PGP-DSBa project report

TSF – Coded Project

**BY**

**Ishaan Shakti Jayaraman**

**PGPDSBA.O.JULY24.A**

Contents

[LIST OF FIGURES 4](#_Toc193058079)

[INTRODUCTION 7](#_Toc193058080)

[1.1 Objective 7](#_Toc193058081)

[1.2 Problem Definition 7](#_Toc193058082)

[1.3 Data Contents 7](#_Toc193058083)

[1.4 Statistical Summary 8](#_Toc193058084)

[1.4.1 Sparkling Wine 8](#_Toc193058085)

[1.4.2 Rose Wine 8](#_Toc193058086)

[EXPLORATORY DATA ANALYSIS – SPARKLING WINE 9](#_Toc193058087)

[2.1 Time Series Trend Chart 9](#_Toc193058088)

[2.2 Yearly Sales Box Plot 10](#_Toc193058089)

[2.3 Monthly Sales Box Plot 10](#_Toc193058090)

[2.4 Monthly Sales across Years 11](#_Toc193058091)

[2.5 Decomposition 11](#_Toc193058092)

[2.5.1 Additive 12](#_Toc193058093)

[2.5.2 Multiplicative 13](#_Toc193058094)

[DATA PRE-PROCESSING – SPARKLING WINE 14](#_Toc193058095)

[3.1 Missing Value Treatment 14](#_Toc193058096)

[3.2 Data Preparation 14](#_Toc193058097)

[MODEL BUILDING – ORIGINAL DATA - SPARKLING WINE 15](#_Toc193058098)

[4.1 Linear Regression Model 15](#_Toc193058099)

[4.2 Simple Average Model 15](#_Toc193058100)

[4.3 Moving Average Model 16](#_Toc193058101)

[4.4 Single Exponential Smoothing Model 18](#_Toc193058102)

[4.5 Double Exponential Smoothing Model 20](#_Toc193058103)

[4.6 Triple Exponential Smoothing Model 21](#_Toc193058104)

[CHECK FOR STATIONARITY – SPARKLING WINE 23](#_Toc193058105)

[MODEL BUILDING – STATIONARY – SPARKLING WINE 24](#_Toc193058106)

[6.1 ACF & PACF Plot 24](#_Toc193058107)

[6.2 Auto ARIMA Model 26](#_Toc193058108)

[6.3 Manual ARIMA Model 28](#_Toc193058109)

[6.4 Auto SARIMA Model 30](#_Toc193058110)

[6.5 Manual SARIMA Model 31](#_Toc193058111)

[COMPARISON OF MODELS – SPARKLING WINE 33](#_Toc193058112)

[7.1 Rebuilding the Model 34](#_Toc193058113)

[ACTIONABLE INSIGHTS & RECOMMENDATIONS – SPARKLING WINE 36](#_Toc193058114)

[EXPLORATORY DATA ANALYSIS – ROSE WINE 37](#_Toc193058115)

[9.1 Time Series Trend Chart 37](#_Toc193058116)

[9.2 Yearly Sales Box Plot 38](#_Toc193058117)

[9.3 Monthly Sales Box Plot 38](#_Toc193058118)

[9.4 Monthly Sales Across Years 39](#_Toc193058119)

[9.5 Decomposition 39](#_Toc193058120)

[9.5.1 Additive Model 40](#_Toc193058121)

[9.5.2 Multiplicative Model 41](#_Toc193058122)

[DATA PRE-PROCESSING – ROSE WINE 42](#_Toc193058123)

[10.1 Missing Value Treatment 42](#_Toc193058124)

[10.2 Data Preparation 42](#_Toc193058125)

[MODEL BUILDING – ORIGINAL DATA – ROSE WINE 42](#_Toc193058126)

[11.1 Linear Regression Model 42](#_Toc193058127)

[11.2 Simple Average Model 43](#_Toc193058128)

[11.3 Moving Average Model 44](#_Toc193058129)

[11.4 Single Exponential Smoothing Model 47](#_Toc193058130)

[11.5 Double Exponential Smoothing Model 48](#_Toc193058131)

[11.6 Triple Exponential Smoothing Model 49](#_Toc193058132)

[CHECK FOR STATIONARITY – ROSE WINE 51](#_Toc193058133)

[MODEL BUILDING – STATIONARY – ROSE WINE 53](#_Toc193058134)

[13.1 ACF & PACF Plot 53](#_Toc193058135)

[13.2 Auto ARIMA Model 55](#_Toc193058136)

[13.3 Manual ARIMA Model 57](#_Toc193058137)

[13.4 Auto SARIMA Model 59](#_Toc193058138)

[13.5 Manual SARIMA Model 60](#_Toc193058139)

[COMPARISON OF MODELS – ROSE WINE 62](#_Toc193058140)

[14.1 Rebuilding the Model 63](#_Toc193058141)

[ACTIONABLE INSIGHTS & RECOMMENDATIONS – ROSE WINE 65](#_Toc193058142)

# LIST OF FIGURES

[Figure 1 - Statistical Summary of Sparkling Wine Data 7](#_Toc193057999)

[Figure 2 - Statistical Summary of Rose Wine Data 7](#_Toc193058000)

[Figure 3 - Sparkling Wine Sales Trend Chart 8](#_Toc193058001)

[Figure 4 - Yearly Sales of Sparkling Wine - Box Plot 9](#_Toc193058002)

[Figure 5 - Monthly Sales of Sparkling Wine - Box Plot 9](#_Toc193058003)

[Figure 6 - Monthly sales across Years - Trend Chart 10](#_Toc193058004)

[Figure 7 - Additive Decomposition of Sparkling Wine Data 11](#_Toc193058005)

[Figure 8 - Multiplicative Decomposition of Sparkling Wine Data 12](#_Toc193058006)

[Figure 9 - Train/Test Split of Sparkling Wine Data 13](#_Toc193058007)

[Figure 10 - Linear Regression Model on Test Data 14](#_Toc193058008)

[Figure 11 - Simple Average Model on Test Data 15](#_Toc193058009)

[Figure 12 – Sample Values of Moving Averages 16](#_Toc193058010)

[Figure 13 - Moving Average Model on Entire Dataset 16](#_Toc193058011)

[Figure 14 - Moving Average Model on Test Data 17](#_Toc193058012)

[Figure 15 - Single Exponential Smoothing Model Parameters 18](#_Toc193058013)

[Figure 16 - Single Exponential Smoothing Model on Test Data 18](#_Toc193058014)

[Figure 17 - Double Exponential Smoothing Model Parameters 19](#_Toc193058015)

[Figure 18 - Double Exponential Smoothing Model on Test Data 19](#_Toc193058016)

[Figure 19 - Triple Exponential Smoothing Model Parameters 20](#_Toc193058017)

[Figure 20 - Triple Exponential Smoothing Model on Test Data 20](#_Toc193058018)

[Figure 21 - Different Parameters for Triple Exponential Smoothing Model 21](#_Toc193058019)

[Figure 22 - Triple Exponential Model on Test Data 21](#_Toc193058020)

[Figure 23 - Dickey-Fuller Test on Normal Data 22](#_Toc193058021)

[Figure 24 - Dickey-Fuller Test on Differenced Data 23](#_Toc193058022)

[Figure 25 - ACF Plot 24](#_Toc193058023)

[Figure 26 - PACF Plot 25](#_Toc193058024)

[Figure 27 - Parameters Combinations & AIC values – ARIMA Model 26](#_Toc193058025)

[Figure 28 - Sorted Parameter and AIC Value - ARIMA Model 26](#_Toc193058026)

[Figure 29 - Auto ARIMA Model 27](#_Toc193058027)

[Figure 30 - Manual ARIMA Model 28](#_Toc193058028)

[Figure 31 - Manual ARIMA Model on Test Data 28](#_Toc193058029)

[Figure 32 - Auto SARIMA Model 29](#_Toc193058030)

[Figure 33 - Auto SARIMA Model on Test Data 30](#_Toc193058031)

[Figure 34 - Manual SARIMA Model 31](#_Toc193058032)

[Figure 35 - Manual SARIMA Model on Test Data 31](#_Toc193058033)

[Figure 36 - Comparion of All Models 32](#_Toc193058034)

[Figure 37 - TES Model - Predicted vs Actual Values 33](#_Toc193058035)

[Figure 38 - TES Model on Test Data 33](#_Toc193058036)

[Figure 39 - TES Best Model 34](#_Toc193058037)

[Figure 40 - TES Model - Forecast values 34](#_Toc193058038)

[Figure 41 - Triple Exponential Smoothing Model – Forecast Trend 34](#_Toc193058039)

[Figure 42 - Rose Wine Sales Trend Chart 36](#_Toc193058040)

[Figure 43 - Yearly Sales of Rose Wine - Box Plot 37](#_Toc193058041)

[Figure 44 - Monthly Sales of Rose Wine - Box Plot 37](#_Toc193058042)

[Figure 45 - Monthly Sales Across Years - Trend Chart 38](#_Toc193058043)

[Figure 46 - Additive Decomposition of Rose Wine Data 39](#_Toc193058044)

[Figure 47 - Multiplicative Decomposition of Rose Wine Data 40](#_Toc193058045)

[Figure 48 - Train/Test Split of Rose Wine Sales 41](#_Toc193058046)

[Figure 49 - Linear Regression Model on Test Data 42](#_Toc193058047)

[Figure 50 - Simple Average Model on Test Data 43](#_Toc193058048)

[Figure 51 - Sample Values of Moving Averages 44](#_Toc193058049)

[Figure 52 - Moving Average Model on Entire Dataset 44](#_Toc193058050)

[Figure 53 - Moving Average Model on Test Data 45](#_Toc193058051)

[Figure 54- Single Exponential Model Parameters 46](#_Toc193058052)

[Figure 55 - Single Exponential Smoothing Model on Test Data 46](#_Toc193058053)

[Figure 56 - Double Exponential Smoothing Model Parameters 47](#_Toc193058054)

[Figure 57 - Double Exponential Smoothing Model on Test Data 48](#_Toc193058055)

[Figure 58 - Triple Exponential Smoothing Model Parameters 48](#_Toc193058056)

[Figure 59 - Triple Exponential Smoothing Model on Test Data 49](#_Toc193058057)

[Figure 60 - Different Parameters for Triple Exponential Smoothing Model 49](#_Toc193058058)

[Figure 61 - Triple Exponential Smoothing Model on Test Data 50](#_Toc193058059)

[Figure 62 - Dickey-Fuller Test 51](#_Toc193058060)

[Figure 63 - Dickey-Fuller Test on Differenced Data 52](#_Toc193058061)

[Figure 64 - ACF Plot 53](#_Toc193058062)

[Figure 65 - PACF Plot 54](#_Toc193058063)

[Figure 66 - Parameters Combinations & AIC values – ARIMA Model 55](#_Toc193058064)

[Figure 67 - Sorted Parameter and AIC Value - ARIMA Model 55](#_Toc193058065)

[Figure 68 - Auto ARIMA Model 56](#_Toc193058066)

[Figure 69 - Manual ARIMA Model 57](#_Toc193058067)

[Figure 70 - Manual ARIMA Model on Test Data 57](#_Toc193058068)

[Figure 71 - Auto SARIMA Model 58](#_Toc193058069)

[Figure 72 - Auto SARIMA Model on Test Data 59](#_Toc193058070)

[Figure 73 - Manual SARIMA Model 60](#_Toc193058071)

[Figure 74 - Manual SARIMA Model on Test Data 60](#_Toc193058072)

[Figure 75 - Comparion of All Models 61](#_Toc193058073)

[Figure 76 - TES Model - Predicted vs Actual Values 62](#_Toc193058074)

[Figure 77 - TES Model on Test Data 62](#_Toc193058075)

[Figure 78 - TES Best Model 63](#_Toc193058076)

[Figure 79 - TES Model - Forecast values 63](#_Toc193058077)

[Figure 80 - Triple Exponential Smoothing Model Forecast Trend 64](#_Toc193058078)

# INTRODUCTION

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. Our objective is to delve into the data, analyze trends, patterns, and factors influencing wine sales over the course of the century. By leveraging data analytics and forecasting techniques, we aim to gain actionable insights that can inform strategic decision-making and optimize sales strategies for the future.

## 1.1 Objective

The primary objective of this project is to analyze and forecast 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.

## 1.2 Problem Definition

The objective is to analyze historical wine sales (sparkling wine & rose wine) data from ABC Estate Wines to identify trends, patterns, and factors influencing sales across different wine varieties throughout the 20th century with the goal being to leverage this analysis to forecast future sales and optimize marketing strategies. By understanding past performance, the company aims to make informed decisions for future growth. The challenge is accurately predicting sales based on historical data while considering external factors. The ultimate objective is to enhance ABC Estate Wines' competitive edge and profitability in the market.

## 1.3 Data Contents

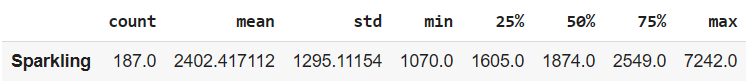
There are 2 datasets (sparkling.csv & rose.csv) consisting of monthly sales data from January 1980 to July 1995.

* Both datasets contain 187 observations.
* There are 2 columns containing object type data and numerical type data.
* “YearMonth” column consists of the date at the end of month the observation has been made from 1990-01 to 1995-07 and the other column contains the sales data of Sparkling wine or Rose wine in the datasets.
* In Sparkling Wine dataset, there are no null values and no missing values.
* In Rose Wine dataset, there are 2 null values and no missing values.

## 1.4 Statistical Summary

### 1.4.1 Sparkling Wine

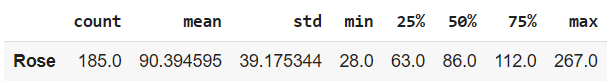
Figure 1 - Statistical Summary of Sparkling Wine Data



* There is an average of 2402 sales of sparkling wines throughout the years 1980-01 to 1995-07.
* The lowest recorded number of sparkling sales is 1070.
* The highest recorded number of sparkling sales is 7242.
* 50% of monthly sales have less than 1874 wine units sold.

### 1.4.2 Rose Wine

Figure 2 - Statistical Summary of Rose Wine Data

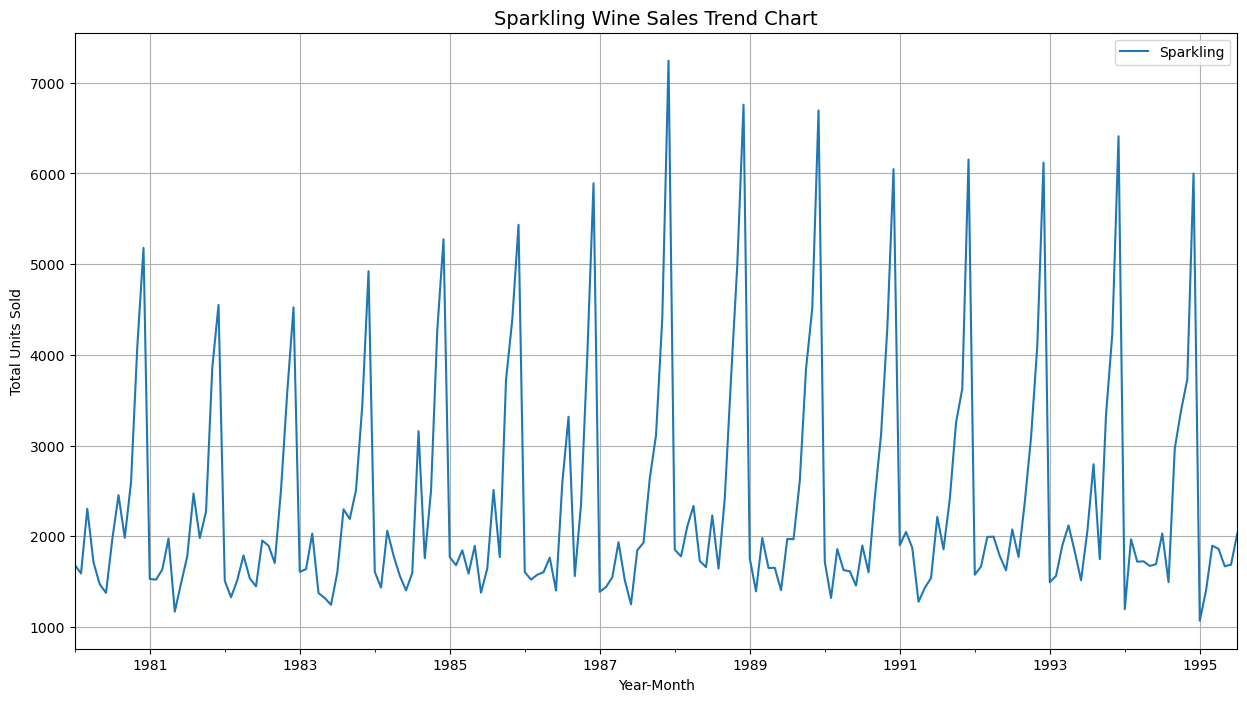


* There are an average rose wine sales of 90 throughout the years 1980-01 to 1995-07.
* The lowest recorded number of rose wine sales is 28 in a given month.
* The highest recorded number of rose wine sales is 267.
* 50% of monthly sales have less than 86 wine units sold.

# EXPLORATORY DATA ANALYSIS – SPARKLING WINE

## 2.1 Time Series Trend Chart

Figure 3 - Sparkling Wine Sales Trend Chart

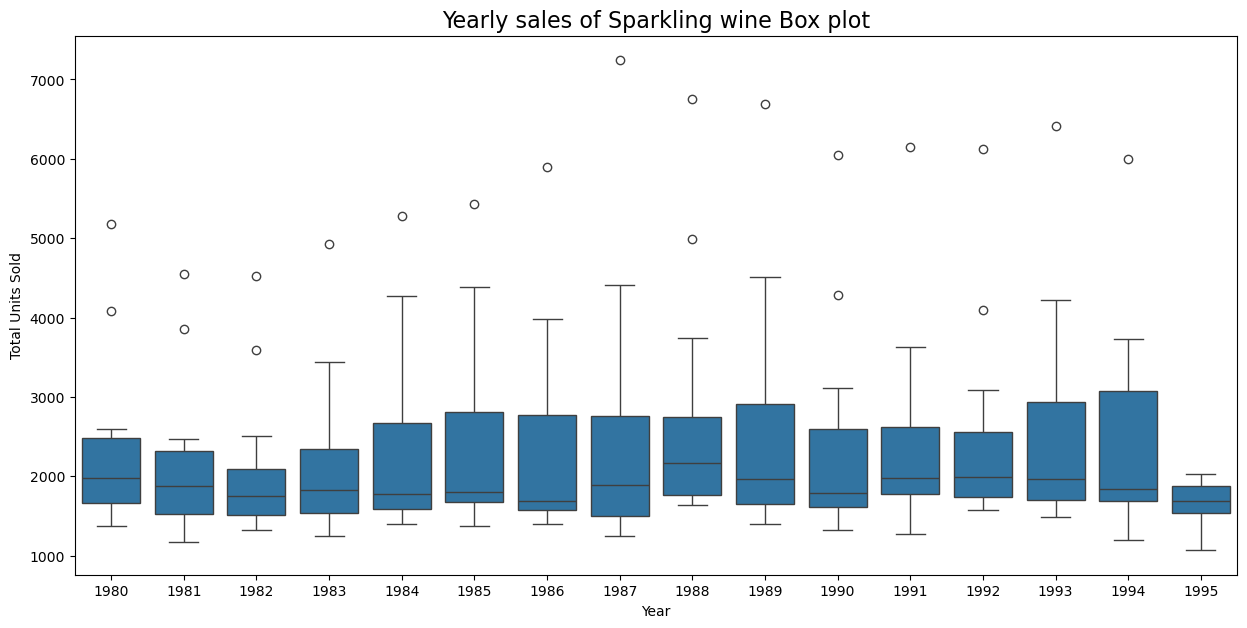


**Insights:**

* There is an increase in general sales after 1985 however it remains more or less constant throughout the years after, we can say that there is no trend.
* There is an increase in the volume of sales in any given year indicating seasonality in the data.
* The peak sales of sparkling wines happened in the year 1988.

## 2.2 Yearly Sales Box Plot

Figure 4 - Yearly Sales of Sparkling Wine - Box Plot

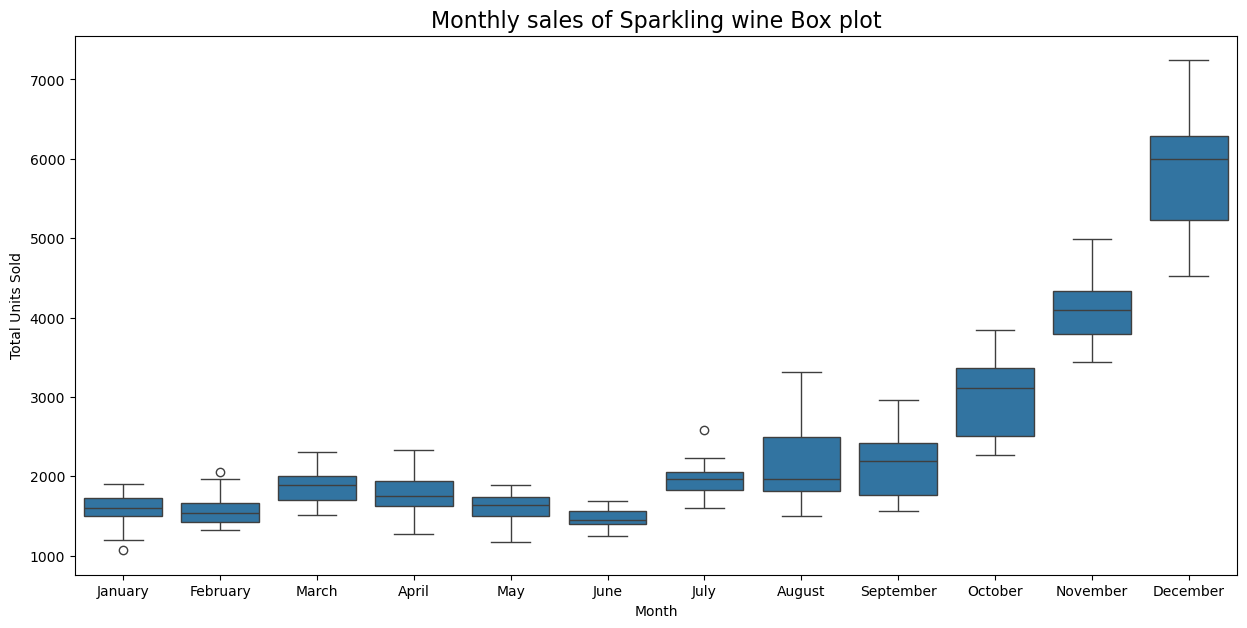


**Insights:**

* The number of sales of sparkling wine reached their lowest point in the year 1995.
* There are outliers present in every year indicating some months have higher than normal sales.
* The median of sales has a peak during the year 1988.

## 2.3 Monthly Sales Box Plot

Figure 5 - Monthly Sales of Sparkling Wine - Box Plot

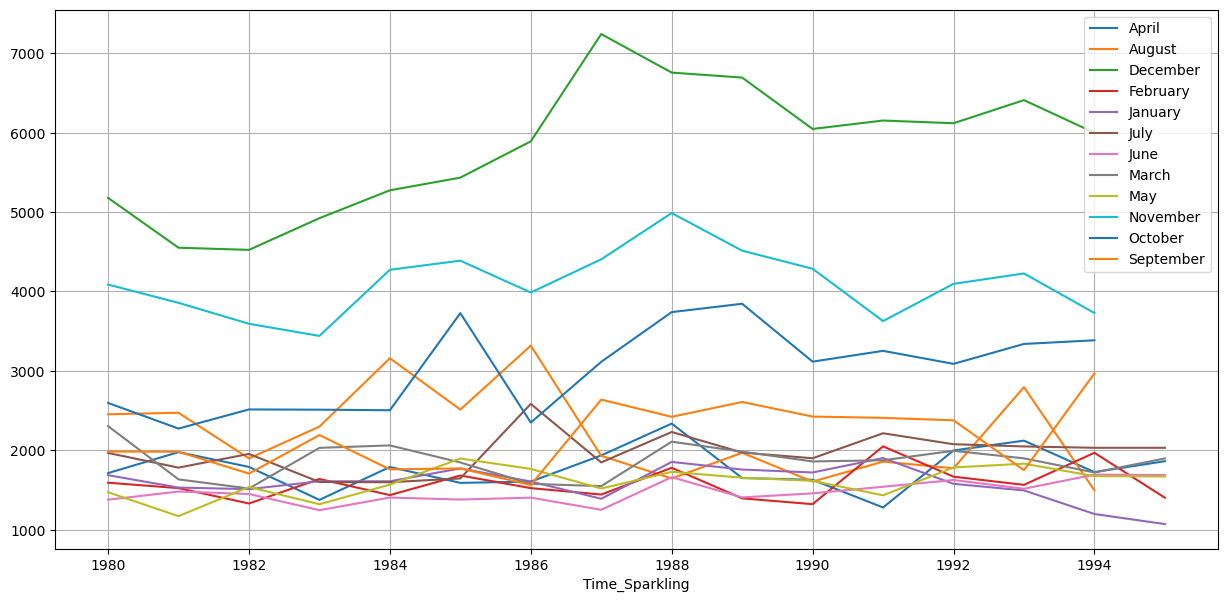


**Insights:**

* The sales of sparkling wines ae the highest during the month of December.
* The sales of sparkling wines are the lowest during January and June.
* The sale of sparkling wines steadily increases from July until December.
* There are few outliers present in month of January, February and July.

