, d, q) based on statistical criteria like AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). The model iteratively evaluates different combinations of parameters to minimize the error and select the best-fit model.

The below plot shows Auto ARIMA Forecast:

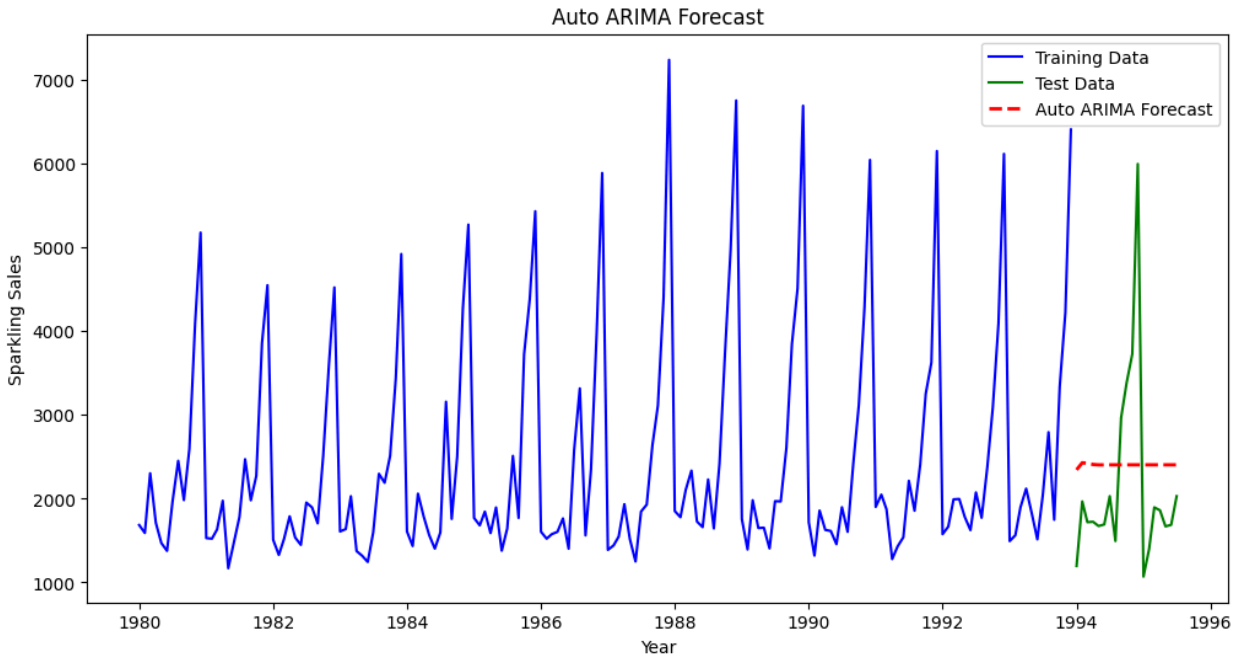

Figure 16. Auto ARIMA Forecast plot

Interpretation:

The Auto ARIMA model was used to predict future values based on the training dataset. The blue line represents the training data, while the green line represents the actual test data. The red dashed line denotes the Auto ARIMA forecast.

From the plot, it is observed that the Auto ARIMA forecast remains relatively constant over time, failing to capture the seasonal patterns present in the data. This suggests that Auto ARIMA may not have fully captured the seasonality or non-linearity in the dataset.

5.2.2 Manual ARIMA Model:

Unlike Auto ARIMA, the Manual ARIMA model requires selecting the (p, d, q) parameters manually. Based on the ACF & PACF plots, we determined the values:

- p (AR terms) = 3
- d (Differencing order) = 1
- q (MA terms) = 2

Using these parameters, we built a Manual ARIMA model and generated forecasts.

