05/01/2025

Time Series Forecasting - ABC Estate Wines

[Rose 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

Rose Dataset

- **YearMonth (Date):** Represents the year and month of the sales data in YYYY-MM format.
- **Rose (Float):** Monthly sales volume of Rose 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	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

Table 1: First 5 rows of the given dataset

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

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   Rose        185 non-null    float64
dtypes: float64(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 Rose column have less entries (less than 187 rows), Indicates that there are two missing values are present in rose column.
- Here, YearMonth column is identified as object data types which needs to convert to datetime format for time-series analysis.
- Rose Column is of float 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
Rose	185.0	NaN	NaN	NaN	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

Table 4: Statistical summary of the data

- The **range** of "Rose" values is large (from 28 to 267), highlighting high variability in the data.
- The mean is higher than the median (86.0), suggesting a right-skewed distribution with some high values.

INFERENCE:

- Missing Values are treated with mean.
- 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.

YearMonth	Rose
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

Table 5: Indexed data for Time Series Analysis

Exploratory Data Analysis

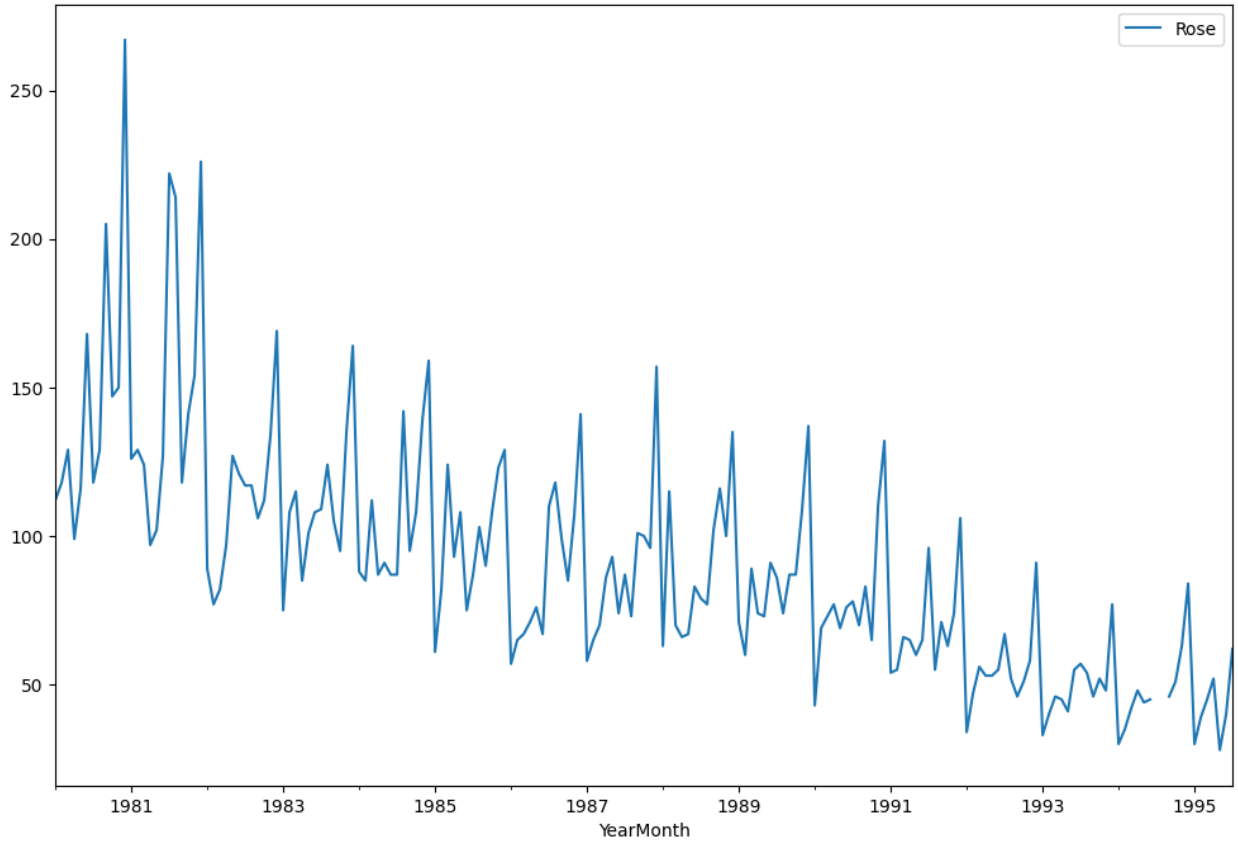

Fig 1: Trend of the Rose wine sales

OBSERVATIONS:

Trend:

- There is an evident downward trend in the data over the years (1981–1995). The values of the series steadily decline, indicating a decrease in the measured variable (e.g., production, sales, or demand of "Rose").
- Peaks in earlier years (1980–1983) are significantly higher than those in later years (1990–1995), suggesting a long-term reduction in magnitude.

Seasonality:

- There appears to be seasonal variation in the data, as evidenced by the recurring peaks and troughs at regular intervals (likely corresponding to specific months or seasons).

- The magnitude of the seasonal peaks decreases over time but retains a somewhat regular pattern, indicating that while the overall value is declining, the seasonal cycles remain consistent in their periodicity.

Variability (peaks and troughs) is higher in the earlier years, but the fluctuations reduce in magnitude over time.