## 2.4 Monthly Sales across Years

Figure 6 - Monthly sales across Years - Trend Chart



**Insights:**

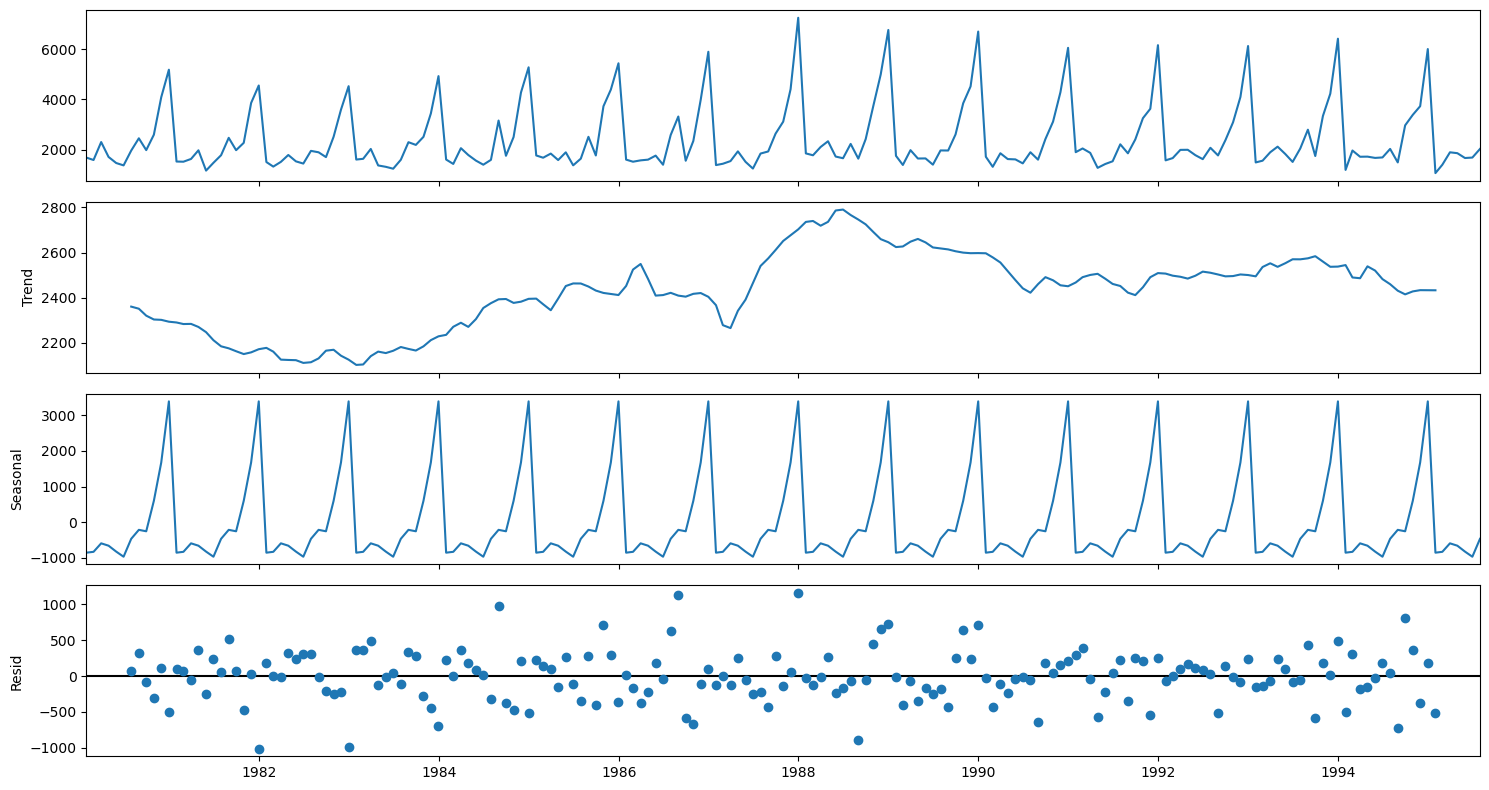
* The chart show that there is seasonality in the sale of sparkling wines as shown by the differences in sales for each month.
* For all years from 1980 to 1994, unit sales of sparkling wines are the highest during December followed by November and October.
* Throughout the years, the rest of the months have sales less than 3000 units of sparkling wines.

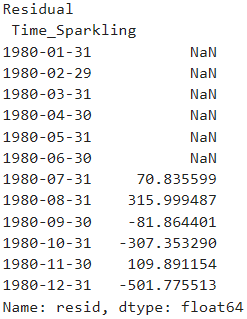
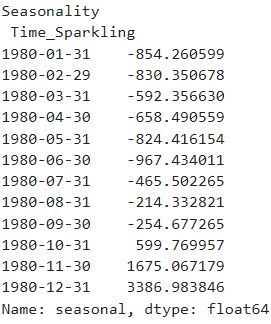
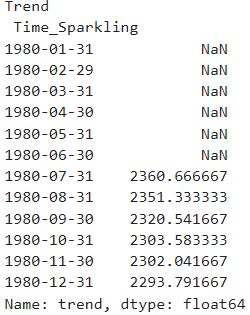
## 2.5 Decomposition

Decomposition is the process of breaking down complex data into smaller, more manageable components or subproblems to better understand the structure of the data, identify patterns, or simplify complex tasks. It involves separating a time series into its underlying components, which are trend, seasonality and residuals.

### 2.5.1 Additive

Figure 7 - Additive Decomposition of Sparkling Wine Data



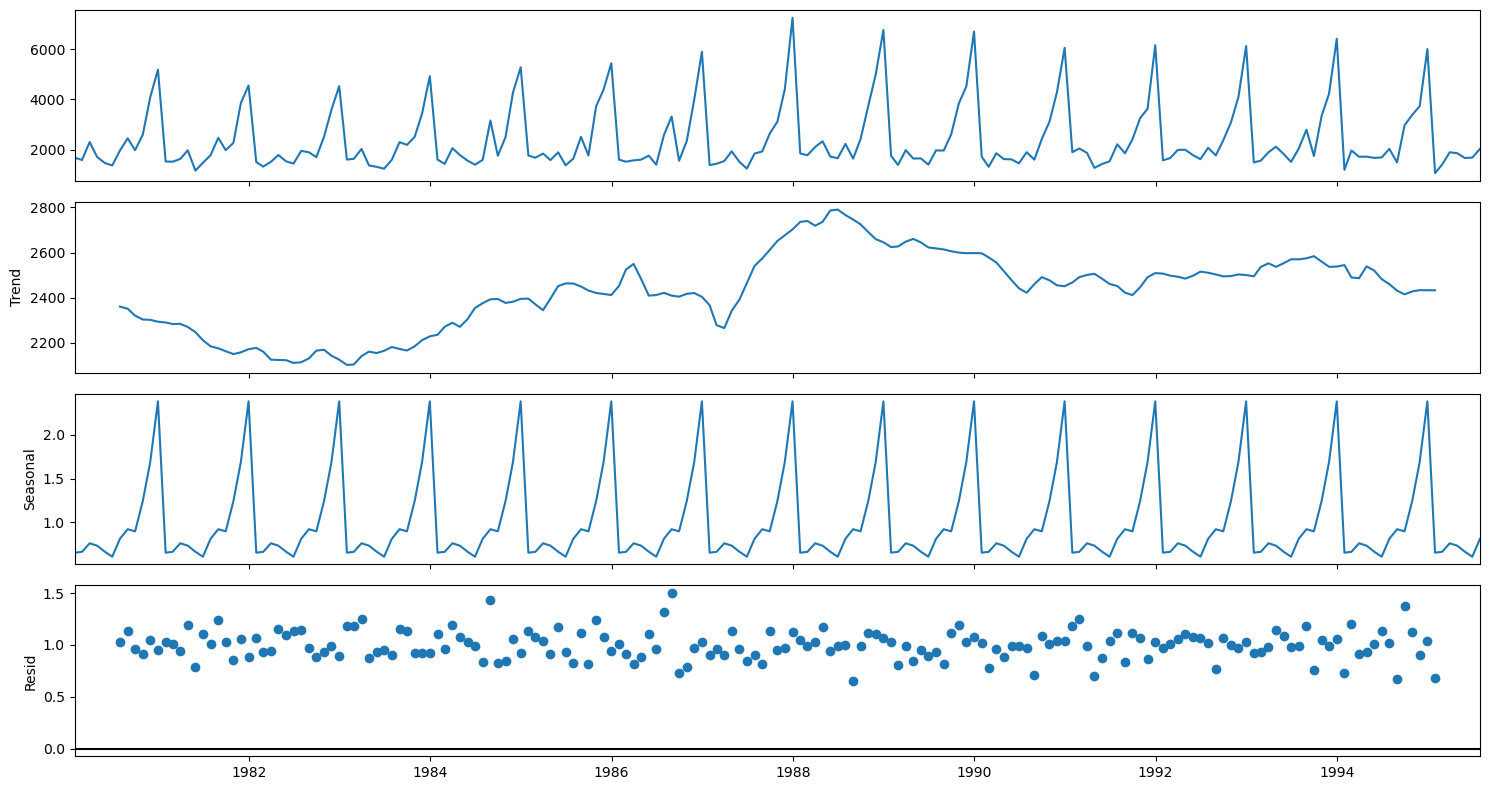


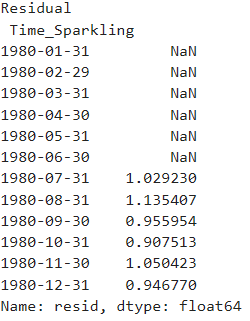
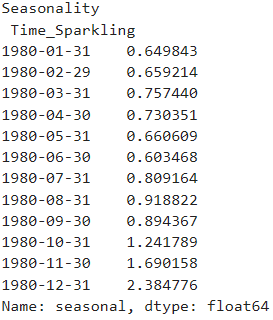
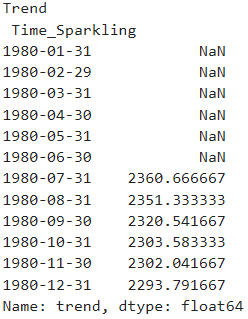
**Insights:**

* The trend shows a peak of sales between the year 1989 and 1990.
* After the year 1990, there has been a slight decline in sales volume.
* Residuals are spread out in additive model which represent seasonality.

### 2.5.2 Multiplicative

Figure 8 - Multiplicative Decomposition of Sparkling Wine Data





**Insights:**

* The trend chart is similar to additive model showing a peak at 1989 and then a slight decline from the year 1990.
* There is seasonality in data shown in the multiplicative model as well.
* The residuals are more stable and in a straight line in multiplicative model.
* The data can be assumed as multiplicative based on the stability of residuals.

# DATA PRE-PROCESSING – SPARKLING WINE

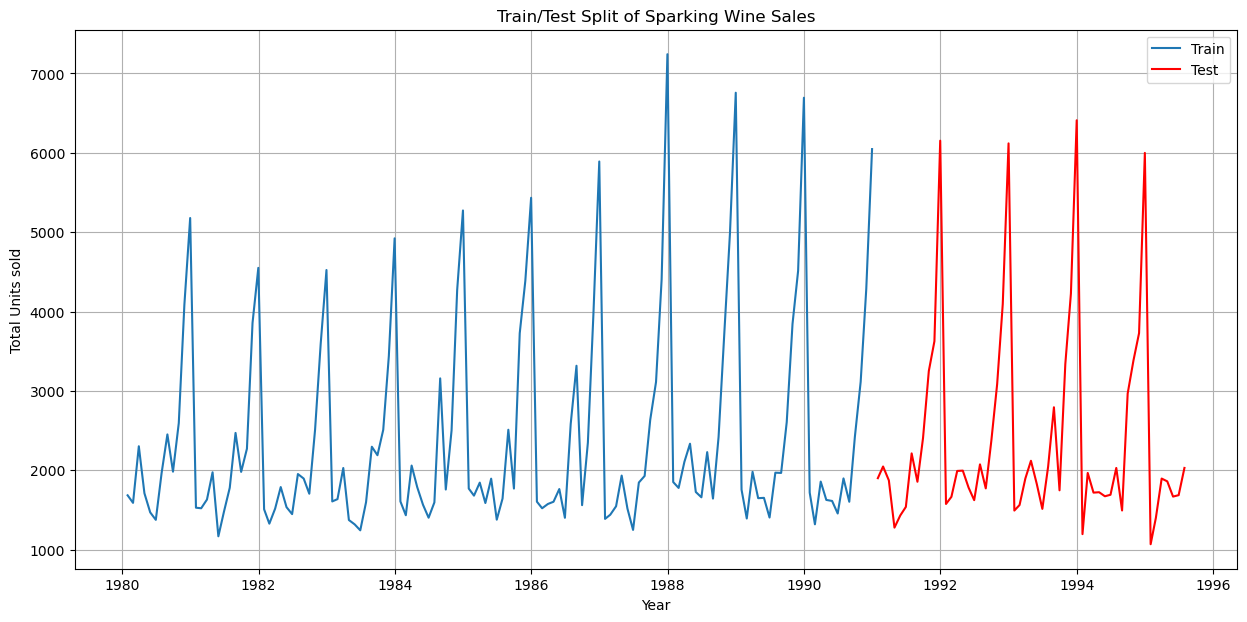
## 3.1 Missing Value Treatment

There are no missing values present in the sparking.csv dataset, therefore no treatment is required.

## 3.2 Data Preparation

The training and testing data are separated for the purpose of building models. All observations prior to the January 1991 will be considered as Training Data, all observations from the January 1991 onwards will be considered as Testing Data.

Figure 9 - Train/Test Split of Sparkling Wine Data



* There are a total of 132 observations present in the training data.
* There are 55 observations in the testing data.

The model will be built using the training data and tested on the test data to check for model efficiency.

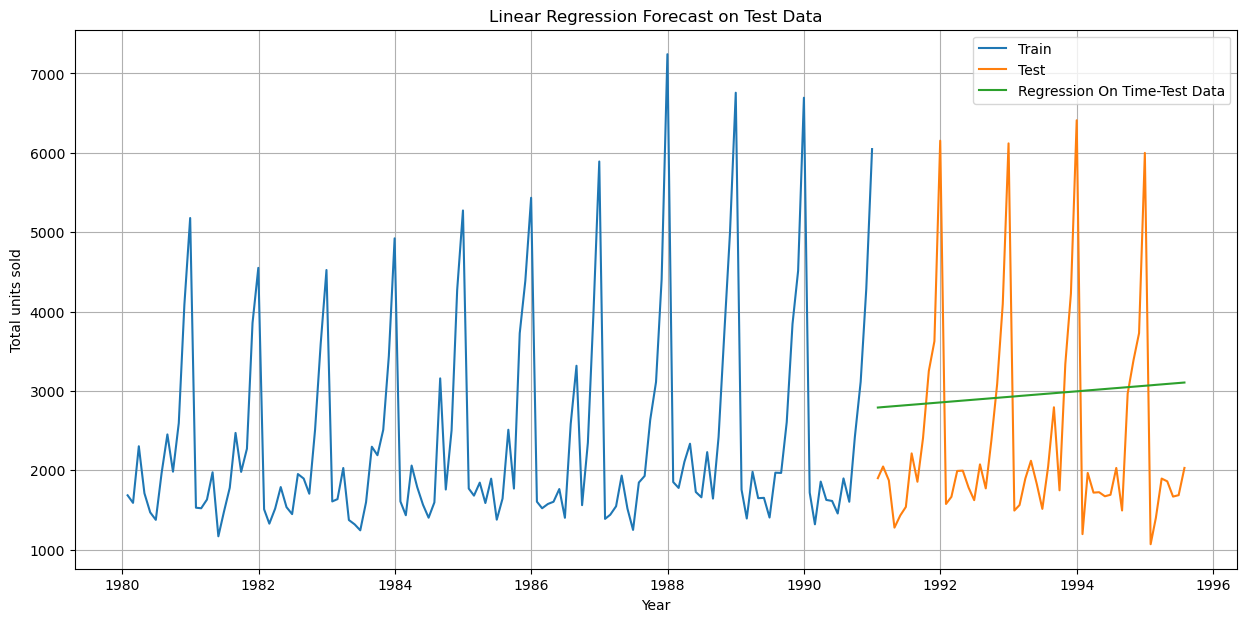
# MODEL BUILDING – ORIGINAL DATA - SPARKLING WINE

## 4.1 Linear Regression Model

Linear regression is a method used to predict future values based on the relationship between past observations, the observations are collected in monthly intervals and the model will be built based on the analysis of trends.

In this dataset the column containing the sales data (“Sparkling”) will the regressed against the order of occurrence.

Figure 10 - Linear Regression Model on Test Data



**Insights:**

* The model built and tested on the test data shows a predicted upward trend.
* It does not show any seasonality in the test data.
* We can also observe the predicted values vary greatly from the actual test values.

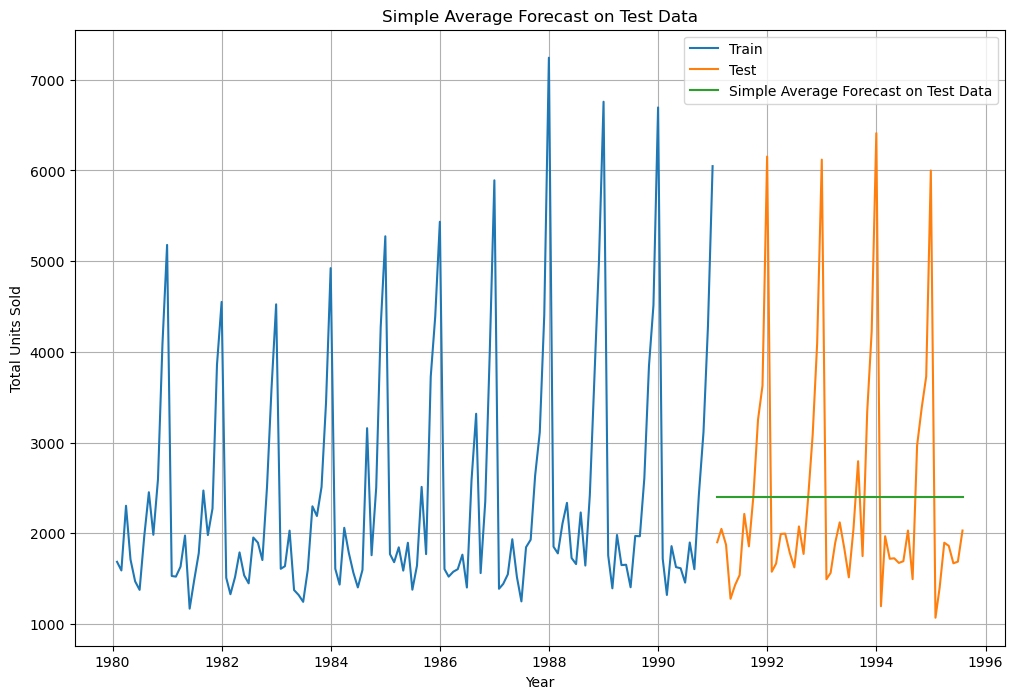
On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **1389.135**.

## 4.2 Simple Average Model

The simple average method in time series forecasting involves using the average of historical data points to predict future values. It assumes that future values will be similar to the overall average of past observations.

In this dataset, the forecast will be based on using the average of train data values.

Figure 11 - Simple Average Model on Test Data



**Insights:**

* The simple average model does not show any trend or seasonality.
* The predicted values vary significantly from the actual test values.

On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **1275.081**.

## 4.3 Moving Average Model

The moving average model in time series forecasting involves averaging a fixed number of past observations to predict future values. The model smooths the data to identify trends by calculating the average of a sliding window of data points.

For this model, the rolling means (or moving averages) will be calculated for different trailing points (2,4,6,9). The best interval will be determined by the maximum accuracy (or the minimum error).

Figure 12 – Sample Values of Moving Averages

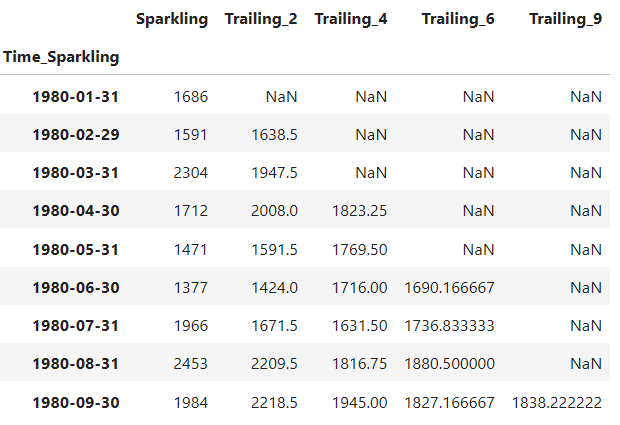


Figure 13 - Moving Average Model on Entire Dataset

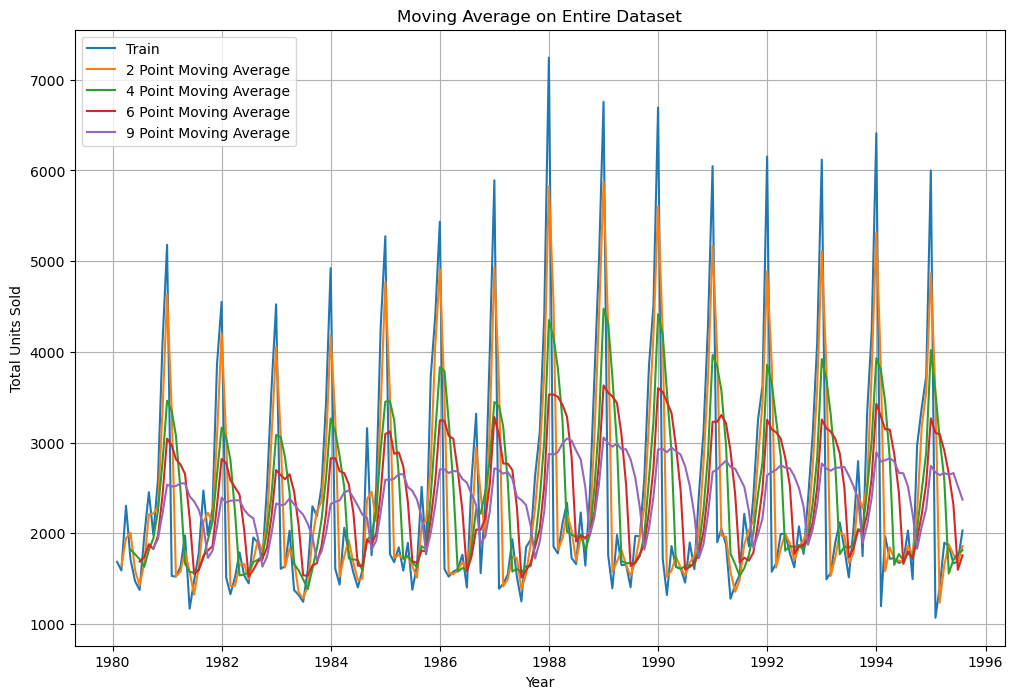
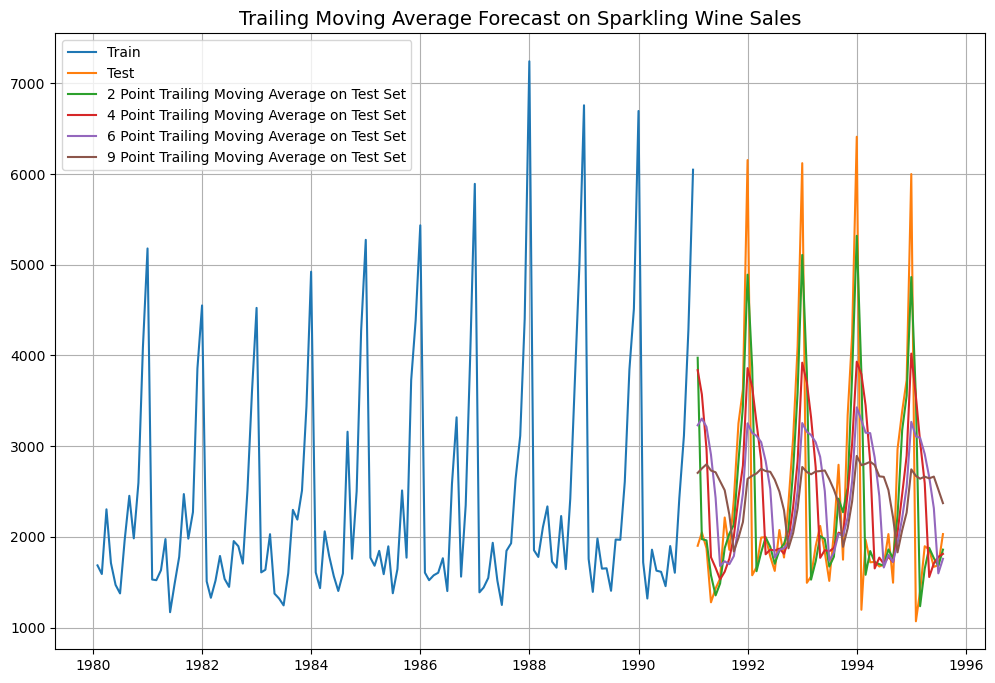


Figure 14 - Moving Average Model on Test Data



**Insights:**

* All trailing point show trend and seasonality unlike the previous models.
* It can be observed that the 2-point trailing moving average model is the most accurate to predict the test data.
* There is better accuracy with lower trailing point average.

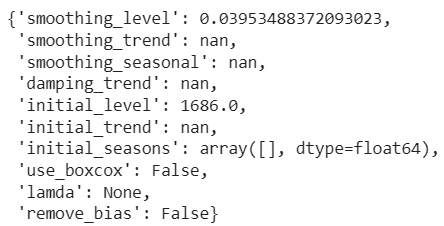
On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score are.

* 2 point moving average model – **813.401**
* 4 point moving average model – **1156.590**
* 6 point moving average model – **1283.927**
* 9 point moving average model – **1346.278**

## 4.4 Single Exponential Smoothing Model

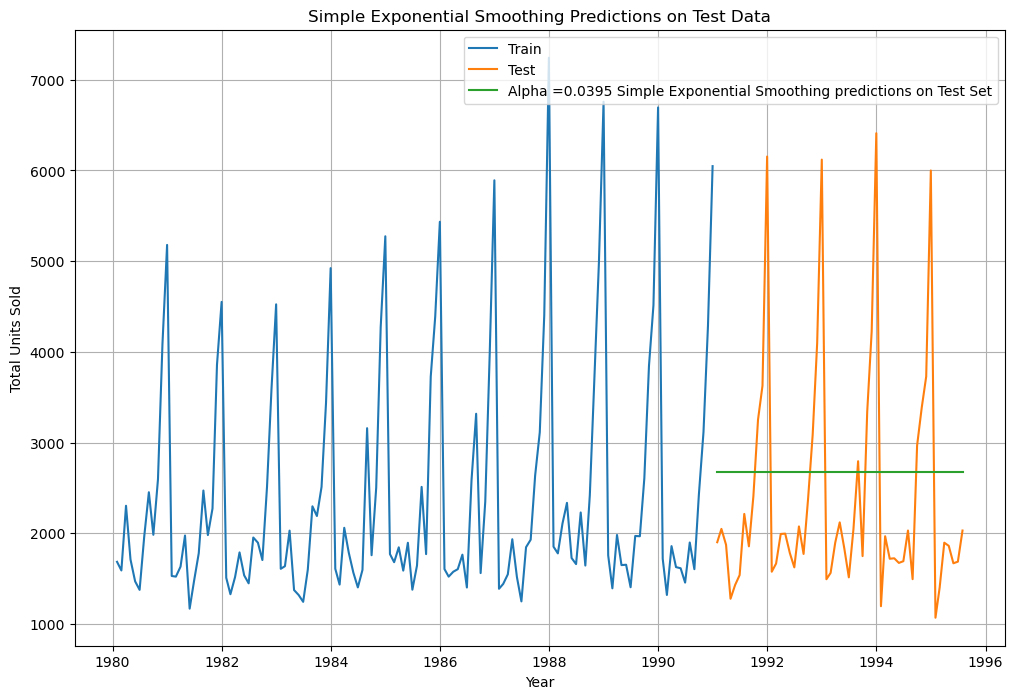
The single exponential smoothing method in time series forecasting is a technique that smooths past observations to make predictions for future values. It gives more weight to recent observations while gradually reducing the weight for older data. The method uses a smoothing constant (α) to control the level of smoothing, where, α is a value between 0 and 1.

