05/01/2025

# Time Series Forecasting - ABC Estate Wines

[Sparkling Wines]

## Business Report (Coded - Project)

NABANKUR RAY
PGP-DSBA

# Contents

# List of figures

# List of Tables

# Problem Statement - UL Project - Coded

## Business 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.

## Data Description

The datasets provided contain monthly sales data for two types of wines—Rose and Sparkling—spanning from January 1980 to November 1995. Each dataset records sales volumes in numeric format, reflecting the demand for each wine type during the specified time period. The data is structured in time-series format, enabling trend analysis, seasonality detection, and forecasting.

## Data Dictionary

### Sparkling Dataset

- **YearMonth (Date):** Represents the year and month of the sales data in YYYY-MM format.
- **Sparkling (Integer):** Monthly sales volume of Sparkling wine.

## Executive Summary:

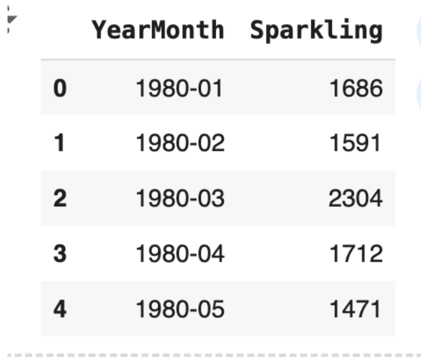
This report analyzes historical sales trends for Rose and Sparkling wines from ABC Estate Wines between 1980 and 1995. The primary goal is to uncover sales patterns, detect seasonal trends, and forecast future demand to support strategic decision-making and enhance business performance. Initial exploration reveals consistent seasonal trends in both datasets, with Sparkling wine demonstrating higher sales volumes and sharper fluctuations compared to Rose wine. Decomposition and statistical forecasting models will further refine these insights to predict future performance and identify growth opportunities.

## Deliverables:

- **Exploratory Data Analysis (EDA):** Visualization and statistical summaries highlighting trends, seasonality, and patterns.
- **Time-Series Decomposition:** Breakdown of sales data into trend, seasonality, and residual components.
- **Forecast Models:** Development of predictive models to forecast future sales based on historical patterns.
- **Business Insights:** Actionable recommendations for optimizing sales strategies, inventory planning, and marketing efforts.
- **Report Documentation:** Comprehensive project report summarizing methodology, findings, and suggestions for future steps.

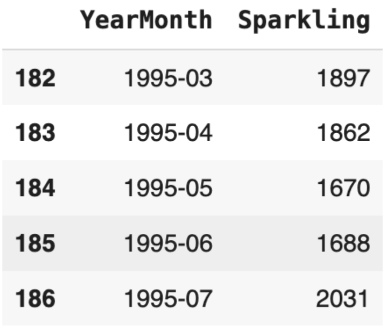
# Understanding the Data

## Data Overview



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

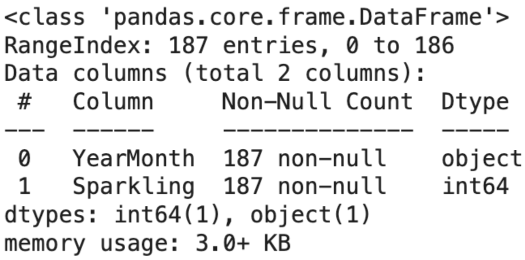
*Table 1: First 5 rows of the given dataset*



|  | YearMonth | Sparkling |
|---|---|---|
| 182 | 1995-03 | 1897 |
| 183 | 1995-04 | 1862 |
| 184 | 1995-05 | 1670 |
| 185 | 1995-06 | 1688 |
| 186 | 1995-07 | 2031 |

