Project 9

- Time Series Forecasting -

- (Demand Forecasting & Predictive Analytics for Rose Wine Sales using ARIMA, SARIMA & Holt-Winters) -

1. Introduction

1.1 Problem Definition

1.1.1 Introduction

Wine sales are influenced by a variety of factors, including consumer preferences, seasonal demand, and broader economic conditions. ABC Estate Wines, a prominent name in the wine industry, has collected extensive historical sales data spanning the 20th century. This dataset presents an opportunity to analyze past sales trends and patterns specific to Rose wine, allowing for data-driven decision-making. By utilizing advanced analytics and forecasting techniques, the goal is to derive valuable insights that can aid in strategic planning and future sales optimization.

1.1.2 Business Problem

In the highly competitive wine industry, understanding historical sales trends is essential for sustained growth. ABC Estate Wines offers a range of wine varieties, each exhibiting distinct demand patterns shaped by factors such as seasonality, consumer behavior, and market dynamics. Rose wine, in particular, has seen fluctuating demand over the years. Without a clear understanding of these variations, challenges arise in forecasting demand, managing inventory, and optimizing production schedules. Inaccurate predictions can result in overstocking, lost sales opportunities, or inefficient resource allocation, ultimately impacting profitability.

1.1.3 Objective

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

- Detect seasonal patterns and long-term sales trends specific to Rose wine.
- Identify key drivers influencing fluctuations in sales.
- Build forecasting models to predict future demand.
- Provide actionable insights to optimize inventory and production planning.

By leveraging historical data and predictive analytics, ABC Estate Wines can make well-informed decisions, enhance operational efficiency, and strengthen its market position within the wine industry.

1.2. Data Background and Contents

1.2.1. Dataset Overview

The dataset consists of historical monthly sales data for Rose 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.
- Rose: Represents the number of Rose wine units sold in that particular month.

There are no missing values in the YearMonth column, ensuring completeness for analysis. However, the Rose column has two missing values, which will require appropriate handling before modeling.

The YearMonth column is currently in an object format and will need to be converted to a datetime format to facilitate time series analysis. The dataset captures the long-term sales trends and seasonal patterns of Rose wine over time.

1.2.2. Statistical Summary

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

- Mean Sales: 90.39 units per month
- Standard Deviation: 39.17 units, indicating significant variation in monthly sales
- Minimum Sales: 28 units in the lowest-selling month
- Maximum Sales: 267 units in the highest-selling month
- 25th Percentile (Q1): 63 units, meaning 25% of the months had sales below this value
- Median (Q2/50th Percentile): 86 units, representing the middle of the sales distribution
- 75th Percentile (Q3): 112 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 267 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

Given that the dataset is time-dependent, structuring it as a proper time series is essential for accurate analysis. This involves converting the date column into a standard datetime format and setting it as an index to facilitate time-based operations. Once structured, the dataset is explored through summary statistics and visualizations to detect trends, patterns, or anomalies.

By establishing a structured time series format, the dataset becomes suitable for deeper exploratory analysis, decomposition, and forecasting, forming the basis for data-driven decision-making.

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, the sales data is analyzed through different visualizations to gain insights into its structure.

1. Monthly Rose Sales Over the Years

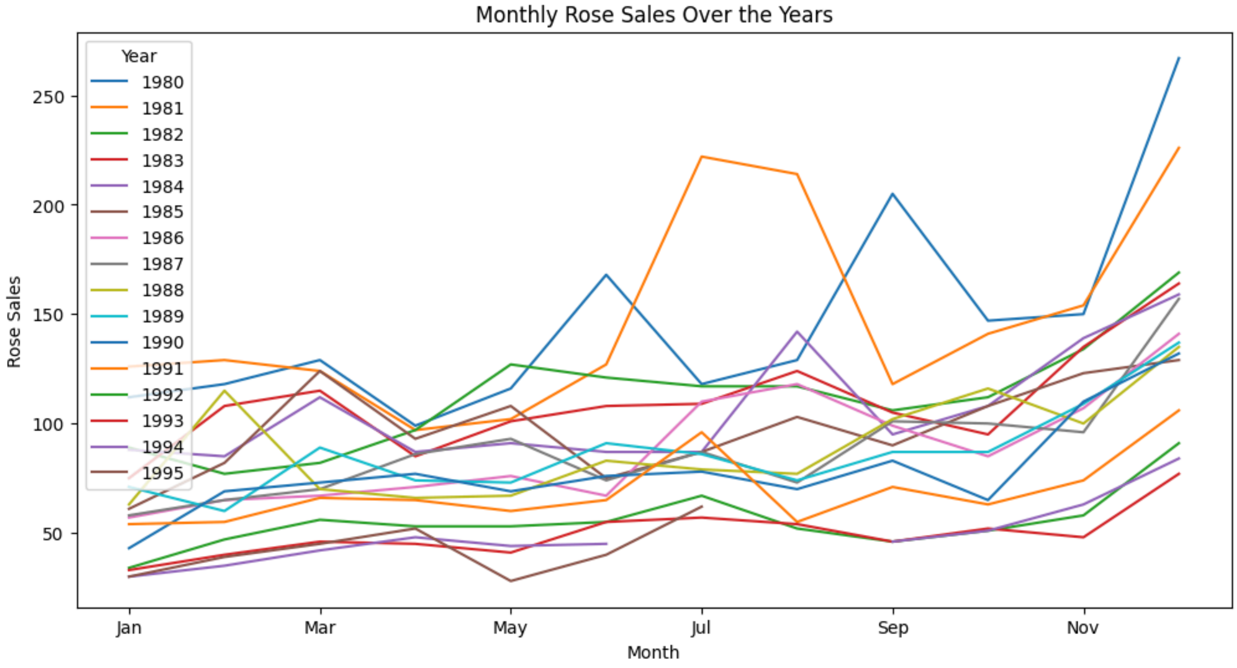

Figure 1. Monthly Rose Sales Over the Years

Interpretation:

The plot illustrates the monthly sales trend of Rose wine over multiple years. A clear seasonal pattern is observed, with relatively stable sales in the first half of the year, followed by a significant rise in sales towards the end of the year. The months of November and December consistently show the highest sales volumes across different years, indicating strong seasonal demand, likely due to holiday and festive purchases. While some years exhibit more fluctuations, the overall trend remains consistent, reinforcing the recurring seasonal nature of sales.

2. Yearly Rose Sales Trend by Month

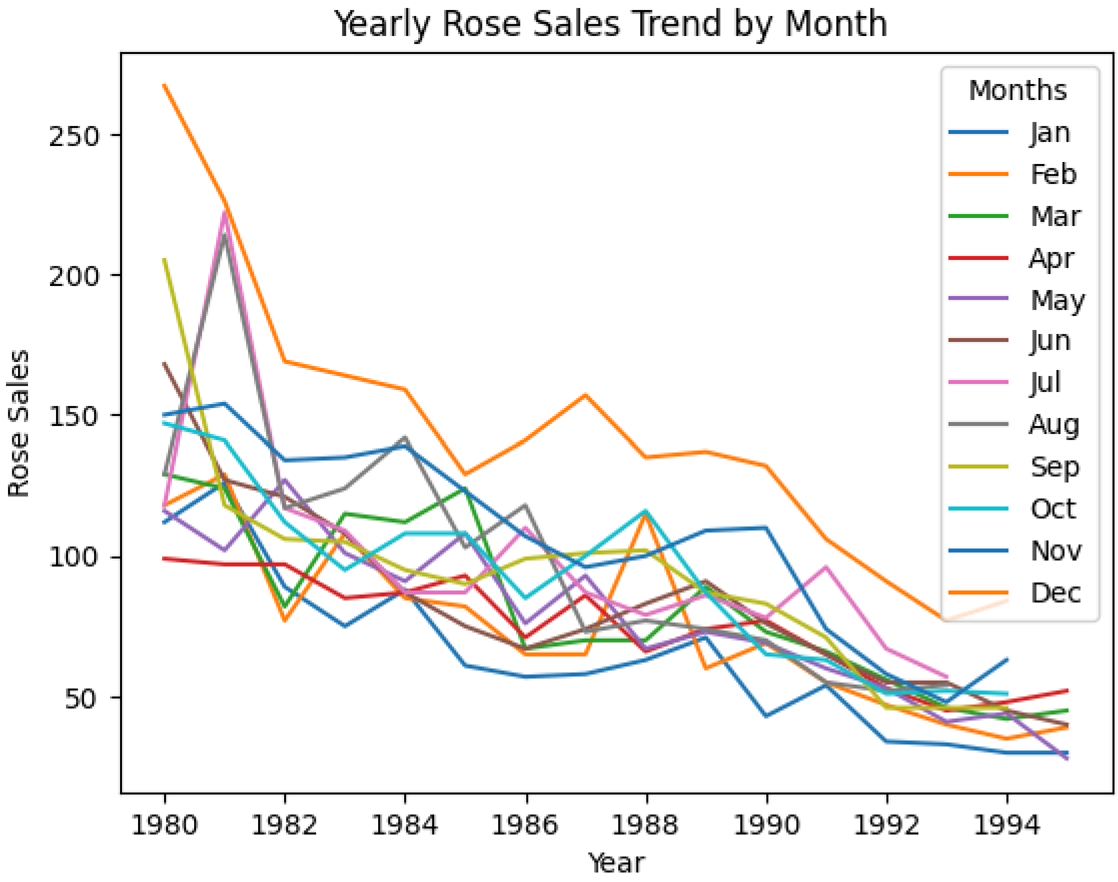

Figure 2. Yearly Rose sales Trend by Month

Interpretation:

This visualization showcases the annual trend in Rose wine sales for each month. The sales figures tend to decline gradually over the years, with noticeable peaks in certain months. The early years (1980s) have significantly higher sales compared to later years, suggesting a possible decline in market demand or a shift in consumer preferences. Additionally, months like November and December continue to show increased sales across years, further highlighting the strong seasonality present in the dataset.