Figure 15 - Single Exponential Smoothing Model Parameters



The simple exponential smoothing is built with optimized parameters, an α value of 0.039 can be observed.

Figure 16 - Single Exponential Smoothing Model on Test Data



**Insights:**

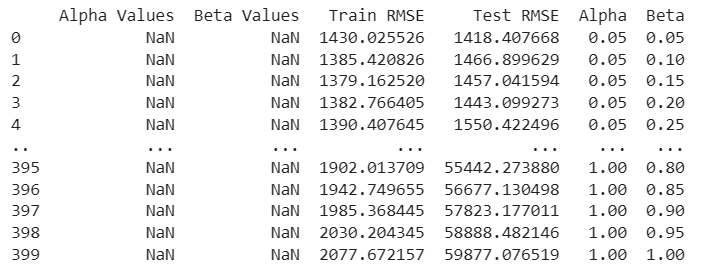
* The predicted values show no trend and seasonality on the test data.
* There is significant variation on the predicted test values and actual test values.

On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **1304.927**.

## 4.5 Double Exponential Smoothing Model

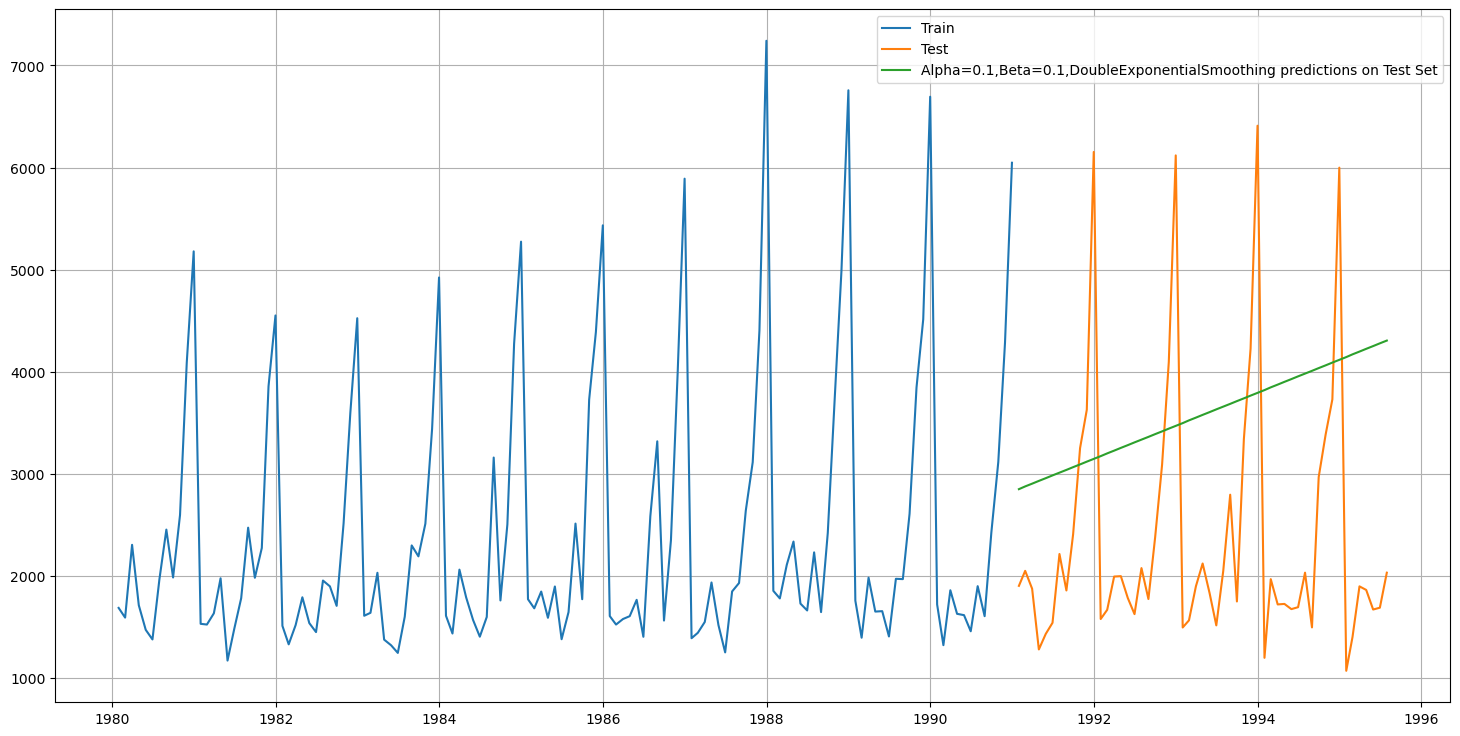
The double exponential smoothing method in time series forecasting extends single exponential smoothing by accounting for trends in the data. It uses two smoothing constants: one for the level (α) and one for the trend (β). This method adjusts both the level and the trend of the series over time, making it suitable for data with a linear trend. The model predicts future values by combining both smoothed levels and trends.

Figure 17 - Double Exponential Smoothing Model Parameters



The double exponential smoothing is built with optimized parameters, an α value of 0.05 and a β value of 0.05 can be observed.

Figure 18 - Double Exponential Smoothing Model on Test Data



**Insights:**

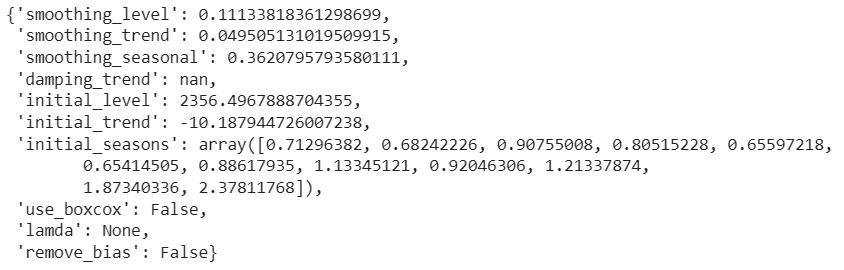
* The predicted values show an increase indicated a trend however it does not account for seasonality.
* The model is only suitable to showcase trend and no seasonality. This model has a higher RMSE score compared to single exponential smoothing model.

On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **1418.407**.

## 4.6 Triple Exponential Smoothing Model

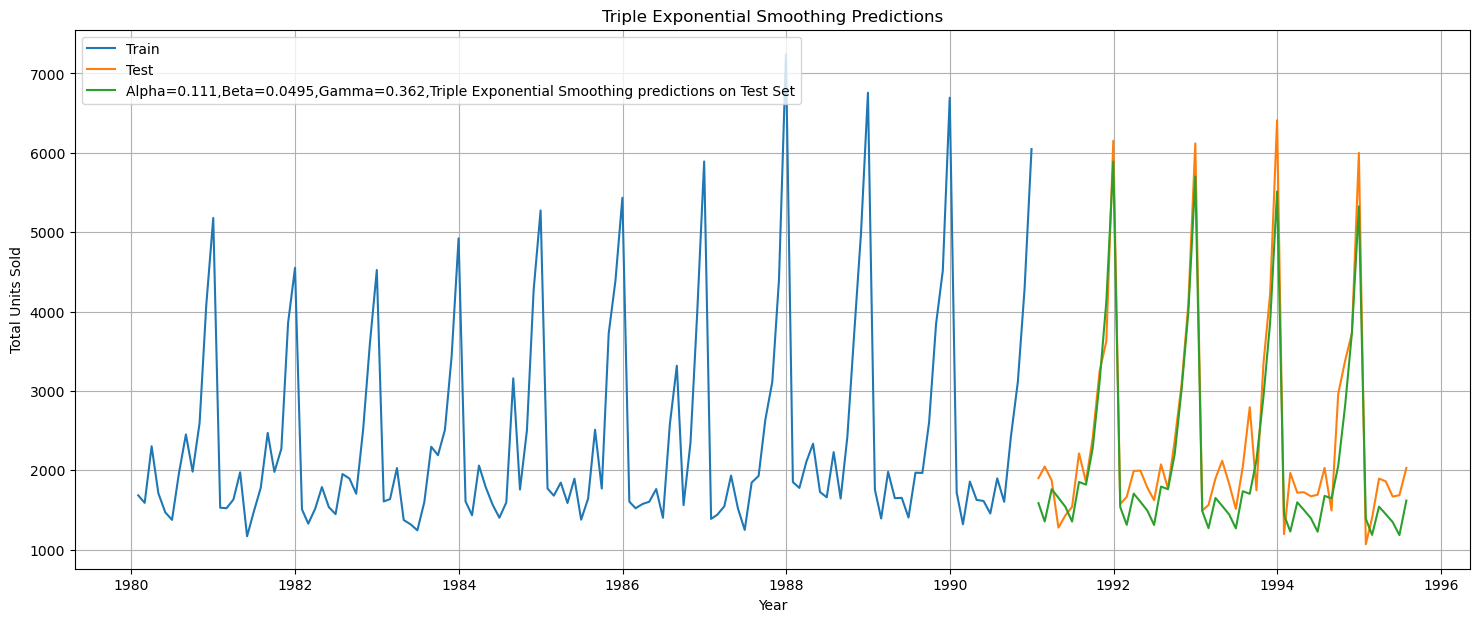
The triple exponential smoothing model (Holt-Winters) is used for time series forecasting with data that includes level, trend, and seasonality. It uses three smoothing parameters: α for the level, β for the trend, and γ for the seasonal component. This model adjusts for seasonal variations in addition to capturing the overall trend in the data.

Figure 19 - Triple Exponential Smoothing Model Parameters



The triple exponential smoothing is built with optimized parameters, an α value of 0.111 and a β value of 0.495 and a γ value of 0.362 can be observed.

Figure 20 - Triple Exponential Smoothing Model on Test Data



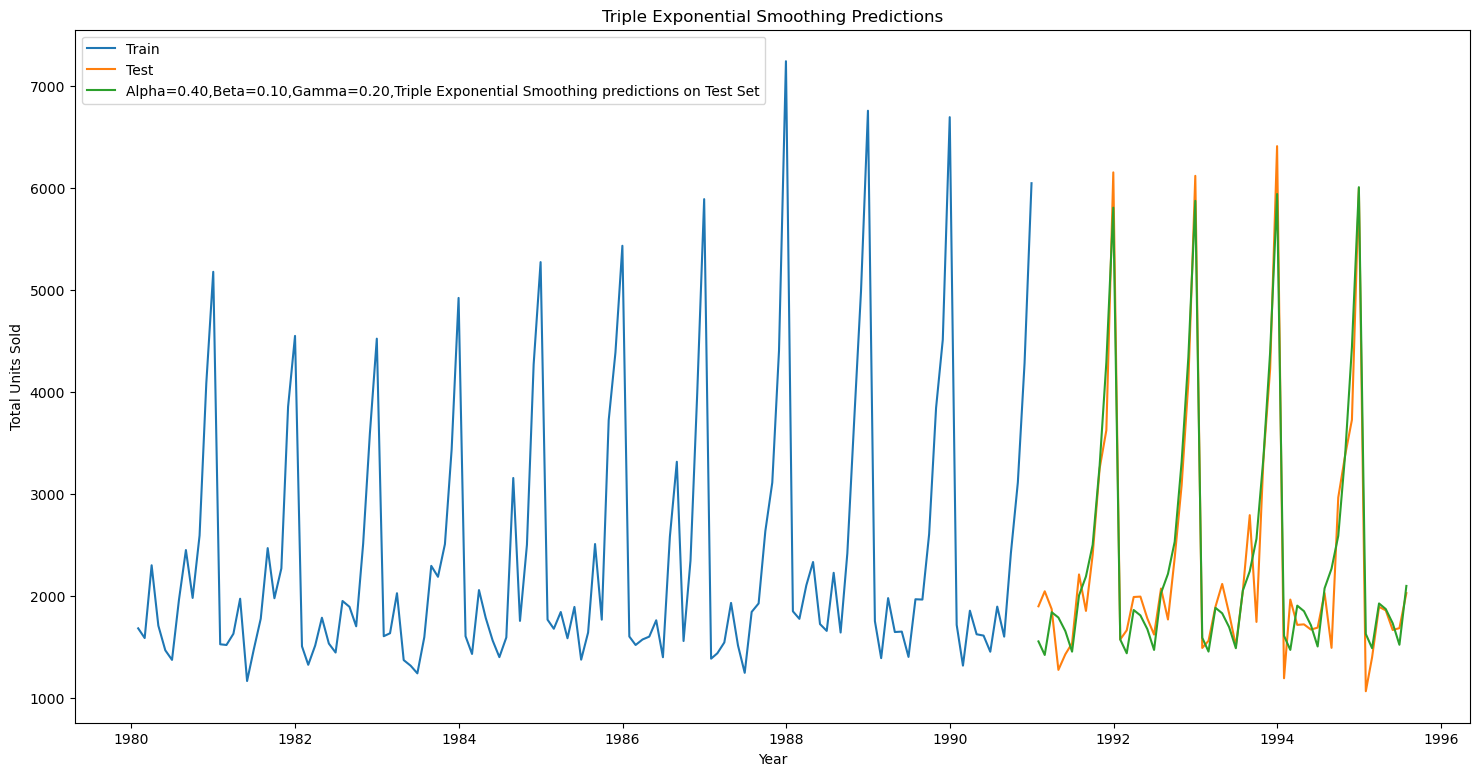
The recent observation has more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.

Figure 21 - Different Parameters for Triple Exponential Smoothing Model



an α value of 0.40 and a β value of 0.10 and a γ value of 0.20

Figure 22 - Triple Exponential Model on Test Data



**Insights:**

* There is both seasonality and trend predicted on the test data.
* The predicted values match the actual test data values to a good degree which makes it the most suitable model for now.
* This model has the lowest RMSE score so far compared to the all the previous built models.

On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **317.434**.

# CHECK FOR STATIONARITY – SPARKLING WINE

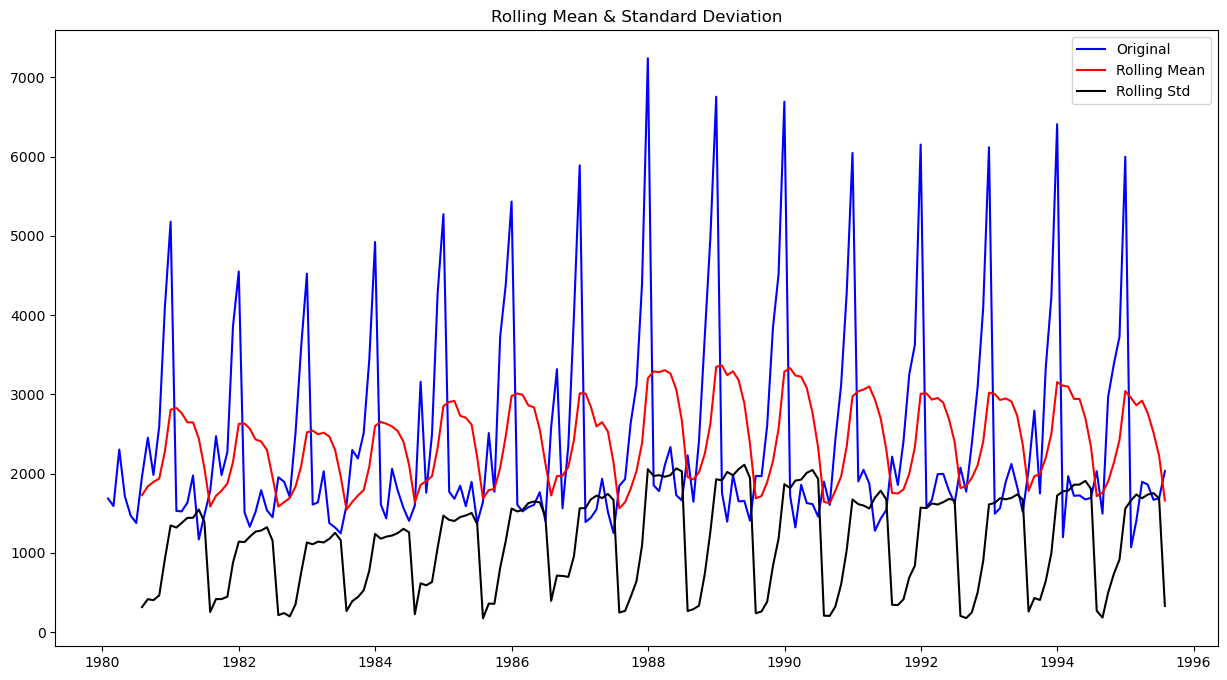
The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

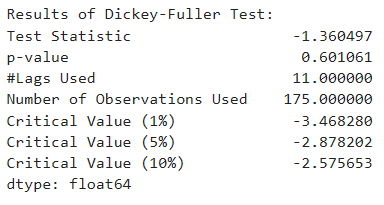
Frame the hypothesis:

H0: The Time Series has a unit root and is thus non-stationary.

H1: The Time Series does not have a unit root and is thus stationary.

Figure 23 - Dickey-Fuller Test on Normal Data

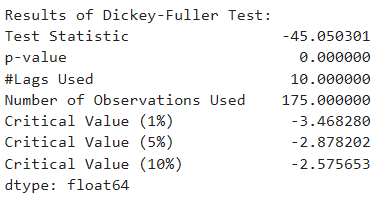




The p-value observed post the dickey-fuller test is 0.6 which is more than the α value of 0.05, so the data is non stationary as we fail to reject the null hypthesis.

Figure 24 - Dickey-Fuller Test on Differenced Data





To make the data stationary, a differencing approach can be used on the dataset with a diff value of 1 and all null values which are generated will be dropped.

We can observe a p-value of 0 which is less than α value of 0.05, therefore we reject the null hypothesis and the data is stationary.

# MODEL BUILDING – STATIONARY – SPARKLING WINE

## 6.1 ACF & PACF Plot

An ACF (Auto Correlation Function) plot is a graphical tool used in time series analysis to measure and visualize the correlation between a time series and its lagged versions. It shows how the values of the series are related to previous time steps (lags). The x-axis represents the lag, and the y-axis represents the correlation at each lag. Peaks in the ACF plot indicate significant autocorrelations at those lags. ACF plots are useful for identifying seasonality, trends, and determining the appropriate model for time series forecasting (e.g., ARIMA).

A PACF (Partial Auto Correlation Function) plot is used in time series analysis to measure the correlation between a time series and its lagged values, after removing the influence of shorter lags. Unlike the ACF plot, which shows total correlations, the PACF isolates the direct relationship between the series and a specific lag. The x-axis represents the lag, and the y-axis represents the partial correlation at each lag. Significant peaks in the PACF help identify the order of autoregressive (AR) components in models like ARIMA. It is especially useful for determining the appropriate number of lags for an AR model.

Figure 25 - ACF Plot

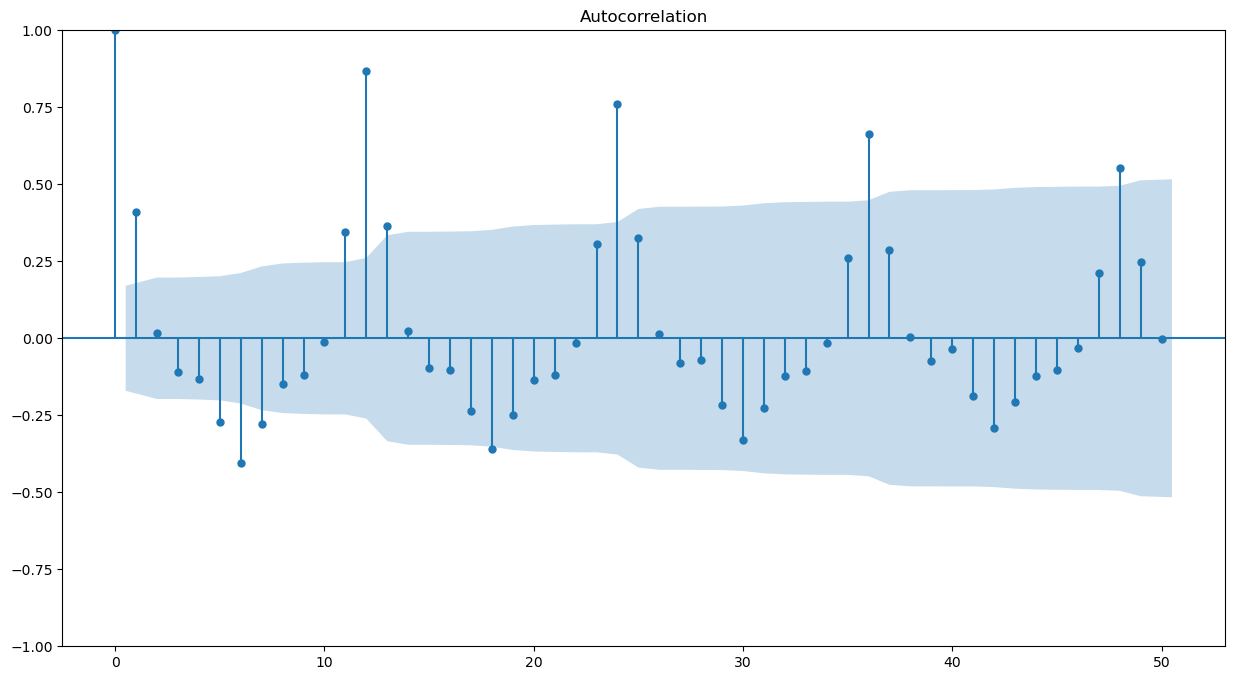
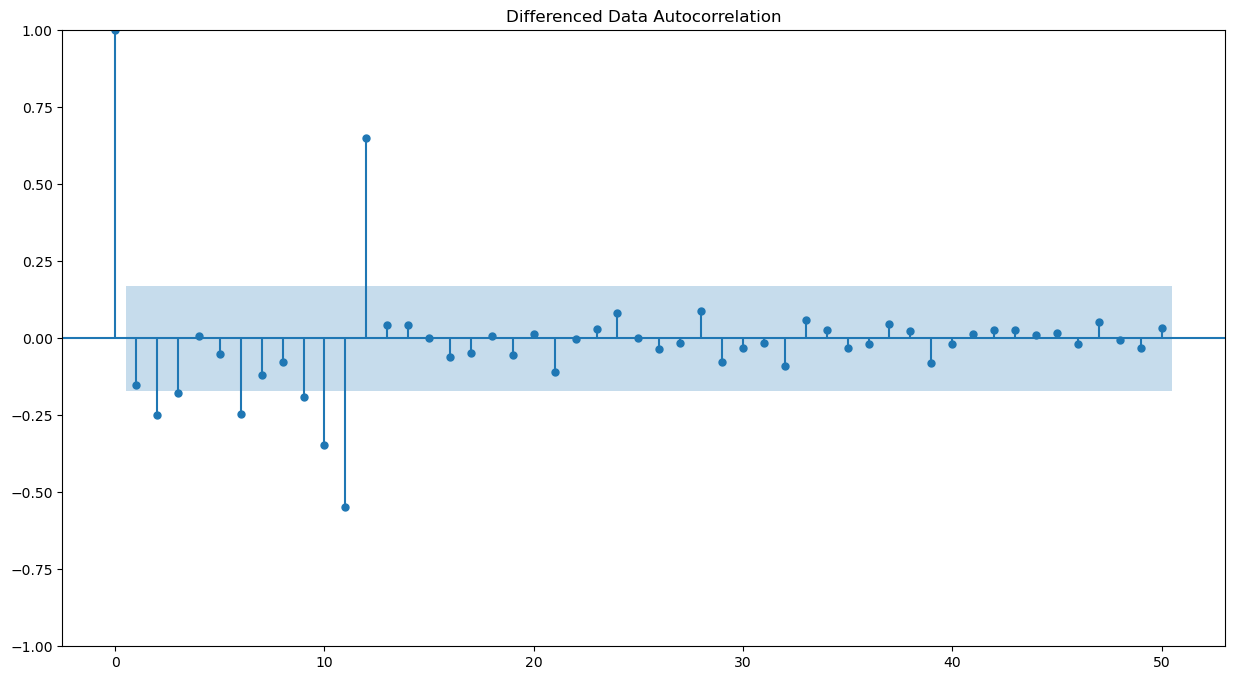


Figure 26 - PACF Plot



The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag after which the PACF plot cuts-off below the confidence interval.

The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off below the confidence interval.

It can be observed that after lag 1, we have few significant lags and hence we can take a value of p=2 and q=1 respectively.

## 6.2 Auto ARIMA Model

The Auto ARIMA model is an automated version of the ARIMA used for time series forecasting. It automatically selects the best combination of parameters (p, d, q) by evaluating different values of autoregressive, differencing, and moving average components. The model uses techniques like AIC or BIC to determine the optimal model that minimizes prediction error. Auto ARIMA simplifies the process of fitting ARIMA models by automating parameter selection. It is particularly useful for users without deep expertise in model tuning.

For the selection criteria of p,d,q the below ARIMA model is built using the automated model parameters with lowest Akaike Information Criteria.

