

**Time Series Forecasting - Business Report**

**By: Aaryani Kadiyala**

**PGP-Data Science and Business Analytics**

**(PGPDSBA.O.JAN24.A)**

**Table of Contents**

1. Problem ----------------------------------------------------------------------------------------------------------------------- 5

**List of Figures:**

Figure 1: First 5 rows Dataset Sample --------------------------------------------------------------------------------------- 6

Figure 2: Last 5 rows Dataset Sample ---------------------------------------------------------------------------------------- 6

Figure 3: Data types of the dataset ------------------------------------------------------------------------------------------- 6

Figure 4: Statistical summary of the dataset ------------------------------------------------------------------------------- 7

Figure 5: Unique categories of categorical variables --------------------------------------------------------------------- 7

Figure 6: Univariant analysis of Age ------------------------------------------------------------------------------------------ 8

Figure 7: Univariant analysis of Work Experience ------------------------------------------------------------------------- 8

Figure 8: Bar Plot of the Gender ----------------------------------------------------------------------------------------------- 9

Figure 9: Bar Plot of the Transport -------------------------------------------------------------------------------------------- 9

Figure 10: Bar Plot between Transport and Gender ---------------------------------------------------------------------- 9

Figure 11: Transport and Gender in numbers ------------------------------------------------------------------------------ 9

Figure 12: Prediction with distance w.r.t Transport --------------------------------------------------------------------- 10

Figure 13: Prediction with Work Experience w.r.t Transport --------------------------------------------------------- 11

Figure 14: Prediction with Distance w.r.t Gender ------------------------------------------------------------------------ 12

Figure 15: Prediction with Age w.r.t Transport --------------------------------------------------------------------------- 13

Figure 16: Correlation of the dataset of the dataset -------------------------------------------------------------------- 13

Figure 17: Before Outlier Treatment ---------------------------------------------------------------------------------------- 14

Figure 18: After Outlier Treatment ------------------------------------------------------------------------------------------ 15

Figure 19: Shape of training & testing dataset ---------------------------------------------------------------------------- 16

Figure 20: Training Dataset --------------------------------------------------------------------------------------------------- 16

Figure 21: Testing Dataset ----------------------------------------------------------------------------------------------------- 16

Figure 22: Classification Report of Training Dataset --------------------------------------------------------------------- 16

Figure 23: Classification Report of Testing Dataset ---------------------------------------------------------------------- 16

Figure 24: Confusion Matrix of training Data Set ------------------------------------------------------------------------ 17

Figure 25: Confusion Matrix of testing Data Set -------------------------------------------------------------------------- 17

Figure 26: Classification Report of Training Dataset --------------------------------------------------------------------- 17

Figure 27: Classification Report of Testing Dataset ---------------------------------------------------------------------- 18

Figure 28: Confusion Matrix of training Data Set ------------------------------------------------------------------------ 18

Figure 29: Confusion Matrix of testing Data Set -------------------------------------------------------------------------- 18

Figure 30: Classification Report of Training Dataset --------------------------------------------------------------------- 19

Figure 31: Classification Report of Testing Dataset ---------------------------------------------------------------------- 19

Figure 32: Confusion Matrix of training Data Set ------------------------------------------------------------------------ 19

Figure 33: Confusion Matrix of testing Data Set --------------------------------------------------------------------------19

Figure 34: Classification Report of Training Dataset --------------------------------------------------------------------- 20

Figure 35: Classification Report of Testing Dataset --------------------------------------------------------------------- 20

Figure 36: Confusion Matrix of training Data Set ------------------------------------------------------------------------ 20

Figure 37: Confusion Matrix of testing Data Set -------------------------------------------------------------------------- 20

Figure 38: Classification Report of Training Dataset --------------------------------------------------------------------- 21

Figure 39: Classification Report of Testing Dataset ---------------------------------------------------------------------- 21

Figure 40: Confusion Matrix of training Data Set ------------------------------------------------------------------------ 21

Figure 41: Confusion Matrix of testing Data Set -------------------------------------------------------------------------- 21

Figure 42: Classification Report of Training Dataset --------------------------------------------------------------------- 22

Figure 43: Classification Report of Testing Dataset ---------------------------------------------------------------------- 22

Figure 44: Confusion Matrix of training Data Set ------------------------------------------------------------------------ 22

Figure 45: Confusion Matrix of testing Data Set -------------------------------------------------------------------------- 23

Figure 46 : Model Comparison for training data ------------------------------------------------------------------------- 24

Figure 47 : Model Comparison for testing data -------------------------------------------------------------------------- 24

Figure 48 : Important features of final model ---------------------------------------------------------------------------- 24

Figure 49: First 5 rows of the data ------------------------------------------------------------------------------------------- 26

Figure 50: Last 5 rows of the data ------------------------------------------------------------------------------------------- 26

Figure 51: Data types of the dataset ---------------------------------------------------------------------------------------- 26

Figure 52: Deal & Description in separate dataset ----------------------------------------------------------------------- 27

Figure 53: Deals Secured ------------------------------------------------------------------------------------------------------ 27

Figure 54: Deals Not Secured ------------------------------------------------------------------------------------------------- 27

Figure 55: No. of characters in each corpus ------------------------------------------------------------------------------- 27

Figure 56: After removing http links ---------------------------------------------------------------------------------------- 28

Figure 57: After words de-contraction of deals secured --------------------------------------------------------------- 28

Figure 58: After words de-contraction of deals not secured ---------------------------------------------------------- 28

Figure 59: After removing numbers of deals secured ------------------------------------------------------------------- 29

Figure 60: After removing numbers of deals not secured ------------------------------------------------------------- 29

Figure 61: After tokenizing of deals secured ------------------------------------------------------------------------------ 29

Figure 62: After tokenizing of deals not secured ------------------------------------------------------------------------- 29

Figure 63: After converting into lower case of deals secured --------------------------------------------------------- 30

Figure 64: After converting into lower case of deals not secured --------------------------------------------------- 30

Figure 65: After removing punctuations of deals secured ------------------------------------------------------------- 30

Figure 66: After removing punctuations of deals not secured -------------------------------------------------------- 30

Figure 67: After removing stopwords of deals secured ---------------------------------------------------------------- 31

Figure 68: After removing stopwords of deals not secured ----------------------------------------------------------- 31

Figure 69: After lemmatization of deals secured ------------------------------------------------------------------------- 31

Figure 70: After Normalization of deals secured ------------------------------------------------------------------------- 31

Figure 71: Wordcloud of deals secured ------------------------------------------------------------------------------------ 32

Figure 72: Wordcloud of deals not secured ------------------------------------------------------------------------------- 33

**Problem :**

### **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. 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.

### **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.

**Solution :**

### **Rose Wine:**

### **Synopsis:**

* Total no. of Entries = 187
* No. of Null Value = 2
* No Duplicates
* Train Set split till 1990 & Test set from 1991.
* Forecasting Models used
  + Linear Regression
  + Simple Average
  + 2 Point Moving Average
  + 4 Point Moving Average
  + 6 Point Moving Average
  + 9 Point Moving Average
  + Single Exponential Smoothing
  + Double Exponential Smoothing (Holt’s Model)
  + Triple Exponential Smoothing (Holt-Winter Model)
  + ARIMA / SARIMA (Auto fitted)
  + ARIMA / SARIMA (Manually fitted)

### **Sparkling Wine:**

### **Synopsis:**

* Total no. of Entries = 187
* No Null values
* No Duplicates
* Train Set split till 1990 & Test set from 1991.
* Forecasting Models used
  + Linear Regression
  + Simple Average
  + 2 Point Moving Average
  + 4 Point Moving Average
  + 6 Point Moving Average
  + 9 Point Moving Average
  + Single Exponential Smoothing
  + Double Exponential Smoothing (Holt’s Model)
  + Triple Exponential Smoothing (Holt-Winter Model)
  + ARIMA / SARIMA (Auto fitted)
  + ARIMA / SARIMA (Manually fitted)

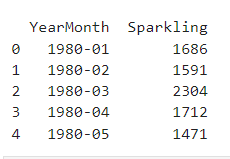
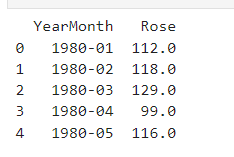


Figure 1: Head of Rose Wine Figure 2: Head of Sparkling Wine

Converted ‘YearMonth’ Column of both Rose & Sparkling wine into datetime.

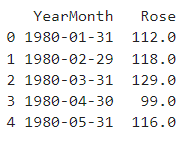
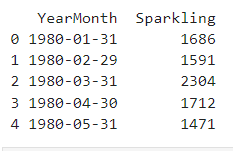
 

Figure 3: Head of Rose Wine after convertion Figure 4: Head of Sparkling Wine after convertion

