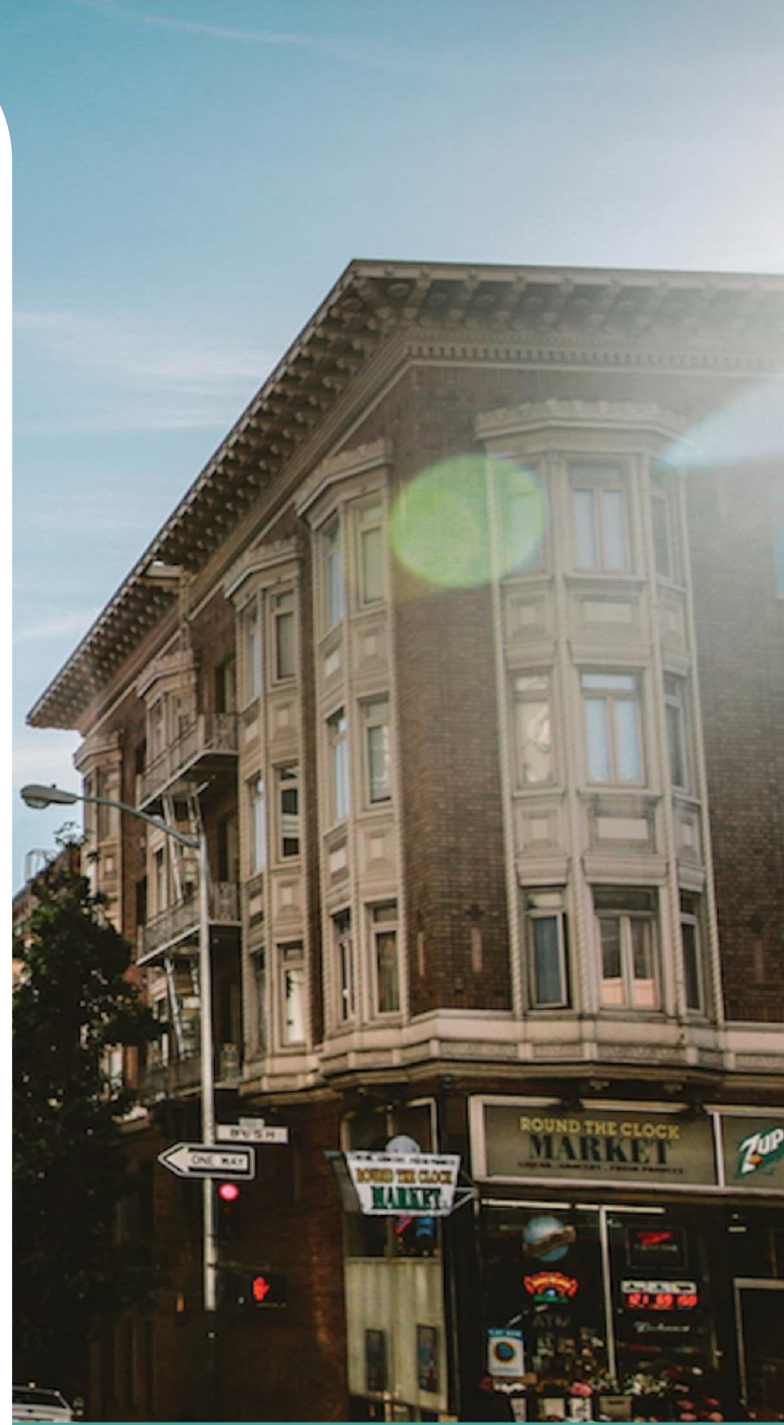
# TIME SERIES FORECASTING - 2024




PGP DSBA PROGRAM

by: ABHISHEK K HIREMATH

Great Learning
POWER AHEAD

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

# List of Figures

# List of Tables

# Problem 1: Sparkling Wine Sales

## Problem Statement:

**As an analyst in the ABC Estate Wines, your task is to analyses and forecast Wine Sales in the 20th century. Data set for the Problem: Sparkling.csv**

Data Dictionary:

Year Month: **Month & Year of the sale**

Sparkling: **Total Number of Sparkling Wine sales in particular Month-Year**

1.1 Read the data as an appropriate Time Series data and plot the data. Read the data as an appropriate Time Series data and plot the data.

Basic Information about the dataset

- Sample of the dataset: **First & last 5 values of the dataset:**

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

**Table 1.1 First 5 Samples of the Dataset**

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

**Table 1.2 Last 5 Samples of the Dataset**

- **Converting the** YearMonth **Column to** Date time Index & dropping **default index. Sample:**

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

**Table 1.3 First 5 Samples of the Converted Dataset**

➢ **Information about the dataset:**

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

**Table.4 Infoof theDataset**

- The DataFrame has 187 entries with a Date time Index ranging from January 1980 to July 1995.
- The 'Sparkling' column is of integer type (int64), and it has 187 non-null values.