Figure 27 - Parameters Combinations & AIC values – ARIMA Model

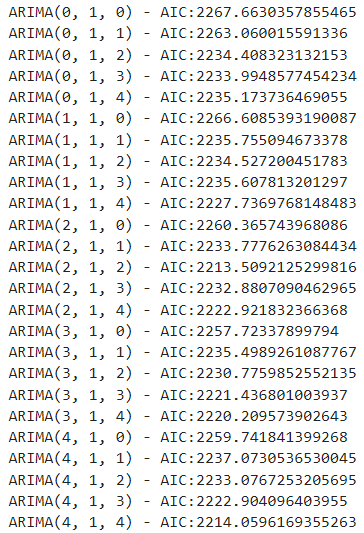
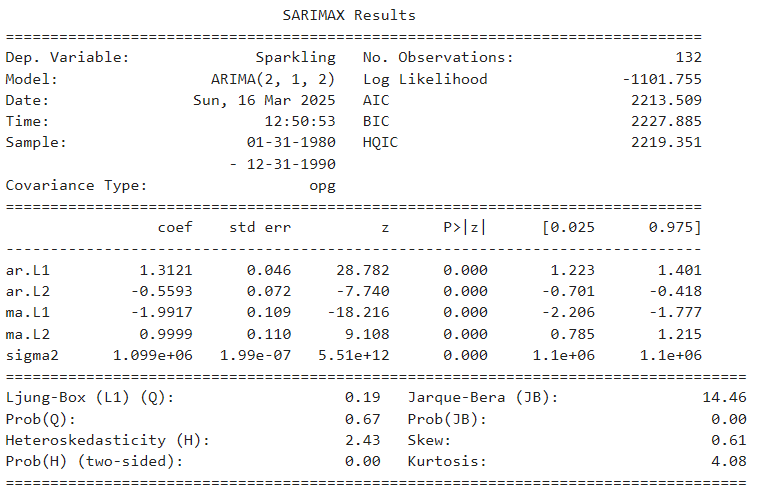


Figure 28 - Sorted Parameter and AIC Value - ARIMA Model



We can observe the AIC is lowest for the combination (2, 1, 2). The model will be built with this parameter.

Figure 29 - Auto ARIMA Model



On evaluation of the model, we can observe a root mean square error (RMSE) score of **1299.979**.

## 6.3 Manual ARIMA Model

The manual ARIMA model involves manually selecting the parameters for autoregressive, differencing, and moving average components based on analysis of the time series data. The process typically includes examining the ACF and PACF plots to determine appropriate values for p and q. The parameter d is chosen based on the need for differencing to make the series stationary. Manual ARIMA requires a deeper understanding of the data and model diagnostics. It can be time-consuming but offers more control over the model specification.

Based on the previous ACF and PACF plot, we can observe a value of p=2 and q=2 and build a model based on those parameters.

Figure 30 - Manual ARIMA Model

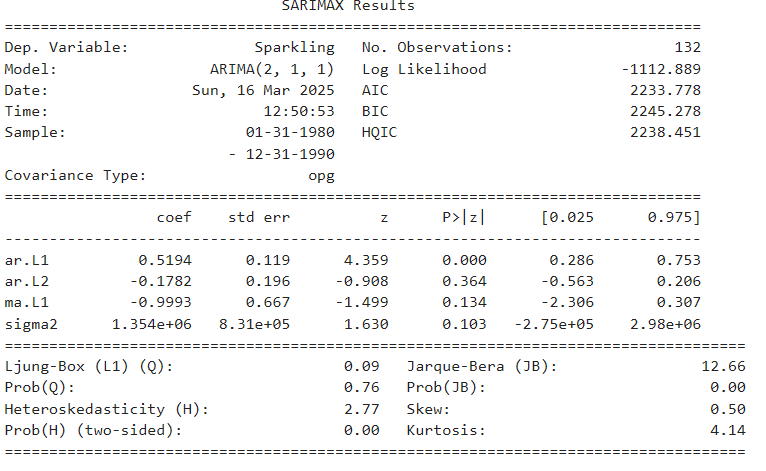
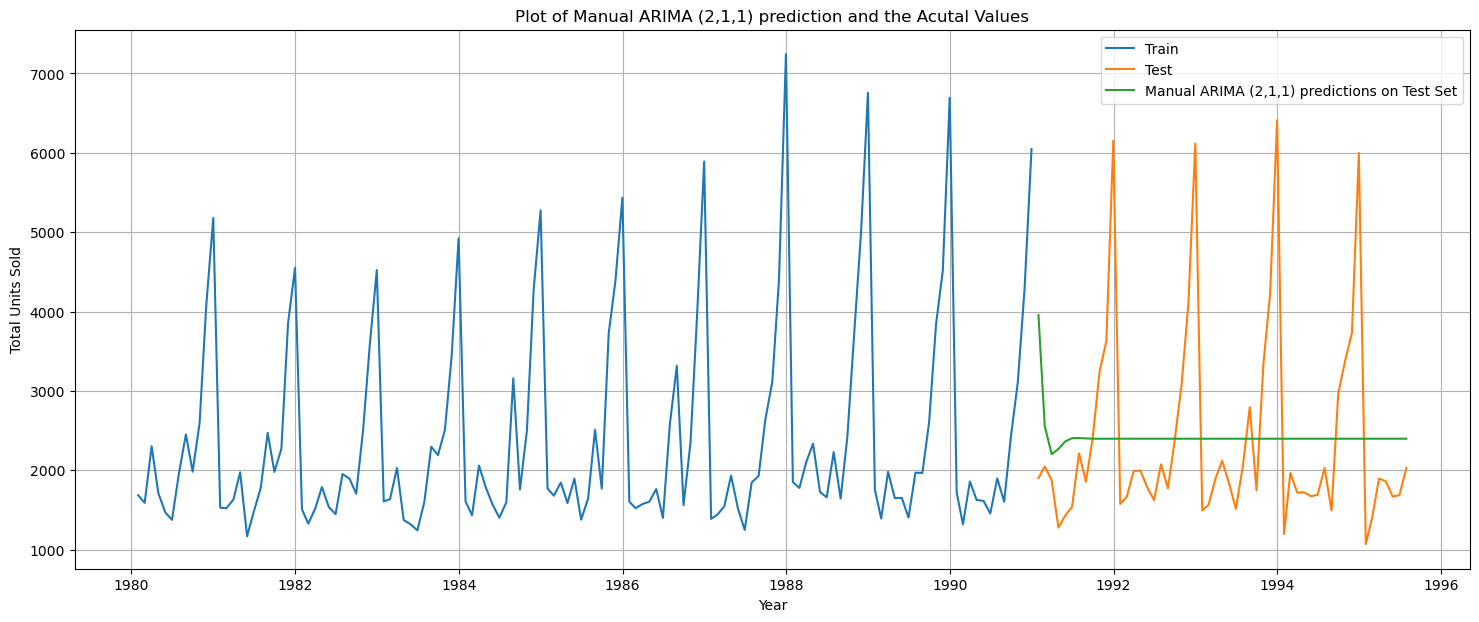


Figure 31 - Manual ARIMA Model on Test Data



**Insights:**

* There is no trend or seasonality shown in the predicted values besides showing an initial decline.
* This model predicted values vary significantly from the actual test values

On evaluation of the model, we can observe a root mean square error (RMSE) score of **1300.721**.

## 6.4 Auto SARIMA Model

The Auto SARIMA is an extension of the Auto ARIMA model that automatically handles seasonal components in time series forecasting. It selects optimal parameters for parts of the model. The model accounts for seasonality by incorporating seasonal differencing and seasonal AR and MA terms. Auto SARIMA automates the process of selecting the best model by evaluating different seasonal configurations. It is useful for forecasting time series data with seasonal patterns.

the optimum AIC which for the combination (3,1,2) (3,0,1,12) is considered due to errors with combinations lower that this. The model will be built with these parameters.

Figure 32 - Auto SARIMA Model

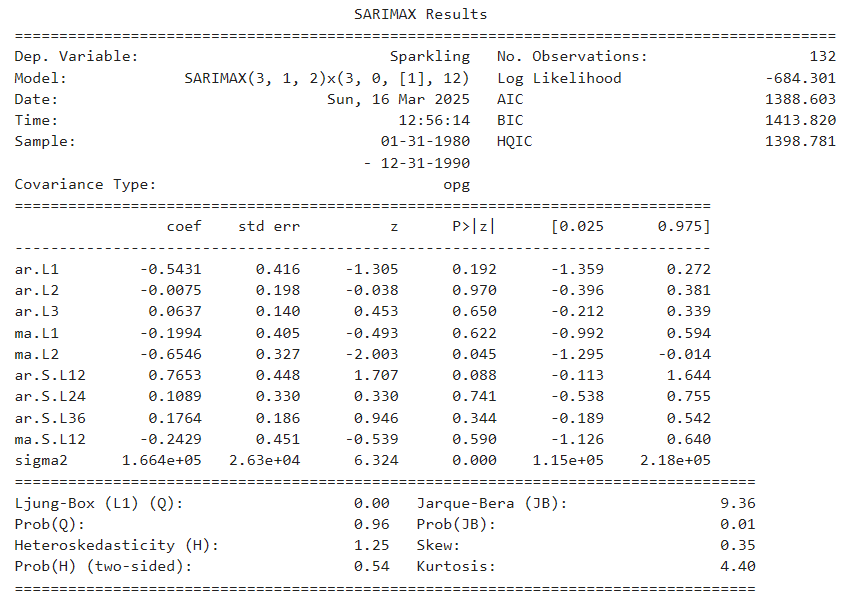
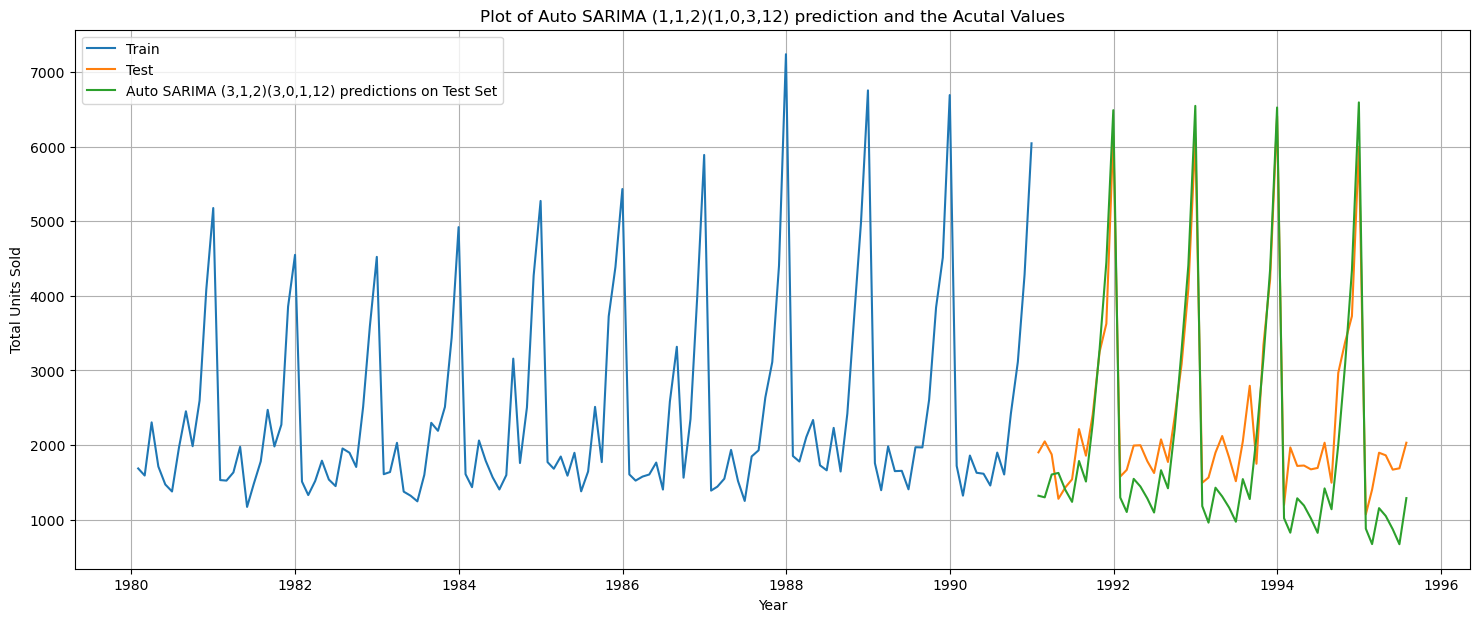


Figure 33 - Auto SARIMA Model on Test Data



**Insights:**

* The model predicts both trend and seasonality on the test data.
* The predicted values are close to the actual test data, making this a viable model for future predictions.

On evaluation of the model, we can observe a root mean square error (RMSE) score of **579.874**.

## 6.5 Manual SARIMA Model

The manual SARIMA model involves manually selecting the parameters for both the non-seasonal and seasonal components of the SARIMA model. This includes choosing values for the autoregressive, differencing, and moving average terms for both non-seasonal and seasonal parts. The seasonal period is also determined based on the observed seasonal patterns in the data. The process typically involves analyzing ACF and PACF plots and checking for stationarity and seasonality.

The SARIMA model will be built with the parameters of are p=4, P=0, q=2 and Q=1 with this combination (4,1,2) (0,1,1,12)

Figure 34 - Manual SARIMA Model

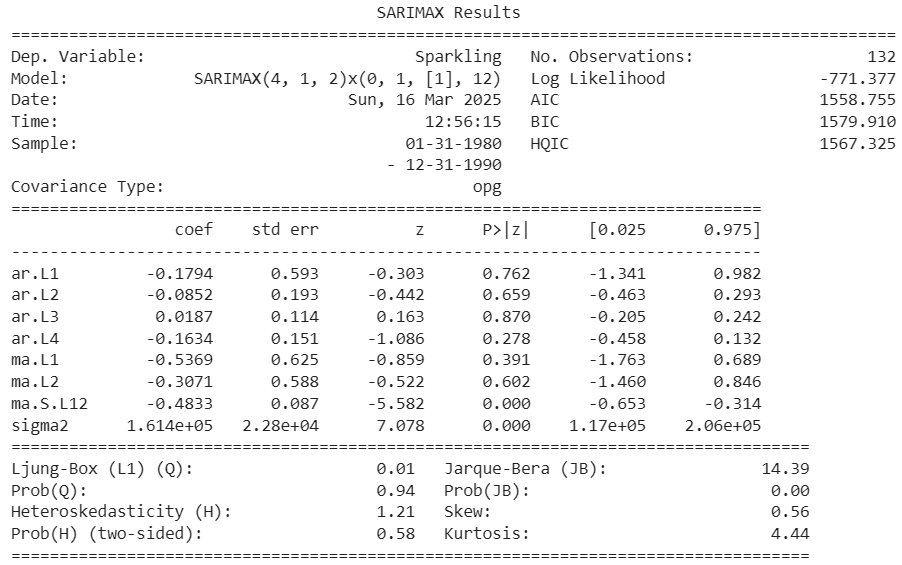
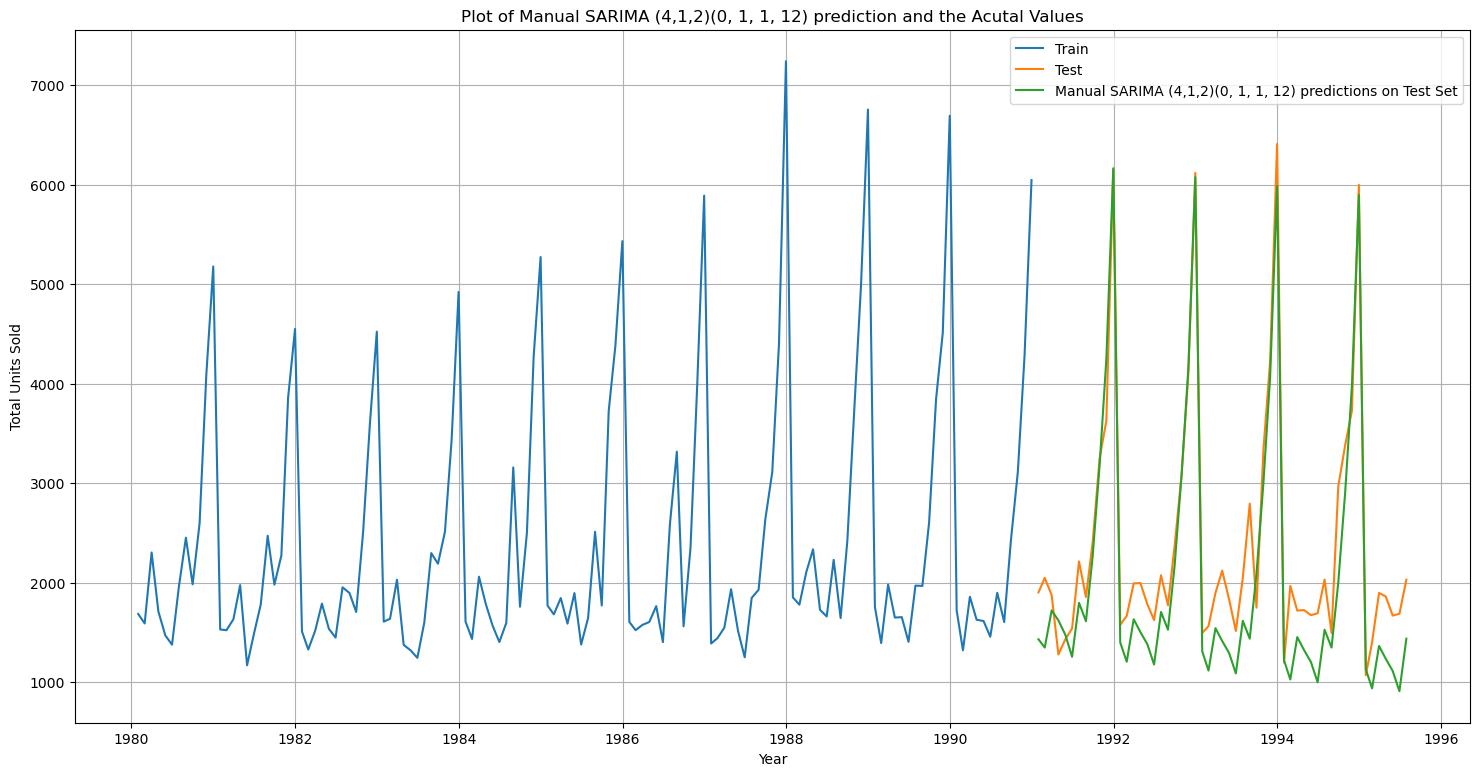


Figure 35 - Manual SARIMA Model on Test Data



**Insights:**

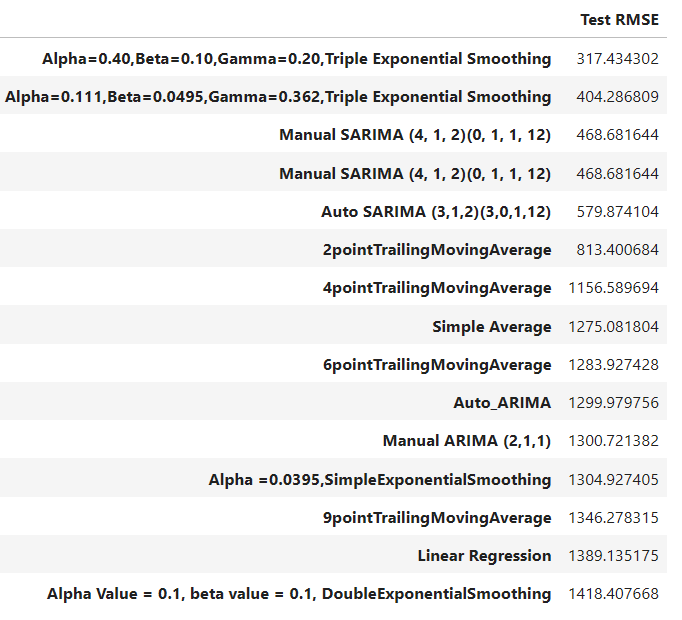
* The model shows both trend and seasonality on the predicted test values.
* The predicted values are close to the actual test data values.
* The model is suitable to use to predict the future values.

On evaluation of the model, we can observe a root mean square error (RMSE) score of **468.681**. The RMSE score of Manual SARIMA is lesser than the RMSE error of the Auto SARIMA Model.

# COMPARISON OF MODELS – SPARKLING WINE

All the models built so far will be compared and the most suitable model will be chosen and used to predict the sale of sparkling wine in the upcoming months of the dataset.

Figure 36 - Comparion of All Models



Based on the figure, we can see that the model with the lowest RMSE score is the Triple exponential smoothing model with α=0.40 and β=0.10 and a γ=0.20. The manual SARIMA and auto SARIMA model also have low RMSE score which are also suitable to be chosen.

The Triple exponential smoothing model with α=0.40 and β=0.10 and a γ=0.20 will the best model and chosen to predict the values for the next months.

## 7.1 Rebuilding the Model

Figure 37 - TES Model - Predicted vs Actual Values

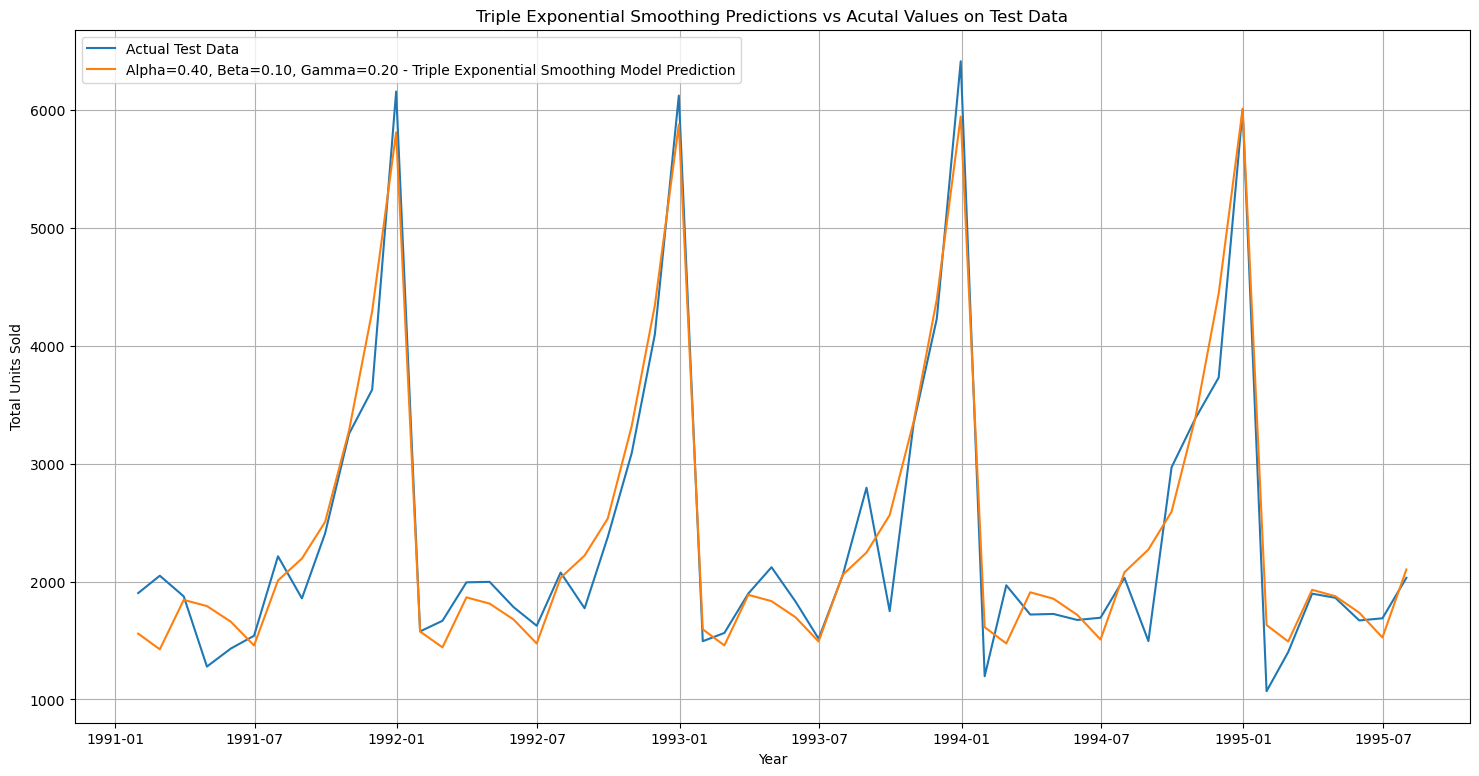
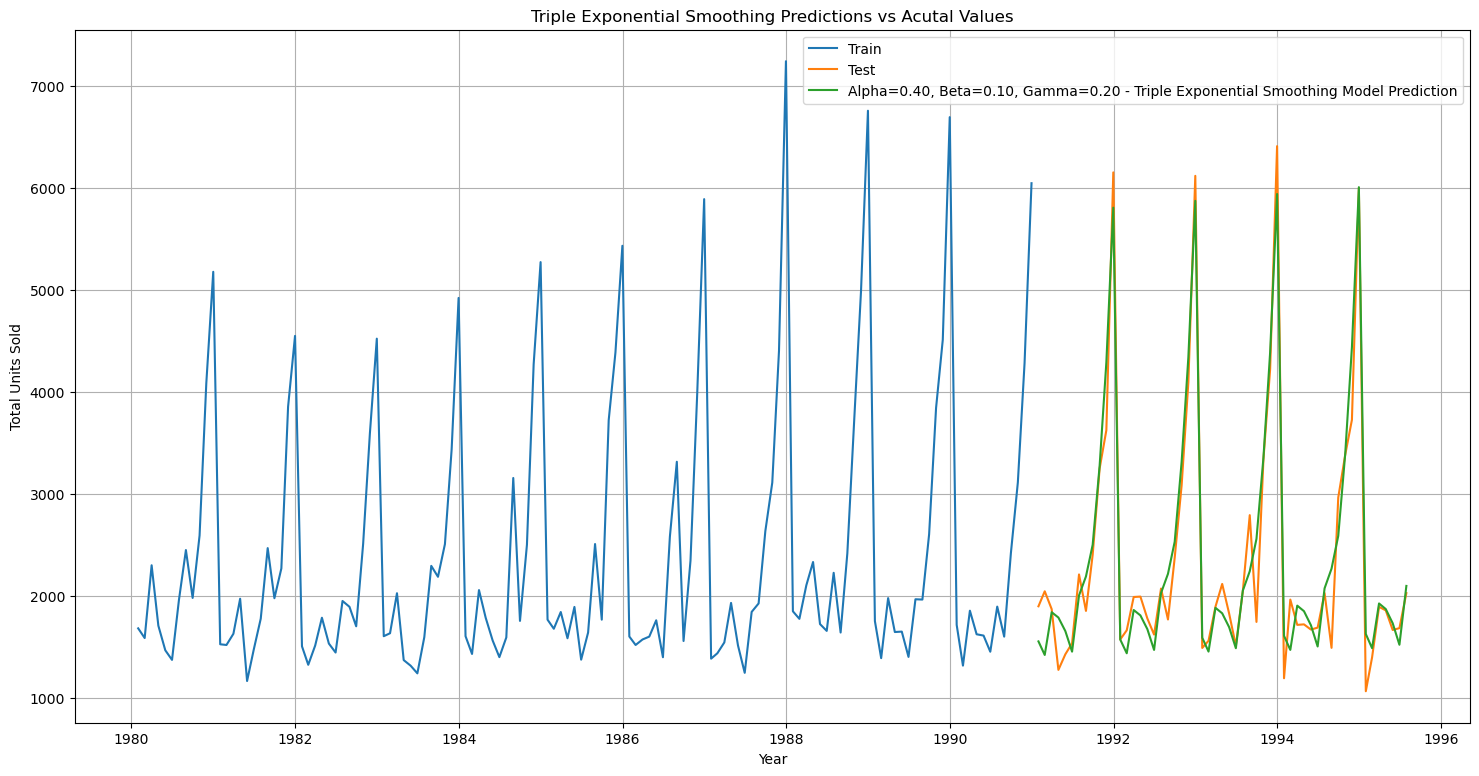


Figure 38 - TES Model on Test Data



The model is tested on test values which shows good performance as the data closely resemble the actual test data values.

