Time Series Forecasting

Issac Abraham

Table of Contents

List of Figures

List of Tables

Vintage Insights: A Time Series Analysis and Forecasting of Wine Sales in the 20th Century

Executive Summary:

The "Vintage Insights" initiative explores past sales information of two different varieties of ABC Estate Wines: Rose and Sparkling. We seek to identify trends and patterns in time series data and use forecasting models to extract useful information for upcoming sales tactics.

Problem:

The data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

Dataset Exploration

We started by reading the Sparkling.csv dataset using the Pandas library, creating a DataFrame named **sparkling_data.** The result showed important columns like "YearMonth" and "Sparkling," giving us a peek into the Sparkling dataset. After that, we started the process of datetime64-formatting the 'YearMonth' column in order to perform temporal analysis. We made sure that our temporal data was prepared correctly for time series analysis by taking this step. Next, in order to see the trends and patterns in the time series data, we plotted the data. The time series plot that resulted provided a visual representation of the sparkling wine sales trends over the course of the analysis.

Simultaneously, we delved into the Rose.csv dataset, following a similar approach of data exploration and visualization.

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

	YearMonth	Rose
0	1980-01-01	112.0
1	1980-02-01	118.0
2	1980-03-01	129.0
3	1980-04-01	99.0
4	1980-05-01	116.0

Fig 1 : Appropriate Formatting to Datetime64 for Sparkling and Rose Datasets

Time Series plots

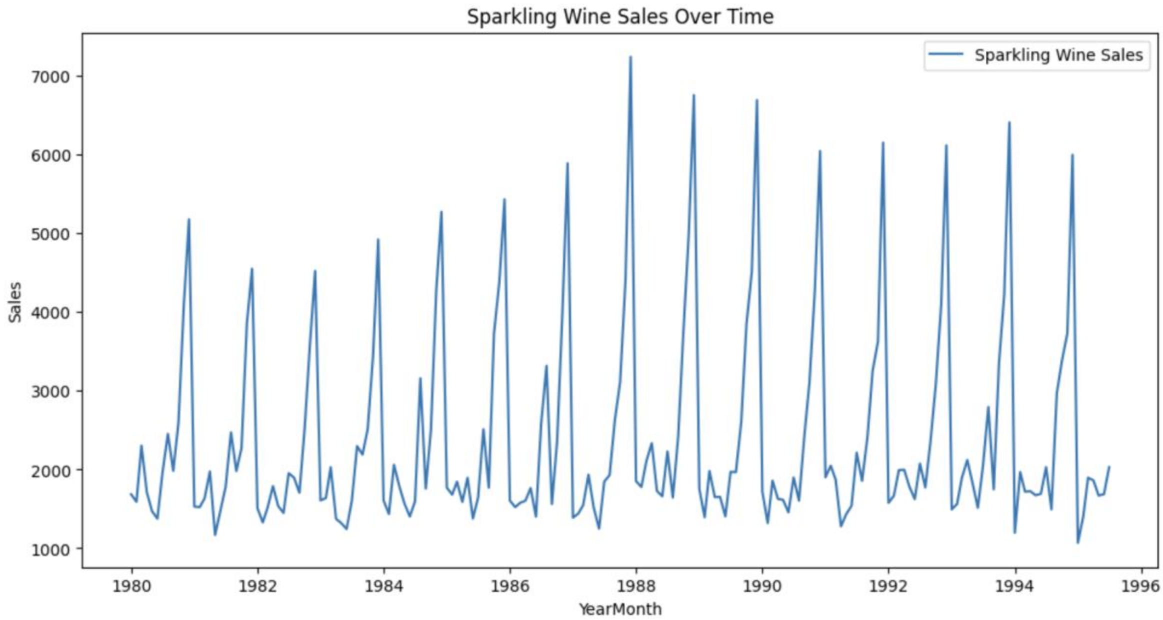

Fig 2: Time Series plot for Sparkling wine (1980-1996)

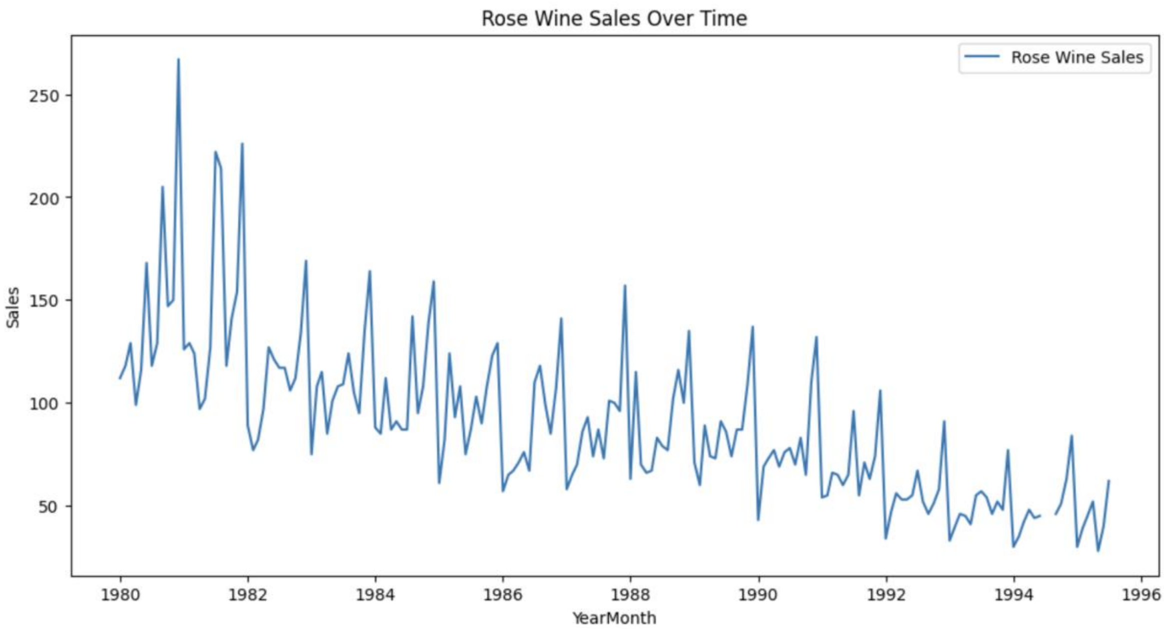

Fig 3 : Time Series plot for Rosy wine (1980-1996)

Handling missing values

Next, we evaluated if there were any missing values in the datasets. There were no missing values found in the Sparkling dataset, indicating that the 'Sparkling' column had a full set of observations. On the other hand, two missing entries in the 'Rose' column were found in the Rose dataset.
To resolve the missing values in the Rose dataset's "Rose" column, we choose to use a moving average imputation method. This approach was chosen due to its efficacy and simplicity, especially when working with time series data.

Seasonal Decomposition of Sparkling Wine and Rose Wine Sales: Additive and Multiplicative Models

We then used both additive and multiplicative models to perform seasonal decomposition on the imputed sales data for Rose wine. Understanding the fundamental elements of a time series, such as trend, seasonality, and residual fluctuations, requires first performing a seasonal decomposition.