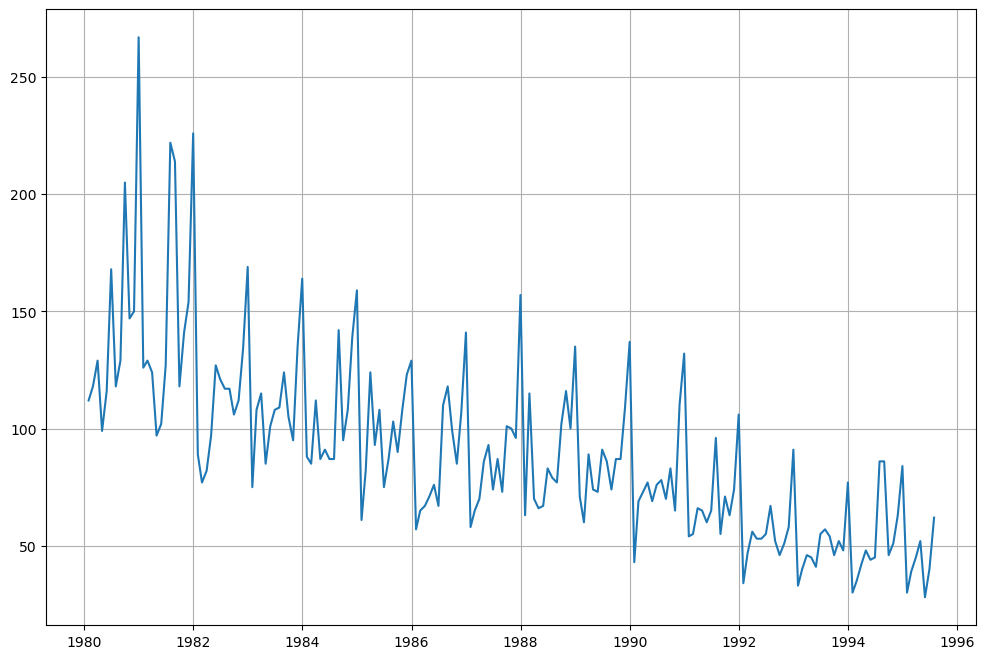


Figure 5 : Rose Wine Plot on Original Data

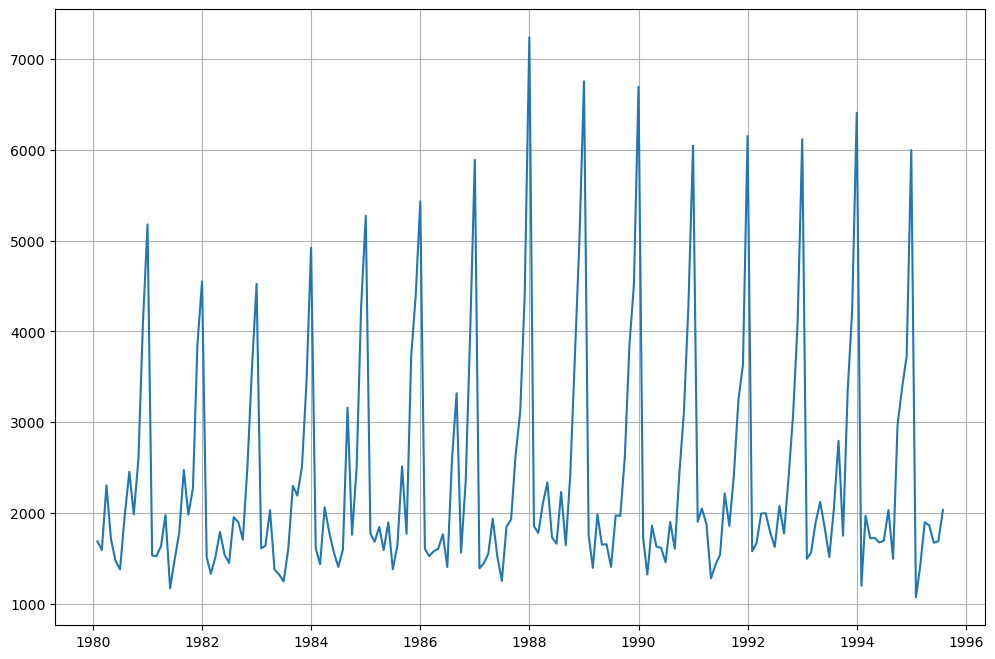


Figure 6: Sparkling Wine Plot on Original Data

### **Exploratory Data Analysis:**

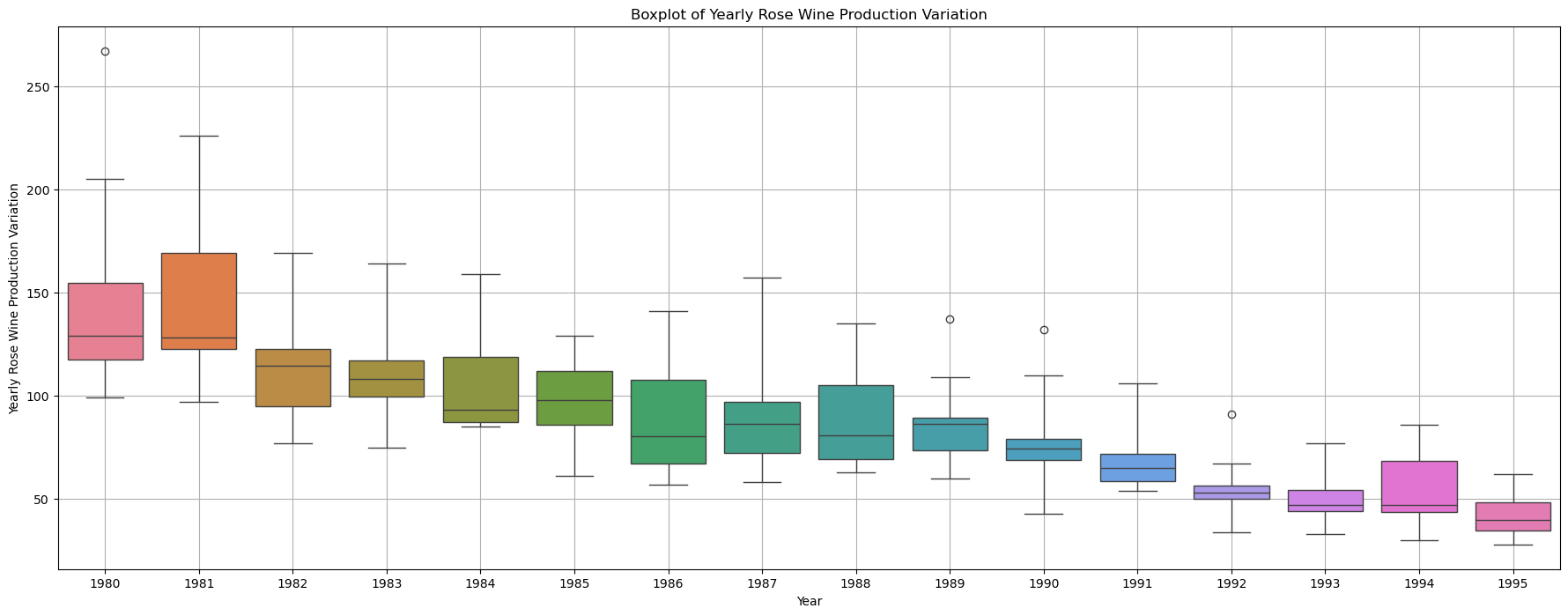


Figure 7: Yearly Production of Rose Wine

* Rose wine has mostly a downward sales trend.
* The highest sales for Rose wine can be observed in 1981 and the lowest sales in 1994.

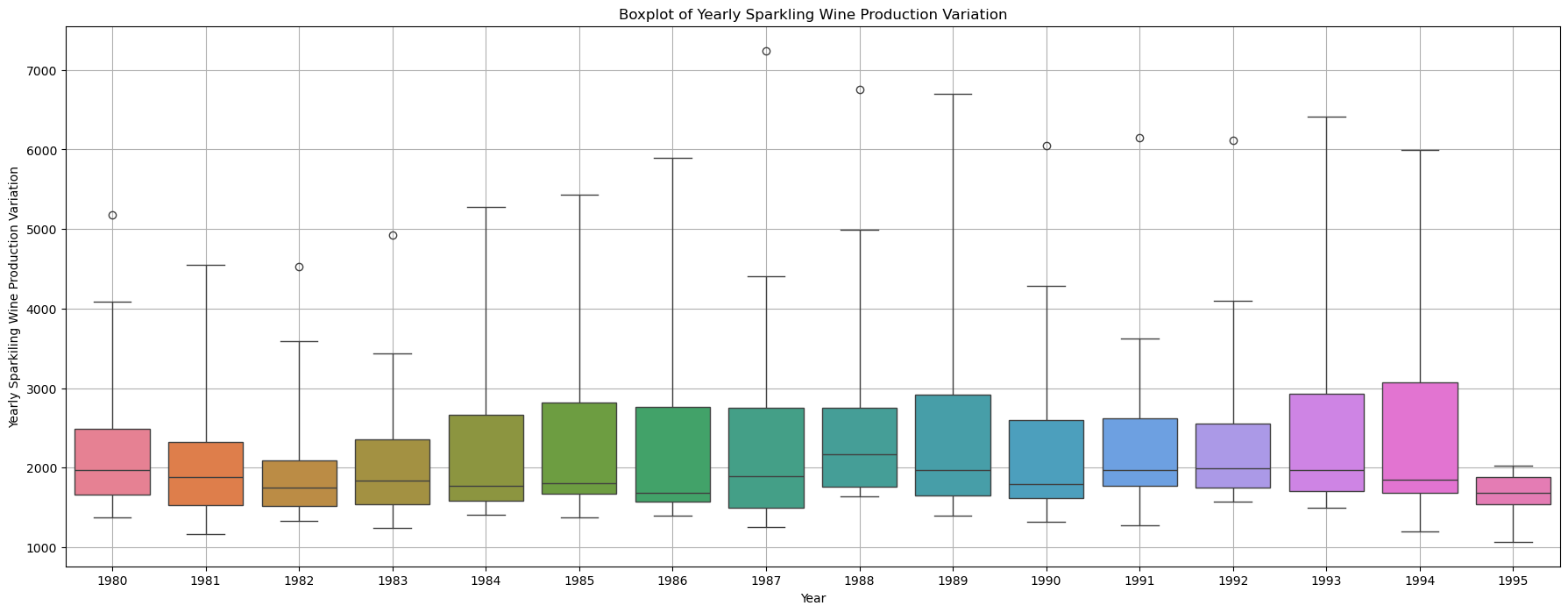


Figure 8: Yearly Production of Sparkling Wine

* Sparkling Wine sales have a variation each year.
* The years 1985 and 1986 seem to be the years with the least variation, so the 2 years show certain consistency in terms of sales.
* The highest sales for Sparkling Wine seems to happen in the year 1994 and the lowest in the year 1982.
* There is clear skewness that can be observed for Sparkling wine sales for all the years, except maybe in 1981.
* There are outliers in the yearly sales data, however as it is a Time Series, we can ignore the outlier data.
* The Sparkling wine sales appear to be going down from the year 1980 and have started increasing from the year 1983.
* The variation in Sparkling Wine sales seem to be increasing for the period 1983-1986, while the highest variation in Sparkling wine sales is in the year 1994.

