Time Series Forecasting

Business Report

Jayeesha Chakraborty

9/14/24

Table of Contents

Data Dictionary of Problem 1

Column Name	Description	Data Type
Year-Month	Monthly Time Stamp from Jan,1980 to July,1995	object
Sparkling	Monthly Sales of Sparkling Wine	int64

Data Dictionary of Problem 2

Column Name	Description	Data Type
Year-Month	Monthly Time Stamp from Jan,1980 to July,1995	object
Rose	Monthly Sales of Rose Wine	int64

Problem 1

Context

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties.

Objective

As an aspiring data scientist, the primary objective of this project is to analyze and forecast sparkling wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

Data Overview:

- **Shape**: There are 187 records and 2 columns.
- **Data Types**: Year-Month is time stamp of the data set hence it is object and Sales is int64
- **Independent & Target Variable**: We need to forecast the wine sales, so target variable is Sparkling
- **Check Duplicates**: There are no duplicate records in the database.
- **Check Missing Values:** There are no missing values in the dataset.
- **Date Time Index:** Year-Month column has been converted to index as it attributes to time-based sales.
- **Statistical Description:**

Metric	count	mean	std	min	25%	50%	75%	max
Sparkling	187	2402.17	1295.11	1070.0	1605	1874	2549	7242

Observations:

- ✓ Mean and Median of the sales are not close, hence there seem to have some extreme variations.
- ✓ Sales ranges from as low as 1070 to max 7242

Exploratory Data Analysis:

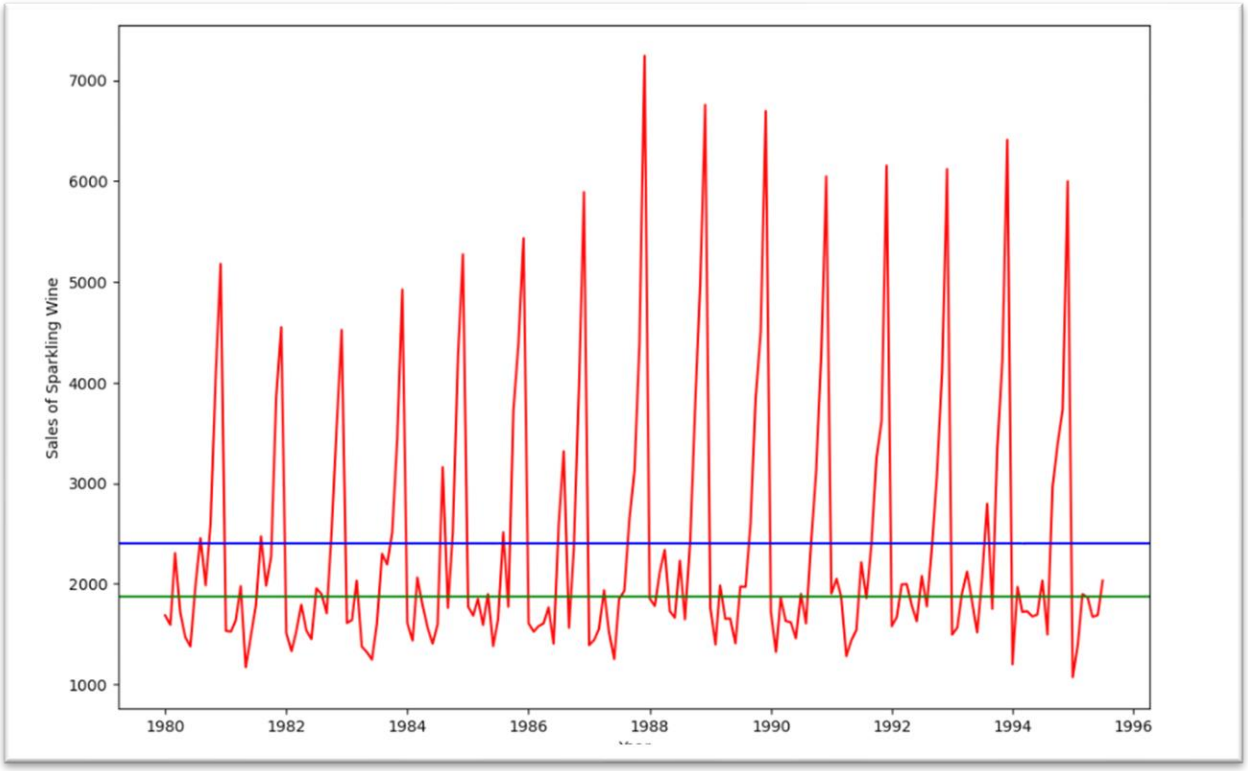

Fig 1. Wine Sales Trend of 20th Century

Observation:

- ✓ There is a hybrid trend in wine sales over the 20th century.
- ✓ There is a strong seasonal pattern observed every year.
- ✓ There is a strong variation in the sales around the avg of overall sales.

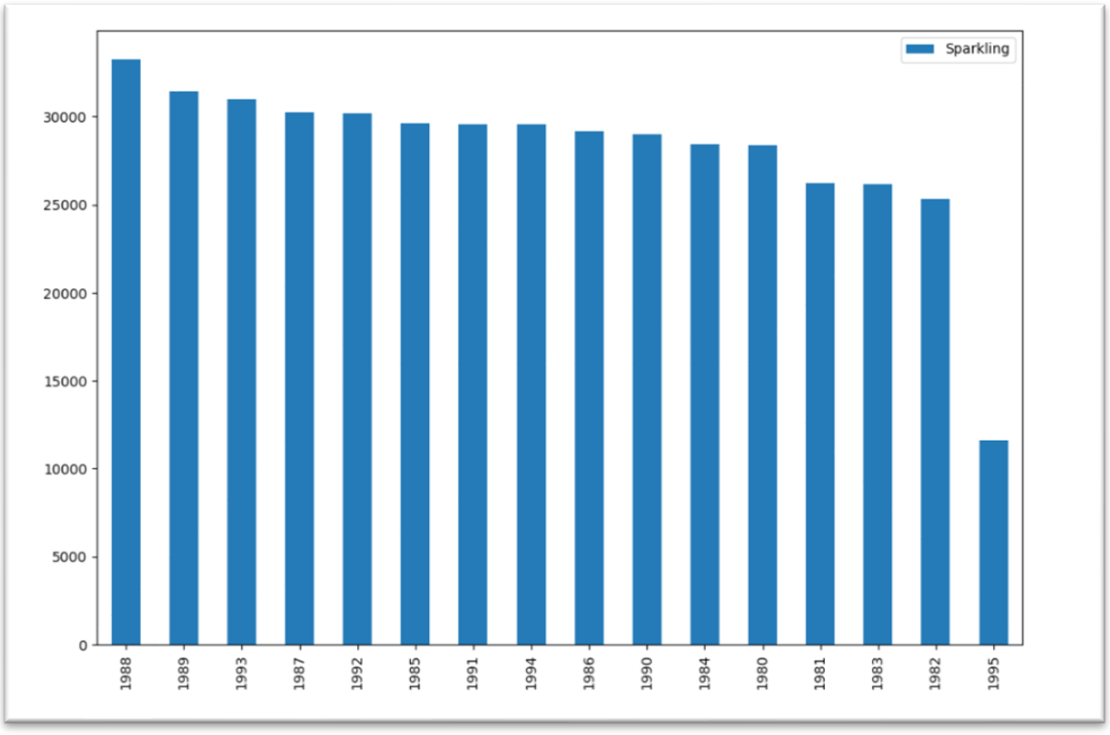

Fig 2. Year Wise Wine Sales of 20th Century

Observations:

- ✓ Maximum wine sales happened in the year of 1988
- ✓ We have data only till July,1995, minimum wine sales year can't be concluded.

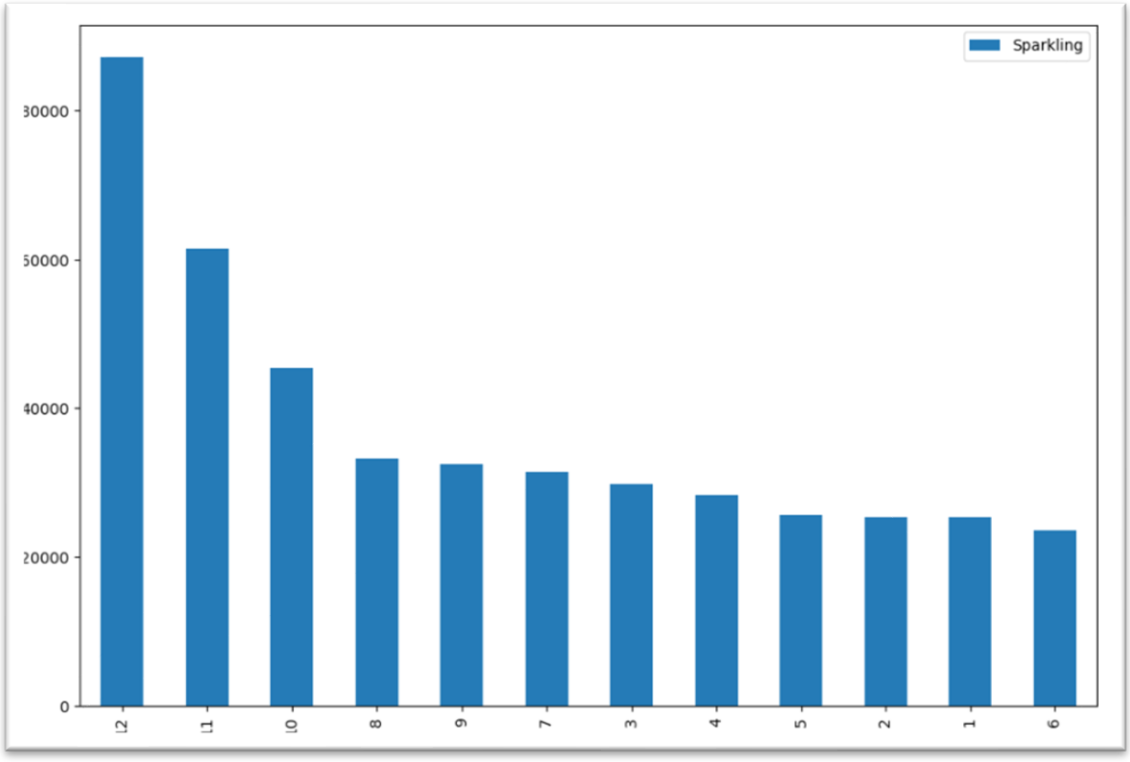

Fig 3. Month Wise Wine Sales of 20th Century

Observations:

- ✓ Wine sales typically see a spike during the holiday season, particularly in December. This trend is consistent across various years**.**
- ✓ Wine sales have reduced significantly in the other months compared to December.

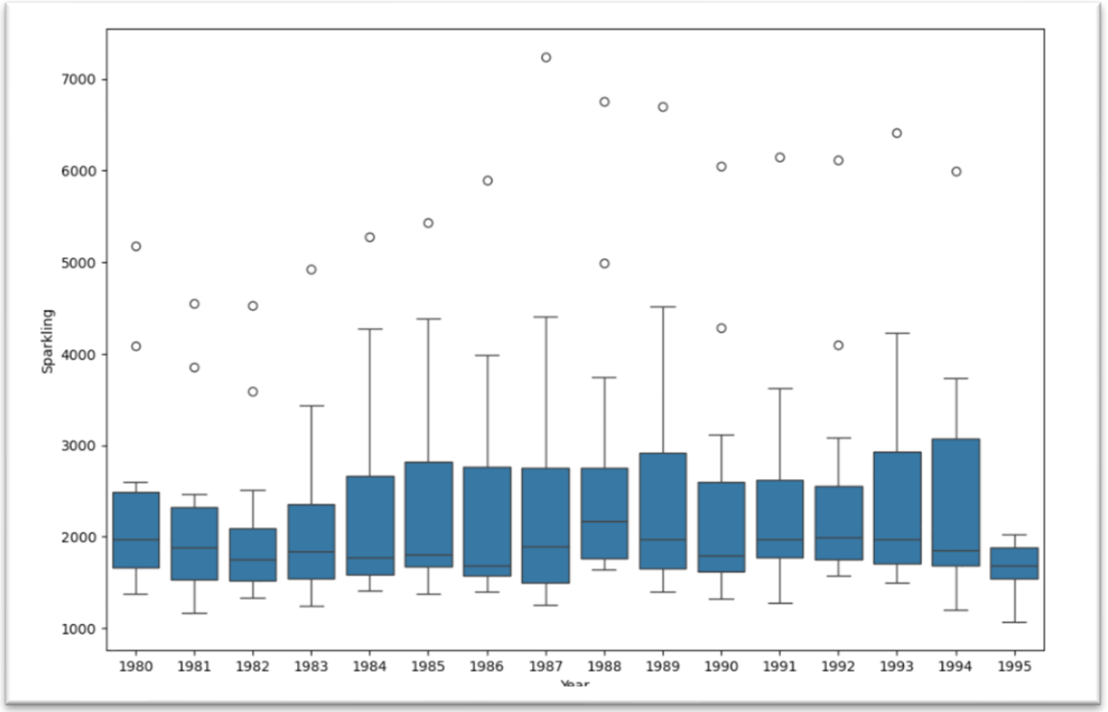

Fig 4. Yearly Sales Pattern in Box Plot

Observations:

- ✓ Minimum wine sales are highest in 1988 compared to other years.
- ✓ Very less sales happened in the year 1982

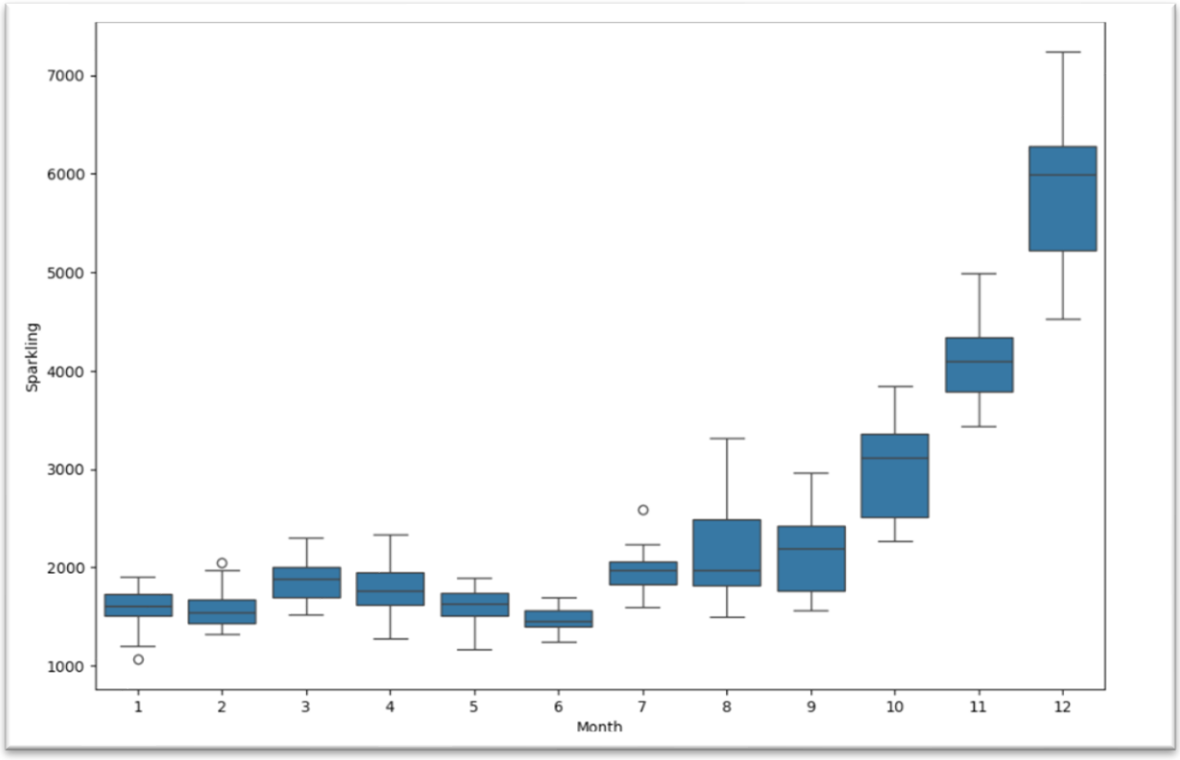

Fig 5. Monthly Sales Pattern in Box Plot

Observations:

- ✓ Sales performance peaks significantly in December, surpassing all other months.
- ✓ June experiences the lowest volume of wine sales.

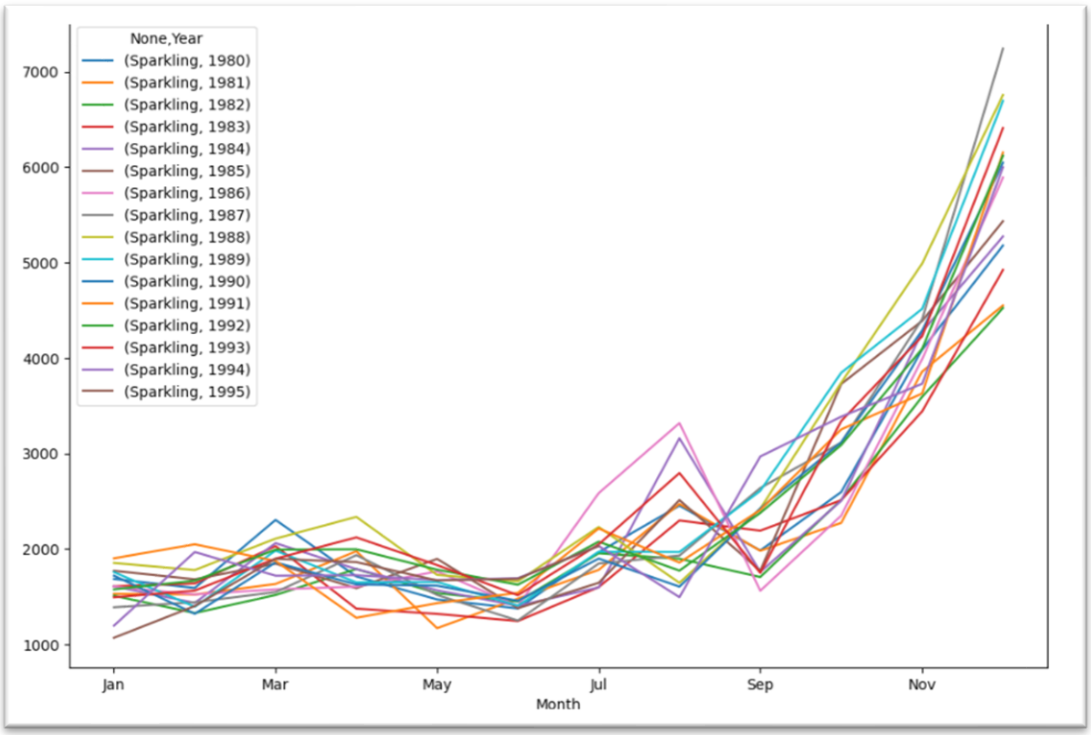

Fig 6. Monthly Sales Trend in 20th Century

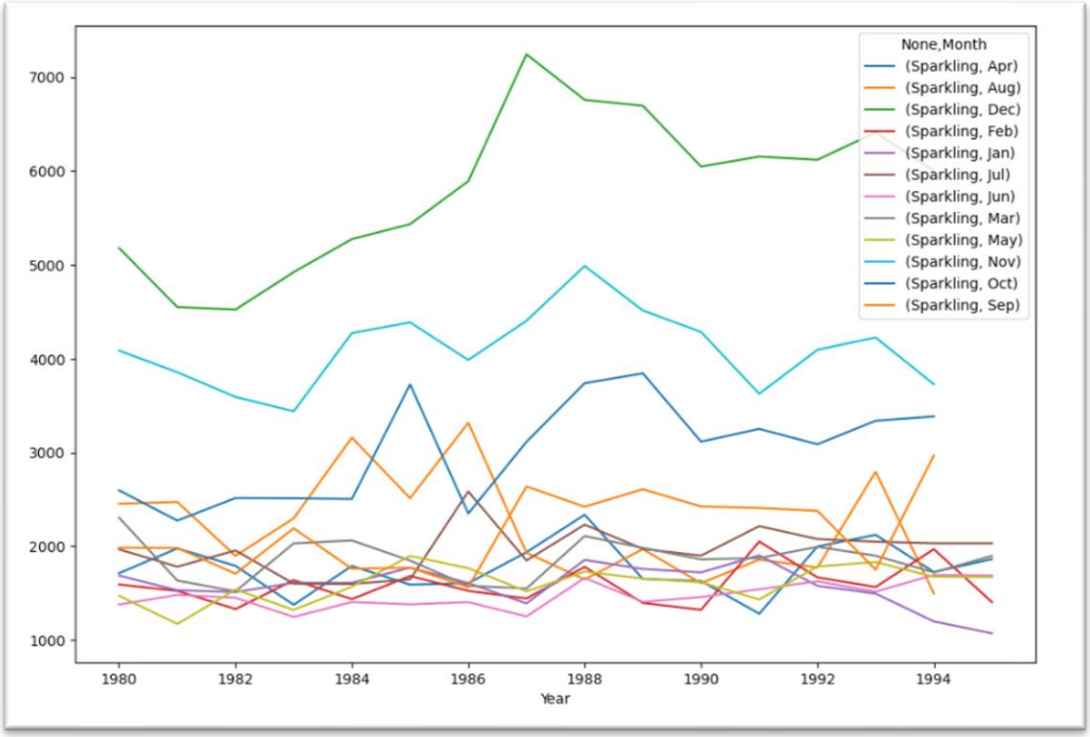

Fig 7. Yearly Sales Trend in 20th Century

Decomposition:

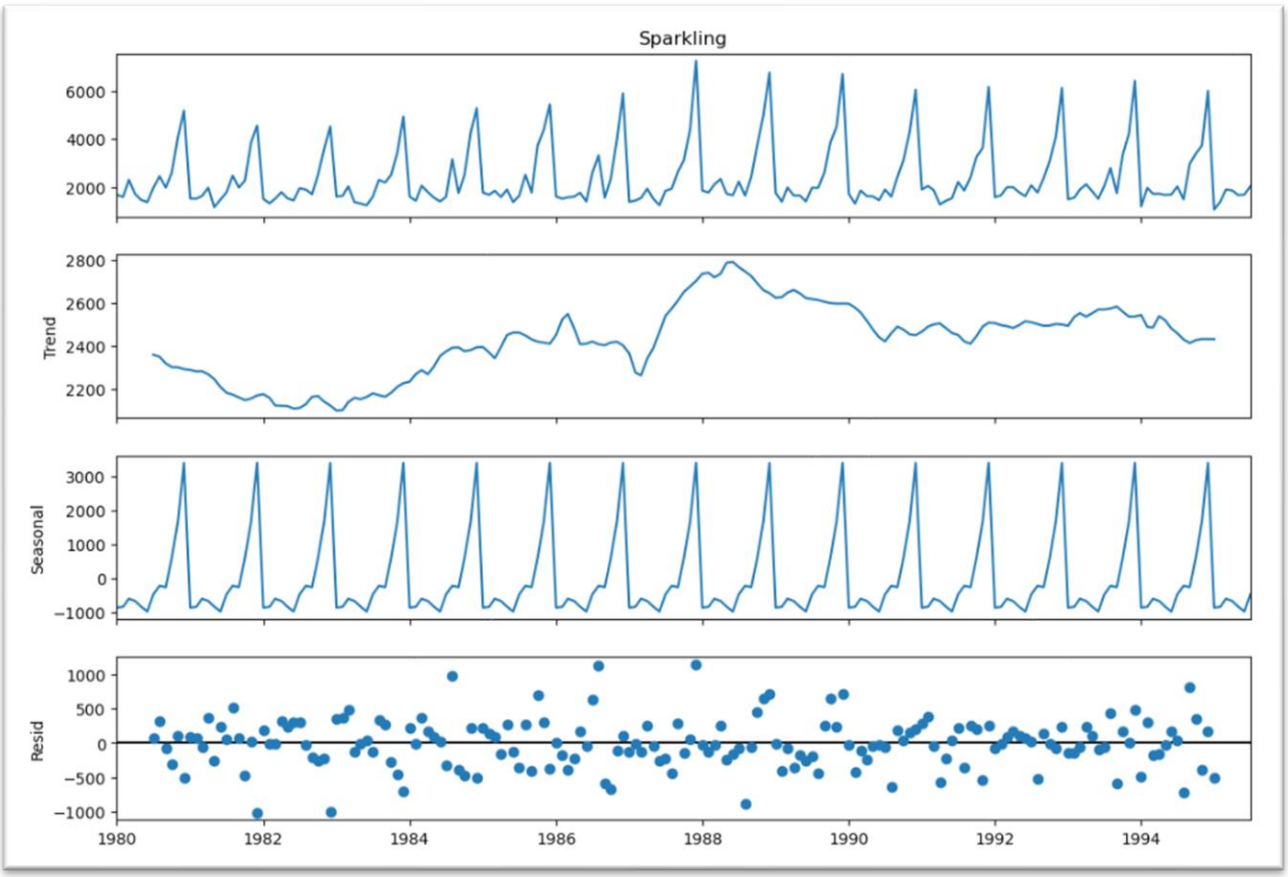

Fig 8. Additive Seasonal Decomposition

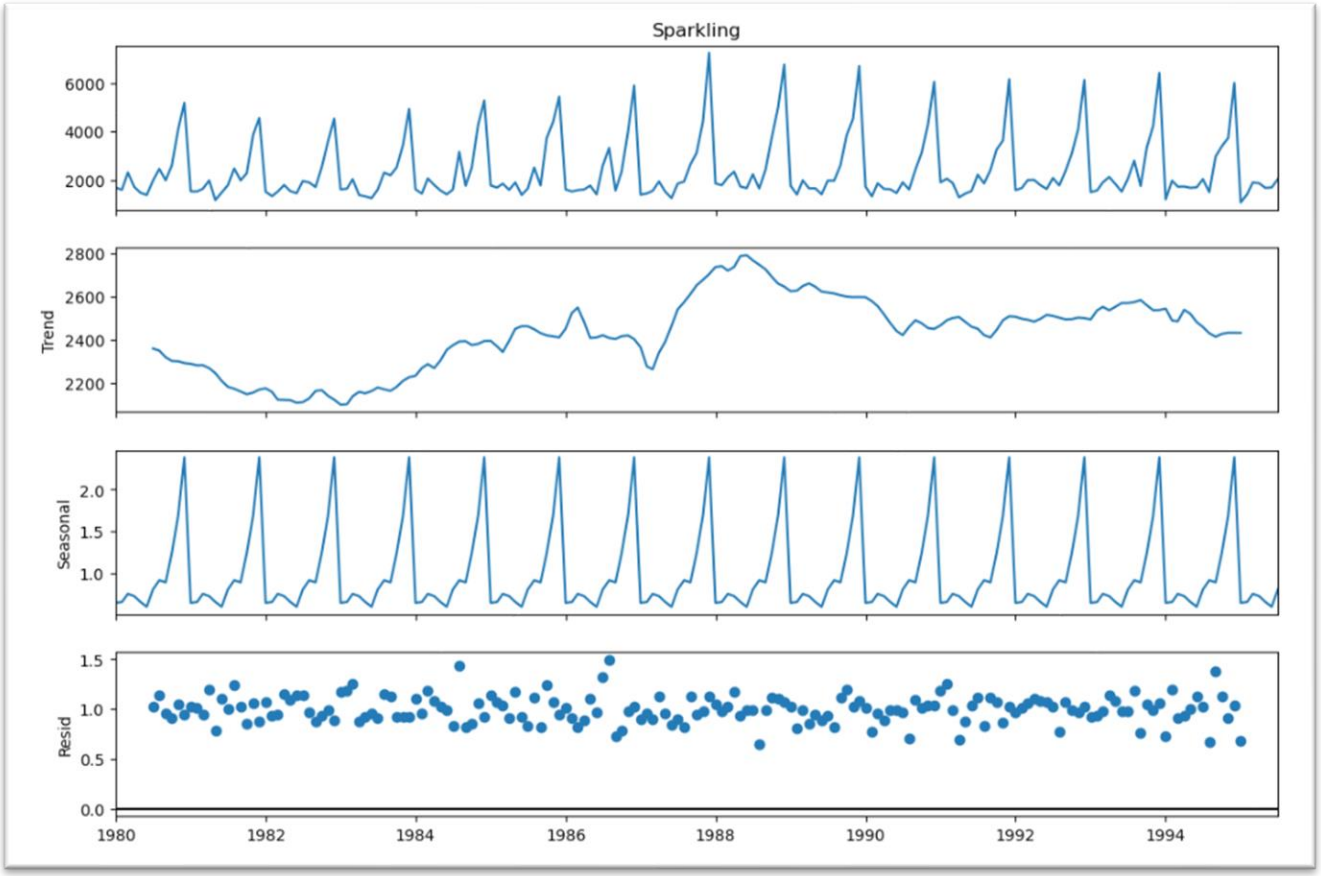

Fig 9. Multiplicative Seasonal Decomposition

Observations:

- ✓ In the decomposition process, the raw data gets split into the trend (the pattern observed over a long period of time), seasonality (pattern repeats in equal interval of time) and noise.
- ✓ As we see the residual pattern is similar in both the process additive and multiplicative. So, we proceed to analyze the data with additive process because of less complexity and computation time.

Data Pre-processing:

Missing Value Treatment:

As we saw earlier there are no missing values in the dataset.

Train Test Split:

- ✓ We split the data into train and test data set with train size of 80% which means 80% data is being
  Used for model training and 20% for testing.
- ✓ There are 149 records in the train data set and 38 records in the test.

Model Building – Original Data:

Linear Regression Model:

We used 80% of the data to train the linear regression model.

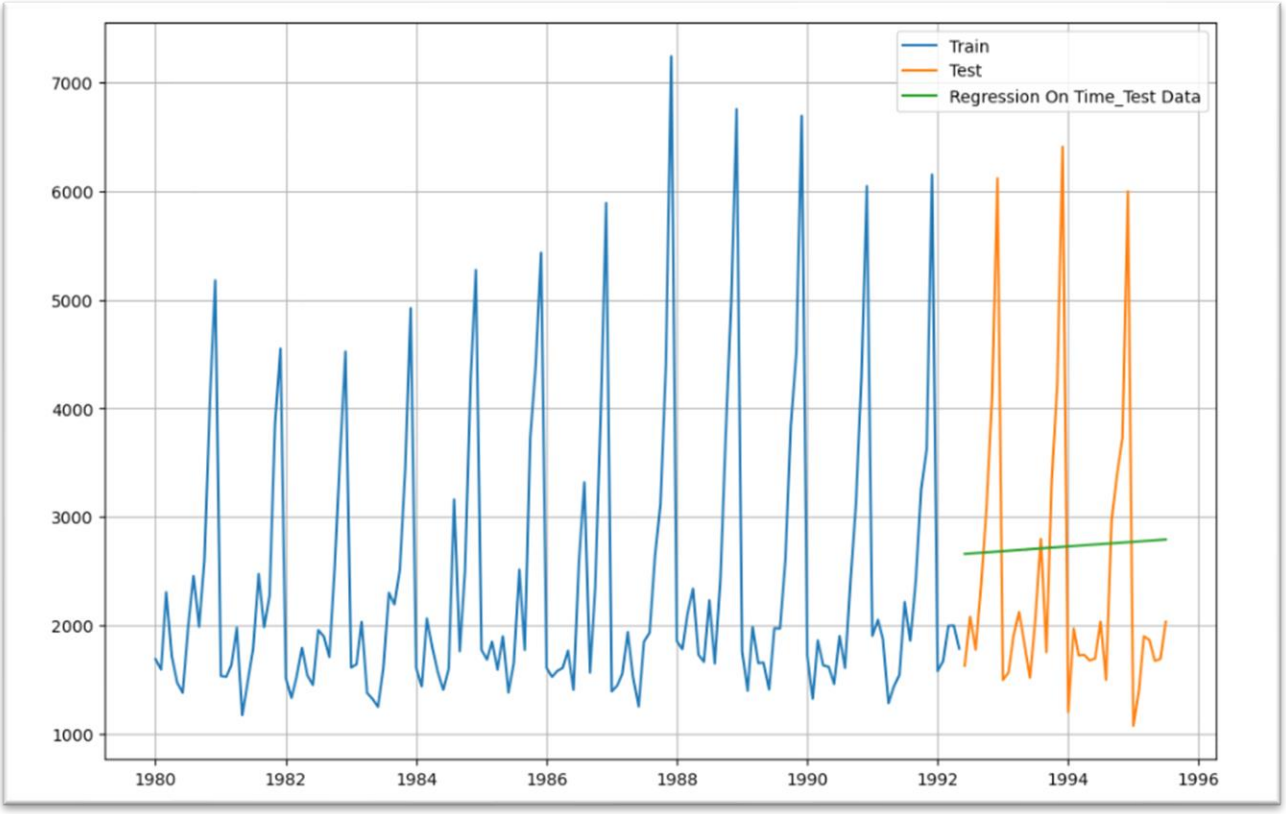

Fig 10. Sales Forecast from 1993 to 1996 using linear regression

Observation*:*

- ✓ As we see, forecasted sale using linear regression model is straight line which fails to predict the variation in the sales.
- ✓ The root mean squared error obtained from the model is 1359.7

Simple Average Model:

- ✓ This model forecast future sales as the average of all the sales made so far.
- ✓ So forecasted sales will be constant over time. It fails to capture the variation in sales in the future as it doesn't consider the same pattern in the train data.
- ✓ The root mean squared error obtained from the model is 1331

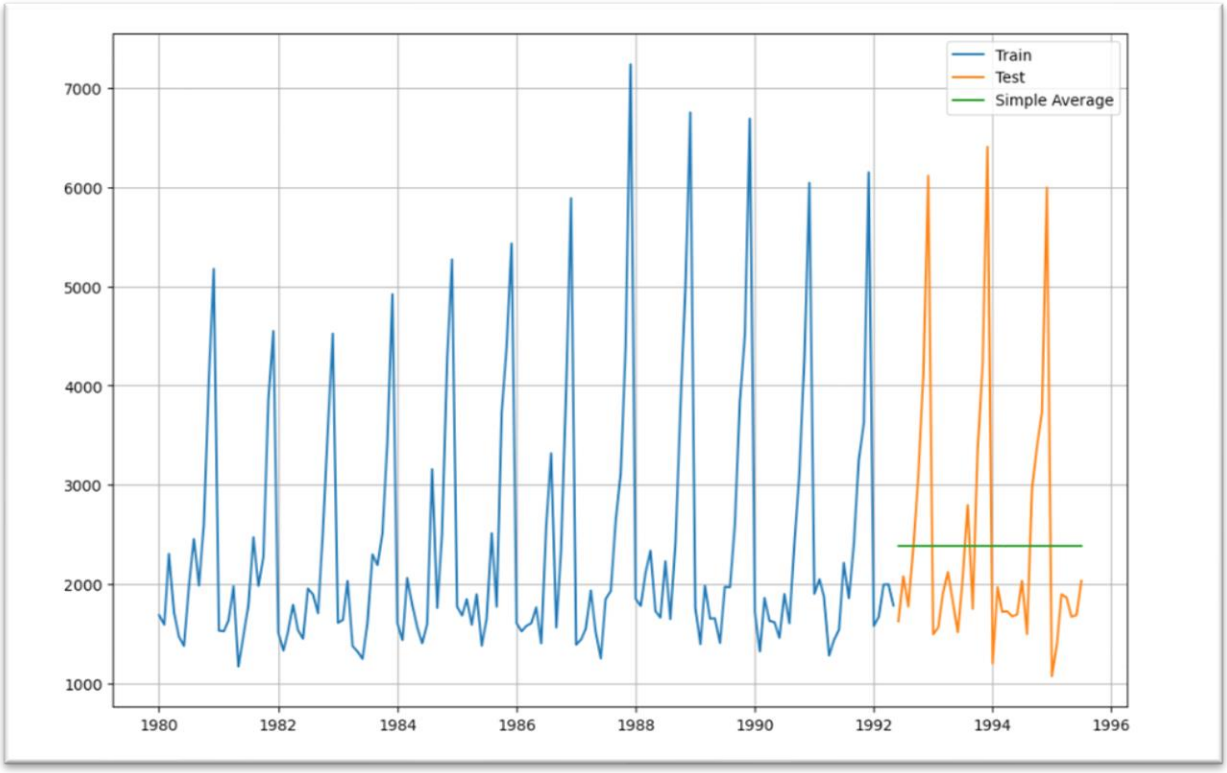

Fig 11. Sales Forecast from 1993 to 1996 using Simple Average

Moving Average Model:

- ✓ This model predicts sales at any given time point by calculating the moving average over the last specific time intervals (e.g., 3, 6, or 9 months).
- ✓ We have calculated a 3-month, 6- and 9-month moving average to forecast future sales.
- ✓ The root mean squared error obtained from 3-month model is 1026.53, 6-month model is 1290.6- and 9-month model is 1375.5
- ✓ As we see in the below plot, the model captures the variation to some extent in the sales in the future.

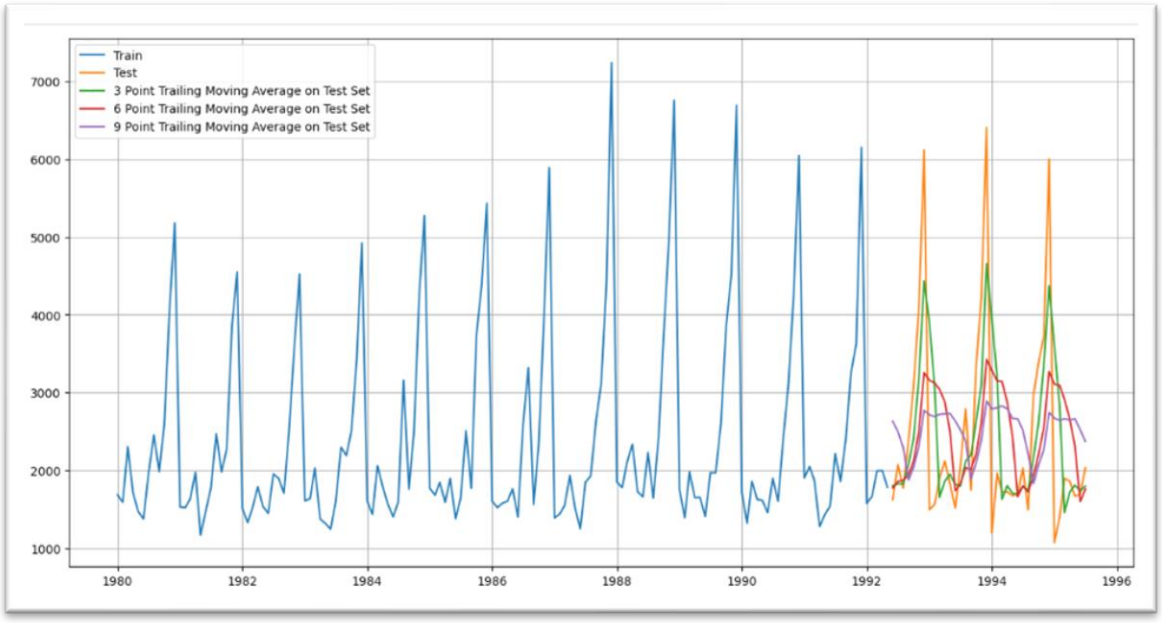

Fig 12. Sales Forecast from 1993 to 1996 using Moving Average

Simple Exponential Smoothing:

- ✓ The model does not consider trend or seasonality in the data set. It forecasts future values by giving **more weight to recent observations** while gradually (exponentially) decreasing the weight of older data points.
- ✓ We iterate through smoothing level factor ranging from 0.01 to 1 in increments of 0.01, selecting the model with the lowest RMSE value as the best fit.
- ✓ The best fit model has alpha (smoothing level factor) 0.02 and RMSE 1279.49
- ✓ As we see in the graph, the model accounts for only level change.

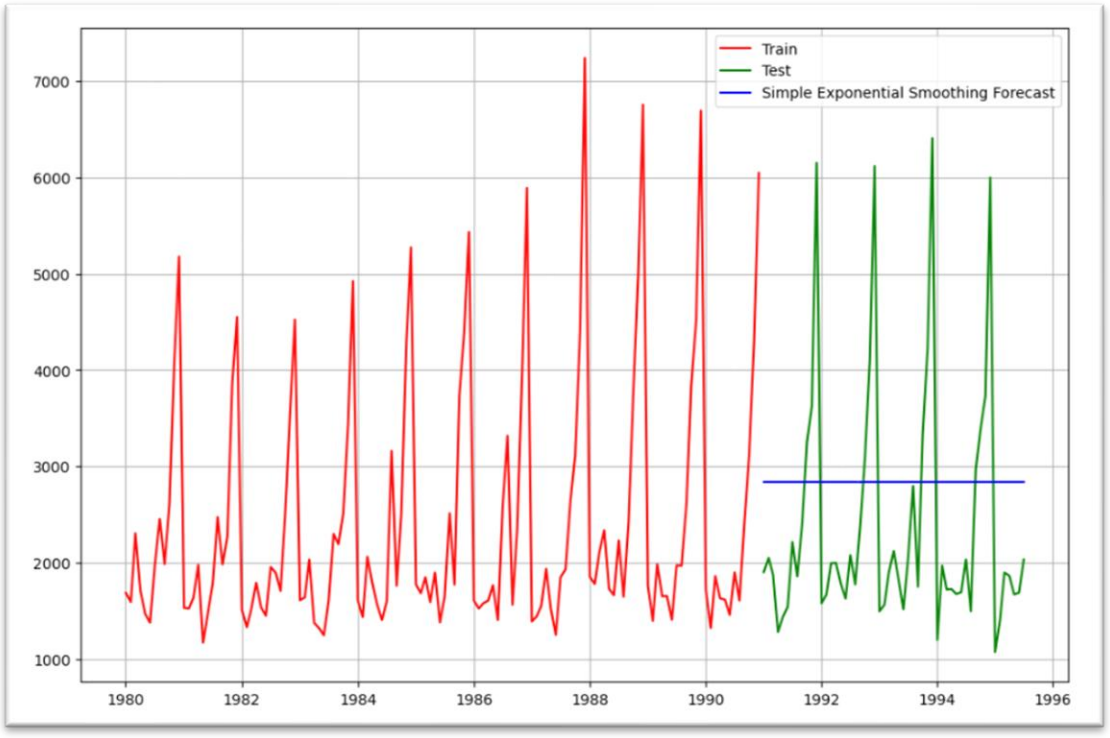

Fig 13. Sales Forecast from 1993 to 1996 using Simple Exponential Smoothing

Double Exponential Smoothing:

- ✓ Double Exponential Smoothing enhances simple exponential smoothing by incorporating a trend component, making it ideal for forecasting time series data with linear trends but no seasonality.
- ✓ We iterate through both smoothing level (alpha) and smoothing trend factor (beta) ranging from 0.01 to 1 in increments of 0.01, selecting the model with the lowest RMSE value as the best fit.
- ✓ The best fit model has alpha 0.02 and beta 0.5 and RMSE 1274.63
- ✓ As we see in the graph, the model could capture the trend (little tilted) in forecasted sales.

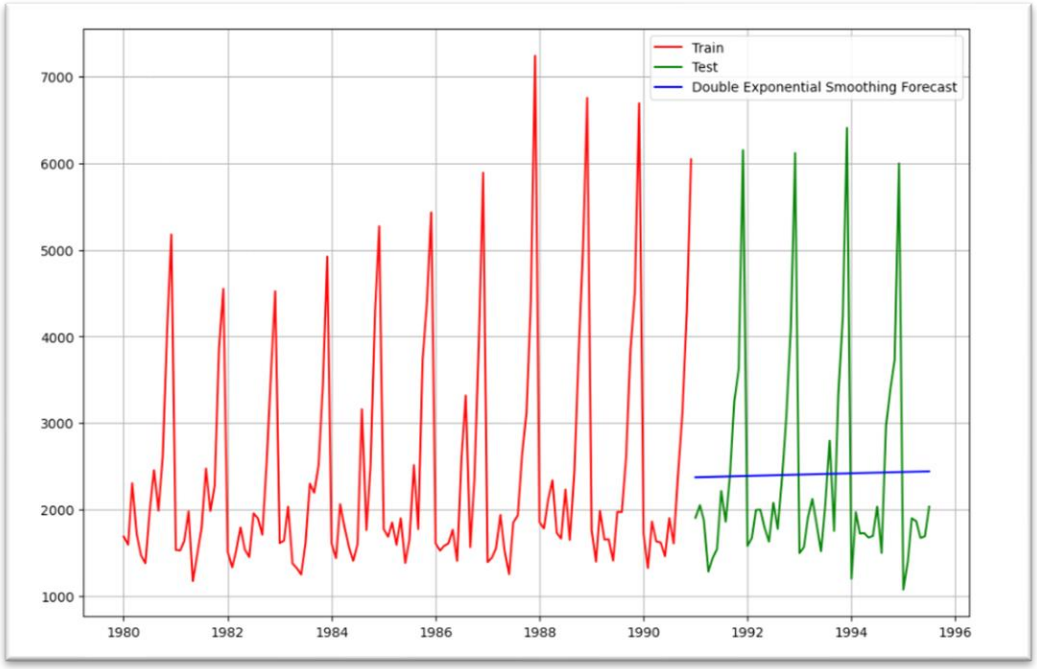

Fig 14. Sales Forecast from 1993 to 1996 using Double Exponential Smoothing

Triple Exponential Smoothing:

- ✓ This model exhibits both a **trend** and **seasonality**. It extends **Double Exponential Smoothing (DES)** by adding a third component to account for the seasonality in the data.
- ✓ We iterate through smoothing level (alpha), smoothing trend factor (beta) and smoothing Seasonal factor (gamma) ranging from 0.01 to 1 in increments of 0.1, selecting the model with the lowest RMSE value as the best fit.
- ✓ The best fit model has alpha 0.01 and beta 0.01, gamma 0.01 and RMSE 1329
- ✓ As we see in the graph, the model has captured trend and seasonality well in the test data.
- ✓ We see the forecasted line is very closely passing through the actual line which might indicate the overfitting.