The below plot shows Manual ARIMA Forecast:

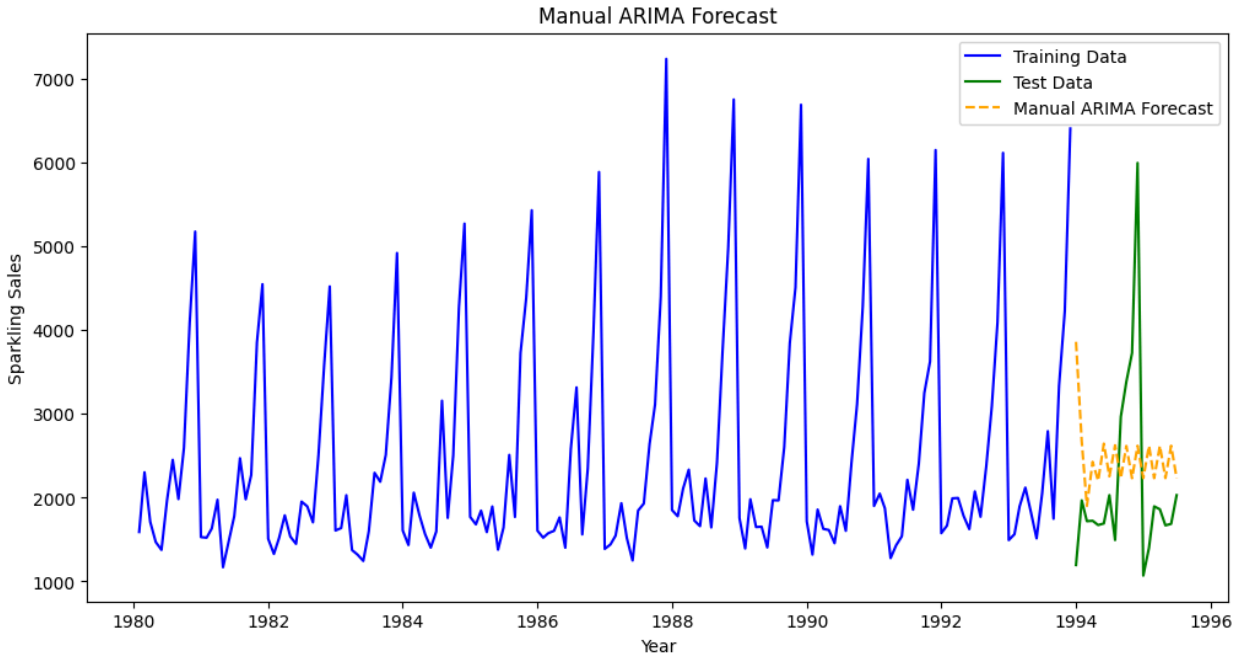

Figure 17. Manual ARIMA Forecast plot

Interpretation:

The above plot illustrates the Manual ARIMA model's forecast. Similar to the previous plot, the blue line represents the training data, while the green line indicates the test data. The forecasted values produced by the manually configured ARIMA model (ARIMA(3,1,2)) are depicted as an orange dashed line.

Compared to the Auto ARIMA forecast, the manually tuned model exhibits better responsiveness to fluctuations in the test data. The forecasted values show variability rather than a flat-line prediction, indicating that the manually chosen parameters (p=3, d=1, q=2) helped capture some underlying patterns in the time series. However, discrepancies remain, particularly in capturing the sharp seasonal peaks and extreme variations in the test data.

Further refinement, such as incorporating seasonal components using SARIMA or testing different AR and MA values, may be required to enhance forecast accuracy.

5.3 Build Different SARIMA Models

SARIMA (Seasonal AutoRegressive Integrated Moving Average) is an extension of the ARIMA model that incorporates seasonality, making it suitable for time series data that exhibits seasonal patterns. It is represented as SARIMA(p, d, q)(P, D, Q, s), where:

- p, d, q: Non-seasonal ARIMA parameters
- P, D, Q, s: Seasonal counterparts, where s is the seasonal period length

In this section, we explore two SARIMA modeling approaches: Auto SARIMA, which automatically determines optimal parameters, and Manual SARIMA, where parameters are selected based on statistical analysis.

5.3.1 Auto SARIMA Model

Auto SARIMA is an automated approach that selects the best SARIMA parameters based on statistical criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). It evaluates multiple combinations of seasonal and non-seasonal parameters to minimize forecast errors and optimize model performance.