Decomposition

Additive Decomposition Summary

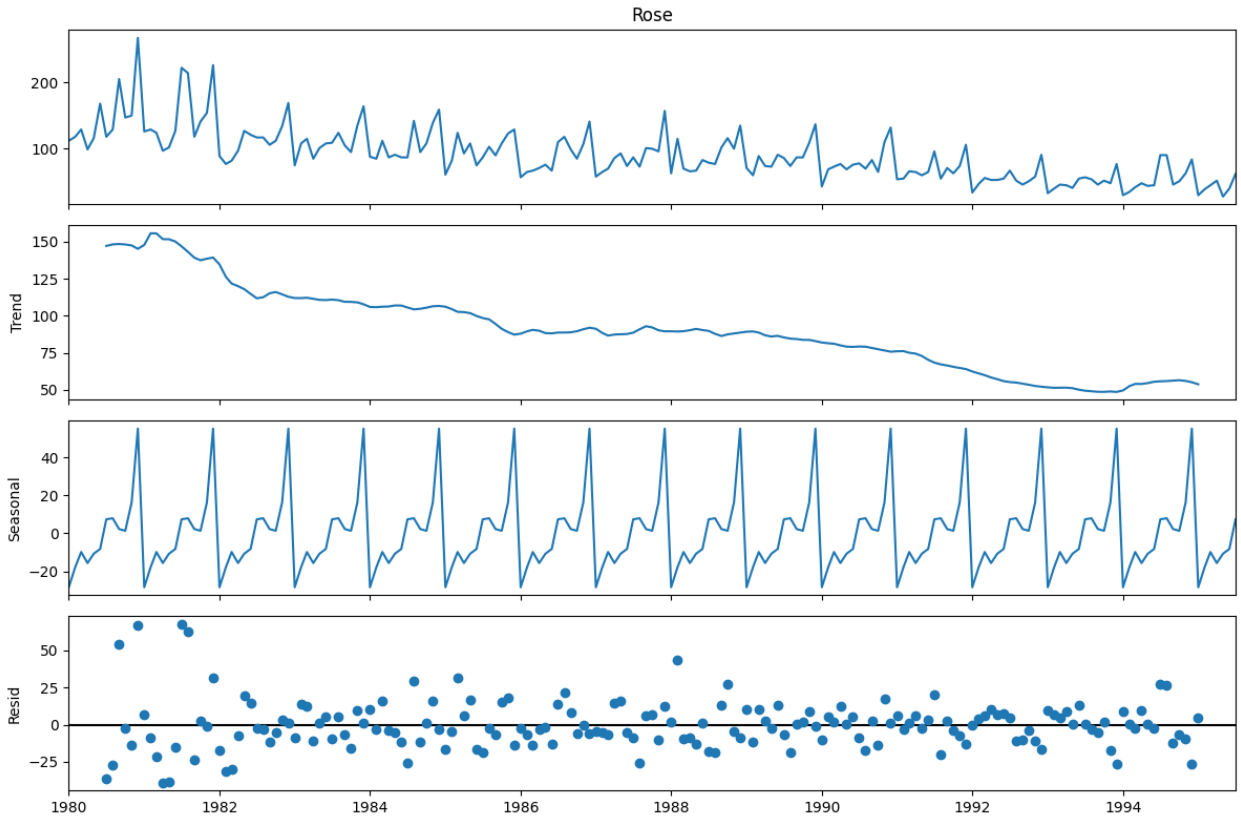

Fig 1: Additive Decomposition Summary

Observed Component: The time series exhibits clear seasonal patterns along with variations in the trend. Peaks and troughs occur periodically, indicating seasonality.

Trend Component: The trend shows a steady decline over time, suggesting that the overall level of the time series is decreasing.

Seasonal Component: The seasonal variation is consistent in amplitude, repeating periodically, independent of the overall trend.

Residual Component: The residuals are scattered around zero, with no clear patterns, suggesting that the additive model captures most of the systematic variation in the data.

