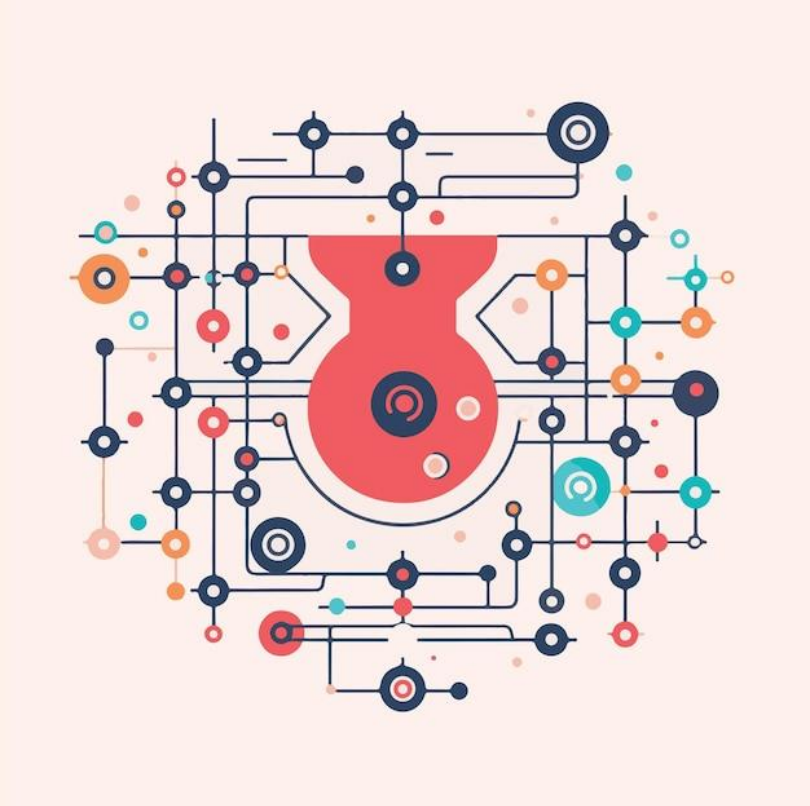


TIME SERIES PROJECT

By Kurt Warren Mario Gilby On September 22nd 2024

Submitted to

Great Learning
POWER AHEAD

As a part of the requirements for completion of PGP-DSBA offered in affiliation with

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Table of Contents

List of Figures

No table of figures entries found.

List of Tables

No table of figures entries found.

List of Equations

No table of figures entries found.

Problem 1

Overview

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, analyse 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 analyse 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.

Questions Asked

Define the problem and perform Exploratory Data Analysis.

Problem definition

We have been given wine sales data for Rose and Sparkling Wine, for one company from 1980 to 1995, we need to look at the historical trends and predict the forecast for future sales for both these varieties of wine.

Read the data as an appropriate time series data

- The data is given to us monthly, from 1980 to 1995.

```
Data Loaded:
            Sparkling
YearMonth
1980-01-01       1686
1980-02-01       1591
1980-03-01       2304
1980-04-01       1712
1980-05-01       1471
Data Loaded:
             Rose
YearMonth
1980-01-01  112.0
1980-02-01  118.0
1980-03-01  129.0
1980-04-01   99.0
1980-05-01  116.0
```

Plotting the Data

- Observations

- Sparkling Sales are much higher than Rose, and shows seasonality but trend looks flat.

- Rose Sales looks flat but could be due to scaling on the same plot we will plot Rose Separately.

- Sperate plot of Rose shows a potential declining trend and some missing values around 1995 which we will have to treat.

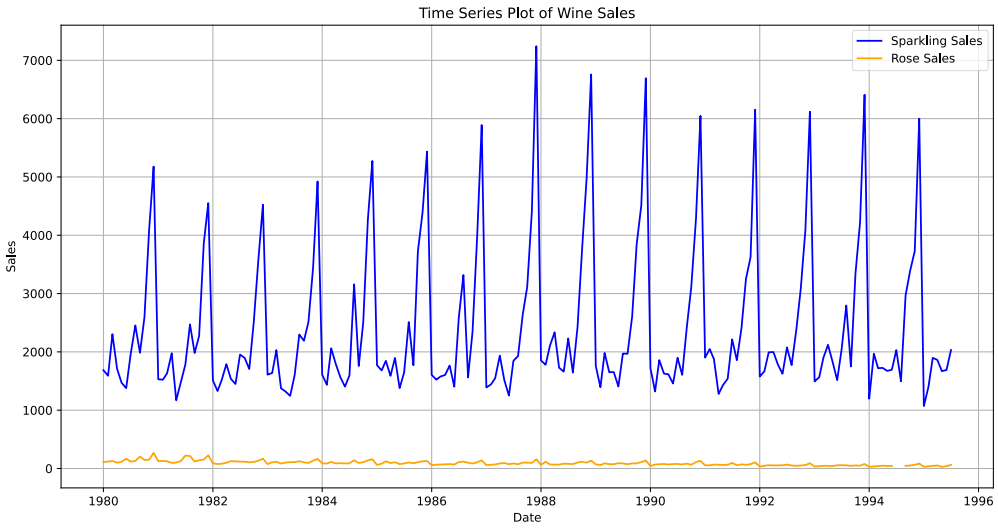


Figure 1 Time Series trend plot for wine Sales

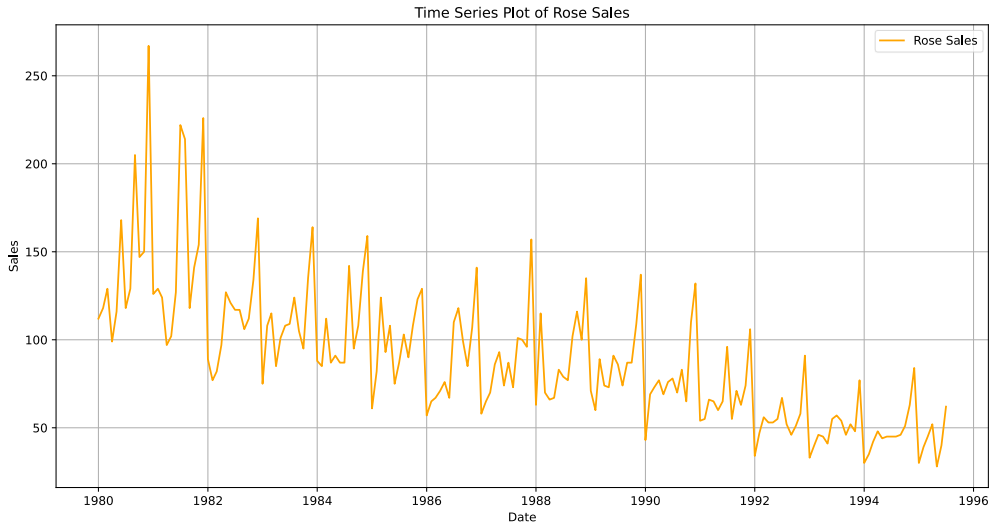


Figure 2 Time Series Plot for Rose Wine

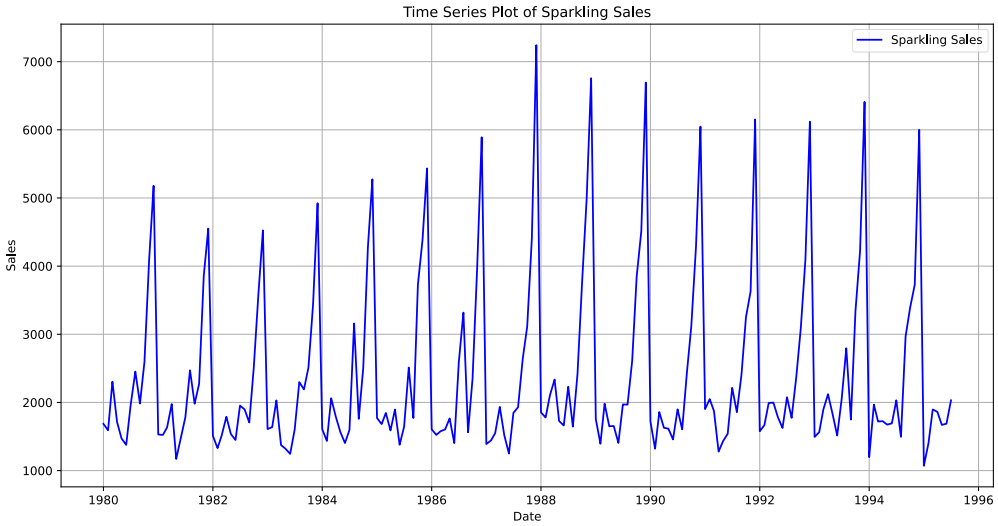


Figure 3 Time Series plot for Sparkling Wine Sales

Perform Exploratory Data Analysis (EDA)

- Let's do a describe, histogram and box plot.