3. Box Plot of Yearly Sales Distribution

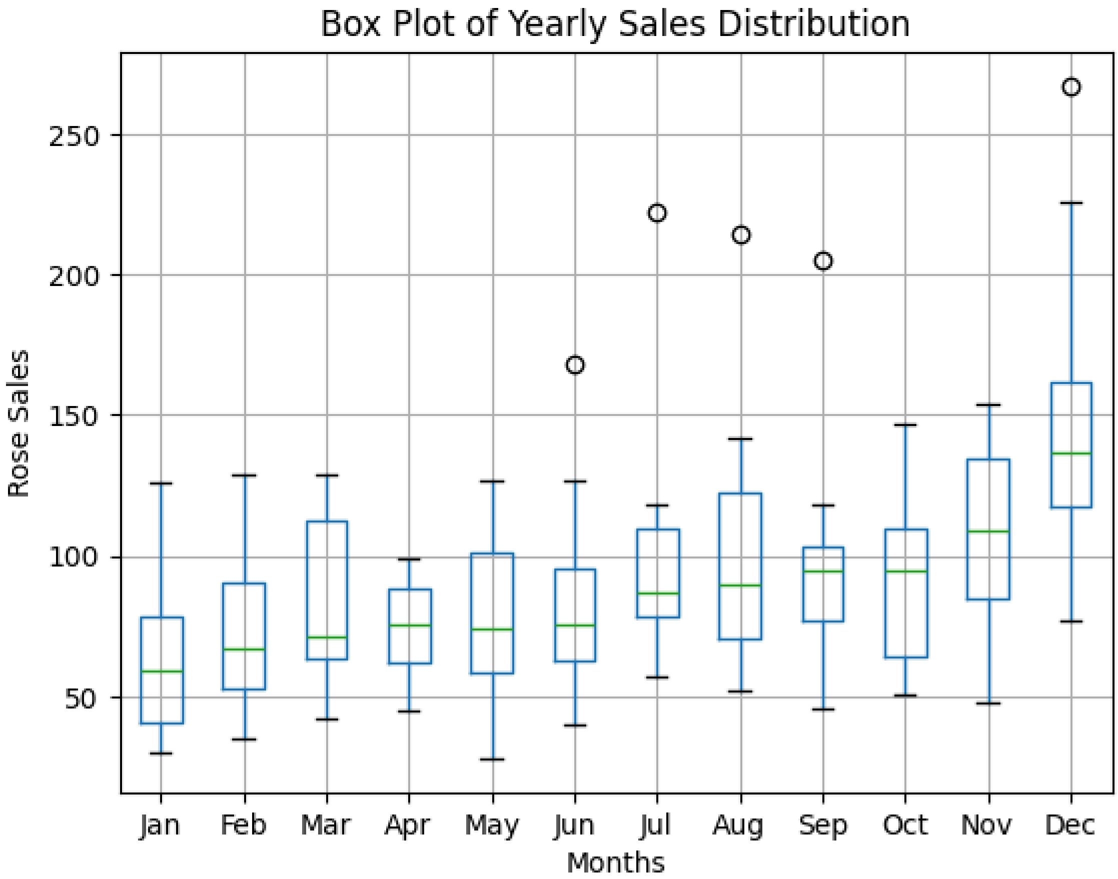

Figure 3. Box Plot of Yearly Sales Distribution

Interpretation:

The box plot illustrates the distribution of Rose wine sales across different months. The median sales values indicate a gradual increase towards the end of the year, with December showing the highest sales levels and variability. The interquartile range (IQR) expands significantly for the later months, reflecting increased fluctuations in demand. The presence of outliers in some months, particularly in the mid-year period, suggests occasional extreme variations in sales, possibly due to promotional events or external factors influencing purchasing behavior.

Insights from Exploratory Data Analysis:

- Sales exhibit a strong seasonal pattern, with peaks occurring consistently toward the end of each year.
- The early years demonstrate higher sales levels compared to later years, indicating potential market shifts.
- Box plot analysis reveals increasing variability in sales, with notable spikes in specific months.
- The presence of outliers suggests occasional demand fluctuations that may be driven by promotional efforts or external economic factors.

This exploratory analysis provides a solid foundation for further time series modeling, helping to identify patterns that can be leveraged for forecasting future sales trends.

1.5. Handling Missing Values

In time series forecasting, handling missing values is crucial to maintaining model accuracy. Upon examining the Rose Sales dataset, missing values were identified in July and August 1994. Since these missing values were within historical data (not future months), they needed to be imputed before applying time series decomposition.

To address this, we applied forward fill imputation, which propagates the last available observation to fill in missing values. This method ensures data continuity and prevents distortions in the decomposition process.

Missing Value Analysis:

- Columns affected: Rose_Sales
- Missing entries: 2 (July 1994, August 1994)
- Imputation method used: Forward fill

After handling the missing values, the cleaned dataset was used for further time series analysis and forecasting.

1.6 Perform Time Series Decomposition

Time series decomposition is an essential technique that helps in breaking down a time series into its core components: trend, seasonality, and residuals. This decomposition provides insights into the structure of the data, making it easier to build accurate forecasting models.

For the Rose Sales dataset, a multiplicative decomposition was chosen since the seasonal variations appear to be proportional to the overall sales trend. This approach allows the seasonal fluctuations to scale with the data, making it a more appropriate choice compared to an additive decomposition.

Decomposition of the Rose Sales Time Series

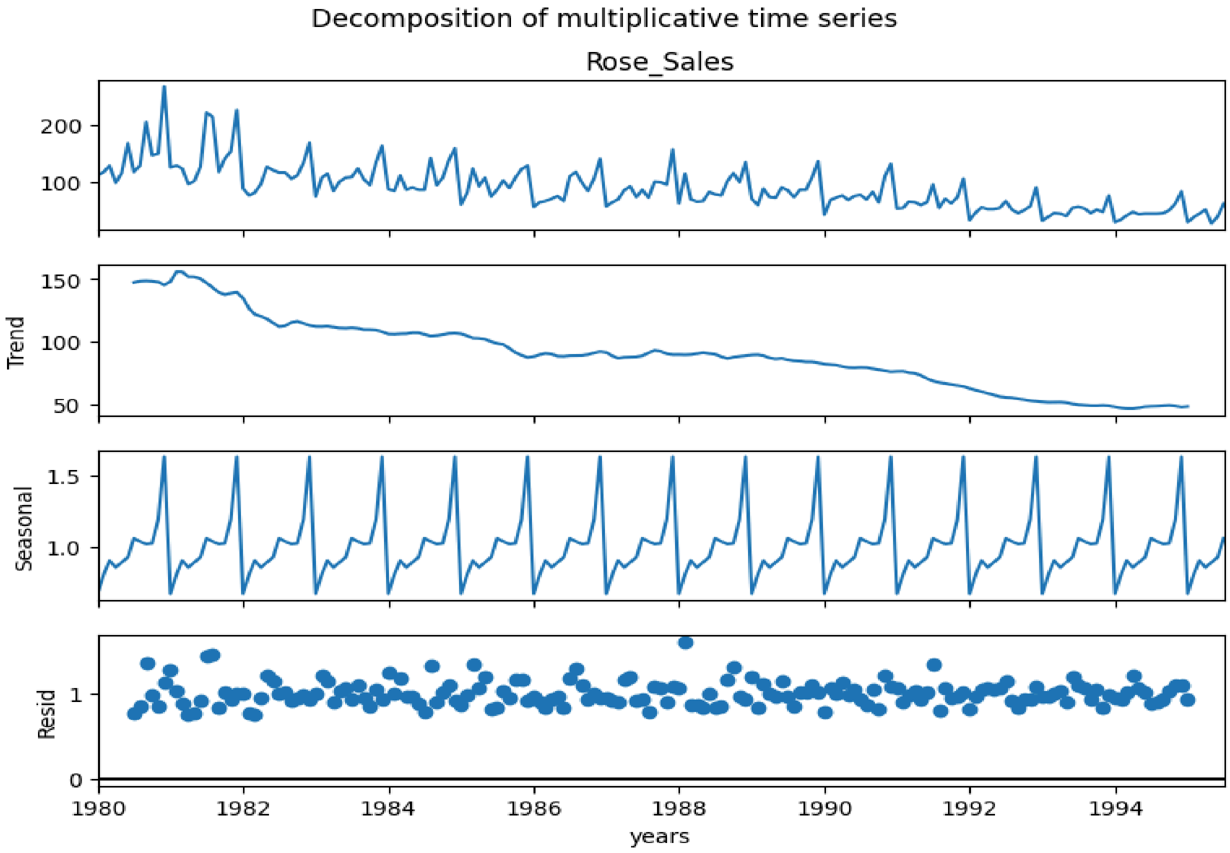

Figure 4. Decomposition of the Rose Sales Time Series

Interpretation:

- Trend Component: The trend shows a declining pattern in sales over time, suggesting that Rose wine sales have decreased steadily from the early 1980s to the mid-1990s.
- Seasonality Component: The seasonal component exhibits a consistent recurring pattern, confirming that sales follow a yearly cycle with peaks and troughs occurring at fixed intervals. This indicates strong seasonal demand, likely influenced by consumer purchasing behaviors during specific months.
- Residual Component: The residuals appear to be randomly distributed, implying that most of the systematic variations in the data have been captured by the trend and seasonal components. This randomness suggests that external factors or unpredictable events contribute to minor fluctuations in sales.

By decomposing the time series, we gain a clearer understanding of its structure, helping us select appropriate forecasting techniques for predicting future sales trends.

2. Data Pre-processing

2.1 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 ensures that missing values have been appropriately addressed and allows us to observe sales trends over time.

The plot below presents the Yearly Rose Sales Trend after data preprocessing:

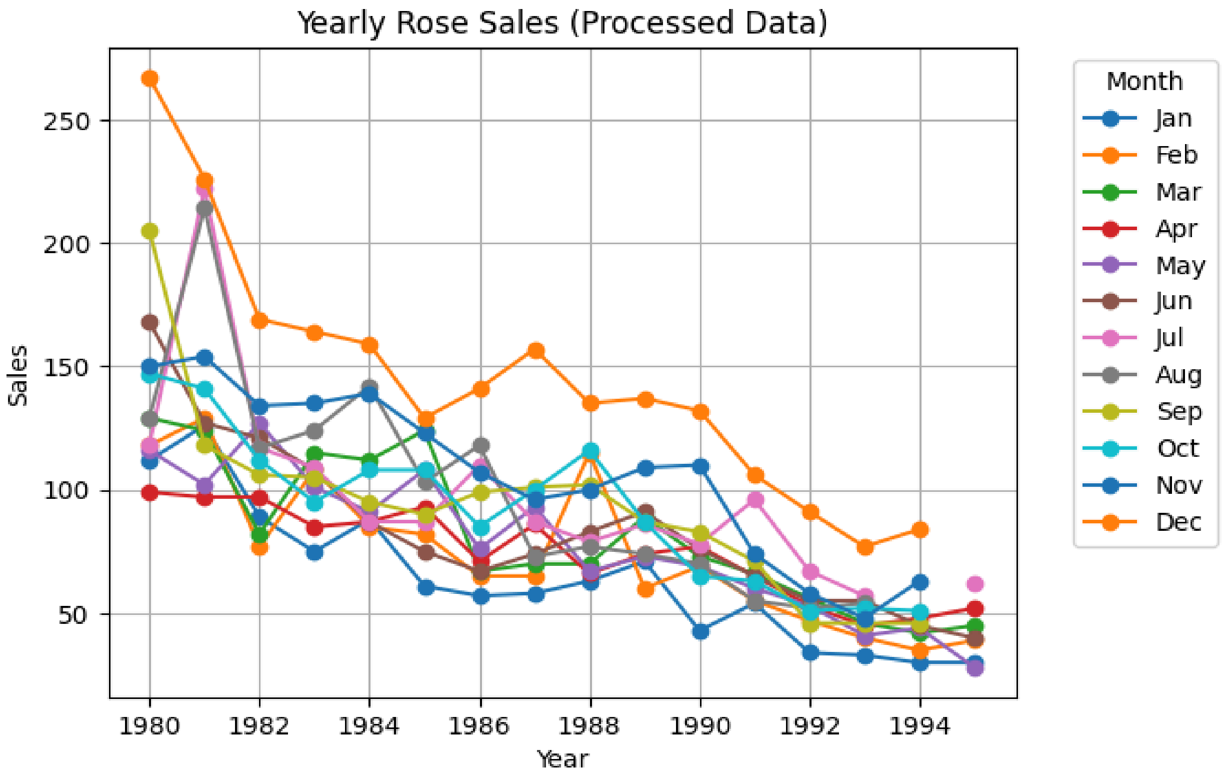

Figure 5. Yearly Rose Sales (Processed Data)

Interpretation:

This line plot illustrates the yearly Rose Sales trend after handling missing values. The overall pattern shows a declining trend in sales, suggesting a gradual decrease in demand over time. Despite the downward trend, seasonal variations remain evident, with noticeable peaks occurring at regular intervals. This confirms the presence of a seasonal effect, which will be considered in forecasting models.