➢ **Time Series Plot:**

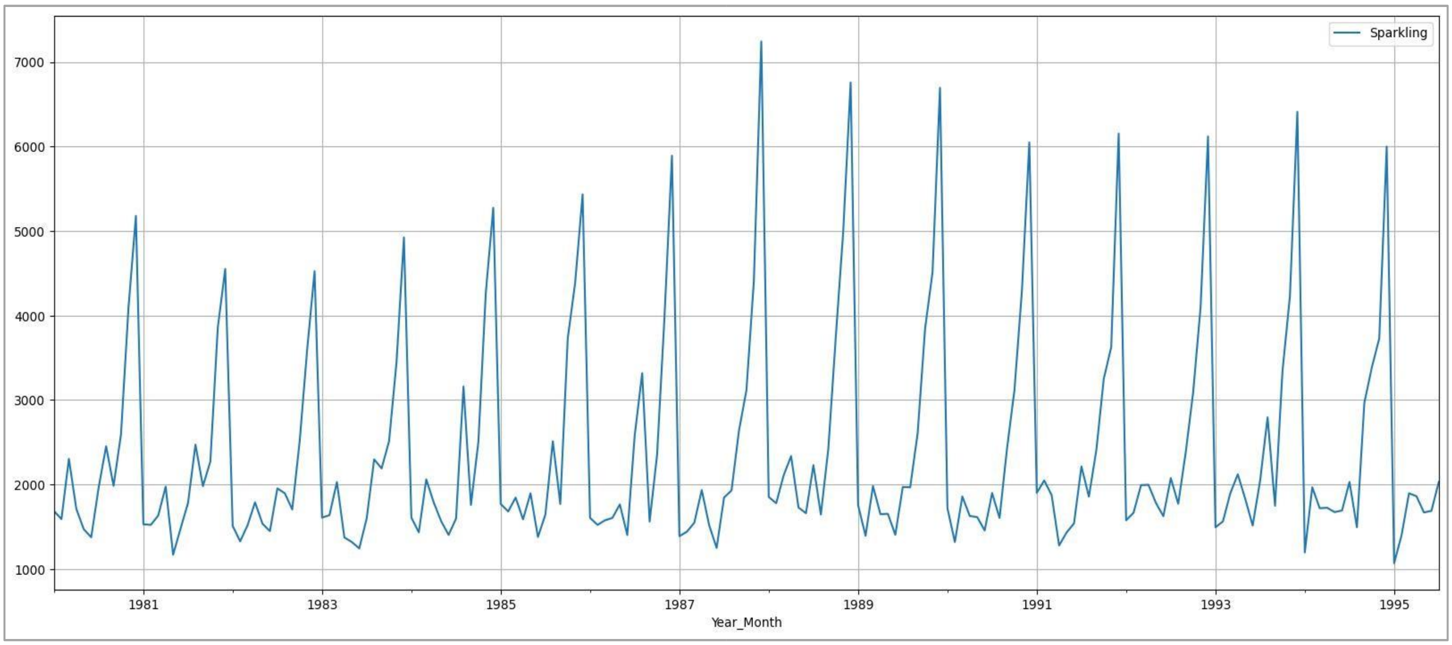


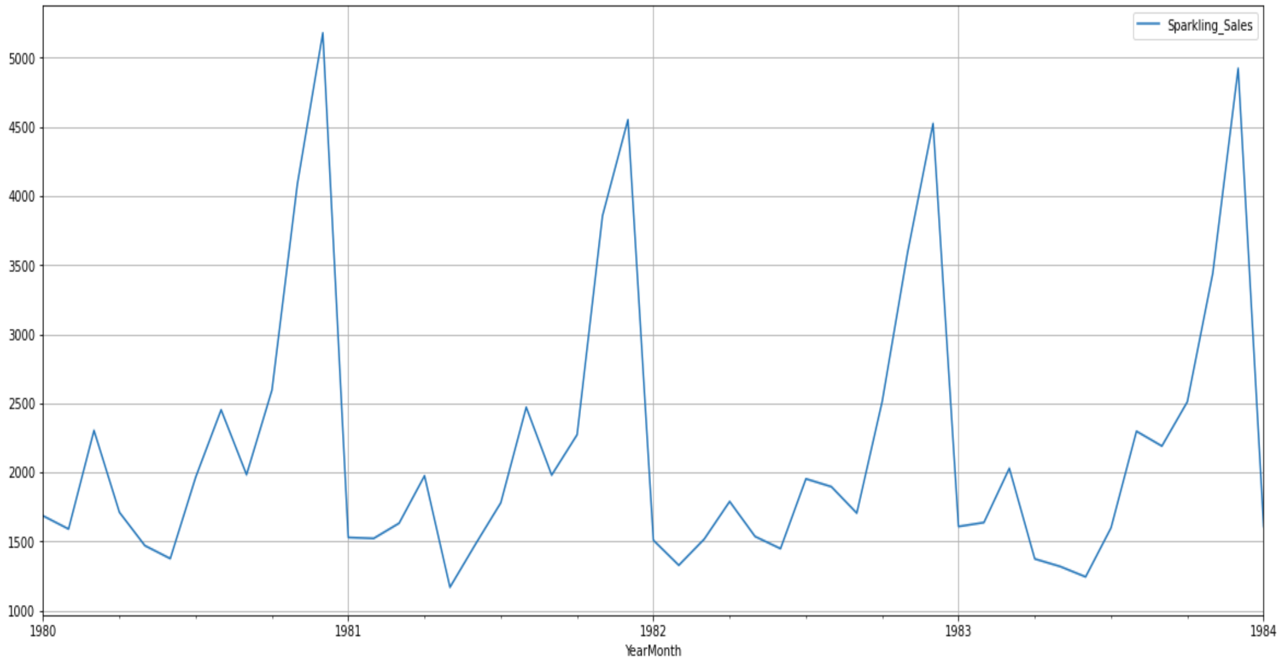

**Fig1.1 Time Series Plot**

- Increasing Trend: The sales of Sparkling wine have been steadily increasing over the years, indicating a positive trend in customer demand.
- Seasonal Patterns: We observe that there are specific periods each year when sales spike, especially during November and December. These peaks might be due to the holiday season, when people tend to buy more