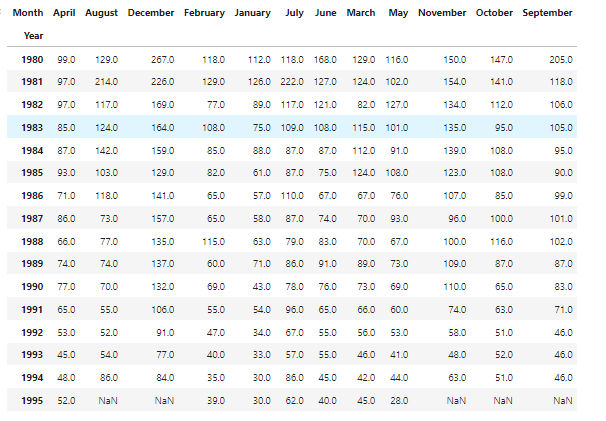


Figure 9: Monthly Sales Across years for Rose Wine

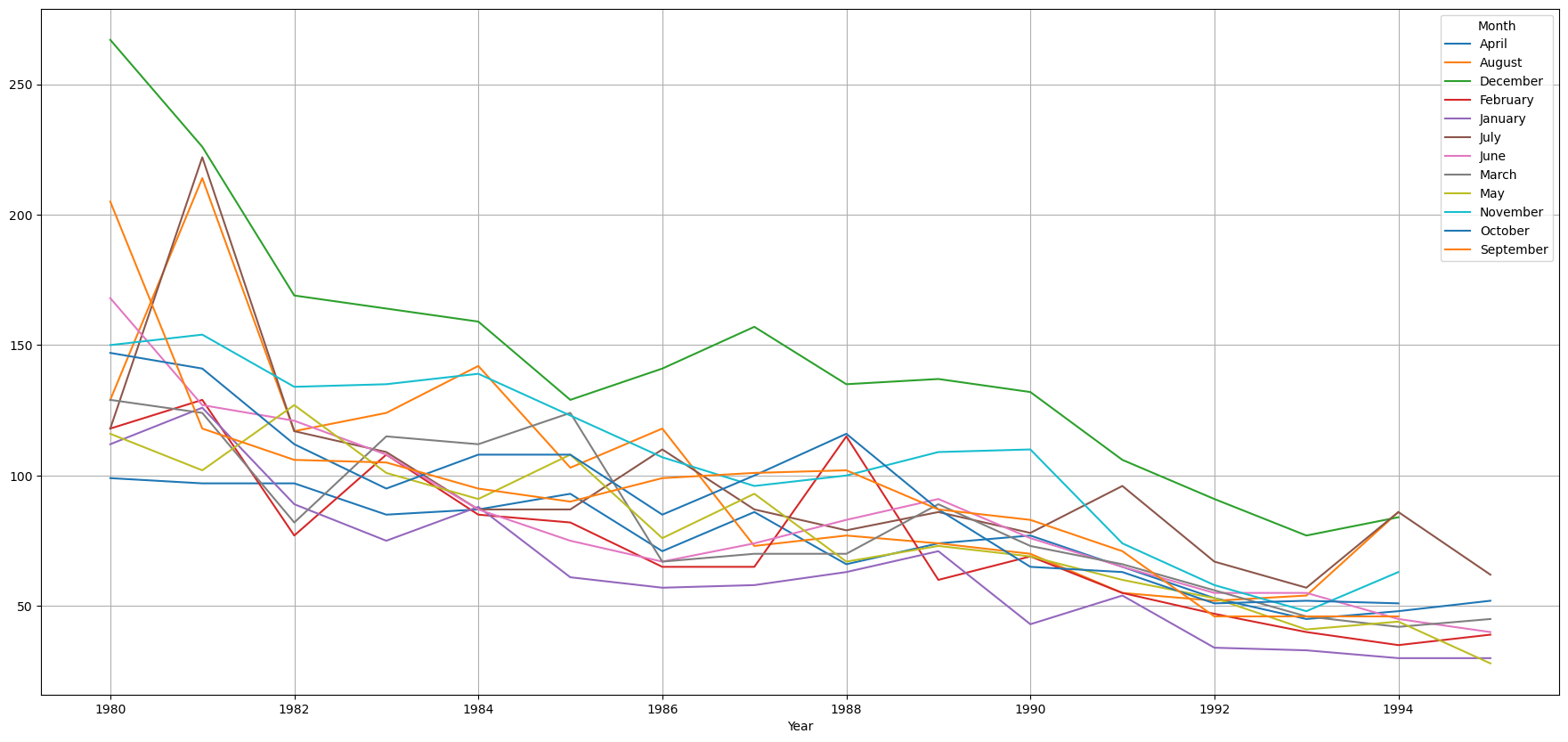


Figure 10: Monthly Production of Rose Wine

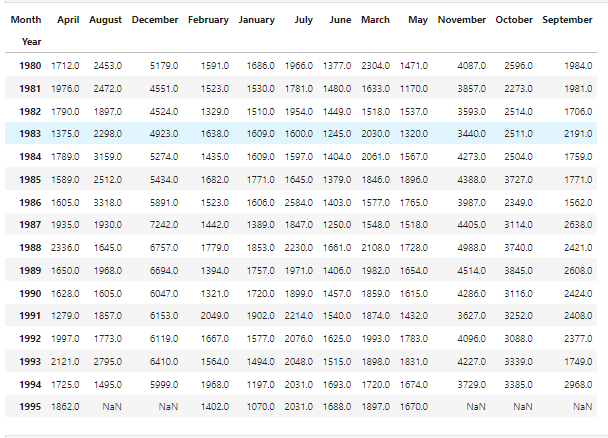


Figure 11: Monthly Sales Across years for Sparkling Wine

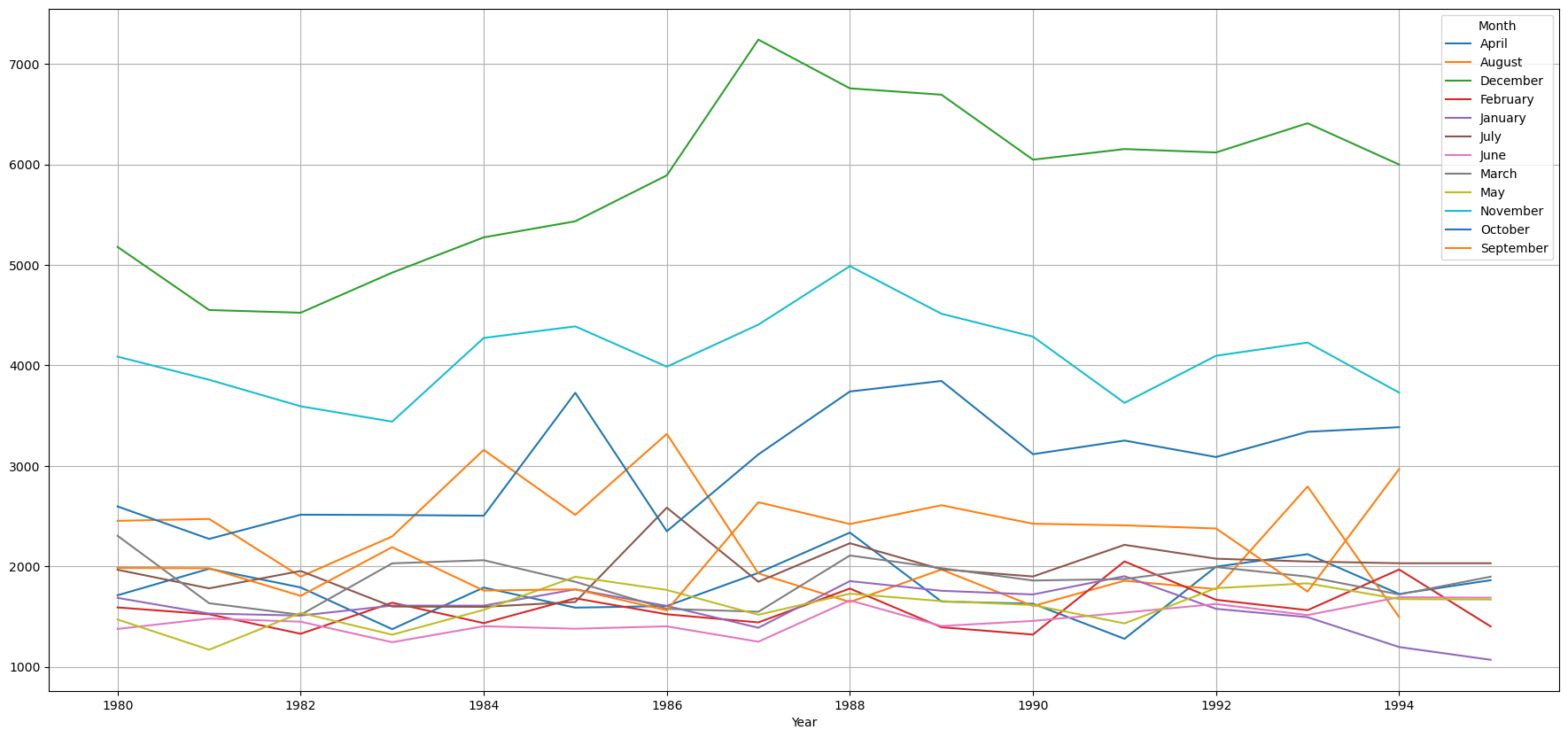


Figure 12: Monthly Production of Sparkling Wine

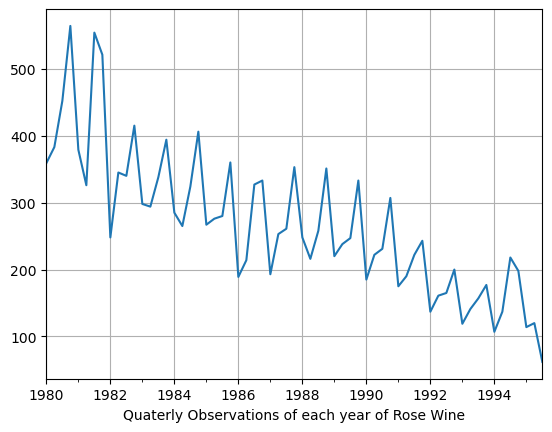


Figure 13: Quarterly Production of Rose Wine

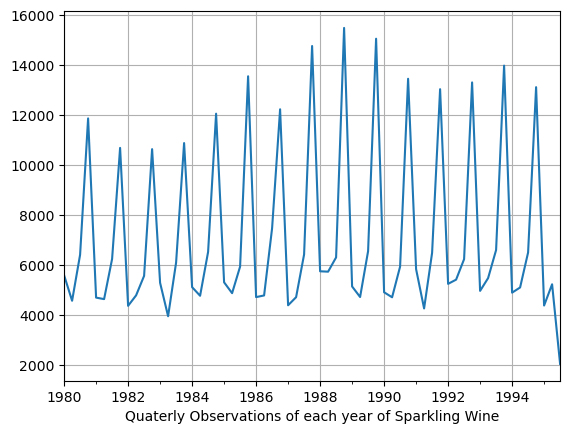


Figure 14: Quarterly Production of Sparkling Wine

* There is a seasonality element visible in both the Rose as well as Sparkling Wine time series datasets.
* The Sparkling wine seem to have a higher seasonality element as compared to Rose Wine.
* The sales have an increasing sales trend in the last quarter of the year, with Sparkling wine observing a steeper rise in sales during last quarter.
* The sales for Rose wine seems to pick up from January month and is more or less consistent till June, observes some stagnancy till September month and then starts to pick up again from October (i.e. last quarter); while for Sparkling wine, the sales is relatively low in first two quarters, slowly picks up pace during the third quarter and goes on a rise till the end of the year.
* Monthly sales data for both the type of wines shows skewness without much exceptions.