2.2. Train-Test Split

To develop an effective forecasting model, we need to split the dataset into training and testing sets. The training set is used to train 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. To align with real-world forecasting scenarios, we divide the dataset as follows:

- Training Set: Data before January 1994
- Test Set: Data from January 1994 onward

This method mimics practical forecasting conditions, where future values are unknown during model training. By assessing model performance on the test set, we can determine how well it generalizes to unseen data.

Training Set Summary:

- Total Entries: 168
- Time Range: January 1980 – December 1993
- Non-null Data: 100%

Test Set Summary:

- Total Entries: 19
- Time Range: January 1994 – July 1995
- Non-null Data: 100%

This structured split ensures that our model is trained on historical data and evaluated on a separate test period, allowing for a robust performance assessment before making future sales predictions.

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 Rose Sales. Each model follows a unique approach to capturing 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:

- Linear Regression – A fundamental predictive model that identifies linear relationships between time and sales.
- Simple Average – A naive method that forecasts future values based on the average of past observations.
- Moving Average – A smoothing technique that reduces short-term fluctuations to highlight underlying trends.
- 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 (Rose 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 a specific period and predicted sales for the future period. The model estimates future sales based on the established trend, helping in understanding long-term growth patterns.

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

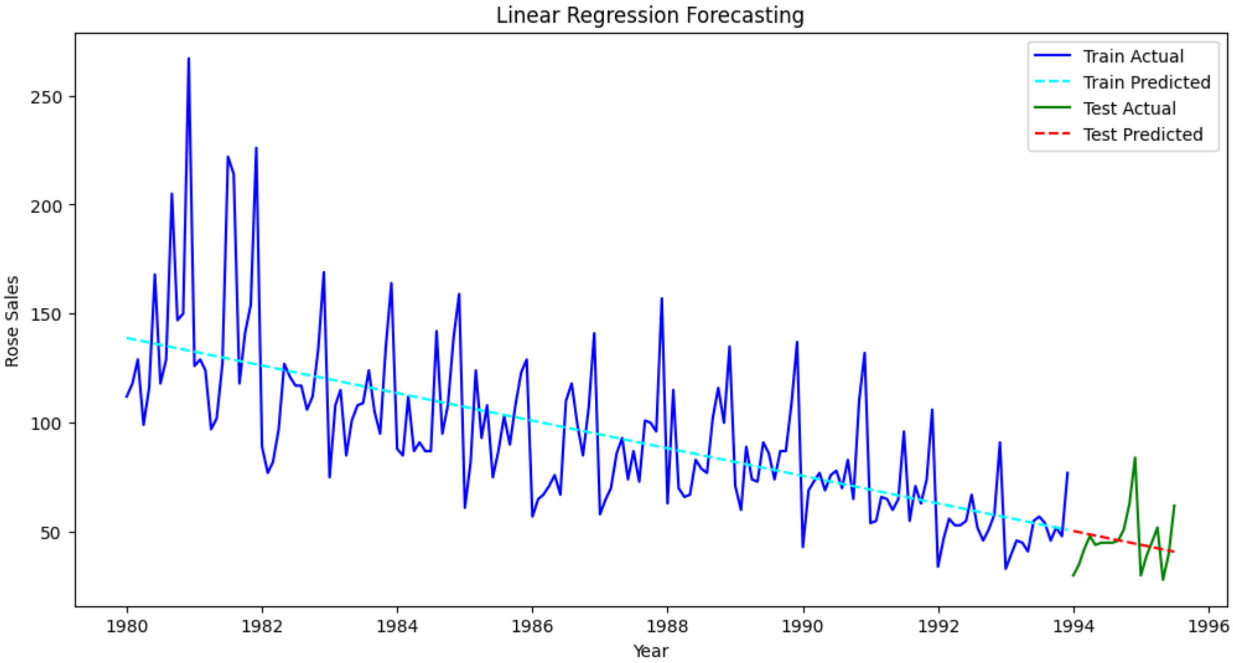

Figure 6. Actual vs predicted sales using Linear Regression

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 Rose 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 Model is a straightforward forecasting technique that assumes future sales will follow the average of past sales values. This method is beneficial when sales fluctuate randomly without a clear trend or seasonality.

For this analysis, the model was trained using historical sales data, computing the mean of past Rose Sales values, and using it to forecast future sales.

The plot below illustrates the Simple Average Forecasting Model.

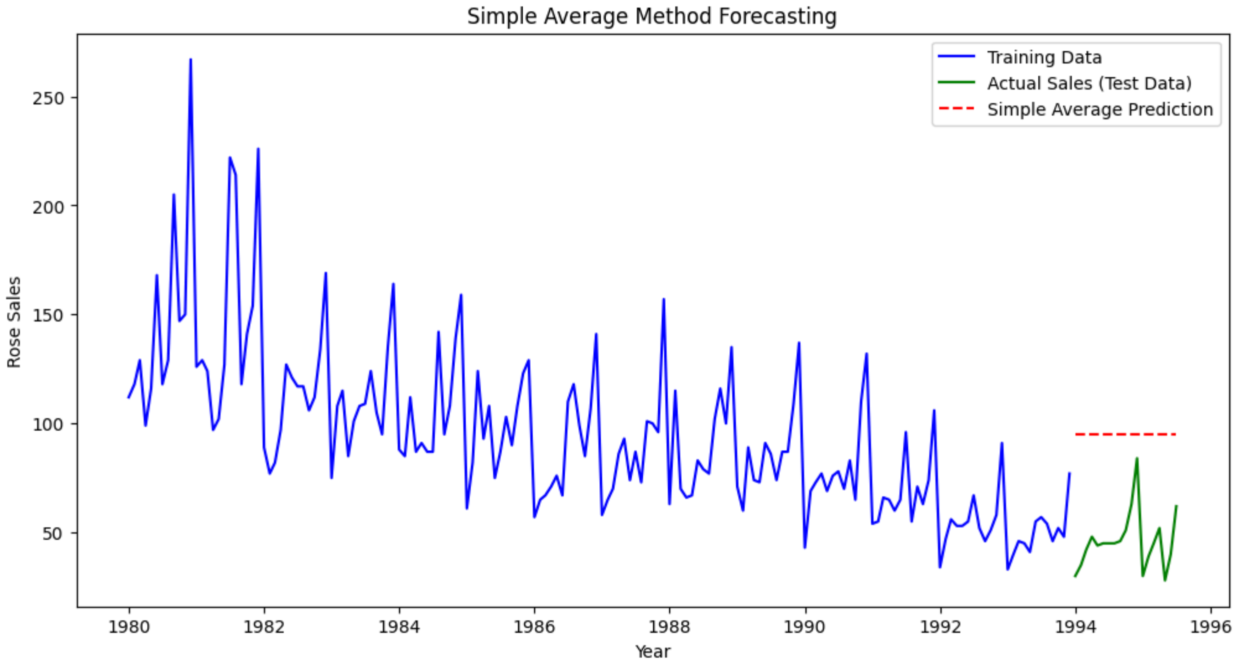

Figure 7. Simple Average Forecasting Model

Interpretation:

- The forecasted sales remain constant, equal to the average of past Rose Sales values.
- This method is useful when there are no apparent trends or seasonality.
- However, since it does not adapt to fluctuations or trends, it may not be suitable for datasets with strong patterns.

3.3 Moving Average Method

The Moving Average Model is a smoothing technique that averages a fixed number of previous observations to forecast future sales. This method reduces short-term fluctuations and highlights the underlying trend in Rose Sales.

For this analysis, we applied a Moving Average Model using a specified window size to smooth past sales data and generate forecasts for future periods.

The plot below illustrates the Moving Average Forecasting Model.

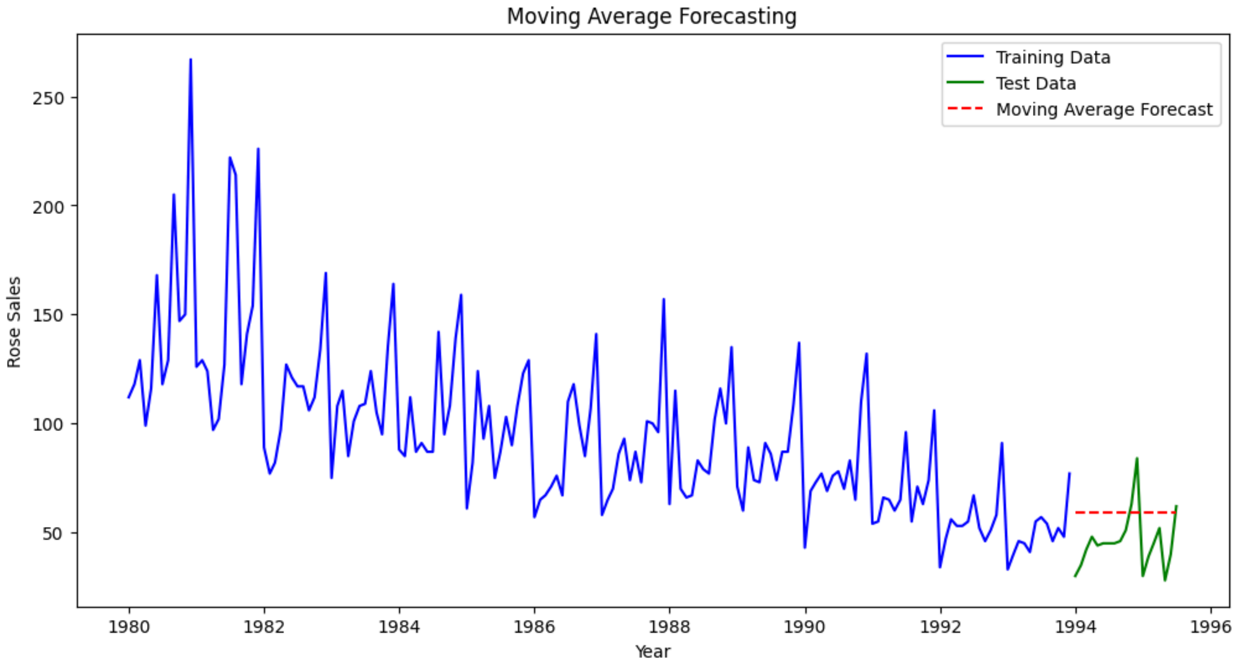

Figure 8. Moving Average Forecasting Model

Interpretation:

- The moving average curve smooths out short-term variations, revealing long-term sales trends.
- A longer window size results in a smoother trend but may lag behind recent changes in sales patterns.
- While effective for detecting trends, this model does not account for seasonal variations.