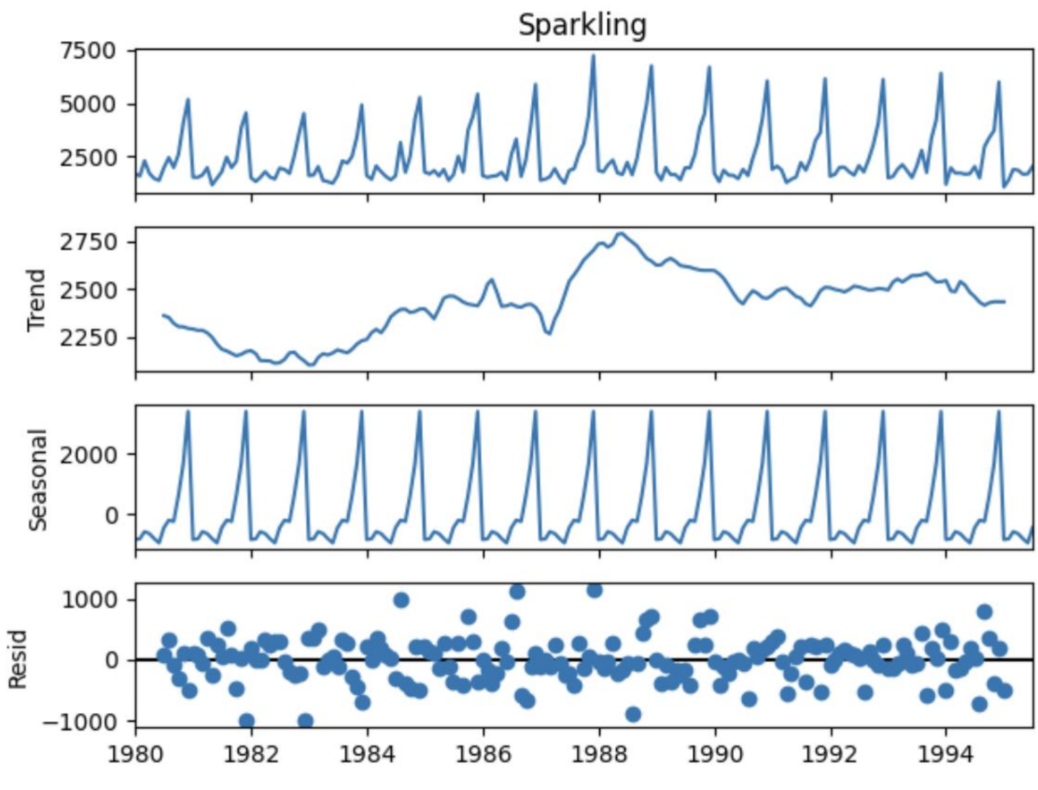

Fig 4 : Seasonal decomposition of Sparkling wine(Additive model)

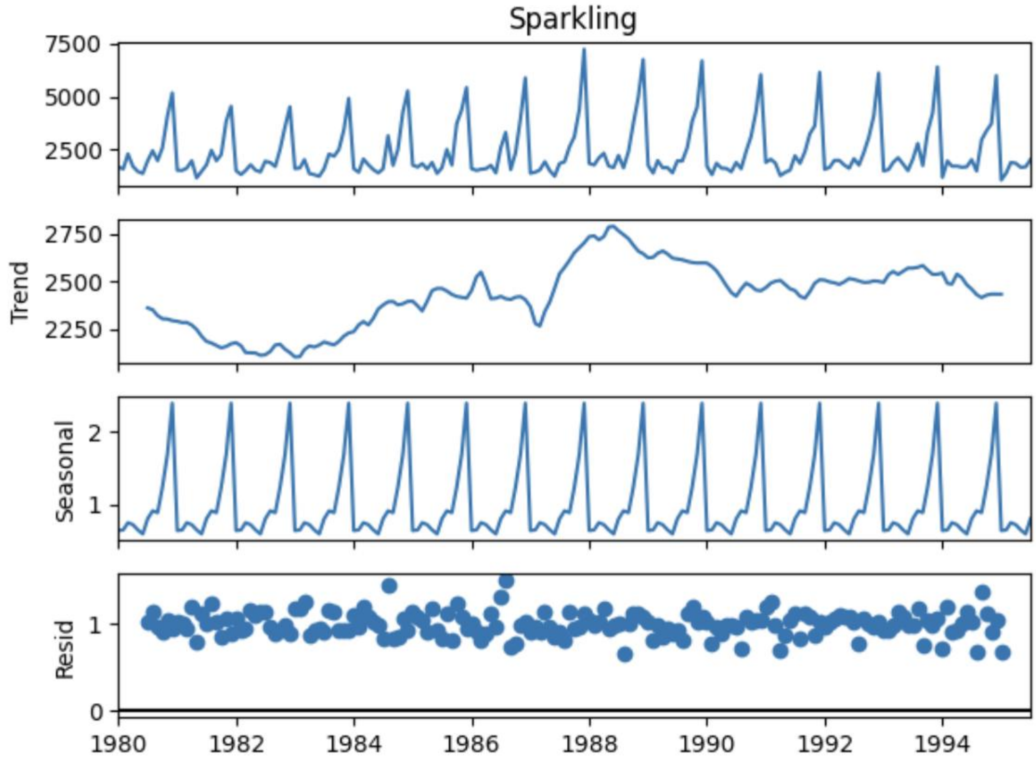

Fig 5: Seasonal decomposition of Sparkling wine (Multiplicative model)

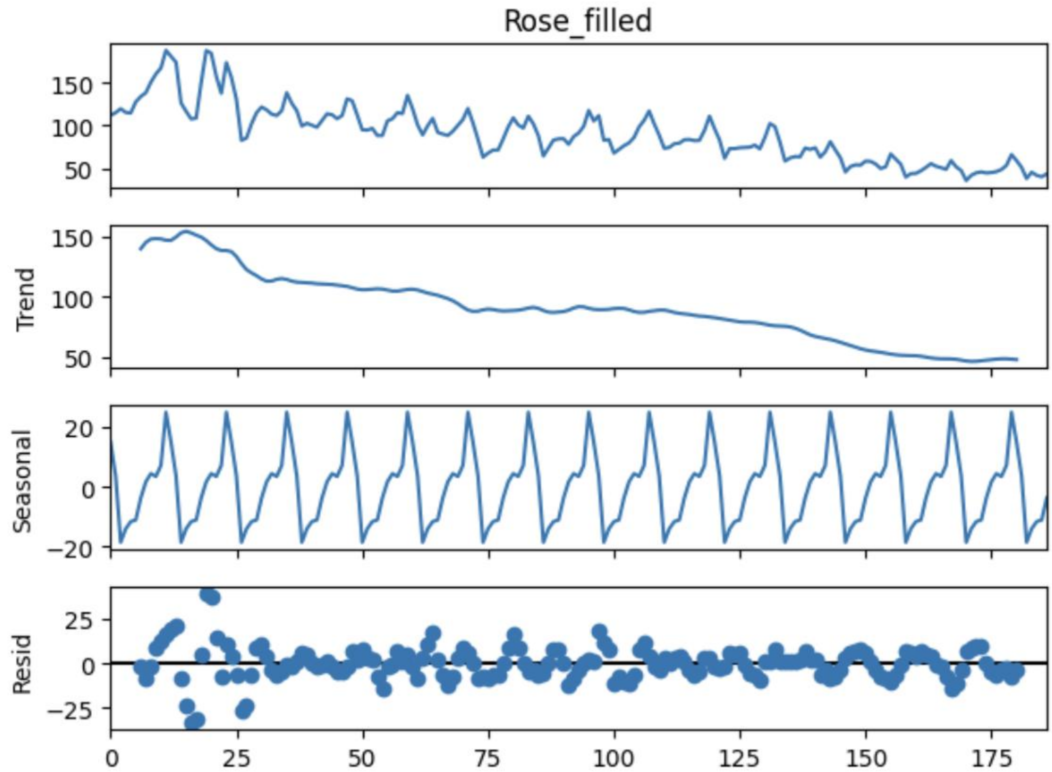

Fig 6: Seasonal decomposition of Rosy wine (Additive model)