*Table 2: Last 5 rows of the given dataset*

## Structure and Types of Data



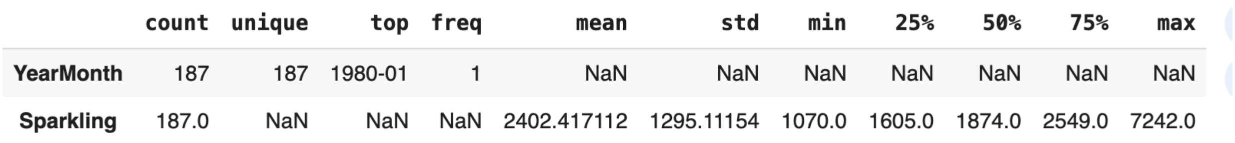
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   YearMonth  187 non-null    object
 1   Sparkling  187 non-null    int64 
dtypes: int64(1), object(1)
memory usage: 3.0+ KB
```

*Table 3: structure and type of data*

### OBSERVATIONS:

- There are **187 rows** and 2 **Columns** are present in the each given datasets (rose.csv & sparkling.csv).
- It can be observed that sparkling column don't have any less entries (less than 187 rows) indicates that there are no missing values.
- Here, YearMonth column is identified as object data types which needs to convert to datetime format for time-series analysis.
- And Sparkling Column is of int data type.

## Statistical summary of the Numerical Data



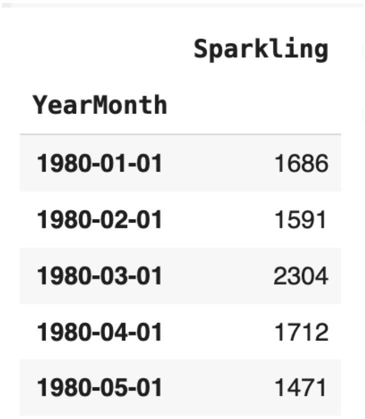
| | count | unique | top | freq | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **YearMonth** | 187 | 187 | 1980-01 | 1 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Sparkling** | 187.0 | NaN | NaN | NaN | 2402.417112 | 1295.11154 | 1070.0 | 1605.0 | 1874.0 | 2549.0 | 7242.0 |

*Table 4: Statistical summary of the data*

- The **range** of "Sparkling" values is large (from 1070 to 7242), highlighting high variability in the data.
- The mean is higher than the median (1874.0), suggesting a right-skewed distribution with some high values.
- Most values fall between the 25th percentile (1605.0) and the 75th percentile (2549.0), but extreme peaks are present.
- The data sets contains sparkling sales data from Timeframe “1980-01 to 1995-07".

## INFERENCE:

- There is no missing values.
- YearMonth column is identified as object data types and converted to DateTime format for time-series analysis. Hence, The data is ready for Time Series Analysis.



| | Sparkling |
|---|---|
| **YearMonth** | |
| **1980-01-01** | 1686 |
| **1980-02-01** | 1591 |
| **1980-03-01** | 2304 |
| **1980-04-01** | 1712 |
| **1980-05-01** | 1471 |

*Table 5: Indexed data for Time Series Analysis*