3.4 Exponential Smoothing

Exponential Smoothing is a widely used forecasting method that applies exponentially decreasing weights to past observations, giving greater importance to recent values. This approach helps to smooth fluctuations and is particularly useful for forecasting Rose Sales when the data exhibits trends and seasonality.

Unlike Simple and Moving Averages, which treat all past observations equally within a window, Exponential Smoothing assigns more weight to recent data points, making it more responsive to changes. There are three main types of Exponential Smoothing Models:

1. Single Exponential Smoothing (SES) – Suitable for data with no clear trend or seasonality.
2. Double Exponential Smoothing (DES) – Useful when a trend is present in the data.
3. Triple Exponential Smoothing (TES) – Also known as Holt-Winters Method, designed for data that exhibits both trend and seasonality.

Each model has been applied to the Rose Sales dataset to evaluate its forecasting effectiveness.

3.4.1 Single Exponential Smoothing (SES)

Single Exponential Smoothing is the simplest form of this method. It predicts future sales based on a weighted average of past observations, where recent data points have higher weights. This method is best suited for time series data with random variations but no trend or seasonality.

For this analysis, we applied SES to the Rose Sales dataset, using an optimal smoothing factor (α) to determine the best weight distribution for past sales data.

The plot below illustrates the Single exponential Forecasting Model.

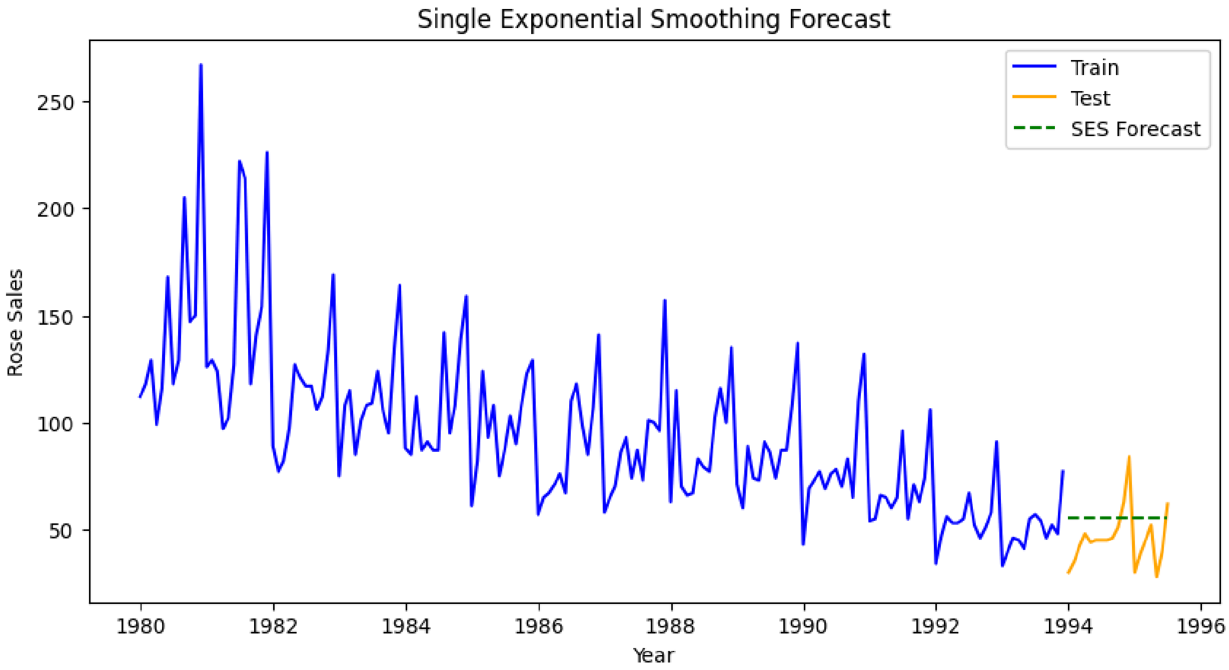

Figure 9. Single exponential Forecasting Model

Interpretation:

- The forecasted line follows a smooth trajectory, closely resembling the overall movement of past Rose Sales values.
- Since this method does not account for trends or seasonality, it may not be suitable for datasets with clear upward or downward patterns.
- If sales remain relatively stable with minor fluctuations, this approach can provide reliable short-term forecasts.

3.4.2 Double Exponential Smoothing (Holt’s Method)

- Double Exponential Smoothing extends Single Exponential Smoothing by incorporating a trend component. It helps in forecasting data that exhibits a clear increasing or decreasing trend over time.
- This method introduces an additional smoothing factor (β) that adjusts for the trend component. The forecasted values are a combination of level smoothing (SES) and trend smoothing, making DES suitable for datasets with a consistent growth or decline in sales.
- For this analysis, we applied DES (Holt’s Method) to capture the underlying trend in Rose Sales.

The below plot shows the Double Exponential Smoothing Forecast:

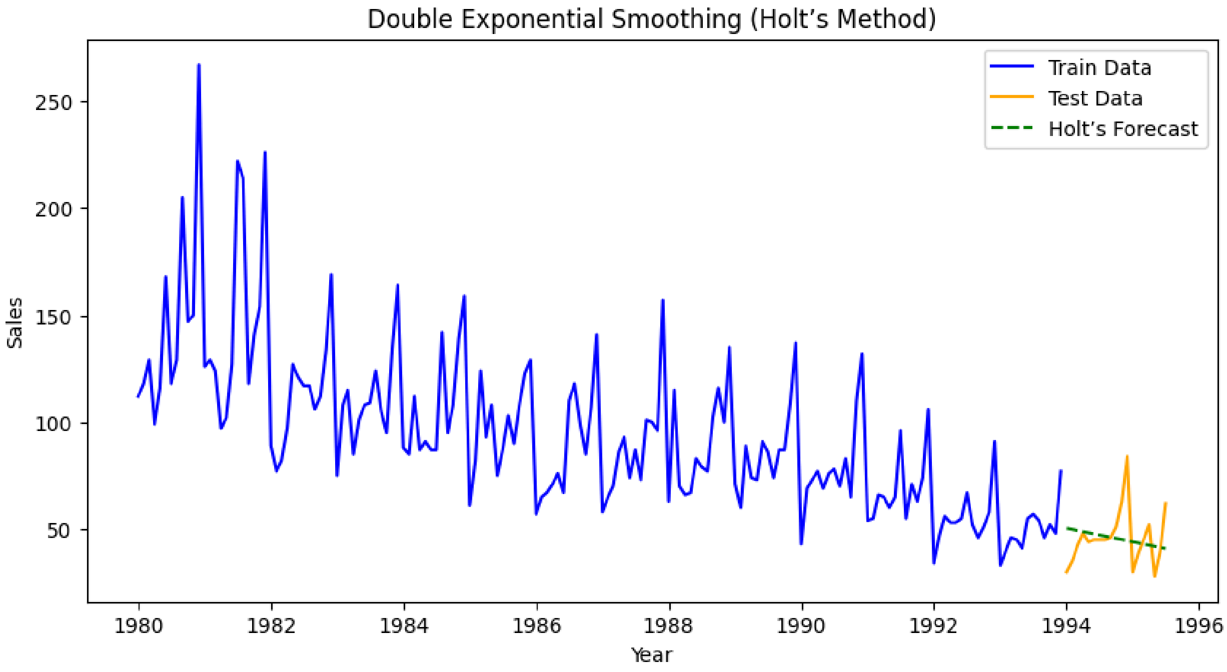

Figure 10. Double Exponential Smoothing Forecast

Interpretation:

- The forecasted sales line captures the upward/downward trend in the data, making it more accurate than SES for trending sales.
- This method effectively accounts for long-term growth or decline, providing a more refined forecast compared to Single Exponential Smoothing.
- However, since DES does not include a seasonal component, it may not accurately predict recurring fluctuations in sales.

3.4.3 Triple Exponential Smoothing (Holt-Winters Method)

Triple Exponential Smoothing, also known as the Holt-Winters Method, is an advanced extension of Exponential Smoothing that incorporates both trend and seasonality. This model is particularly useful when forecasting data that exhibits a repeating seasonal pattern over time.

This method introduces a third smoothing parameter (γ) that adjusts for seasonal variations in the data. It can be applied in two variations:

- Additive Model – Used when the seasonal variations remain constant over time.
- Multiplicative Model – Used when seasonal variations increase or decrease proportionally with sales growth.

For this analysis, we applied TES (Holt-Winters Method) to Rose Sales using an optimal seasonal period and smoothing parameters.