### **Decomposition:**

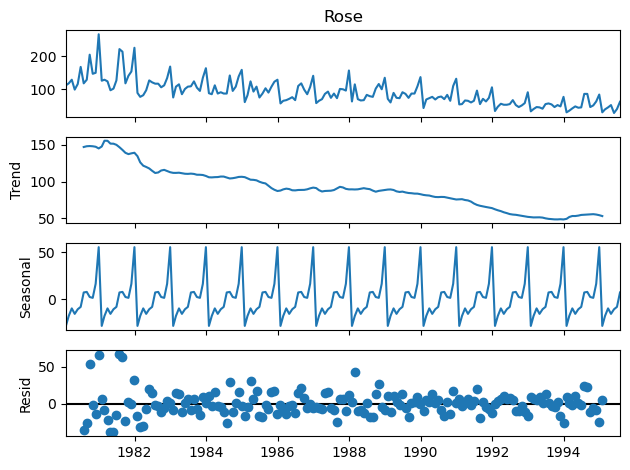


Figure 14: Additive Decomposition of Rose Wine

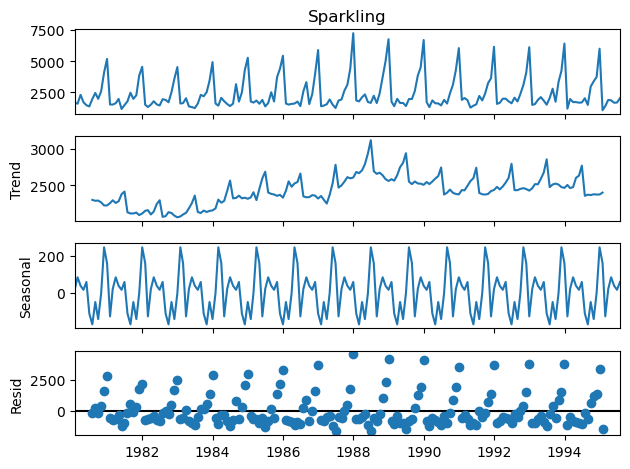


Figure 15: Additive Decomposition of Sparkling Wine

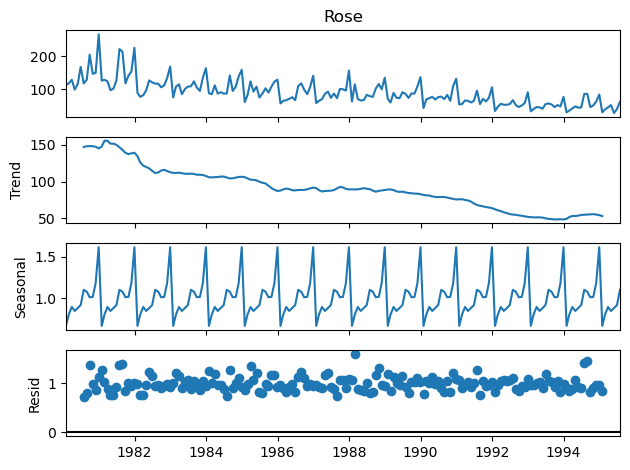


Figure 16: Multiplicative Decomposition of Rose Wine

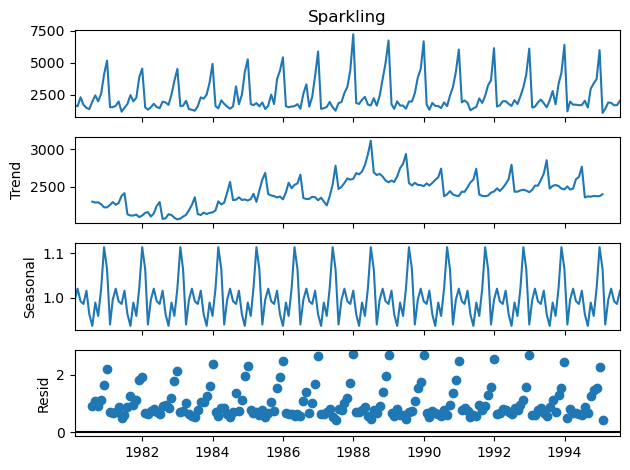


Figure 17: Multiplicative Decomposition of Sparkling Wine

* Here by just observing the Residual patterns of Additive and Multiplicative models of Rose and Sparkling datasets. It seems that
  + Rose is Multiplicative
  + Sparkling is Additive
* The plots above clearly indicate that the Wine sales are unstable and not uniform, and they have an apparent seasonality trend. Moreover, the seasonal variation seems to be more in the case of Sparkling wine as compared to the Rose wine; while the sales variation seems to be more in case of Rose wine as compared to Sparkling wine.

### **Train & Test Data Split:**

Rose & Sparkling wine data has been split into test and train datasets for further model evaluations.

* Train data for both rose & sparkling wine is till 1990
* Test data for both rose & sparkling wine is from 1991

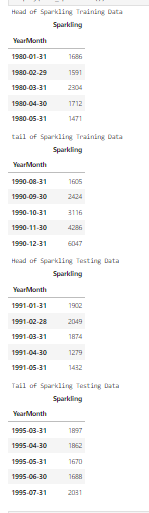
 

Figure 18: Train & Test Split of Rose Wine Figure 19: Train & Test Split of Sparkling Wine

* Shape of Train Rose & Sparkling wine dataset : (132, 1)
* Shape of Test Rose & Sparkling wine dataset : (55, 1)