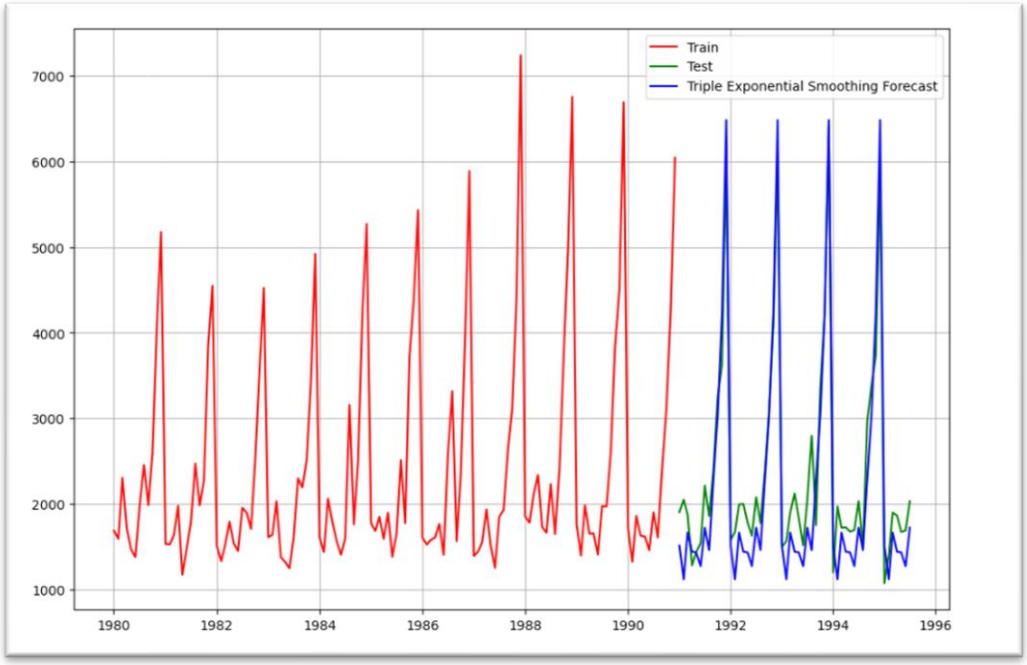

Fig 14. Sales Forecast from 1993 to 1996 using Triple Exponential Smoothing

Comparing the models:

Model Name	Test RMSE
Linear Regression	1359.7
Simple Average	1331
Moving Average	1026, 1290, 1375
Single Exponential Smoothing	1324.7
Double Exponential Smoothing	1274.6
Triple Exponential Smoothing	1286

Insights:

- ✓ **Moving Average (3-window)**: This model performs the best, indicating that a shorter window (3-period) Moving Average captures the underlying patterns in the data effectively.
- ✓ **Double Exponential Smoothing (DES)** performs better than other models like Linear Regression, Simple Average, and Single Exponential Smoothing, but it still doesn't beat the 3-window Moving Average.
- ✓ **Moving Average (6-window, 9-window)** The performance worsens as the window size increases, showing that a larger window smooths too much of the data and misses out on recent fluctuations, leading to higher errors.
- ✓ **Triple Exponential Smoothing** captures the fluctuations well, but lesser RMSE than DES indicates seasonality is not a strong factor in the data.

Model Building – Stationary Data

Check for Stationarity:

- ✓ Any data set is called stationary when the statistical parameters like mean, standard deviation and auto-correlation structure don't vary much over a period.
- ✓ Data needs to be stationary to train any time series model on that.
- ✓ We apply dicky fuller test to verify if the data is stationary or not. If p-value from the test Is less than 0.05 then we conclude that the data is stationary.
- ✓ Here we obtained p-value 0.6 which means that behavior of data changes as we move through different periods.
- ✓ We take 1 lag difference of the data set and perform the test again. The P- value obtained from this process is less than 0.05. Thus, we achieve stationary and build model on top of that.

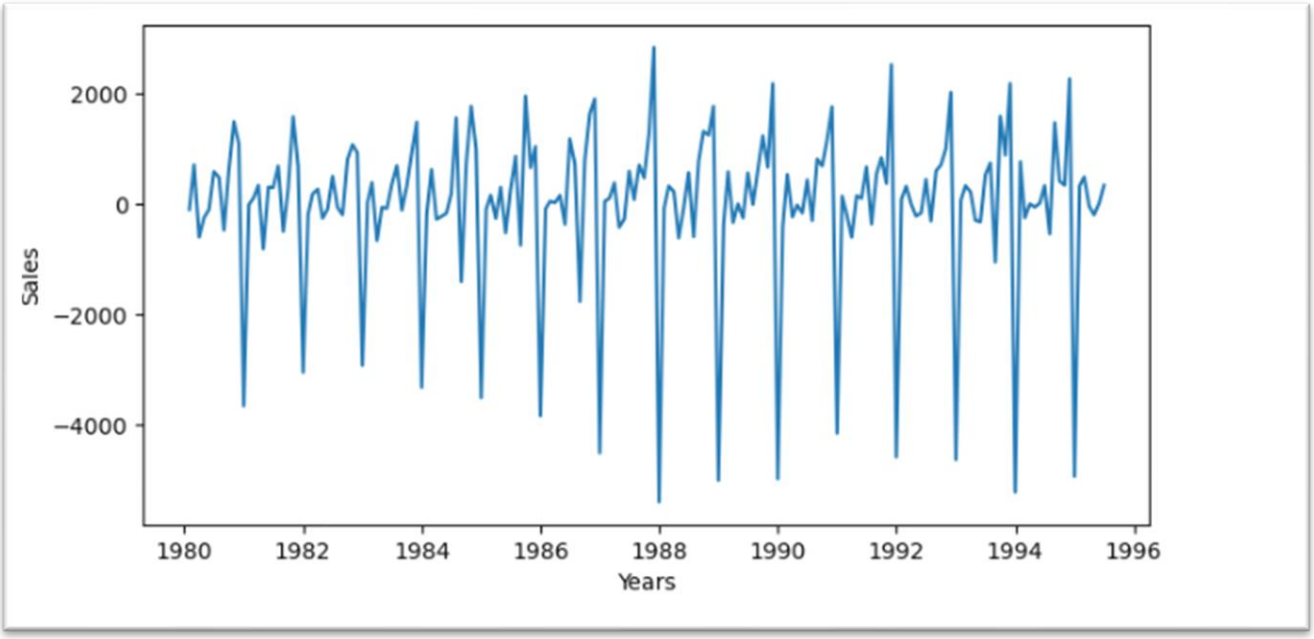

Fig 15. Trend of 1 lag Sales Difference of 20th century

ACF and PACF Plot

- ✓ ACF represents the autocorrelation of current data with its past values. By plotting it, we can analyze how present sales patterns are influenced by historical sales trends.
- ✓ PACF represents the partial autocorrelation of current data with its past values, isolating the direct relationship between present sales and specific past sales points. By plotting it, we can determine the direct influence of past sales on current sales.
- ✓ We can determine the autocorrelation coefficient (q) from the ACF plot, indicating how many past periods affect current sales, and the partial autocorrelation coefficient (p) from the PACF plot, isolating the influence of specific past points.

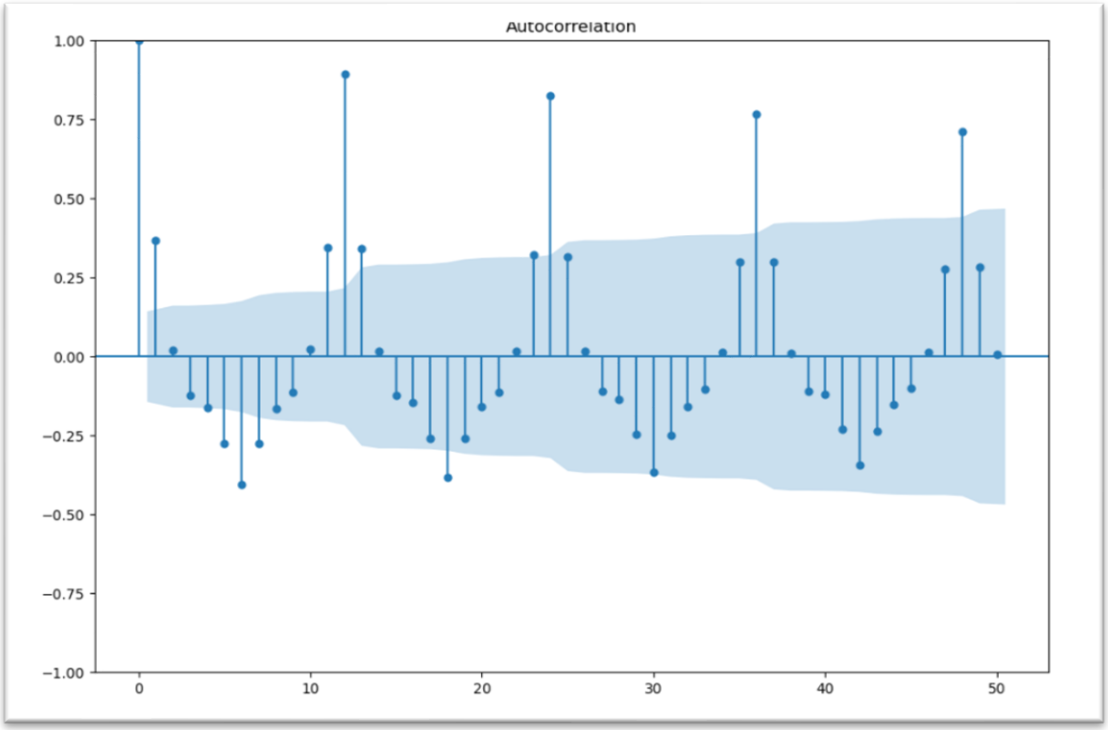

Fig 16. Autocorrelation of Sparkling Sales

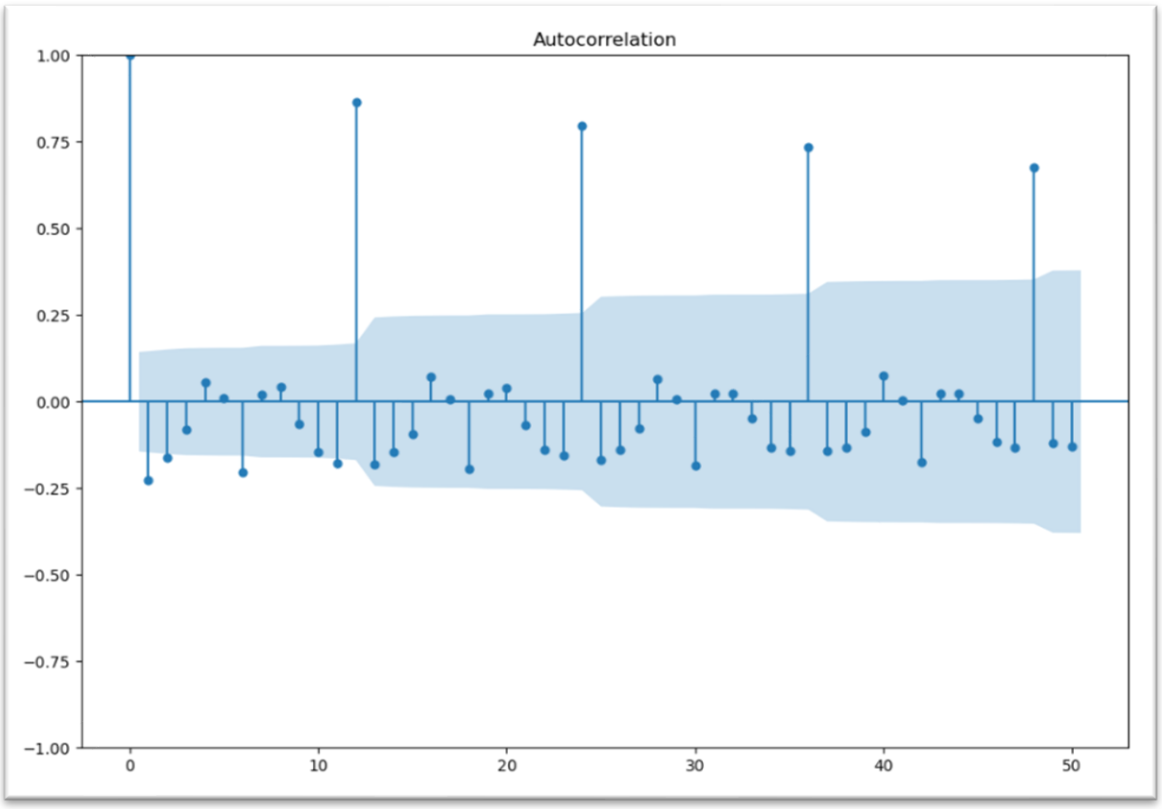

Fig 17. Autocorrelation of Sparkling Sales of 1 month difference

Observation:

- ✓ To ensure stationarity, we will consider the 1-month sales difference for the ACF plot.
- ✓ Based on the plot, we estimate a q-value of 2, as two data points exceed the confidence interval.

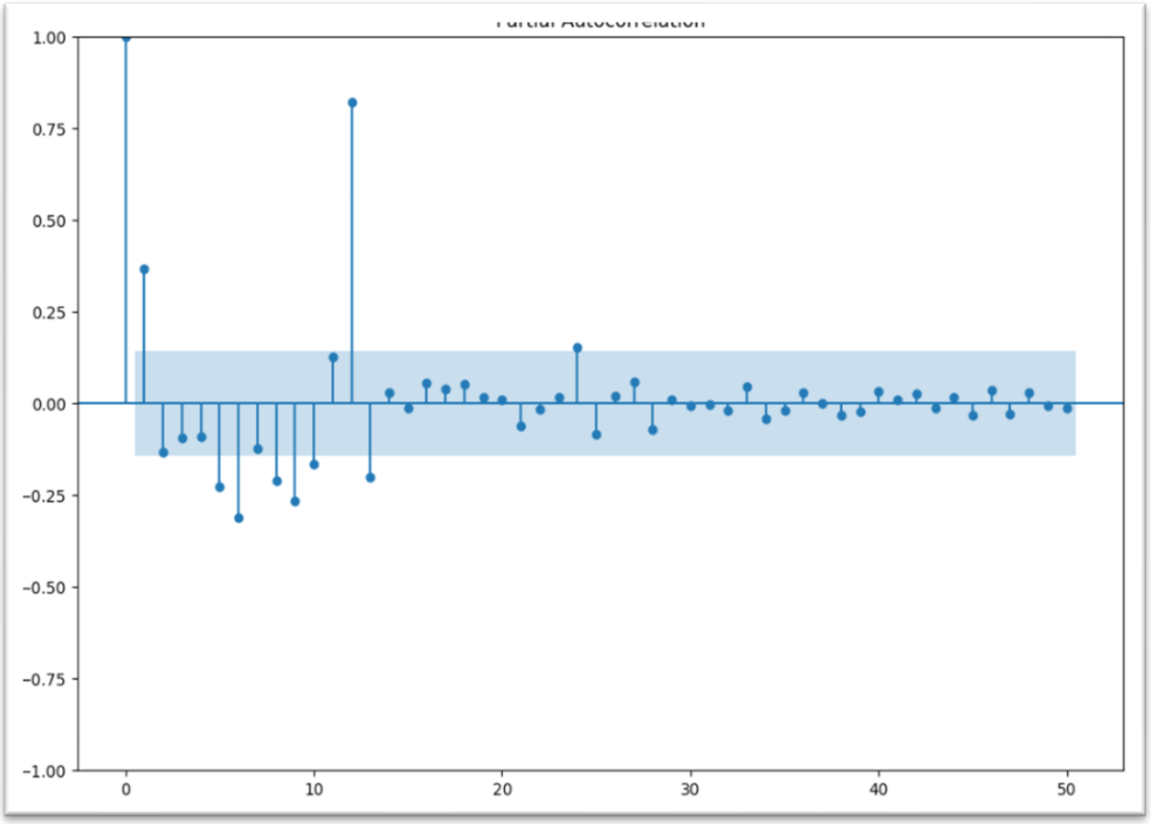

Fig 18. Partial Autocorrelation of Sparkling Sales

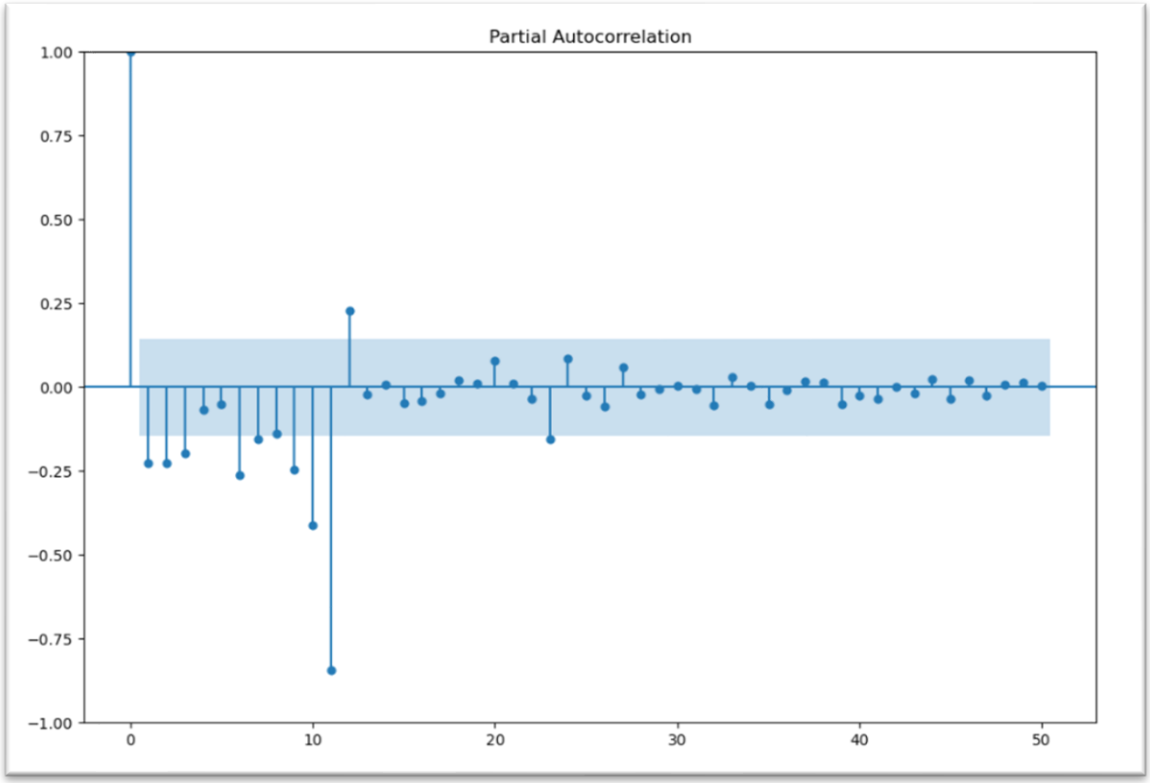

Fig 19. Partial Autocorrelation of Sparkling Sales of 1 month difference

Observations:

- ✓ To ensure stationarity, we will consider the 1-month sales difference for PACF plot also.
- ✓ Based on the plot, we estimate a p-value of 3, as three data points exceed the confidence interval.

ARIMA Model

Manual ARIMA Model:

- ✓ We split the data into train and test set with 80% train set size.
- ✓ For the ARIMA model we choose **p=3, q=2, d=1** from the above observations.
- ✓ We fit the model with train data and analyze the performance.
- ✓ From the model summary we observe that the past 3 months' sales and second last month's forecast error significantly contribute to the current forecast.
- ✓ Residual diagnostics indicate potential issues with autocorrelation (Ljung-Box) and heteroskedasticity, which could affect forecasting accuracy.
- ✓ The root mean square error on the test data is 1286.2
- ✓ From the plot below, we observe that while the forecast initially aligns with the actual data, it struggles to capture the variations over extended periods

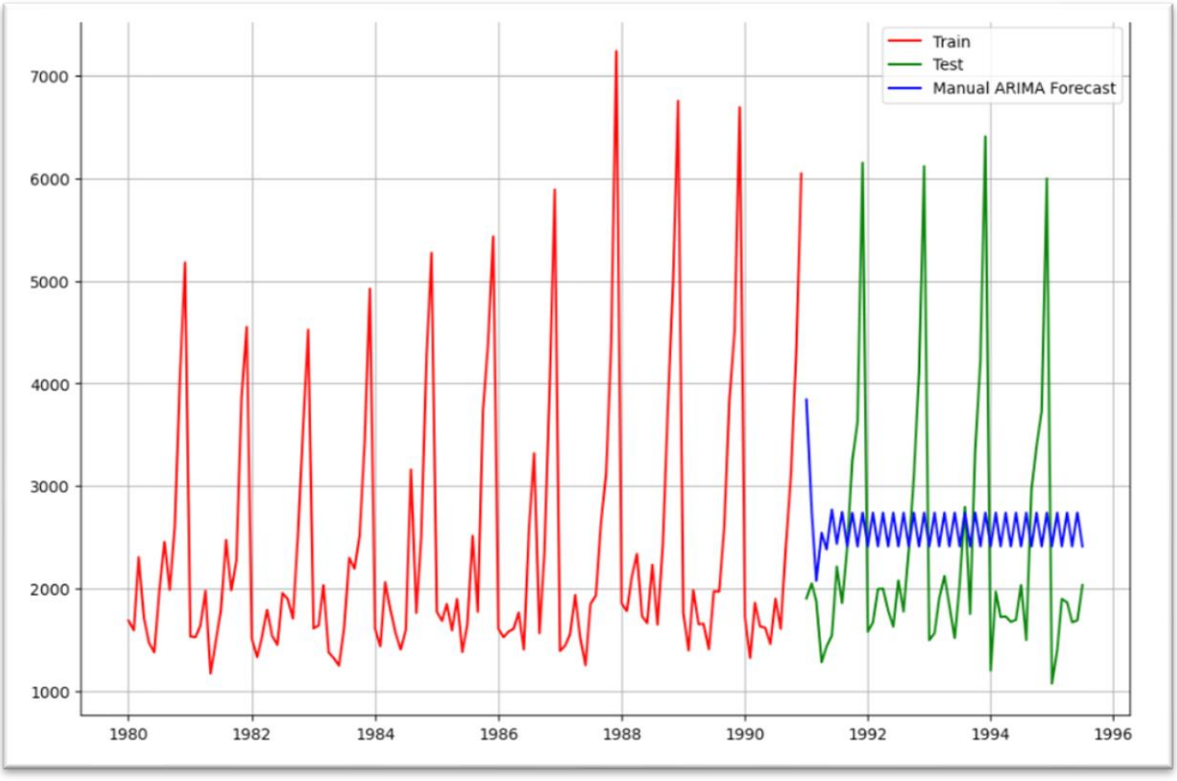

Fig 20. Manual ARIMA Forecasting

Auto ARIMA Model:

- ✓ We split the data into train and test set with 80% train set size.
- ✓ We evaluate the model across a range of p and q values while keeping d fixed at 1 and select the model with the lowest AIC value as the best fit.
- ✓ From the model summary we observe that the past 4 months' sales and forth last month's forecast error significantly contribute to the current forecast.
- ✓ Residual diagnostics indicate the residuals are well-behaved, with no significant autocorrelation and approximate normality.
- ✓ The root mean square error on the test data is 1204.8
- ✓ From the plot below, we observe that this time forecast is much more accurate than the previous ARIMA model, it can capture the variation to some extent.