The below plot shows Auto SARIMA Forecast:

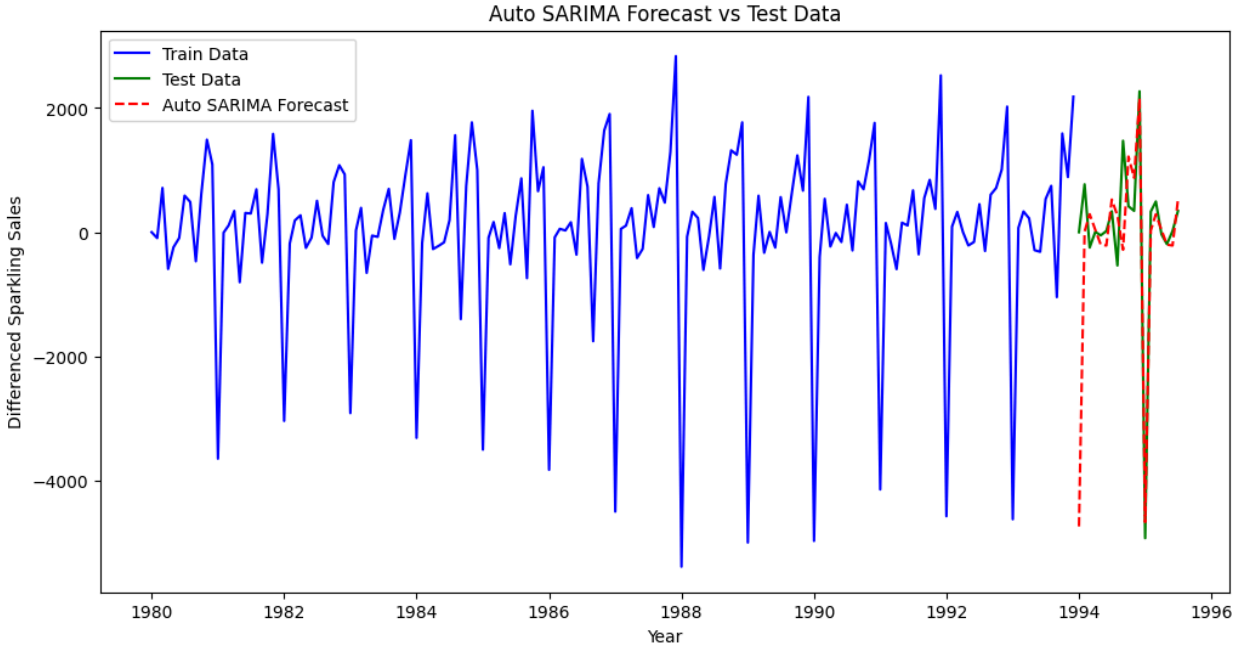

Figure 18. Auto SARIMA Forecast plot

Interpretation:

The Auto SARIMA model was applied to the dataset for forecasting future values. The plot displays:

- Blue line: Represents the training data.
- Green line: Represents the actual test data.
- Red dashed line: Represents the Auto SARIMA forecast.

From the visualization, it is evident that Auto SARIMA captures some of the seasonal fluctuations present in the data. However, certain variations in the actual test data are not fully reflected in the forecast. This indicates that while the model accounts for seasonality to some extent, additional tuning may be necessary for improved accuracy. The results suggest that a manual SARIMA approach could be explored to fine-tune the seasonal parameters for a better fit.

5.3.2 Manual SARIMA Model

In contrast to Auto SARIMA, the Manual SARIMA model involves manually selecting the best-fit (p, d, q) and (P, D, Q, s) parameters based on statistical techniques, including:

- ACF (AutoCorrelation Function) and PACF (Partial AutoCorrelation Function) analysis
- Stationarity testing (e.g., Augmented Dickey-Fuller test)
- Seasonal decomposition

By manually determining these values, we can fine-tune the model to better capture the underlying seasonal and trend patterns in the dataset.

Stationarity Testing and Differencing

To ensure stationarity, the dataset was subjected to seasonal differencing. The Augmented Dickey-Fuller (ADF) test was used to validate stationarity. Initially, the data was found to be non-stationary after the first round of seasonal differencing (D=1). A second round of seasonal differencing (D=2) was applied, which successfully rendered the data stationary, as confirmed by the ADF test.

Manual SARIMA Model Implementation

The Manual SARIMA model was implemented using the following parameters:

- Non-seasonal order (p, d, q): (3, 1, 2)
- Seasonal order (P, D, Q, s): (2, 2, 2, 12)

These parameters were chosen based on ACF and PACF analysis, ensuring the model effectively captures both seasonal and non-seasonal components of the data.