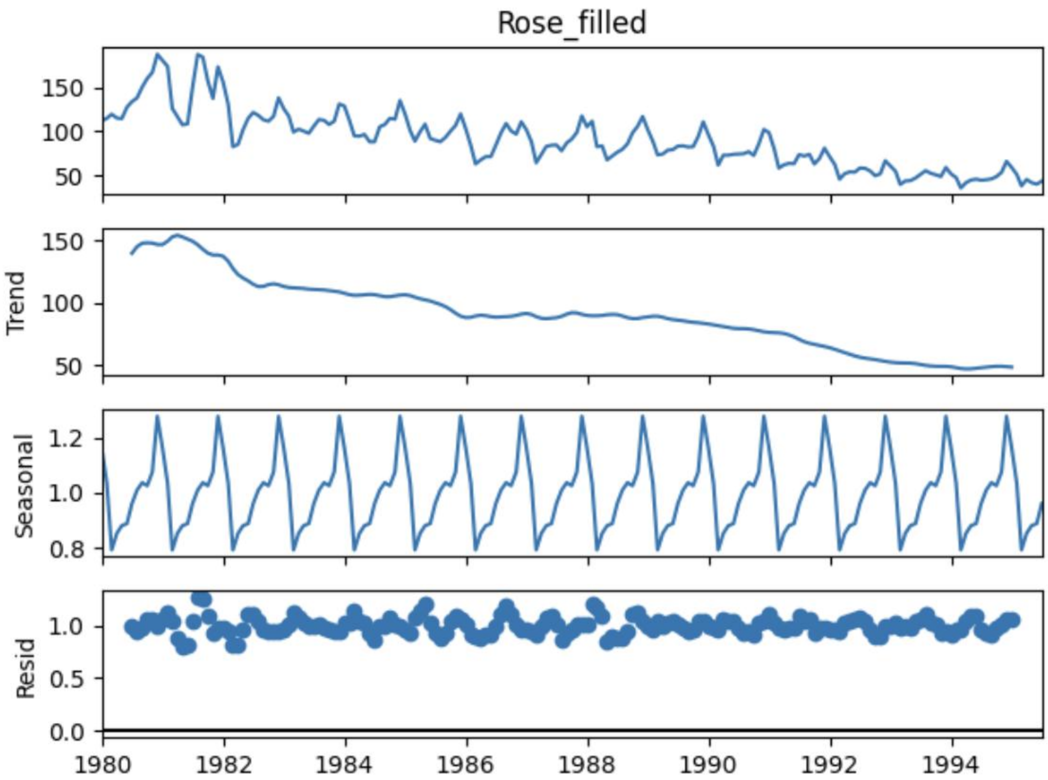

Fig 7: Seasonal decomposition of Rosy wine (Multiplicative model)

For Sparkling wines, we see there in an increasing trend over the years in sales and for Rosy wines we see there is a decreasing trend over the years. The time series decomposition for the Rosy wines is done on the imputed data.

Data Splitting for Sparkling and Rose Wine Sales: Training and Test Sets

We proceeded to split the imputed Rose wine sales data into training and test sets. This division was crucial for assessing the performance of forecasting models. We specifically set the test data to start from the year 1991, allowing us to evaluate the models' accuracy on more recent sales data.
For Rose wine, the training set has 132 entries with 2 features, and the test set includes 55 entries with 2 features. In the case of Sparkling wine, the training set consists of 132 entries with 1 feature, and the test set has 55 entries with 1 feature.

Model Building and Evaluation:

We started our investigation of sparkling wine sales with Simple Exponential Smoothing (SES), which offered a baseline projection based on a weighted average of historical data. That being said, the Root Mean Squared Error (RMSE) of 1304.93 suggests that SES's accuracy was just moderate. After that, we looked at Holt's Double Exponential Smoothing, an extension of SES that includes a trend component. With an RMSE of 5291.88, the model did not perform well and might not be appropriate for this dataset. Next, multiplicative and additive Holt-Winters models were used. Better results were obtained with an RMSE of

378.95 for the Holt-Winters model with additive seasonality, and an RMSE of 403.41 for the multiplicative variant.
For Rose wine sales, SES provided a baseline forecast with a moderate accuracy, as indicated by an RMSE of 48.75. Holt's Double Exponential Smoothing was applied, resulting in a relatively poor performance with an RMSE of 46.31. The Holt-Winters model with additive seasonality demonstrated moderate accuracy, producing an RMSE of 17.65.

Below are the Time series graphs with predictions for all the exponential smoothing models built.

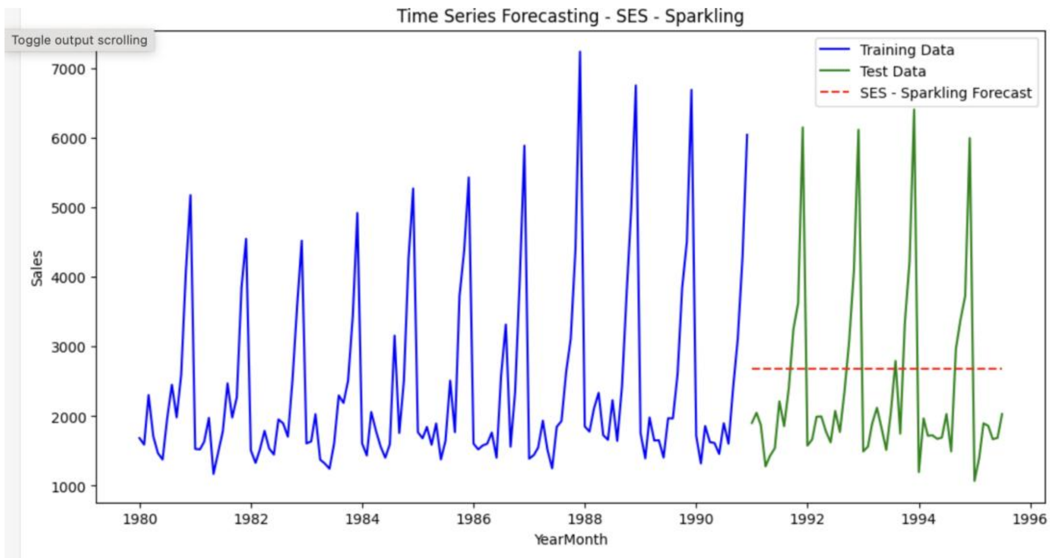

Fig 8: Simple Exponential Smoothing (Sparkling)

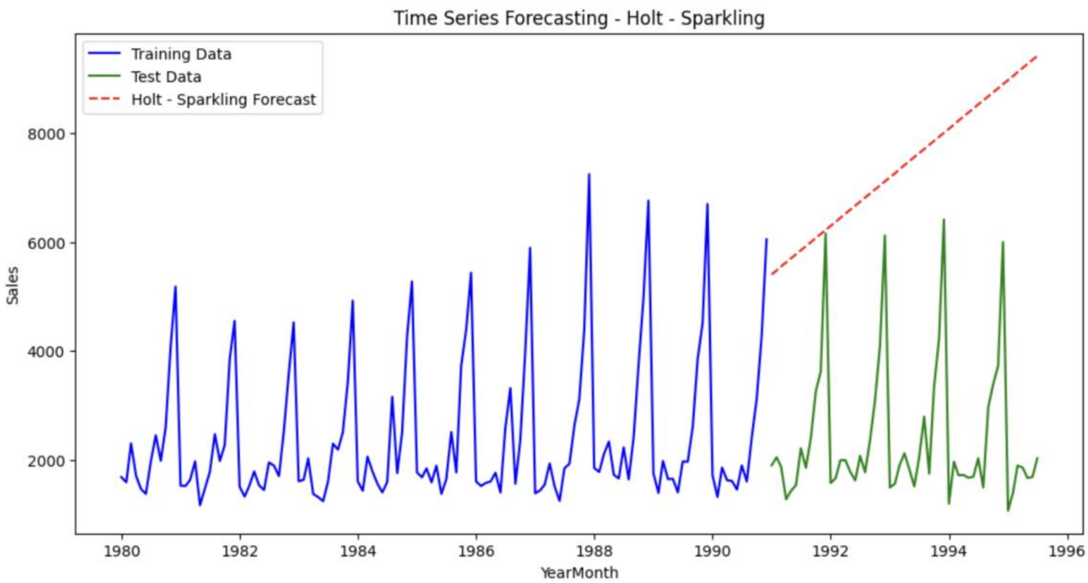

Fig 9: Holt's Double Exponential Smoothing (Sparkling)

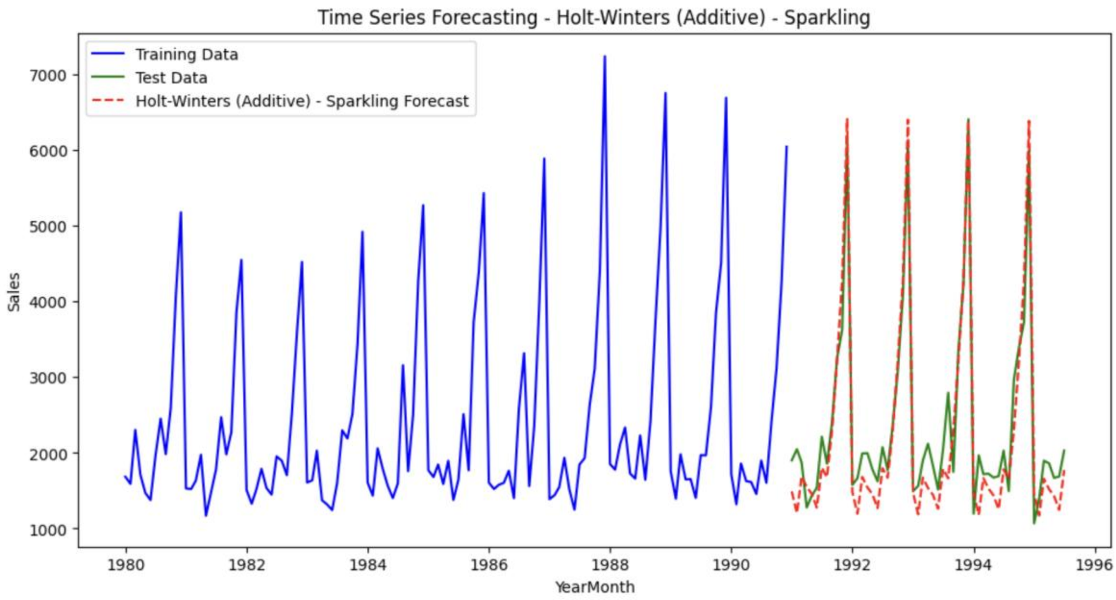

Fig 10: Holt Winters -Additive (Sparkling)