Multiplicative Decomposition Summary

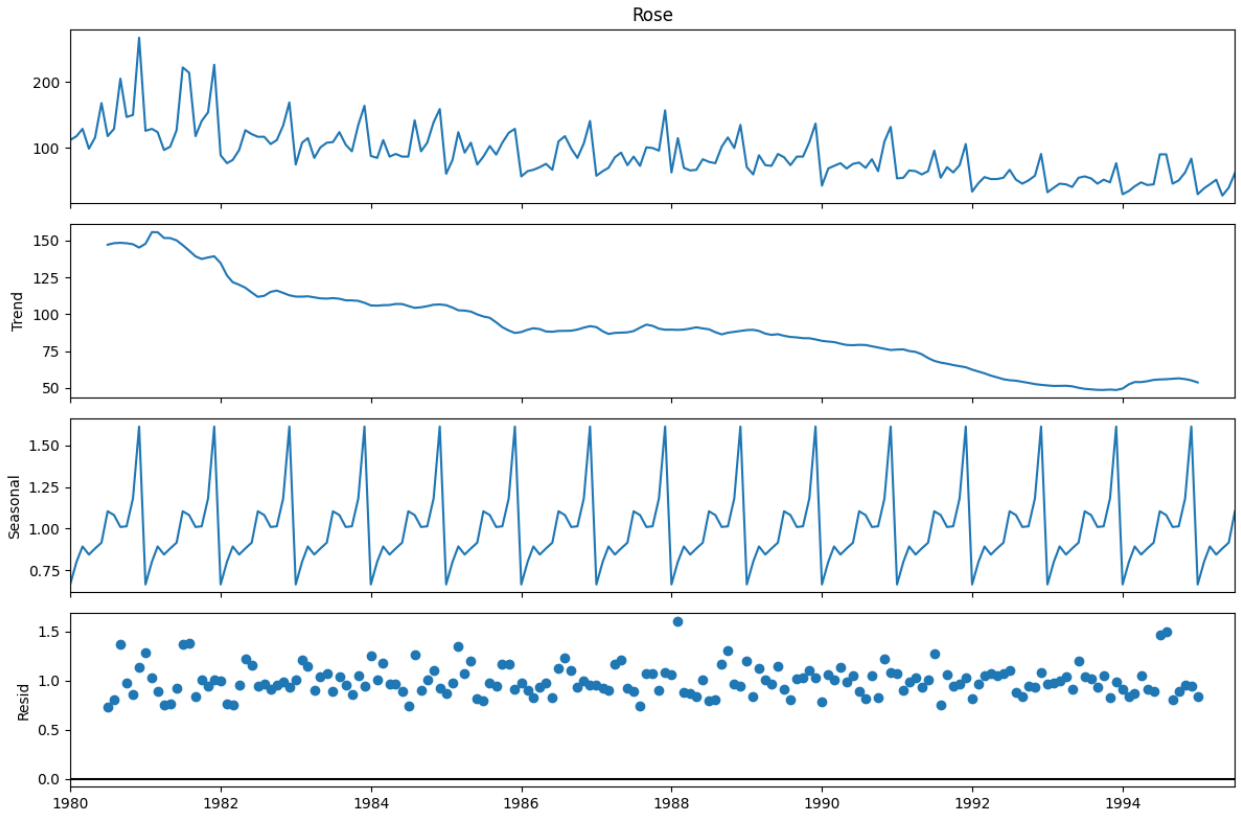

Fig 1: Multiplicative Decomposition Summary

Observed Component: Similar to the additive decomposition, the time series has strong periodic seasonal patterns, but the amplitude of seasonality seems proportional to the trend level.

Trend Component: The trend component also shows a decreasing pattern over time, similar to the additive case.

Seasonal Component: The seasonal effect is expressed as a ratio or proportion of the trend, showing periodic patterns with values that are greater than 1 during peak seasons and less than 1 during low seasons.

Residual Component: The residuals are scattered and proportional, which indicates that the multiplicative model effectively captures the proportional nature of seasonality in the data.