- Also, box plot between years and same months access years to see if there are any trends

- Describe

  - Sparkling: has much higher sales with 1874 as the median, with long tails 1070 min and 7242 max, and possible outliers

  - Rose: has more modest sales with 86 as the median, it too seems to have long tails and possible outliers

-Histogram, Boxplots and Trends:

- Sparkling: Histogram is uniform across the months and show that the distribution is not normal, boxplot between the years shows there was a dip between 1980 to 1982, 1983 to 1985 is pretty flat with a fall in 1986 followed by a climb to 1988 and a fall again till 1990, flat from 1991 to 1993, followed but a fall in 1944 and 1995, there seems to be trend every three years, boxplot between the month across years, shows a slight rise in March and April, followed by a step rise from July to December, The line plot by month split on years shows the similar trend to the box plots but more clearly, The rolling mean and standard deviation are flat across the years.

- Rose: Histogram is uniform across the months and show that the distribution is not normal, boxplot between the years clearly shows a steady decline in sales through the years with a steep fall since 1990, boxplot between the month across years, shows sales remain flat in the mid months from April to June, with a rise in July to the end of the year and also a rise from January to March, The line plot by month split by year show no distinct trend in the recent years, The rolling mean and standard deviation are in a downward trend across the years.

```
Sparkling Summary Statistics:
         Sparkling
count   187.000000
mean   2402.417112
std    1295.111540
min    1070.000000
25%    1605.000000
50%    1874.000000
75%    2549.000000
max    7242.000000

Rose Summary Statistics:
             Rose
count  185.000000
mean    90.394595
std     39.175344
min     28.000000
25%     63.000000
50%     86.000000
75%    112.000000
max    267.000000
```