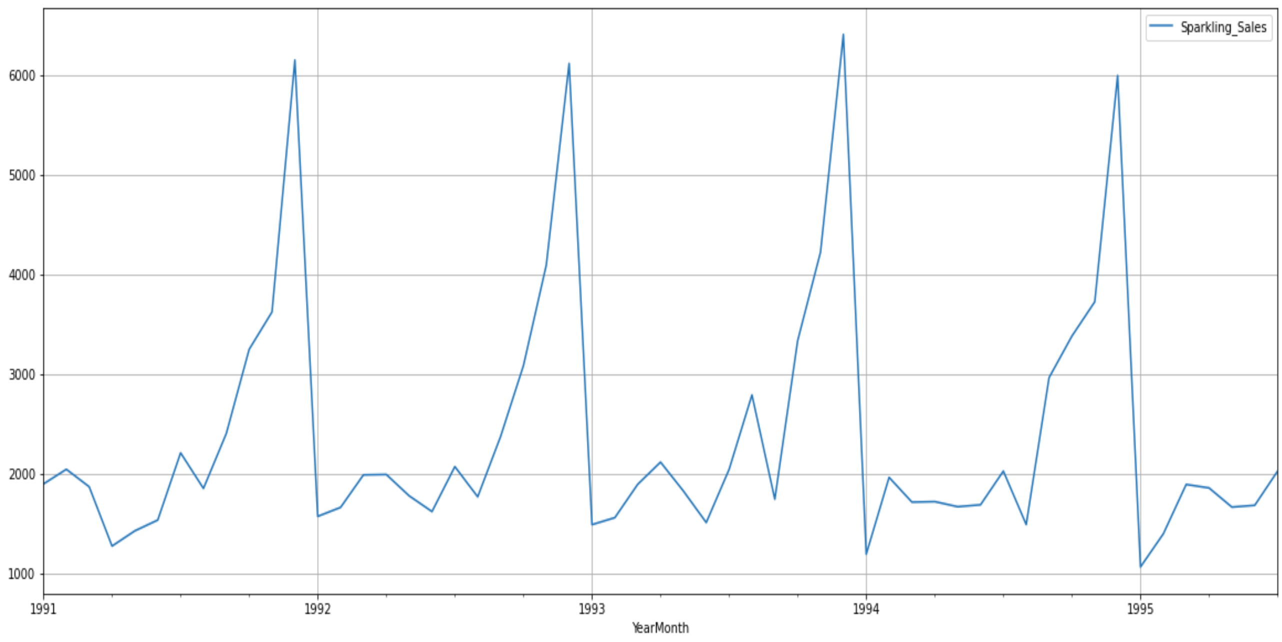
  Sparkling products for celebrations

Historical data will enable accurate forecasting for better planning.

The first 4 years, starting January 1980:




The last few years, starting January 1991:




We will explore the data in detail in the next section.

## 1.2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

- **Missing Values:** o **There are no missing values in the dataset ➢ Duplicate Values:** o **The dataset shows 11 Duplicates rows for Sparkling wine sales, however, when checked further, these were the same no of sales at different year. Hence, we conclude that there are no duplicate values ➢ Descriptive Statistics:**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Sparkling** | 187.0 | 2402.417112 | 1295.11154 | 1070.0 | 1605.0 | 1874.0 | 2549.0 | 7242.0 |

**Table 1.5 Descriptive Statistics**

- The dataset contains 187 observations of sparkling wine sales.
- On average, there are approx. 2402 sales
- Half of the sales fall below 1874, indicating a balanced distribution around the median value.
- The sales data has a moderate level of variability, with a standard deviation of about 1295.
- The minimum recorded sales for sparkling products are 1070, while the maximum is 7242.