The below plot shows the Triple Exponential Smoothing Forecast:

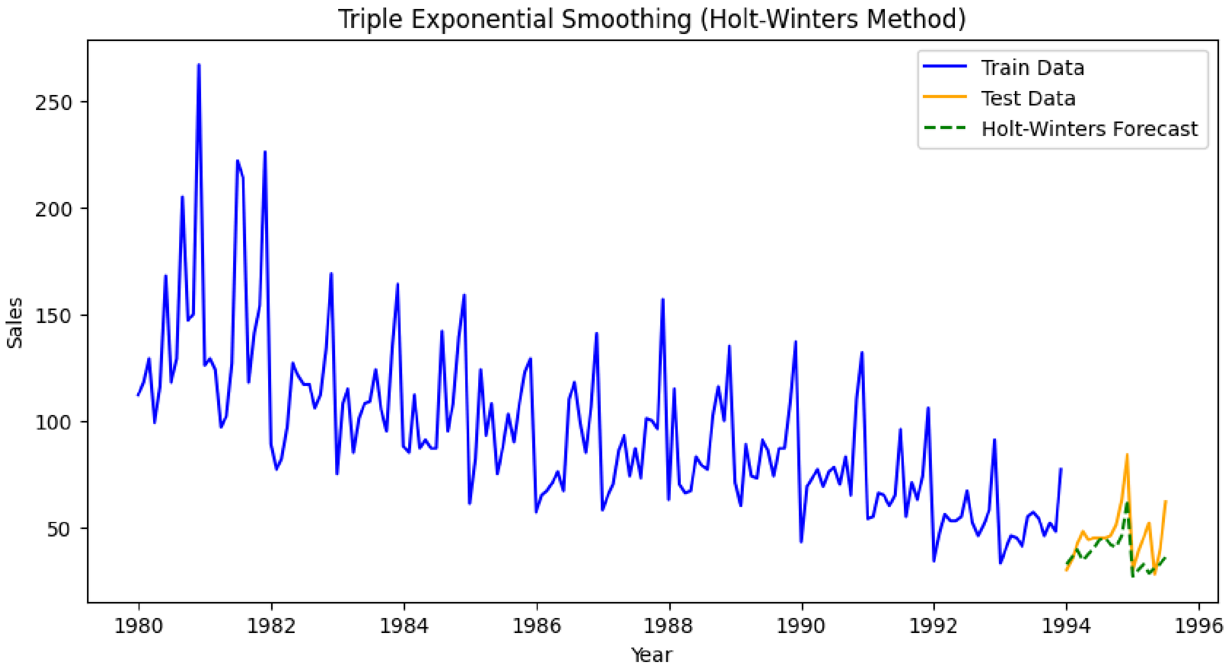

Figure 11. Triple Exponential Smoothing Forecast

Interpretation:

- The forecasted sales line accurately captures both trend and seasonal patterns, making it highly effective for Rose Sales, where demand may fluctuate seasonally.
- The model adjusts dynamically to changes in trend and seasonality, ensuring more reliable long-term forecasts.
- Compared to SES and DES, this method provides a superior forecasting accuracy, especially when clear seasonal patterns exist.

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, as it suggests that the predicted values are closer to the actual sales figures.

3.5.Model Evaluation using Mean Squared Error (MSE)

The MSE results for each model are as follows:

- Holt-Winters Method achieved the lowest MSE of 140.45, making it the best-performing model.
- Linear Regression followed closely, with an MSE of 188.48, indicating moderate accuracy.
- Holt's Method also produced an MSE of 188.48, showing similar performance to Linear Regression.
- Simple Exponential Smoothing had a higher error of 258.04, suggesting limitations in capturing trends effectively.
- Moving Average resulted in an MSE of 333.21, showing poor predictive capability.
- Simple Average performed the worst, with an MSE of 2552.98, 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.

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 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 Rose Sales dataset, we employ both visual inspection and a statistical test.

The plot below shows the Rolling Statistics - Stationarity check:

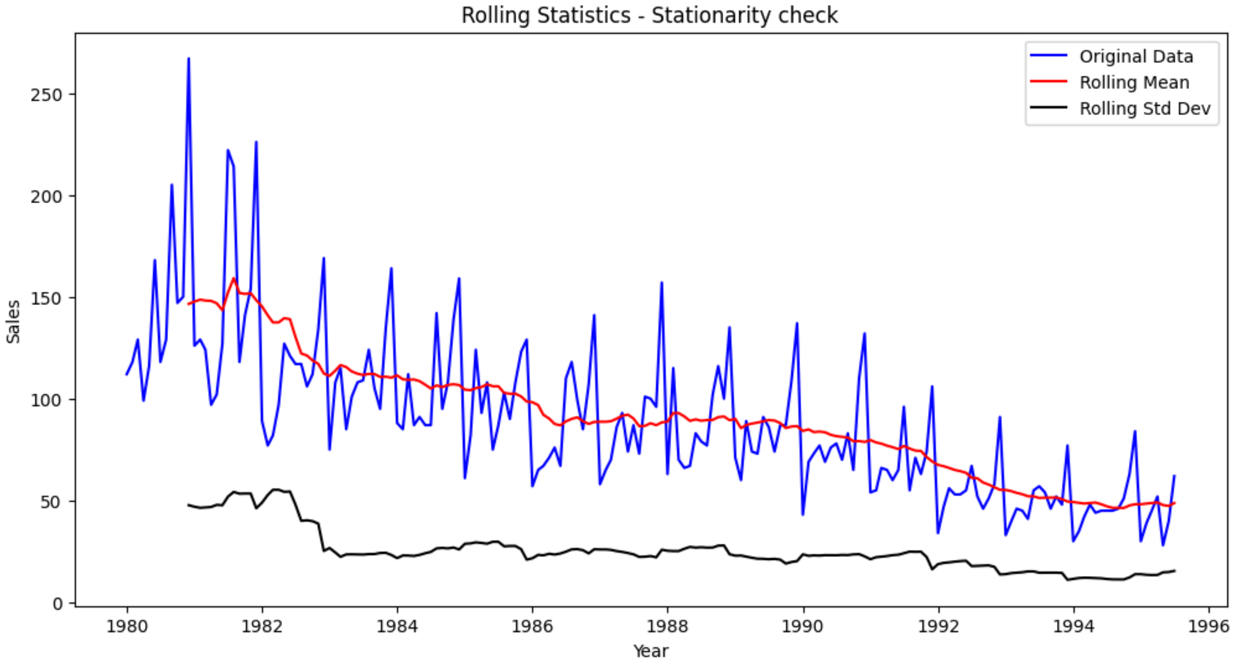

Figure 12. Rolling Statistics - Stationarity Check

Interpretation:

- The rolling mean (red line) shows a gradual downward trend, suggesting a declining sales pattern over time.
- The rolling standard deviation (black line) fluctuates, indicating that the variance of the dataset is not constant.
- The original time series (blue line) exhibits strong seasonal patterns, with periodic spikes and variations in sales.
- Since both the mean and standard deviation are changing over time, the dataset is non-stationary and requires transformation 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.

Hypothesis of the ADF Test

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

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:

- ADF Statistic: -1.8748
- p-value: 0.3440
- Critical Values:
  - 1%: -3.4687
  - 5%: -2.8784
  - 10%: -2.5758

Since the p-value (0.3440) is greater than 0.05, we fail to reject the null hypothesis, confirming that the time series is non-stationary. The results from both rolling statistics and the ADF test confirm that the Rose Sales dataset is non-stationary. This means that further transformations will be required to make the data suitable for forecasting models.

4.2 Transforming Data to Achieve Stationarity

After confirming that the Rose Sales dataset was non-stationary, transformations were applied to stabilize the mean and remove trends. One of the most effective methods for achieving stationarity is first-order differencing, which calculates the difference between consecutive observations.

By applying first-order differencing, we eliminate trend components, ensuring that fluctuations in the data remain consistent over time. This transformation helps make the series suitable for time series modeling, particularly for forecasting methods that assume stationarity.

The plot below illustrates the differenced data, highlighting the stability of the transformed time series.

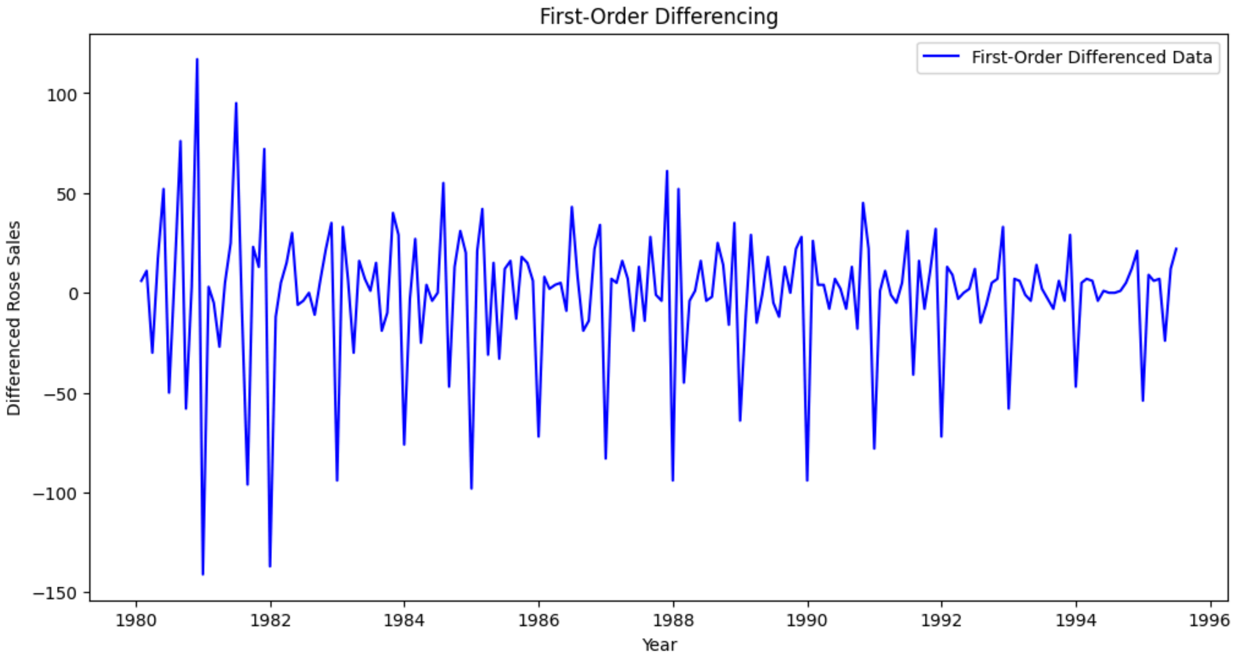

Figure 13. First-Order Differencing Plot

Interpretation:

- The trend observed in the original data has been removed, and the values now fluctuate around a relatively constant mean.
- The variability in the dataset is more stable, with no clear increasing or decreasing trends.
- The fluctuations are more consistent over time, suggesting that the series is moving toward stationarity.

Augmented Dickey-Fuller (ADF) Test After First-Order Differencing

To confirm whether the first-order differenced series is stationary, we perform the Augmented Dickey-Fuller (ADF) test again.

The results of the ADF test are as follows:

- ADF Statistic: -8.0441
- p-value: 1.81×10^{-12}
- Critical Values:
    - 1%: -3.4687
    - 5%: -2.8784
    - 10%: -2.5758

Since the p-value is significantly lower than 0.05, we reject the null hypothesis of non-stationarity. Additionally, the ADF test statistic (-8.0441) is lower than all critical values, confirming that the time series is now stationary. The results from both visual inspection and the ADF test confirm that after first-order differencing, the dataset has achieved stationarity. This transformation ensures that the time series meets the key assumption required for models like ARIMA and SARIMA, allowing for accurate forecasting.