# Exploratory Data Analysis

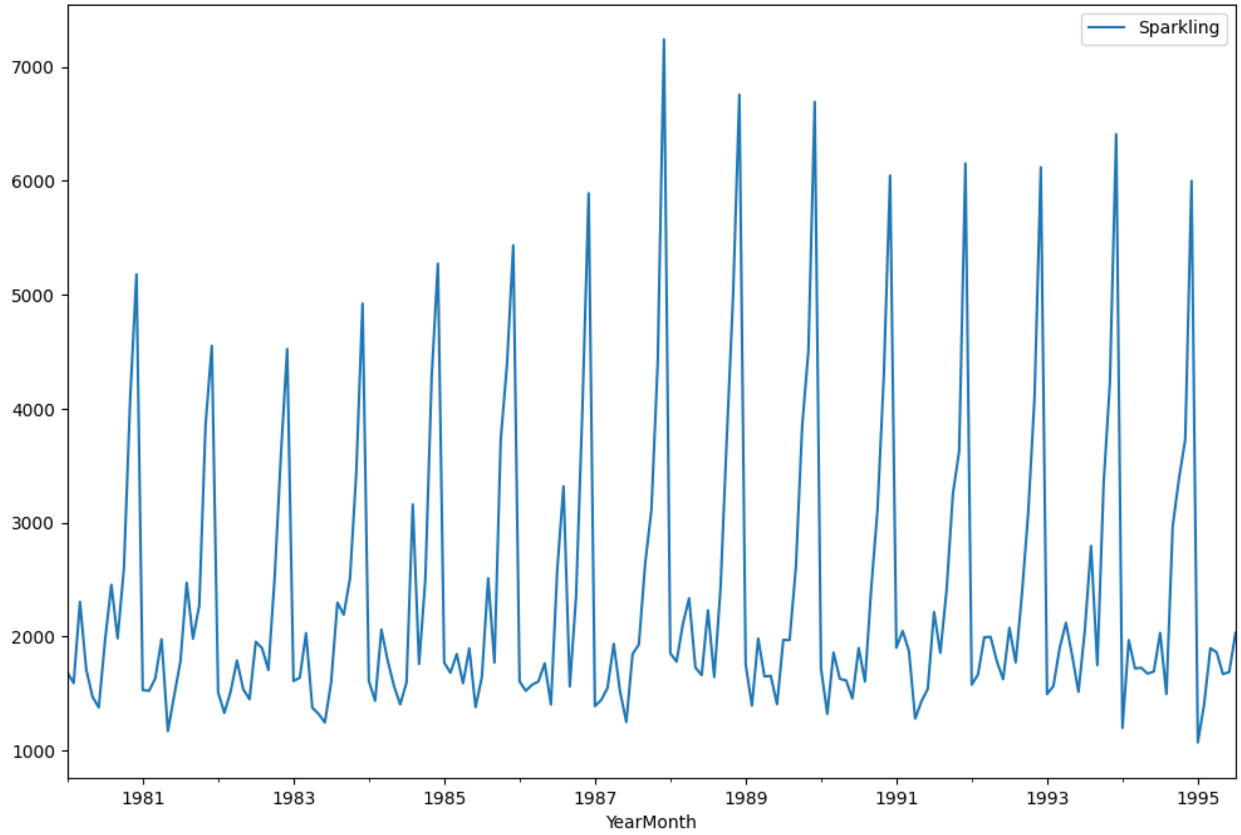


*Fig 1: Trend of the Sparkling wine sales*

### OBSERVATIONS:

**Trend:**

The general trend appears to be slightly increasing over time. While there is no sharp rise, the peaks seem to grow taller as the years progress. Between 1981 and 1995, there is a gradual increase in the amplitude of the values, suggesting growth in the "Sparkling" metric over time.

**Seasonality:**

A clear seasonal pattern is present, with sharp peaks recurring at regular intervals, likely on an annual basis. These peaks indicate specific times of the year when the "Sparkling" value is significantly higher, possibly corresponding to seasonal demand or production cycles. The valleys between the peaks are consistent, further reinforcing the presence of strong seasonality.

# Decomposition

## Additive Decomposition Summary

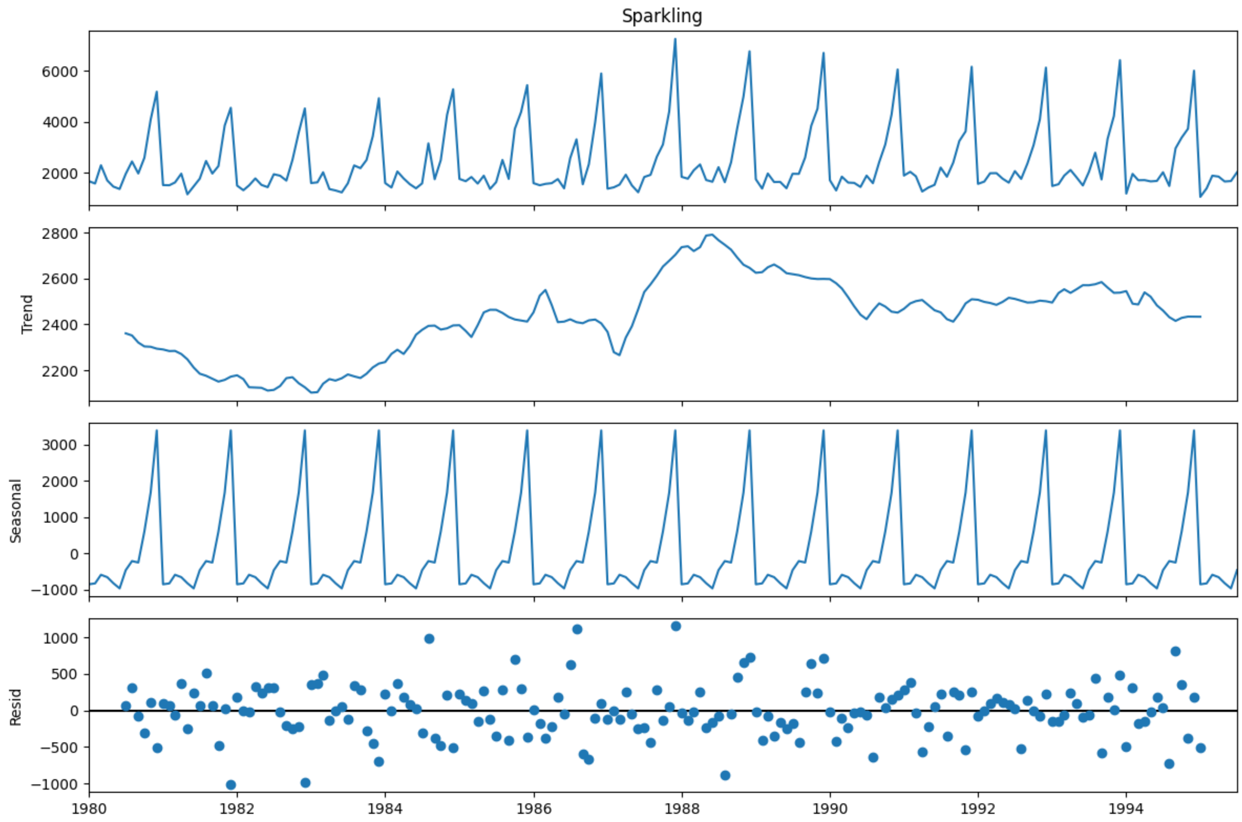


*Fig 2: Additive Decomposition Summary*

**The trend** follows a similar trajectory as in the multiplicative decomposition: a gradual decline initially, a rise until the late 1980s, and a slight decrease towards the end.

**The seasonal** component remains constant in amplitude over time. Peaks and troughs occur at regular intervals, but the scale is not influenced by the overall magnitude of the data.

**The residuals** show consistent variability across the timeline, with no noticeable proportionality to the trend or seasonality. The spread is more uniform compared to the multiplicative case.

## Multiplicative Decomposition Summary

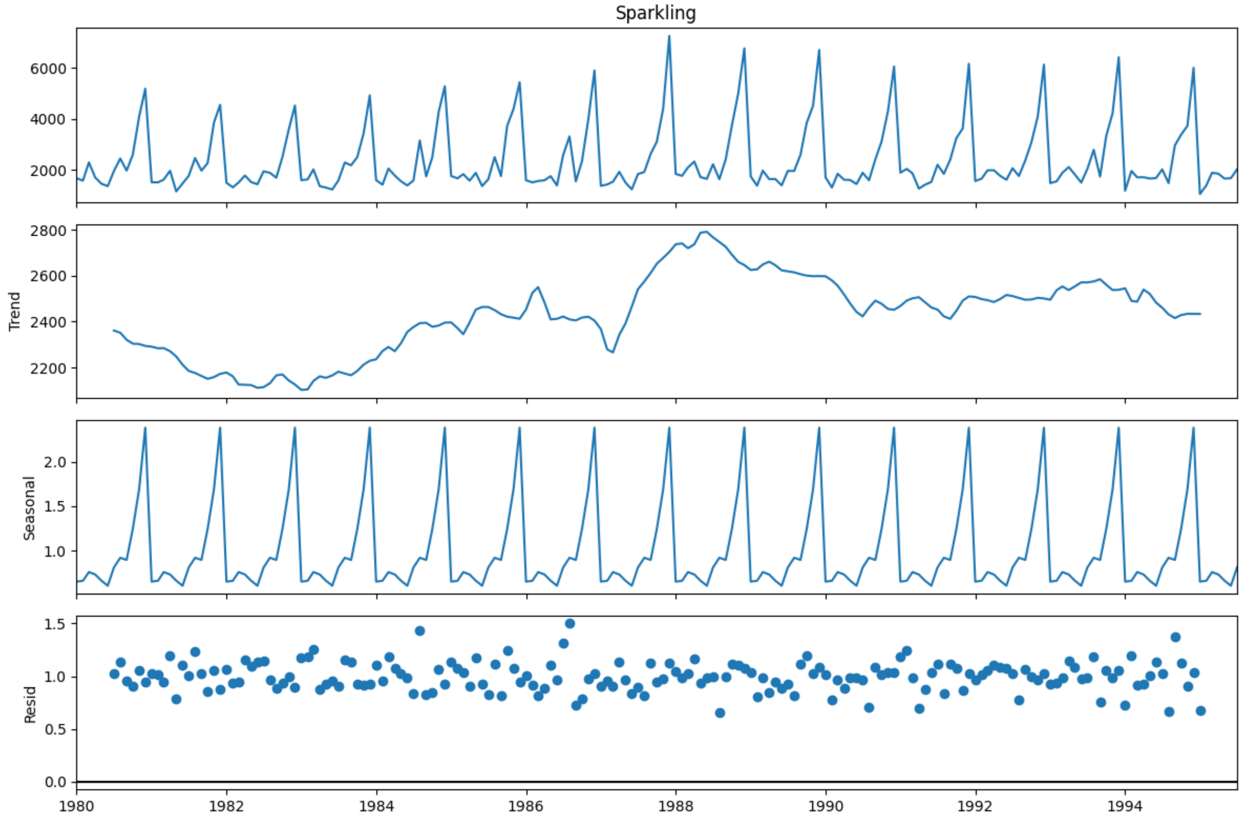


*Fig 3: Multiplicative Decomposition Summary*

**The trend** shows a gradual decline in the earlier years, followed by a steady increase until the peak around the late 1980s, and then stabilizes or slightly declines. This indicates a long-term change in the base level of the data.

**The seasonal** component exhibits a proportional pattern relative to the magnitude of the data. Peaks and troughs become more prominent as the trend rises, reflecting multiplicative seasonality.