- **Histogram & Boxplot**

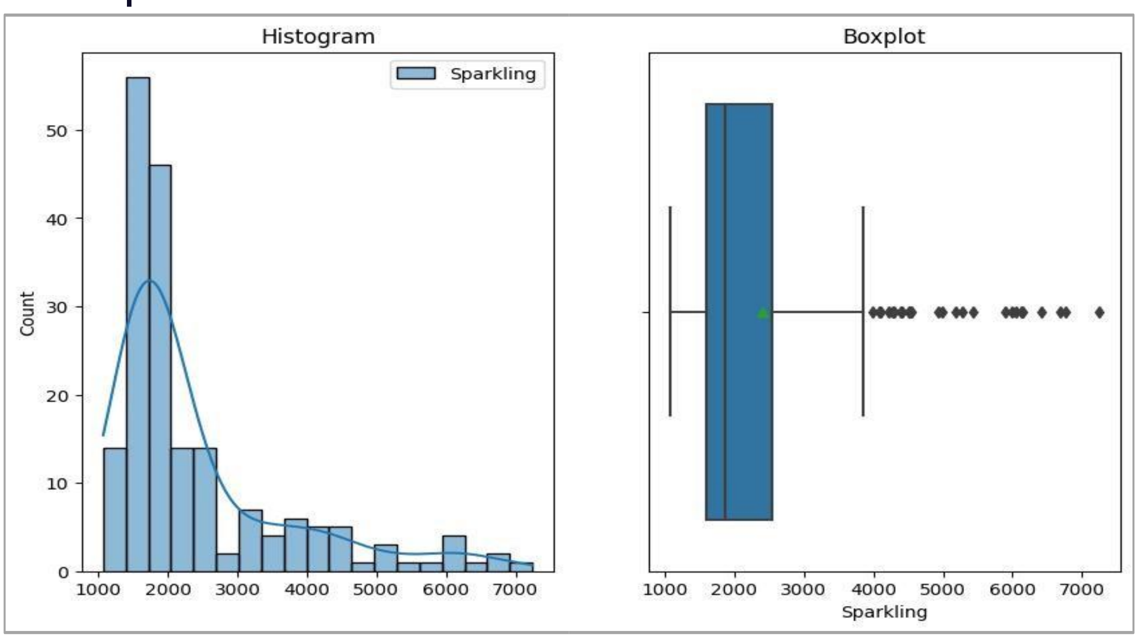



**Fig1.2 Histogram&Boxplot**

- the dataset is right skewed with the presence of outliers on the right tail

➢ **Spread of Sales: Year-on-Year Boxplot**

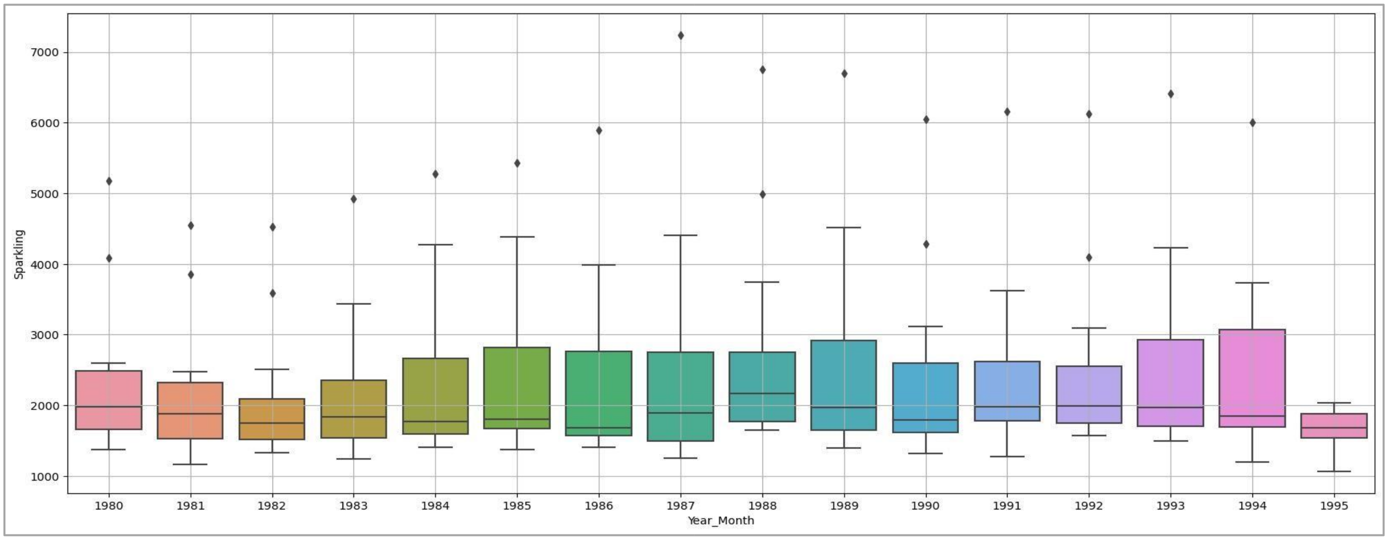



**Fig1.3 SpreadofSale Across Different Year**

- Sales of Sparkling wine has been increasing, with median sales rising over time.
- the dataset is skewed with the presence of outliers on the right tail.
- There is significant variability in sales from year to year, with a large interquartile range.
- There are a few outliers in the data, which could be due to special promotions, holidays or other occasions