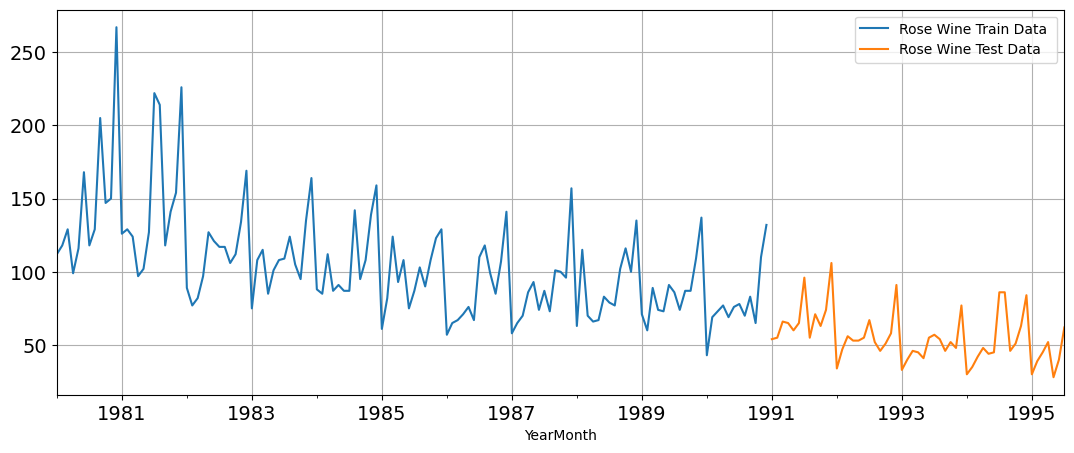


Figure 20: Train & Test Set of Rose Wine

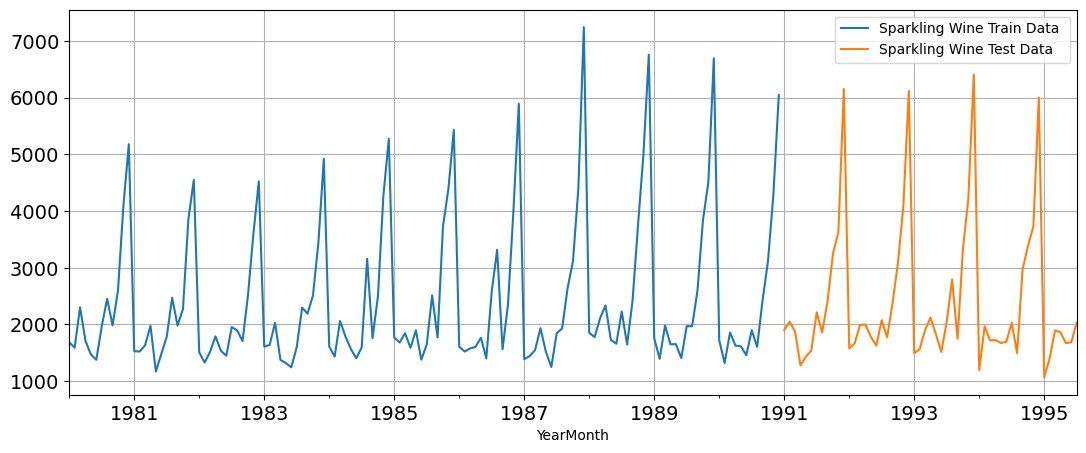


Figure 21: Train & Test Set of Sparkling Wine

### **Model Building:**

### **Linear Regression:**

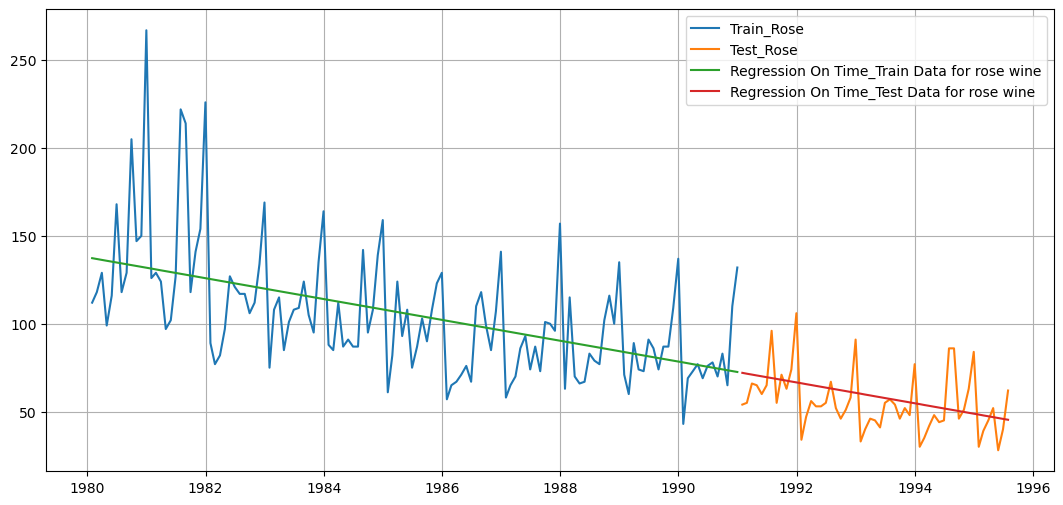


Figure 22: Linear Regression of Rose Wine

Linear Regression RSME on the Train Data of Rose Wine = 30.72

Linear Regression RSME on the Test Data of Rose Wine = 16.63

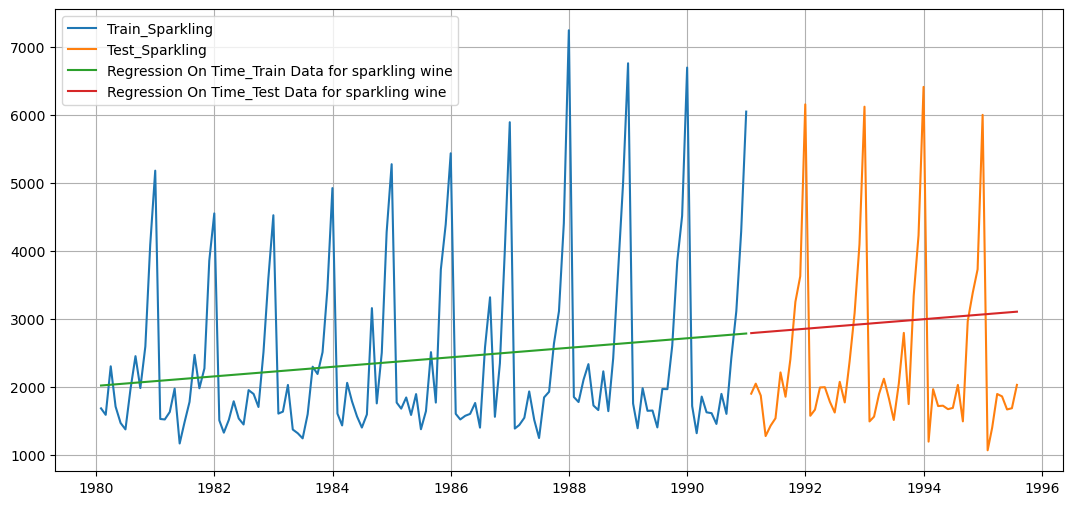


Figure 23: Linear Regression of Sparkling Wine

Linear Regression RSME on the Train Data of Sparkling Wine = 1279.32

Linear Regression RSME on the Test Data of Sparkling Wine = 1389.14

**Simple Average:**

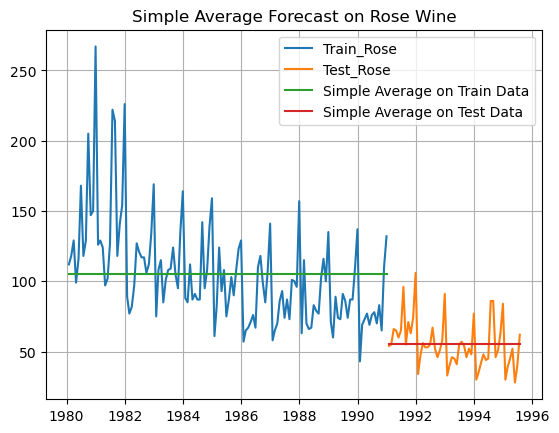


Figure 24: Simple Average of Rose Wine

RSME Simple Average forecast on the Train Data for Rose Wine = 36.034

RSME Simple Average forecast on the Test Data for Rose Wine = 16.770

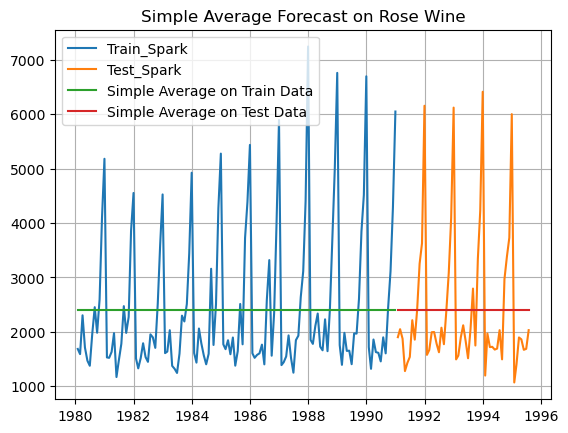


Figure 25: Simple Average of Sparkling Wine

RMSE Simple Average forecast on the Train Data for Sparkling Wine = 1298.484

RMSE Simple Average forecast on the Test Data for Sparkling Data = 1275.073

### **Moving Average:**

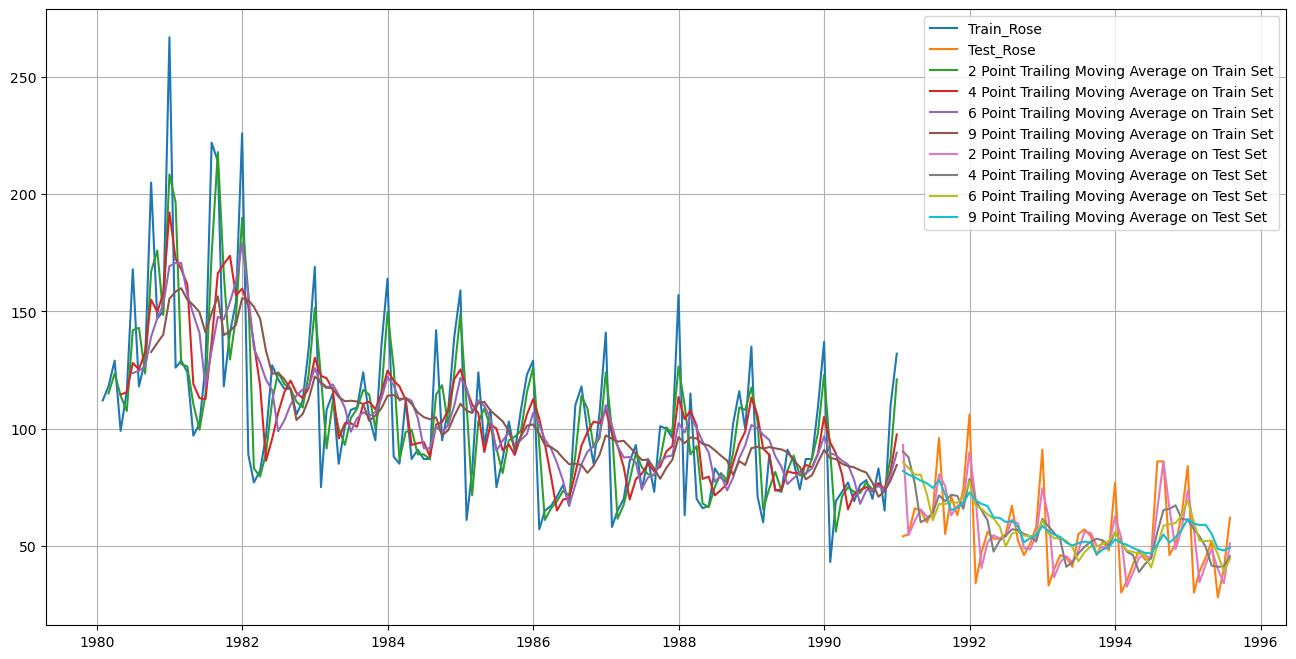


Figure 26: Moving Average of Rose Wine

For 2 point Moving Average Model forecast on the Test Data, RMSE is 12.159

For 4 point Moving Average Model forecast on the Test Data, RMSE is 15.572

For 6 point Moving Average Model forecast on the Test Data, RMSE is 15.688

For 9 point Moving Average Model forecast on the Test Data, RMSE is 16.161

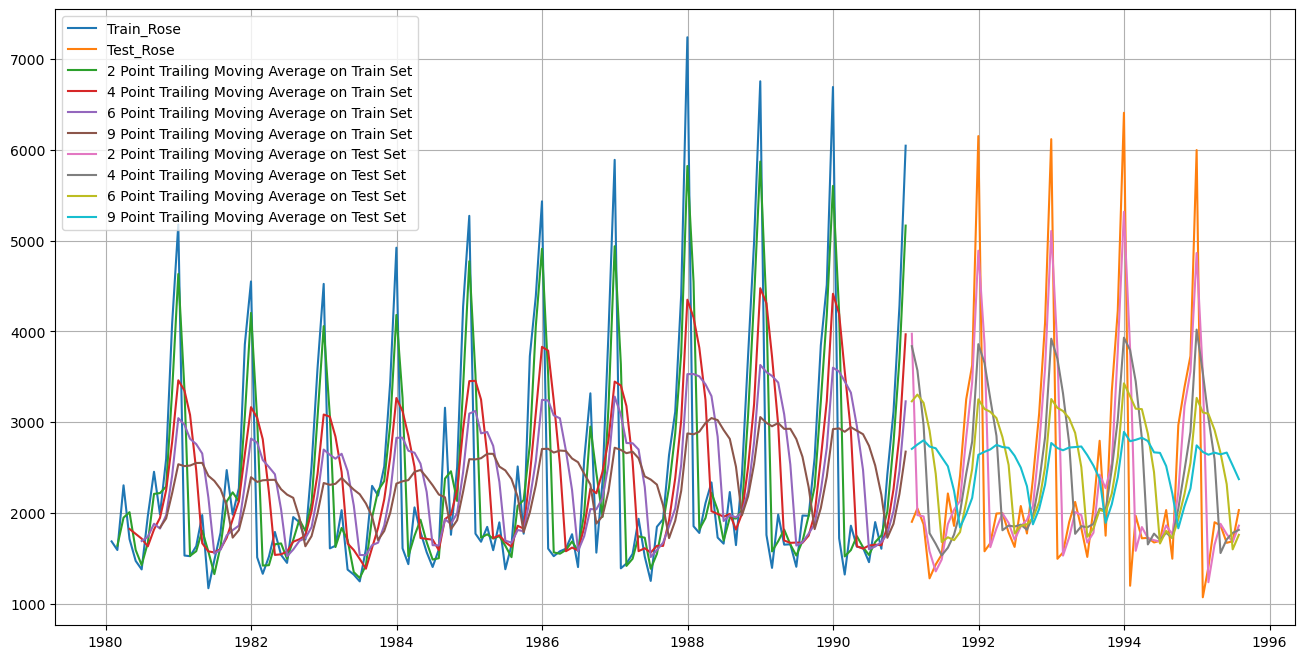


Figure 27: Moving Average of Sparkling Wine

For 2 point Moving Average Model forecast on the Test Data, RMSE is 813.401

For 4 point Moving Average Model forecast on the Test Data, RMSE is 1156.590