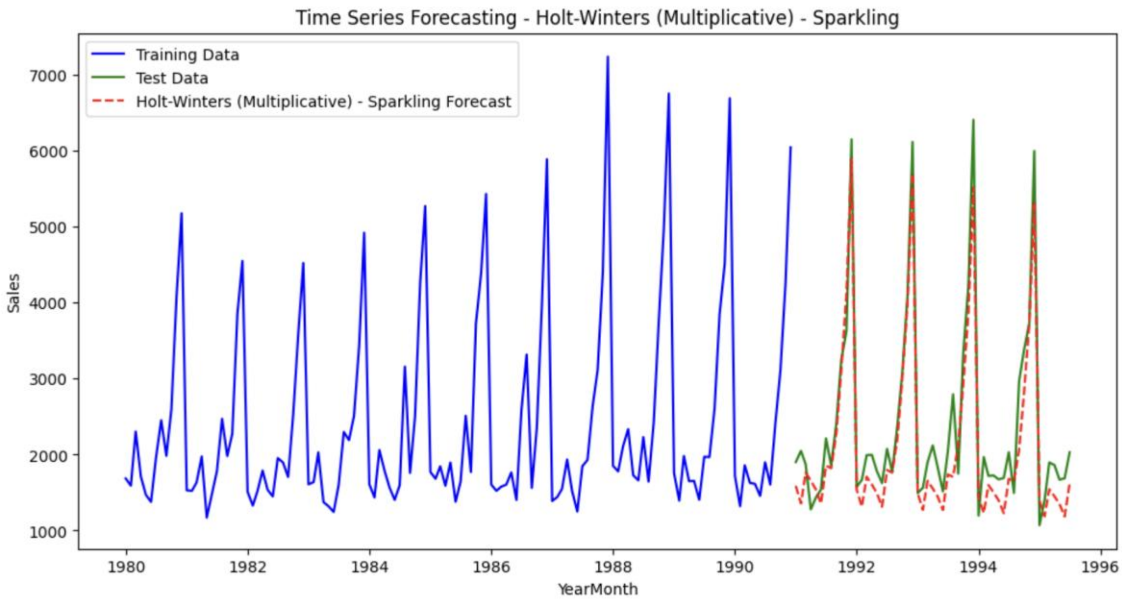

Fig 11: Hot Winters- -Multiplicative (Sparkling)

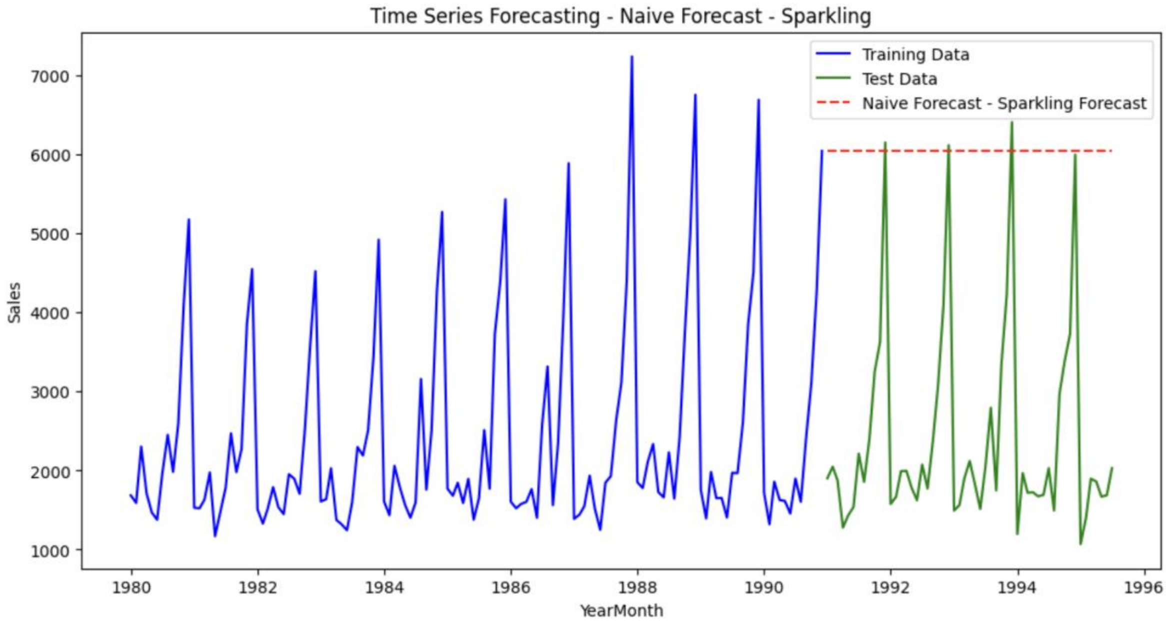

Fig 12: Naïve Forecast (Sparkling)

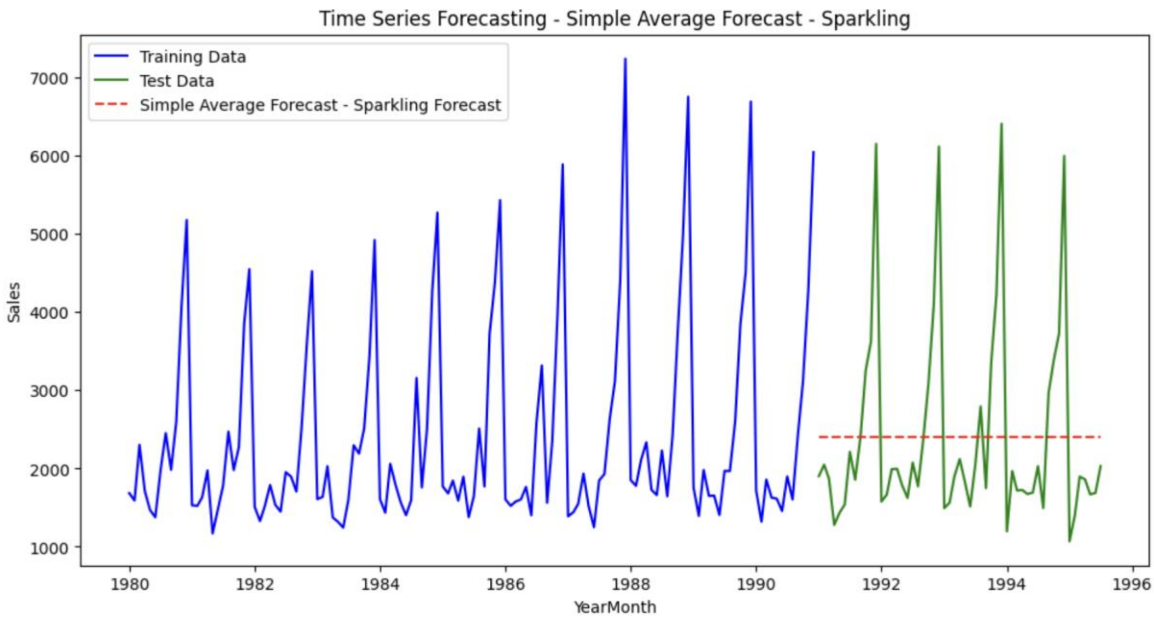

Fig 13: Simple Average forecasting (Sparkling)