➢ **Spread of Sales: Month-on- Month**

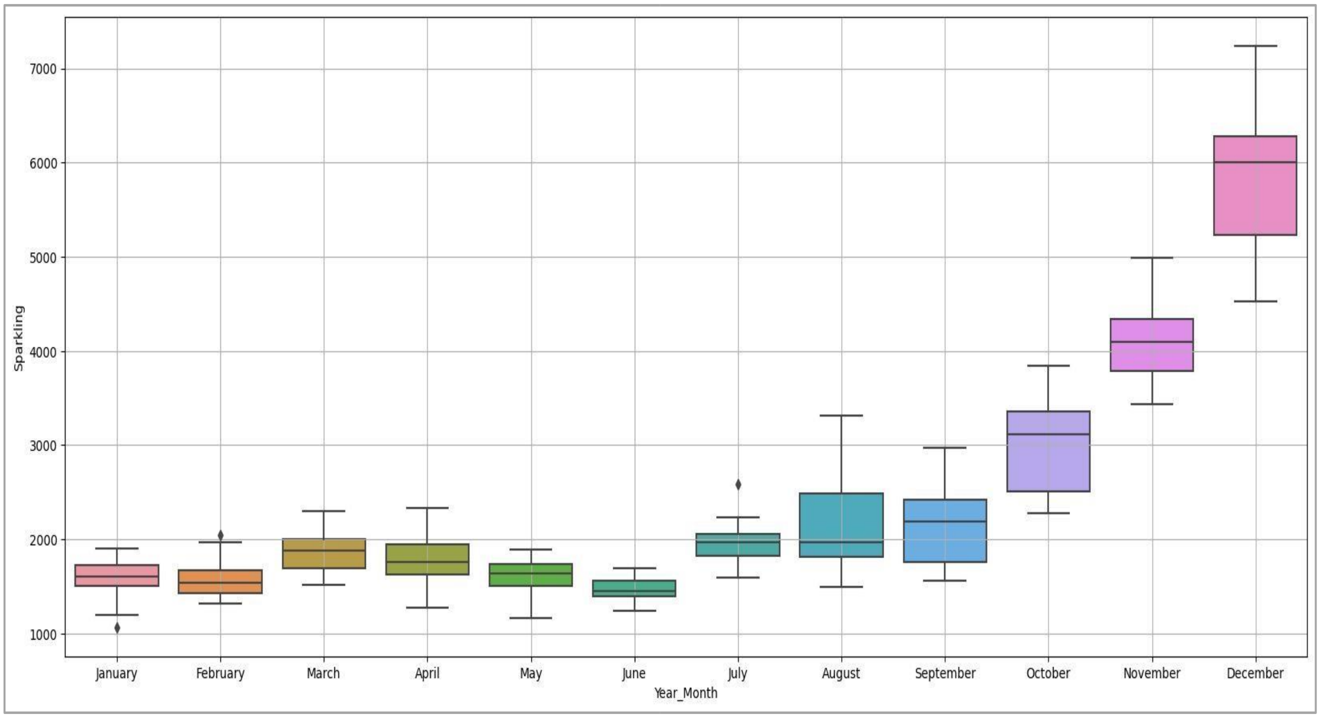



**Fig 1.4 Spread of Sales Across Different Months**

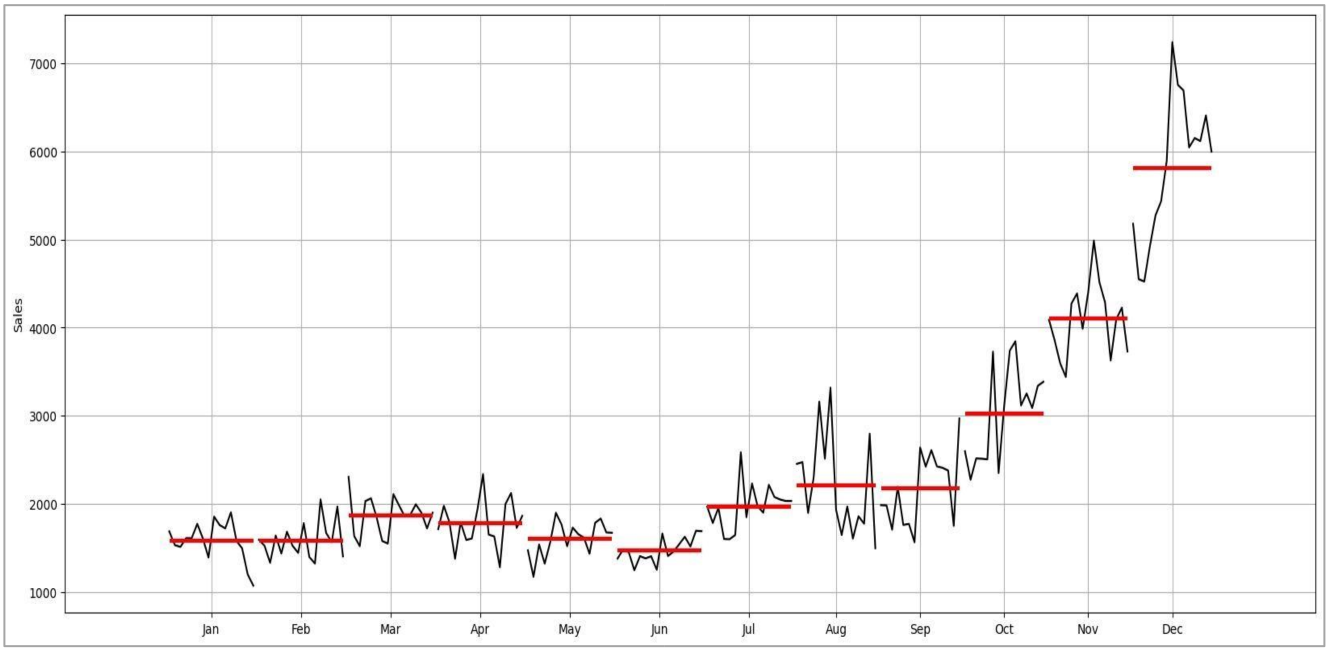



**Fig1.5 Distributionof timeseries acros different months**

- The Month Plot provides insights into the distribution of the Time Series data across different months, with the red line representing the median value.
- Upon analyzing the Month Plot, we can observe a constant (stable) trend that remains consistent across all the years for each month, except for December. Additionally, the plot shows clear seasonal patterns across the months.
- In December, the sales initially dip, then experience an increase, followed by another drop, eventually stabilizing over the years. This pattern in December's sales aligns with the overall trend observed in the entire time series.
- The evidence suggests that December might be a significant month influencing the trend in the time series, given that it has the maximum sales compared to other months.