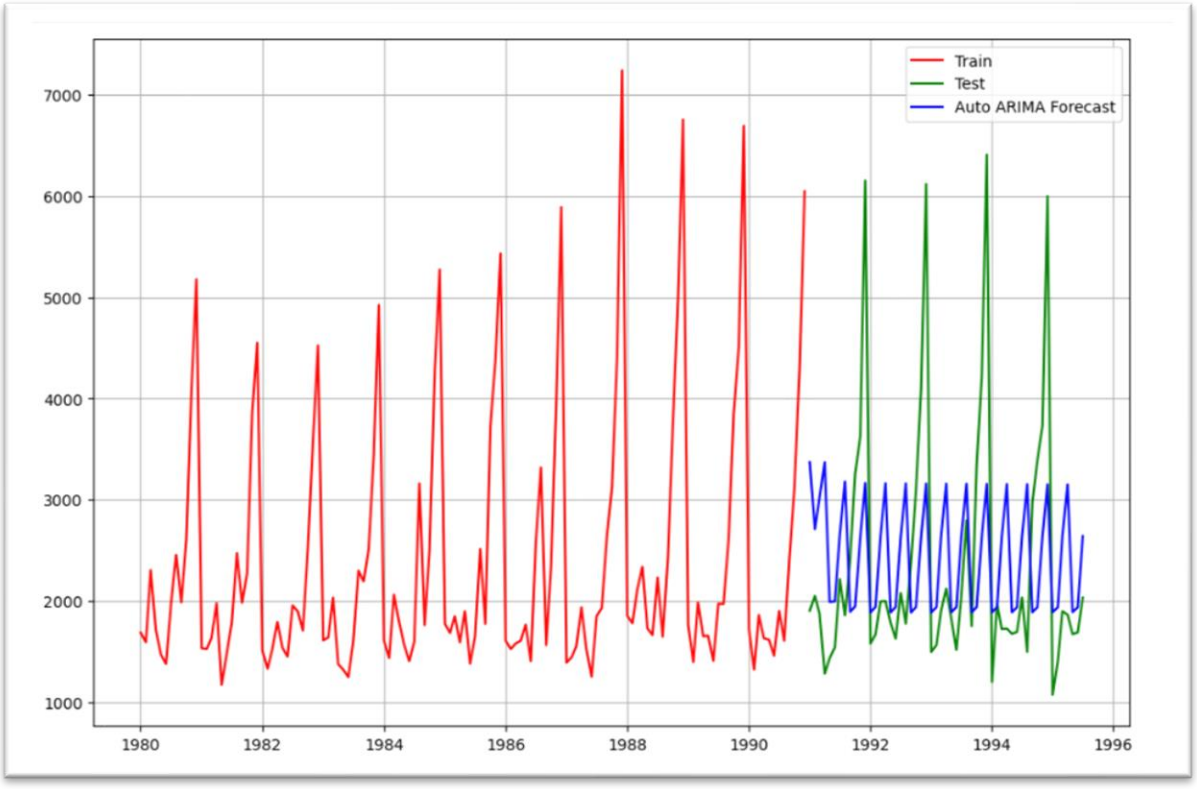

Fig 21. Auto ARIMA Forecasting

Manual SARIMA Model:

- ✓ We split the data into train and test set with 80% train set size.
- ✓ For the SARIMA model we choose **p=3, q=2, d=1** as analyzed before.
- ✓ We see from ACF and PACF plots that the pattern repeats itself every 1 year. So the seasonal period is 12 month.
- ✓ We perform the dicky fuller test to ensure the stationarity of the seasoned data. P-value obtained is less than 0.05 indicates the seasoned data is stationary.so D is 0.
- ✓ We plot ACF and PACF on seasoned and identify **P and Q value as 1** for both
- ✓ We fit the model with train data with above parameters and analyze the performance.
- ✓ From the model summary we observe that last month's sales, second last month's

forecast error, last season (12 months prior) sales and last season forecast error significantly contribute to the current forecast.

- ✓ Residual diagnostics indicate there is no autocorrelation (Ljung-Box) in the residuals but there are issues of heteroskedasticity, and it is not normally distributed.
- ✓ The root mean square error on the test data is 613.2 which is reduced significantly.
- ✓ From the plot below, we observe that forecast pretty much follows the actual trend.

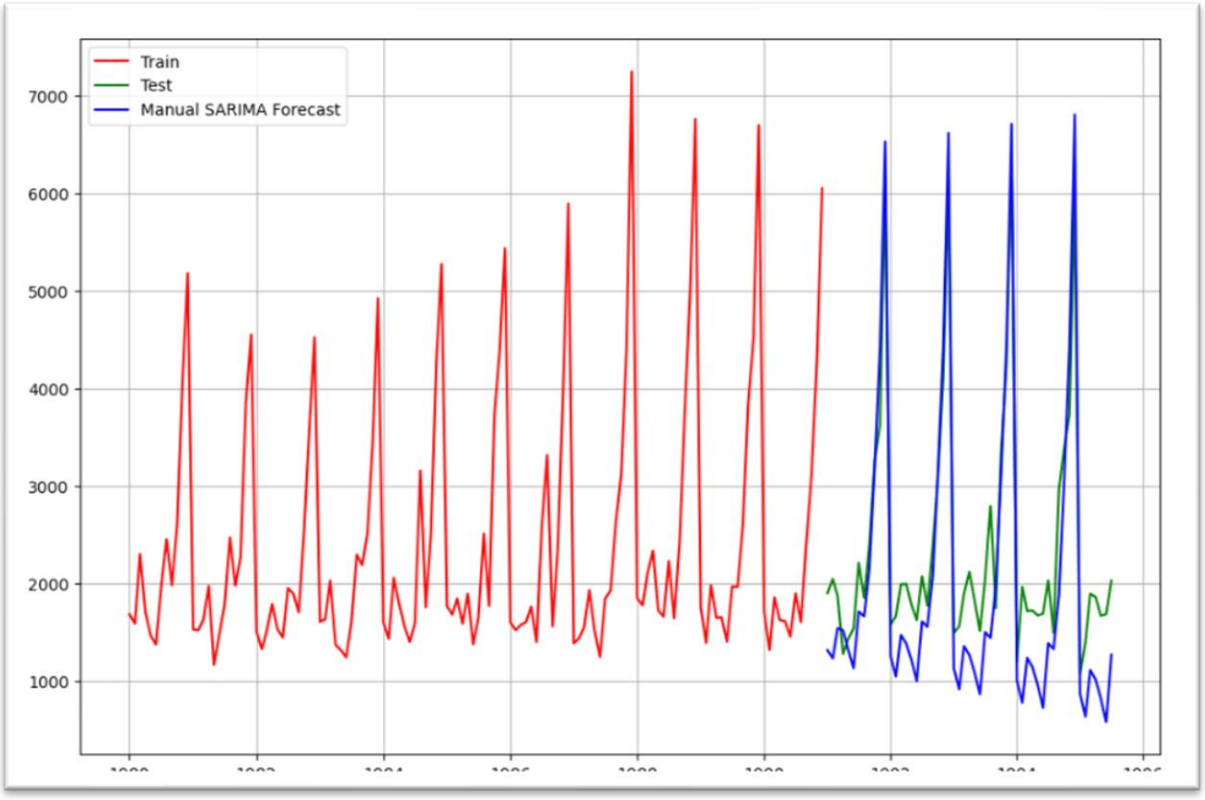

Fig 22. Manual SARIMA Forecasting

AUTO SARIMA Model:

- ✓ We split the data into train and test set with 80% train set size.
- ✓ We evaluate the model across a range of p and q values while keeping d fixed at 1 and a range of seasonal parameters P and Q while keeping D fixed at 0 with seasonal period of 12 month and select the model with the lowest AIC value as the best fit.
- ✓ From the model summary we observe that last month's sales, second last month's forecast error, last season sales (12 months prior) and last season forecast error significantly contribute to the current forecast.
- ✓ Residual diagnostics indicate the residuals are well-behaved, with no significant autocorrelation and heteroskedasticity but the residuals do not follow a normal distribution.
- ✓ The root mean square error on the test data is 528.6
- ✓ From the plot below, we observe that this time forecast is much more accurate than the previous SARIMA model.

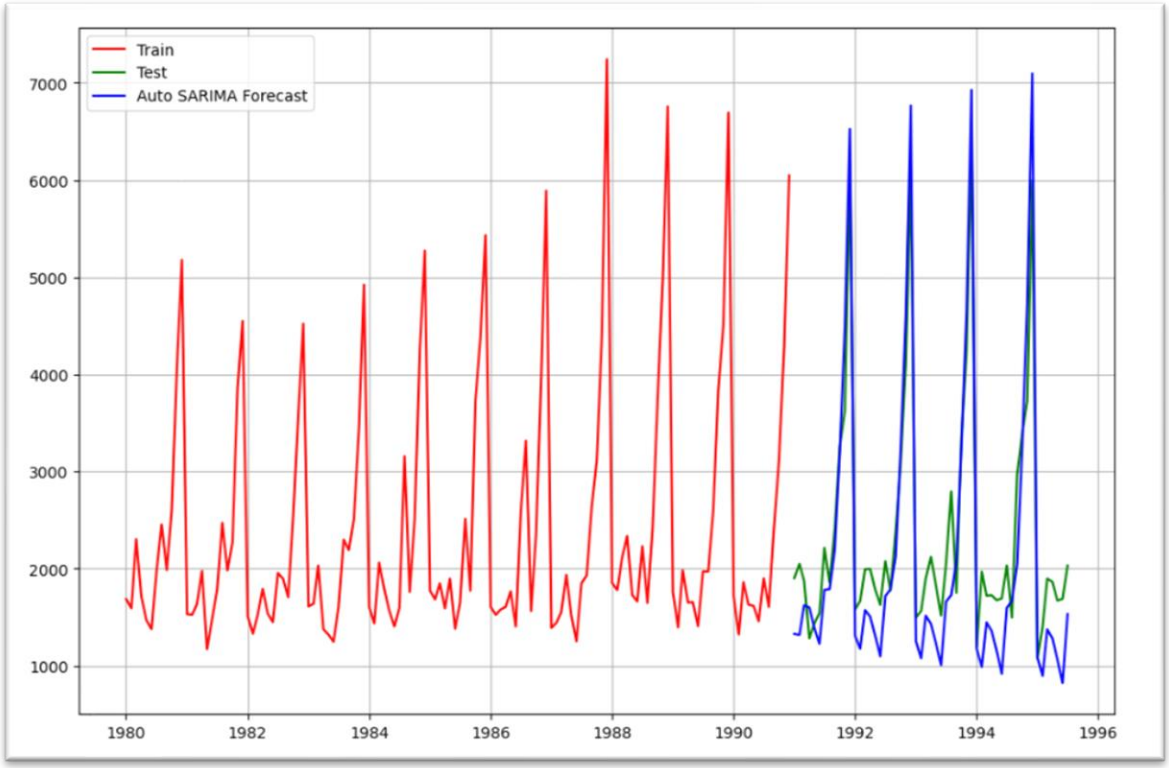

Fig 22. Auto SARIMA Forecasting

Check Performance of the Models:

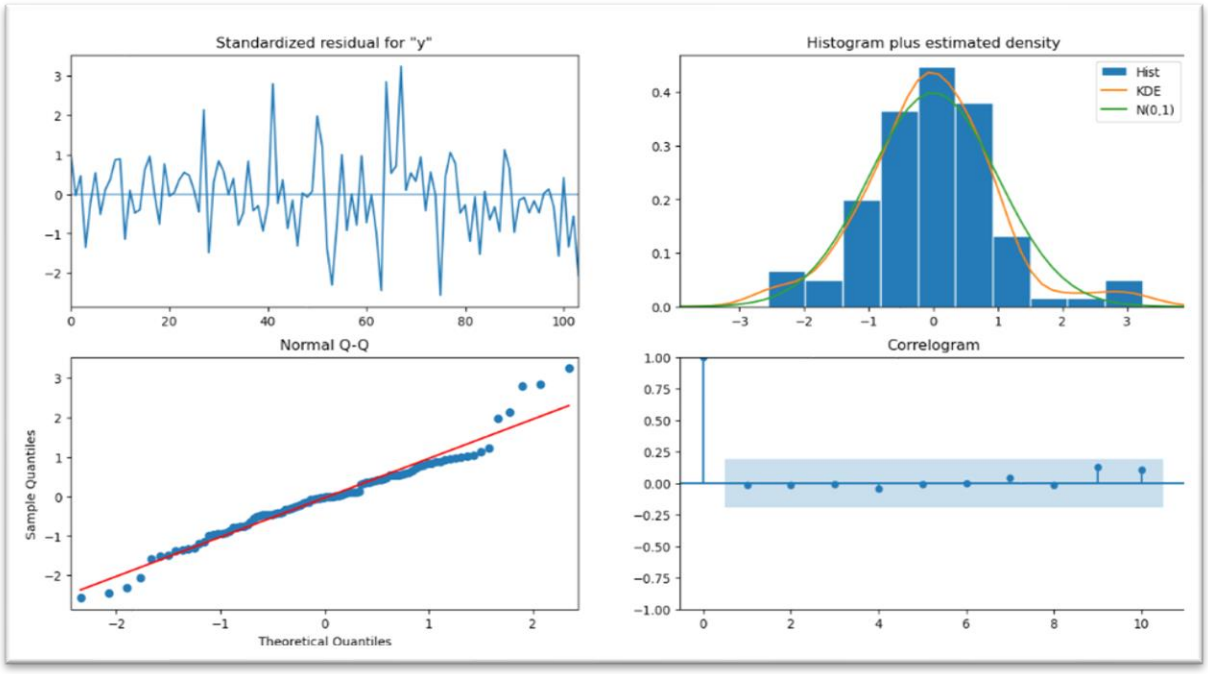

Fig 23. Residual Plots of Auto SARIMA Model

Observations:

- ✓ **Standardized Residuals**: Appear to be randomly scattered, suggesting no obvious patterns.

- ✓ **Histogram and KDE:** The residuals do not perfectly follow a normal distribution, as indicated by deviations from the normal curve.
- ✓ **Normal Q-Q Plot:** Shows some deviations from the 45-degree line, indicating non-normality in the residuals.
- ✓ **Correlogram:** No significant autocorrelation is observed, suggesting that the residuals are approximately white noise.

Comparing Model Performance:

Model	Test RMS	Autocorrelation in Residuals (Ljung-Box Test)	Normal Distribution of the Residuals (Jarque-Bera Test:)	Heteroskedasticity in the residual
Manual ARIMA (3,1,2)	**1286.2**	**Doesn't exist**	**Exists**	**Exists**
Auto ARIMA (4,1,4)	**1204.838**	**Doesn't exist**	**Exists**	**Exists**
Manual SARIMA (3,1,2) (1,0,1,12)	**613.152**	**Doesn't exist**	**Doesn't Exists**	**Exists**
Auto SARIMA (1,1,2)(1,0,2,12)	**528.584**	**Doesn't exist**	**Doesn't Exists**	**Doesn't Exists**

Fig 24. Model Comparison

Observation:

- The **Auto SARIMA (1,1,2)(1,0,2,12)** model is the **best choice based** on the following reasons:
  - ✓ **Lowest Test RMS**: This model has the lowest Test RMS value (528.584), indicating the best fit to the data among the four models.
  - ✓ **No Autocorrelation in Residuals**: The Ljung-Box test shows no significant autocorrelation in the residuals, suggesting that the model has effectively captured the time series structure.
  - ✓ **No Heteroskedasticity**: The absence of heteroskedasticity means that the variance of the residuals is constant over time, which is a desirable property for a well-fitted model.
- While residual is not fully following the normal distribution in the Auto SARIMA model, the significant reduction in Test RMS and the absence of autocorrelation and heteroskedasticity make it a strong candidate.
- Log transformation can be a possible solution to make the residual normally distributed and optimize the model performance further.

Building the model with full Data:

- ✓ We developed a SARIMA model with the parameters (1,1,2)(1,0,2,12) using a dataset of 187 records.
- ✓ The last month's sales, second last month's forecast error, last season's sales (12 months prior) and last season's forecast error significantly contribute to the current forecast.
- ✓ Residual diagnostic is as follows:

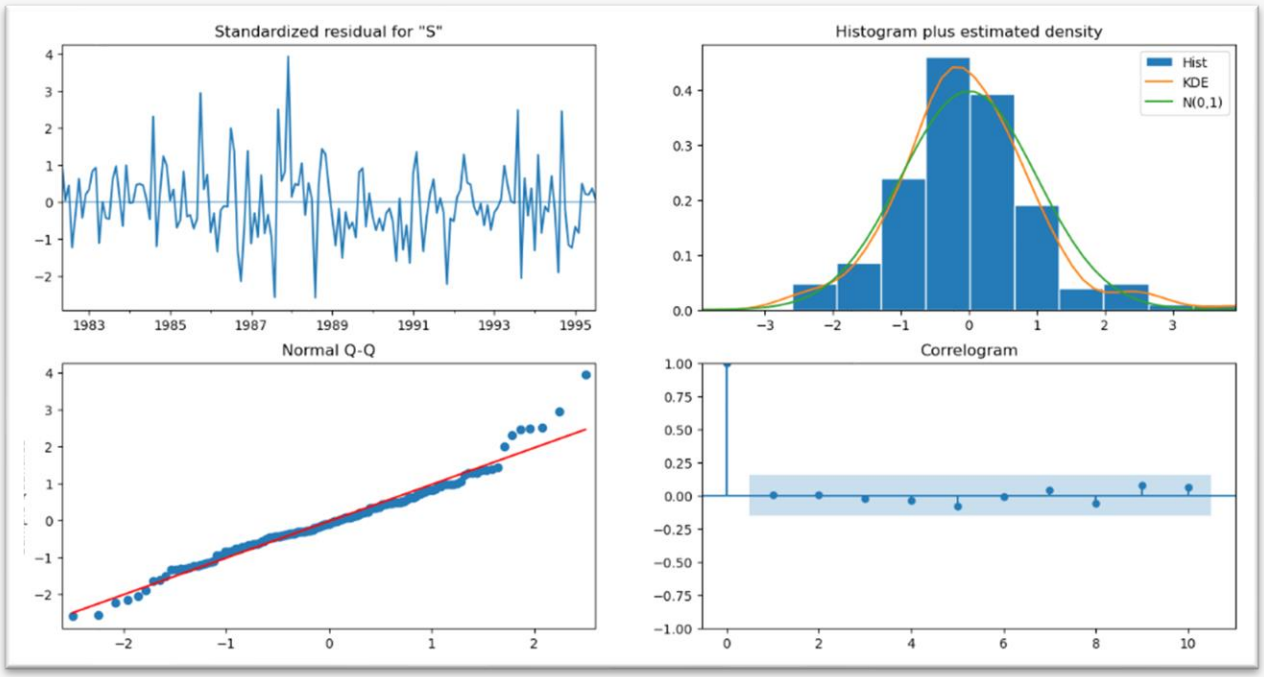

Fig 25. Residual Plots of Final SARIMA Model

Forecast for the next 12 months:

	Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01		1836.377934	379.702232	1092.175234	2580.580634
1995-09-01		2489.573090	384.466572	1736.032455	3243.113725
1995-10-01		3324.576401	384.572277	2570.828589	4078.324214
1995-11-01		4020.221876	386.330167	3263.028662	4777.415090
1995-12-01		6289.986888	386.384809	5532.686578	7047.287199
1996-01-01		1244.689638	387.294980	485.605425	2003.773851
1996-02-01		1533.155476	387.523494	773.623386	2292.687567
1996-03-01		1821.699301	388.150436	1060.938426	2582.460177
1996-04-01		1788.492563	388.490622	1027.064936	2549.920190
1996-05-01		1627.570071	389.009697	865.125075	2390.015067
1996-06-01		1563.325880	389.405708	800.104716	2326.547043
1996-07-01		2000.703118	389.880299	1236.551773	2764.854462

Fig 26. Sales Forecast of Next 12 months

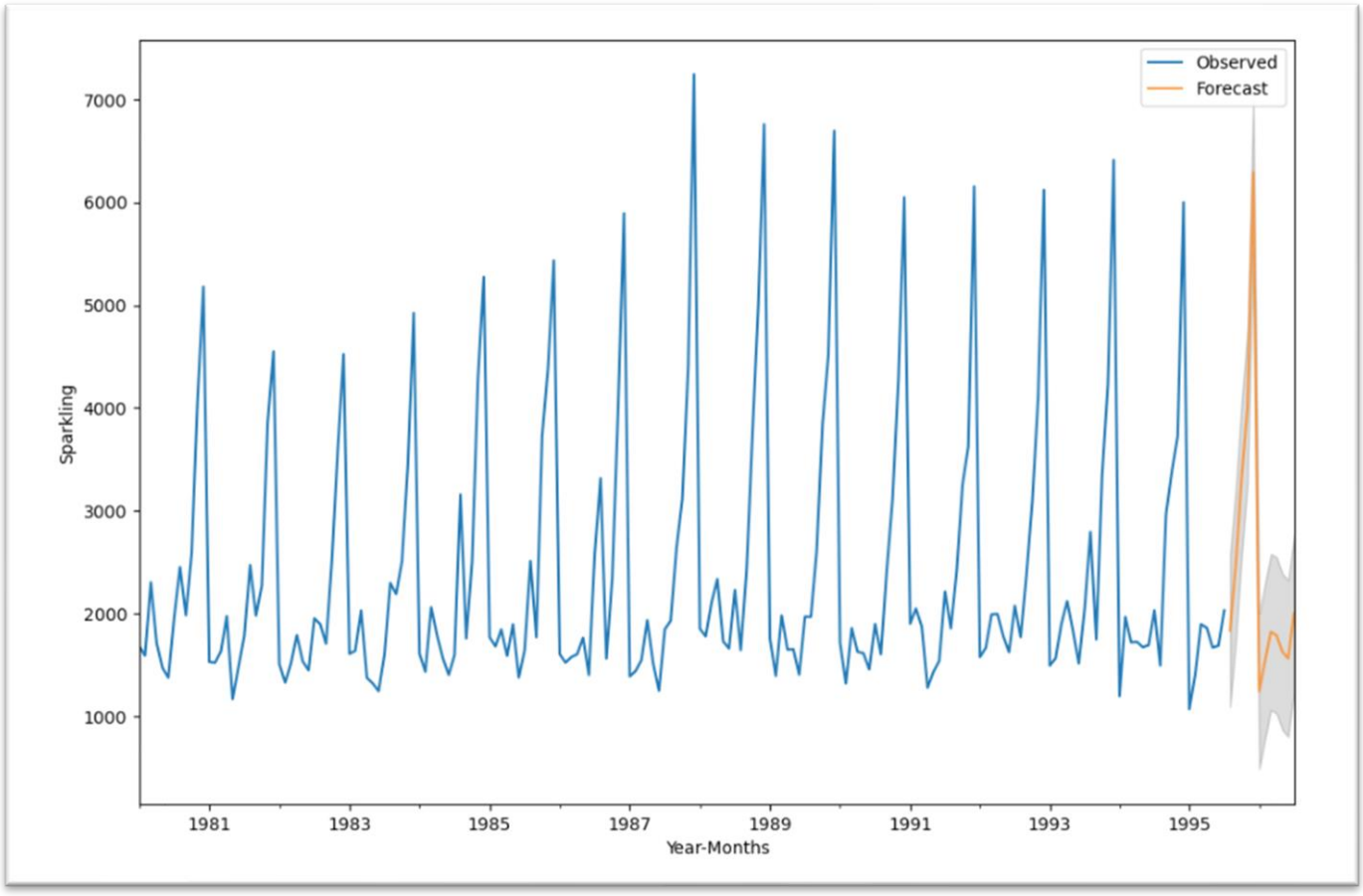

Fig 27. Forecast Trend from Aug,1995 to July 1996 with confidence interval

Key Insights:

- ✓ We developed several models starting with linear regression to different combination of auto-regression and moving average models integrating the differencing methodology along with seasonal variation to forecast the sparkling wine sales most accurately for the future.
- ✓ The key metrics used to achieve the optimal model for forecasting are Root mean square error (RMSE) computed on most recent dataset (test data). Lesser RMSE indicates better model performance.
- ✓ After fine-tuning the contributing parameters, the SARIMA model emerged as the best fit compared to the other models for predicting sparkling wine sales. This model effectively captures the seasonal patterns and trends in the data, providing more accurate and reliable forecasts for future sales.
- ✓ As we see in the above graph, the observed data has several sharp peaks and troughs, which are not fully captured by the forecasted values.
- ✓ The forecasted line tends to smooth out these extreme variations, providing a more stable prediction.
- ✓ Though SARIMA model provides a reasonable forecast, including more data points or external variables might help improve the accuracy of the forecasts.

Problem 2

Context

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties.

Objective

As an aspiring data scientist, the primary objective of this project is to analyze and forecast rose wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

Data Overview:

- **Shape**: There are 187 records and 2 columns.
- **Data Types**: Year-Month is time stamp of the data set hence it is object and Sales is int64
- **Independent & Target Variable**: We need to forecast the rose wine sales, so target variable is Rose
- **Check Duplicates**: There are no duplicate records in the database.
- **Check Missing Values:** There are 2 missing values in the dataset which we need to handle.
- **Date Time Index:** Year-Month column has been converted to index as it attributes to time-based sales.
- **Statistical Description:**

Metric	count	mean	std	min	25%	50%	75%	max
Sparkling	185	90.34	39.17	28	63	86	112	267

Observations:

- ✓ Mean and Median of the sales are close, hence there seem to be no extreme variations.
- ✓ Sales range from as low as 28 to max 267.