The below plot shows Manual SARIMA Forecast Plot:

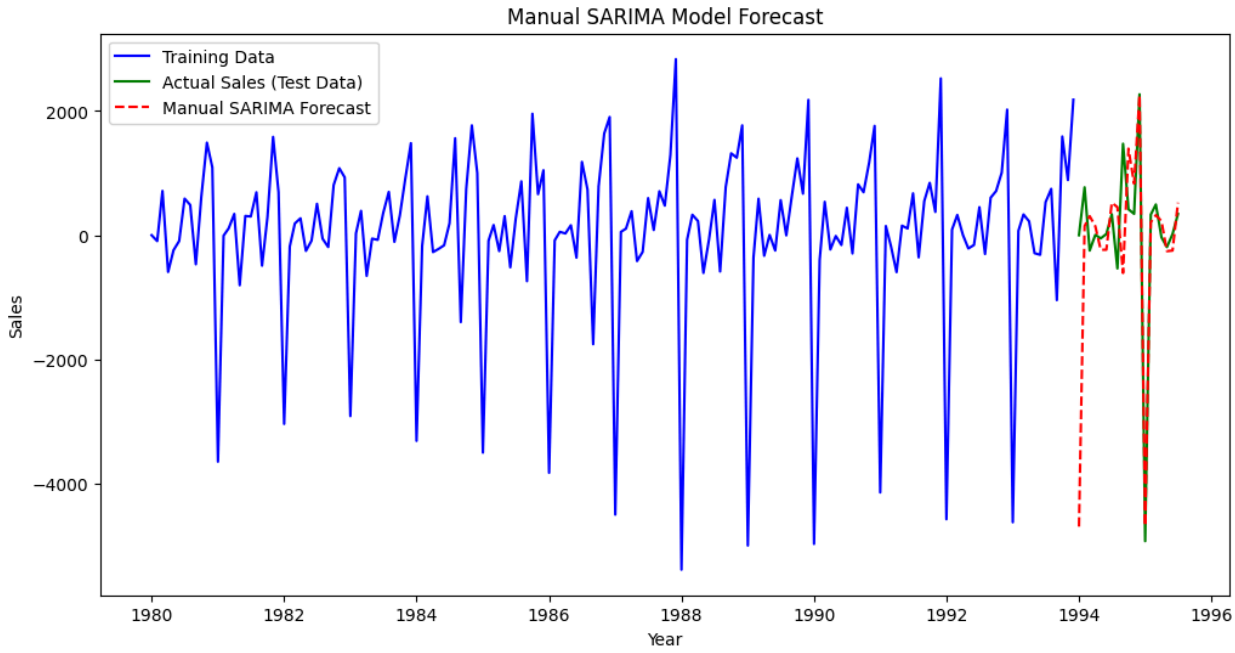

Figure 19. Manual SARIMA Forecast Plot

Interpretation:

The plot showcases the results of the manually tuned SARIMA model:

- Blue line: Represents the training data.
- Green line: Represents the actual test data.
- Red dashed line: Represents the Manual SARIMA forecast.

Both Auto SARIMA and the manually tuned SARIMA model produce similar forecasts, with only minor differences in their predictions. Auto SARIMA already provides a well-optimized solution by automatically selecting the best parameters, making it a convenient and efficient choice. This indicates that for datasets with well-defined seasonal patterns, automated approaches like Auto SARIMA can be just as effective as manual selection in capturing trends and variations.

5.4 Evaluating the Performance of ARIMA and SARIMA Models

To evaluate the accuracy of the different ARIMA and SARIMA models, we use the Root Mean Squared Error (RMSE) metric. RMSE measures the average magnitude of errors in the predictions, with lower values indicating better model performance. The RMSE values for all four models are analyzed to determine which model provides the most accurate forecasts.

RMSE Results

The RMSE values for the ARIMA and SARIMA models are as follows:

- Auto SARIMA: 1222.92
- Manual SARIMA: 1249.77
- Auto ARIMA: 2707.56
- Manual ARIMA: 2778.79

Interpretation:

- The Auto SARIMA model has the lowest RMSE of 1222.92, followed closely by the Manual SARIMA model with an RMSE of 1249.77.
- In contrast, the Auto ARIMA model has an RMSE of 2707.56, while the Manual ARIMA model has the highest RMSE at 2778.79.

The results indicate that both SARIMA models significantly outperform ARIMA models, as they effectively capture the seasonality present in the dataset. Among them, the Auto SARIMA model is the most accurate. The Auto ARIMA and Manual ARIMA models show poor performance, suggesting that ARIMA models are not well-suited for this dataset. Given these findings, the Auto SARIMA model will be used for final forecasting and business recommendations.