➢ **Spread of Sales: Year-on-Year & Monthly Comparison:**

| Year_Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year_Month | | | | | | | | | | | | |
| 1980 | 1686.0 | 1591.0 | 2304.0 | 1712.0 | 1471.0 | 1377.0 | 1966.0 | 2453.0 | 1984.0 | 2596.0 | 4087.0 | 5179.0 |
| 1981 | 1530.0 | 1523.0 | 1633.0 | 1976.0 | 1170.0 | 1480.0 | 1781.0 | 2472.0 | 1981.0 | 2273.0 | 3857.0 | 4551.0 |
| 1982 | 1510.0 | 1329.0 | 1518.0 | 1790.0 | 1537.0 | 1449.0 | 1954.0 | 1897.0 | 1706.0 | 2514.0 | 3593.0 | 4524.0 |
| 1983 | 1609.0 | 1638.0 | 2030.0 | 1375.0 | 1320.0 | 1245.0 | 1600.0 | 2298.0 | 2191.0 | 2511.0 | 3440.0 | 4923.0 |
| 1984 | 1609.0 | 1435.0 | 2061.0 | 1789.0 | 1567.0 | 1404.0 | 1597.0 | 3159.0 | 1759.0 | 2504.0 | 4273.0 | 5274.0 |
| 1985 | 1771.0 | 1682.0 | 1846.0 | 1589.0 | 1896.0 | 1379.0 | 1645.0 | 2512.0 | 1771.0 | 3727.0 | 4388.0 | 5434.0 |
| 1986 | 1606.0 | 1523.0 | 1577.0 | 1605.0 | 1765.0 | 1403.0 | 2584.0 | 3318.0 | 1562.0 | 2349.0 | 3987.0 | 5891.0 |
| 1987 | 1389.0 | 1442.0 | 1548.0 | 1935.0 | 1518.0 | 1250.0 | 1847.0 | 1930.0 | 2638.0 | 3114.0 | 4405.0 | 7242.0 |
| 1988 | 1853.0 | 1779.0 | 2108.0 | 2336.0 | 1728.0 | 1661.0 | 2230.0 | 1645.0 | 2421.0 | 3740.0 | 4988.0 | 6757.0 |
| 1989 | 1757.0 | 1394.0 | 1982.0 | 1650.0 | 1654.0 | 1406.0 | 1971.0 | 1968.0 | 2608.0 | 3845.0 | 4514.0 | 6694.0 |
| 1990 | 1720.0 | 1321.0 | 1859.0 | 1628.0 | 1615.0 | 1457.0 | 1899.0 | 1605.0 | 2424.0 | 3116.0 | 4286.0 | 6047.0 |
| 1991 | 1902.0 | 2049.0 | 1874.0 | 1279.0 | 1432.0 | 1540.0 | 2214.0 | 1857.0 | 2408.0 | 3252.0 | 3627.0 | 6153.0 |
| 1992 | 1577.0 | 1667.0 | 1993.0 | 1997.0 | 1783.0 | 1625.0 | 2076.0 | 1773.0 | 2377.0 | 3088.0 | 4096.0 | 6119.0 |
| 1993 | 1494.0 | 1564.0 | 1898.0 | 2121.0 | 1831.0 | 1515.0 | 2048.0 | 2795.0 | 1749.0 | 3339.0 | 4227.0 | 6410.0 |
| 1994 | 1197.0 | 1968.0 | 1720.0 | 1725.0 | 1674.0 | 1693.0 | 2031.0 | 1495.0 | 2968.0 | 3385.0 | 3729.0 | 5999.0 |
| 1995 | 1070.0 | 1402.0 | 1897.0 | 1862.0 | 1670.0 | 1688.0 | 2031.0 | NaN | NaN | NaN | NaN | NaN |