The model is now rebuilt on the entire dataset and the next 12 months values will be predicted.

Figure 39 - TES Best Model

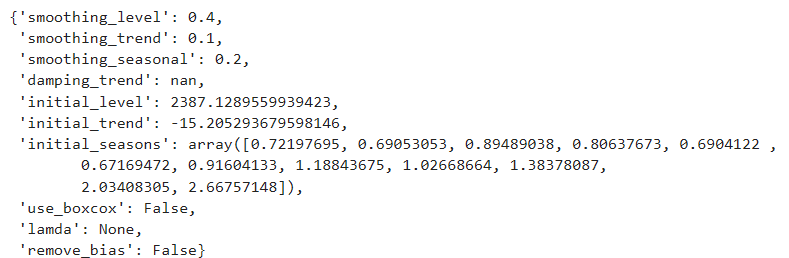


Figure 40 - TES Model - Forecast values

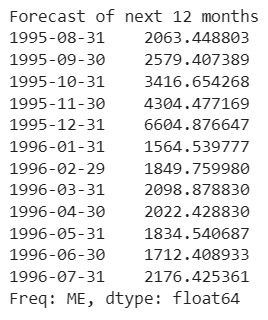
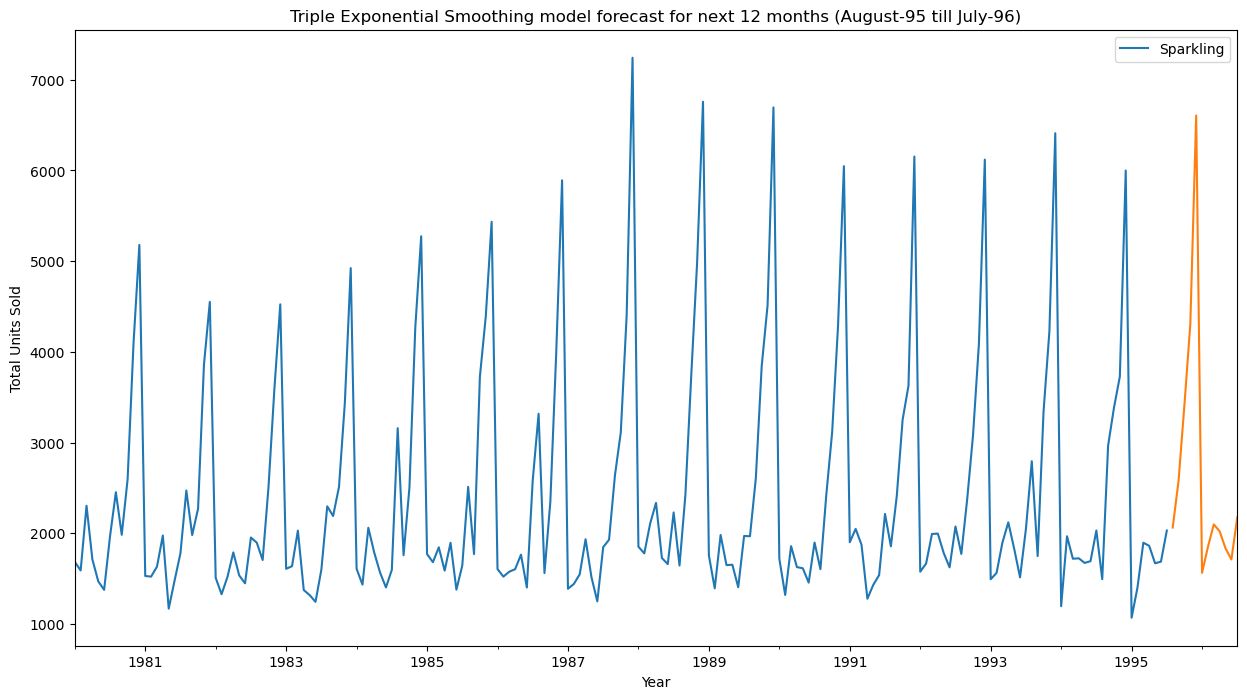


Figure 41 - Triple Exponential Smoothing Model – Forecast Trend



The predicted values for the next 12 months (Aug-1995 to Jul-1996) showcases trend and seasonality showing a peak sales of 6604 during December and a low amount of sales of 1564 in January 1996.

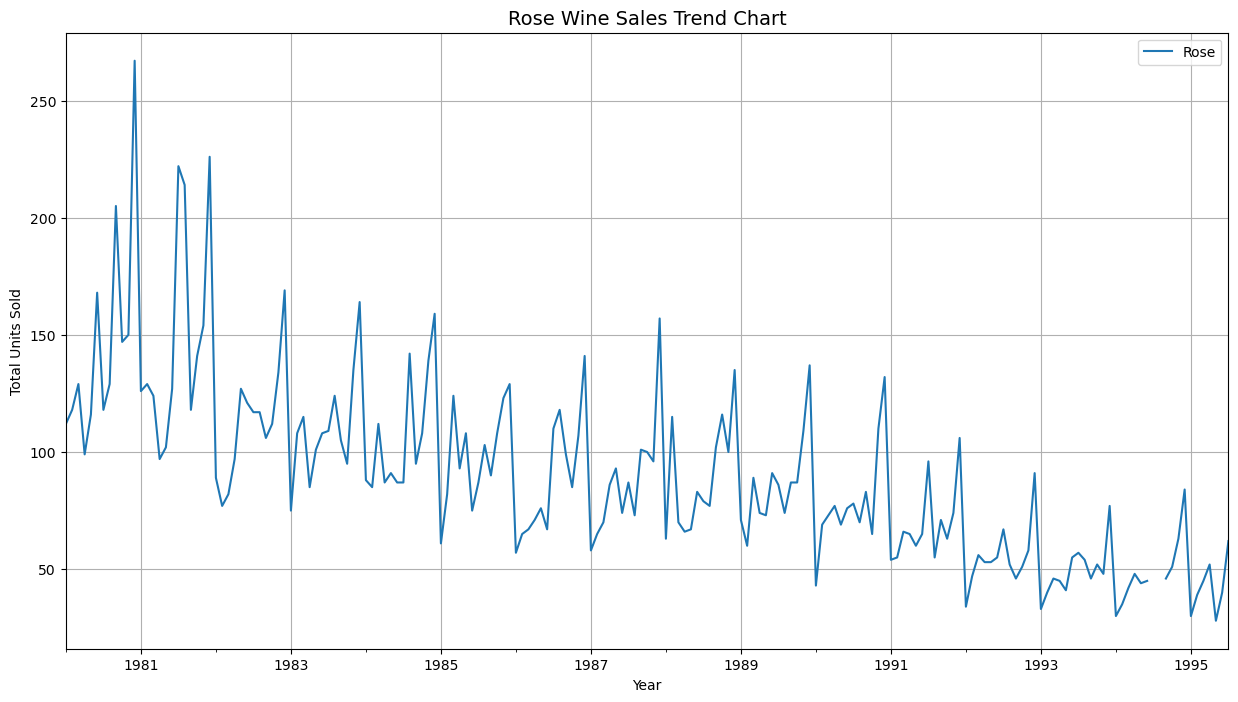
# ACTIONABLE INSIGHTS & RECOMMENDATIONS – SPARKLING WINE

* Sparkling wine sales show a stable trend with a clear seasonal pattern, peaking towards the end of the year.
* SARIMA and Triple Exponential Smoothing models provide the most accurate forecasts due to their ability to capture trend and seasonality.
* The forecast predicts a 10% increase in average monthly sales, with a significant rise in minimum sales and a slight decrease in maximum sales.
* Implement promotions and discounts during the holiday season (September to December) to boost sales.
* Offer free shipping or discounts on bulk purchases to attract more customers during festive periods.
* Provide gift incentives for customers to enhance the user experience and increase sales.
* Run targeted marketing campaigns for different customer demographics to expand the product’s reach.
* Introduce new product variations to address the sales dip from January to June.
* Conduct in-depth market research to identify factors affecting sales and improve forecast accuracy.
* Focus on bulk sales as many customers buy sparkling wine for events and gifting.
* Customize pricing strategies to align with seasonal demand fluctuations, especially during holidays.
* Invest in e-commerce campaigns and competitions to increase product visibility and attract more customers.

# EXPLORATORY DATA ANALYSIS – ROSE WINE

## 9.1 Time Series Trend Chart

Figure 42 - Rose Wine Sales Trend Chart



**Insights:**

* A peak of rose wines sales is observed around the year 1980-81.
* There is a steady decline overtime for rose wine sales.
* There is trend and seasonality in the data as seen by the varying ups and downs in the sales during the years.

## 9.2 Yearly Sales Box Plot

Figure 43 - Yearly Sales of Rose Wine - Box Plot

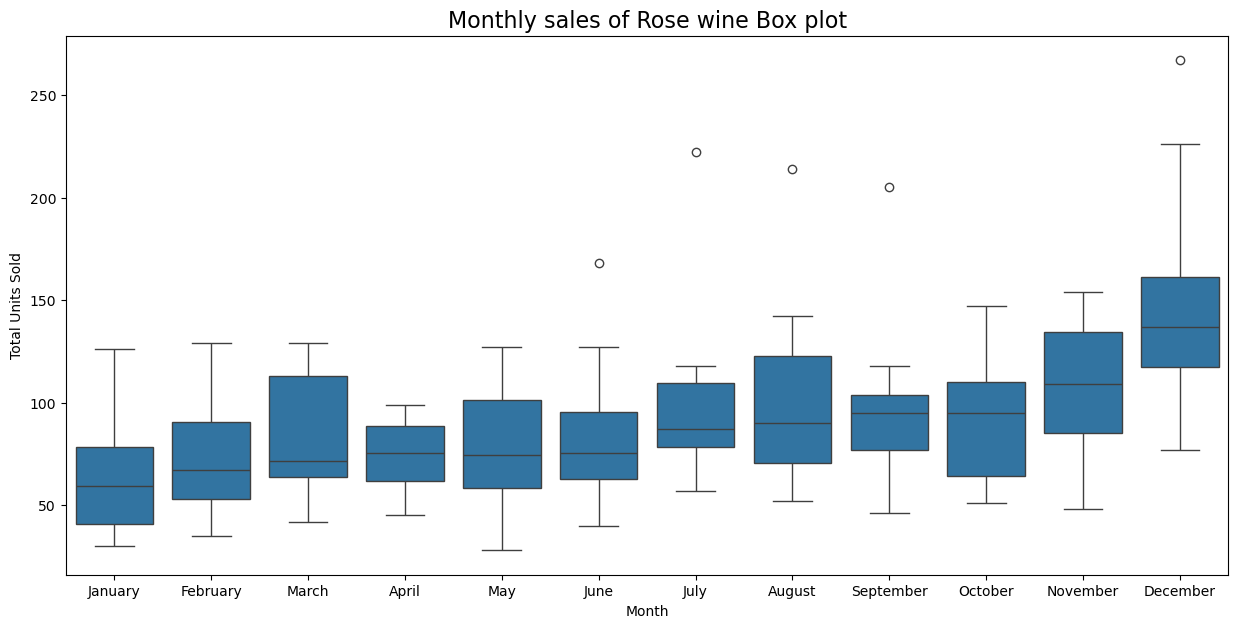


**Insights:**

* The last 3 years show the lowest median of unit sales rose wine with 1995 being the lowest.
* There are outliers present in few of the years.
* There is a decline of rose wine sales throughout the years.

## 9.3 Monthly Sales Box Plot

Figure 44 - Monthly Sales of Rose Wine - Box Plot

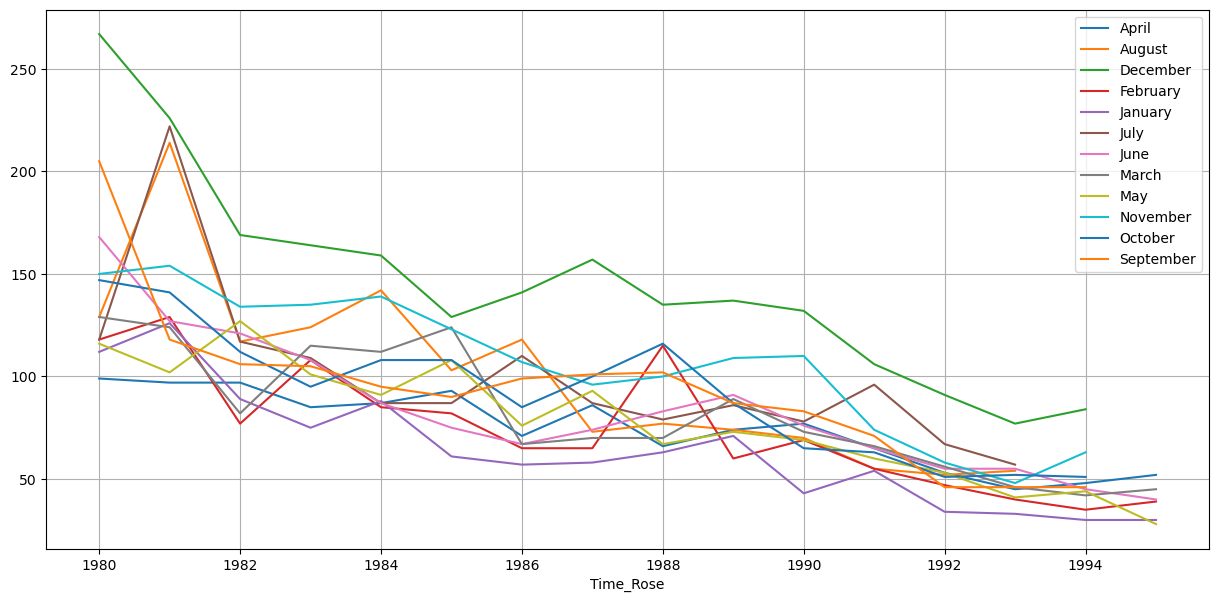


**Insights:**

* During a year, there is an increase in rose wine sales with a peak at the end of year in December.
* January observes the lowest amount of rose wine sales.
* There are outliers present in some of the months of the dataset.

## 9.4 Monthly Sales Across Years

Figure 45 - Monthly Sales Across Years - Trend Chart



**Insights:**

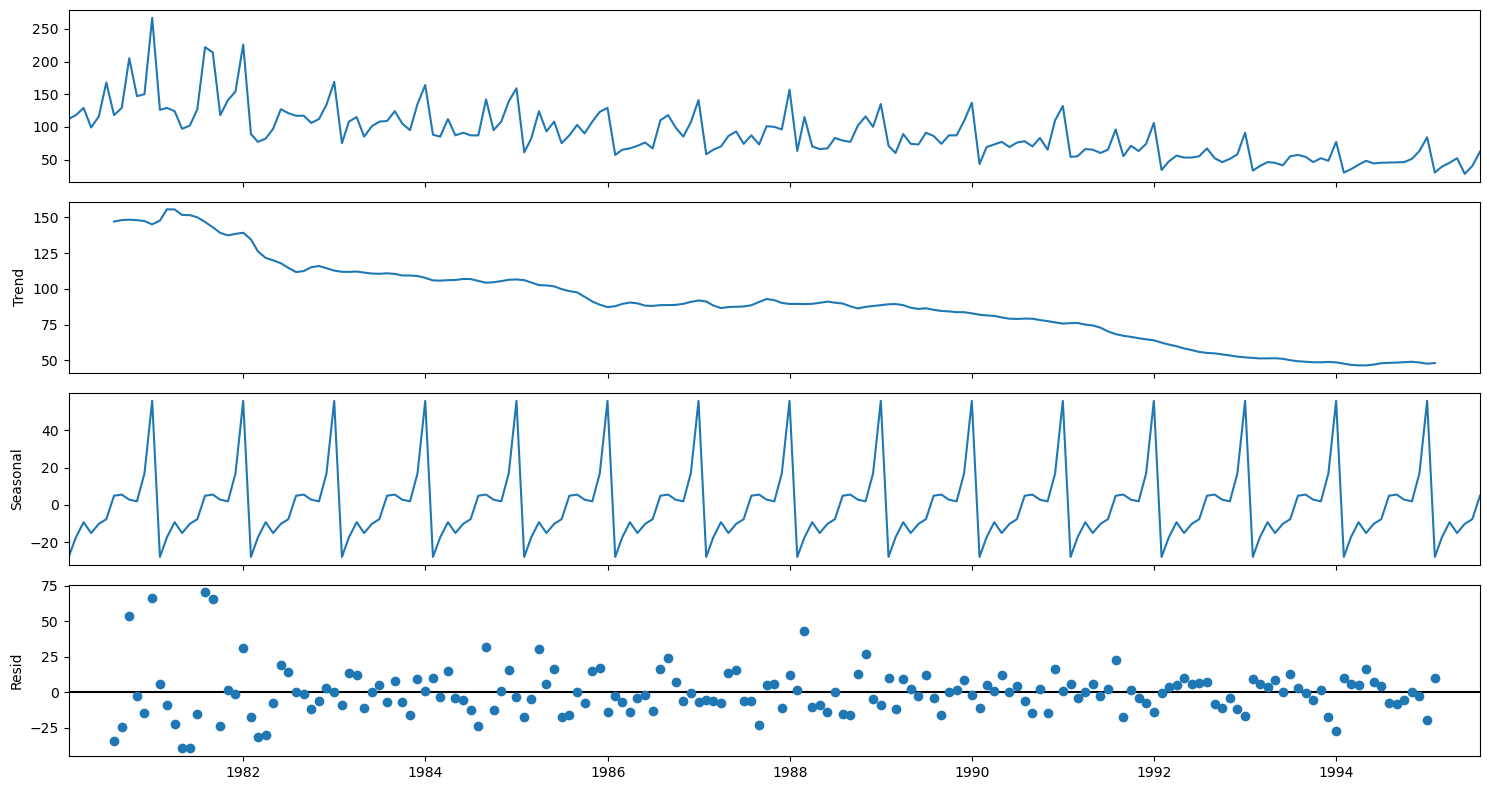
* Throughout the years, December shows the highest amount of rose wine sales followed by November.
* January has the lowest number of sales.
* There is an overall decrease in sales of rose wines across all months throughout the years.

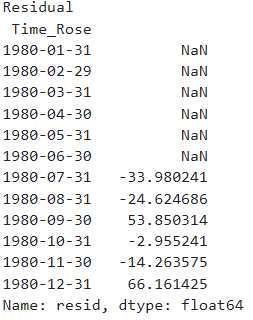
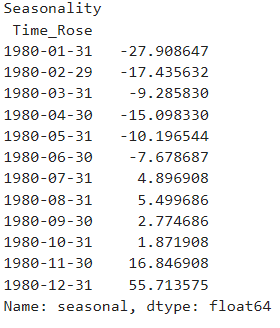
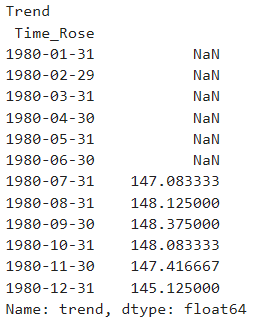
## 9.5 Decomposition

Decomposition is the process of breaking down complex data into smaller, more manageable components or subproblems to better understand the structure of the data, identify patterns, or simplify complex tasks. It involves separating a time series into its underlying components, which are trend, seasonality and residuals.

## 9.5.1 Additive Model

Figure 46 - Additive Decomposition of Rose Wine Data



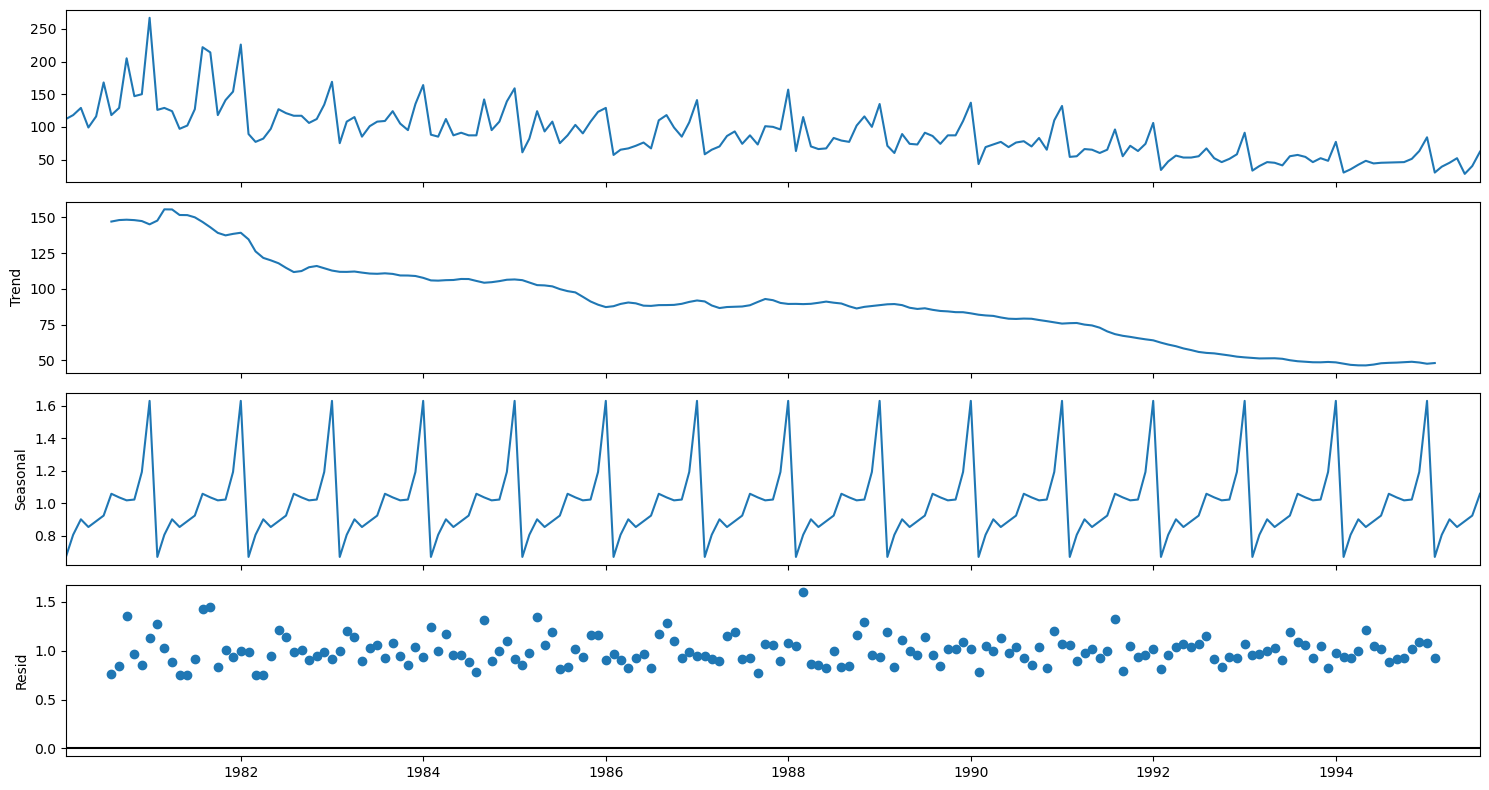


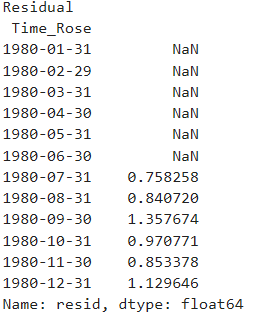
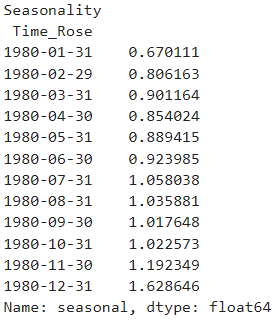
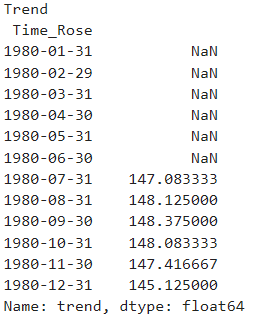
**Insights:**

* The trend shows a peak of sales initially during 1991 and then a constant decline of sales.
* Residuals are spread out in additive model which represent seasonality.
* Overall, there is both trend and seasonality.

## 9.5.2 Multiplicative Model

Figure 47 - Multiplicative Decomposition of Rose Wine Data





**Insights:**

* The trend chart is similar to additive model showing a peak at 1981 and then constant decline in sales.
* There is seasonality in data shown in the multiplicative model as well.
* The residuals are more stable and in a straight line in multiplicative model.
* The data can be assumed as multiplicative based on the stability of residuals.