Exploratory Data Analysis:

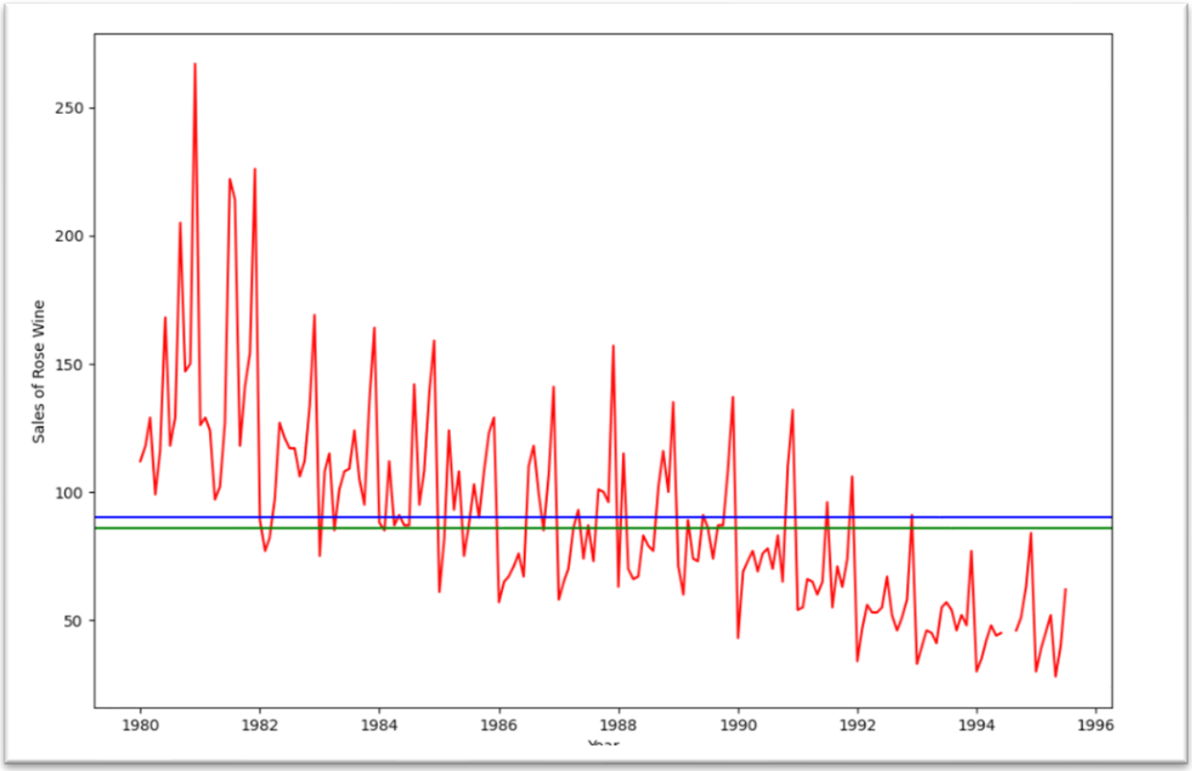

Fig 28. Wine Sales Trend of 20th Century

Observation:

- ✓ There is a downward trend in rose wine sales over the 20th century.
- ✓ There is a strong seasonal pattern observed every year.
- ✓ There is a strong variation in the sales around the avg of overall sales.

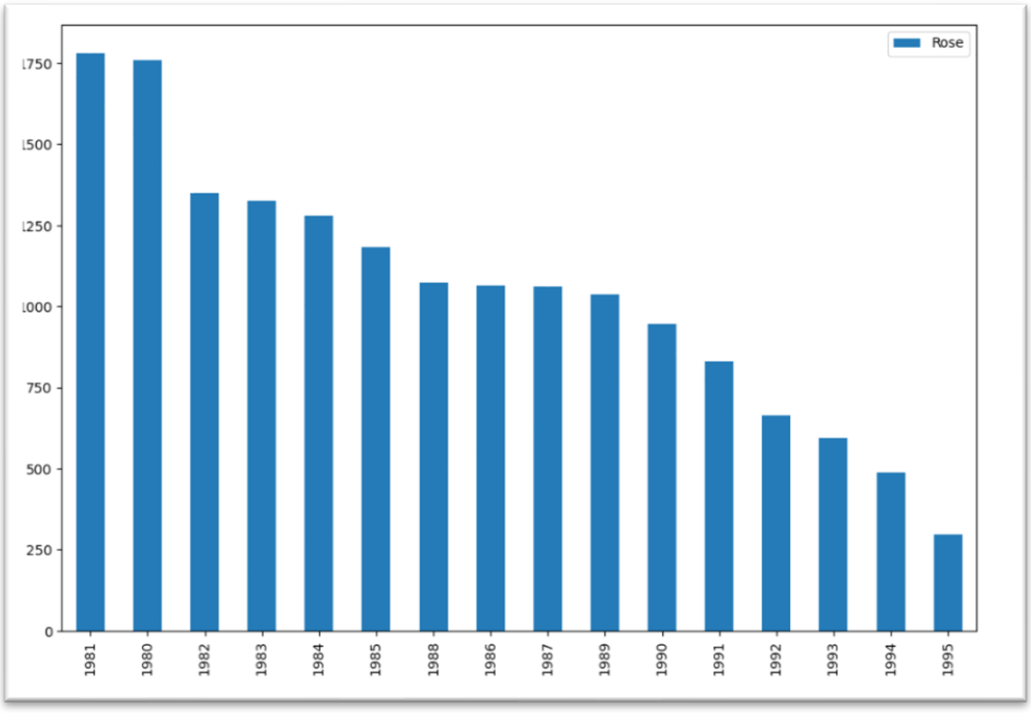

Fig 29. Year Wise Wine Sales of 20th Century

Observations:

- ✓ Maximum wine sales happened in the year of 1981
- ✓ We have data only till July,1995, minimum wine sales year can't be concluded.

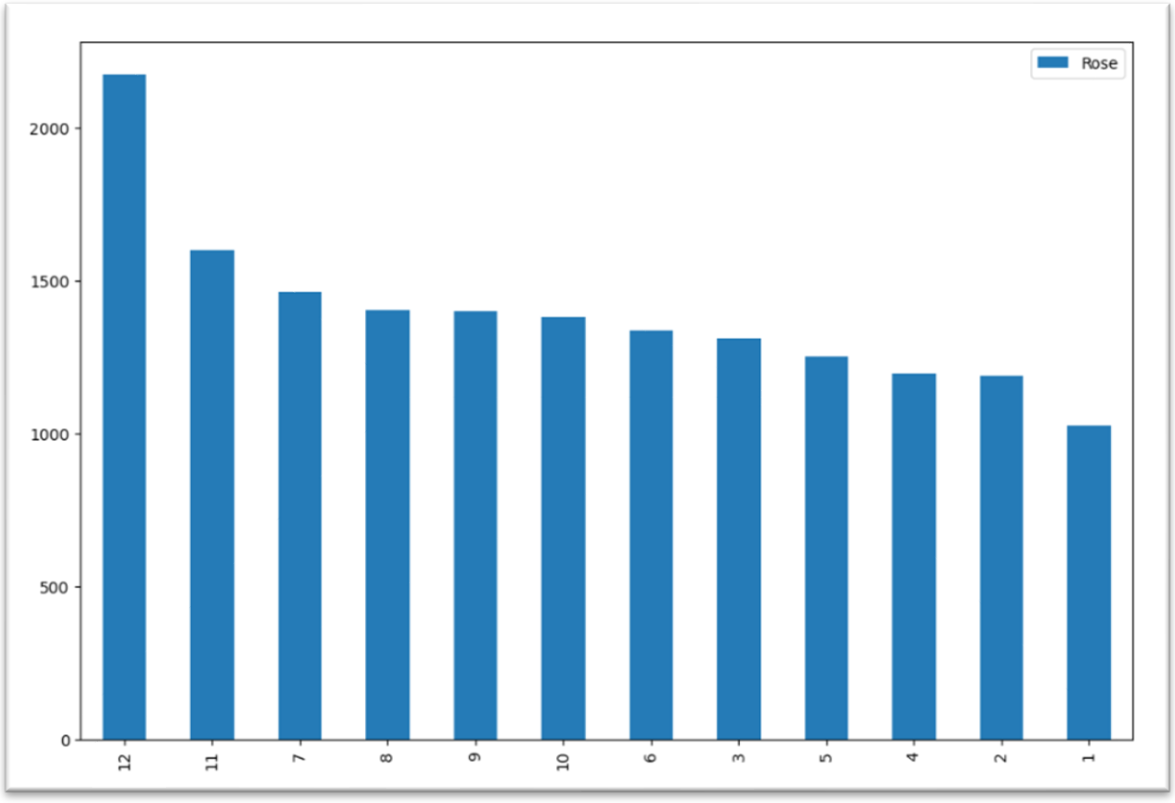

Fig 30. Month Wise Wine Sales of 20th Century

Observations:

- ✓ Wine sales typically see a spike during the holiday season, particularly in December. This trend is consistent across various years**.**
- ✓ Wine sales have reduced significantly in the other months compared to December.

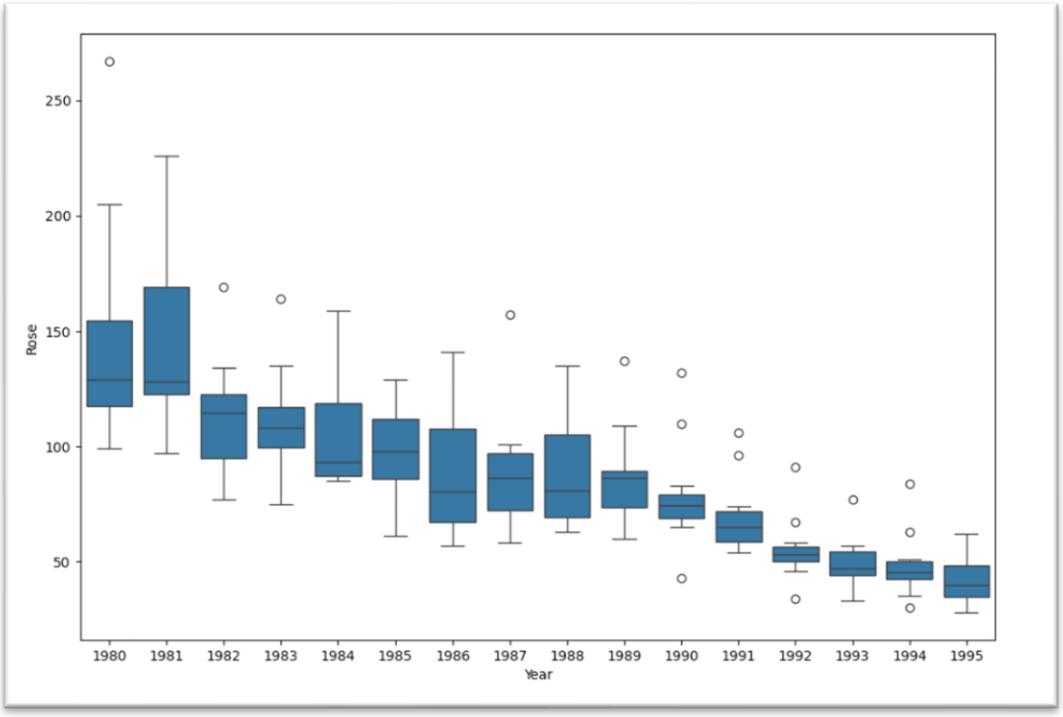

Fig 31. Yearly Sales Pattern in Box Plot

Observations:

- ✓ Minimum wine sales are highest in 1984 compared to other years.
- ✓ Very less sales happened in the year 1992

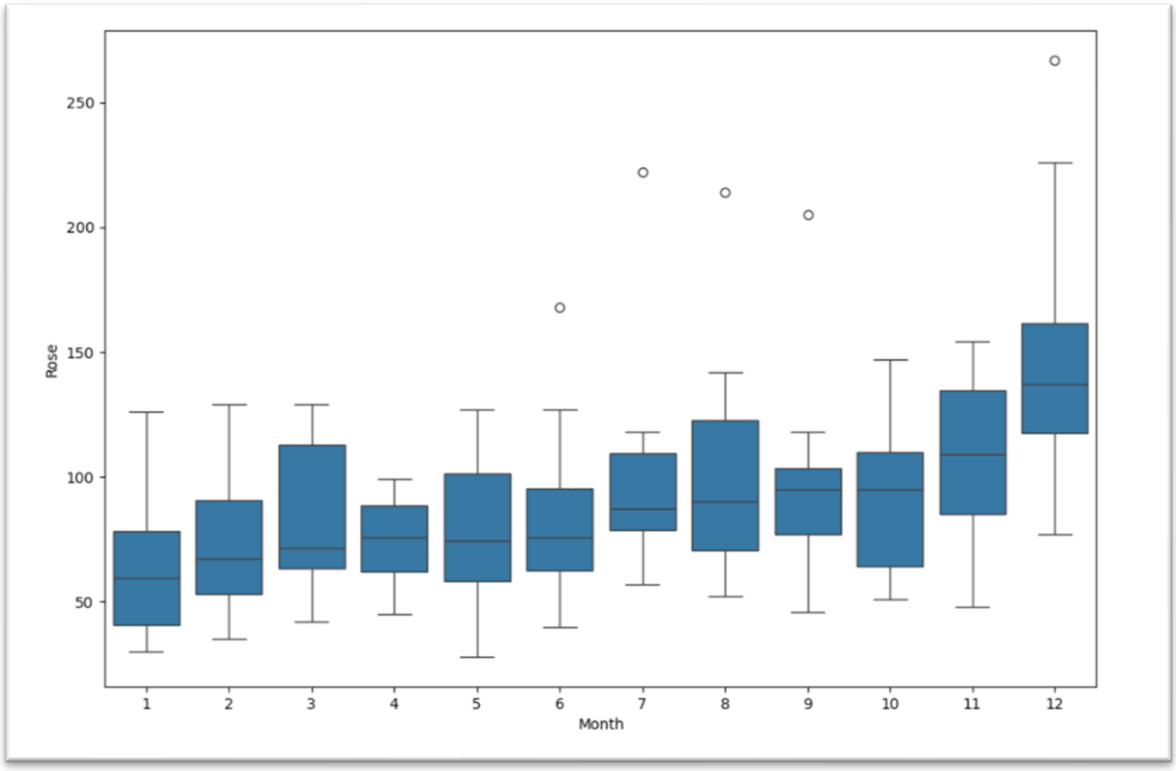

Fig 32. Monthly Sales Pattern in Box Plot

Observations:

- ✓ Sales performance peaks significantly in December, surpassing all other months.
- ✓ April and September experience the lowest volume of wine sales.

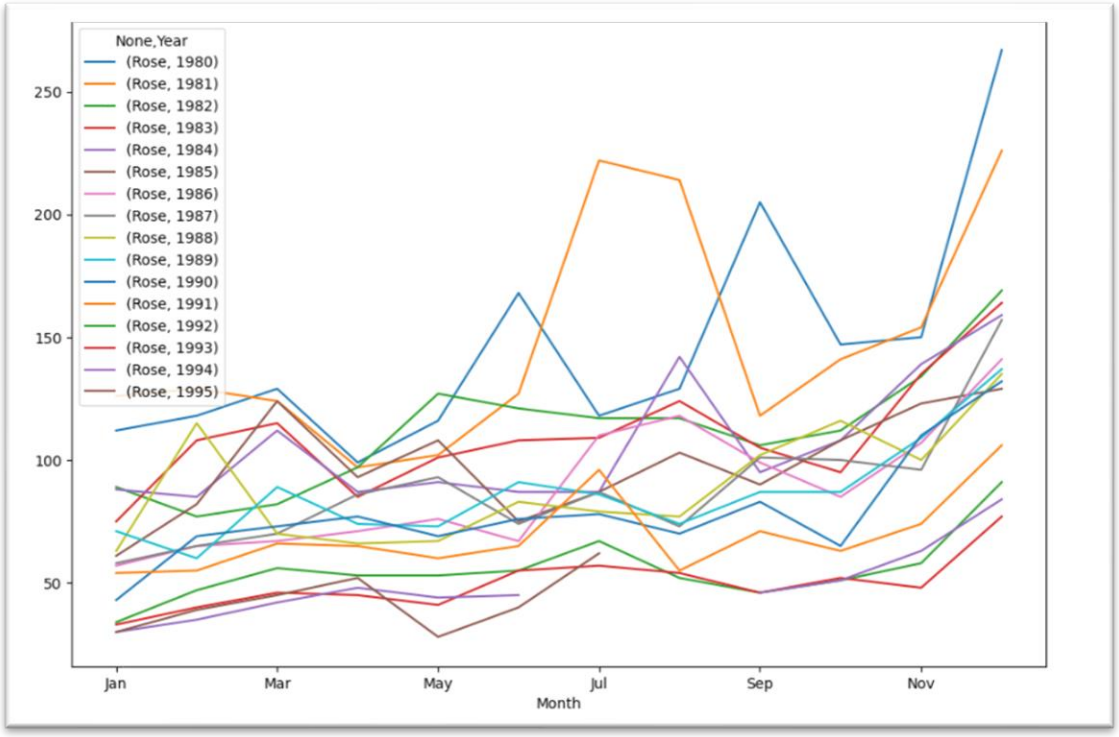

Fig 33. Monthly Sales Trend in 20th Century

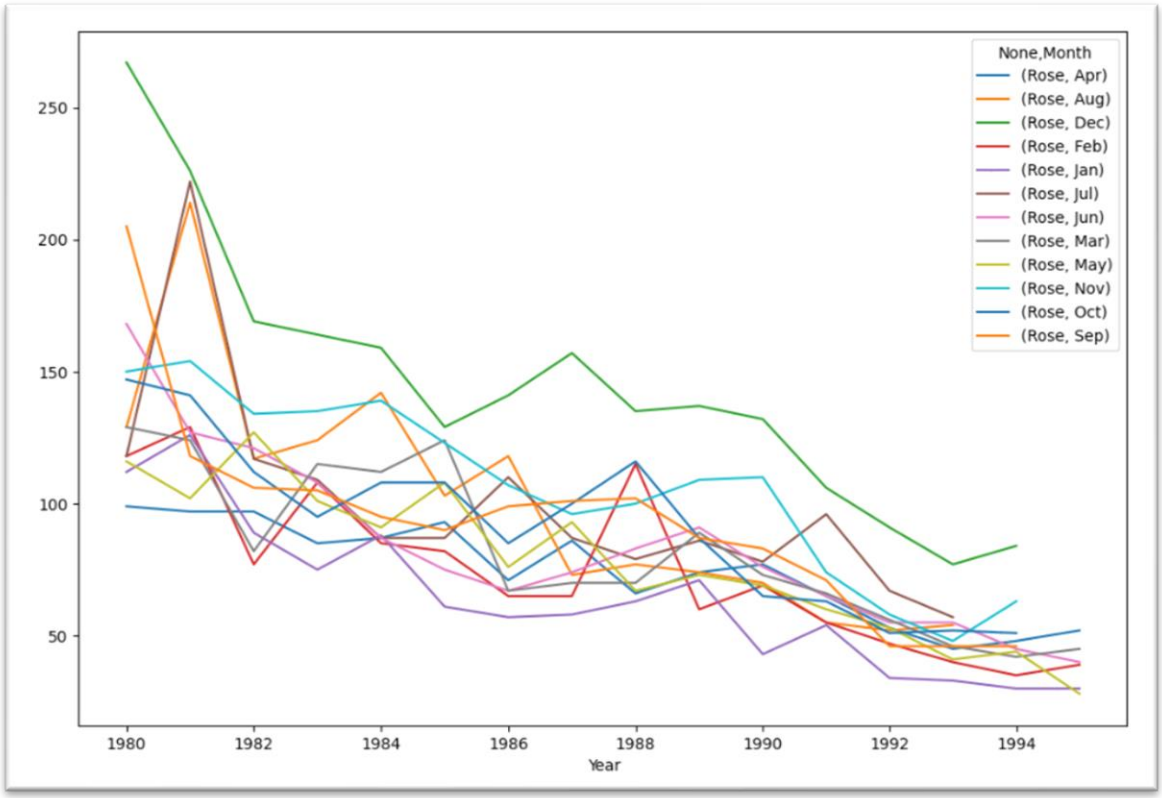

Fig 34. Yearly Sales Trend in 20th Century

Decomposition:

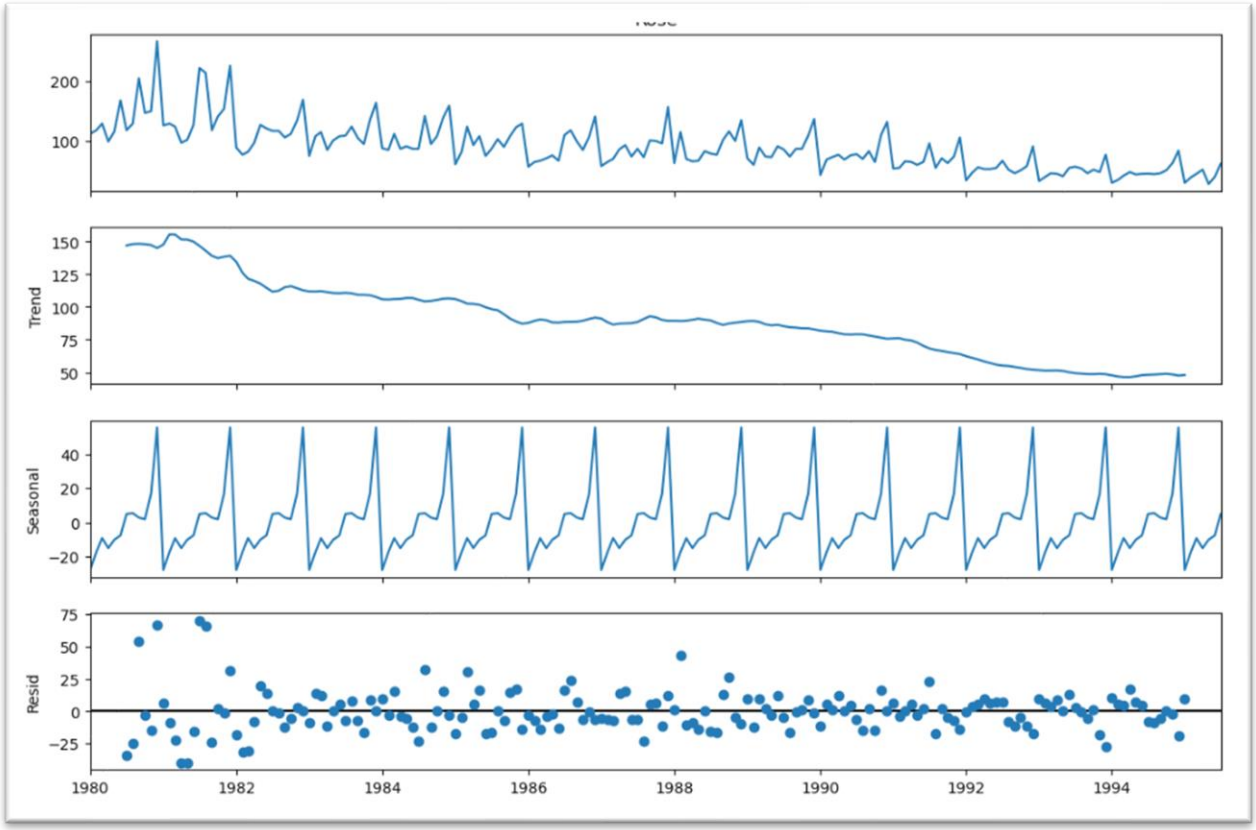

Fig 35. Additive Seasonal Decomposition

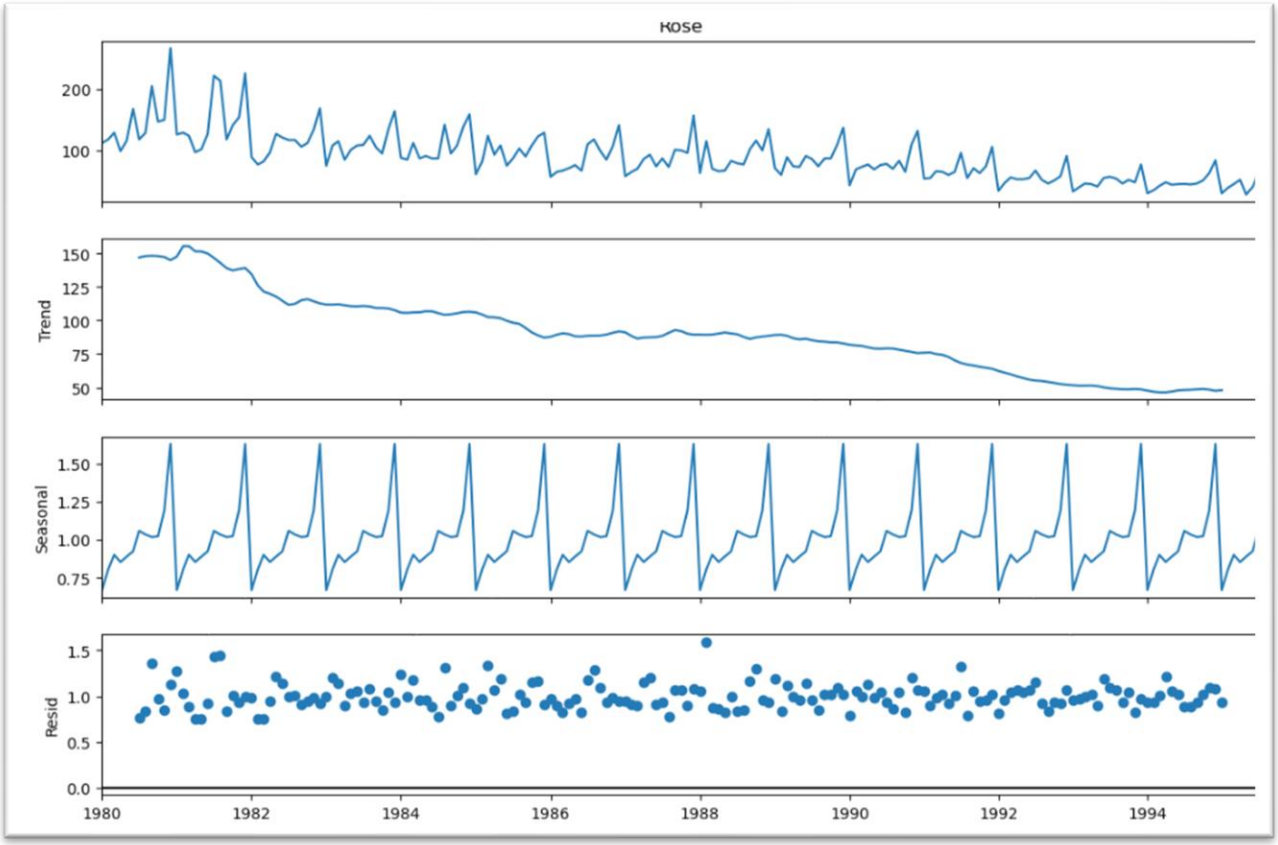

Fig 36. Multiplicative Seasonal Decomposition

Observations:

- ✓ In the decomposition process, the raw data gets split into the trend (the pattern observed over a long period of time), seasonality (pattern repeats in equal interval of time) and noise.
- ✓ As we see the residual pattern is similar in both the process additive and multiplicative. So, we proceed to analyze the data with additive process because of less complexity and computation time.

Data Pre-processing:

Missing Value Treatment:

There are 2 missing values at time stamp 1994-07-01 and 1994-08-01. We used spline interpolation to impute the missing values as it can capture the smallest variation in trend.

Train Test Split:

- ✓ We split the data into train and test data set with train size of 80% which means 80% data is being Used for model training and 20% for testing.
- ✓ There are 149 records in the train data set and 38 records in the test.

Model Building – Original Data:

Linear Regression Model:

- ✓ We used 80% of the data to train the linear regression model.
- ✓ As we see, forecasted sale using linear regression model is straight downward trend line which fails to predict the variation in the sales.
- ✓ The root mean squared error obtained from the model is 13.73

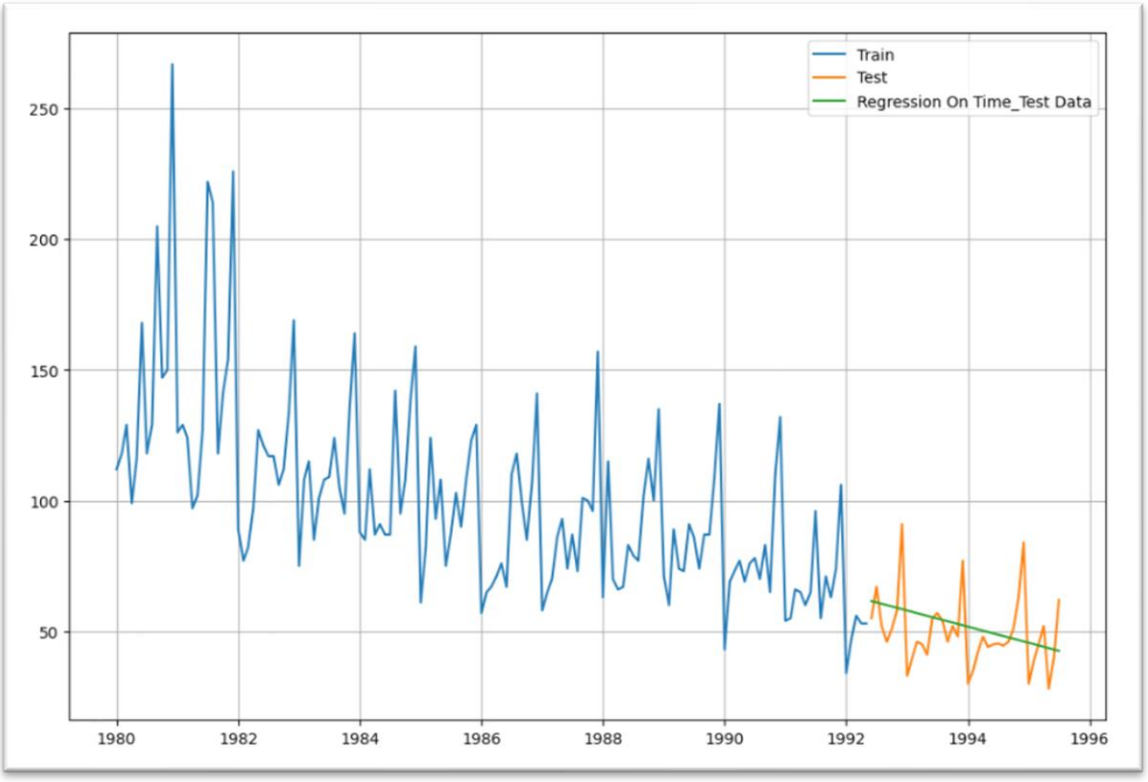

Fig 37. Sales Forecast from 1993 to 1996 using linear regression

Simple Average Model:

- ✓ This model forecast future sales as the average of all the sales made so far.
- ✓ So forecasted sales will be constant over time. It fails to capture the variation in sales in the future as it doesn't consider the same pattern in the train data.
- ✓ The root mean squared error obtained from the model is 52.24

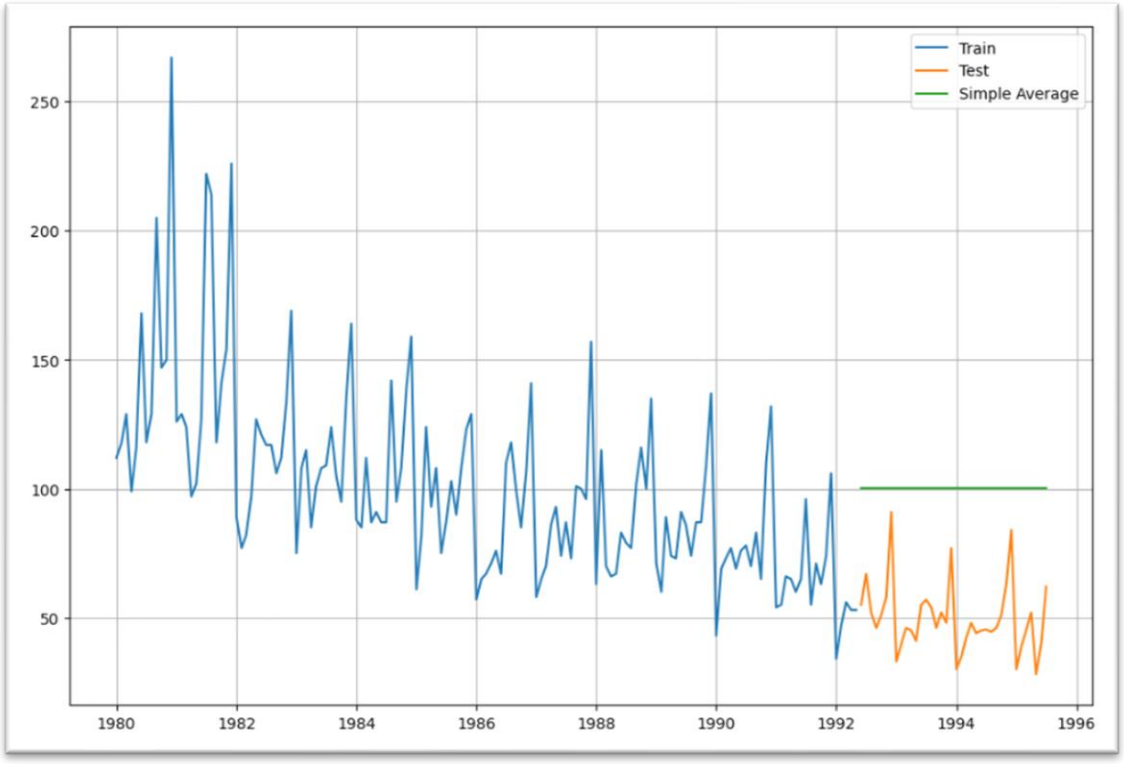

Fig 38. Sales Forecast from 1993 to 1996 using Simple Average

Moving Average Model:

- ✓ This model predicts sales at any given time point by calculating the moving average over the last specific time intervals (e.g., 3, 6, or 9 months).
- ✓ We have calculated a 3-month, 6- and 9-month moving average to forecast future sales.
- ✓ The root mean squared error obtained from 3-month model is 11.5, 6-month model is 12.4- and 9-month model is 12.6
- ✓ As we see in the below plot, the model captures the variation to some extent in the sales in the future,3-month moving average is the best fit out of 3 models.

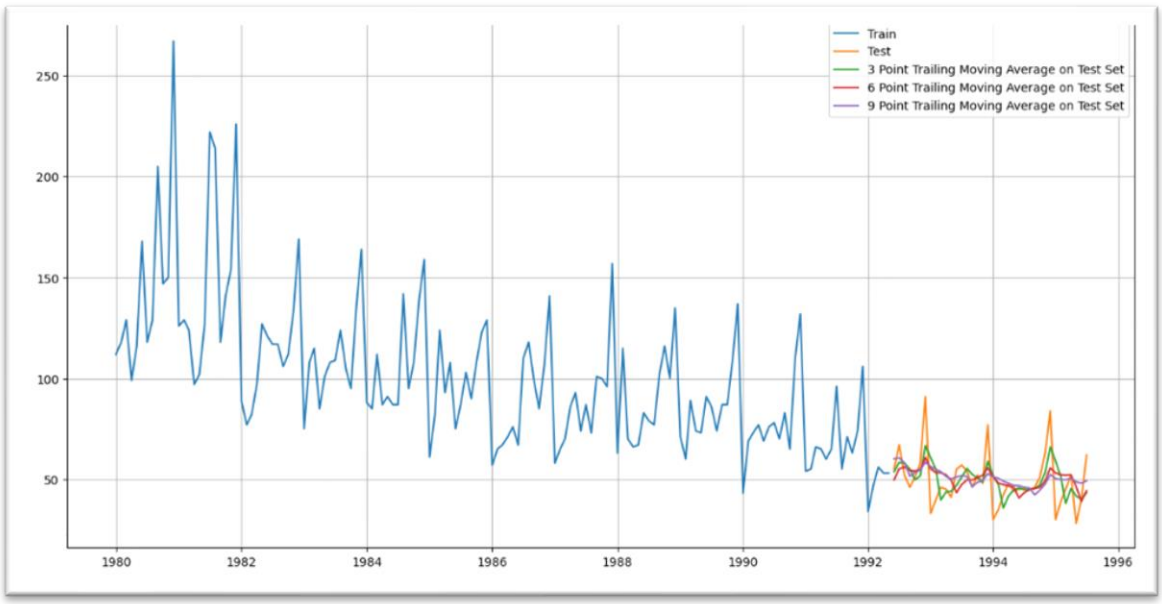

Fig 39. Sales Forecast from 1993 to 1996 using Moving Average

Simple Exponential Smoothing:

- ✓ The model does not consider trend or seasonality in the data set. It forecasts future values by giving **more weight to recent observations** while gradually (exponentially) decreasing the weight of older data points.
- ✓ We iterate through smoothing level factor ranging from 0.01 to 1 in increments of 0.01, selecting the model with the lowest RMSE value as the best fit.
- ✓ The best fit model has alpha (smoothing level factor) 0.07 and RMSE 36.45
- ✓ As we see in the graph, the model accounts for only level change.

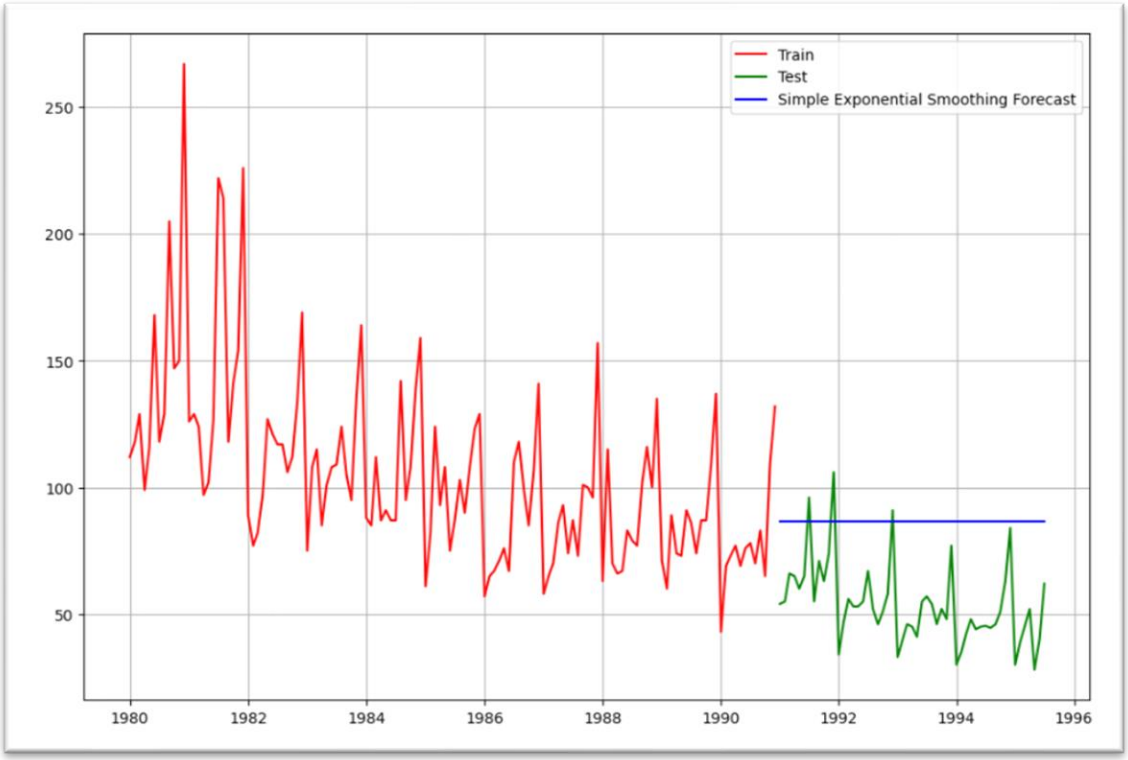

Fig 40. Sales Forecast from 1993 to 1996 using Simple Exponential Smoothing

Double Exponential Smoothing:

- ✓ Double Exponential Smoothing enhances simple exponential smoothing by incorporating a trend component, making it ideal for forecasting time series data with linear trends but no seasonality.
- ✓ We iterate through both smoothing level (alpha) and smoothing trend factor (beta) ranging from 0.01 to 1 in increments of 0.01, selecting the model with the lowest RMSE value as the best fit.
- ✓ The best fit model has alpha 0.04 and beta 0.47 and RMSE 14.56
- ✓ As we see in the graph, the model could capture the downward trend in forecasted sales.

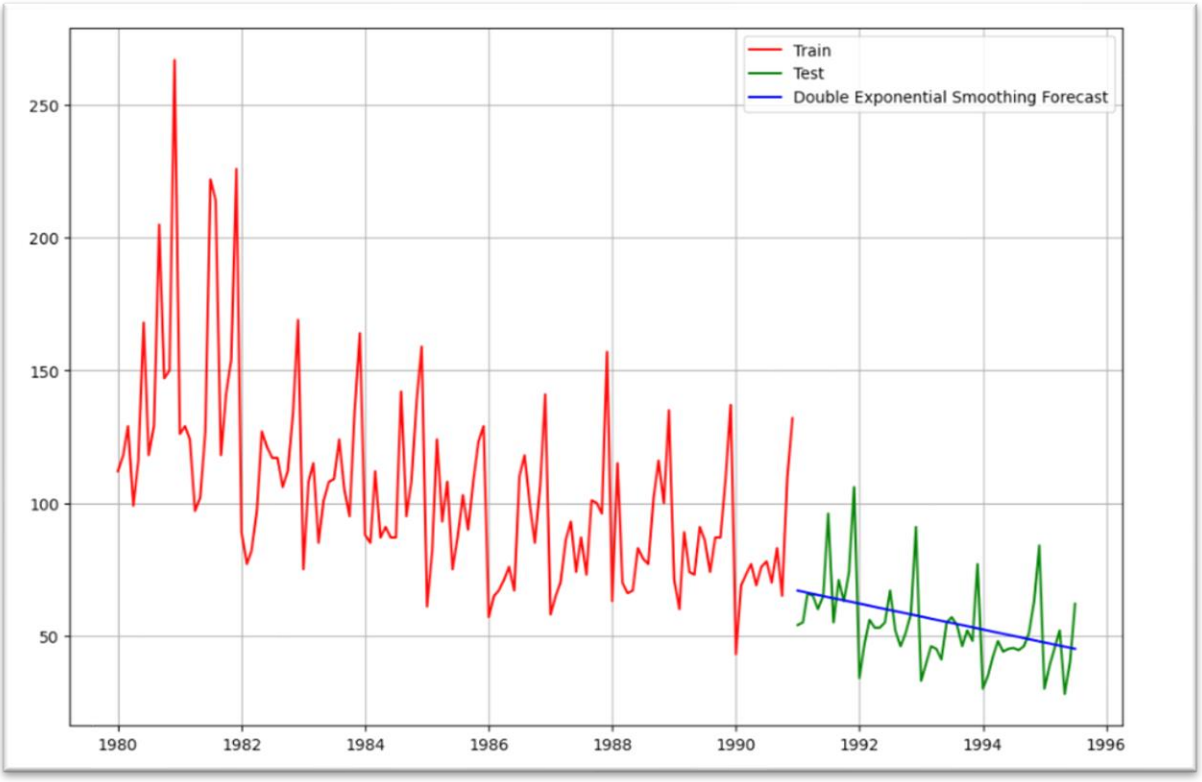

Fig 41. Sales Forecast from 1993 to 1996 using Double Exponential Smoothing

Triple Exponential Smoothing:

- ✓ This model exhibits both a **trend** and **seasonality**. It extends **Double Exponential Smoothing (DES)** by adding a third component to account for the seasonality in the data.
- ✓ We iterate through smoothing level (alpha), smoothing trend factor (beta) and smoothing Seasonal factor (gamma) ranging from 0.01 to 1 in increments of 0.1, selecting the model with the lowest RMSE value as the best fit.
- ✓ The best fit model has alpha 0.91 and beta 0.81, gamma 0.01 and RMSE 19.8
- ✓ As we see in the graph, the model has captured trend and seasonality well in the test data.
- ✓ We see the forecasted line is very closely passing through the actual line which might indicate the overfitting.

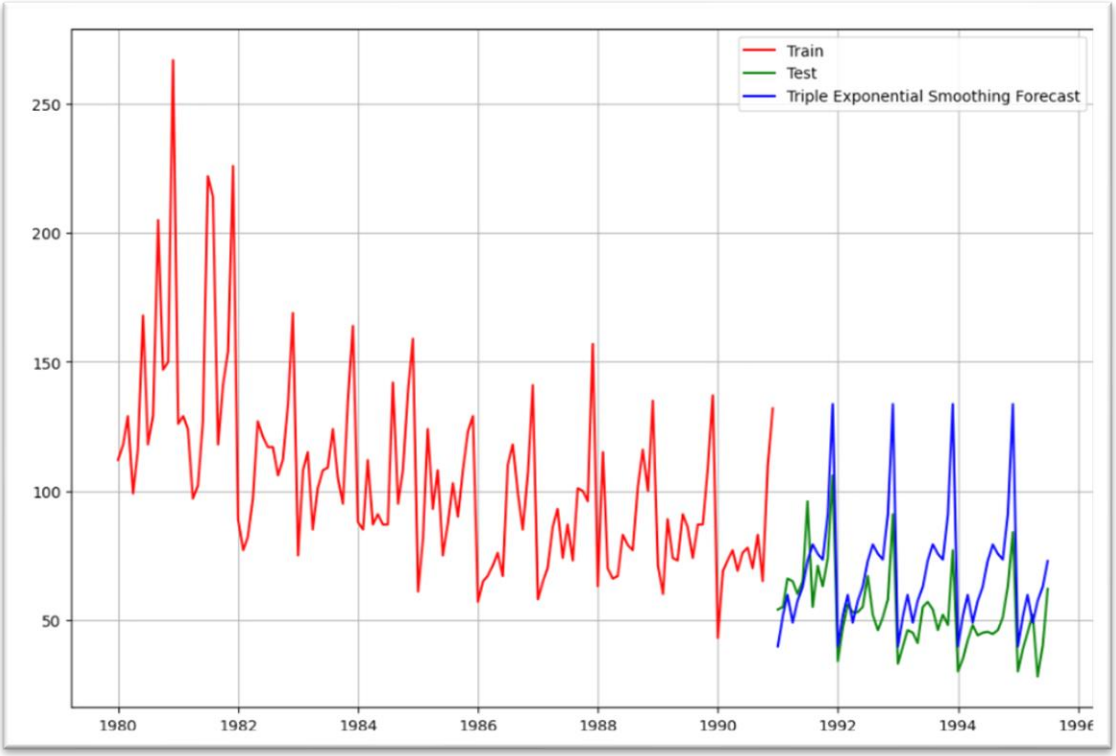

Fig 42. Sales Forecast from 1993 to 1996 using Triple Exponential Smoothing

Comparing the models:

Model Name	Test RMSE
Linear Regression	13.73
Simple Average	52.24
Moving Average	11.5,12.4,12.6
Single Exponential Smoothing	36.4
Double Exponential Smoothing	14.6
Triple Exponential Smoothing	19.8

Insights:

- ✓ **Moving Average (3-window)**: This model performs the best as RMSE is lowest for this, indicating that a shorter window (3-period) Moving Average captures the underlying patterns in the data effectively.
- ✓ **Double Exponential Smoothing (DES)** and **Linear Regression** performance is pretty much the same. The models can capture the trend quite good and not seasonality.
- ✓ **Moving Average (6-window, 9-window)** The performance worsens as the window size increases, showing that a larger window smooths too much of the data and misses out on recent fluctuations, leading to higher errors.
- ✓ **Triple Exponential Smoothing** captures the fluctuations well, but it looks like overfitting the model from the above graph.