**Table1.6 Year-on-MonthlvSale**

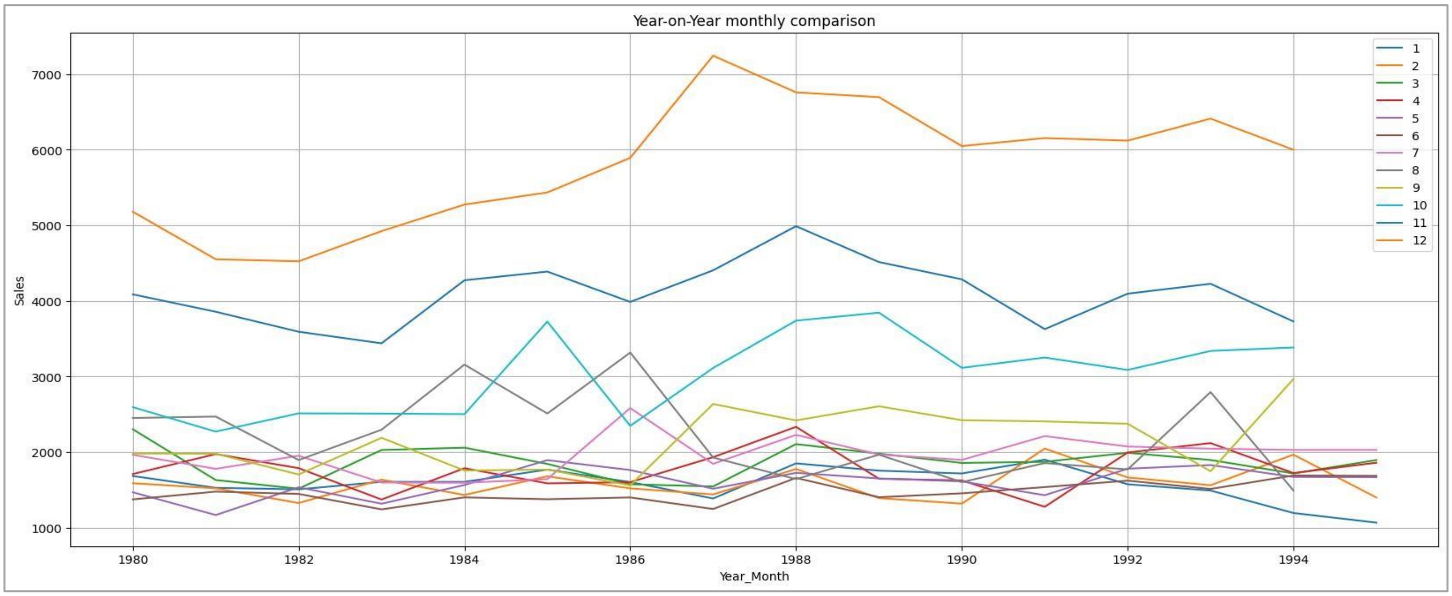



**Fig1.6 Year-on- Monthly Comparison**

- The above plots show us the behavior of the Sparkling sales across various months.
- The sales are highest in December. The sales appear to drop in the month of January and are stable till July, with seasonal patterns across the years.
- There is significant variability in sales from month to month
- There are a few outliers in the data, which could be due to special promotions, holidays or other occasions

➢ **Empirical Cumulative Distribution Plot**

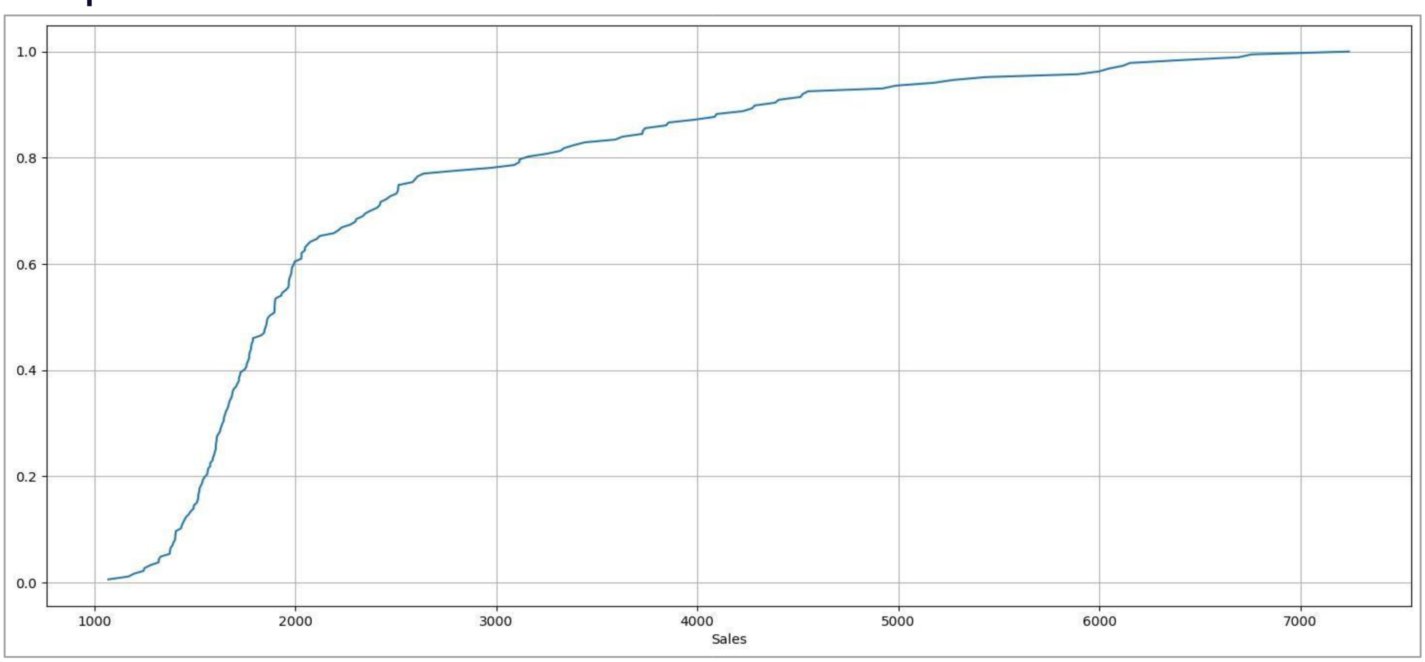



**Fig1.7 Empirical Cumulative Distribution Plot**

- This graph tells us what percentage of data points refer to what number of Sales.
- The distribution of sales is skewed to the right. This means that there are more sales at the lower end of the distribution than at the higher end.