Data Preprocessing

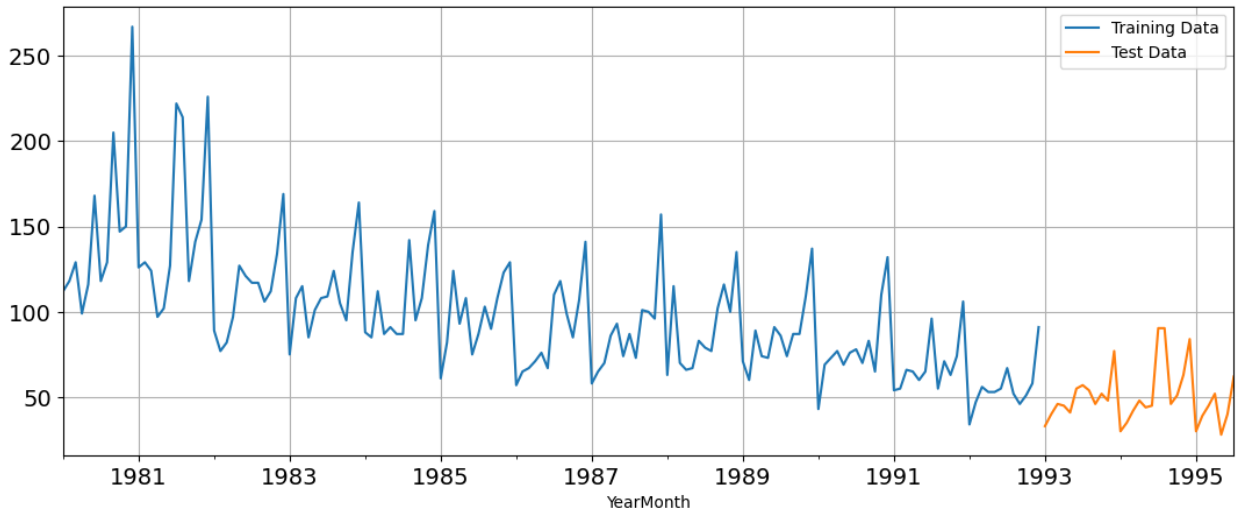

Fig 4: Plot of Train and Test Data

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

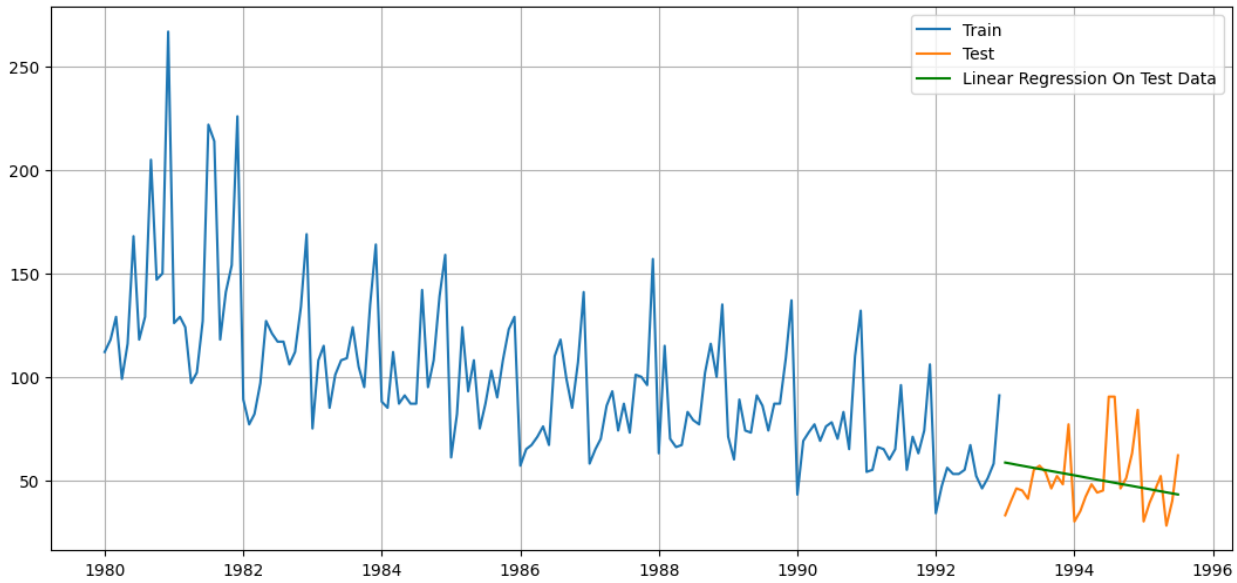

Fig 5: Linear Regression On Test Data

- RMSE on Linear Regression Test data - **296.574259**

Moving Average

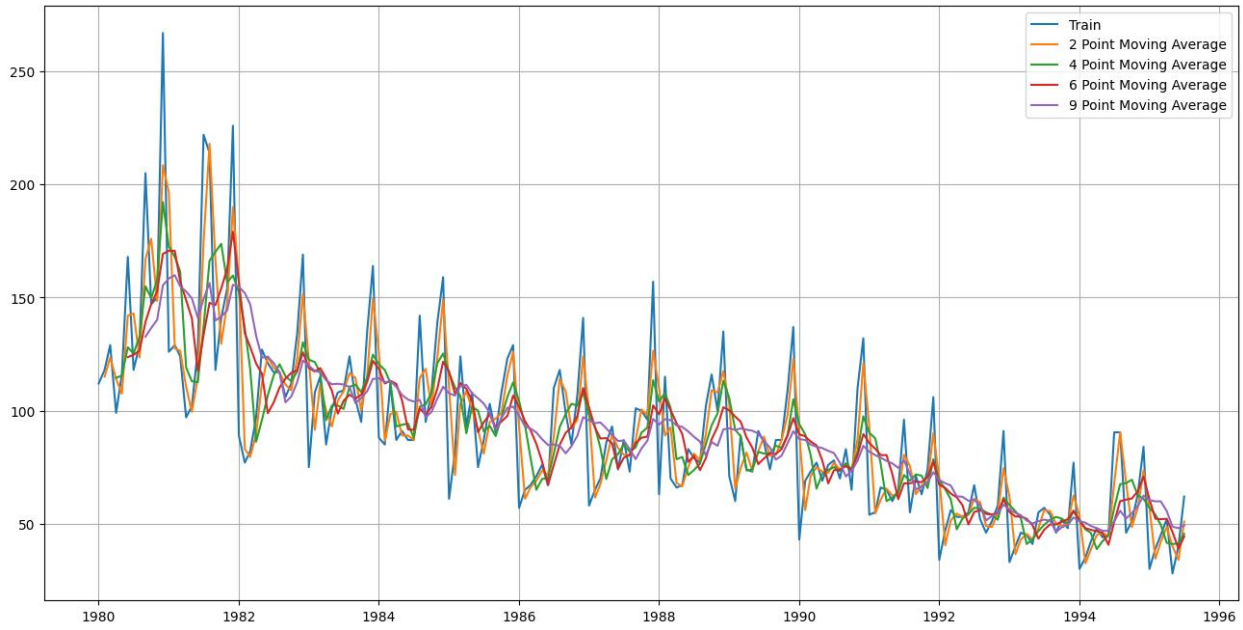

Fig 6: Trailing Moving Average of 2, 4, 6, 9 points

For **2 point Moving Average Model** forecast on the Testing Data, RMSE is **128.189913**

For **4 point Moving Average Model** forecast on the Training Data, RMSE is **220.594570**

For **6 point Moving Average Model** forecast on the Training Data, RMSE is **227.387686**

For **9 point Moving Average Model** forecast on the Training Data, RMSE is **254.211371**

- 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.
- The 2-point Trailing Moving Average has the lowest RMSE (Root Mean Squared Error) value of 128.189913, indicating it performs best among the models tested.