Model Building – Stationary Data

Check for Stationarity:

- ✓ Any data set is called stationary when the statistical parameters like mean, standard deviation and auto-correlation structure don't vary much over a period.
- ✓ Data needs to be stationary to train any time series model on that.
- ✓ We apply dicky fuller test to verify if the data is stationary or not. If p-value from the test Is less than 0.05 then we conclude that the data is stationary.
- ✓ Here we obtained p-value 0.34 which means that behavior of data changes as we move through different periods.
- ✓ We take 1 lag difference of the data set and perform the test again. The P- value obtained from this process is less than 0.05. Thus, we achieve stationary and build model on top of that.

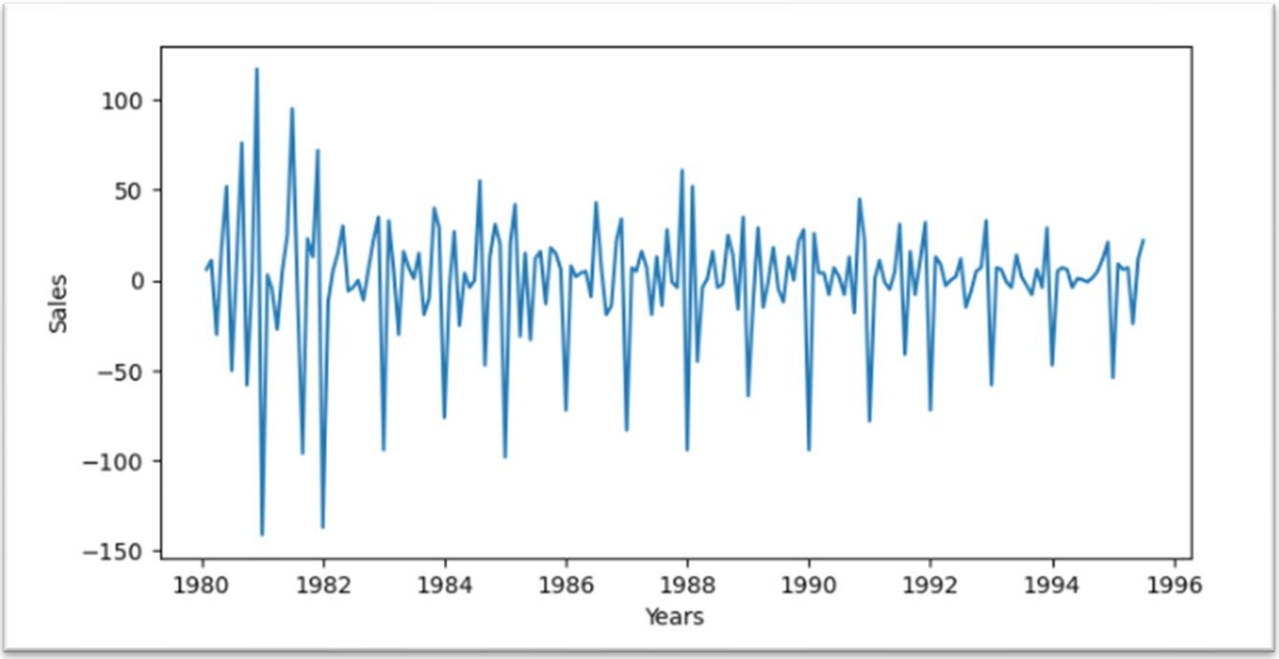

Fig 43. Trend of 1 lag Sales Difference of 20th century

ACF and PACF Plot

- ✓ ACF represents the autocorrelation of current data with its past values. By plotting it, we can analyze how present sales patterns are influenced by historical sales trends.
- ✓ PACF represents the partial autocorrelation of current data with its past values, isolating the direct relationship between present sales and specific past sales points. By plotting it, we can determine the direct influence of past sales on current sales.
- ✓ We can determine the autocorrelation coefficient (q) from the ACF plot, indicating how many past periods affect current sales, and the partial autocorrelation coefficient (p) from the PACF plot, isolating the influence of specific past points.

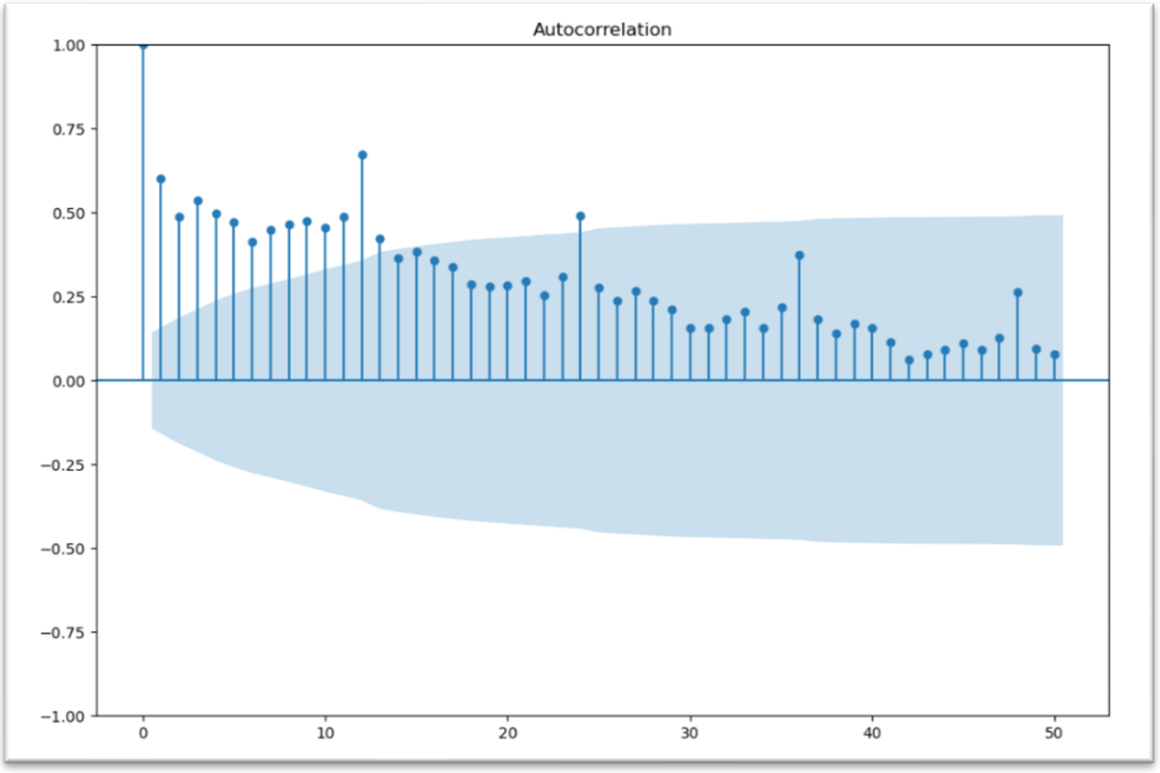

Fig 44. Autocorrelation of Rose wine Sales

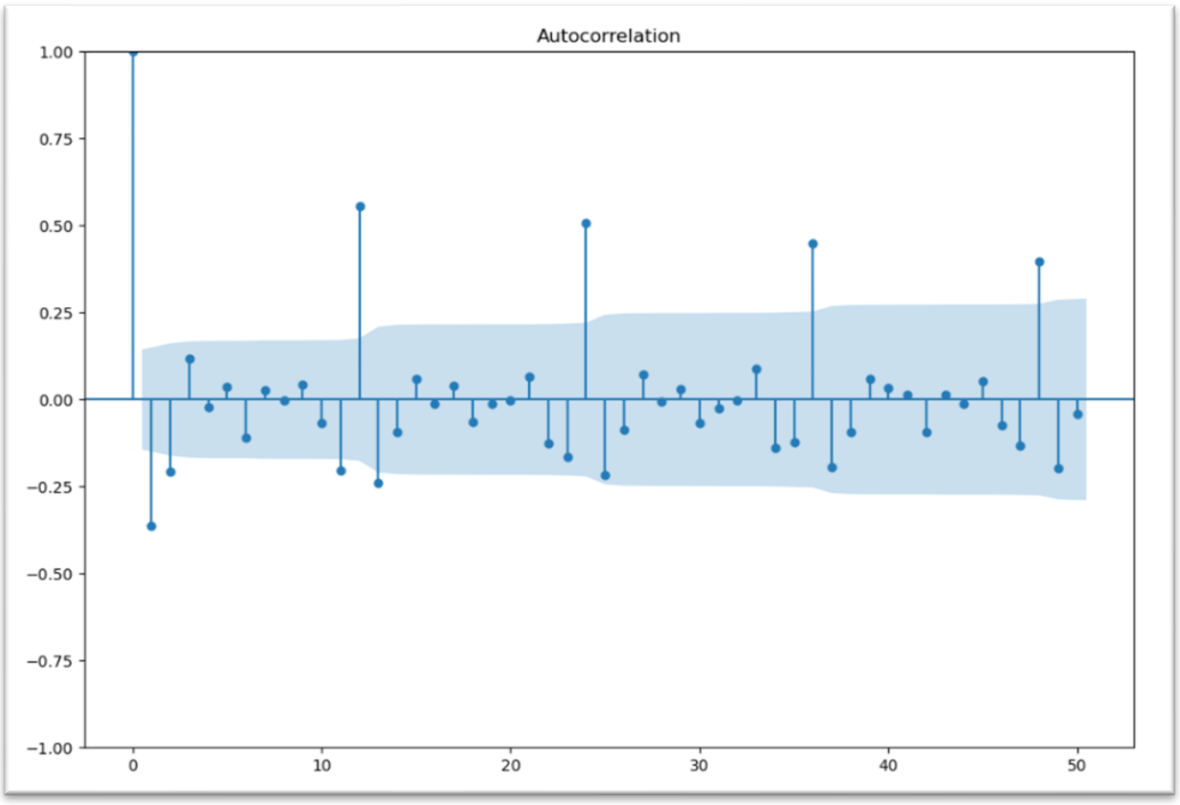

Fig 45. Autocorrelation of Rose wine Sales of 1 month difference

Observation:

- ✓ To ensure stationarity, we will consider the 1-month sales difference for the ACF plot.
- ✓ Based on the plot, we estimate a q-value of 2, as two data points exceed the confidence

interval.

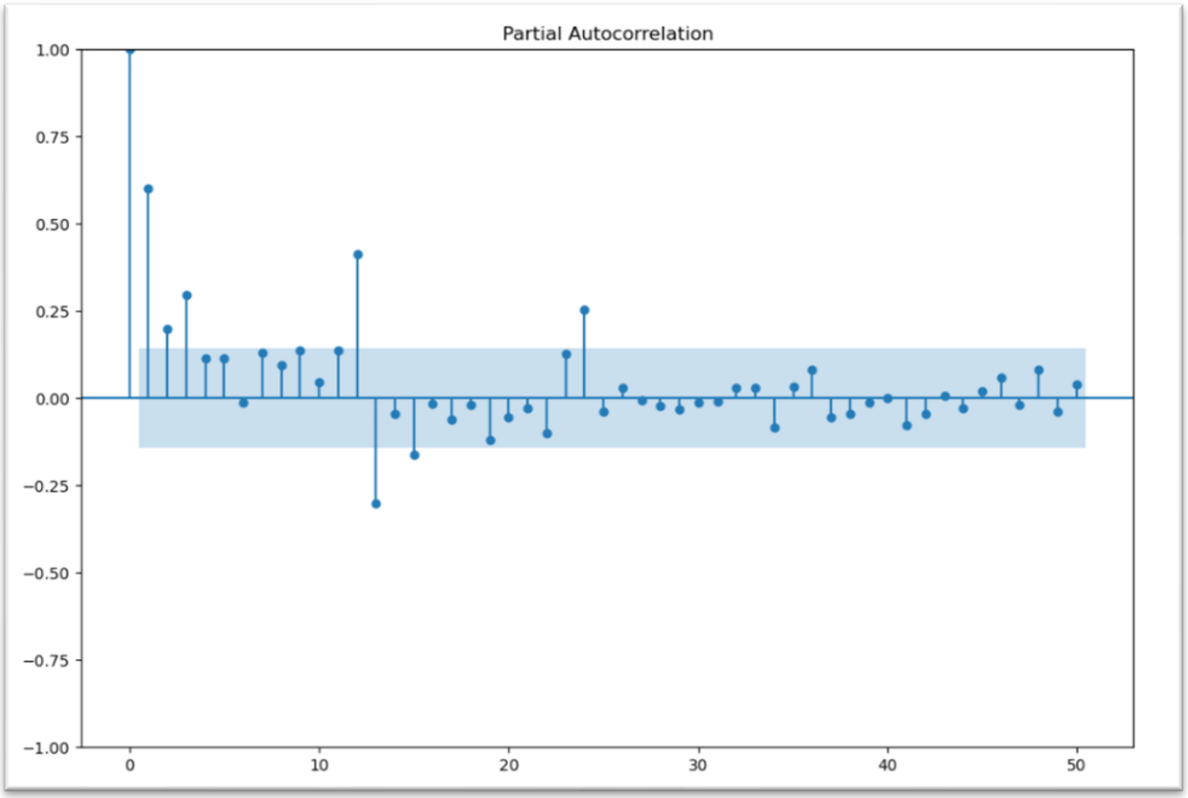

Fig 46. Partial Autocorrelation of Rose Wine Sales

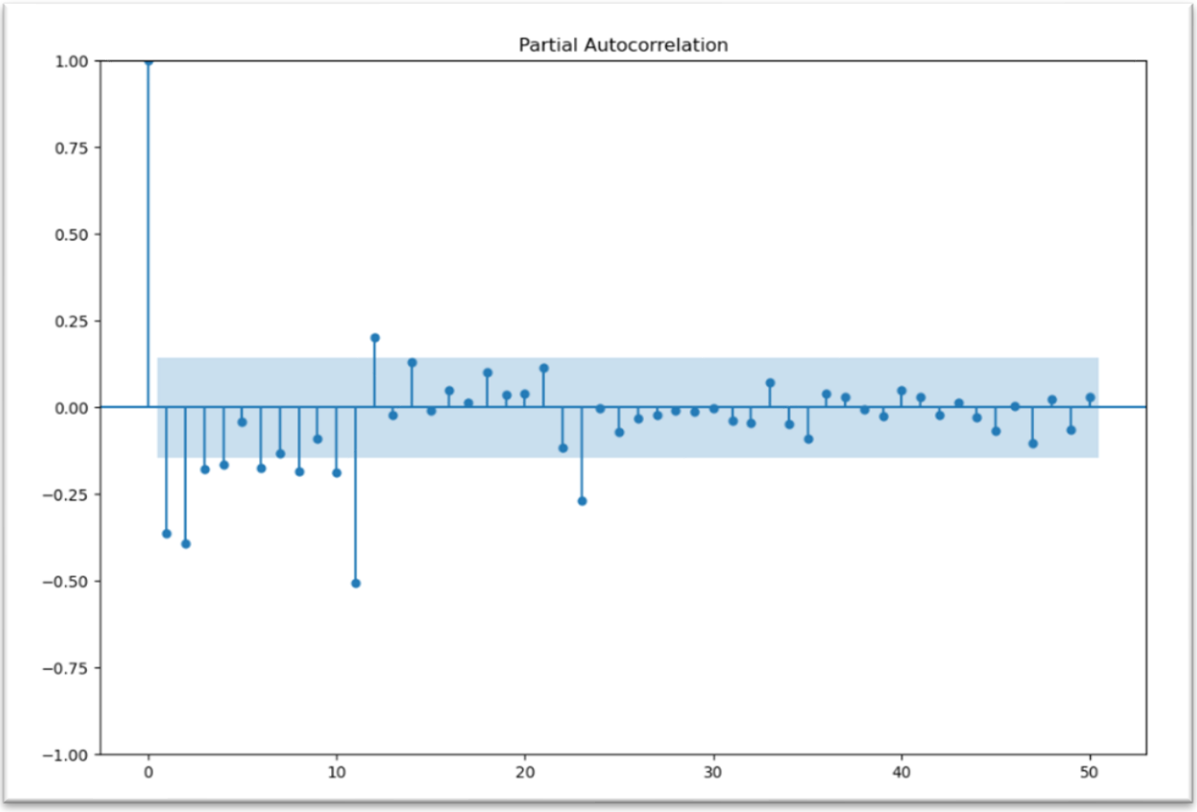

Fig 47. Partial Autocorrelation of Rose Wine Sales of 1 month difference

Observations:

- ✓ To ensure stationarity, we will consider the 1-month sales difference for PACF plot also.
- ✓ Based on the plot, we estimate a p-value of 4, as three data points exceed the confidence interval.

ARIMA Model

Manual ARIMA Model:

- ✓ We split the data into train and test set with 80% train set size.
- ✓ For the ARIMA model we choose **p=4, q=2, d=1** from the above observations.
- ✓ We fit the model with train data and analyze the performance.
- ✓ From the model summary we observe that the past 4 months' sales and past 2 months' forecast error contribute to the current forecast but are not very significant.
- ✓ Residual diagnostics indicate autocorrelation is not present (Ljung-Box) but having heteroskedasticity, which could affect forecasting accuracy.
- ✓ The root mean square error on the test data is 37.06
- ✓ From the plot below, we observe that the forecast fails to capture the variation in the data.

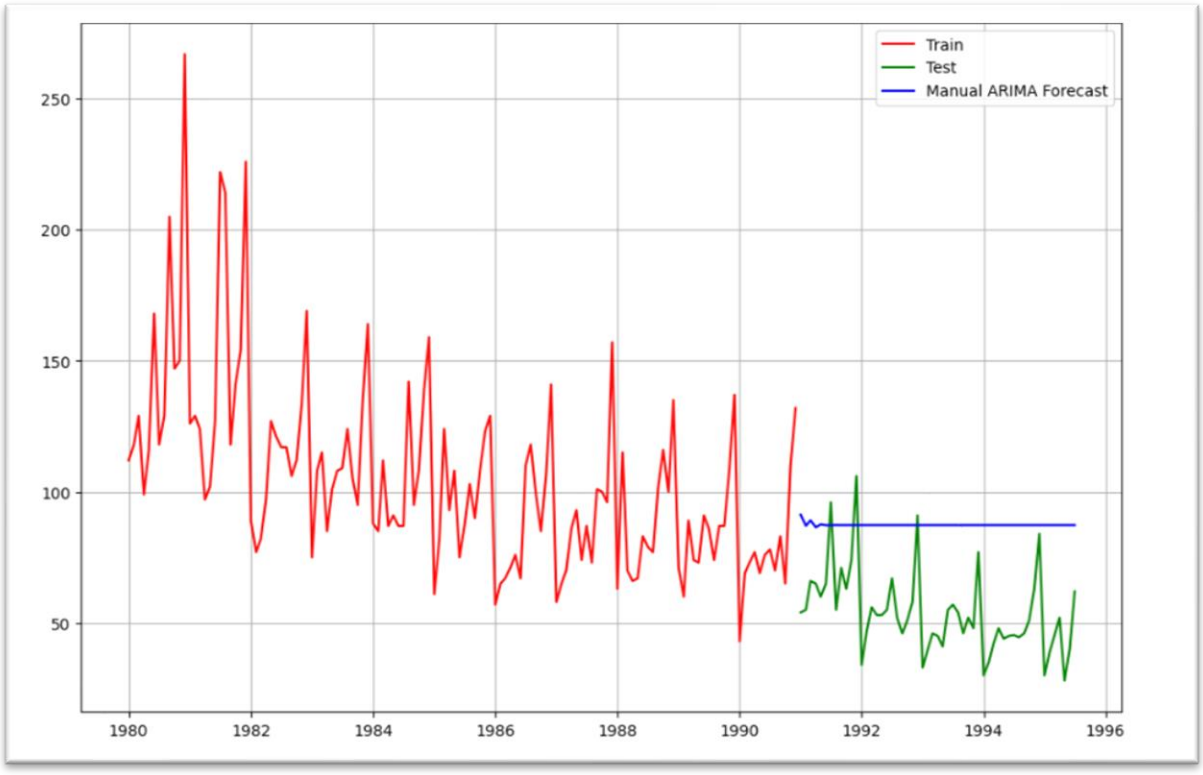

Fig 48. Manual ARIMA Forecasting

Auto ARIMA Model:

- ✓ We split the data into train and test set with 80% train set size.
- ✓ We evaluate the model across a range of p and q values while keeping d fixed at 1 and select the model with the lowest AIC value as the best fit.
- ✓ From the model summary we observe that the past 2 months' sales and second last month's forecast error significantly contribute to the current forecast.
- ✓ Residual diagnostics indicate autocorrelation is not present (Ljung-Box) but having heteroskedasticity, which could affect forecasting accuracy.
- ✓ The root mean square error on the test data is 36.8
- ✓ From the plot below, we observe that this time forecast is better than the previous ARIMA model, it can capture the variation initially but fails in longer period.

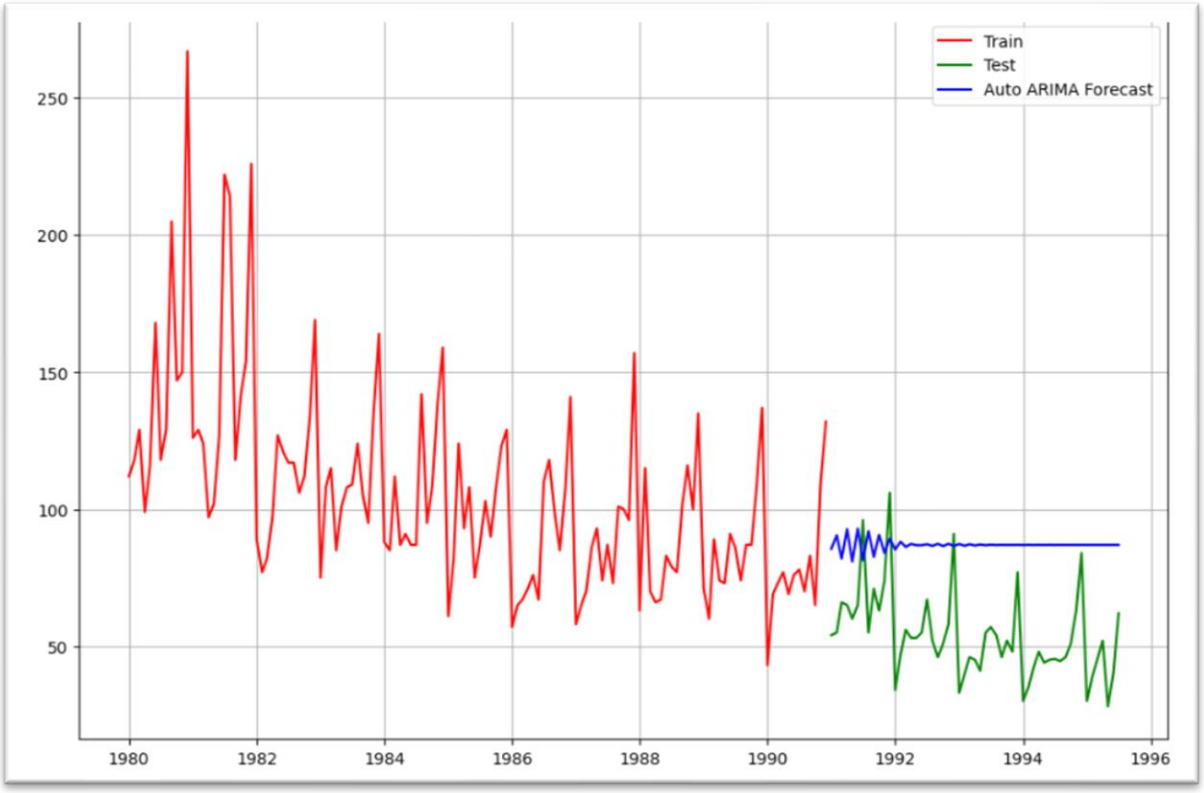

Fig 49. Auto ARIMA Forecasting

Manual SARIMA Model:

- ✓ We split the data into train and test set with 80% train set size.
- ✓ For the SARIMA model we choose **p=4, q=2, d=1** as analyzed before.
- ✓ We see from ACF and PACF plots that the pattern repeats itself every 1 year. So, the seasonal period is 12 months.
- ✓ We perform the dicky fuller test to ensure the stationarity of the seasoned data. P-value obtained is less than 0.05 indicates the seasoned data is stationary. So D is 0.
- ✓ We plot ACF and PACF on seasoned and identify **P and Q value as 2** for both
- ✓ We fit the model with train data with above parameters and analyze the performance.