# DATA PRE-PROCESSING – ROSE WINE

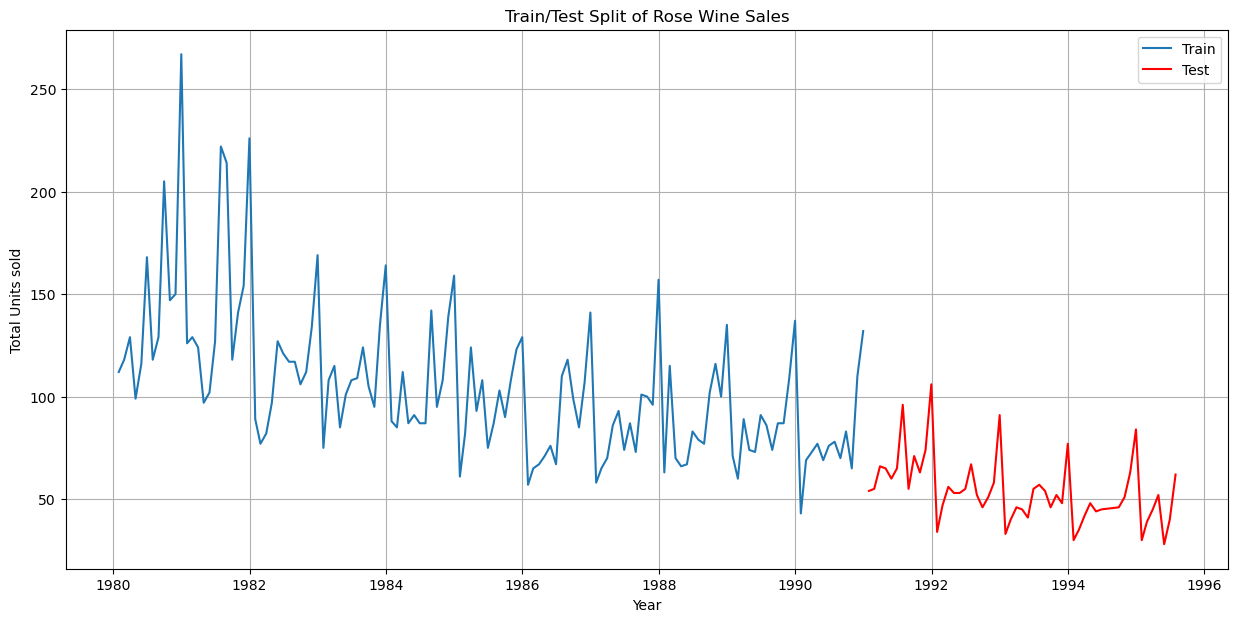
## 10.1 Missing Value Treatment

There are missing values present in the dataset (rose.csv). There are a total of 2 missing values and are treated by linear interpolation method.

## 10.2 Data Preparation

The training and testing data are separated for the purpose of building models. All observations prior to the January 1991 will be considered as Training Data, all observations from the January 1991 onwards will be considered as Testing Data.

Figure 48 - Train/Test Split of Rose Wine Sales



* There are a total of 132 observations present in the training data.
* There are 55 observations in the testing data.

The model will be built using the training data and tested on the test data to check for model efficiency.

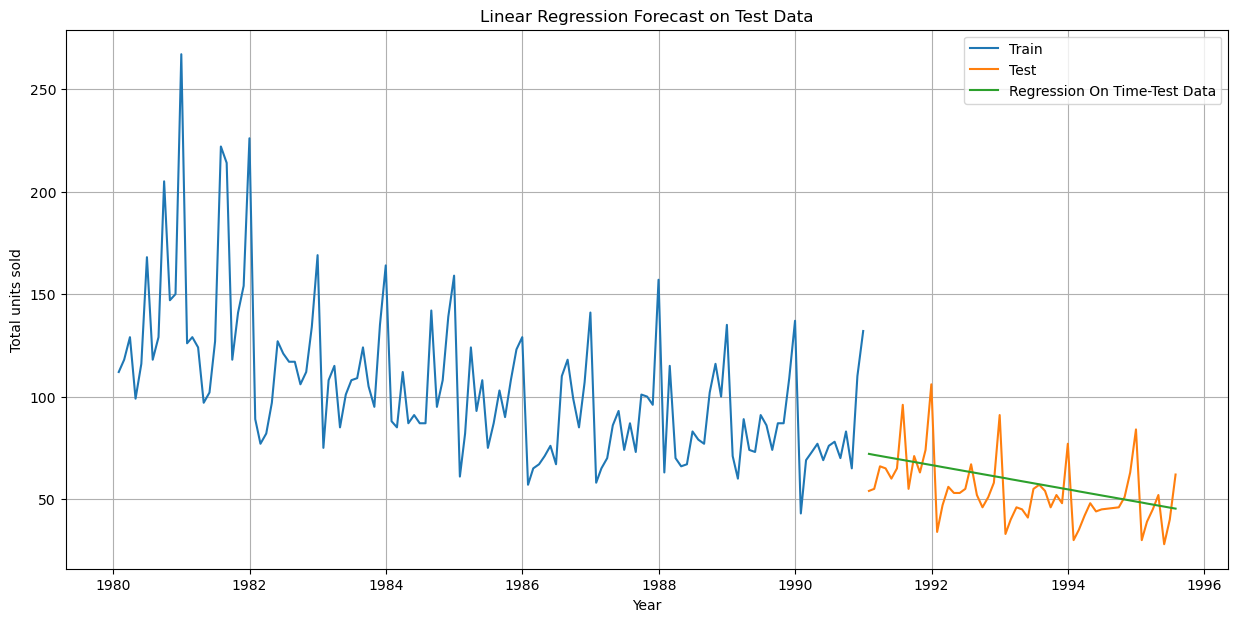
# MODEL BUILDING – ORIGINAL DATA – ROSE WINE

## 11.1 Linear Regression Model

Linear regression is a method used to predict future values based on the relationship between past observations, the observations are collected in monthly intervals and the model will be built based on the analysis of trends.

In this dataset the column containing the sales data (“Rose”) will the regressed against the order of occurrence.

Figure 49 - Linear Regression Model on Test Data



**Insights:**

* The model built and tested on the test data shows a predicted downward trend.
* It does not show any seasonality in the test data.

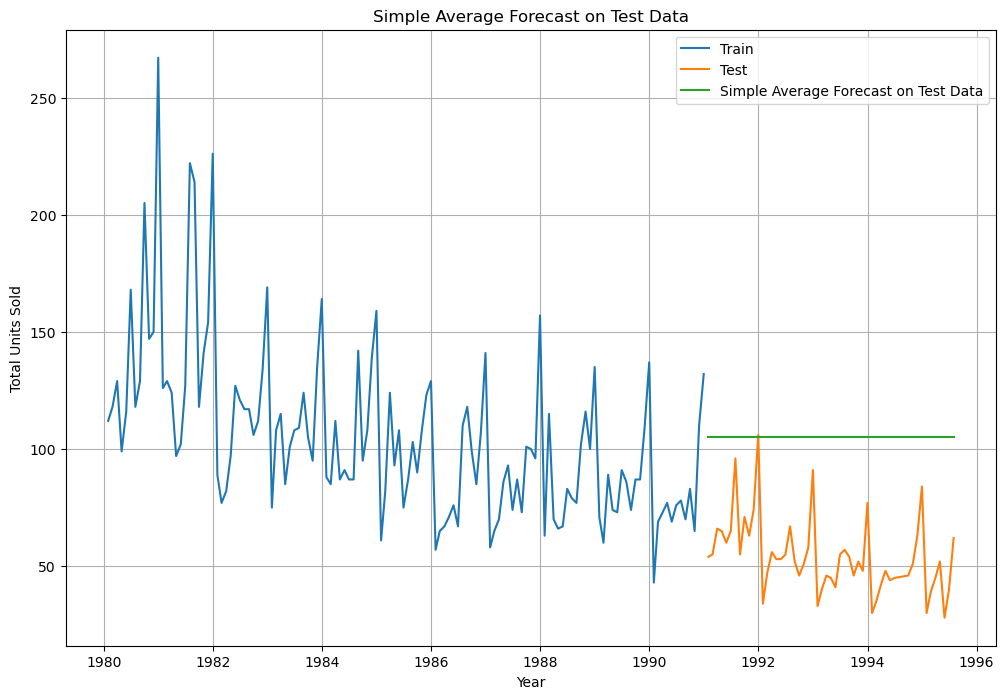
On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **15.269**.

## 11.2 Simple Average Model

The simple average method in time series forecasting involves using the average of historical data points to predict future values. It assumes that future values will be similar to the overall average of past observations.

In this dataset, the forecast will be based on using the average of train data values.

Figure 50 - Simple Average Model on Test Data



**Insights:**

* The predicted values show no trend or seasonality.
* The predicted values also vary slightly from the actual test values.

On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **53.460**.

## 11.3 Moving Average Model

The moving average model in time series forecasting involves averaging a fixed number of past observations to predict future values. The model smooths the data to identify trends by calculating the average of a sliding window of data points.

For this model, the rolling means (or moving averages) will be calculated for different trailing points (2,4,6,9). The best interval will be determined by the maximum accuracy (or the minimum error).

Figure 51 - Sample Values of Moving Averages

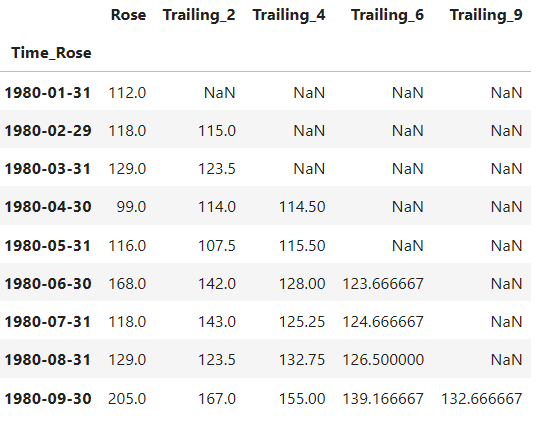


Figure 52 - Moving Average Model on Entire Dataset

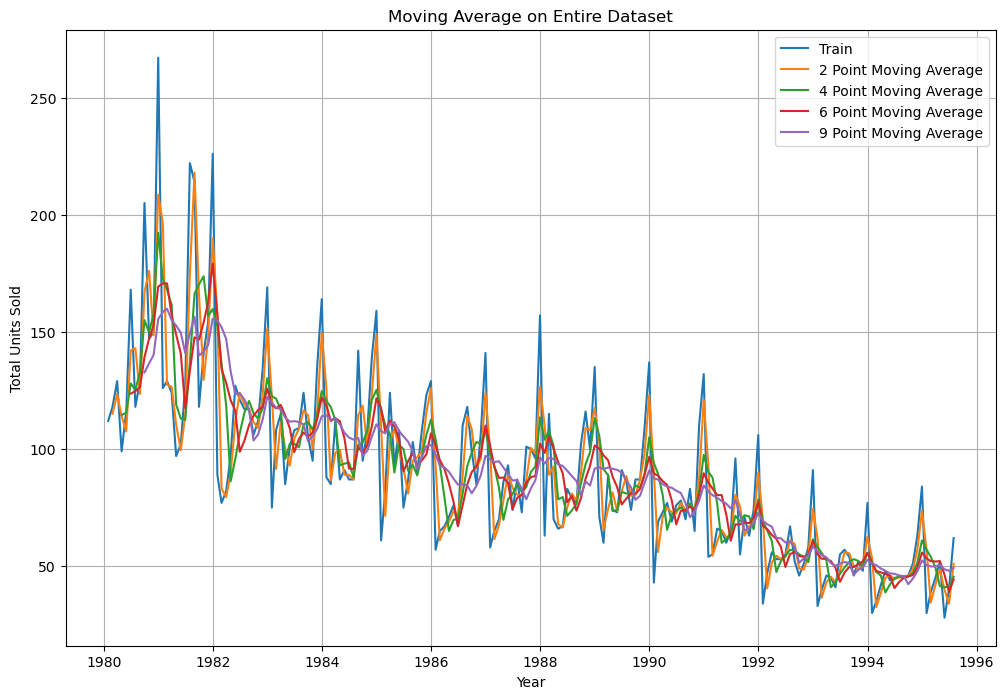
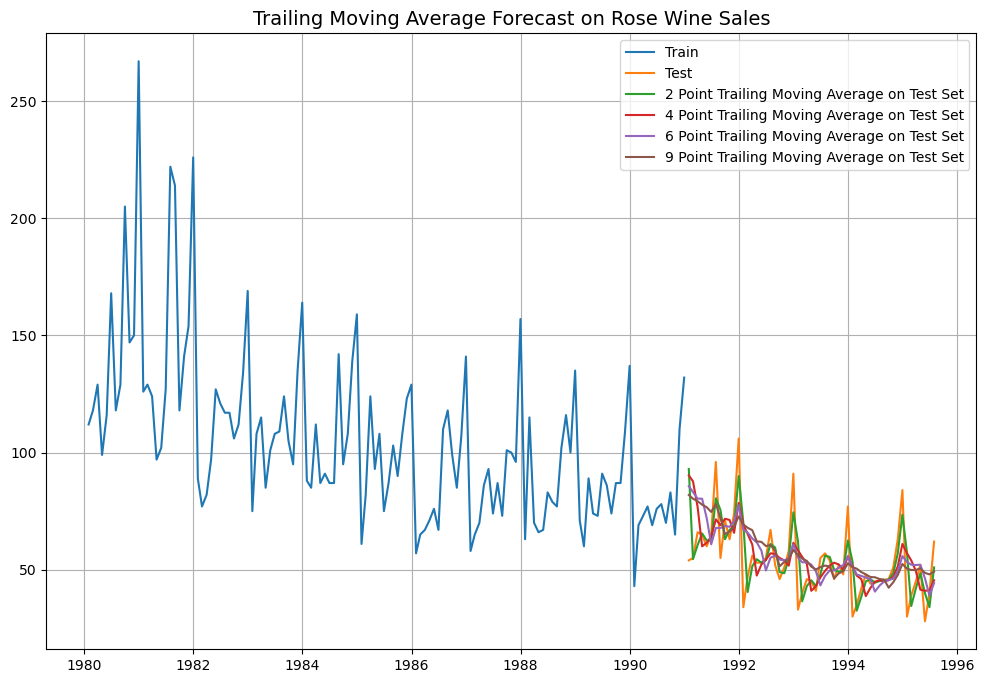


Figure 53 - Moving Average Model on Test Data



**Insights:**

* All trailing point predicted values show trend and seasonality unlike the previous models.
* It can be observed that the 2-point trailing moving average model is the most accurate to predict the test data.
* There is better accuracy with lower trailing point average.

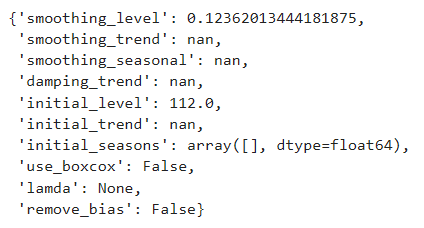
On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score are.

* 2 point moving average model – **11.529**
* 4 point moving average model – **14.451**
* 6 point moving average model – **14.566**
* 9 point moving average model – **14.728**

## 11.4 Single Exponential Smoothing Model

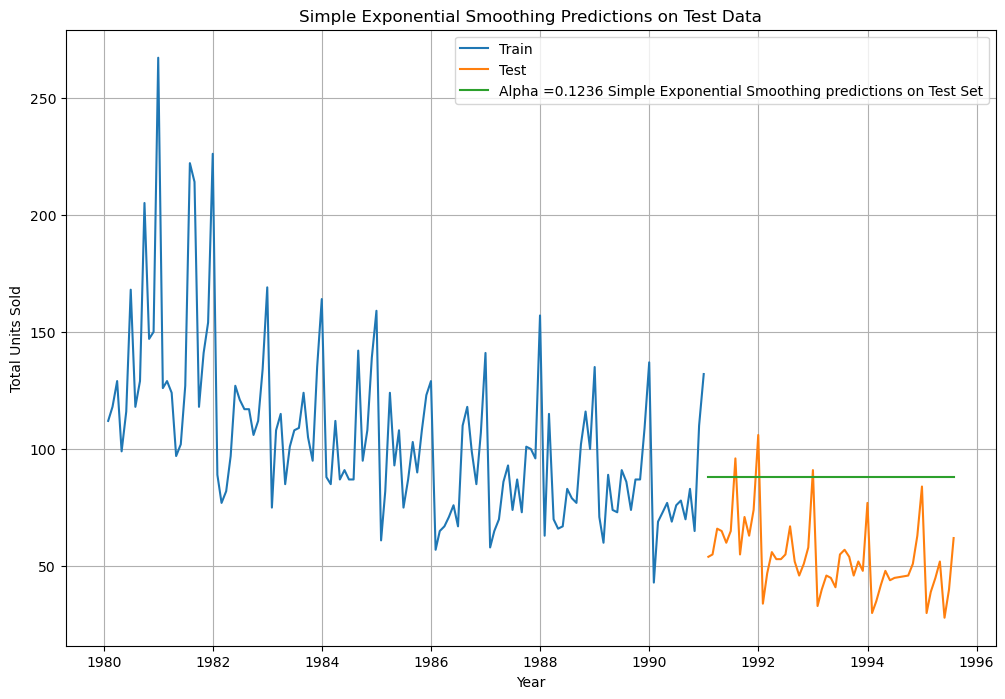
The single exponential smoothing method in time series forecasting is a technique that smooths past observations to make predictions for future values. It gives more weight to recent observations while gradually reducing the weight for older data. The method uses a smoothing constant (α) to control the level of smoothing, where, α is a value between 0 and 1.

Figure 54- Single Exponential Model Parameters



The simple exponential smoothing is built with optimized parameters, an α value of 0.1236 can be observed.

Figure 55 - Single Exponential Smoothing Model on Test Data



**Insights:**

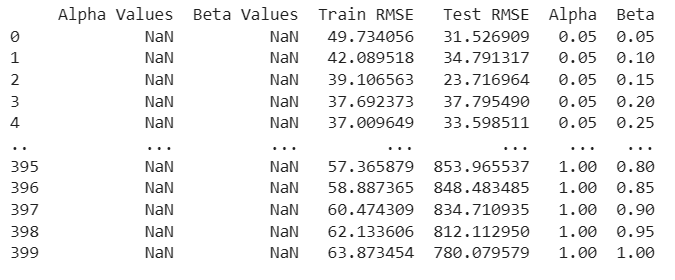
* The predicted values show no trend and seasonality on the test data.
* There is variation on the predicted test values and actual test values.

On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **37.592**.

## 11.5 Double Exponential Smoothing Model

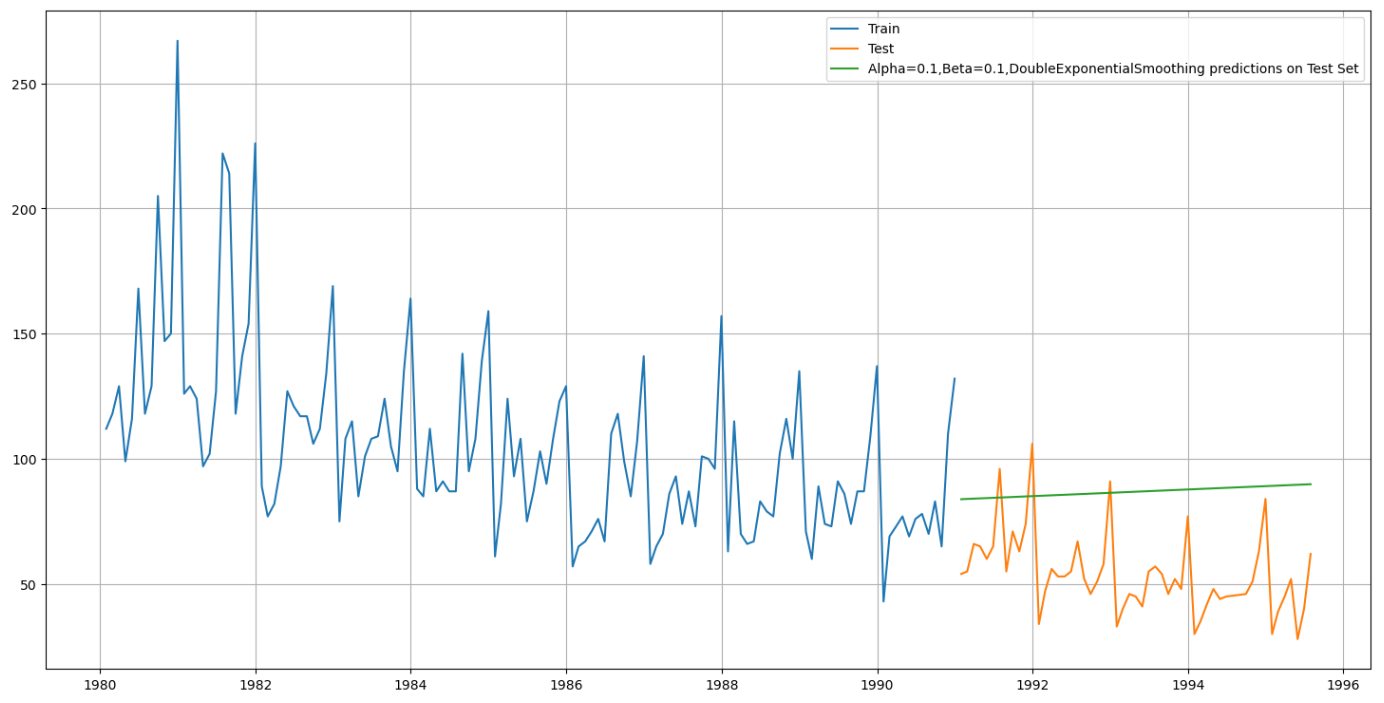
The double exponential smoothing method in time series forecasting extends single exponential smoothing by accounting for trends in the data. It uses two smoothing constants: one for the level (α) and one for the trend (β). This method adjusts both the level and the trend of the series over time, making it suitable for data with a linear trend. The model predicts future values by combining both smoothed levels and trends.

Figure 56 - Double Exponential Smoothing Model Parameters



The double exponential smoothing is built with optimized parameters, an α value of 0.05 and a β value of 0.05 can be observed.

Figure 57 - Double Exponential Smoothing Model on Test Data



**Insights:**

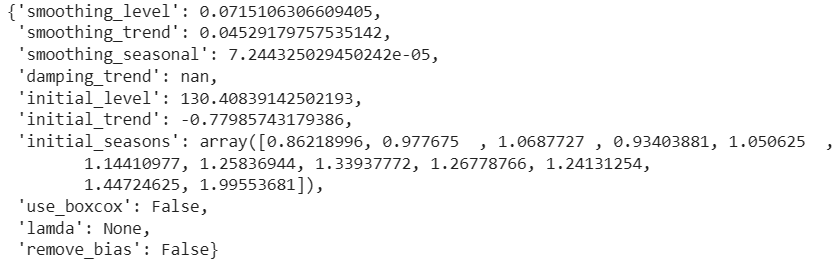
* The predicted values show an increase indicated a trend however it does not account for seasonality.
* The model is only suitable to showcase trend and no seasonality.

On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **16.329**.

## 11.6 Triple Exponential Smoothing Model

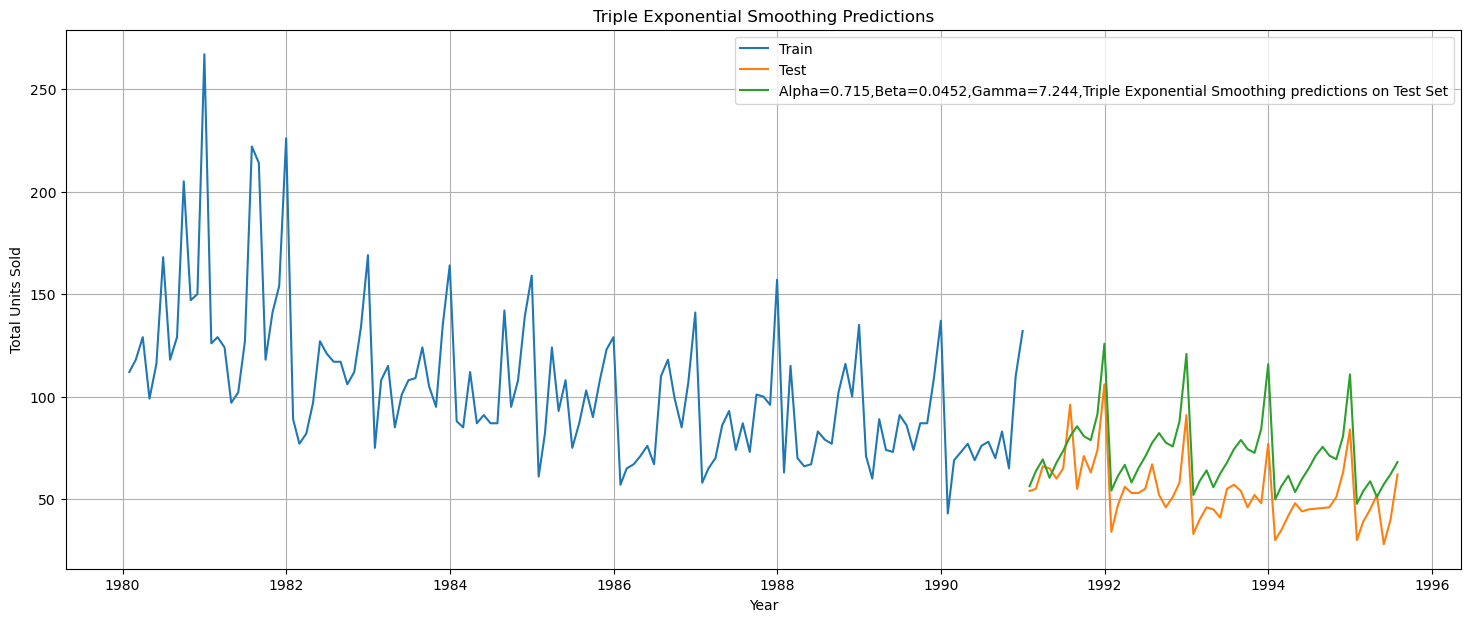
The triple exponential smoothing model (Holt-Winters) is used for time series forecasting with data that includes level, trend, and seasonality. It uses three smoothing parameters: α for the level, β for the trend, and γ for the seasonal component. This model adjusts for seasonal variations in addition to capturing the overall trend in the data.