Figure 4 Describe on the two data sets

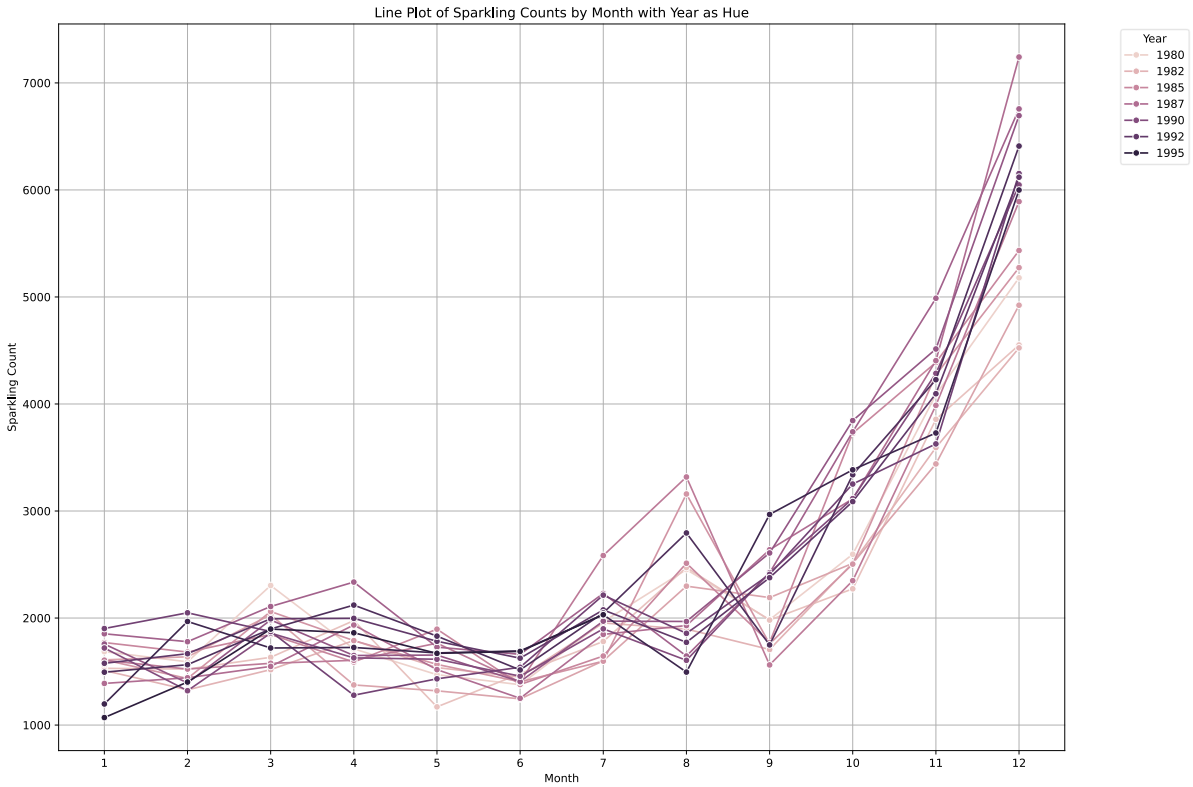


Figure 5 sparkling line plot by Sales by month split by Years

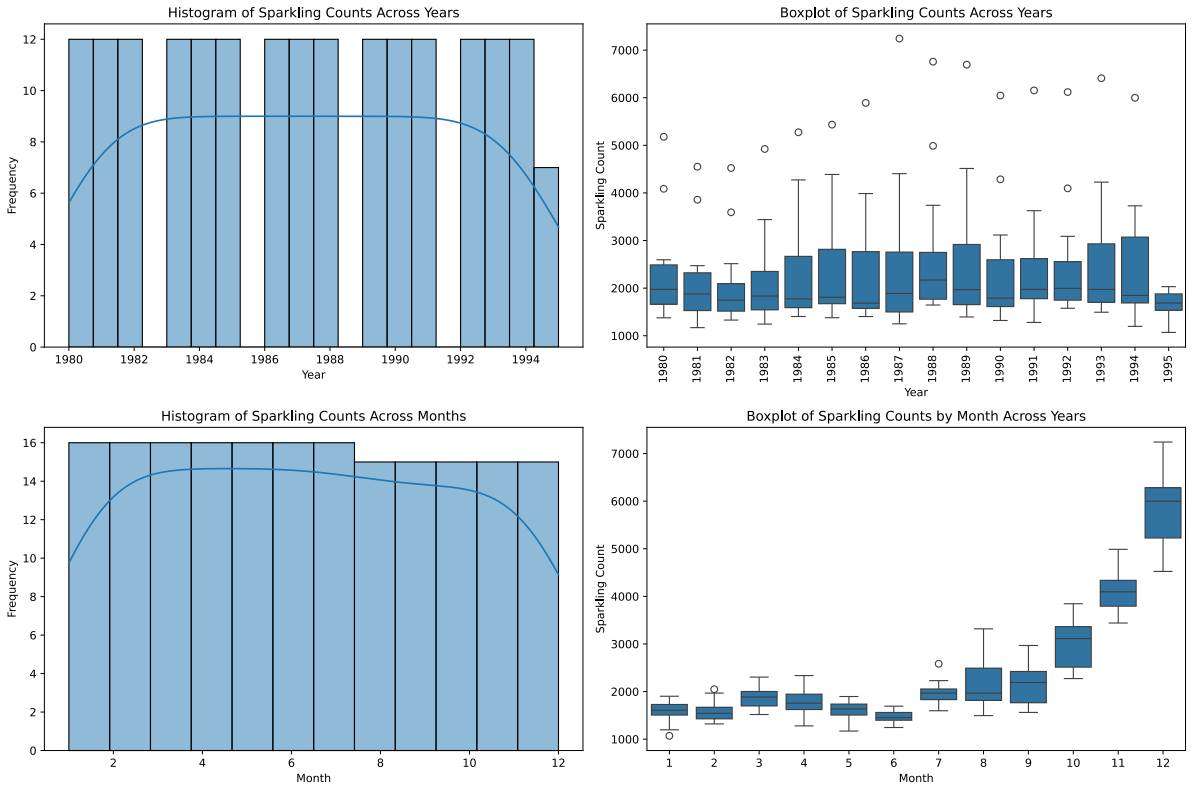


Figure 6 Sparkling wine Sales Histogram and box plots

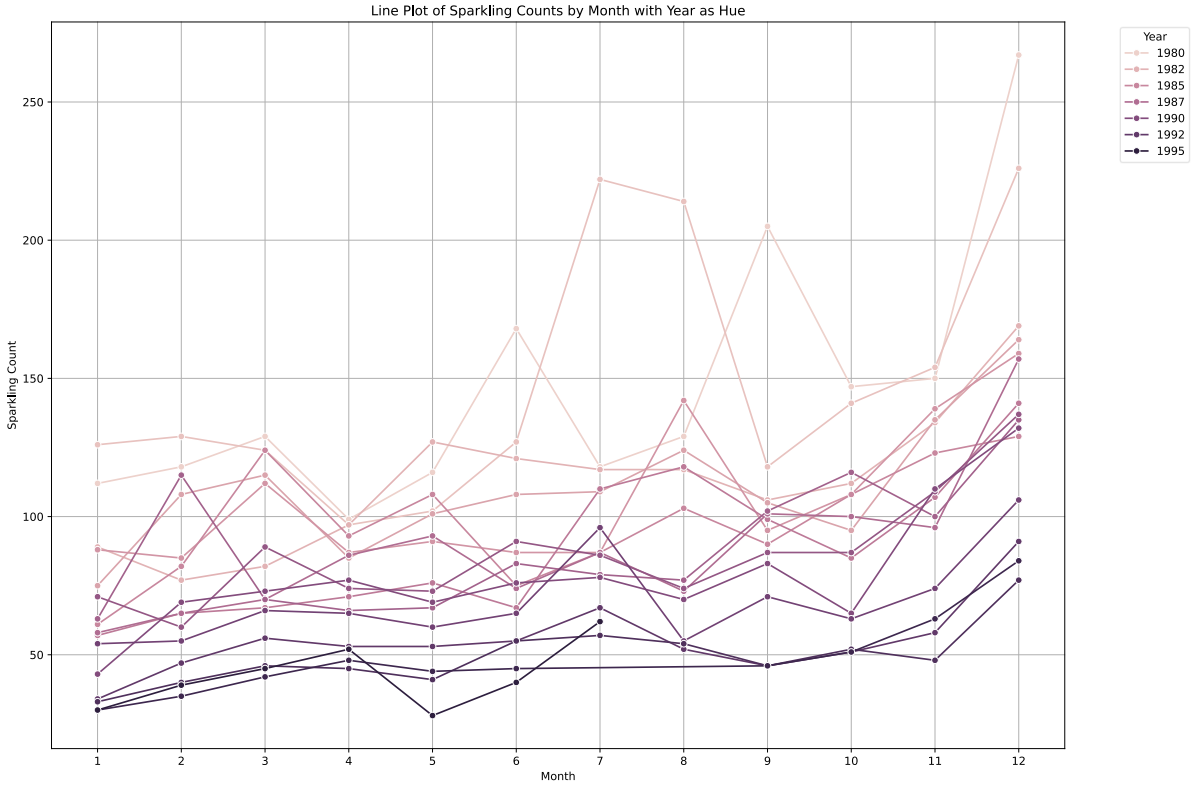


Figure 7 Rose Wine Sales Trend by Months Split by Years

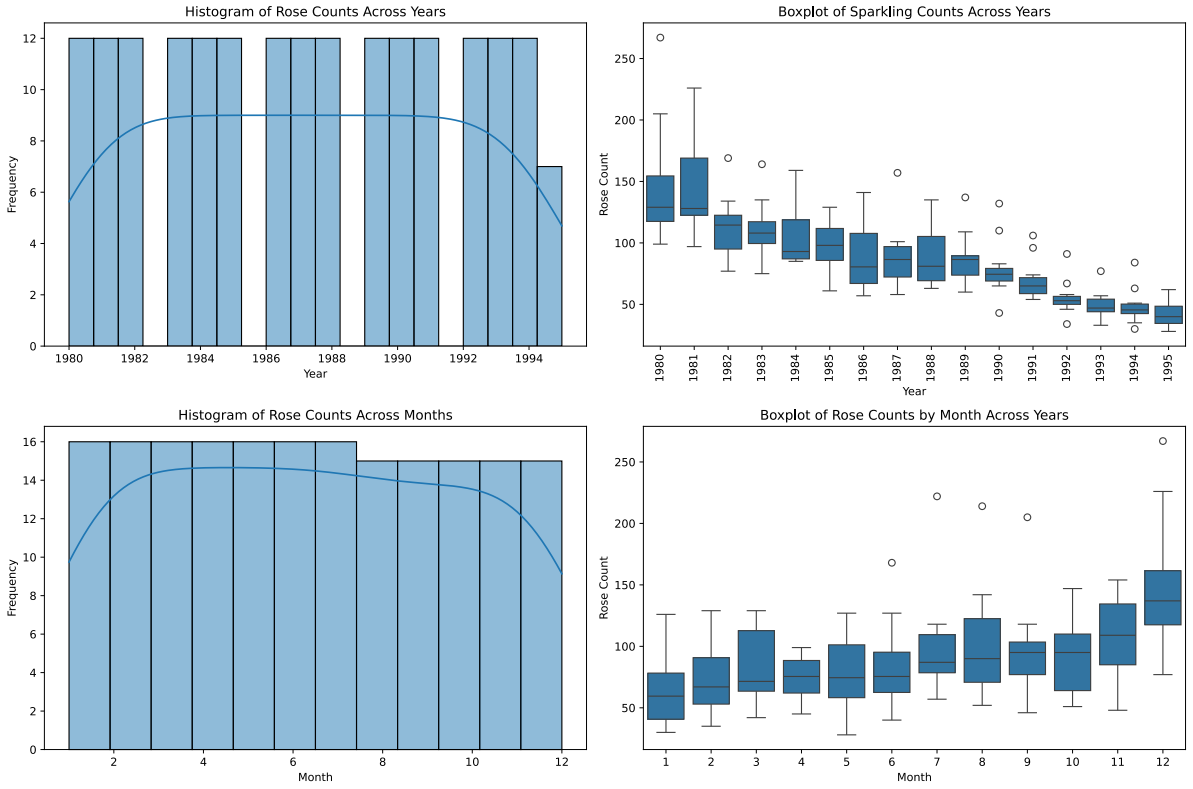


Figure 8 Rose Wine Sales Histogram and Box plots

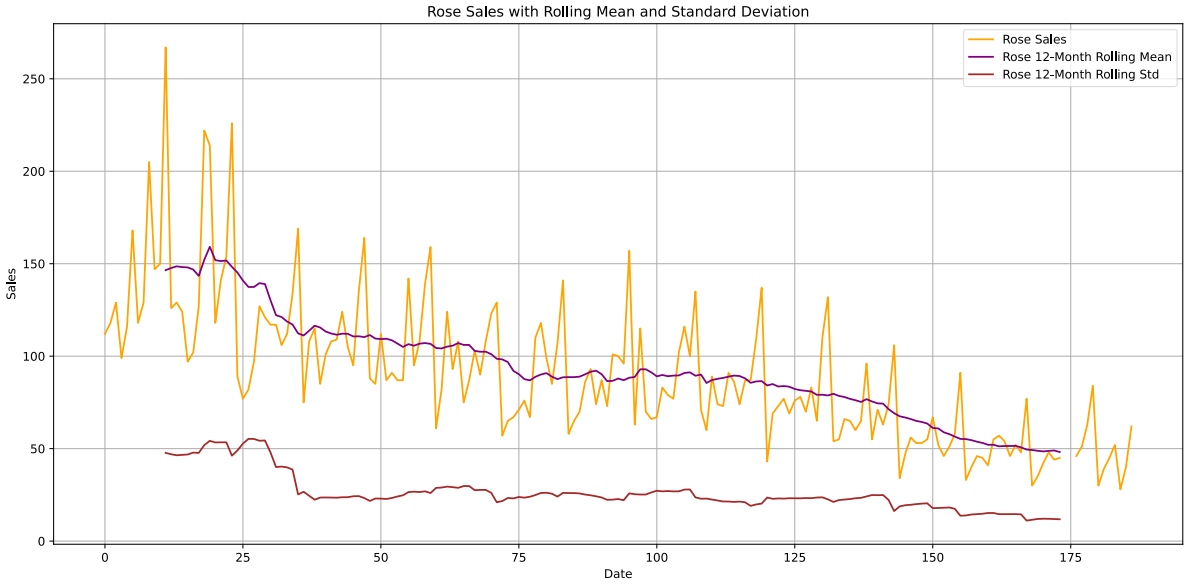