6. Model Comparison and Final Forecasting

After building multiple forecasting models, it is essential to compare their performance and determine which model best captures the underlying patterns in the data. The evaluation is primarily based on the Root Mean Squared Error (RMSE), which measures the difference between the predicted and actual values. A lower RMSE indicates a better-performing model. In this section, we will analyze the results of all the models, select the most accurate one, and use it to generate the final forecast.

6.1 Overall Model Performance Comparison

The RMSE values for all the models built are as follows:

- Holt-Winters Method: 383.77
- Simple Average: 1153.91
- Auto SARIMA: 1222.92
- Manual SARIMA: 1249.77
- Linear Regression: 1283.97
- Simple Exponential Smoothing: 1740.36
- Auto ARIMA: 2707.56
- Moving Average: 2732.05
- Manual ARIMA: 2778.79
- Holt's Method: 4152.95

Interpretation:

- The Holt-Winters Method (383.77), Auto SARIMA (1222.92), and Manual SARIMA (1249.77) achieved the lowest RMSE values, making them the most accurate models.
- Conversely, Holt's Method (4152.95) and Manual ARIMA (2778.79) exhibited the highest errors, suggesting that they were less effective in capturing the underlying patterns of the data.

6.2 Selection of the Best Model

Among all the models evaluated, the Holt-Winters Method demonstrated the best performance with the lowest RMSE of 383.77. This model effectively captures trend and seasonality, making it well-suited for forecasting in time-series data with recurring patterns. Although the SARIMA models (Auto and Manual) also exhibited strong predictive power, the Holt-Winters Method outperformed them with a lower error margin. Additionally, Holt-Winters is computationally efficient and does not require extensive parameter tuning, making it a more practical choice for forecasting.

Given these observations, the Holt-Winters Method is selected as the best forecasting model due to its superior accuracy, ability to capture seasonality, and ease of implementation. This model provides a reliable framework for generating future forecasts with minimal error, ensuring more informed decision-making.

6.3. Optimizing the Best Model & Forecasting Sales for the Next 12 Months:

Based on the evaluation of multiple forecasting models, the Holt-Winters Method was identified as the best-performing model due to its lowest RMSE of 383.77. This model was rebuilt using the entire dataset to generate a 12-month forecast for the upcoming period from August 1995 to July 1996.

The forecasted sales values exhibit seasonal variations, consistent with historical trends. Sales are expected to peak in December 1995 (6118.52 units) before experiencing a decline in the first quarter of 1996, reflecting a seasonal sales pattern. The lowest forecasted sales occur in January 1996 (1262.65 units), followed by a gradual stabilization in subsequent months.

12-Month Forecast Plot

The following plot visualizes both historical sales data and forecasted sales for the next 12 months using the Holt-Winters Method.