5. Stationary Data Model Building & Evaluation

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

To construct an effective ARIMA model, it is essential to determine the values of the Auto-Regressive (AR) term (p) and the Moving Average (MA) term (q). This identification is achieved 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 illustrating how past values impact the present value across different lag periods.

The below plot represents the ACF analysis:

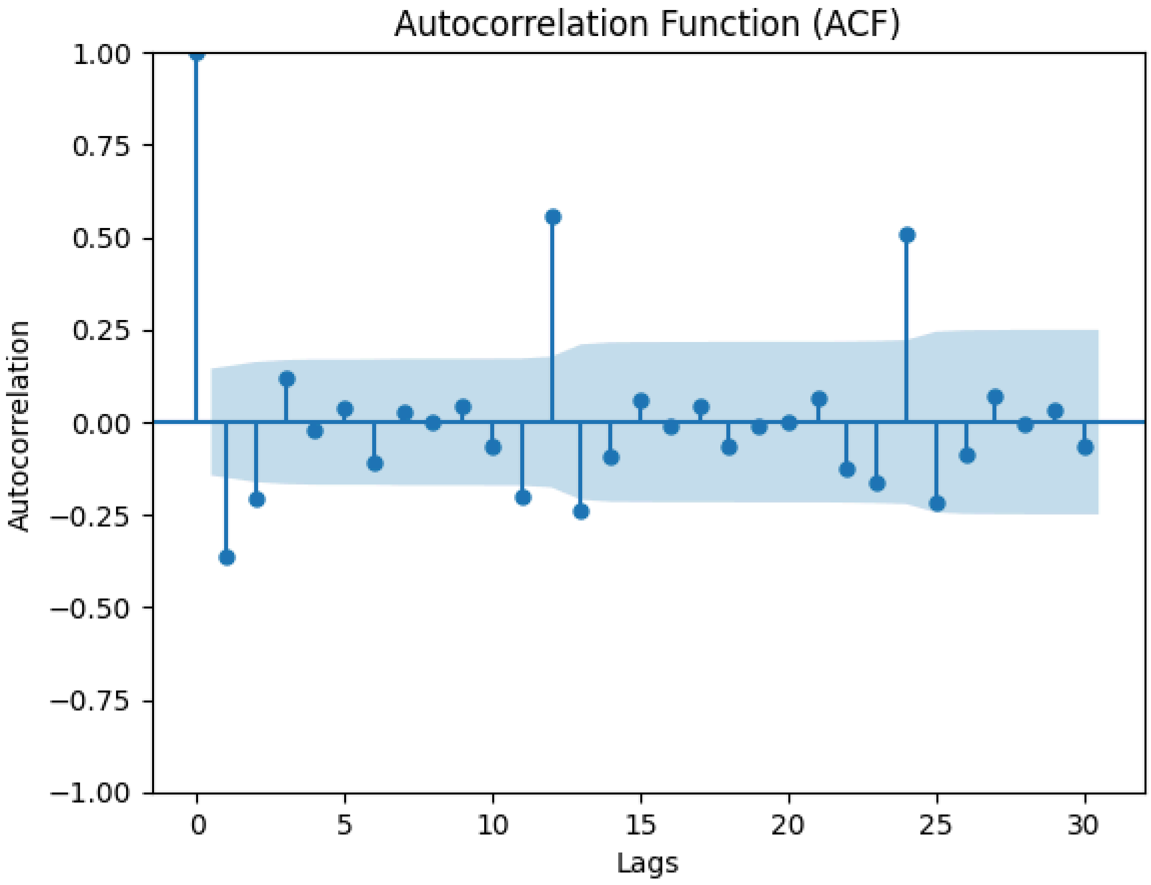

Figure 14. ACF Plot

Interpretation:

- The first spike at lag 0 is always at 1, representing the correlation of the series with itself.
- The blue-shaded region indicates the confidence interval (CI), where values within this range are considered statistically insignificant.
- Significant spikes beyond the CI indicate meaningful correlations at specific lags.
- In this case, the first two spikes beyond lag 0 (at lag 1 and lag 2) exceed the CI, suggesting that past errors up to lag 2 influence 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 accounting for intermediate effects.

The below plot represents the PACF analysis:

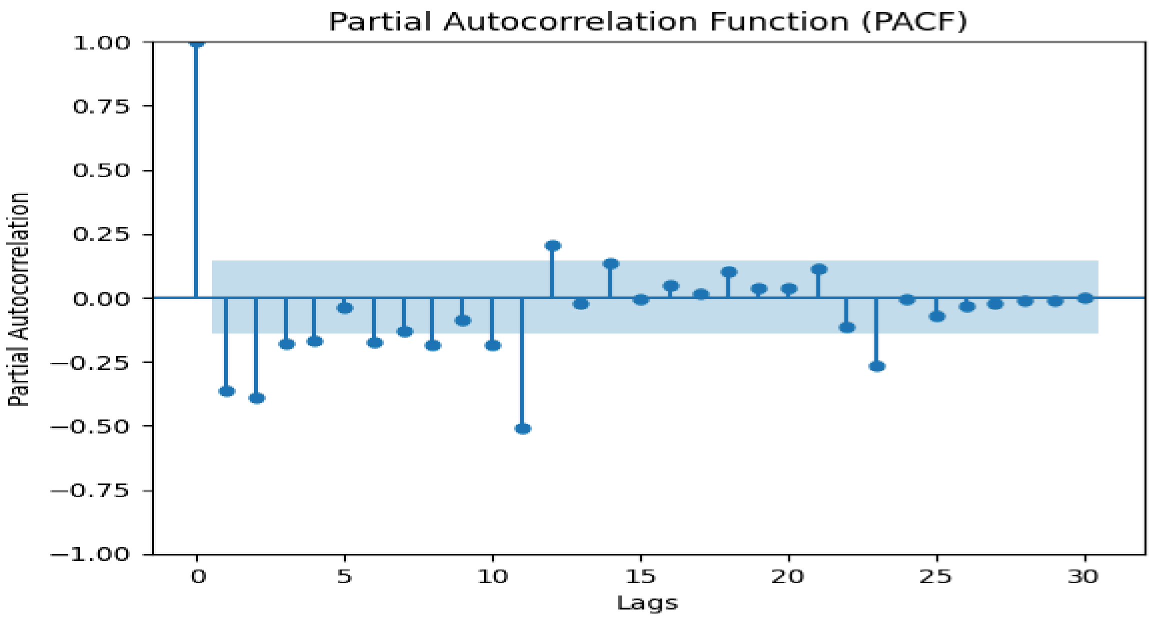

Figure 15. PACF Plot

Interpretation:

- Similar to 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 four spikes beyond lag 0 (at lag 1, 2, 3, and 4) exceed the CI, indicating a significant direct correlation of these past values with the present value.
- As a result, we determine p = 4 for the Auto-Regressive term.

With these values of p = 4 and q = 2, the next step involves selecting an appropriate ARIMA model for forecasting Rose Sales.

5.2 Build Different ARIMA Models

5.2.1 Auto ARIMA Model

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

The selected Auto ARIMA model for forecasting Rose Sales is SARIMAX(2,1,3), based on the lowest AIC value of 1754.161.

The below plot represents the Auto ARIMA Forecast:

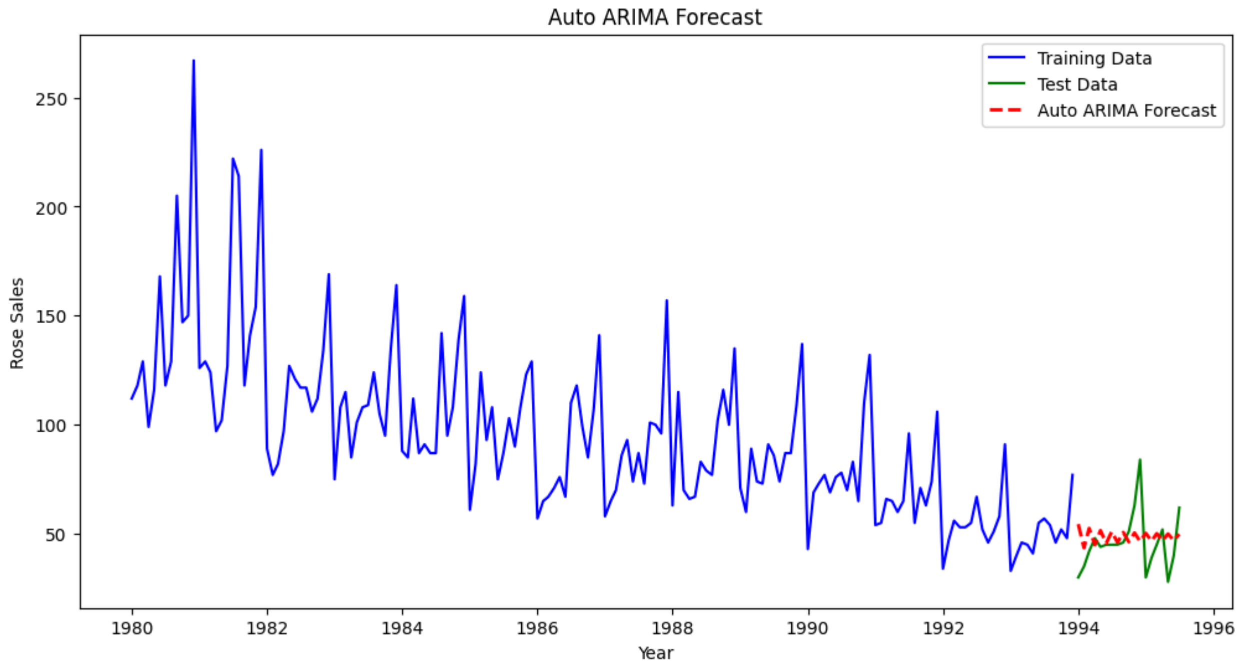

Figure 16. Auto ARIMA Forecast plot

Interpretation:

- The blue line represents the training data, while the green line represents the actual test data.
- The red dashed line denotes the forecasted values generated by the Auto ARIMA model.
- The forecast remains relatively stable over time, showing limited responsiveness to fluctuations in the actual test data.
- The model parameters suggest that two Auto-Regressive (AR) terms and three Moving Average (MA) terms were identified as optimal.
- However, the forecasted values do not fully capture the seasonal patterns in the dataset, indicating that Auto ARIMA might not have accounted for seasonality or underlying non-linearity effectively.