Figure 9 Rose Rolling Means and Standard Deviation

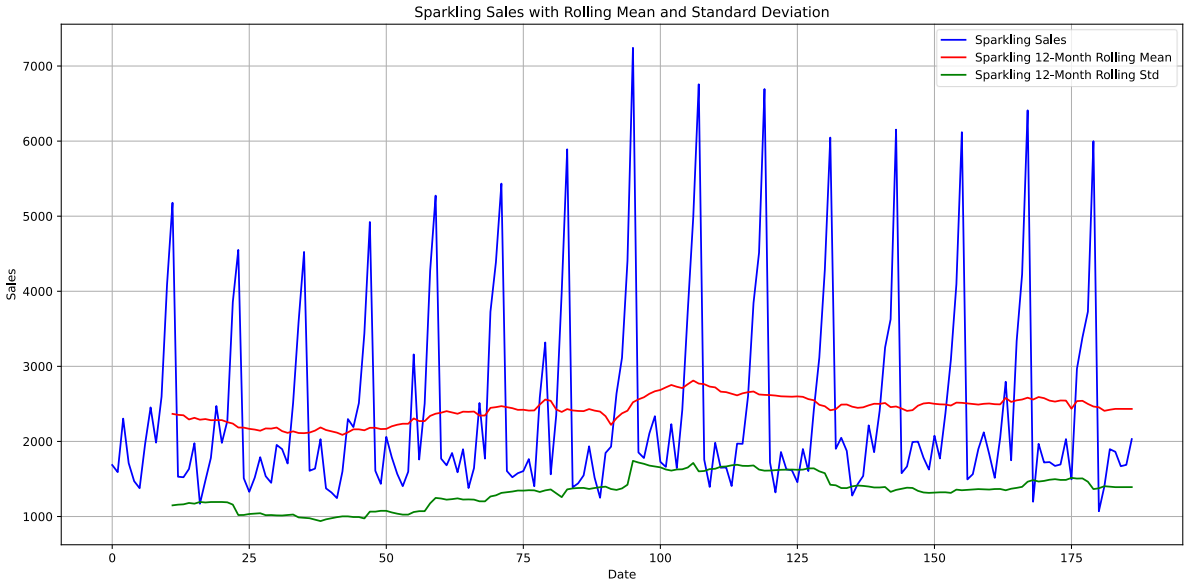


Figure 10 Sparkling Rolling means and Standard Deviation

Perform Decomposition

- Decompose the time series to analyse trend, seasonality, and residual components for both datasets

  - Sparkling the Trend seem to be additive, at a period of 12

  - Rose the has missing values we would need to treat the same and then do decomposition.

  - Post handling of missing values when we do decomposition, it looks to be additive