**The residuals** seem relatively consistent across the timeline, but their variation appears proportional to the level of the data. This suggests that the errors also scale with the magnitude of the time series.

# Data Preprocessing

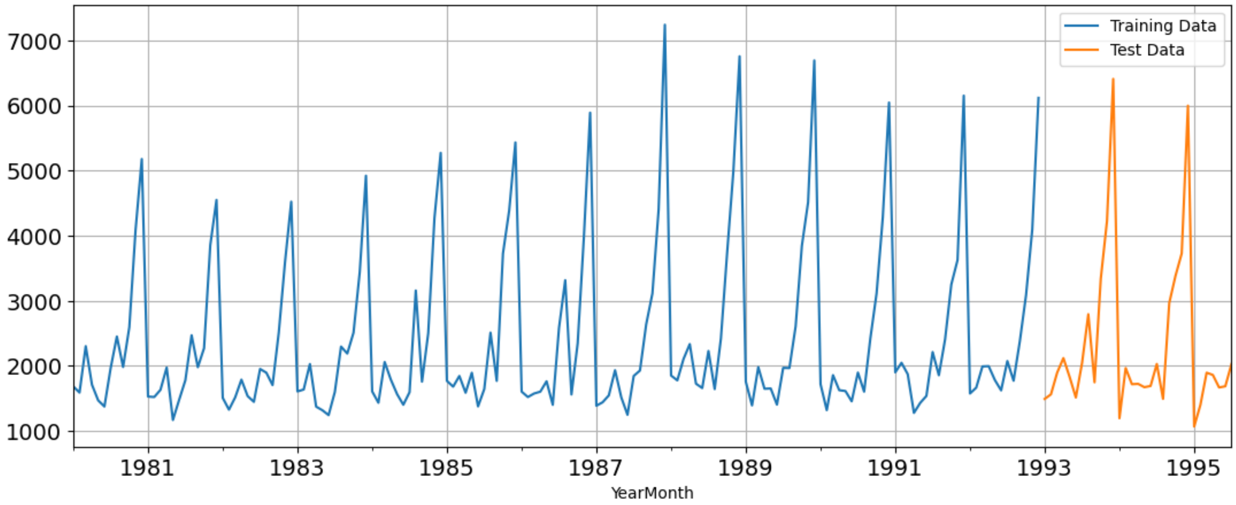


*Fig 4:*

### INFERENCE

- The given data is splitted into train set consisting data until 1992 and test set from 1993 onwards.
- There are 156 data on train set and 31 data on test set.
- There are no missing values.
- The data is ready for model building.

# Model Building – Original Data

## Linear Regression

- RMSE on Linear Regression Test data - **1.803630e+06**

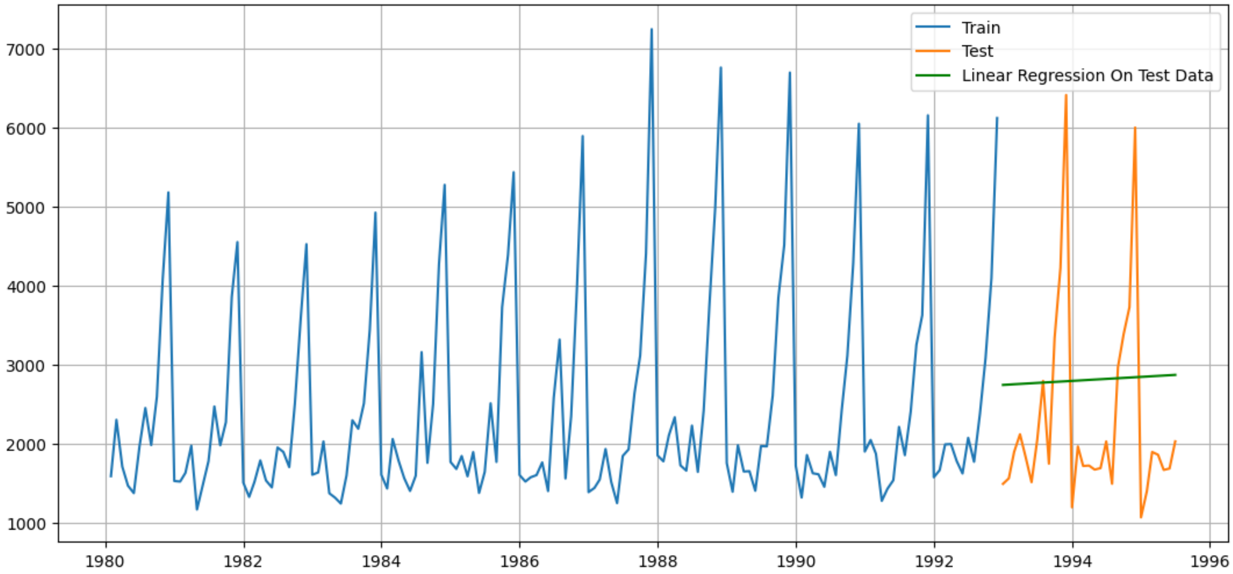


*Fig 5: Linear Regression On Test Data*

## Moving Average

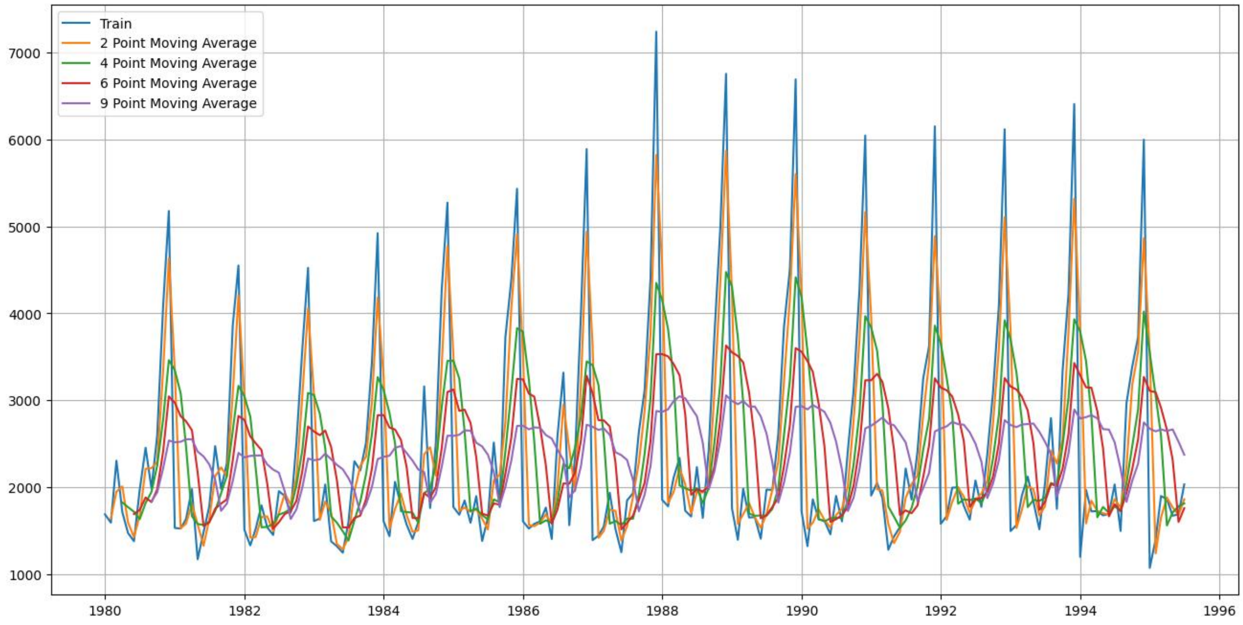


*Fig 6: Trailing Moving Average of 2, 4, 6, 9 points*

For **2 point Moving Average Model** forecast on the Testing Data, RMSE is **7.452916e+05**

For **4 point Moving Average Model** forecast on the Training Data, RMSE is **1.418502e+06**

For **6 point Moving Average Model** forecast on the Training Data,  RMSE is **1.656642e+06**

For **9 point Moving Average Model** forecast on the Training Data,  RMSE is **1.764540e+06**

- The 2-point Trailing Moving Average has the lowest RMSE (Root Mean Squared Error) value of 7.45e+05, indicating it performs best among the models tested.
- As the number of points in the trailing moving average increases (from 2 to 9 points), the RMSE tends to increase. This indicates that shorter window sizes (e.g., 2 points) capture the data's patterns more accurately in this case.
- Simpler models like the 2-point Trailing Moving Average perform better for this dataset, possibly due to the strong seasonal components, while more complex models like Linear Regression fail to capture the underlying time series patterns effectively.

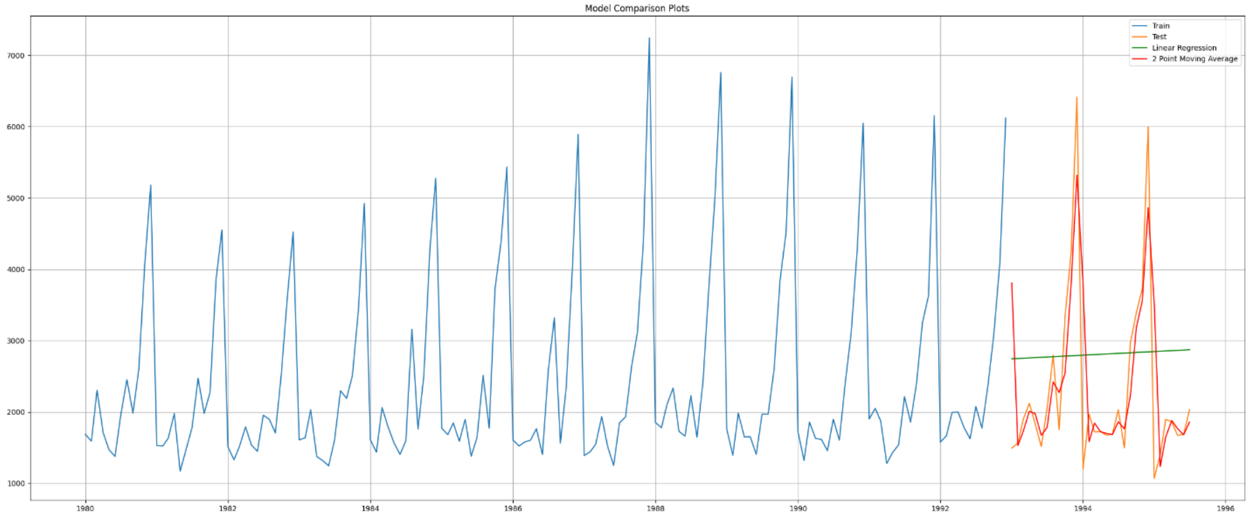
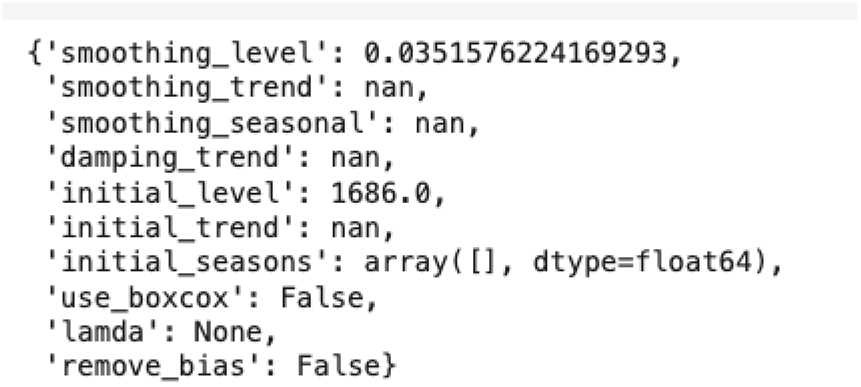


*Fig 7: Model Comparision plot*

## Simple Exponential Smoothening Models