Figure 58 - Triple Exponential Smoothing Model Parameters



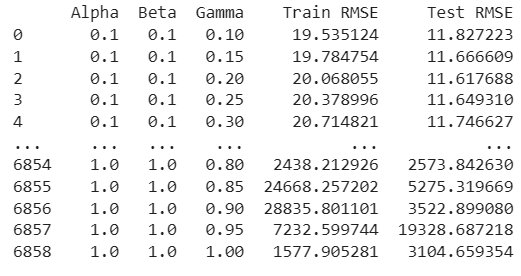
The triple exponential smoothing is built with optimized parameters, an α value of 0.715 and a β value of 0.452 and a γ value of 7.244 can be observed.

Figure 59 - Triple Exponential Smoothing Model on Test Data



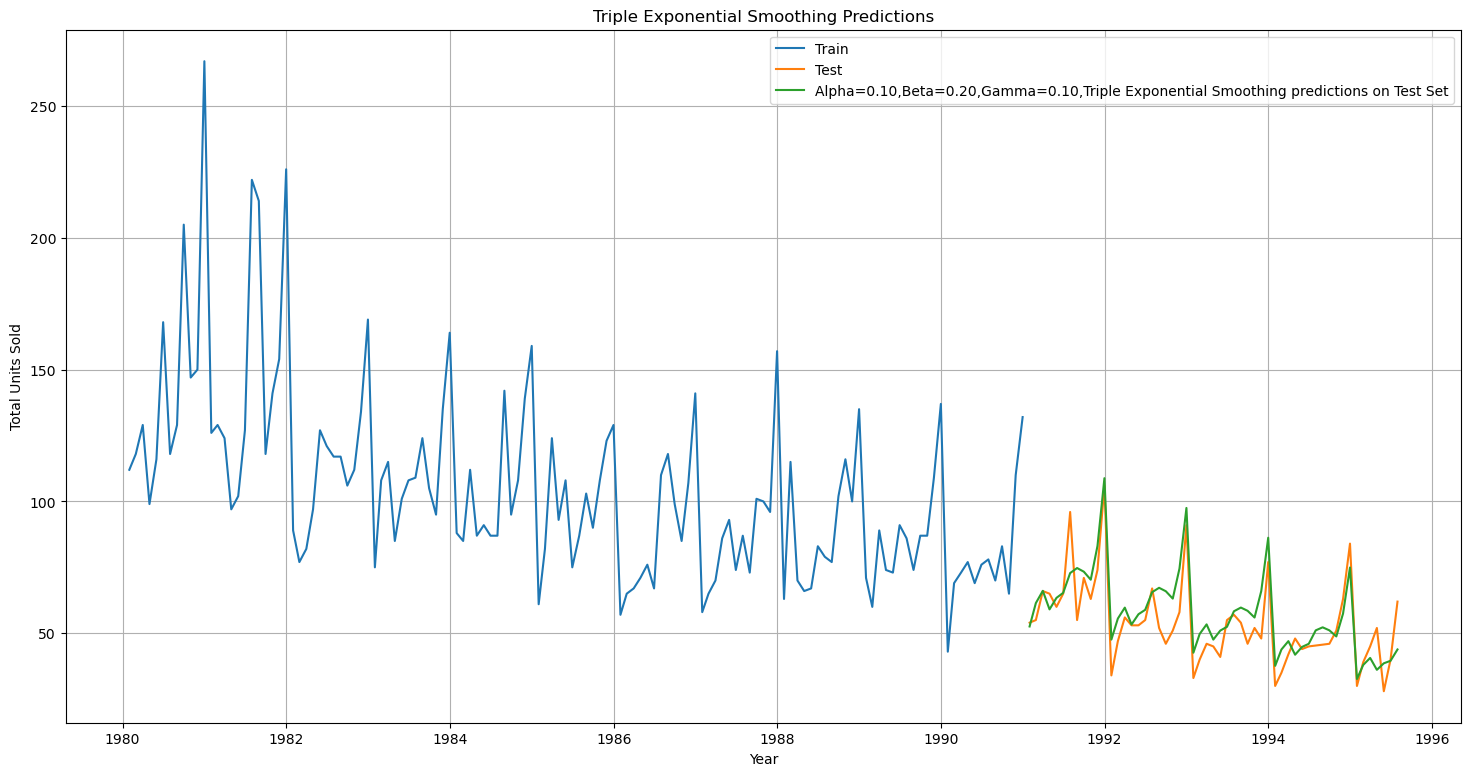
The recent observation has more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.

Figure 60 - Different Parameters for Triple Exponential Smoothing Model



an α value of 0.10 and a β value of 0.10 and a γ value of 0.10

Figure 61 - Triple Exponential Smoothing Model on Test Data



**Insights:**

* There is both seasonality and trend predicted on the test data.
* The predicted values match the actual test data values to a good degree which makes it the most suitable model for now.
* This model has the lowest RMSE score so far compared to the all the previous built models.

On evaluation of the model, it is observed that there is a Root mean square error (RMSE) metric score of **9.223**.

# CHECK FOR STATIONARITY – ROSE WINE

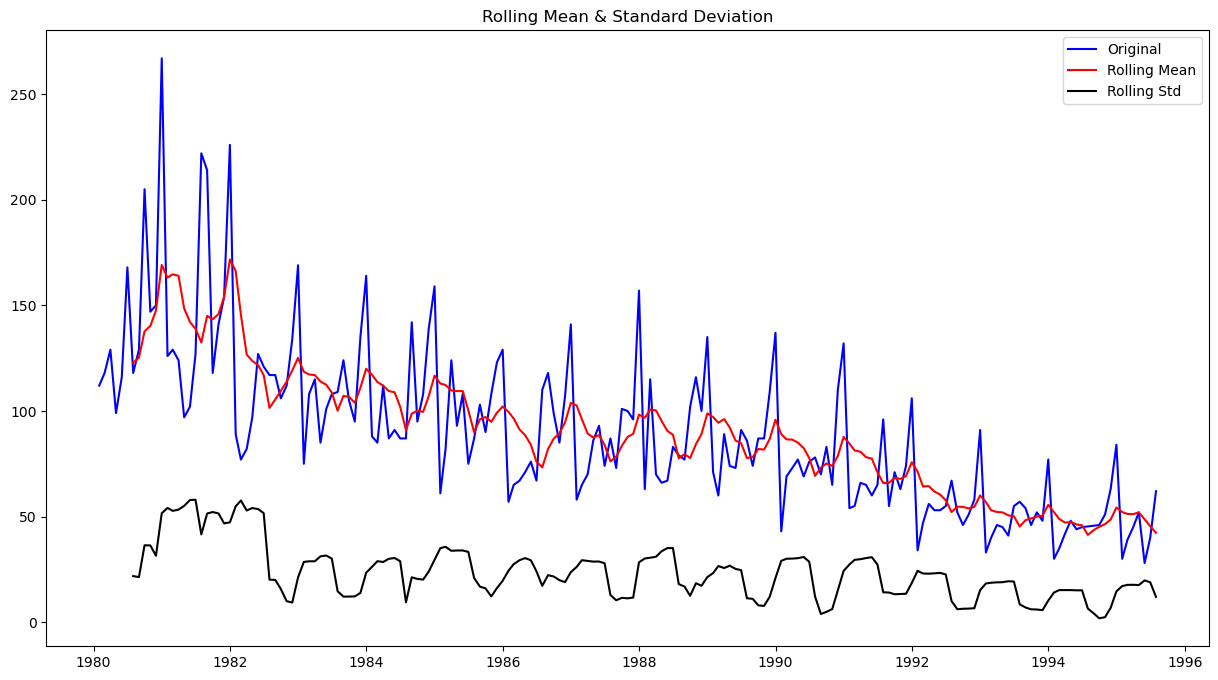
The Augmented Dickey-Fuller test is a unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

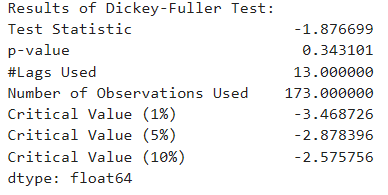
Frame the hypothesis:

H0: The Time Series has a unit root and is thus non-stationary.

H1: The Time Series does not have a unit root and is thus stationary.

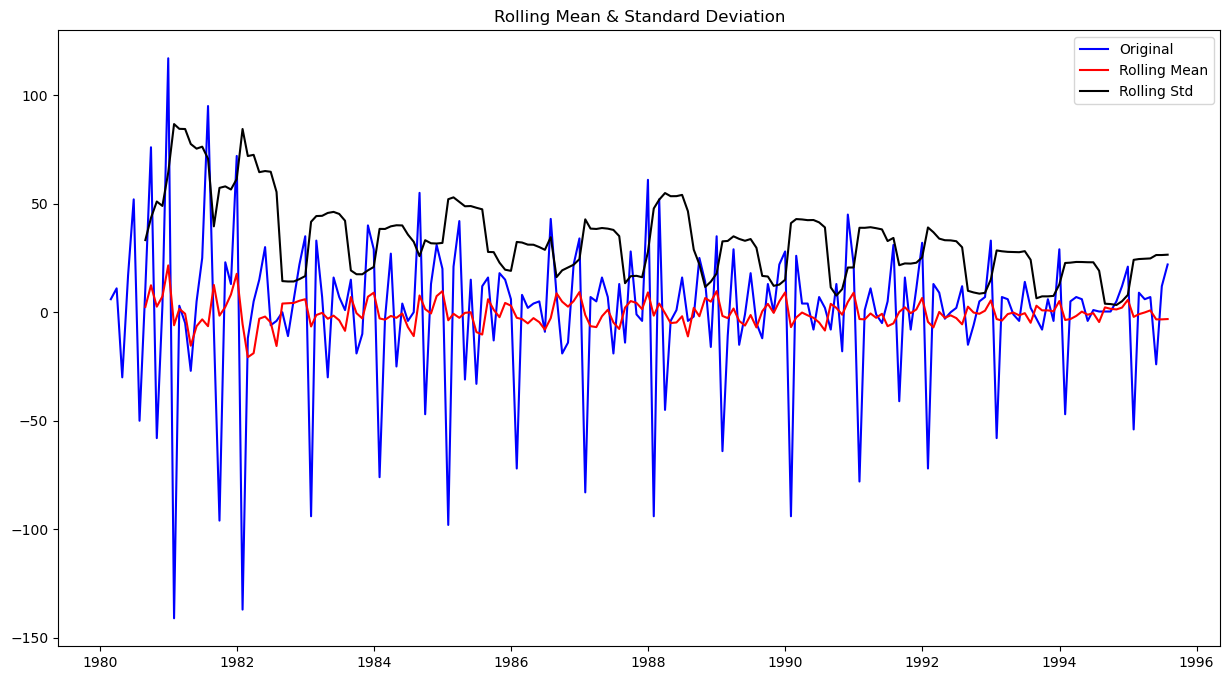
Figure 62 - Dickey-Fuller Test

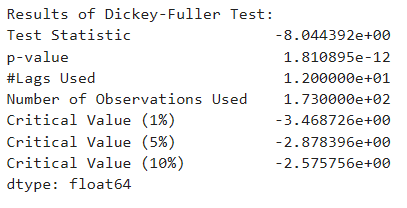




The p-value observed post the dickey-fuller test is 0.343 which is more than the α value of 0.05, so the data is non stationary as we fail to reject the null hypothesis.

Figure 63 - Dickey-Fuller Test on Differenced Data





To make the data stationary, a differencing approach can be used on the dataset with a diff value of 1 and all null values which are generated will be dropped.

We can observe a p-value of 1.810895e-12 which is less than α value of 0.05, therefore we reject the null hypothesis and the data is stationary.

# MODEL BUILDING – STATIONARY – ROSE WINE

## 13.1 ACF & PACF Plot

An ACF (Auto Correlation Function) plot is a graphical tool used in time series analysis to measure and visualize the correlation between a time series and its lagged versions. It shows how the values of the series are related to previous time steps (lags). The x-axis represents the lag, and the y-axis represents the correlation at each lag. Peaks in the ACF plot indicate significant autocorrelations at those lags. ACF plots are useful for identifying seasonality, trends, and determining the appropriate model for time series forecasting (e.g., ARIMA).

A PACF (Partial Auto Correlation Function) plot is used in time series analysis to measure the correlation between a time series and its lagged values, after removing the influence of shorter lags. Unlike the ACF plot, which shows total correlations, the PACF isolates the direct relationship between the series and a specific lag. The x-axis represents the lag, and the y-axis represents the partial correlation at each lag. Significant peaks in the PACF help identify the order of autoregressive (AR) components in models like ARIMA. It is especially useful for determining the appropriate number of lags for an AR model.

Figure 64 - ACF Plot

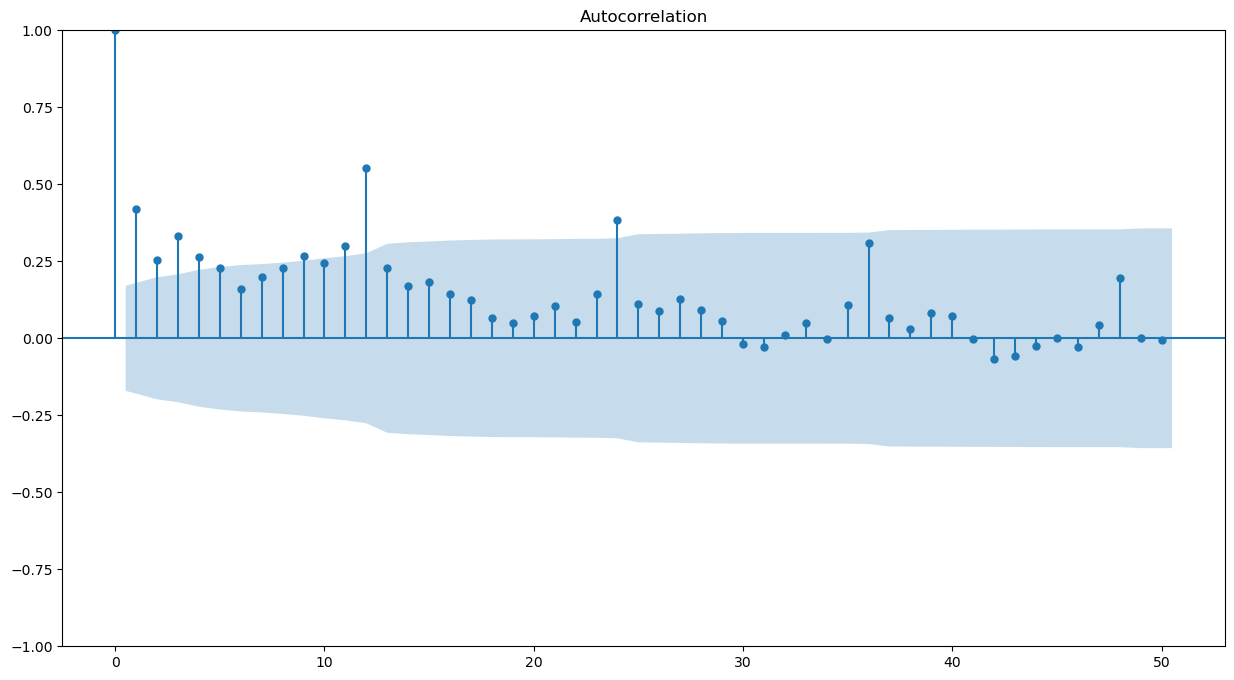
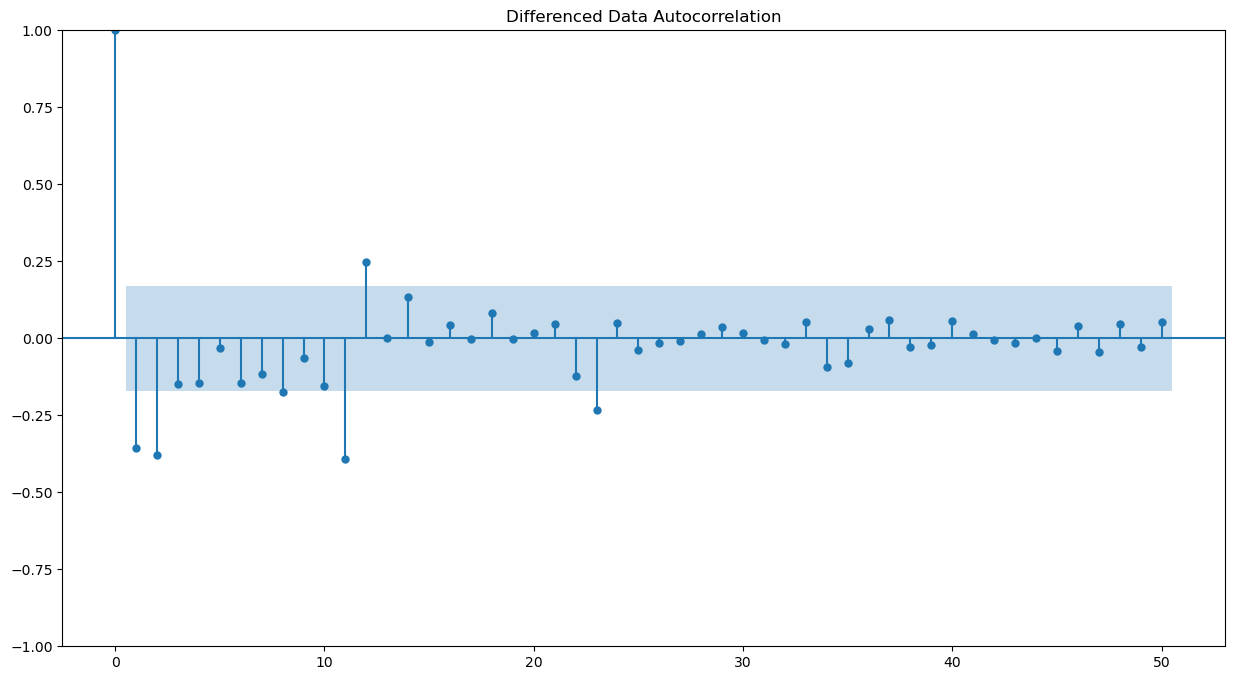


Figure 65 - PACF Plot



The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag after which the PACF plot cuts-off below the confidence interval.

The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off below the confidence interval.

It can be observed that after lag 1, we have few significant lags and hence we can take a value of p=2 and q=2 respectively.

## 13.2 Auto ARIMA Model

The Auto ARIMA model is an automated version of the ARIMA used for time series forecasting. It automatically selects the best combination of parameters (p, d, q) by evaluating different values of autoregressive, differencing, and moving average components. The model uses techniques like AIC or BIC to determine the optimal model that minimizes prediction error. Auto ARIMA simplifies the process of fitting ARIMA models by automating parameter selection. It is particularly useful for users without deep expertise in model tuning.

For the selection criteria of p,d,q the below ARIMA model is built using the automated model parameters with lowest Akaike Information Criteria.

Figure 66 - Parameters Combinations & AIC values – ARIMA Model

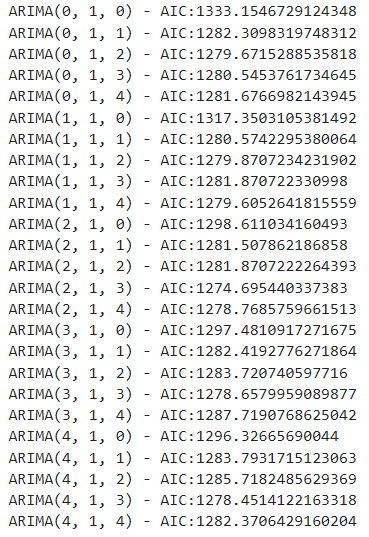
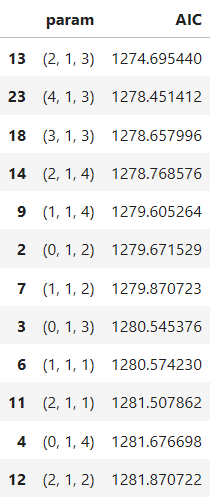
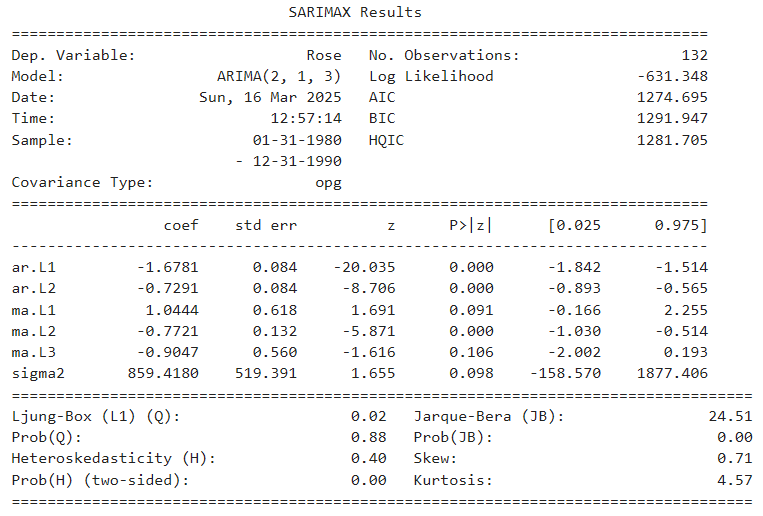


Figure 67 - Sorted Parameter and AIC Value - ARIMA Model



We can observe the AIC is lowest for the combination (2, 1, 3). The model will be built with this parameter.

Figure 68 - Auto ARIMA Model



On evaluation of the model, we can observe a root mean square error (RMSE) score of **36.812**.

## 13.3 Manual ARIMA Model

The manual ARIMA model involves manually selecting the parameters for autoregressive, differencing, and moving average components based on analysis of the time series data. The process typically includes examining the ACF and PACF plots to determine appropriate values for p and q. The parameter d is chosen based on the need for differencing to make the series stationary. Manual ARIMA requires a deeper understanding of the data and model diagnostics. It can be time-consuming but offers more control over the model specification.

Based on the previous ACF and PACF plot, we can observe a value of p=2 and q=2 and build a model based on those parameters.

Figure 69 - Manual ARIMA Model

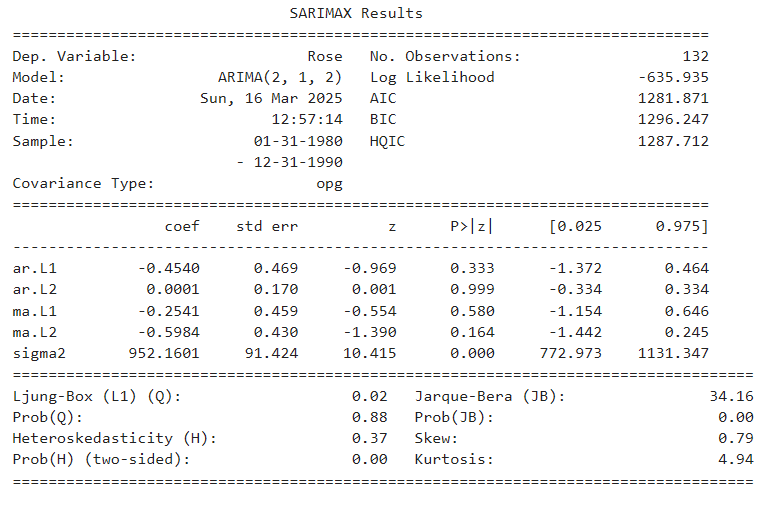
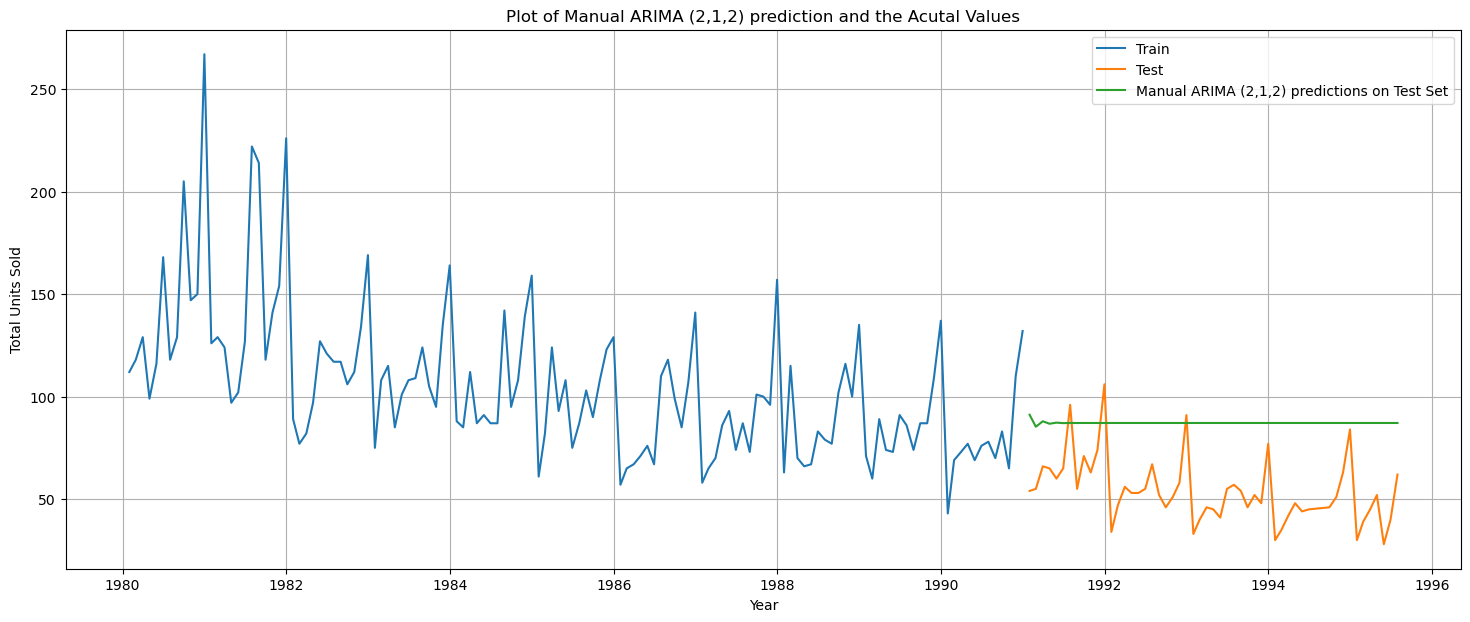


Figure 70 - Manual ARIMA Model on Test Data



**Insights:**

* There is no trend or seasonality shown in the predicted value.
* This model predicted values vary significantly from the actual test values

On evaluation of the model, we can observe a root mean square error (RMSE) score of **36.817**.