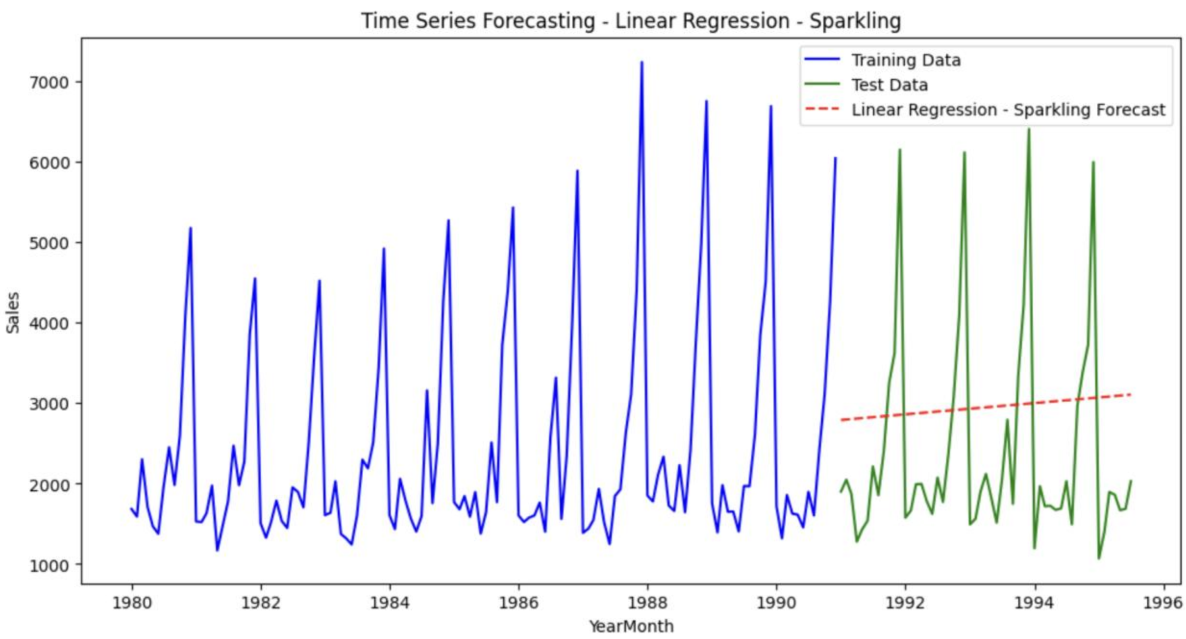

Fig 14: Linear Regression Forecast (Sparkling)

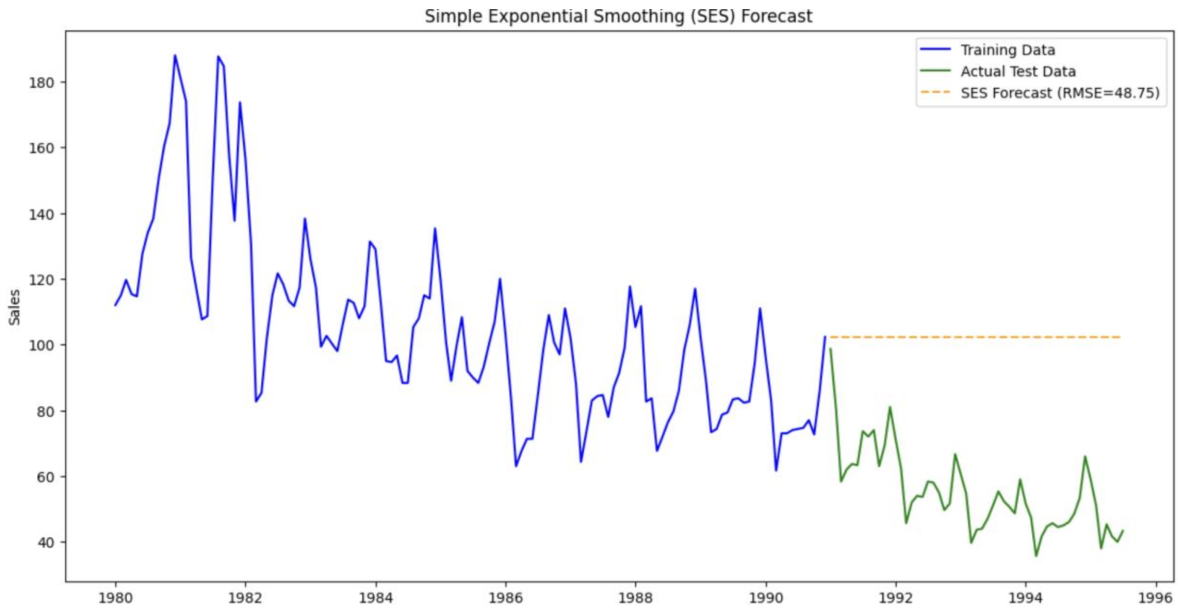

Fig 15: Simple Exponential Smoothing (Rose)

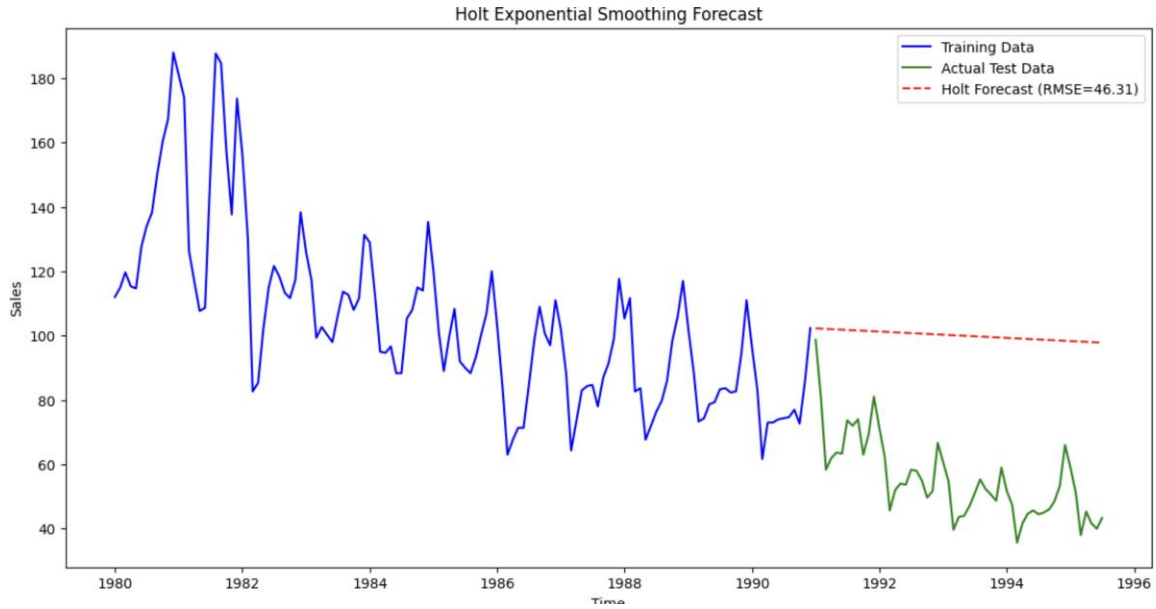

Fig 16: Holt Winters -Additive (Rose)