5.2.2 Manual ARIMA Model

Unlike Auto ARIMA, the Manual ARIMA model requires selecting the (p, d, q) parameters manually. Based on the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, the selected values are:

- p (AR terms) = 4
- d (Differencing order) = 1
- q (MA terms) = 2

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

The below plot represents the Manual ARIMA Forecast:

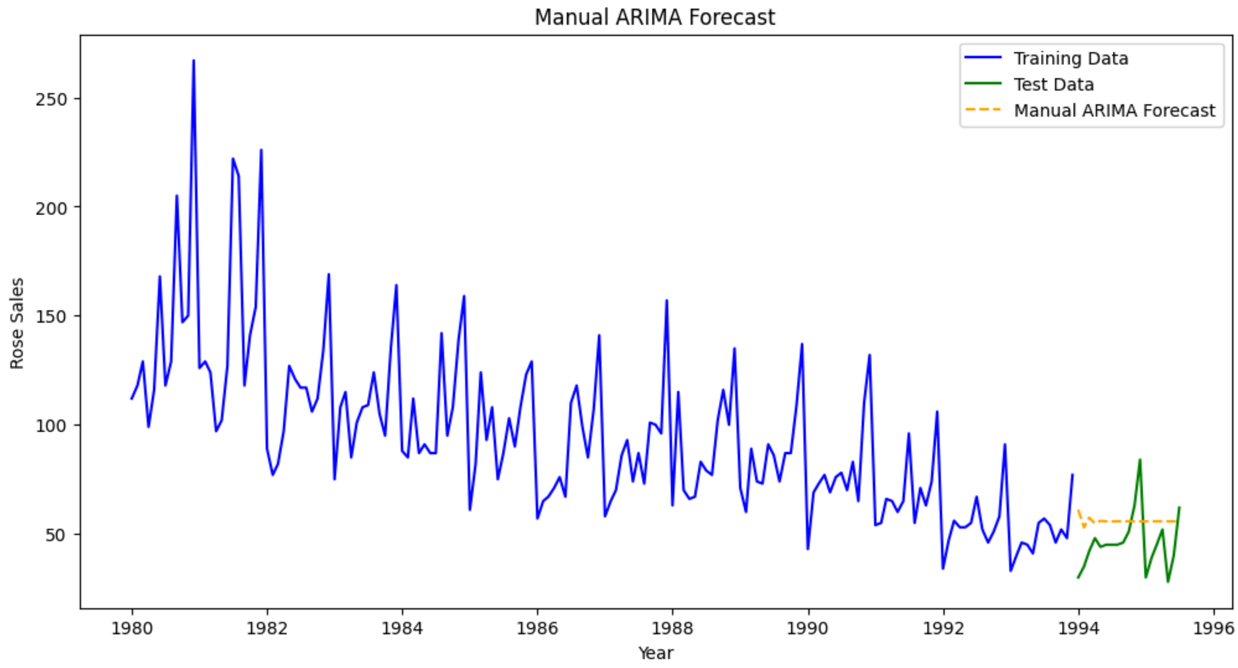

Figure 17. Manual ARIMA Forecast plot

Interpretation:

- The blue line represents the training data, while the green line indicates the actual test data.
- The orange dashed line denotes the forecasted values generated by the manually configured ARIMA model (ARIMA(4,1,2)).
- Compared to the Auto ARIMA forecast, the manually tuned model exhibits better responsiveness to fluctuations in the test data.
- The forecasted values show some variability rather than a flat-line prediction, suggesting that the manually chosen parameters have helped in capturing certain underlying patterns in the time series.
- However, discrepancies remain, particularly in capturing the sharp seasonal peaks and extreme variations in the test data.

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 exhibiting 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 represents the seasonal period length

In this section, we explore two SARIMA modeling approaches:

- Auto SARIMA, which automatically determines optimal parameters based on statistical criteria.
- Manual SARIMA, where parameters are selected through statistical analysis and fine-tuning.

Auto SARIMA is an automated approach that selects the most optimal SARIMA parameters based on statistical criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). This method systematically evaluates multiple combinations of seasonal and non-seasonal parameters, minimizing forecast errors to improve model accuracy.

The below plot illustrates the Auto SARIMA Forecast:

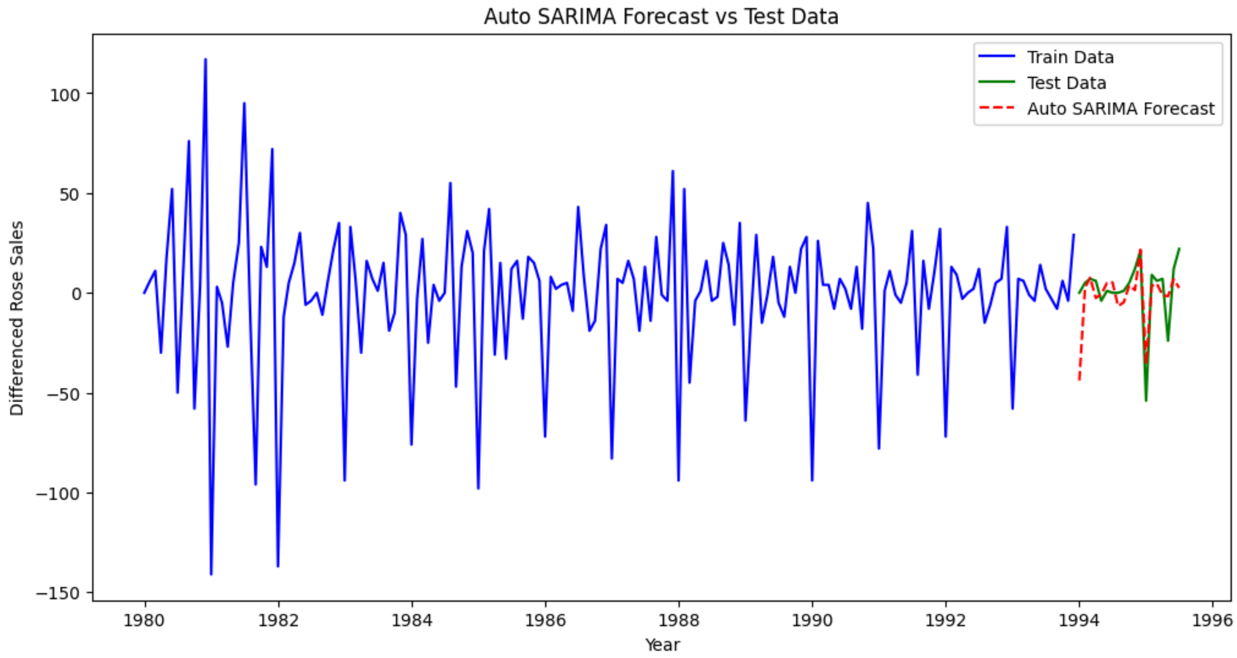

Figure 18. Auto SARIMA Forecast plot

Interpretation:

- The blue line represents the training data.
- The green line represents the actual test data.
- The red dashed line represents the forecasted values generated using the Auto SARIMA model.

From the visualization, it is evident that Auto SARIMA successfully captures certain seasonal fluctuations within the dataset. However, some variations in the actual test data are not entirely reflected in the forecast, suggesting room for further refinement. While Auto SARIMA provides a strong baseline model, manual tuning of seasonal and non-seasonal parameters may lead to improved forecasting accuracy.

5.3.2 Manual SARIMA Model

Unlike Auto SARIMA, which automatically selects parameters, the Manual SARIMA model requires a systematic selection of (p, d, q) and (P, D, Q, s) parameters using statistical techniques such as:

- ACF (AutoCorrelation Function) and PACF (Partial AutoCorrelation Function) Analysis
- Stationarity Testing (Augmented Dickey-Fuller Test)
- Seasonal Decomposition

By carefully tuning these parameters, the model can better capture underlying seasonal trends and fluctuations.

Stationarity Testing and Differencing

To achieve stationarity, the dataset underwent seasonal differencing. The Augmented Dickey-Fuller (ADF) test was conducted to validate stationarity:

- Initially, the dataset was non-stationary after applying a first-order seasonal differencing (D=1).
- A second round of seasonal differencing (D=2) was applied, successfully rendering the data stationary, as confirmed by the ADF test results.

Manual SARIMA Model Implementation

The selected SARIMA parameters for the manual model are:

- Non-seasonal Order (p, d, q): (4, 1, 2)
- Seasonal Order (P, D, Q, s): (2, 2, 2, 12)

These values were determined through ACF and PACF analysis, ensuring the model effectively captures both seasonal and non-seasonal components.

Manual SARIMA Forecast Plot

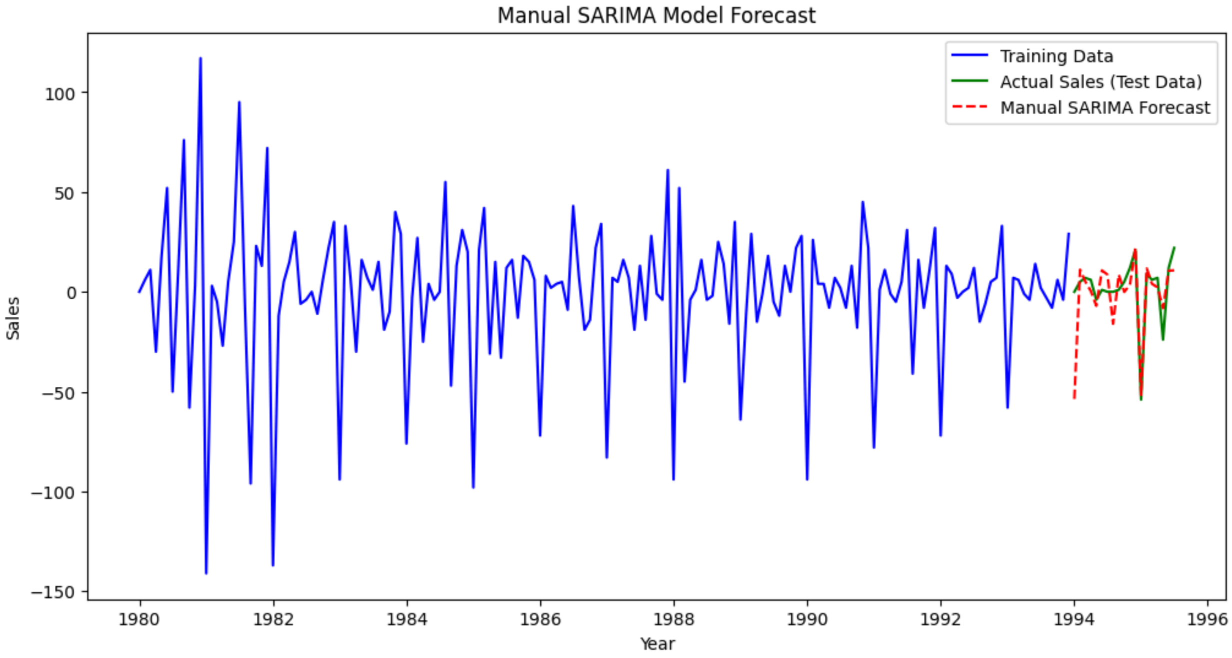

Figure 19. Manual SARIMA Forecast Plot

Interpretation:

- The blue line represents the training data.
- The green line represents the actual test data.
- The red dashed line represents the Manual SARIMA forecast.