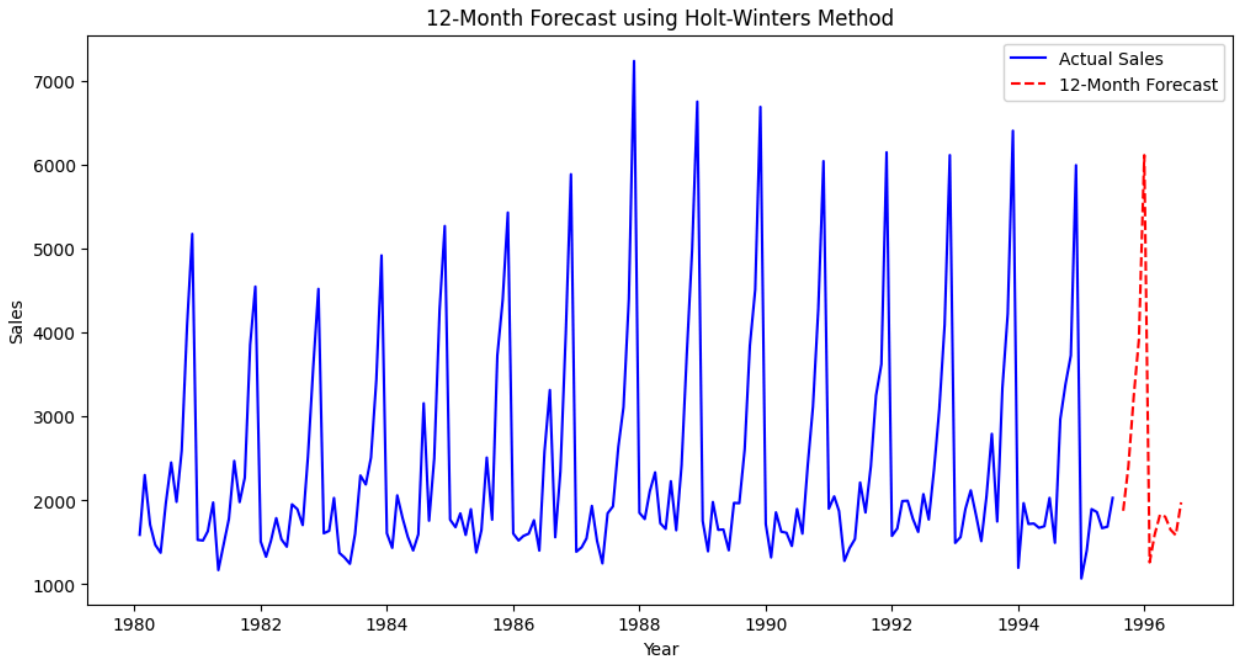

Figure 20. 12-Month Forecast Plot

Interpretation of the Forecast

The plot above illustrates the historical sales data (blue line) and forecasted values (red dashed line) for the next 12 months. The following key insights can be observed:

1. Seasonal Pattern: The forecast aligns with previous seasonality, showing peaks in December and declines in early months of the year.
2. December Sales Surge: A significant spike in sales is expected in December 1995 (6118.52 units), indicating a strong seasonal demand, likely due to holiday or festive season effects.
3. Post-Holiday Decline: Sales drop sharply in January 1996 (1262.65 units), consistent with historical patterns where demand slows after peak months.
4. Stable Growth: From February to July 1996, sales fluctuate moderately, suggesting a period of stable demand.

This forecast enables businesses to make data-driven decisions regarding inventory planning, marketing strategies, and resource allocation. By anticipating demand fluctuations, the company can optimize its operations and capitalize on peak sales periods.

7. Actionable Insights & Recommendations

7.1.Key Observations from Forecast

The 12-month sales forecast using the Holt-Winters Method provides a clear projection of expected sales trends. The forecast captures the strong seasonal pattern observed in historical data, indicating that sales peak at specific intervals. Additionally, a notable increase in projected sales is observed towards the end of the forecast period, suggesting a potential growth phase. However, some fluctuations remain, which may be influenced by external market conditions or operational factors. These insights allow for strategic planning and proactive measures to sustain business growth.

Business Implications

- The forecasted peak sales periods can help businesses optimize inventory and staffing to meet anticipated demand.
- Identifying seasonal fluctuations enables better financial planning, ensuring adequate cash flow management.
- The potential growth trend towards the end of the forecast period suggests an opportunity for market expansion and strategic investments.
- Understanding demand patterns can aid in refining marketing strategies to capitalize on high-demand periods.

7.2.Actionable Recommendations

1. Inventory Management: Ensure adequate stock levels before peak sales periods to prevent shortages or delays.
2. Marketing Strategy: Align promotional campaigns with high-sales periods to maximize revenue potential.
3. Workforce Planning: Adjust staffing levels to accommodate seasonal fluctuations and maintain service quality.
4. Supply Chain Optimization: Strengthen relationships with suppliers to ensure a steady flow of goods during high-demand months.
5. Data-Driven Decision Making: Continuously refine forecasting models with updated sales data to improve accuracy and adaptability.

7.3.Conclusion:

The 12-month sales forecast generated using the Holt-Winters Method offers valuable insights into future sales trends, capturing seasonal peaks and potential growth phases. By identifying these trends, businesses can proactively plan for peak periods, optimize inventory, adjust staffing levels, and manage cash flow effectively. The forecast also highlights opportunities for market expansion and strategic investments, particularly towards the end of the period. Additionally, the insights derived from this analysis can inform marketing strategies, ensuring promotions align with high-demand periods. In conclusion, the forecast provides actionable recommendations for businesses to enhance operational efficiency, refine decision-making, and sustain long-term growth through continuous model refinement and data-driven strategies.