## 13.4 Auto SARIMA Model

The Auto SARIMA is an extension of the Auto ARIMA model that automatically handles seasonal components in time series forecasting. It selects optimal parameters for parts of the model. The model accounts for seasonality by incorporating seasonal differencing and seasonal AR and MA terms. Auto SARIMA automates the process of selecting the best model by evaluating different seasonal configurations. It is useful for forecasting time series data with seasonal patterns.

the optimum AIC which for the combination (3,1,1) (3,0,2,12) is considered due to errors with combinations lower thans this. The model will be built with these parameters.

Figure 71 - Auto SARIMA Model

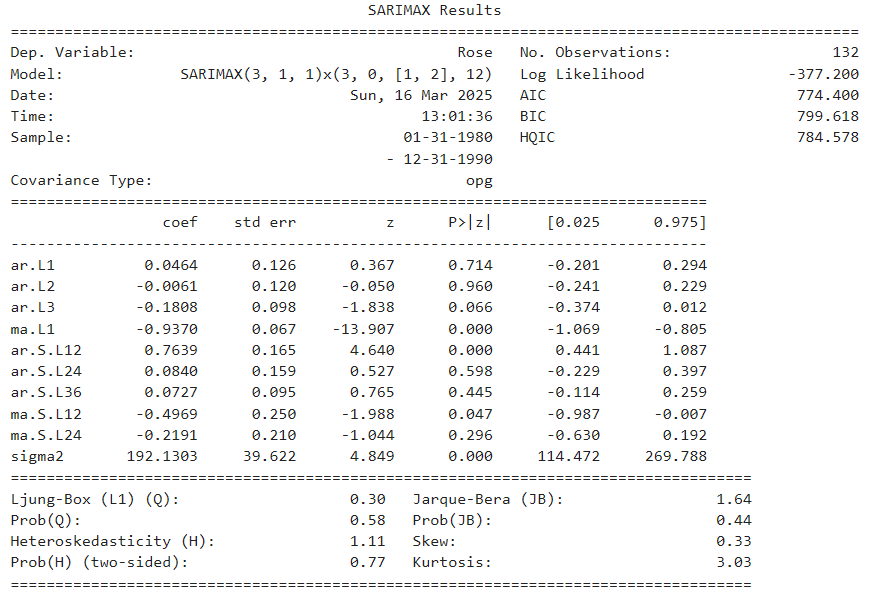
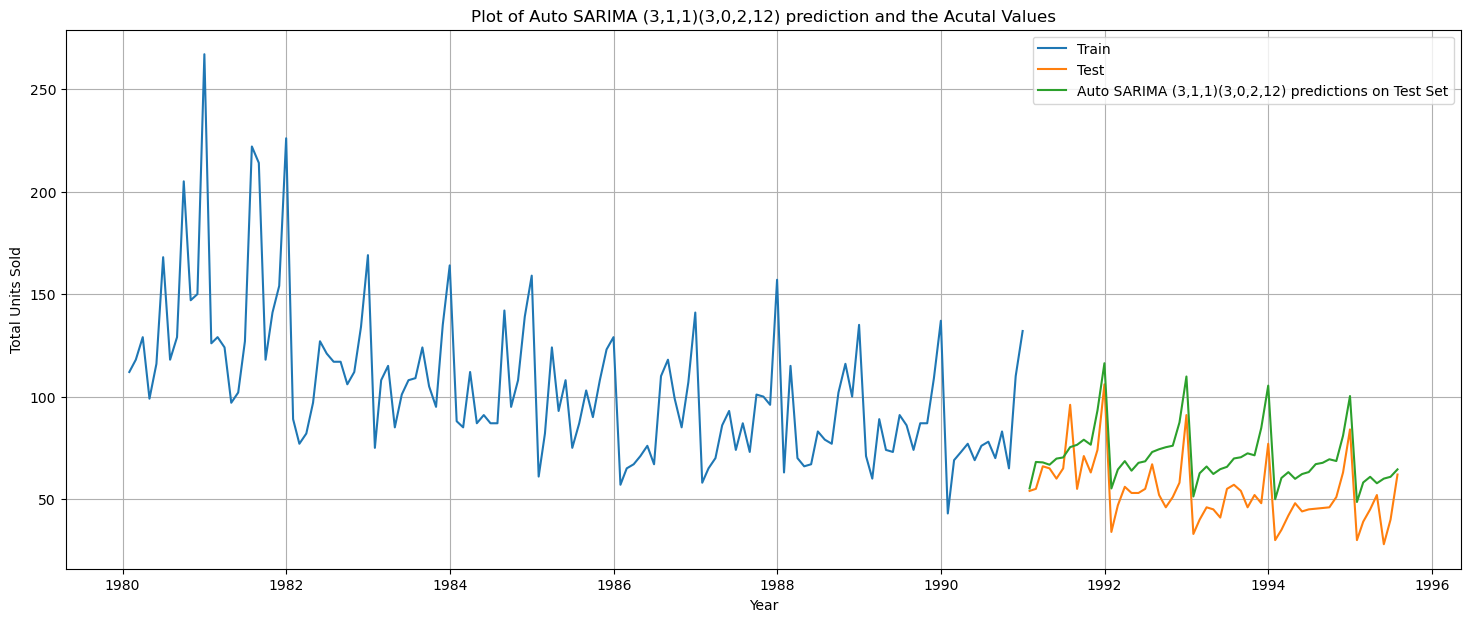


Figure 72 - Auto SARIMA Model on Test Data



**Insights:**

* The model predicts both trend and seasonality on the test data.
* The predicted values are slightly higher to the actual test data, but the low RMSE score makes it a viable model to choose.

On evaluation of the model, we can observe a root mean square error (RMSE) score of **18.881**.

## 13.5 Manual SARIMA Model

The manual SARIMA model involves manually selecting the parameters for both the non-seasonal and seasonal components of the SARIMA model. This includes choosing values for the autoregressive, differencing, and moving average terms for both non-seasonal and seasonal parts. The seasonal period is also determined based on the observed seasonal patterns in the data. The process typically involves analyzing ACF and PACF plots and checking for stationarity and seasonality.

The SARIMA model will be built with the parameters of are p=4, P=0, q=2 and Q=1 with this combination (4,1,2) (0,1,1,12)

Figure 73 - Manual SARIMA Model

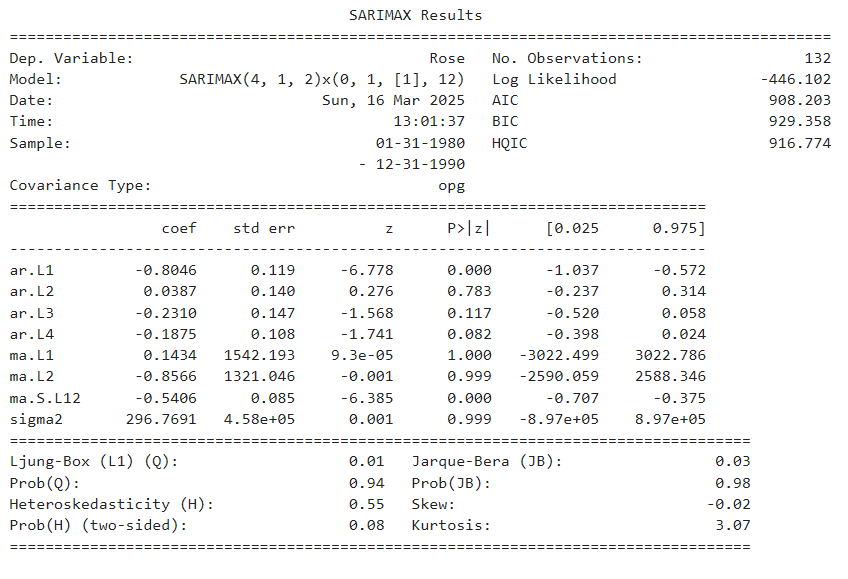
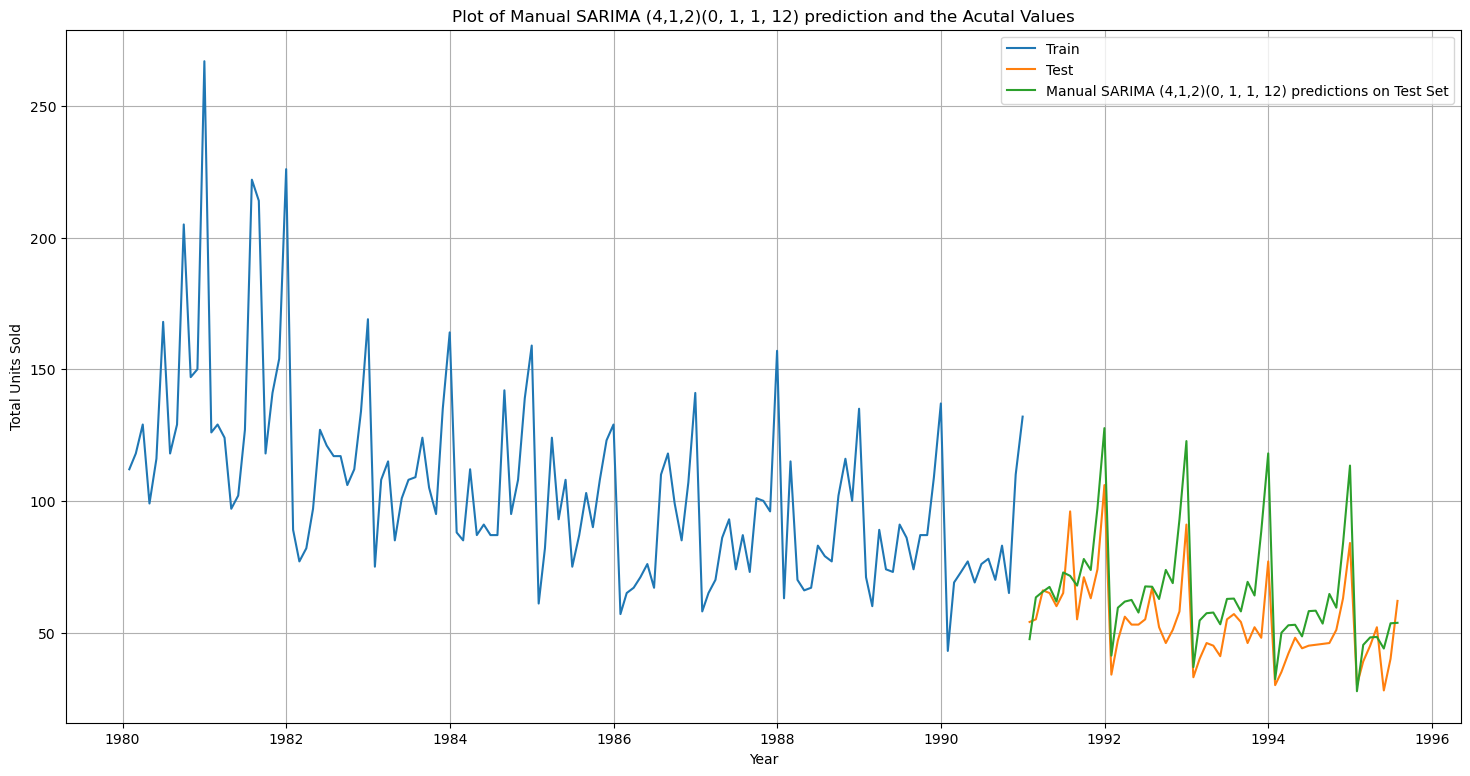


Figure 74 - Manual SARIMA Model on Test Data



**Insights:**

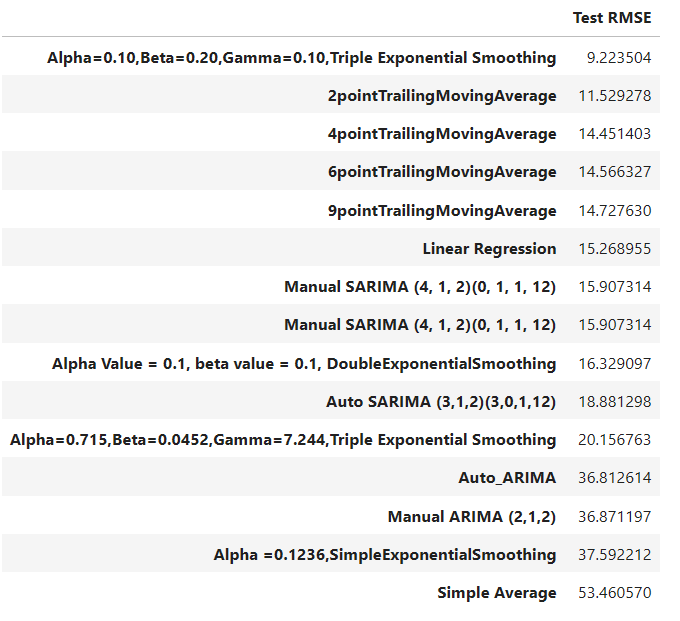
* The model shows both trend and seasonality on the predicted test values.
* The predicted values are close to the actual test data values.
* The model is suitable to use to predict the future values.

On evaluation of the model, we can observe a root mean square error (RMSE) score of **15.907**. The RMSE score of Manual SARIMA is lesser than the RMSE error of the Auto SARIMA Model.

# COMPARISON OF MODELS – ROSE WINE

All the models built so far will be compared and the most suitable model will be chosen and used to predict the sale of rose wine in the upcoming months of the dataset.

Figure 75 - Comparion of All Models



Based on the figure, we can see that the model with the lowest RMSE score is the Triple exponential smoothing model with α=0.10 and β=0.20 and a γ=0.10. The 2-point moving average model is also a suitable model to consider.

The Triple exponential smoothing model with α=0.10 and β=0.20 and a γ=0.10 will the best model and chosen to predict the values for the next months.

## 14.1 Rebuilding the Model

Figure 76 - TES Model - Predicted vs Actual Values

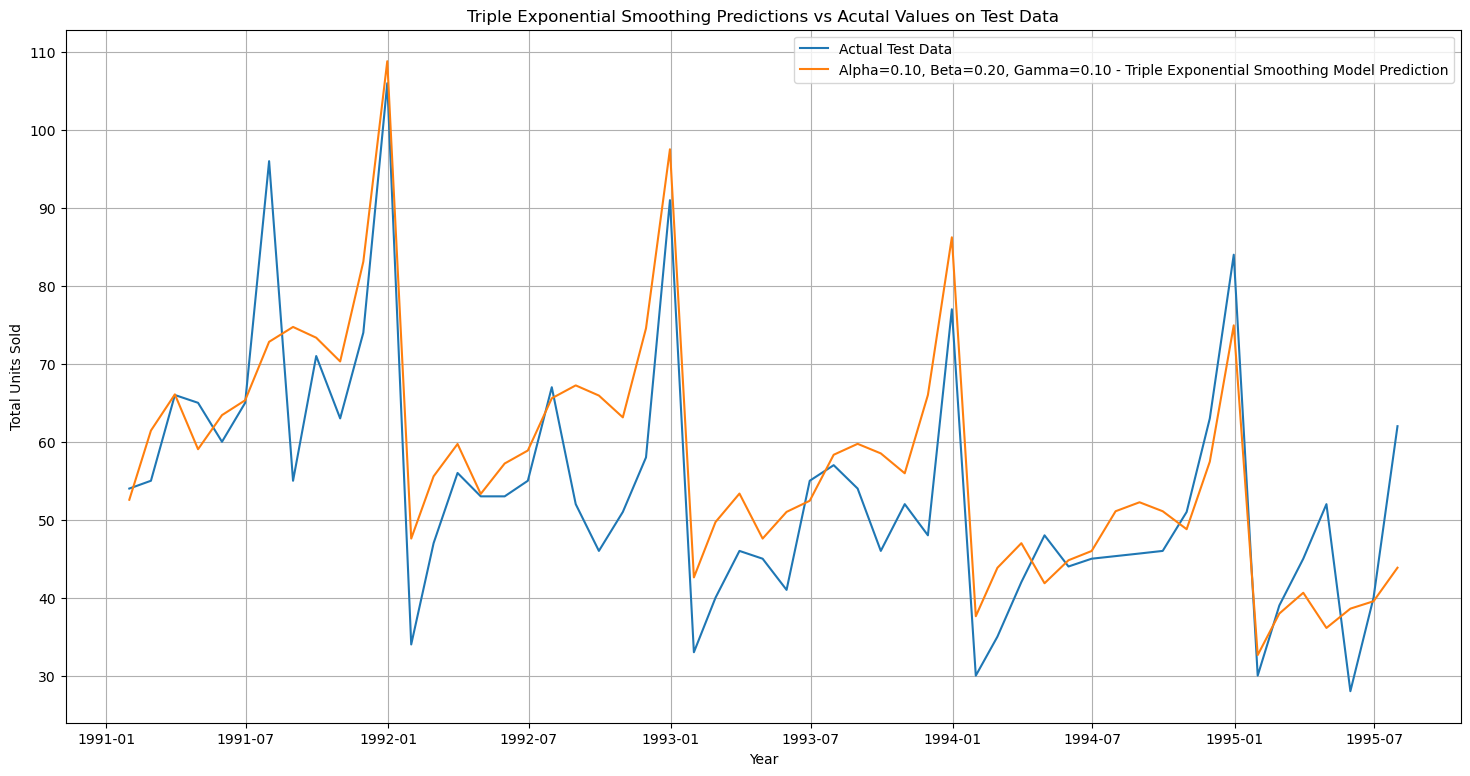
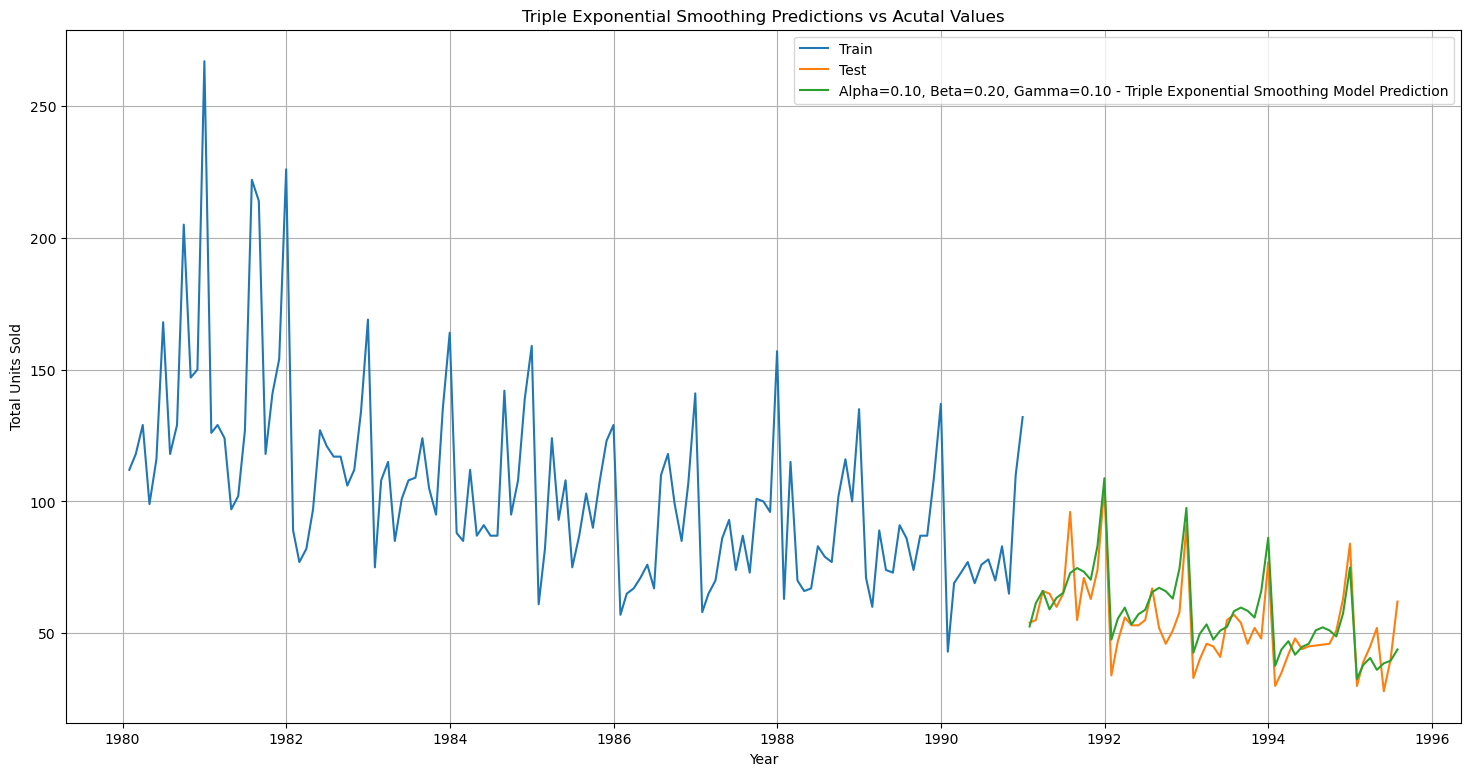


Figure 77 - TES Model on Test Data



The model is tested on test values which shows good performance as the data closely resemble the actual test data values.

The model is now rebuilt on the entire dataset and the next 12 months values will be predicted.

Figure 78 - TES Best Model

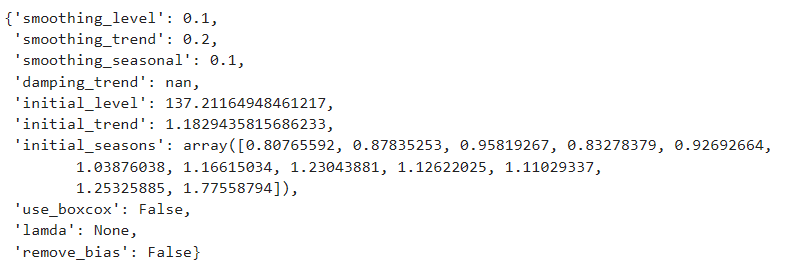


Figure 79 - TES Model - Forecast values

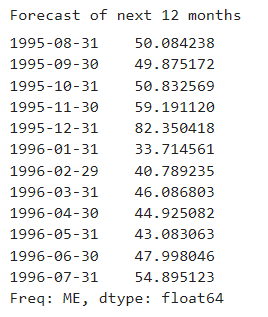
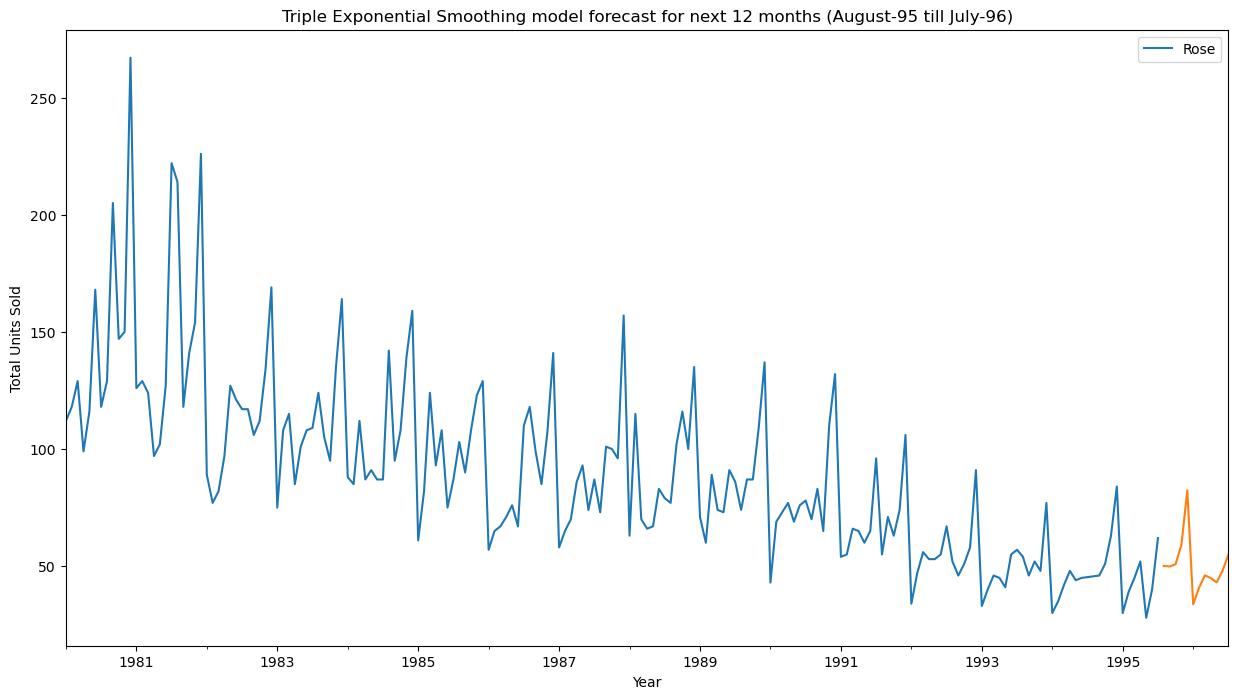


Figure 80 - Triple Exponential Smoothing Model Forecast Trend



The predicted values for the next 12 months (Aug-1995 to Jul-1996) showcase trend and seasonality showing a peak sale of 82 during December and a low number of sales of 33 in January 1996.

# ACTIONABLE INSIGHTS & RECOMMENDATIONS – ROSE WINE

* Triple Exponential Smoothing model provides the most accurate sales forecast, with the lowest RMSE and characteristics closely matching the historical data, making it the ideal model for future predictions
* Sales are highly seasonal, with peaks occurring in September, October, November, and December, driven by festive events and holiday celebrations, accounting for 40% of the total sales forecast.
* Significant decline in sales forecast, with the average sales dropping by 50%, and a sharp decrease in maximum sales, indicating potential challenges in future sales performance.
* Historical sales show steady growth towards the end of the year, with December consistently being the highest sales month, highlighting the importance of marketing strategies during this period.
* The months from January to June show low sales, suggesting the need for targeted promotions, new product variations, or rebranding efforts to boost sales in this challenging period.
* Leverage holiday seasons (September to December) with targeted promotional offers, discounts, and bulk sales strategies to boost revenue during peak sales months.
* Introduce a market-friendly version of the rose wine during the off-season (January to June) to appeal to cost-conscious consumers and mitigate sales declines.
* Enhance marketing campaigns by offering free gifts with significant purchases and running e-Commerce competitions to attract more customers and increase brand visibility.
* Conduct in-depth market research to identify key factors influencing sales and incorporate those insights into future forecasting models for more accurate predictions.
* Rebrand the rose wine product to create a fresh appeal and break the declining sales trend, particularly during the slower sales period from January to June.