For 6 point Moving Average Model forecast on the Test Data, RMSE is 1283.928

For 9 point Moving Average Model forecast on the Test Data, RMSE is 1346.279

* We have built 4 models till now for both Rose and Sparkling Wine datasets
* We fitted various models to the Train split and Tested it on Test split. Accuracy metrics used is Root Mean Squared Error (RMSE) on Test data
* **Model 1 - Linear Regression**
  + We regressed variables ‘Rose’ and ‘Sparkling’ against their individual time instances
  + We modified the datasets and tagged individual sales to their time instances
  + TEST RMSE ROSE = 16.63 | TEST RMSE SPARKLING = 1389.14
* **Model 2 - Simple Average**
  + All future predictions are the same as the simple average of all data till today
  + TEST RMSE ROSE = 16.77 | TEST RMSE SPARKLING = 1275.073
* **Model 3 - Moving Average (MA)**
  + We calculate rolling means (Moving averages) over different intervals for the whole train data
  + 2 Point Moving Average:

TEST RMSE ROSE = 12.159 | TEST RMSE SPARKLING = 813.401

* + 4 Point Moving Average:

TEST RMSE ROSE = 15.572 | TEST RMSE SPARKLING = 1156.590

* + 6 Point Moving Average:

TEST RMSE ROSE = 15.688 | TEST RMSE SPARKLING = 1283.928

* + 9 Point Moving Average:

TEST RMSE ROSE = 16.161 | TEST RMSE SPARKLING = 1346.279\

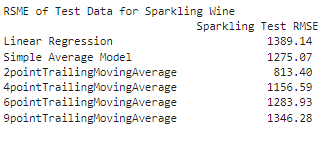
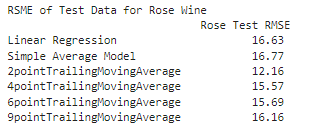


Figure 28: Test RSME of all 3 models of Rose & Sparkling Wine

* Till now, Best Model which gives lowest RMSE score for both Rose and Sparkling is 2 Point Moving Average Model
* We’ll continue to forecast using Exponential Smoothing Models for both datasets of Rose and Sparkling Wine Sales
* We’ll build following Exponential Smoothing Models –
  + Simple Exponential Smoothing
  + Double Exponential Smoothing (Holt’s Model)
  + Triple Exponential Smoothing (Holt-Winter Model)