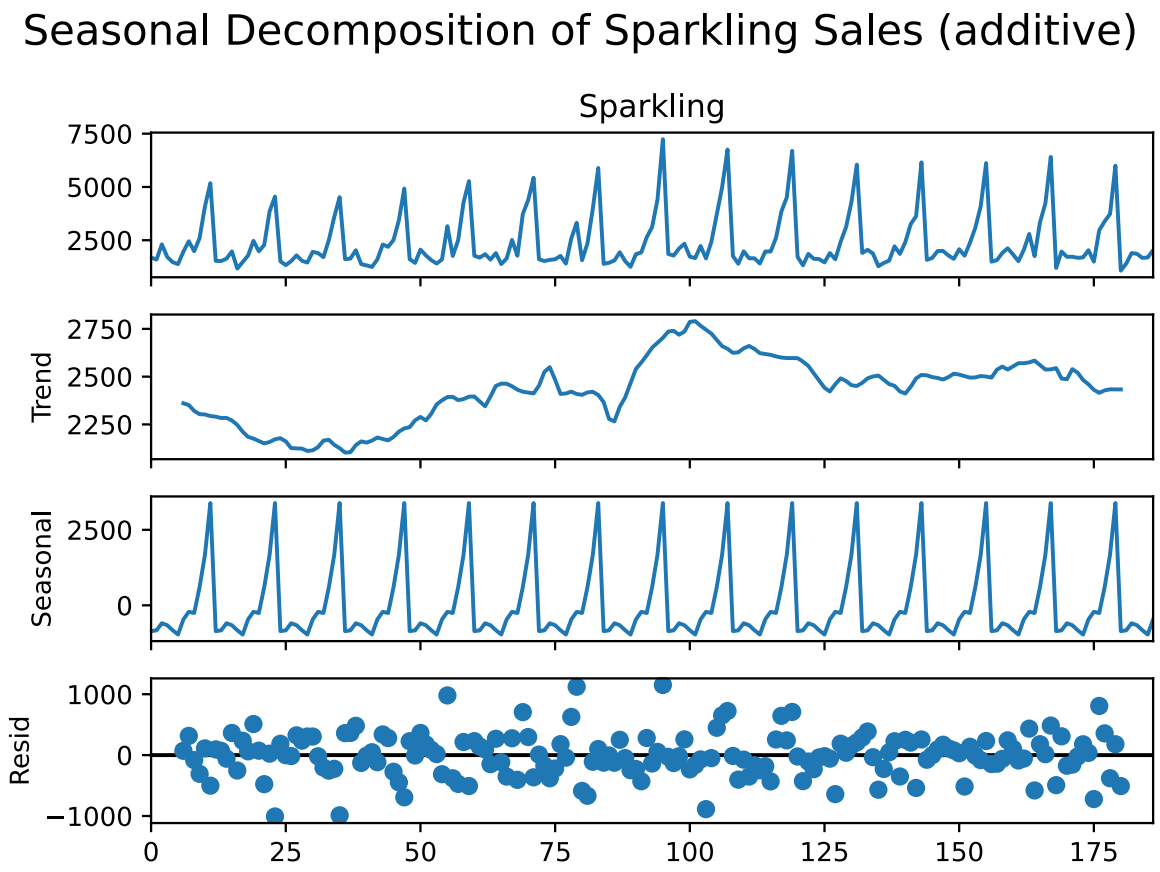


Figure 11 Sparkling Wine Sales decomposition additive

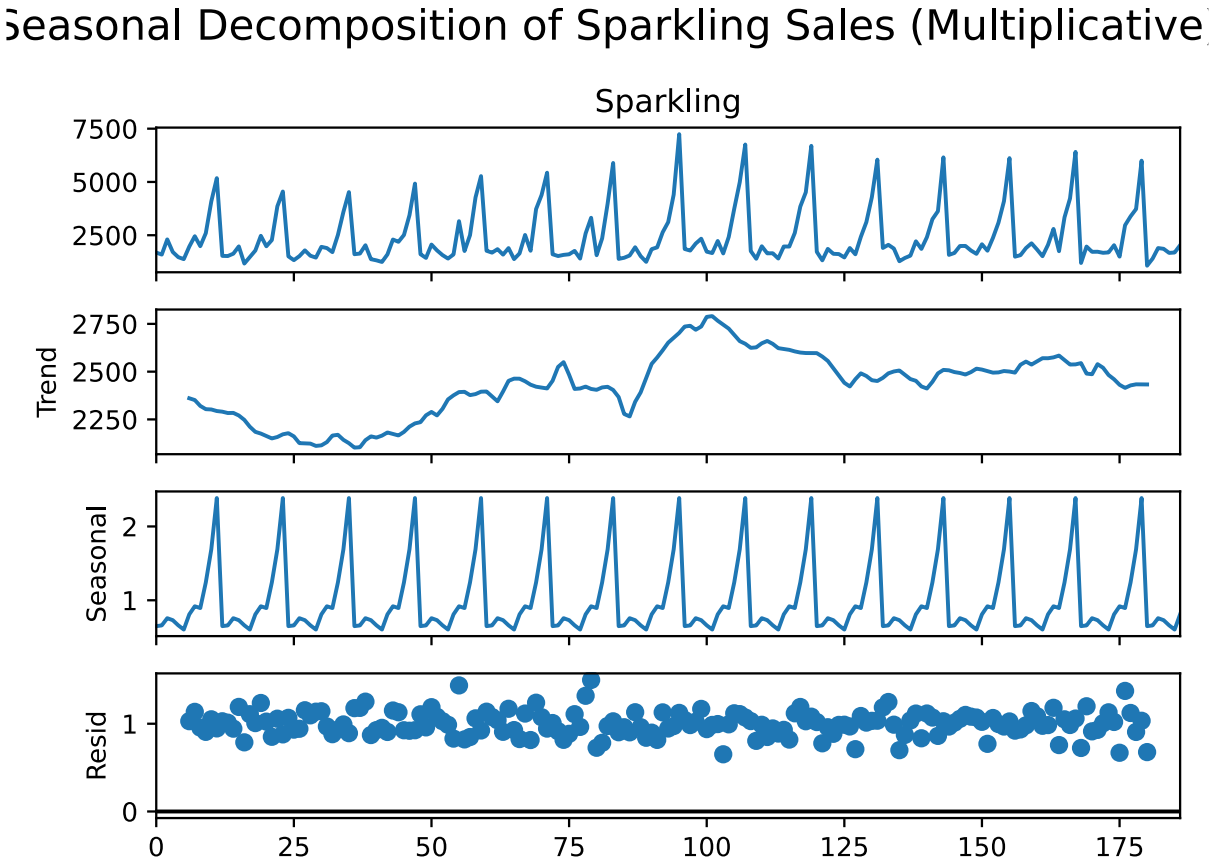


Figure 12 Sparkling Wine Sales Decomposition Multiplicative

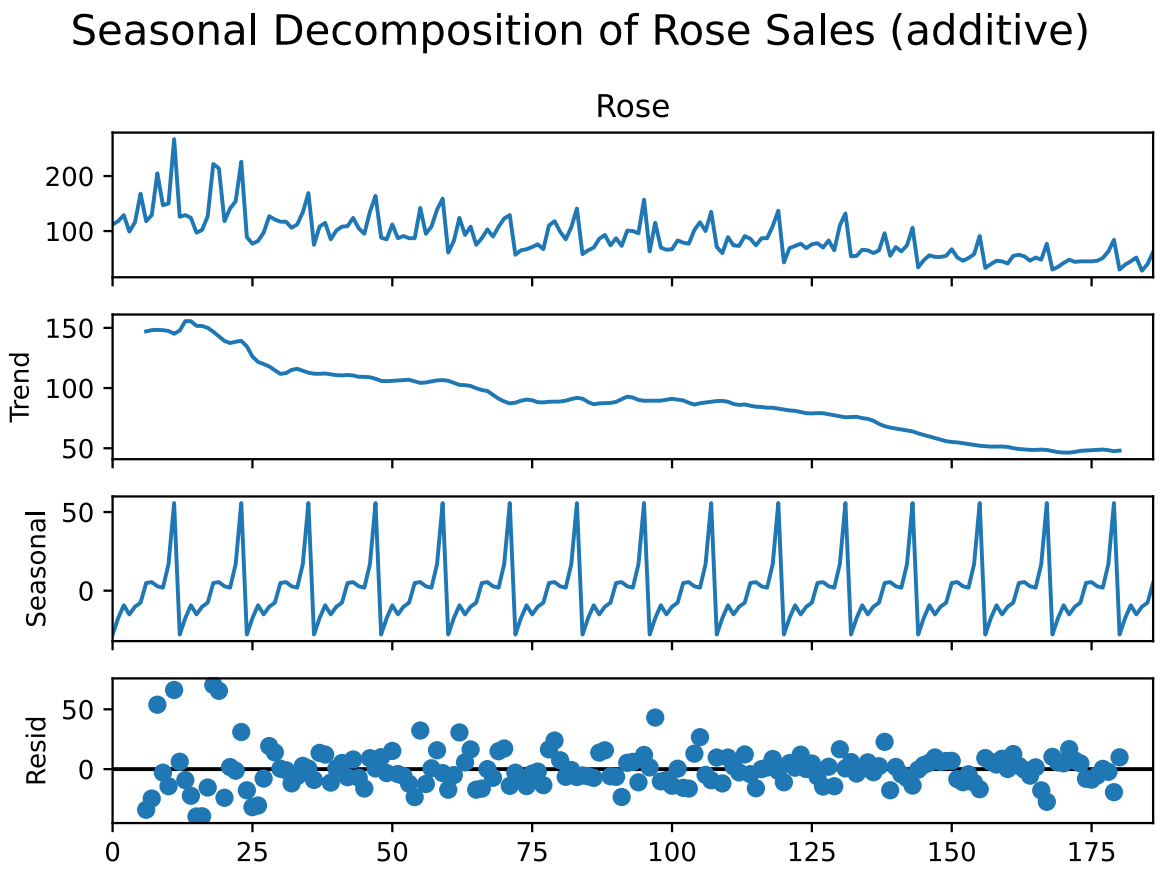


Figure 13 Rose Wine Sales Decomposition Additive

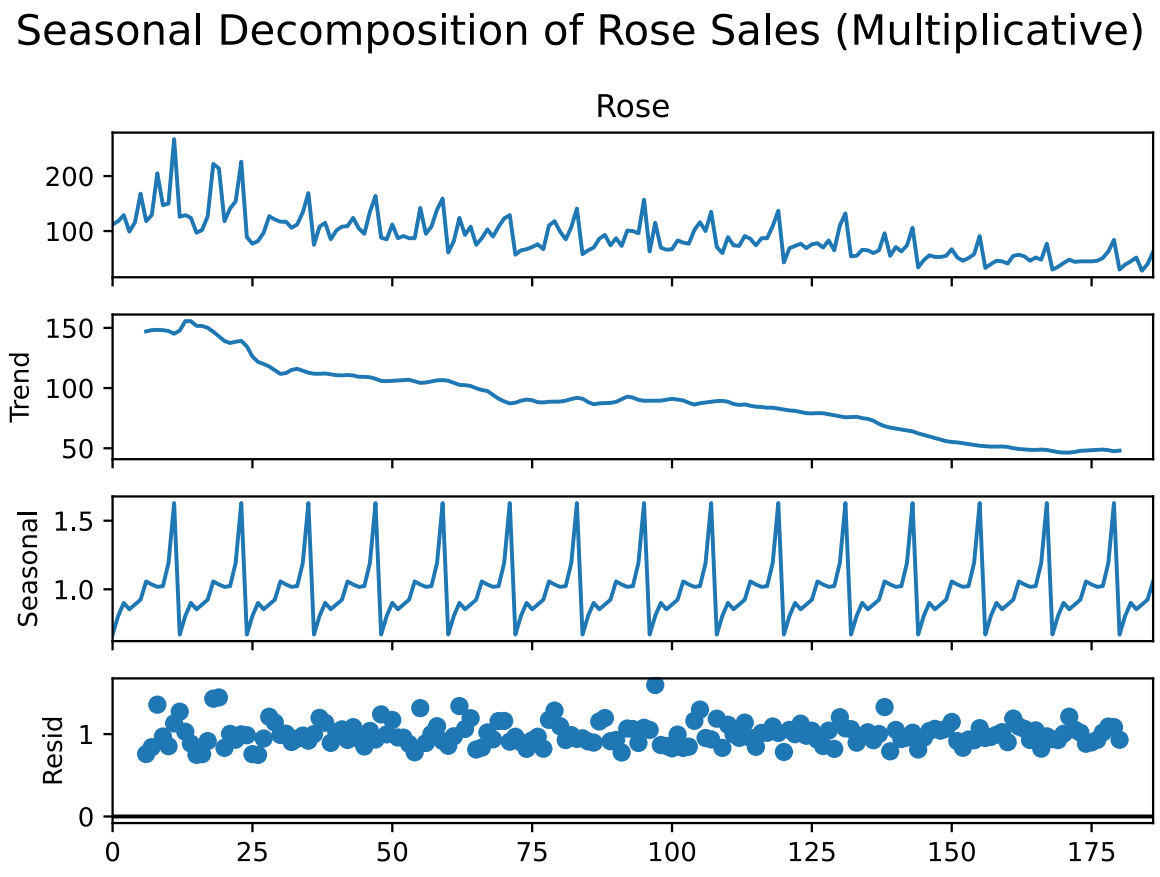


Figure 14 Rose Wine Sales Decomposition Multiplicative

Question 2: Data Preprocessing

Missing Values Treatment

- Sparkling does not have any missing values for sales.

- Rose has missing values for sales we will use ffill to correct this and then run, decomposition.

Visualize the processed data

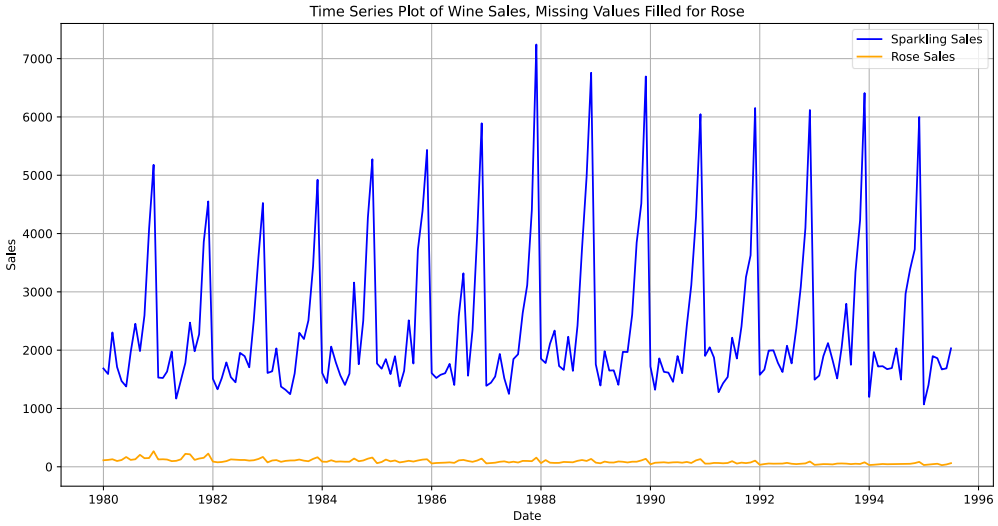


Figure 15 Wine Sales Post Data Missing Correction for Rose

Train-Test Split

-We split the dataset in a 70-30 ratio along the timeline.