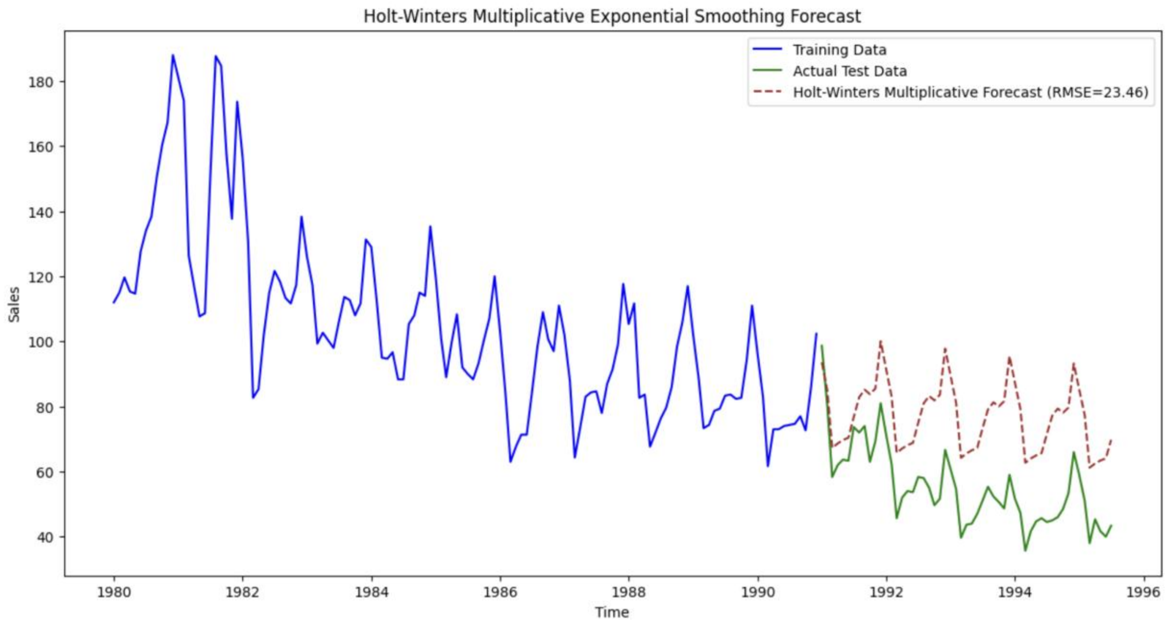

Fig 17: Holt Winters -Multiplicative (Rose)

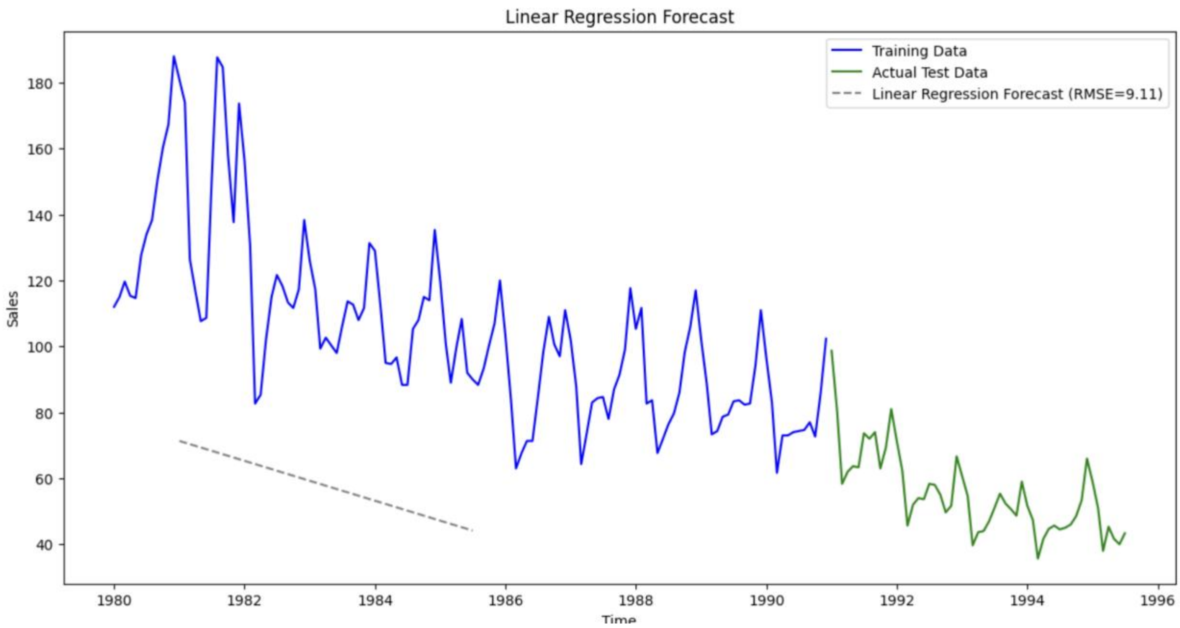

Fig 18: Linear Regression (Rose)

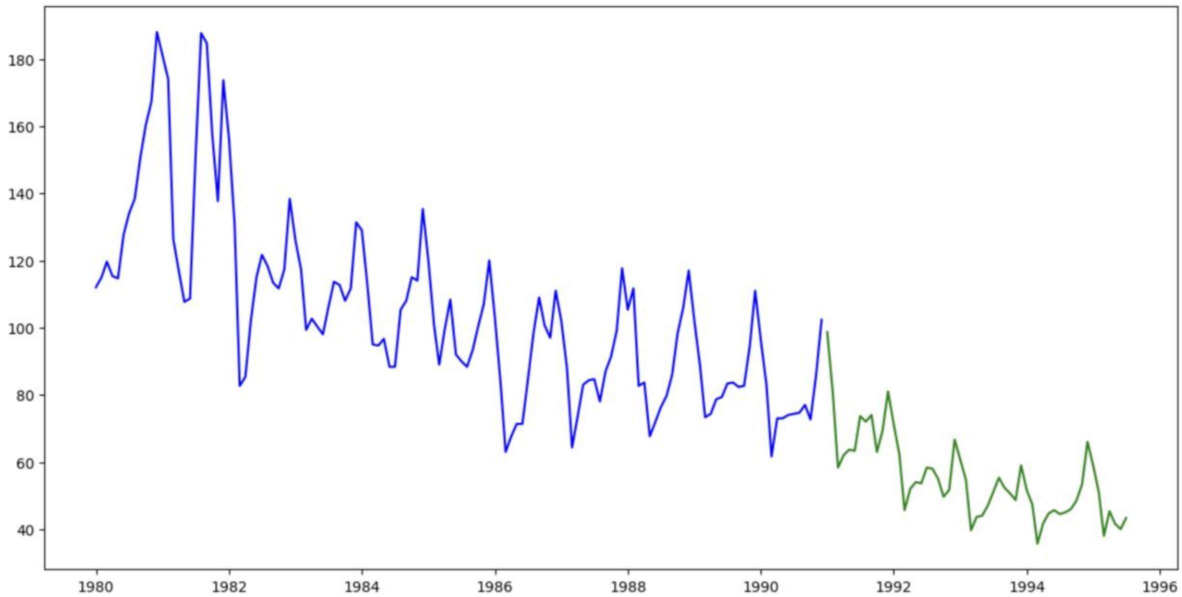

Fig 19: Naïve Forecast (Rose)

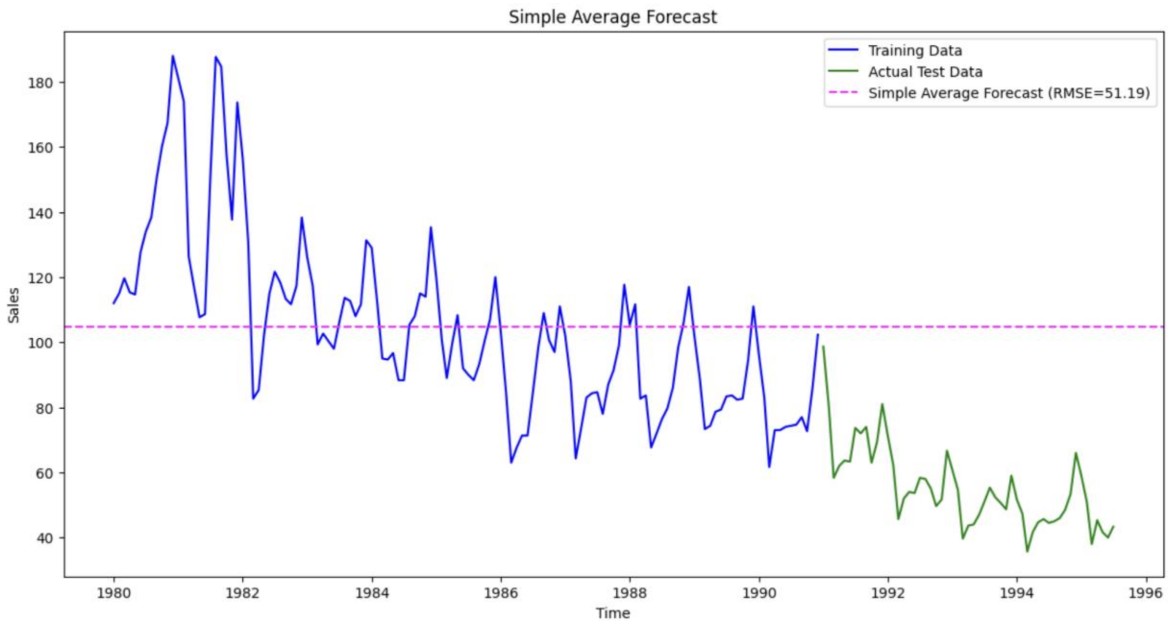

Fig 20: Simple Average Forecast (Rose)

Achieving Stationarity in Time Series Data

Stationarity is a crucial prerequisite for reliable time series modelling. The Augmented Dickey-Fuller (ADF) test is employed to assess and enhance stationarity. This section discusses the application of the ADF test and the subsequent steps taken to achieve stationarity in Sparkling and Rose wine sales data.