### **Single Exponential Smoothing:**

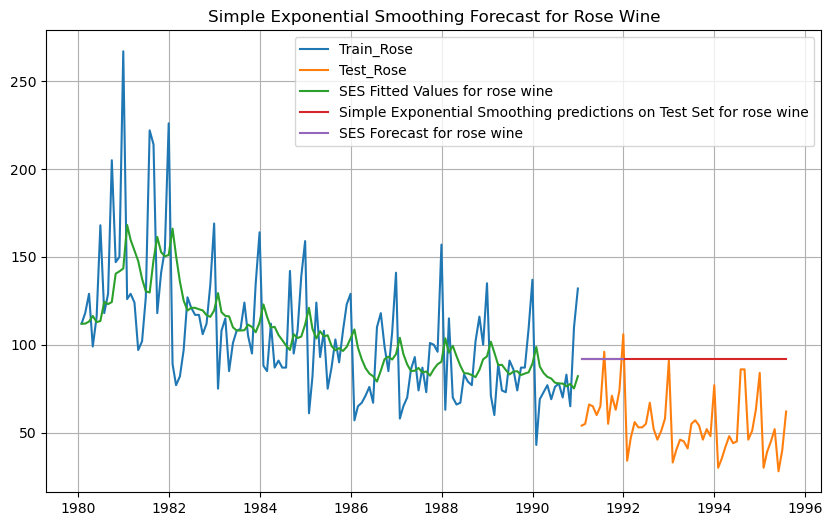


Figure 29: Simple Exponential Smoothing of Rose Wine

Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 40.413

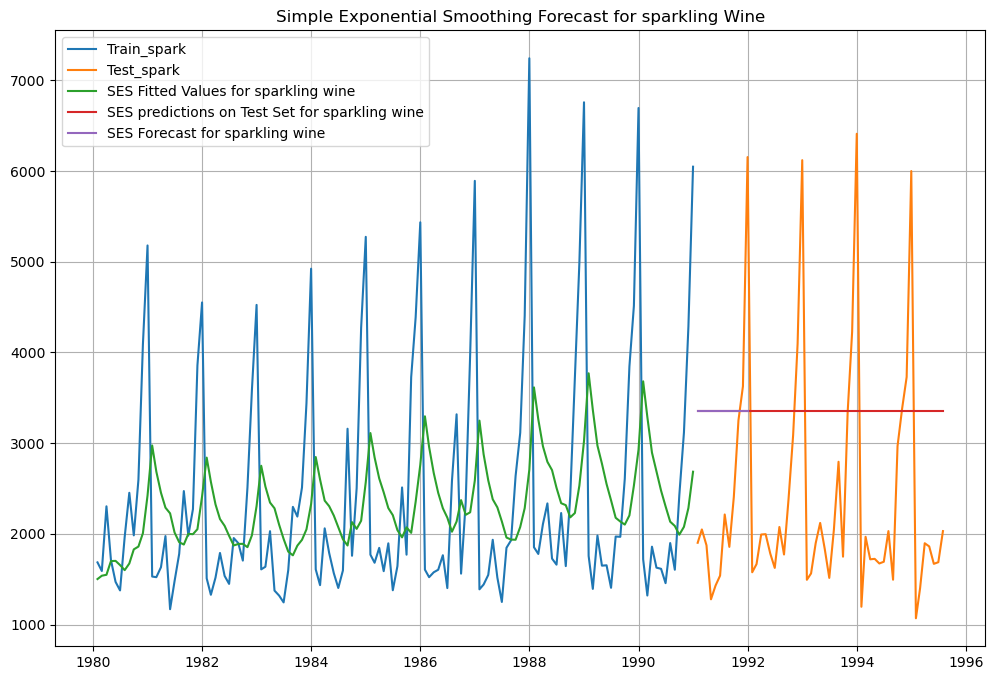


Figure 30: Simple Exponential Smoothing of Sparkling Wine

Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1595.207

Best Model till now for Rose and Sparkling is 2 Point Moving Average Model

### **Double Exponential Smoothing (Holt’s Model):**

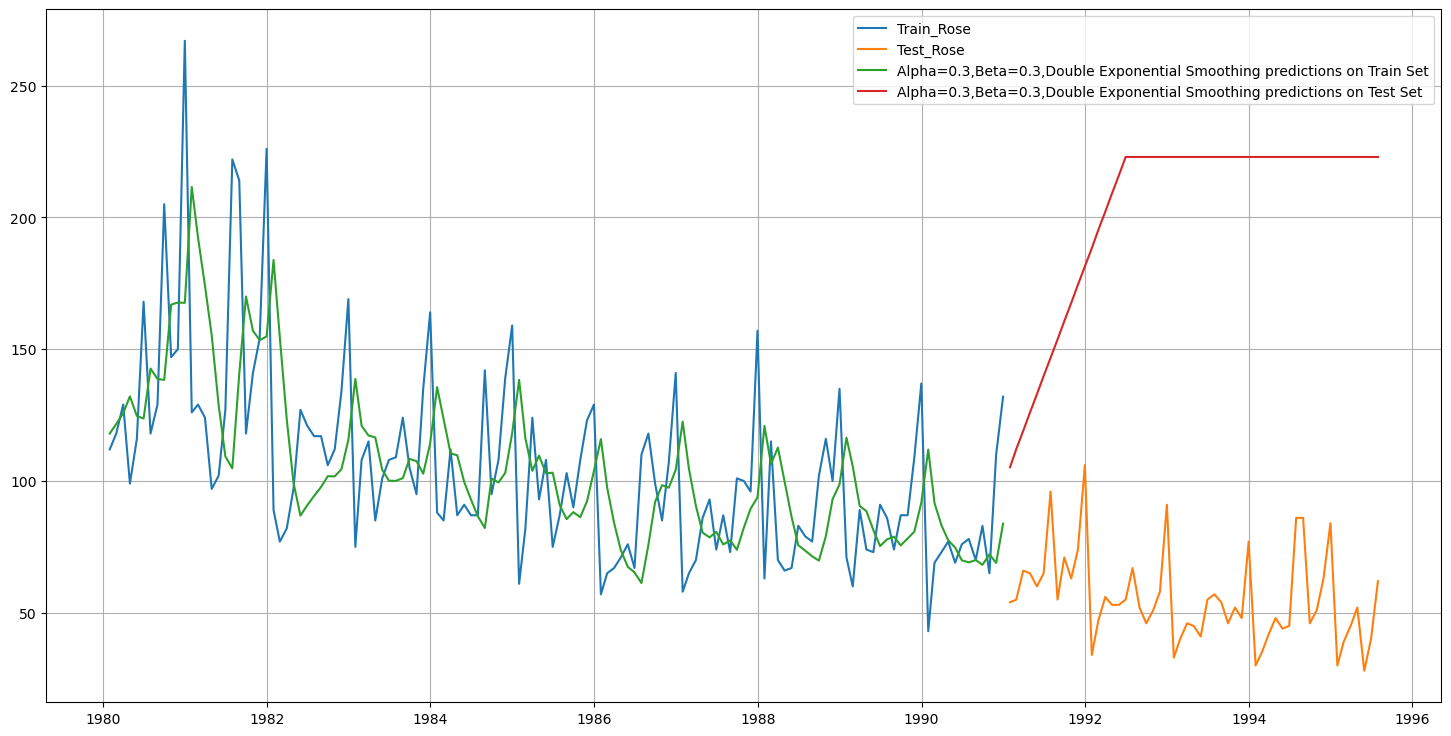


Figure 31: Double Exponential Smoothing of Rose Wine

Double Exponential Smoothing Model forecast on the Test Data, RMSE is 154.323

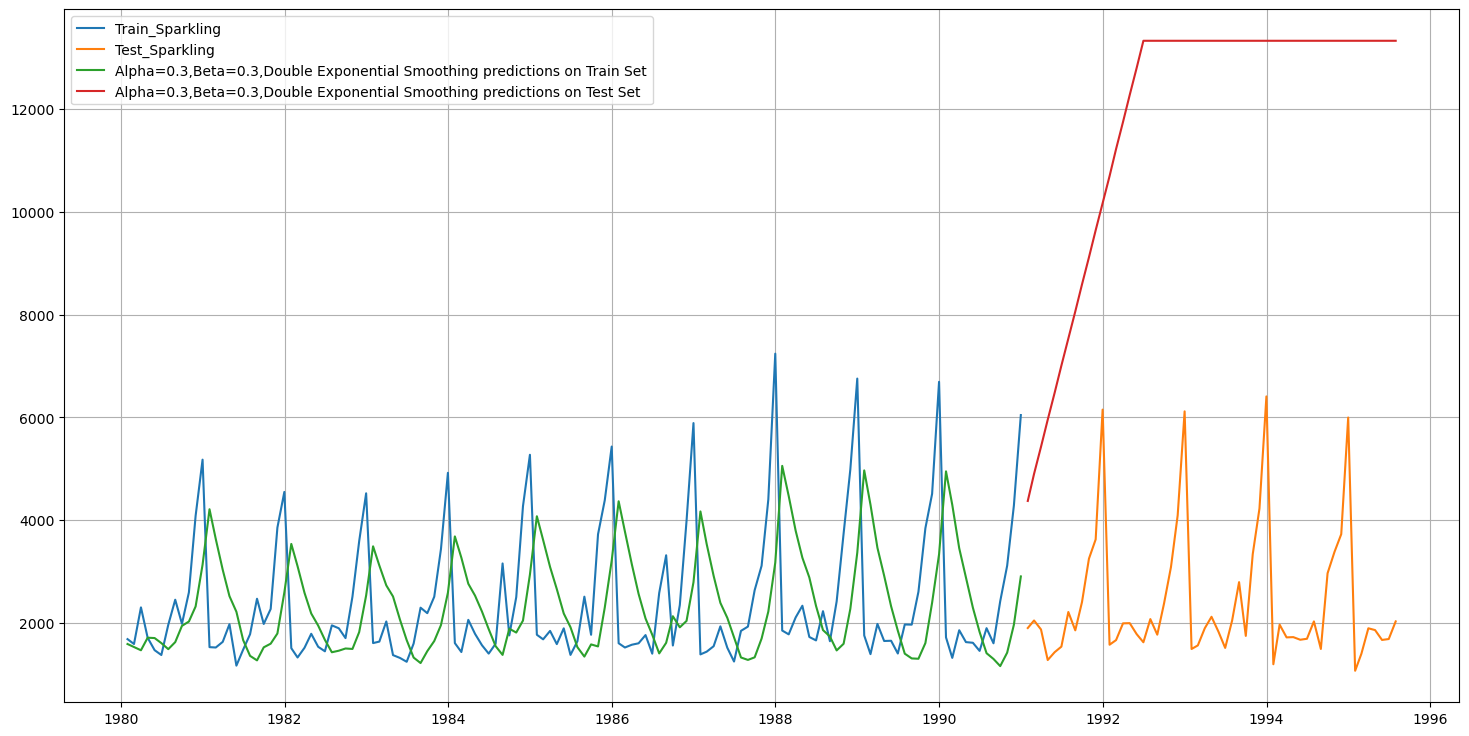


Figure 32: Double Exponential Smoothing of Sparkling Wine

Double Exponential Smoothing Model forecast on the Test Data, RMSE is 9866.648

* In Rose - DES has picked up the trend well. DES seems to perform better than SES here
* In Sparkling - DES shows a non-existent trend. DES does not perform well here
* Best Model till now for Rose and Sparkling is 2 Point Moving Average Model