```
{'smoothing_level': 0.0351576224169293,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1686.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

*Table 6: Simple Exponential Smoothening Model Summary*

Here, Smoothing level is 0.035

For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is **1.667686e+06**

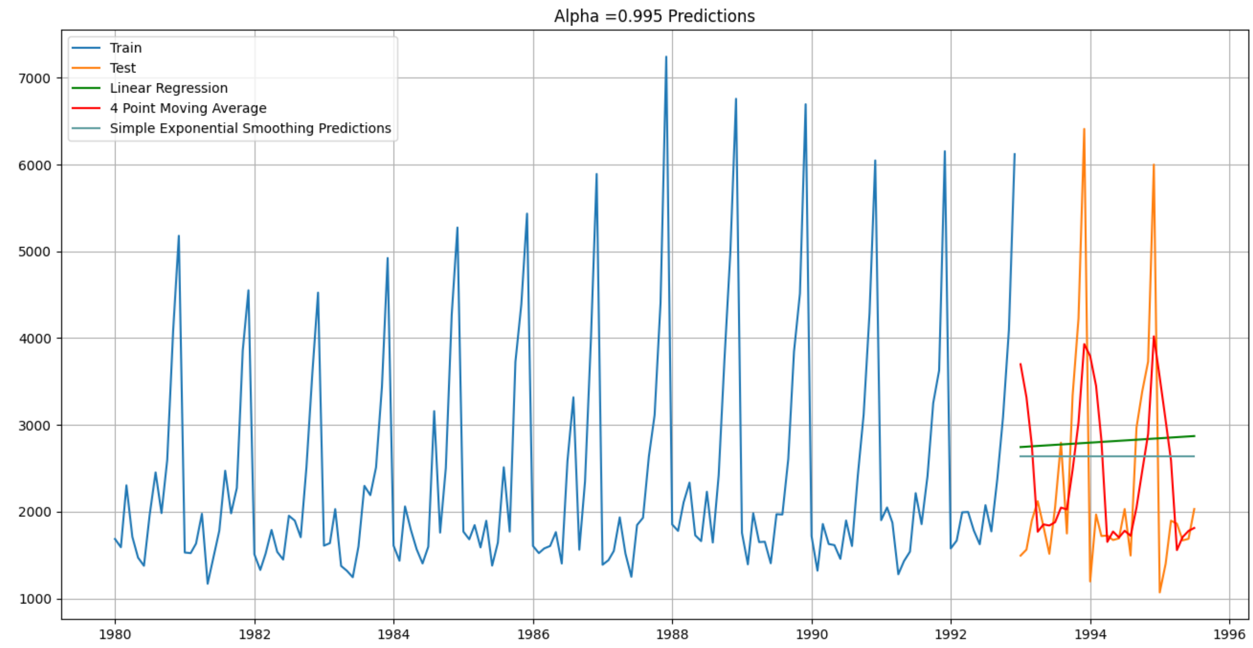


*Fig 8: Model Comparision plot*

## Double Exponential Smoothening (Holt's Model)

For Alpha=0.9, Beta=0.3 Double Exponential Smoothing Model forecast on the Test Data, RMSE is **1.635052e+09**

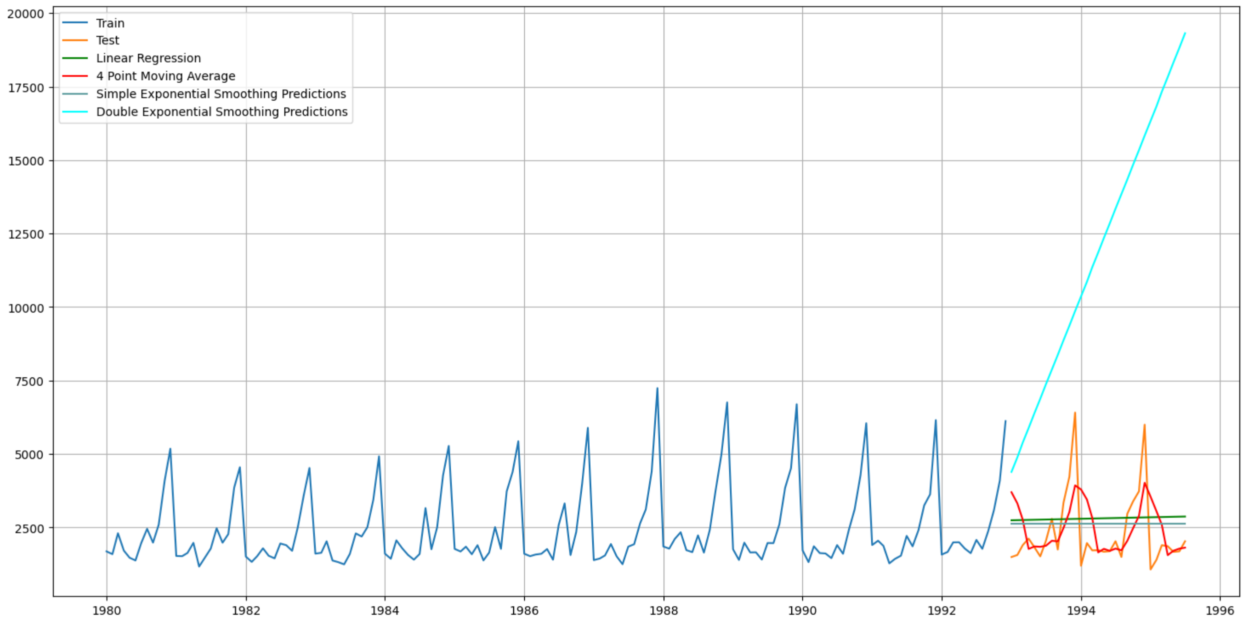
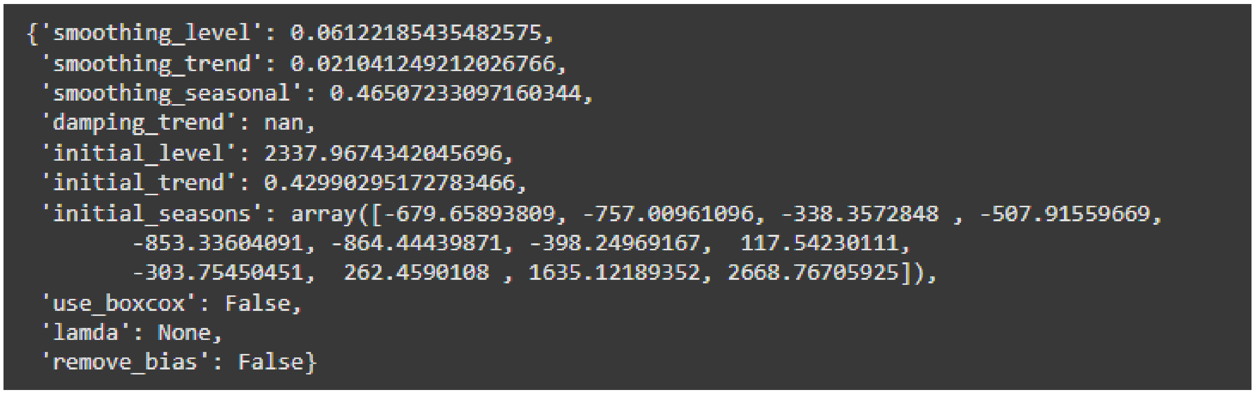