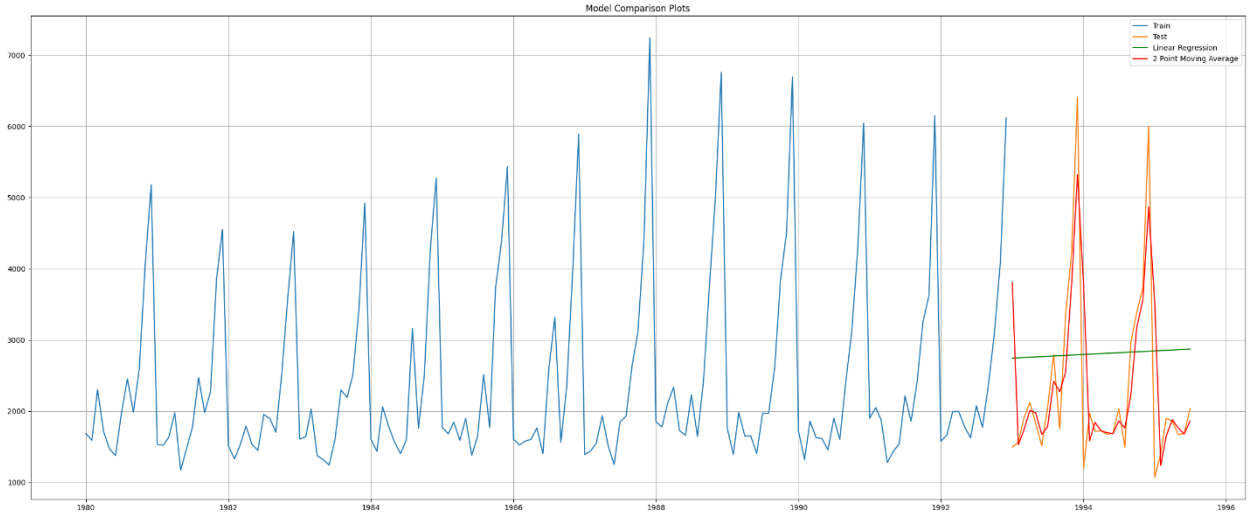

Fig 7: Model Comparision plot

Simple Exponential Smoothening Models

```
{'smoothing_level': 0.12948030833097124,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 112.0,
 '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 0.129

For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is **405.965108**

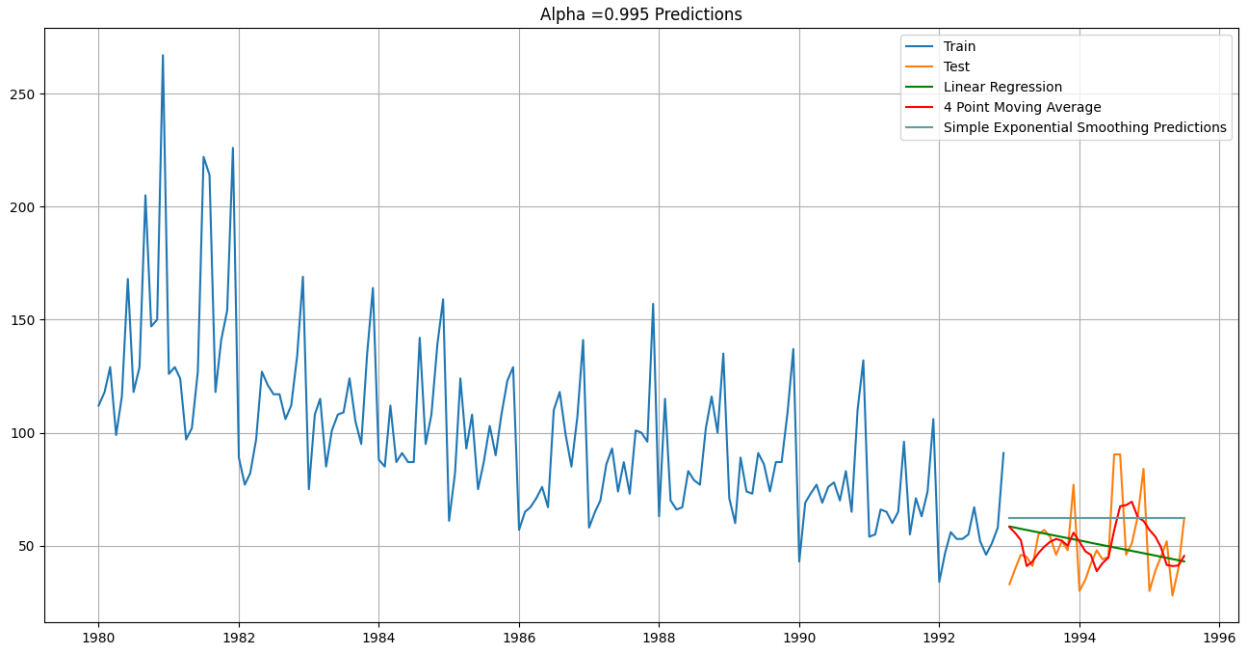

Fig 8: Model Comparision plot

Double Exponential Smoothening (Holt's Model)

For Alpha=0.9, Beta=0.3 Double Exponential Smoothing Model forecast on the Test Data, RMSE is **409696.775695**

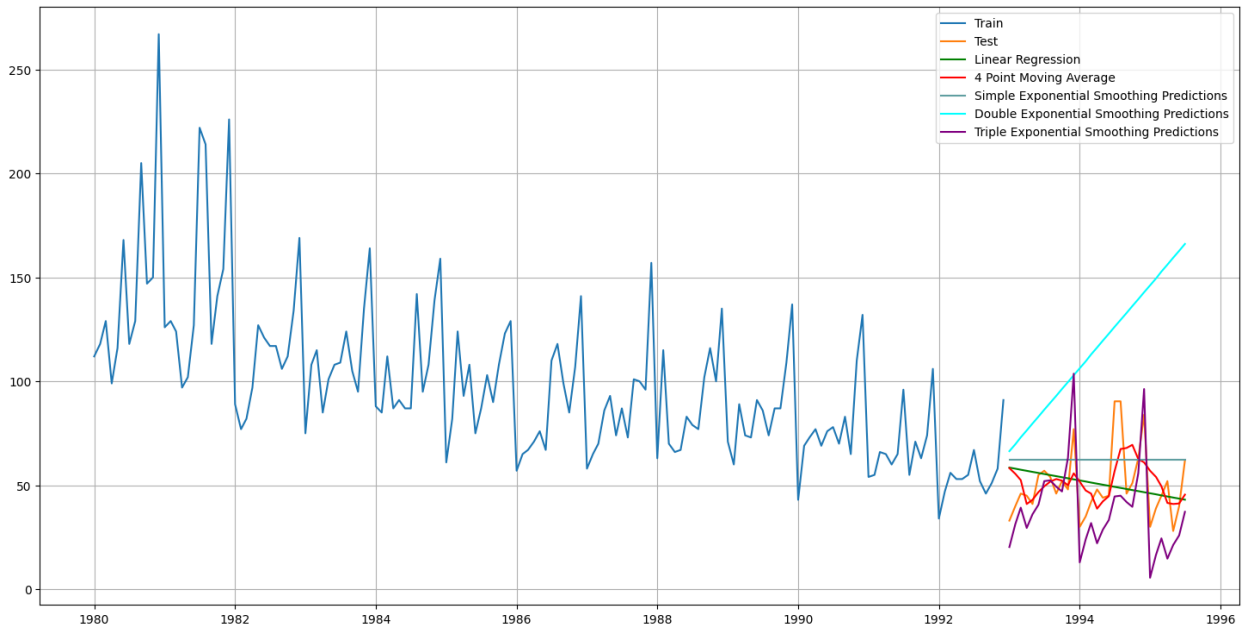

Fig 9: Model Comparision plot

Triple Exponential Smoothing (Holt - Winter's Model)