The initial ADF test on the original Sparkling data indicated non-stationarity, with an ADF statistic of approximately -1.21 and a p-value of 0.67. To address this, differencing was applied to the Sparkling data, resulting in an ADF statistic of -8.01 and a p-value below 0.05, signifying the successful attainment of stationarity through differencing.

In the case of Rose wine sales, the ADF test on the original Rose data suggested initial stationarity, with an ADF statistic of -1.74 and a p-value of 0.41. However, further analysis on the differenced Rose data, which produced an ADF statistic of -3.62 and a p-value of 0.005, emphasized the significance of differencing to enhance stationarity.

Data Type	**ADF Statistic**	**P-Value**	**Stationarity**
Original	-1.21	0.67	No (Fail to reject null hypothesis)
Differenced	-8.01	< 0.05	Yes (Reject null hypothesis)

Table 1: Stationarity Assessment Results: Sparkling Wine Sales

Data Type	ADF Statistic	P-Value	Stationarity
Original	-1.74	0.41	Yes (Fail to reject null hypothesis)
Differenced	-3.62	0.005	No (Reject null hypothesis)

Table 2: Stationarity Assessment Results: Rose Wine Sales

ARIMA Model Selection for Wine Sales Forecasting

For the Sparkling and Rose wine datasets in the time series study of wine sales, an automated ARIMA modeling approach was used. The approach makes use of the pmdarima library's auto_arima function, which does a stepwise search to find the ARIMA parameters that produce the lowest Akaike Information Criteria (AIC) on the training set. The model takes seasonality into account and uses a trace to show the stepwise selection procedure.

Sparkling Wine Sales: Automated ARIMA

For Sparkling wine sales, the auto_arima function determined the optimal ARIMA order to be (0, 0, 1)(0, 1, 1)[12] with intercept. This automated selection, based on AIC minimization, ensures a robust and data-driven forecasting model. The resulting Root Mean Squared Error (RMSE) of 355.93 demonstrates the model's accuracy in predicting Sparkling wine sales, offering valuable insights for industry decision-makers.

Rose Wine Sales: Automated ARIMA

Similarly, the automated ARIMA modelling for Rose wine sales identified the optimal ARIMA order as (0, 1, 1)(1, 0, 1)[12] with intercept. By minimizing AIC through a stepwise search, the model achieved an RMSE of 17.24, indicating its effectiveness in capturing the temporal patterns of Rose wine sales. The automated ARIMA approach contributes to accurate forecasting and strategic decision-making in the Rose wine market.

Wine Type	Model	ARIMA Order	RMSE
Sparkling	Automated ARIMA	(0, 0, 1)(0, 1, 1)	355.93
Rose	Automated ARIMA	(0, 1, 1)(1, 0, 1)	17.24

Table 3: ARIMA Model Results: Wine Sales

Time series graphs were generated to visually represent the ARIMA model results. These plots showcase the training data, actual test data, and ARIMA forecast, providing a clear and intuitive visualization of the wine sales trends over time.

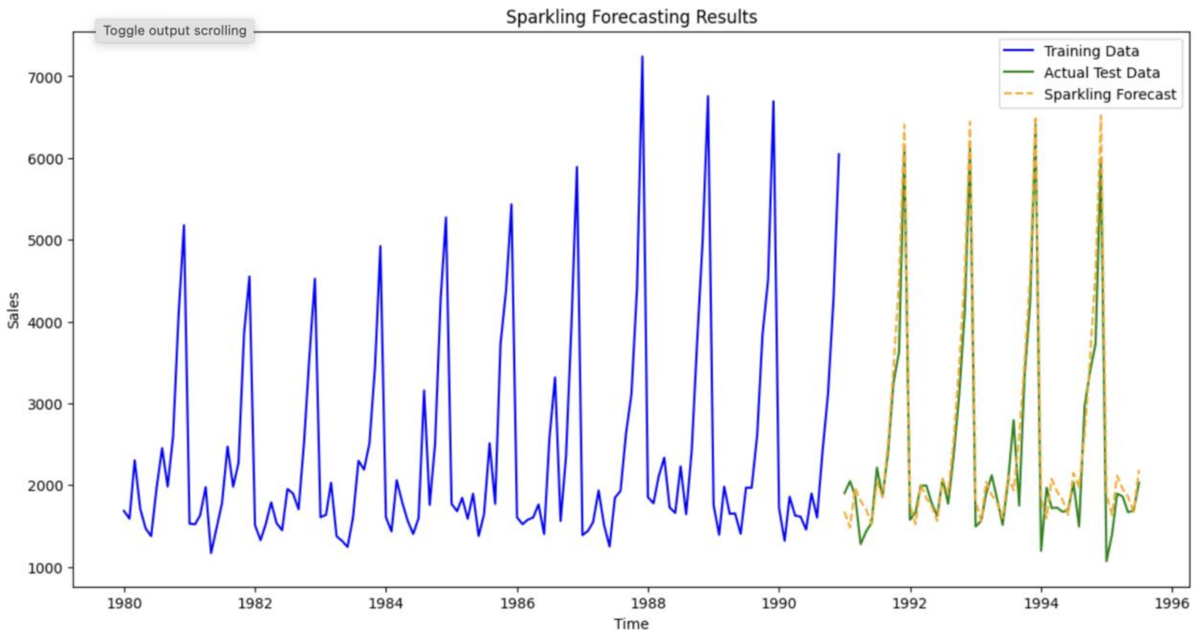

Fig 20: ARIMA model Time series graph for Sparkling

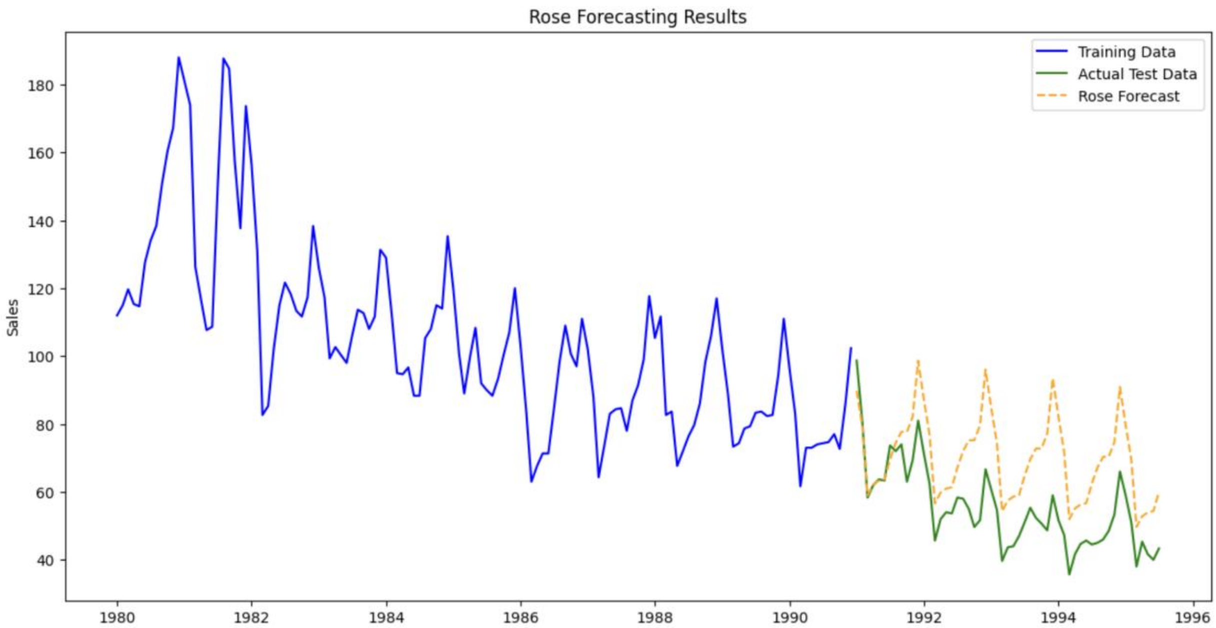

Fig 21: ARIMA model Time series graph for Rose

Sparkling Wine Sales Model Results:

In this analysis, various time series forecasting models were applied to predict Sparkling wine sales. The table below presents the models, their respective parameters, and the Root Mean Squared Error (RMSE) on the test data:

Model	Parameters	RMSE
Exponential Smoothing (Simple)	-	1304.93
Exponential Smoothing (Double)	-	5291.88
Holt-Winters Additive	Additive Seasonal, Trend, Error	378.95
Holt-Winters Multiplicative	Multiplicative Seasonal, Trend, Error	403.41
Naive	-	3864.28
Simple Average	-	1275.08
ARIMA	Manual ARIMA	355.93

Table 4: Sparkling Wine Sales Model Comparison

The results indicate that the ARIMA model achieved the lowest RMSE among the applied models, demonstrating its effectiveness in forecasting Sparkling wine sales.