-This gives us 130 observations in train data and 57 observations in test data

Question 3: Model Building - Original Data

Linear Regression Model

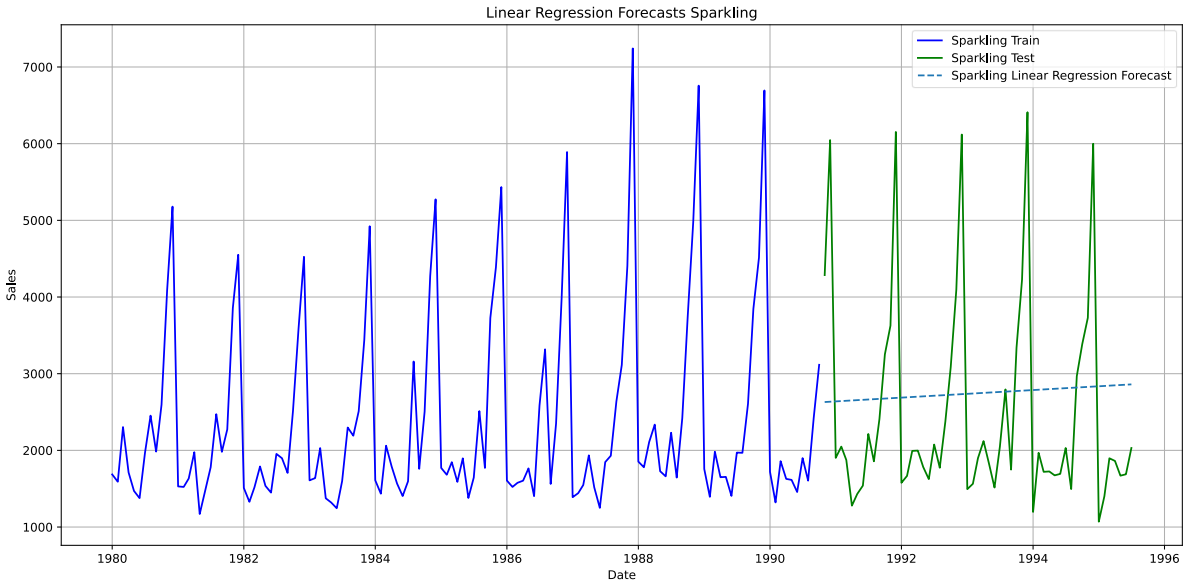


Figure 16 Sparkling Sales Forecast on Test using LR

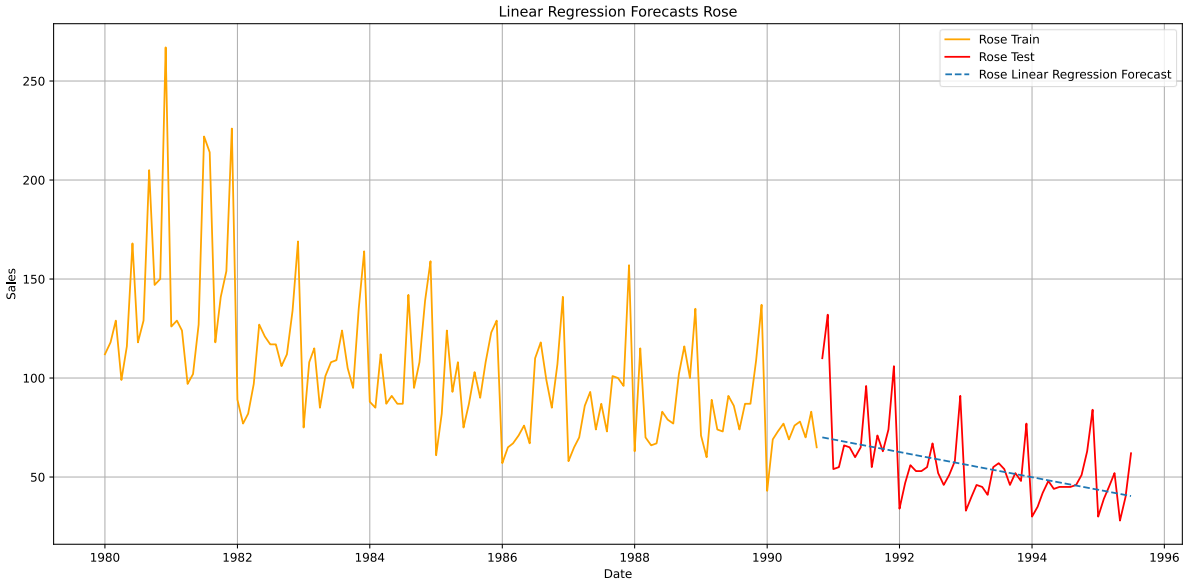


Figure 17 Rose Sales Forecast on Test using LR

Simple Average Model

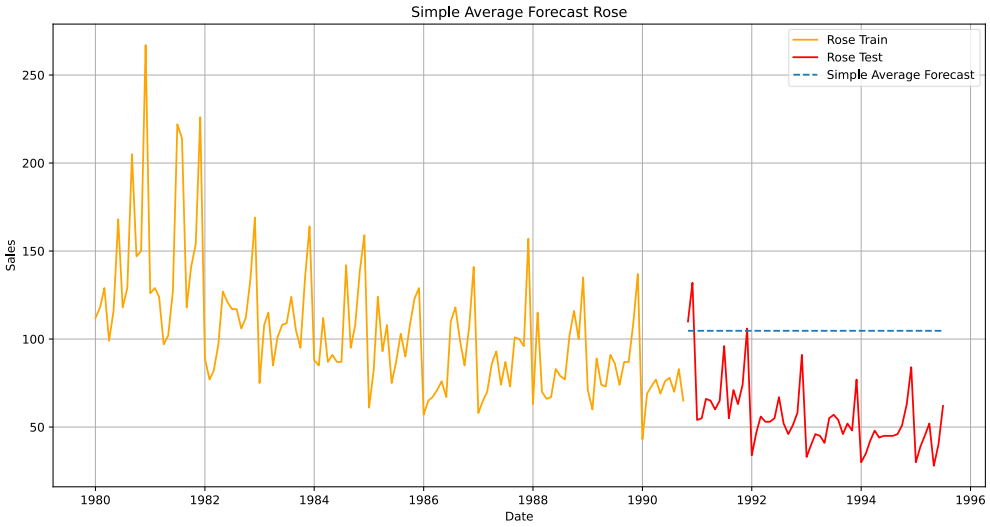


Figure 18 Rose Sales Forecast on Test using Simple Average

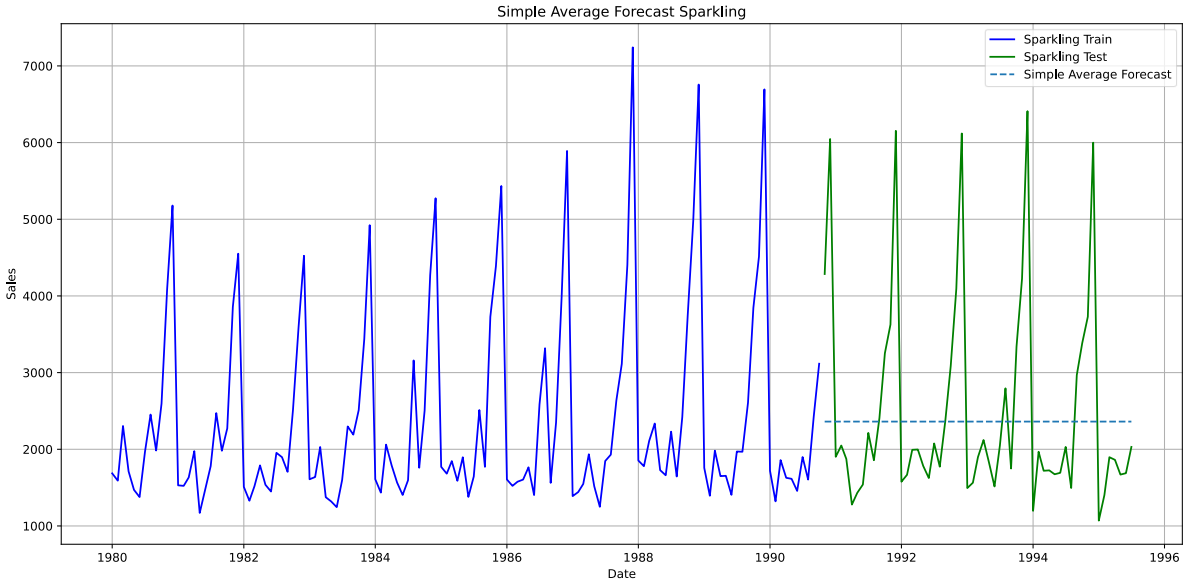


Figure 19 Sparkling Sales Forecast on Test using Simple Average