### **Triple Exponential Smoothing (Holt - Winter Model):**

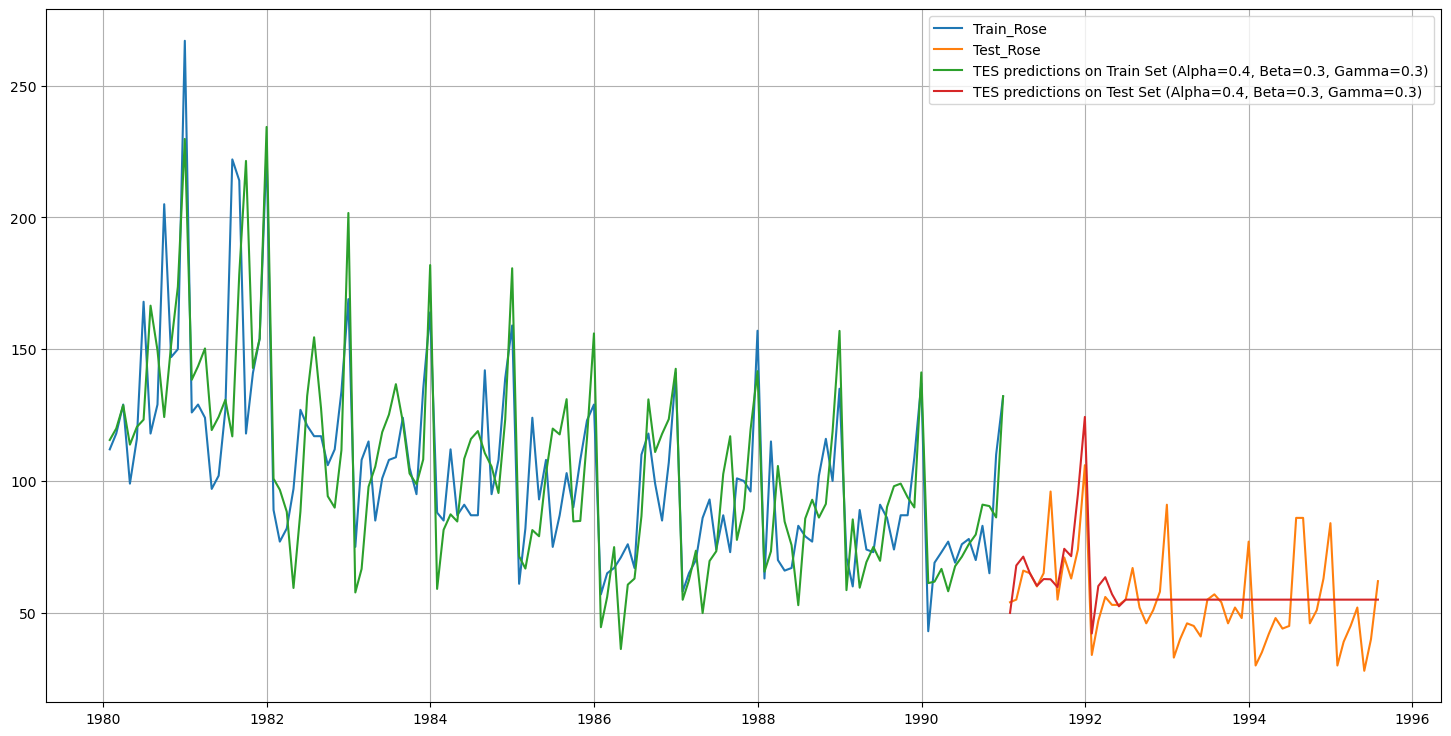


Figure 33: Triple Exponential Smoothing of Rose Wine

Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 25.44

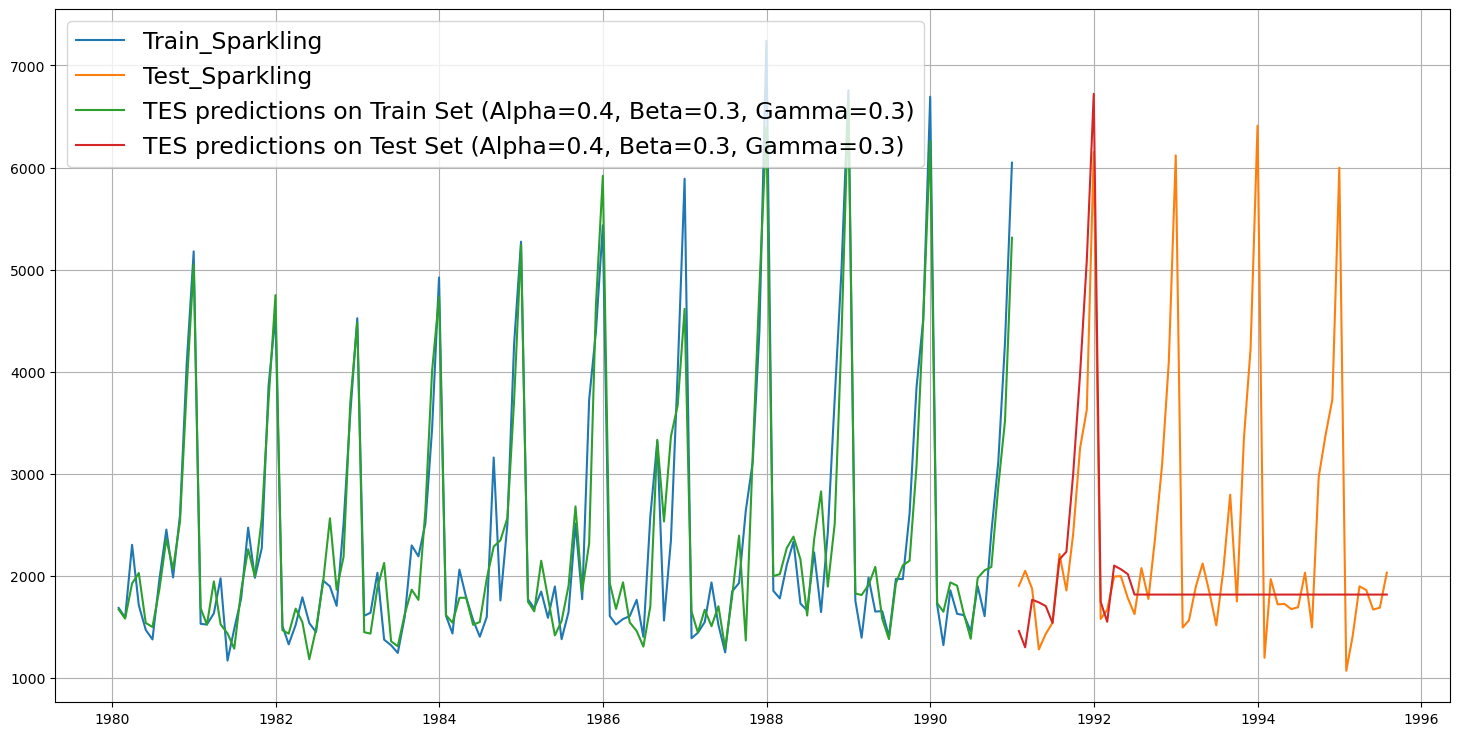


Figure 34: Triple Exponential Smoothing of Sparkling Wine

Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 2766.83

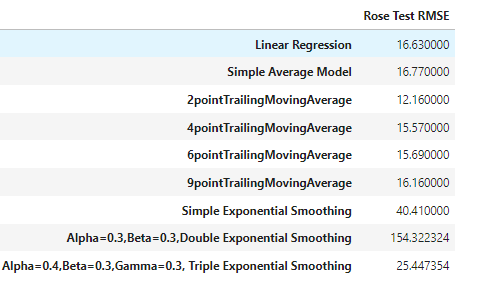


Figure 34: Test RSME of all models of Rose Wine

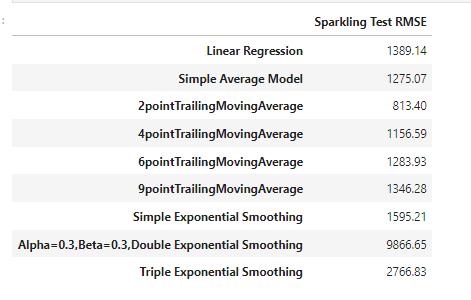


Figure 35: Test RSME of all models of Sparkling Wine

Best Model till now for Rose and Sparkling is 2 Point Moving Average Model

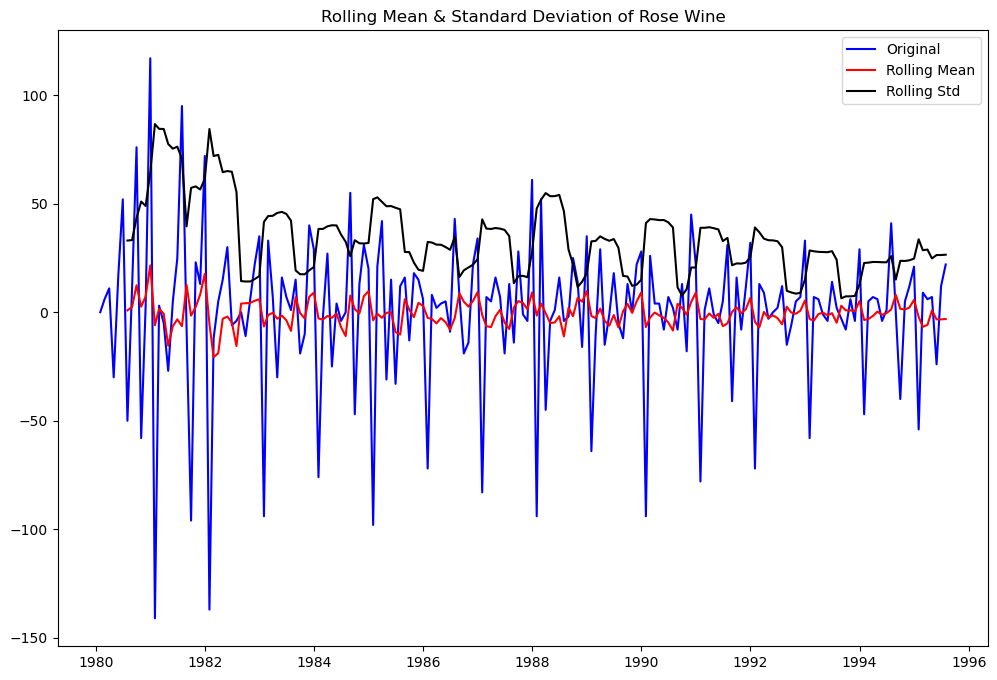
### **Stationary Test:**

Dickey-Fuller Test - Dicky Fuller Test on the timeseries is run to check for stationarity of data.

* Null Hypothesis H0: Time Series is non-stationary.
* Alternate Hypothesis 𝑯𝒂: Time Series is stationary.
* So in ADF Test, if p-value < 0.05 - We reject the Null Hypothesis and hence conclude that given Time Series is Stationary
* So in ADF Test, if p-value > 0.05 - We fail to reject the Null Hypothesis and hence conclude that given Time Series is Not Stationary
* If Time Series is not Stationary then we apply one level of differencing and check for Stationarity again.
* Again, if the Time Series is still not Stationary, we apply one more level of differencing and check for Stationarity again
* Generally, with max 2 levels of differencing, Time Series becomes Stationary • Once the Time Series is Stationary then we are ready to apply ARIMA / SARIMA models

**Stationarity of Rose Wine Dataset:**

* Augmented Dicky-Fuller Test was applied to the whole Rose dataset
  + We found, p-value = 0.3166
  + Here, p-value > 0.05
  + We fail to reject the Null Hypothesis and hence conclude that Rose Wine Time Series is Not Stationary
  + We take 1 level of differencing and check again for Stationarity
  + Now, p-value = 3.755773e-12
  + Now, p-value < 0.05
  + Now, we reject the Null Hypothesis and conclude that Rose Time Series is Stationary with a lag of 1



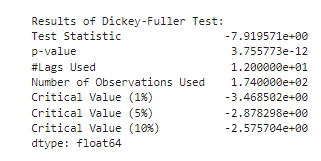
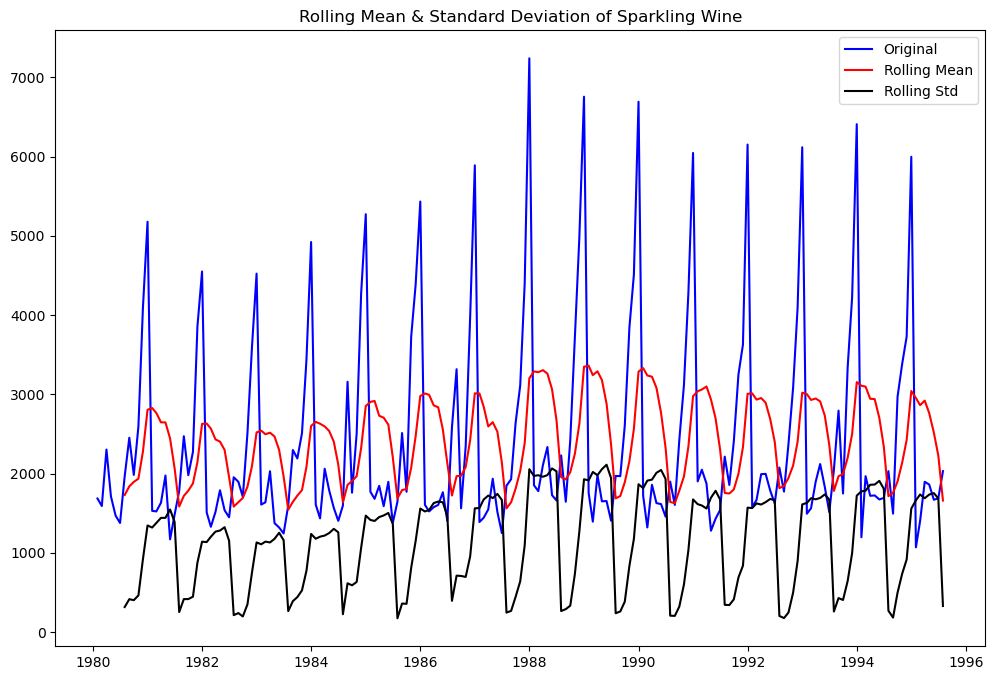


Figure 36: Stationary of Rose Wine

**Stationarity of Sparkling Wine Dataset:**

* Augmented Dicky-Fuller Test was applied to the whole Sparkling dataset
  + We found, p-value = 0.60106
  + Here, p-value > 0.05
  + We fail to reject the Null Hypothesis and hence conclude that Sparkling Wine Time Series is Not Stationary
  + We take 1 level of differencing and check again for Stationarity
  + Now, p-value = 0.000000
  + Now, p-value < 0.05
  + Now, we reject the Null Hypothesis and conclude that Sparkling Time Series is Stationary with a lag of 1



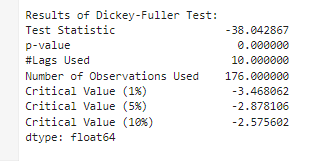


Figure 37: Stationary of Sparkling Wine

### **Auto ARIMA:**

* We create a grid of all possible combinations of (p, d, q)
* Range of p = Range of q = 0 to 3, Constant d = 1

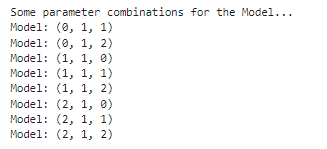


Figure 38: Model Combinations of Rose & Sparkling Wine

* We fit ARIMA models to each of these combinations for both datasets
* We choose the combination with the least Akaike Information Criteria (AIC)
* We fit ARIMA to this combination of (p, d, q) to the Train set and forecast on the Test set
* Finally, we check the accuracy of this model by checking RMSE of Test set

### **Manual ARIMA:**

### **Auto SARIMA:**

### **Manual SARIMA:**