*Fig 9: Model Comparision plot*

## Triple Exponential Smoothing (Holt - Winter's Model)



```
{'smoothing_level': 0.06122185435482575,
 'smoothing_trend': 0.021041249212026766,
 'smoothing_seasonal': 0.46507233097160344,
 'damping_trend': nan,
 'initial_level': 2337.9674342045696,
 'initial_trend': 0.42990295172783466,
 'initial_seasons': array([-679.65893809, -757.00961096, -338.3572848 , -507.91559669,
        -853.33604091, -864.44439871, -398.24969167,  117.54230111,
        -303.75450451,  262.4590108 , 1635.12189352, 2668.76705925]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

*Table 7: Triple Exponential Smoothening Model Summary*

For Alpha=0.676, Beta=0.088, Gamma=0.323 Triple Exponential Smoothing Model forecast on the Test Data, RMSE is **1.078671e+05**

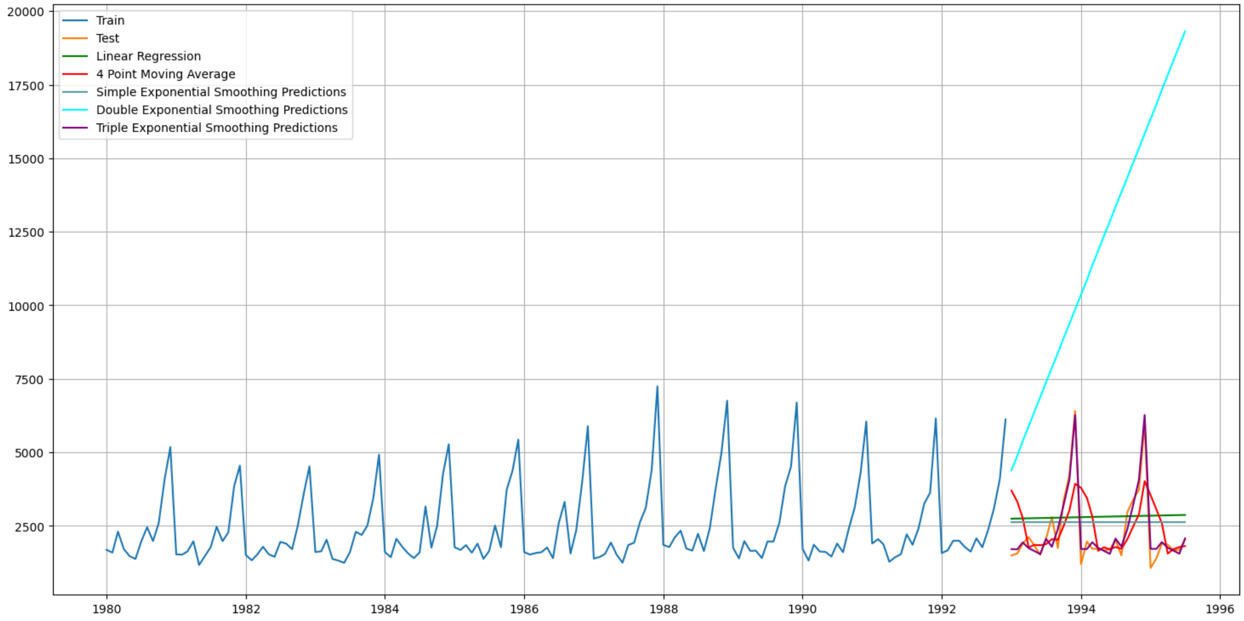


*Fig 10: Model Comparision plot*

## Check for stationarity of the whole Time Series data.

Test for stationarity on the data

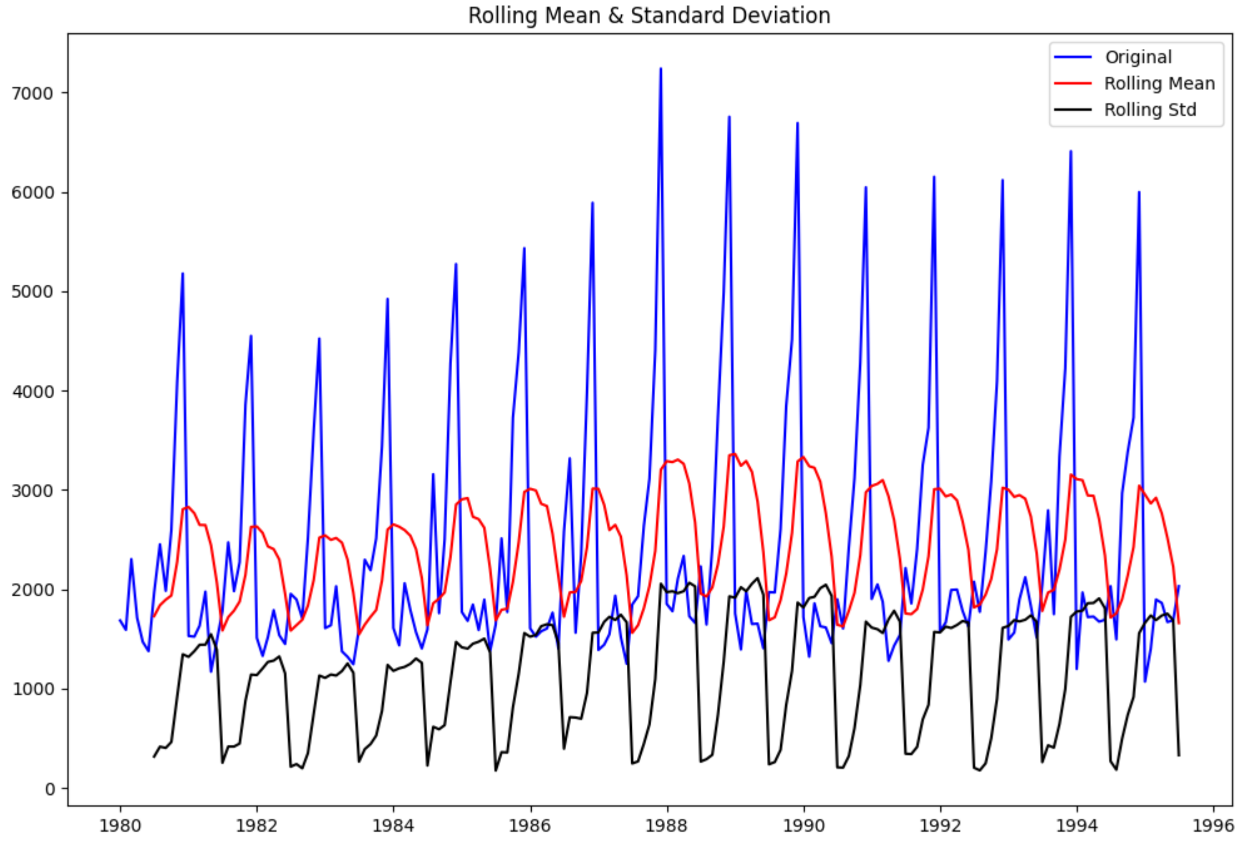
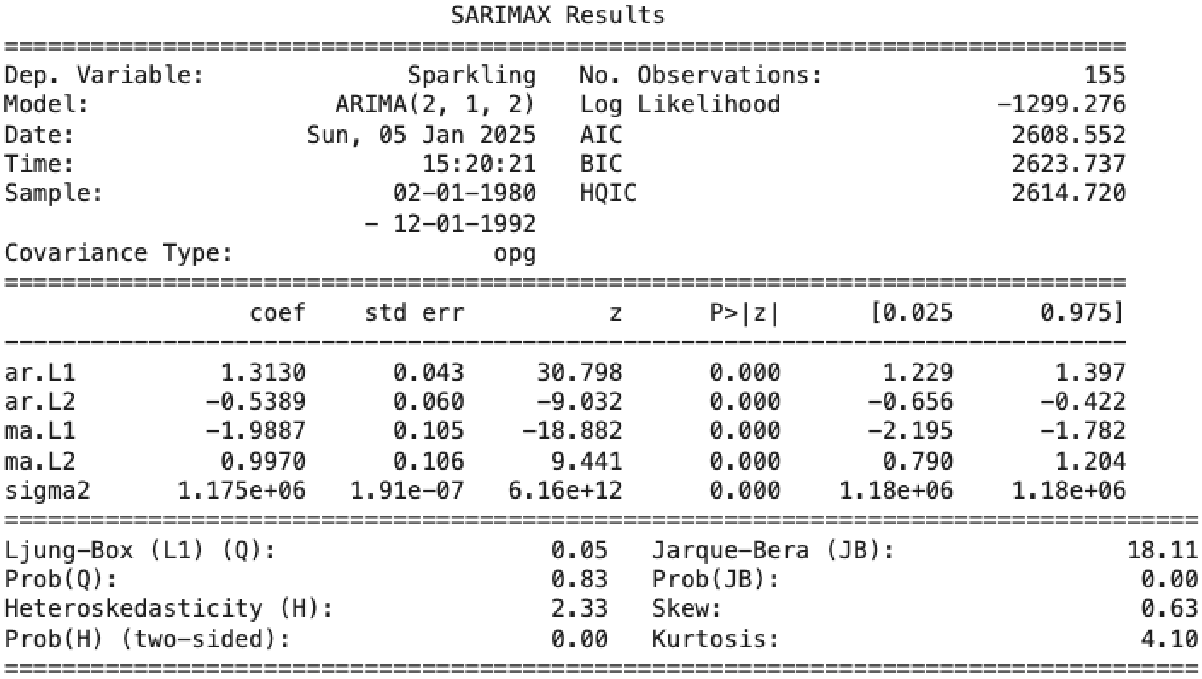


*Fig 11: Test for stationarity on the data*

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
```

# Model Building - Stationary Data

## Auto ARIMA Model



```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  155
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1299.276
Date:                Sun, 05 Jan 2025   AIC                           2608.552
Time:                        15:20:21   BIC                           2623.737
Sample:                    02-01-1980   HQIC                          2614.720
                         - 12-01-1992
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3130      0.043     30.798      0.000       1.229       1.397
ar.L2         -0.5389      0.060     -9.032      0.000      -0.656      -0.422
ma.L1         -1.9887      0.105    -18.882      0.000      -2.195      -1.782
ma.L2          0.9970      0.106      9.441      0.000       0.790       1.204
sigma2      1.175e+06   1.91e-07   6.16e+12      0.000    1.18e+06    1.18e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                18.11
Prob(Q):                              0.83   Prob(JB):                         0.00
Heteroskedasticity (H):               2.33   Skew:                             0.63
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.10
===================================================================================
```