```
{'smoothing_level': 0.09509083303858493,
 'smoothing_trend': 7.769745611839437e-05,
 'smoothing_seasonal': 0.002521115063116996,
 'damping_trend': nan,
 'initial_level': 146.9050470558056,
 'initial_trend': -0.6151877914985002,
 'initial_seasons': array([-30.02031629, -18.32351441,  -9.8563535 , -18.9937688 ,
        -11.8086869 ,  -6.56456019,   5.38204   ,   6.36920981,
          3.94321345,   2.25447711,  19.12137587,  60.09787995]),
 '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 **374.063568**

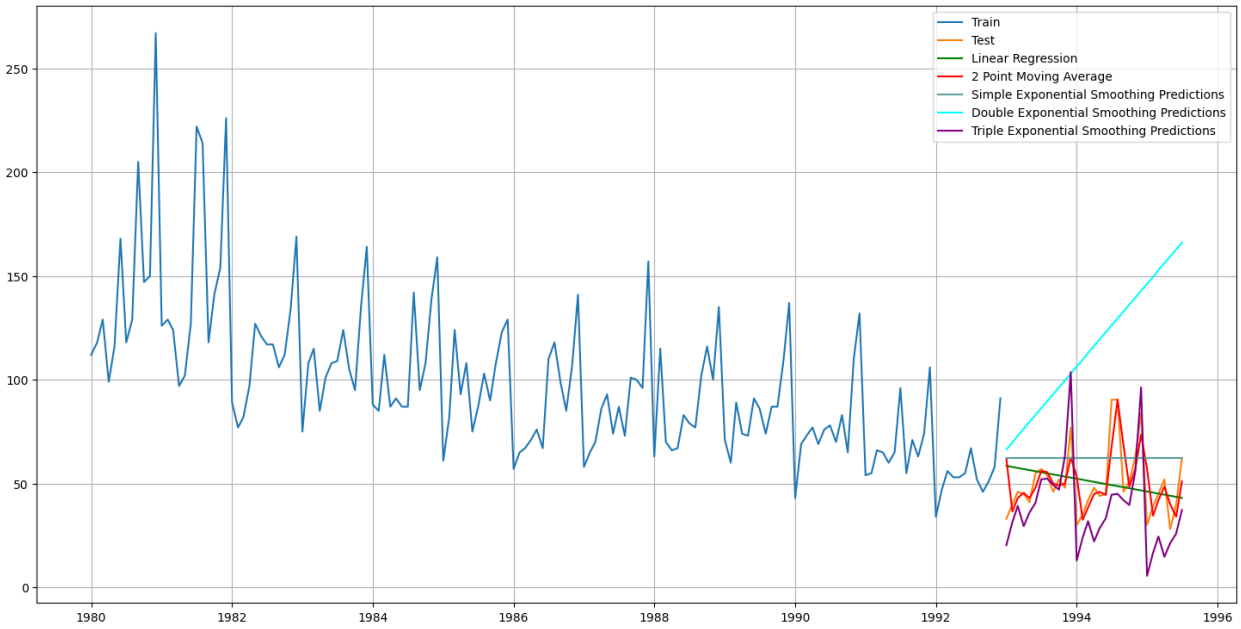

Fig 10: Model Comparision plot

Check for stationarity of the whole Time Series data.