- ✓ From the model summary we observe that last month's sales, last season (12 months prior) sales and last to last season (24 months prior) sales significantly contribute to the current forecast.
- ✓ Residual diagnostics indicate there is no autocorrelation (Ljung-Box) and no issues of heteroskedasticity present in the residuals, and it is normally distributed.
- ✓ The root mean square error on the test data is 27.6 which is reduced significantly.
- ✓ From the plot below, we observe that forecast pretty much follows the actual trend.

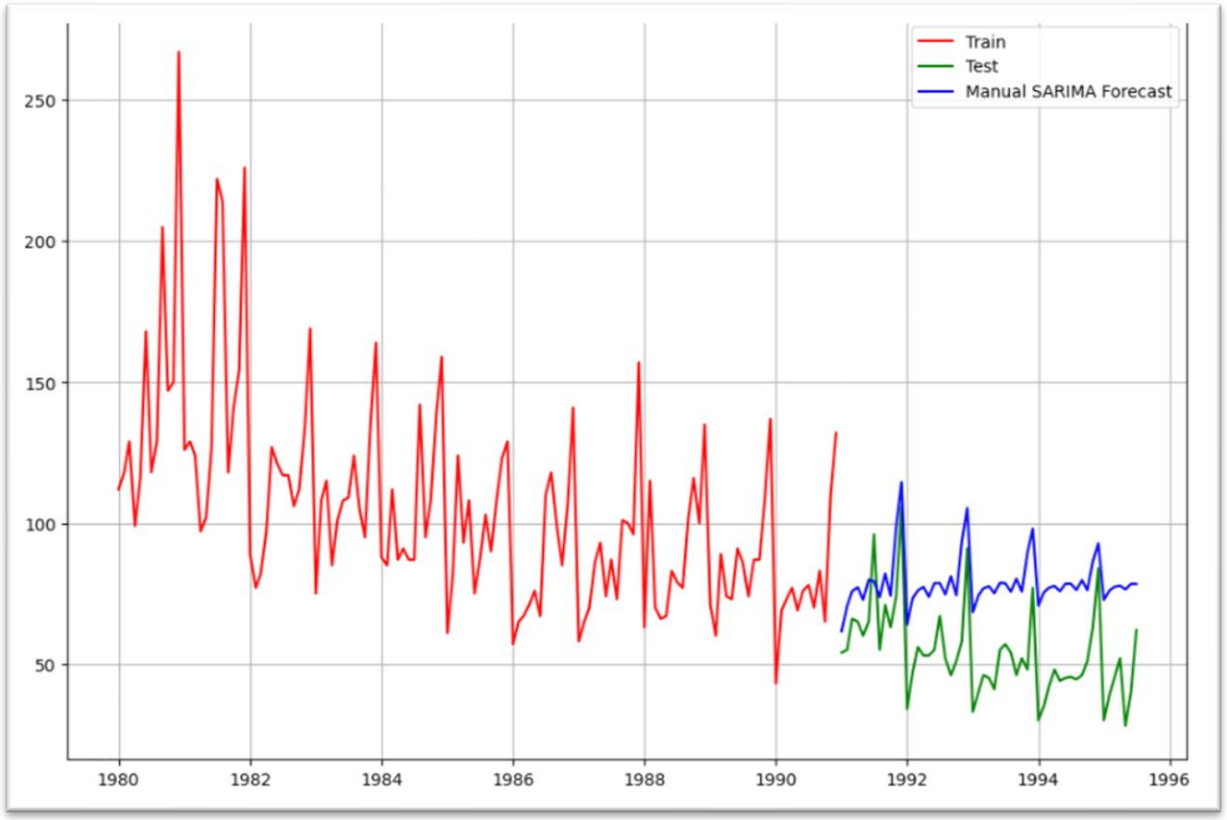

Fig 50. Manual SARIMA Forecasting

AUTO SARIMA Model:

- ✓ We split the data into train and test set with 80% train set size.
- ✓ We evaluate the model across a range of p and q values while keeping d fixed at 1 and a range of seasonal parameters P and Q while keeping D fixed at 0 with seasonal period of 12 month and select the model with the lowest AIC value as the best fit.
- ✓ From the model summary we observe that last season's sales (12 months prior) and last to last season's sales significantly contribute to the current forecast.
- ✓ Residual diagnostics indicate there is no autocorrelation (Ljung-Box) and no issues of heteroskedasticity present in the residuals, and it is normally distributed.
- ✓ The root mean square error on the test data is 26.9
- ✓ From the plot below, we observe that this time forecast is pretty much similar to manual SARIMA model.

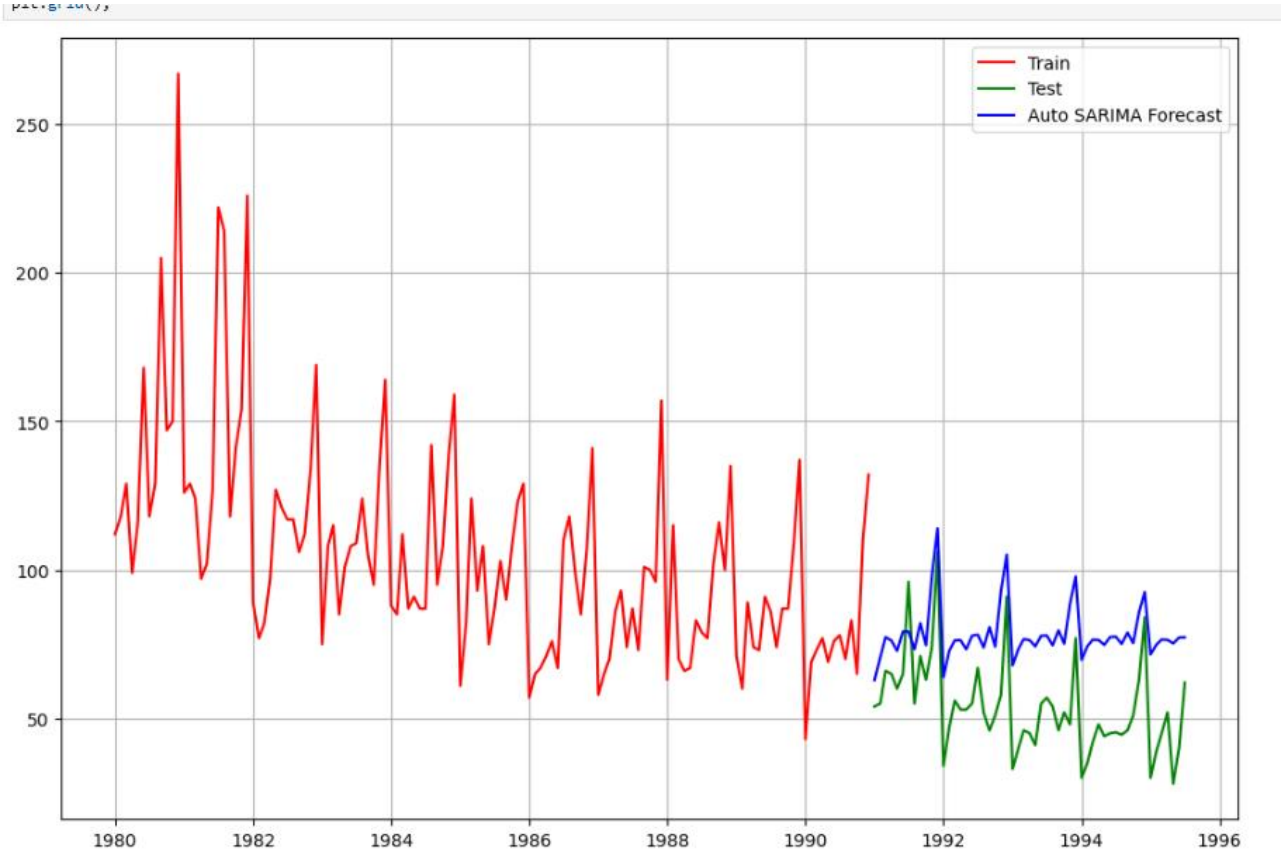

Fig 51. Auto SARIMA Forecasting

Check Performance of the Models:

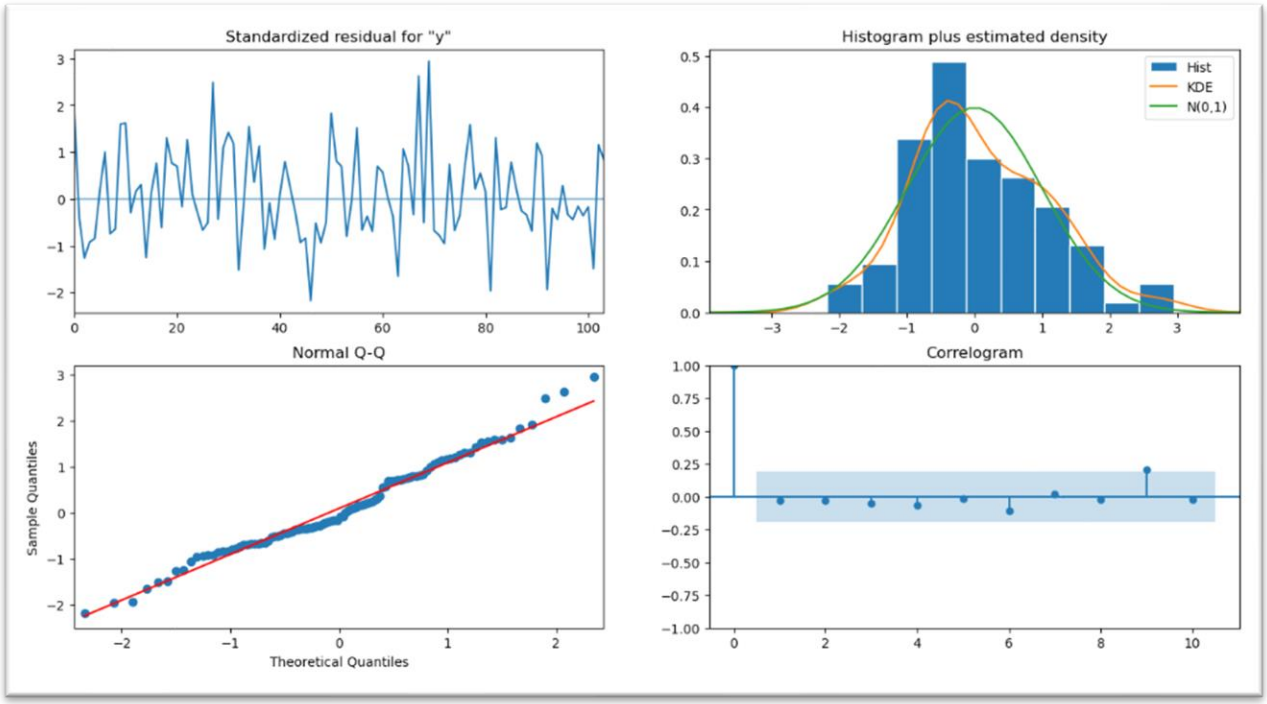

Fig 52. Residual Plots of Auto SARIMA Model

Observations:

- ✓ **Standardized Residuals**: Appear to be randomly scattered, suggesting no obvious patterns.
- ✓ **Histogram and KDE:** The residuals almost follow a normal distribution.
- ✓ **Normal Q-Q Plot:** Shows very little deviations from the 45-degree line, indicating a slight non-normality in the residuals.
- ✓ **Correlogram:** No significant autocorrelation is observed, suggesting that the residuals are approximately white noise.

Comparing Model Performance:

Model	Test RMS	Autocorrelation in Residuals (Ljung-Box Test)	Normal Distribution of the Residuals (Jarque-Bera Test:)	Heteroskedasticity in the residual
Manual ARIMA (4,1,2)	**37.06**	**Doesn't exist**	**Doesn't Exists**	**Exists**
Auto ARIMA (2,1,3)	**36.8**	**Doesn't exist**	**Doesn't Exists**	**Exists**
Manual SARIMA (2,1,2) (2,0,2,12)	**27.6**	**Doesn't exist**	**Exists**	**Doesn't Exists**
Auto SARIMA (0,1,2)(2,0,2,12)	**26.9**	**Doesn't exist**	**Exists**	**Doesn't Exists**

Fig 53. Model Comparison

Observation:

- The **Auto SARIMA (0,1,2)(2,0,2,12)** model is the **best choice based** on the following reasons:
  - ✓ **Lowest Test RMS**: This model has the lowest Test RMS value (26.9), indicating the best fit to the data among the four models.
  - ✓ **No Autocorrelation in Residuals**: The Ljung-Box test shows no significant autocorrelation in the residuals, suggesting that the model has effectively captured the time series structure.
  - ✓ **No Heteroskedasticity**: The absence of heteroskedasticity means that the variance of the residuals is constant over time, which is a desirable property for a well-fitted model.

- While residual is almost following the normal distribution in the Auto SARIMA model, the significant reduction in Test RMS and the absence of autocorrelation and heteroskedasticity make it a strong candidate.

Building the model with full Data:

- ✓ We developed a SARIMA model with the parameters (0,1,2)(2,0,2,12) using a dataset of 187 records.
- ✓ The last month's sales, last season's sales (12 months prior) and last season to season's(24 months prior) sales significantly contribute to the current forecast.
- ✓ Residual diagnostic is as follows:

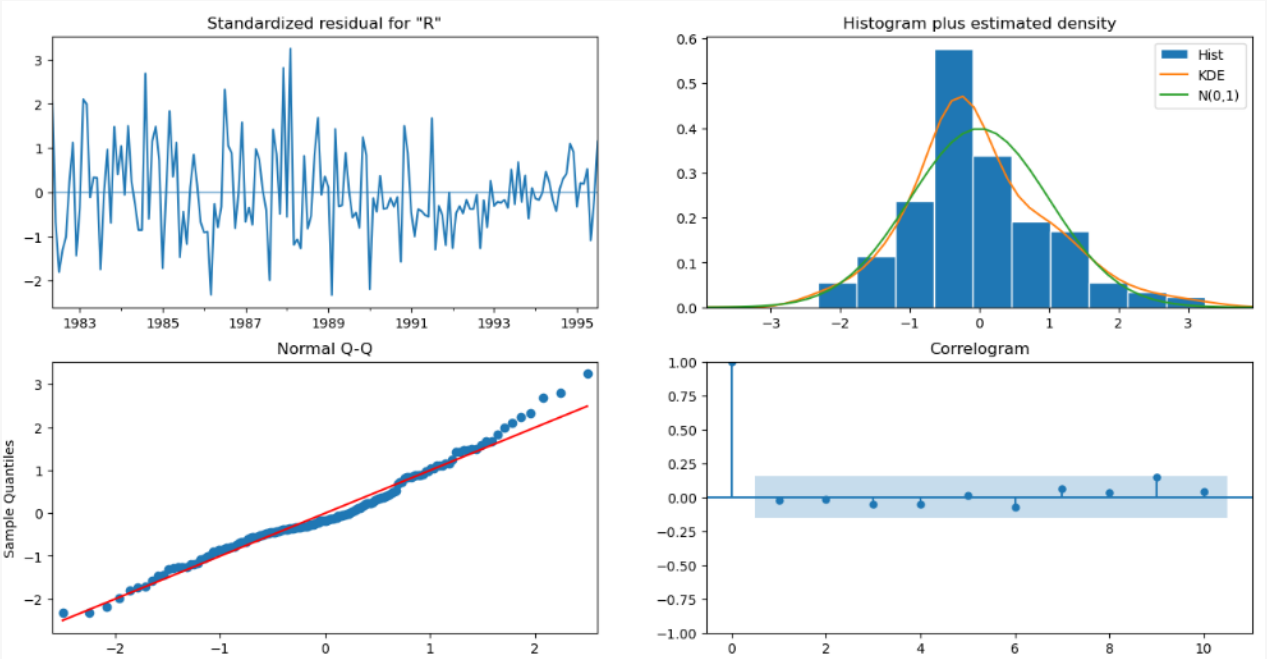

Fig 54. Residual Plots of Final SARIMA Model

Forecast for the next 12 months:

	Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01		46.901483	14.114419	19.237731	74.565236
1995-09-01		44.057391	14.497394	15.643022	72.471761
1995-10-01		47.154458	14.559916	18.617547	75.691370
1995-11-01		52.482041	14.622171	23.823112	81.140970
1995-12-01		69.585574	14.684162	40.805145	98.366003

Fig 55. Sales Forecast of Next 12 months using SARIMA – 1st 5 records

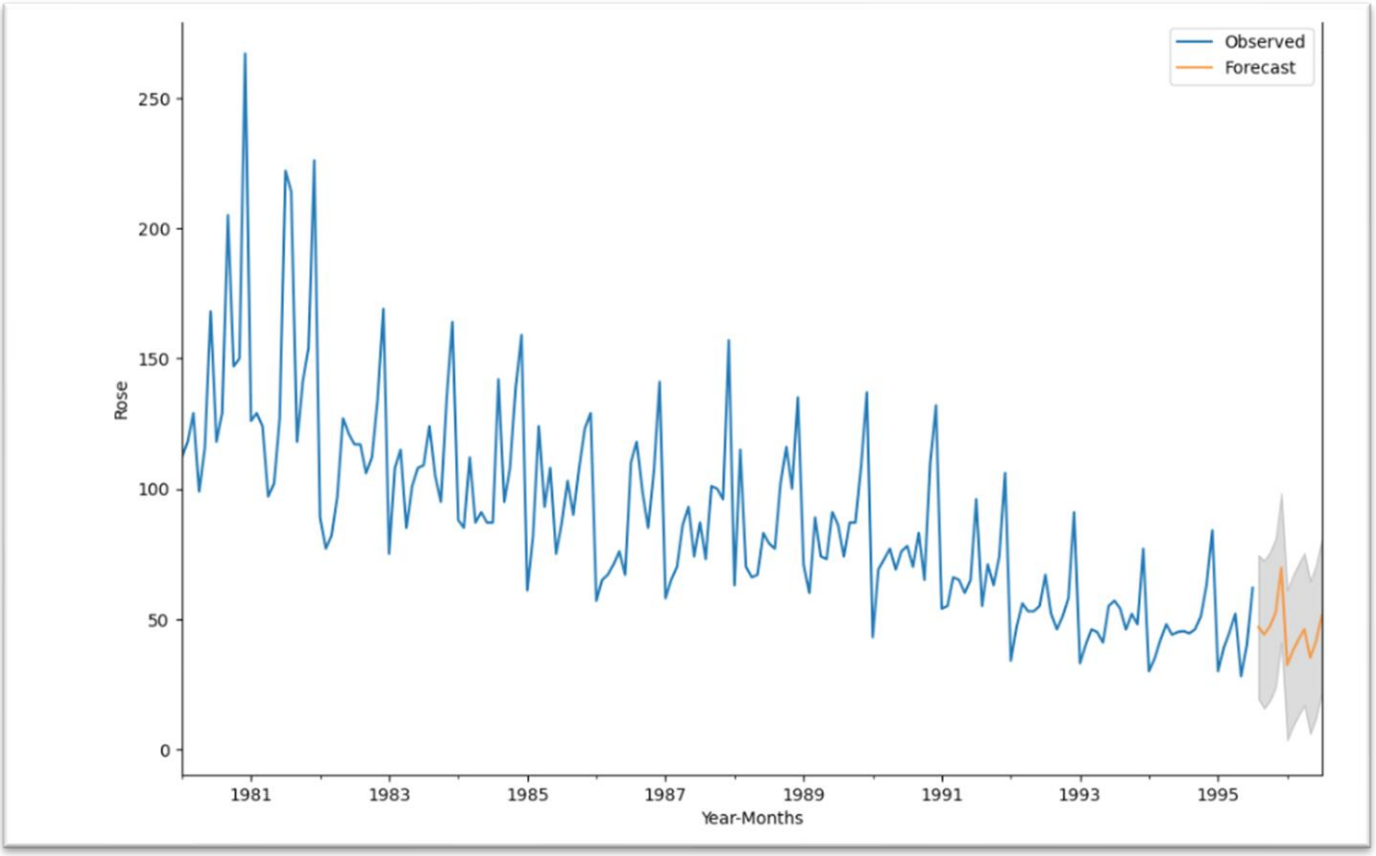

Fig 56. Forecast Trend from Aug,1995 to July 1996 with confidence interval

Summary:

- ✓ We developed several models starting with linear regression to different combination of auto-regression and moving average models integrating the differencing methodology along with seasonal variation to forecast the sparkling wine sales most accurately for the future.
- ✓ The key metrics used to achieve the optimal model for forecasting are Root mean square error (RMSE) computed on most recent dataset (test data). Lesser RMSE indicates better model performance.
- ✓ After fine-tuning the contributing parameters, the SARIMA model emerged as the best fit compared to the other models for predicting sparkling wine sales. This model effectively captures the seasonal patterns and trends in the data, providing more accurate and reliable forecasts for future sales.
- ✓ Though there are extreme variations present in the past data, pattern gets smoothed out in the recent past, and the model can capture the smoothness pretty well.
- ✓ Though SARIMA model provides a reasonable forecast, including more data points or external variables might help improve the accuracy of the forecasts.

Business Insights:

1. **Seasonal Trends:**

   The observed data's fluctuations for both the Rose and Sparkling Wine sales likely correspond to seasonal trends. Understanding these patterns can help in planning inventory and marketing strategies.

2. **Stabilization:**

   The forecasted data suggests a stabilization in future trends. The Rose wine sales trend get much smoother than Sparkling. Sudden Spike is still present in Sparkling wine forecast. This could indicate even though the rose wine sales trend goes downward, its market is getting matured gradually. On the other hand, sparkling wine sales yet to get the stability in the market.

3. **Demand Planning:**

   The forecasted trend can be used for demand planning, ensuring that inventory levels are aligned with expected future demand, reducing the risk of overstocking or stockouts.

4. **Resource Allocation:**

   With a smoother forecasted trend, businesses can better allocate resources, such as staffing and promotional efforts, to match anticipated demand.

5. **Strategic Decisions**:

   The insights from the forecast can inform strategic decisions, such as expanding into new markets, launching new products, or adjusting pricing strategies.