Moving Average Model

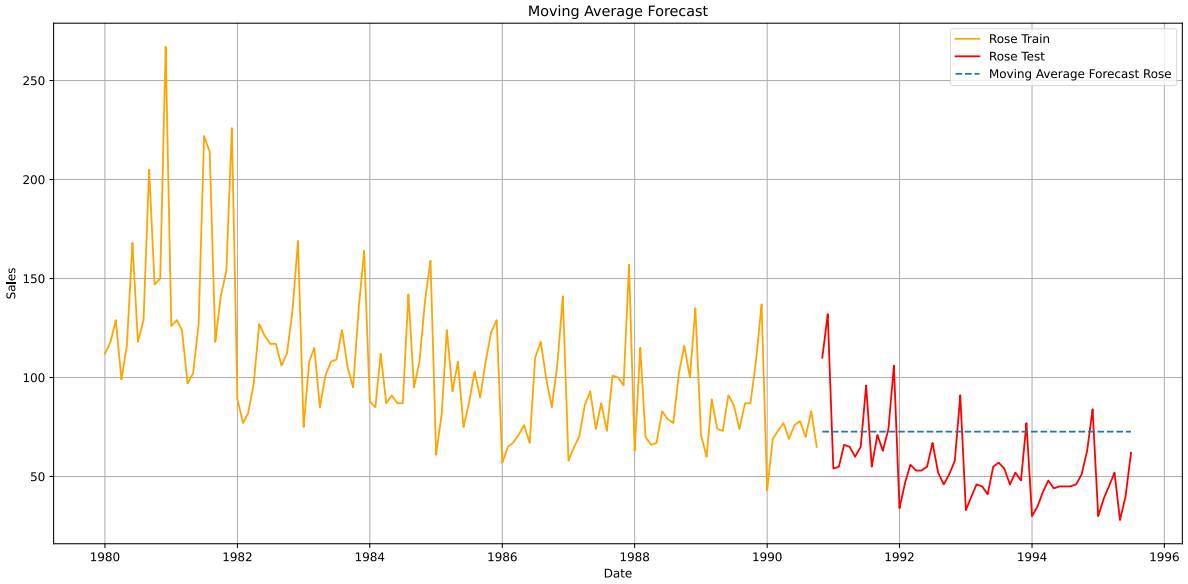


Figure 20 Rose Sales Forecast on Test using Moving Average

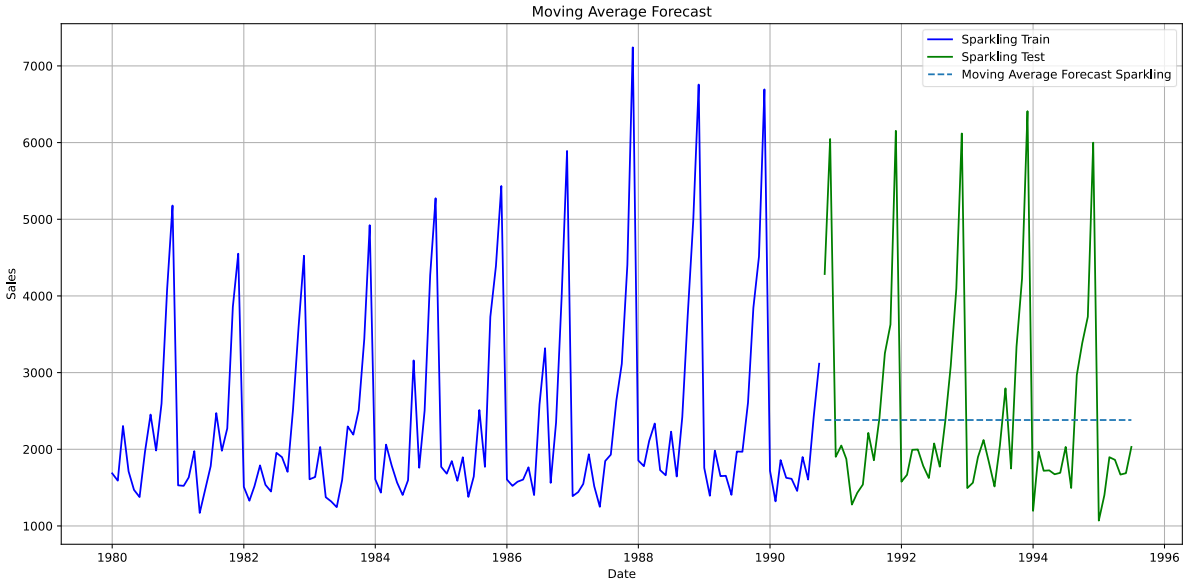


Figure 21 Sparkling Sales Forecast on Test using Moving Average

Exponential Smoothing Models (Single, Double, Triple)

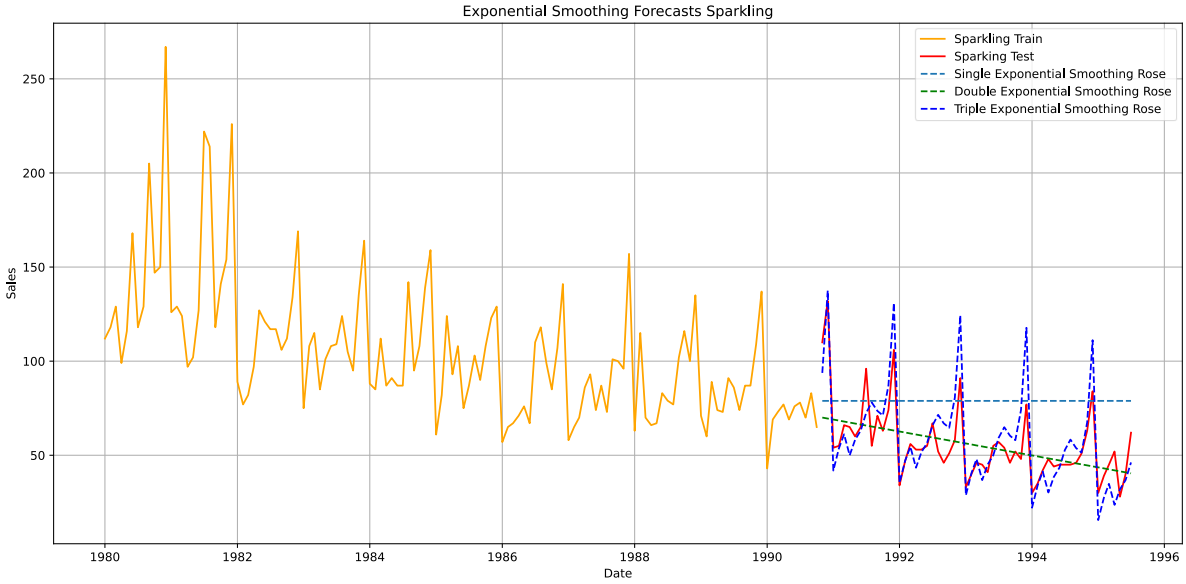


Figure 22 Rose Sales Forecast on Test using Exponential Smoothing

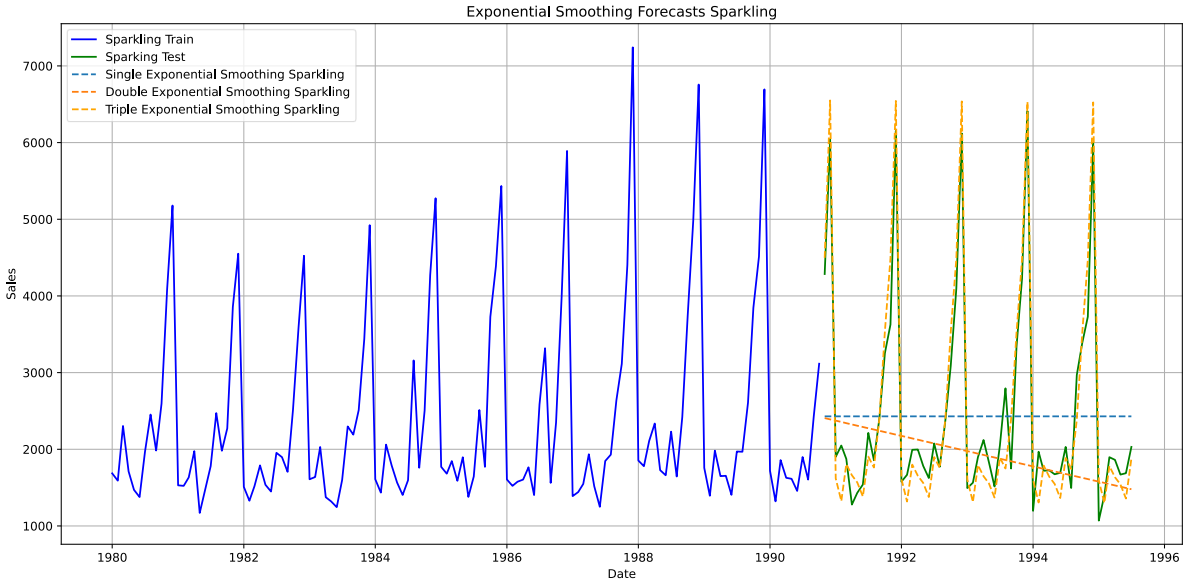


Figure 23 Sparkling Sales Forecast on Test using Exponential Smoothing

Above Model Comparison:

- Linear Regression for Sparking Wine. Performance: MSE=1938884.4329, MAE=1148.3683
- Simple Average for Sparking Wine. Performance: MSE=1873467.5749, MAE=980.8915
- Moving Average for Sparking Wine. Performance: MSE=1868379.3626, MAE=988.5673
- Single Exponential Smoothing for Sparking Wine. Performance: MSE=1859686.7840, MAE=1008.6860
- Double Exponential Smoothing for Sparking Wine. Performance: MSE=2167530.5824, MAE=896.3781
- Triple Exponential Smoothing for Sparking Wine. Performance: MSE=134585.6401, MAE=292.0846

Question 4: Check for Stationarity

- For Sparkling Sales
- ADF Statistic: -1.3604974548123347
- p-value: 0.6010608871634865
- p-value is over 0.05, The Sparking Wine Sales is non-stationary

- For Rose Sales
- ADF Statistic: -1.8748555417199857
- p-value: 0.34398071933430585
- p-value is over 0.05, The Rose Wine Sales is non-stationary

Make the data stationary (if needed)

- For Sparkling Sales post differencing
- ADF Statistic (after differencing): -45.05030093619526
- p-value (after differencing): 0.0
- p-value is less 0.05, The Sparkling Wine Sales is stationary with first-order difference.

- For Rose Sales port differencing
- ADF Statistic (after differencing): -3.8450642364446996
- p-value (after differencing): 0.002478477502772498
- p-value is less 0.05, The Sparkling Wine Sales is stationary with 7th-order difference.

Question 5: Model Building - Stationary Data

Generate ACF & PACF Plot and Find AR, MA values

- ACF Plot Shows spikes at: 1,2,6,11,12,13,14,18 for Sparkling
- PACF Plot Shows spikes at: 10,11,12,18,19 for Sparkling

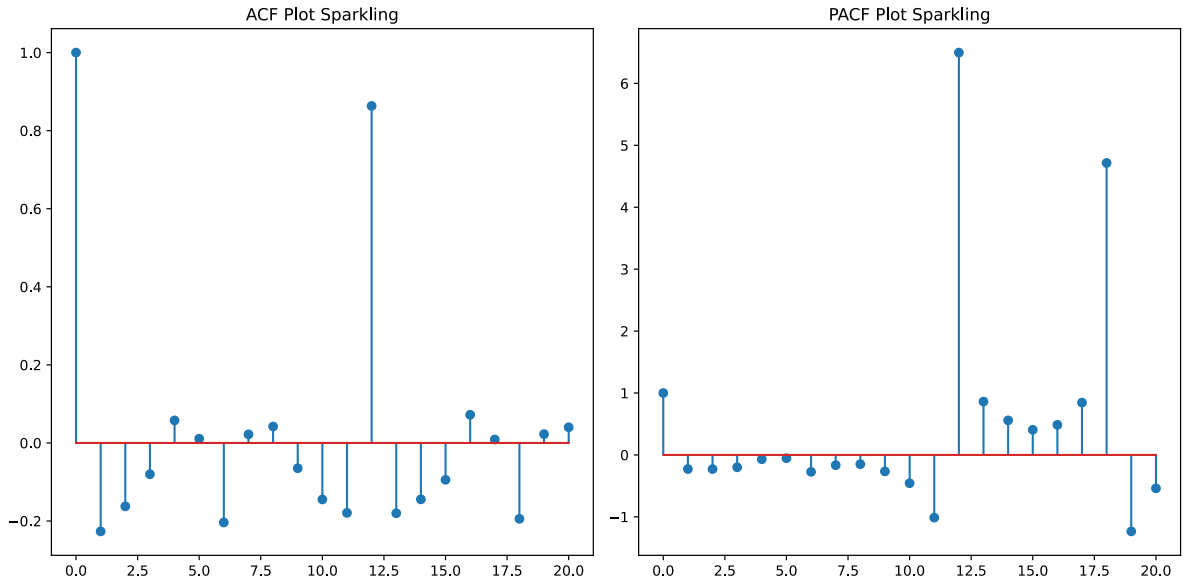
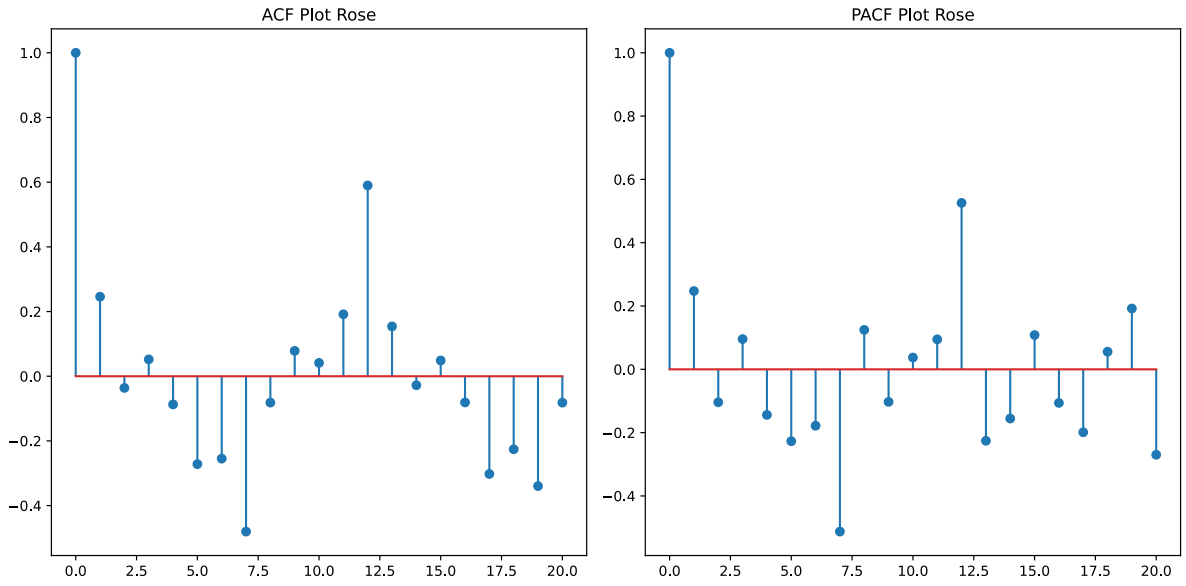


Figure 24 ACF and PACF Sparkling Sales

- ACF Plot Shows spikes at: 1,5,6,7,11,12,17,18,19 for Rose
- PACF Plot Shows spikes at: 1,5,6,7,12,13,14,17,19,20 for Rose



- *Figure 25 ACF and PACF Rose Sales*

Build different ARIMA models

- Built the models
- Linear Regression
- Simple Average
- Moving Average
- Single Exponential Smoothing
- Double Exponential Smoothing
- Triple Exponential Smoothing
- Auto ARIMA model
- Manual ARIMA model
- Auto SARIMA model
- Could not complete Manual SARIMA

Question 6: Compare the performance of the models

- The Double Exponential Smoothing model is good for Rose Sales
- Triple Exponential Smoothing is good for the Sparkling as these have the least MSE, amount the models that ran.

Model	Sparkling MSE	Rose MSE
Linear Regression	1938884.433	301.2628
Simple Average	1873467.575	2749.1122
Moving Average	1868379.363	668.3099
Single Exponential Smoothing	1859686.784	912.5146
Double Exponential Smoothing	2167530.582	301.2628
Triple Exponential Smoothing	134585.6401	196.617
Auto ARIMA model	1857117.255	956.2551
Manual ARIMA model	1860406.753	112859.4726
Auto SARIMA model	1857117.255	956.2551
Manual SARIMA model	Did not complete	Did not complete

The Double Exponential Smoothing model is good for Rose and the Triple Exponential Smoothing is good for the Sparkling as these have the least MSE, amount the models that ran

Figure 26 Models MSE Compare