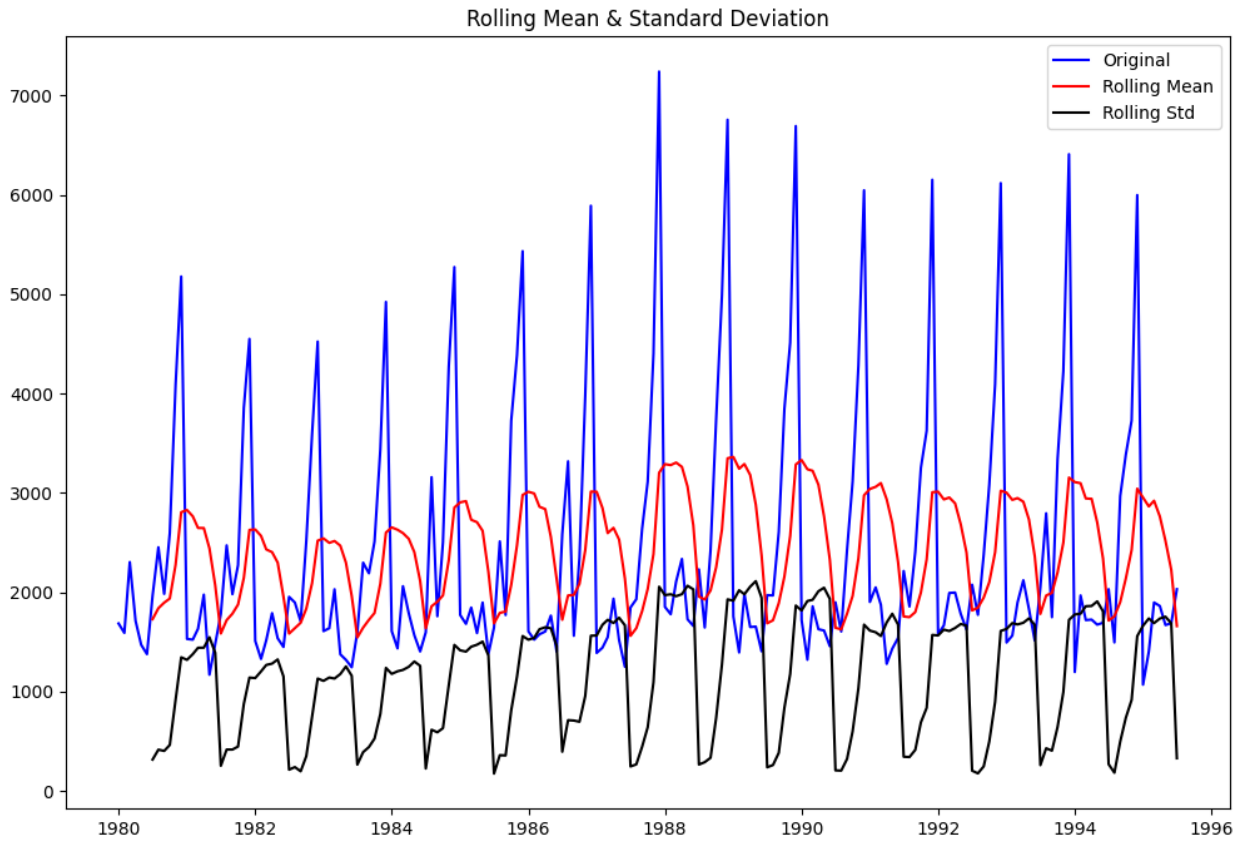

Fig 12: Test for stationarity on the data

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:	Rose	No. Observations:	156
Model:	ARIMA(2, 1, 2)	Log Likelihood	-744.822
Date:	Sun, 05 Jan 2025	AIC	1499.644
Time:	17:46:03	BIC	1514.861
Sample:	01-01-1980	HQIC	1505.825
	- 12-01-1992		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4831	0.370	-1.306	0.191	-1.208	0.242
ar.L2	-0.0140	0.139	-0.101	0.920	-0.286	0.258
ma.L1	-0.2281	0.362	-0.629	0.529	-0.938	0.482
ma.L2	-0.6063	0.338	-1.791	0.073	-1.270	0.057
sigma2	864.9709	74.266	11.647	0.000	719.413	1010.529

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	50.85
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	0.32	Skew:	0.84
Prob(H) (two-sided):	0.00	Kurtosis:	5.25

Table 8: Summary of Auto Arima Model

	param	AIC
5	(1, 1, 2)	1497.663180
2	(0, 1, 2)	1498.350577
4	(1, 1, 1)	1499.367354
8	(2, 1, 2)	1499.643992
7	(2, 1, 1)	1499.825823
1	(0, 1, 1)	1501.042766
6	(2, 1, 0)	1521.176885
3	(1, 1, 0)	1544.681234
0	(0, 1, 0)	1563.960752

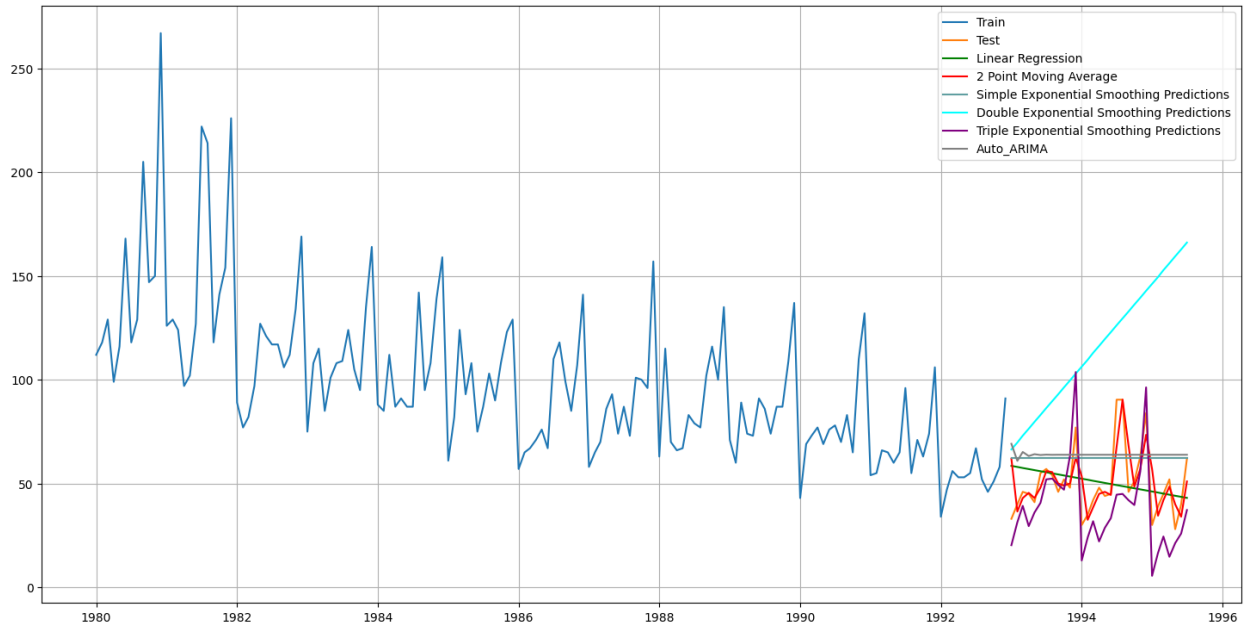

Fig 12: Model Comparision plot

RMSE for Auto Arima Model is **450.24892**

ARIMA Model

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	156
Model:	ARIMA(2, 1, 2)	Log Likelihood	-744.822
Date:	Sun, 05 Jan 2025	AIC	1499.644
Time:	17:55:51	BIC	1514.861
Sample:	01-01-1980	HQIC	1505.825
	- 12-01-1992		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4831	0.370	-1.306	0.191	-1.208	0.242
ar.L2	-0.0140	0.139	-0.101	0.920	-0.286	0.258
ma.L1	-0.2281	0.362	-0.629	0.529	-0.938	0.482
ma.L2	-0.6063	0.338	-1.791	0.073	-1.270	0.057
sigma2	864.9709	74.266	11.647	0.000	719.413	1010.529

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	50.85
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	0.32	Skew:	0.84
Prob(H) (two-sided):	0.00	Kurtosis:	5.25