*Table 8: Summary of Auto Arima Model*

RMSE for Auto Arima Model is **1.652181e+06**

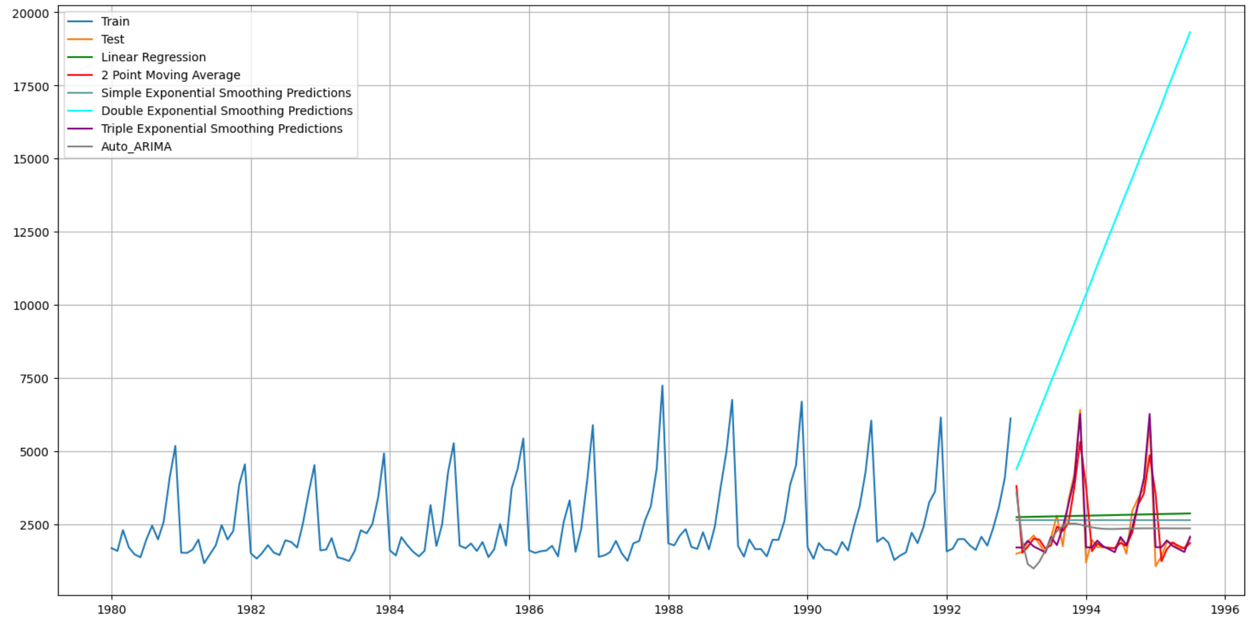
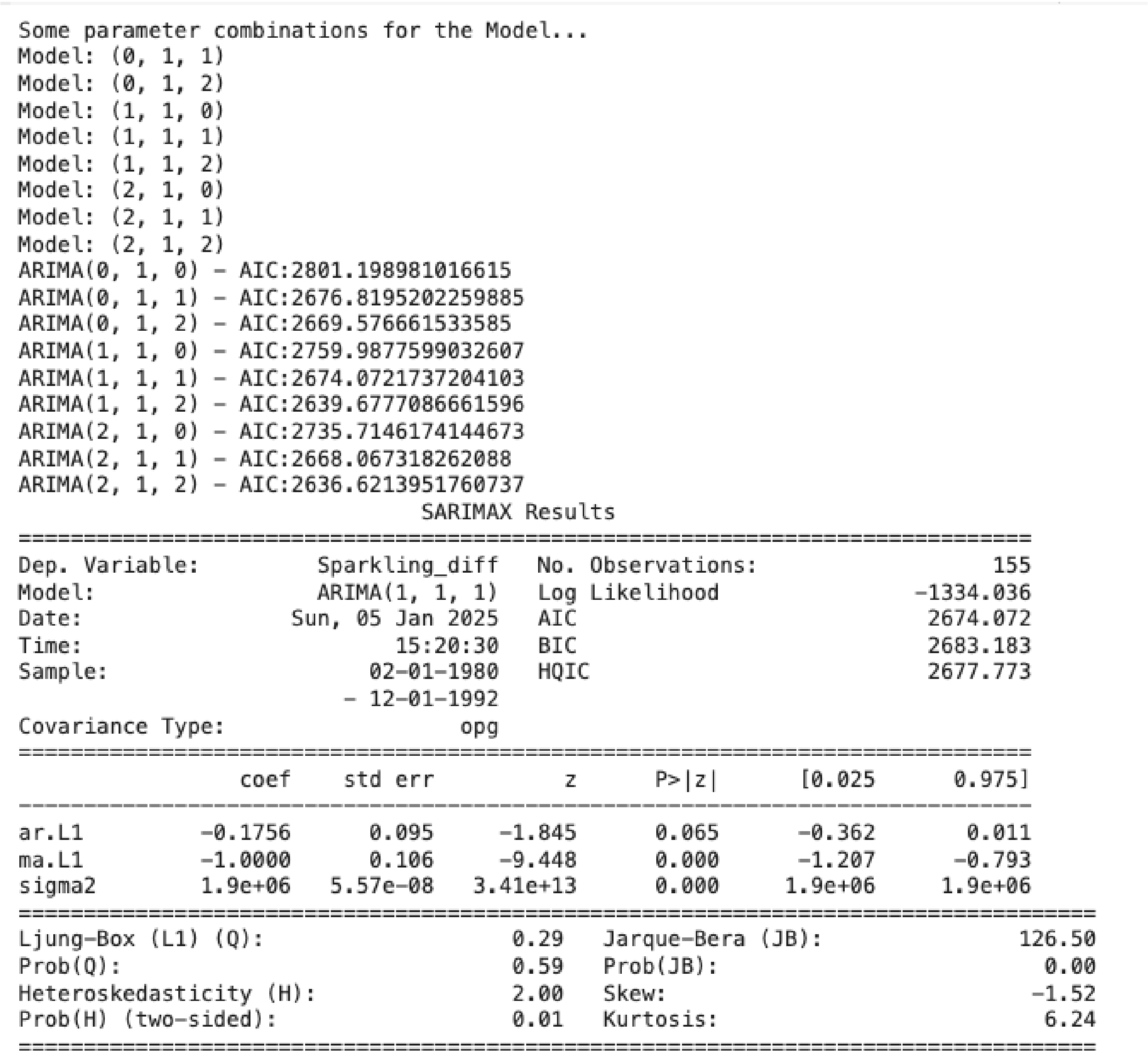


*Fig 12: Model Comparision plot*

## ARIMA Model



```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
ARIMA(0, 1, 0) - AIC:2801.198981016615
ARIMA(0, 1, 1) - AIC:2676.8195202259885
ARIMA(0, 1, 2) - AIC:2669.576661533585
ARIMA(1, 1, 0) - AIC:2759.9877599032607
ARIMA(1, 1, 1) - AIC:2674.0721737204103
ARIMA(1, 1, 2) - AIC:2639.6777086661596
ARIMA(2, 1, 0) - AIC:2735.7146174144673
ARIMA(2, 1, 1) - AIC:2668.067318262088
ARIMA(2, 1, 2) - AIC:2636.6213951760737
                               SARIMAX Results
==============================================================================
Dep. Variable:          Sparkling_diff   No. Observations:                  155
Model:                  ARIMA(1, 1, 1)   Log Likelihood               -1334.036
Date:                 Sun, 05 Jan 2025   AIC                           2674.072
Time:                         15:20:30   BIC                           2683.183
Sample:                     02-01-1980   HQIC                          2677.773
                          - 12-01-1992
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1756      0.095     -1.845      0.065      -0.362       0.011
ma.L1         -1.0000      0.106     -9.448      0.000      -1.207      -0.793
sigma2         1.9e+06   5.57e-08   3.41e+13      0.000     1.9e+06     1.9e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.29   Jarque-Bera (JB):               126.50
Prob(Q):                              0.59   Prob(JB):                         0.00
Heteroskedasticity (H):               2.00   Skew:                            -1.52
Prob(H) (two-sided):                  0.01   Kurtosis:                         6.24
===================================================================================
```