Both Auto SARIMA and Manual SARIMA models generate similar forecast results, with only minor variations. Since Auto SARIMA automatically selects the best-fit parameters, it serves as a reliable and efficient choice for forecasting. However, manual tuning allows for greater flexibility in refining the model based on deeper statistical analysis, making it beneficial for datasets with unique seasonal structures.

To assess the accuracy of different forecasting models, the Root Mean Squared Error (RMSE) metric is used. RMSE measures the average magnitude of errors in predictions, with lower values indicating better model performance. The RMSE values of all models are analyzed to determine the most effective forecasting approach.

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: 13.76 (Best-performing model)
- Auto ARIMA: 13.94
- Manual SARIMA: 14.44
- Manual ARIMA: 16.46 (Least accurate model)

Interpretation:

- The Auto SARIMA model achieves the lowest RMSE (13.76), making it the most accurate forecasting model.
- Auto ARIMA performs slightly worse with an RMSE value of 13.94, but it still provides relatively accurate forecasts.
- Manual SARIMA shows a moderate performance with an RMSE of 14.44, indicating it captures the data's seasonality reasonably well, but not as effectively as Auto SARIMA.
- Manual ARIMA has the highest RMSE (16.46), suggesting that it is the least accurate model among those tested.

The results clearly highlight that Auto SARIMA outperforms the other models, effectively capturing the underlying trends and seasonality in the data. Given its superior performance, it will be used for final forecasting and business recommendations.

6. Model Comparison and Final Forecasting

After developing multiple forecasting models, it is essential to compare their performance and determine the most effective one. The evaluation is based on the Root Mean Squared Error (RMSE), which quantifies the difference between predicted and actual values. A lower RMSE indicates a better-performing model. This section presents a comparative analysis of all models, selects the most accurate one, and uses it for the final forecast.

6.1 Overall Model Performance Comparison

The RMSE values for all models are as follows:

- Holt-Winters Method: 11.85 (Best-performing model)
- Linear Regression: 13.73
- Holt’s Method: 13.73
- Auto SARIMA: 13.76
- Manual SARIMA: 14.44
- Simple Exponential Smoothing: 16.06
- Moving Average: 18.25
- Auto ARIMA: 49.96
- Simple Average: 50.53
- Manual ARIMA: 56.61 (Least accurate model)

Interpretation of Results:

- The Holt-Winters Method achieves the lowest RMSE (11.85), making it the most accurate forecasting model.
- Auto SARIMA (13.76) and Manual SARIMA (14.44) also perform well, indicating their ability to capture seasonality.
- Linear Regression (13.73) and Holt’s Method (13.73) show similar performance, ranking among the top models.
- Simple Exponential Smoothing (16.06) and Moving Average (18.25) offer moderate accuracy.

- Auto ARIMA (49.96) and Manual ARIMA (56.61) show the highest errors, suggesting that they do not effectively capture the data's underlying patterns.
- The Simple Average model (50.53) performs poorly, highlighting its inefficiency in handling trends and seasonality.

The Holt-Winters Method will be used for final forecasting and business recommendations, as it provides the most accurate and reliable predictions.

6.2 Selection of the Best Model

After evaluating multiple forecasting models, the Holt-Winters Method demonstrated the best performance, achieving the lowest RMSE of 11.85. This model effectively captures both trend and seasonality, making it well-suited for forecasting sales data with recurring patterns.

Although other models, such as Linear Regression (RMSE: 13.73) and Auto SARIMA (RMSE: 13.75), also exhibited strong predictive accuracy, 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 time series forecasting.

Given these observations, the Holt-Winters Method is selected as the best forecasting model due to:

- Superior Accuracy: Achieved the lowest RMSE, ensuring minimal forecasting error.
- Seasonality Capture: Effectively models seasonal variations in sales data.
- Ease of Implementation: Requires fewer manual adjustments compared to ARIMA and SARIMA models.

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

Based on the model evaluation, the Holt-Winters Method was identified as the most effective forecasting approach. To enhance its accuracy, the model was rebuilt using the entire dataset and used to generate a 12-month forecast from August 1995 to July 1996.

The forecasted sales values indicate seasonal variations, reflecting historical trends. Notably:

- December Sales Peak: The highest forecasted sales occur in December 1995, consistent with past seasonal demand surges.
- January Drop: A significant decline in sales is expected in January 1996, following the holiday season.
- Stable Growth Period: Sales gradually stabilize from February to July 1996, showing moderate fluctuations.

12-Month Forecast Using Holt-Winters Method

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

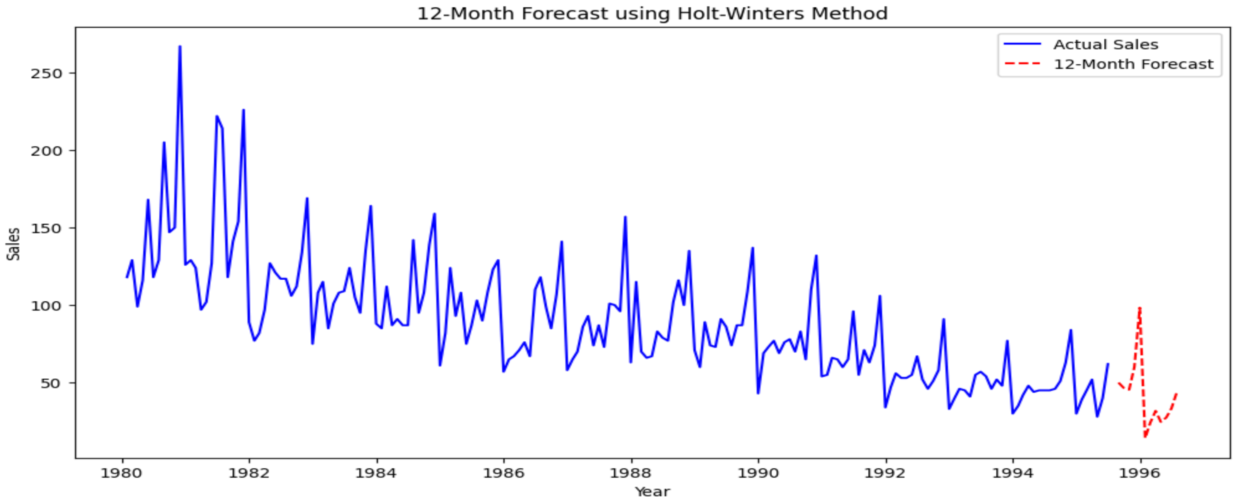

Figure 20: 12-Month Forecast Plot

Interpretation of the Forecast

The plot above illustrates the historical sales data (blue line) and the forecasted values (red dashed line) for the next 12 months. Key insights from the forecast include:

- Seasonal Pattern: The forecast follows a recurring pattern, reflecting past seasonality.
- December Sales Surge: A notable increase in sales is expected in December, aligning with historical seasonal peaks.
- Post-Holiday Decline: Sales show a sharp decline in January, consistent with past trends where demand slows after peak months.
- Stable Growth: From February to July, sales demonstrate moderate fluctuations, suggesting a period of stable demand.

This forecast provides valuable insights for decision-making regarding inventory management, marketing strategies, and resource allocation. By anticipating seasonal demand fluctuations, the company can optimize its operations and maximize profitability

7. Actionable Insights & Recommendations

7.1.Key Observations from the Forecast

The 12-month sales forecast using the Holt-Winters Method provides a clear projection of expected sales trends and accurately reflects the historical seasonal pattern. The forecast indicates:

- Recurring seasonal peaks in sales, with a significant increase observed in December 1995, followed by a sharp decline in January 1996.
- Moderate fluctuations throughout the rest of the forecast period, suggesting a period of stabilized demand.
- Potential growth opportunities towards mid-1996, indicating a shift in demand dynamics that could be leveraged for business expansion.

These insights provide a strong foundation for strategic decision-making, allowing the business to optimize operations, mitigate risks, and capitalize on peak sales periods.

Business Implications:

The insights derived from the forecast have direct implications for business operations, financial planning, and growth strategies:

- Optimized Inventory Management: Forecasted sales peaks allow for better inventory planning, ensuring stock availability during high-demand months.
- Enhanced Financial Planning: Understanding seasonal fluctuations enables businesses to manage cash flow efficiently and allocate resources effectively.
- Strategic Market Expansion: The potential growth trend towards the end of the forecast period suggests opportunities for market expansion and targeted investments.
- Refined Marketing Strategies: Aligning promotional campaigns with high-sales periods can maximize revenue potential and customer engagement.

7.2.Actionable Recommendations:

To leverage the insights gained from the forecast, businesses can implement the following strategic actions:

1. Inventory Management
   - Ensure sufficient stock levels before peak sales periods (e.g., December) to prevent shortages.
   - Avoid overstocking during months with low projected demand to optimize storage costs.
2. Marketing Strategy
   - Align advertising and promotional campaigns with periods of increased sales activity to maximize impact.
   - Offer seasonal discounts or bundled promotions to maintain steady sales during slower months.
3. Workforce Planning
   - Adjust staffing levels to accommodate seasonal demand shifts, ensuring efficient customer service and operational efficiency.
   - Schedule training programs for employees ahead of peak seasons to improve productivity.
4. Supply Chain Optimization
   - Strengthen relationships with key suppliers to maintain a steady supply chain during high-demand months.
   - Explore alternative suppliers to mitigate potential disruptions in procurement.
5. Data-Driven Decision Making
   - Continuously update forecasting models with new sales data to improve prediction accuracy.
   - Leverage real-time analytics to monitor demand fluctuations and adjust strategies accordingly.

7.3.Conclusion

The 12-month sales forecast using the Holt-Winters Method provides valuable insights into future sales trends, capturing seasonal peaks, demand shifts, and potential growth opportunities. Businesses can proactively plan for high-demand periods, optimize inventory management, adjust workforce levels, and implement strategic marketing campaigns to enhance profitability.

Additionally, the forecast highlights opportunities for market expansion and investment, particularly towards mid-1996. By adopting a data-driven approach, businesses can refine forecasting models, ensure financial stability, and sustain long-term growth through continuous optimization and strategic planning.