Table 9: Summary of Arima Model

	param	AIC
8	(2, 1, 2)	2636.621395
17	(2, 1, 2)	2636.621395
5	(1, 1, 2)	2639.677709
14	(1, 1, 2)	2639.677709
7	(2, 1, 1)	2668.067318
16	(2, 1, 1)	2668.067318
2	(0, 1, 2)	2669.576662
11	(0, 1, 2)	2669.576662
4	(1, 1, 1)	2674.072174
13	(1, 1, 1)	2674.072174
10	(0, 1, 1)	2676.819520
1	(0, 1, 1)	2676.819520
6	(2, 1, 0)	2735.714617
15	(2, 1, 0)	2735.714617
3	(1, 1, 0)	2759.987760
12	(1, 1, 0)	2759.987760
9	(0, 1, 0)	2801.198981
0	(0, 1, 0)	2801.198981

RMSE for Arima Model is also **450.24892**

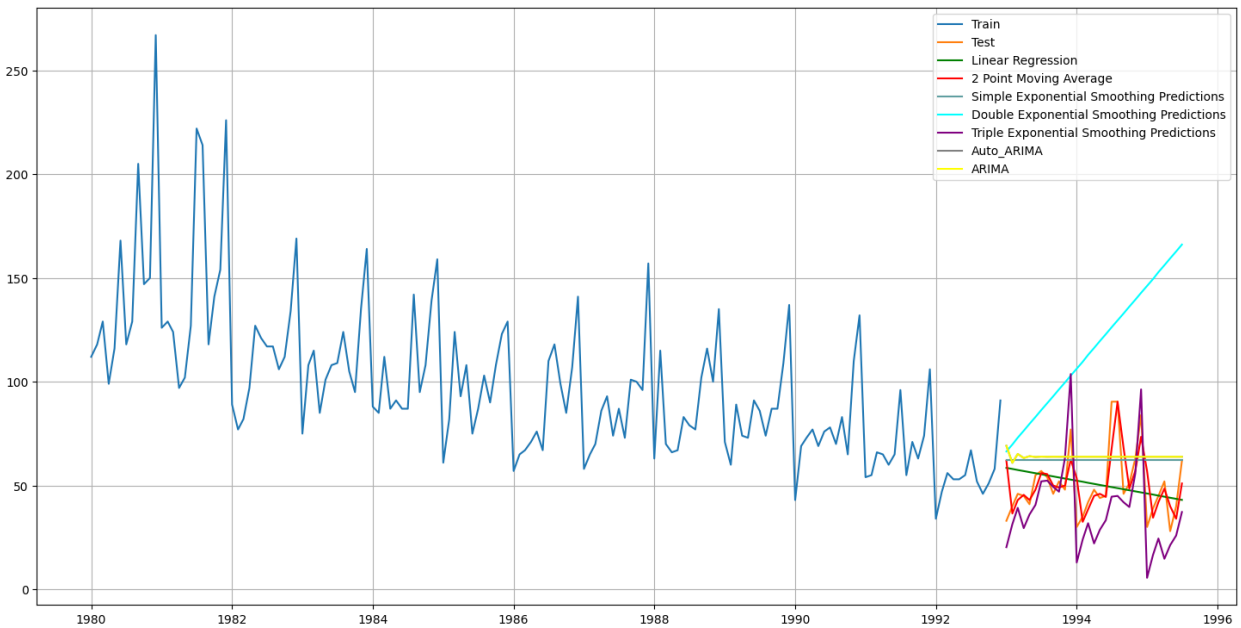

Fig 13: Model Comparision plot

Comparing the Performance of the model

	Test RMSE
Linear Regression	296.574259
2pointTrailingMovingAverage	128.189913
4pointTrailingMovingAverage	220.594570
6pointTrailingMovingAverage	227.387686
9pointTrailingMovingAverage	254.211371
Alpha=0.995,SimpleExponentialSmoothing	405.965108
Alpha=0.9,Beta=0.3,DoubleExponentialSmoothing	409696.775695
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	374.063568
Alpha=0.8,Beta=0.5,Gamma=0.5,TripleExponentialSmoothing	205.066163

2-point Trailing Moving Average has the lowest RMSE of 128.189913, making it the best-performing model among the tested approaches.

Linear Regression has a higher RMSE of 296.574259, indicating less accuracy compared to moving average models.

Other moving average models (4-point, 6-point, and 9-point) perform moderately well, but their RMSE values (220.549570, 227.387686, and 254.211371, respectively) are higher than the 2-point Trailing Moving Average.

Simple Exponential Smoothing (Alpha = 0.995) has an RMSE of 405.965108, which is less accurate than all moving average models.

Double Exponential Smoothing (Alpha = 0.9, Beta = 0.3) performs the worst, with the highest RMSE of 40996.775695, likely due to poor parameter tuning or unsuitability for the dataset.

Triple Exponential Smoothing with Alpha = 0.8, Beta = 0.5, and Gamma = 0.5 performs relatively well, with an RMSE of 205.066163, but it is still not as effective as the 2-point Trailing Moving Average.

Business Recommendations & Insights

- The 2-point Trailing Moving Average model is the most accurate model based on the RMSE value and should be preferred for forecasting or predictions on this dataset.

- Introduce promotional discounts and bundled offers during high-demand seasons to maximize revenue and attract more customers.

- Explore new wine variants or complementary products to capitalize on existing trends and cater to evolving consumer preferences.

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

Conclusion

The analysis of the Rose wine dataset reveals steady growth in sales with moderate seasonal variations. While the sales patterns are less volatile than Sparkling wine, they still exhibit periodic trends that may be linked to seasonal preferences or marketing campaigns. The decomposition analysis highlights a stable trend component, suggesting consistent demand over time. Forecasting models indicate gradual growth, providing opportunities for sustained marketing efforts and inventory planning. Recommendations include focusing on targeted promotions to boost sales during identified seasonal peaks, optimizing distribution channels to maintain steady supply, and leveraging data analytics to identify emerging market trends for further expansion.

THE END