*Table 9: Summary of Arima Model*

RMSE for Arima Model is also **1.652181e+06**

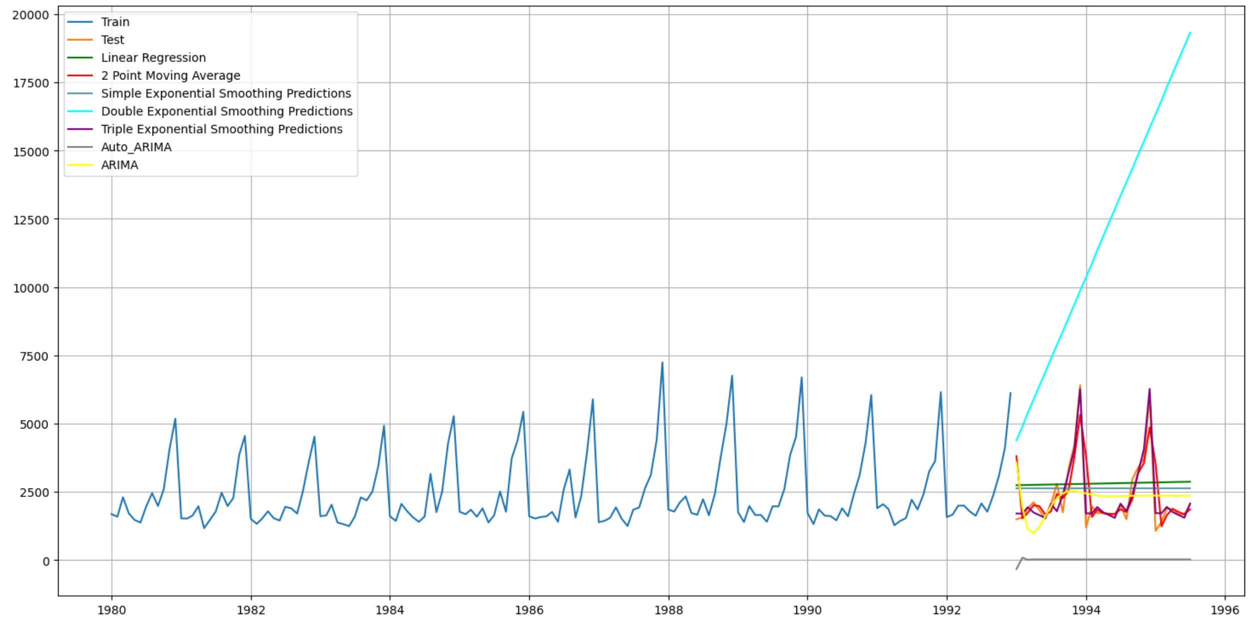
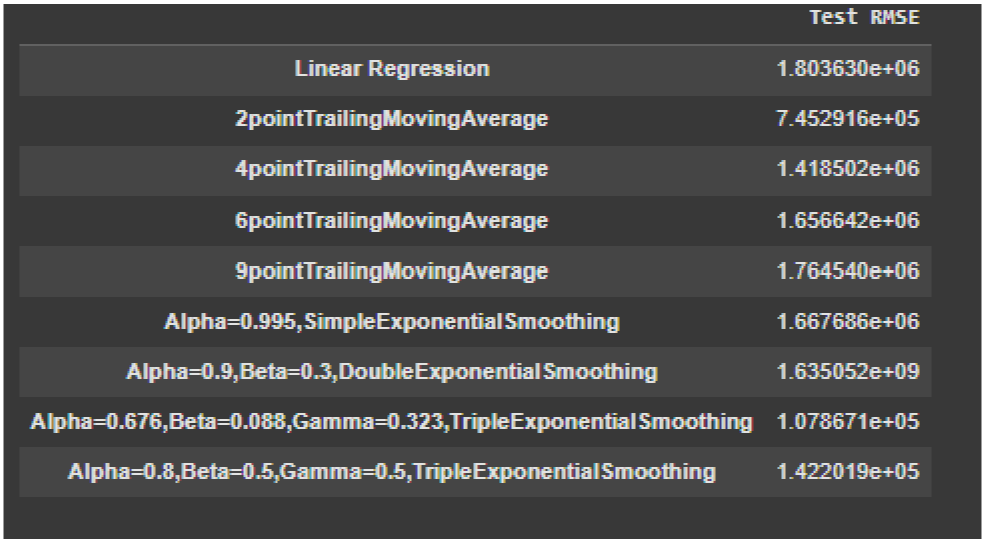


*Fig 13: Model Comparision plot*

## Comparing the Performance of the model



| | Test RMSE |
|---|---|
| Linear Regression | 1.803630e+06 |
| 2pointTrailingMovingAverage | 7.452916e+05 |
| 4pointTrailingMovingAverage | 1.418502e+06 |
| 6pointTrailingMovingAverage | 1.656642e+06 |
| 9pointTrailingMovingAverage | 1.764540e+06 |
| Alpha=0.995,SimpleExponentialSmoothing | 1.667686e+06 |
| Alpha=0.9,Beta=0.3,DoubleExponentialSmoothing | 1.635052e+09 |
| Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing | 1.078671e+05 |
| Alpha=0.8,Beta=0.5,Gamma=0.5,TripleExponentialSmoothing | 1.422019e+05 |

Linear Regression has the highest RMSE value of 1.80e+06, suggesting it is the least effective model for this time series data.

2-point Trailing Moving Average has the lowest RMSE of 745,291.6, making it the most accurate model for this dataset.

Linear Regression has a significantly higher RMSE of 1,803,630, indicating that it is less accurate compared to moving average models.

Among the other moving average models:

- 4-point Trailing Moving Average (RMSE: 1,418,502) is the second-best moving average model.
- 6-point and 9-point Trailing Moving Averages have higher RMSE values (1,656,642 and 1,764,542, respectively), showing decreasing accuracy as the number of points increases.

Simple Exponential Smoothing (Alpha = 0.995) has an RMSE of 1,667,686, which is better than linear regression but worse than moving averages.

Double Exponential Smoothing (Alpha = 0.9, Beta = 0.3) performs poorly, with an extremely high RMSE of 1,635,052,000, making it unsuitable for this dataset.

Triple Exponential Smoothing:

- Model with Alpha = 0.676, Beta = 0.088, Gamma = 0.323 has a low RMSE of 1,078,671, making it the second-best model overall.
- Model with Alpha = 0.8, Beta = 0.5, Gamma = 0.5 has an RMSE of 1,422,019, which is moderately accurate but not as effective as the 2-point Trailing Moving Average.

## Business Recommendations & Insights

- The **2-point Trailing Moving Average** is the best model based on its **lowest RMSE** value of **745,291.6**, indicating superior accuracy for forecasting or prediction tasks on this dataset.

- Utilize the observed seasonal patterns to design targeted marketing campaigns during peak sales months, such as holidays and festive seasons.

- Align inventory planning with forecasted seasonal demand to avoid stockouts during peak periods and minimize holding costs during off-peak months.

- Focus on digital marketing and social media promotions to attract younger demographics and expand customer reach beyond traditional sales channels.

## Conclusion

The analysis of the Sparkling wine dataset indicates significant seasonal trends and periodic fluctuations in sales volumes. Peaks in sales correspond to specific times of the year, likely driven by holidays or celebrations. The decomposition analysis highlights a strong seasonal component, emphasizing the need for inventory and marketing strategies that align with peak demand periods. Forecasting models predict continued growth with sustained seasonal patterns, offering opportunities for strategic promotions and supply chain optimization. Recommendations include leveraging data-driven insights to target marketing campaigns, ensuring stock availability during high-demand periods, and exploring market expansion opportunities to capitalize on consistent growth trends.

THE END