Rose Wine Sales Model Results:
Similar to the Sparkling wine sales analysis, various forecasting models were employed to predict Rose wine sales. The following table summarizes the models, their parameters, and RMSE on the test data:

Model	Parameters	RMSE
Exponential Smoothing (Simple)	-	48.75
Exponential Smoothing (Double)	-	46.31
Holt-Winters Additive	Additive Seasonal, Trend, Error	17.65
Linear Regression	Linear Regression Forecast	9.11
Naive	-	48.75
Simple Average	-	51.19
ARIMA	ARIMA Forecast	17.24

Table 5: Rose Wine Sales Model Comparison

Among these models, the Holt-Winters Additive and ARIMA models displayed particularly low RMSE, indicating their effectiveness in capturing the patterns of Rose wine sales.

ARIMA Forecasting: Optimized Model

The capability of the ARIMA (AutoRegressive Integrated Moving Average) model to identify long-term patterns as well as short-term trends in wine sales data led to its selection. It was chosen because of its exceptional prediction ability, which regularly produced lower Root Mean Squared Error (RMSE) compared to alternative models.

12 Month Future Forecast:

The best ARIMA model was constructed by utilizing the entire dataset. The wine sales for the upcoming 12 months were forecasted using this approach. With the proper confidence intervals included, the forecasts offer a reliable estimate of the expected sales trajectory.

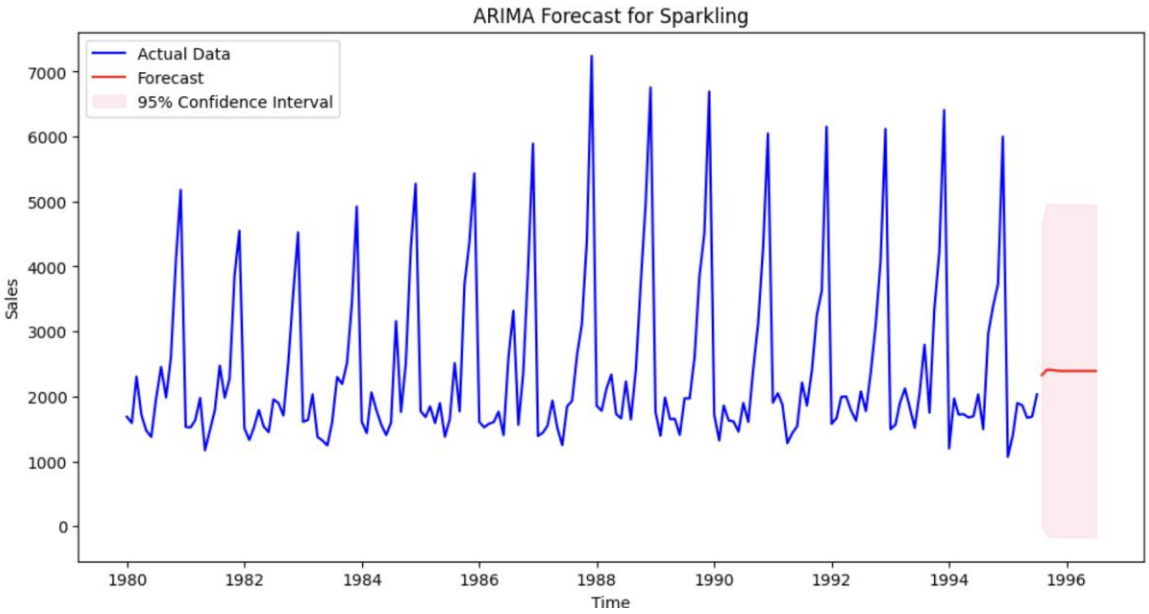

Fig 22: ARIMA Forecast: 12-Month Time Series Projection (Sparkling)

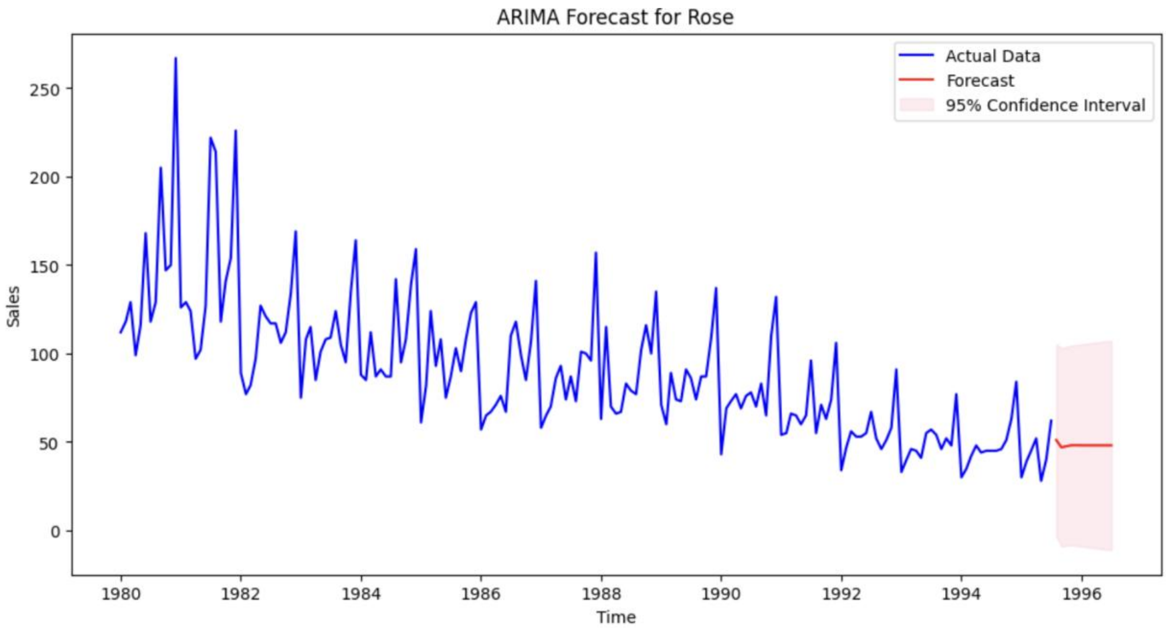

Fig 23**:** ARIMA Forecast: 12-Month Time Series Projection(Rose)

Business Insights and Recommendations:

The ARIMA model proved to be a dependable tool for accurately forecasting wine sales in both short and long terms. Its precision provides a solid basis for strategic planning, helping the company align resources effectively. Furthermore, the inclusion of confidence intervals around predictions offers valuable insights, aiding the company in anticipating and preparing for potential sales variations and uncertainties. Leveraging these insights, the business can adopt a proactive and resilient approach to navigate the dynamic wine market landscape.

Recommendations for Future Strategies:

Monitoring and Adjustment: Regularly monitor actual sales against forecasted values and be prepared to adjust model parameters in response to significant deviations.

Data Refinement: Continuously improve the model by incorporating the latest and most relevant data to enhance its predictive accuracy.

Market Analysis Integration: Explore the integration of market trends, economic indicators, or promotional activities into the model for a more comprehensive approach.

Scenario Planning: Develop scenario-based forecasts to anticipate sales under different market conditions, enabling proactive decision-making.

END OF REPORT