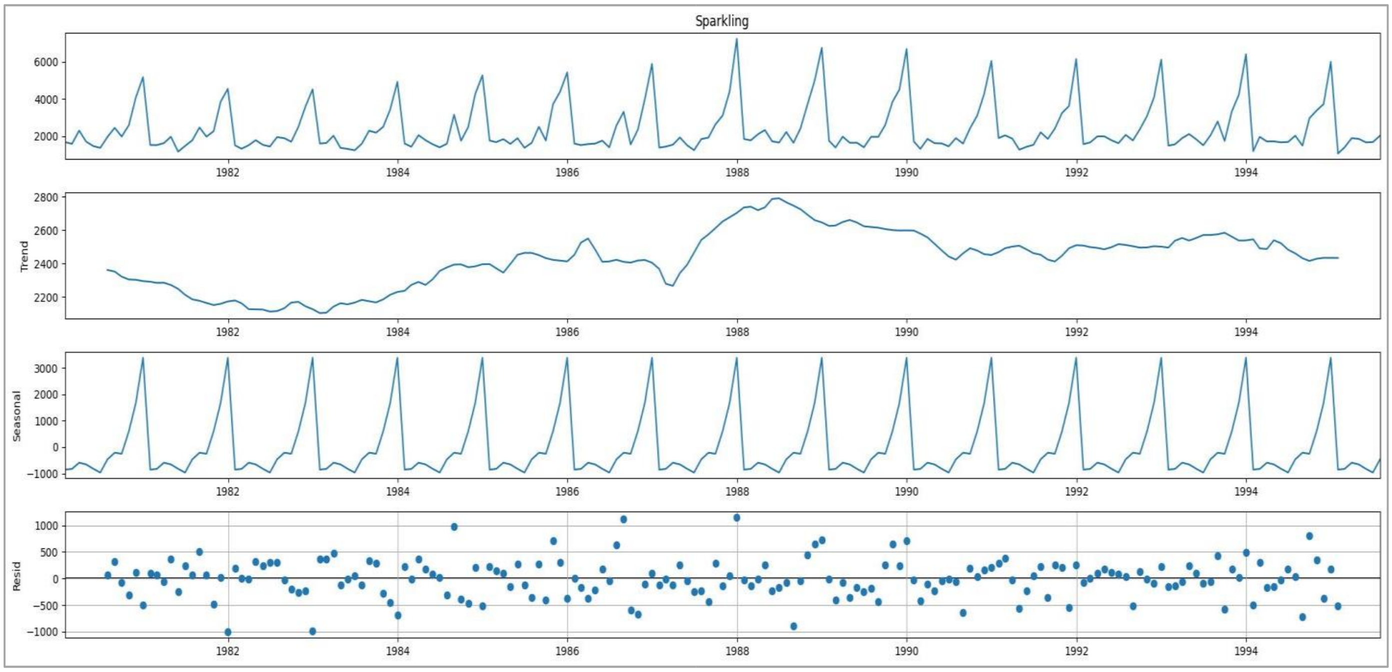
➢ **Average Sparkling Sale**


**Fig1.8 Year-on-YearAverage Sale**

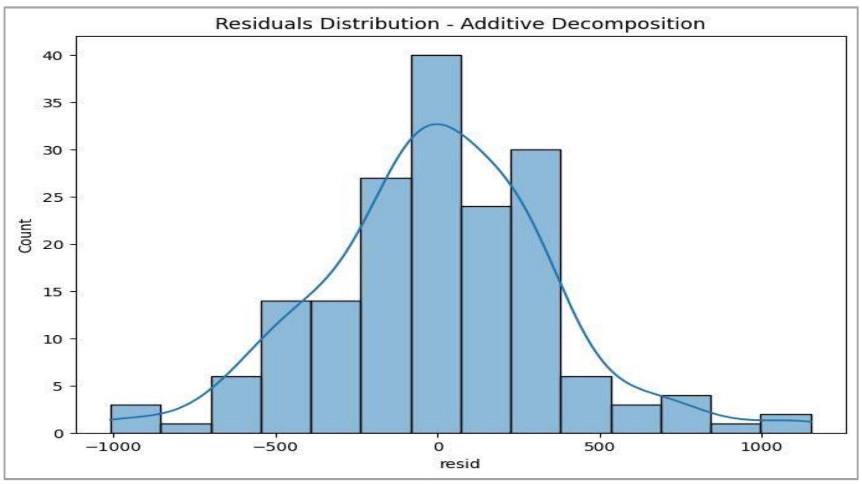
- The average sales are increasing year-on-year. This is evident from the fact that the line graph is generally increasing.
- The year-on-year percentage change in sales is positive for most years, but there are some years with negative changes increasing, but there are some periods where it decreases.
- here is a seasonal pattern in sales, with sales being highest in December.

## Decomposition:

➢ **Additive Decomposition:**




**Fi 1.9 DecomposedTime Series Additive**




**Fi 1.10 Residual Histogram Additive Decomposition**

- **Test for Normality**

➢ We will use the Shapiro Wilk Test for Normality. Let's define the Null & alternate hypothesis: -

➢ Ho: The residuals are normally distributed

➢ Ha: The residuals are not normally distributed

➢ p–value of the Shapiro-Wilk Test on the residuals = 0.03

➢ Since the p-value < 0.05 - We Reject the null hypothesis.

➢ Hence Residuals are not normally distributed at 95% confidence level. The time series is not an additive time series.

➢ **Multiplicative Decomposition:**

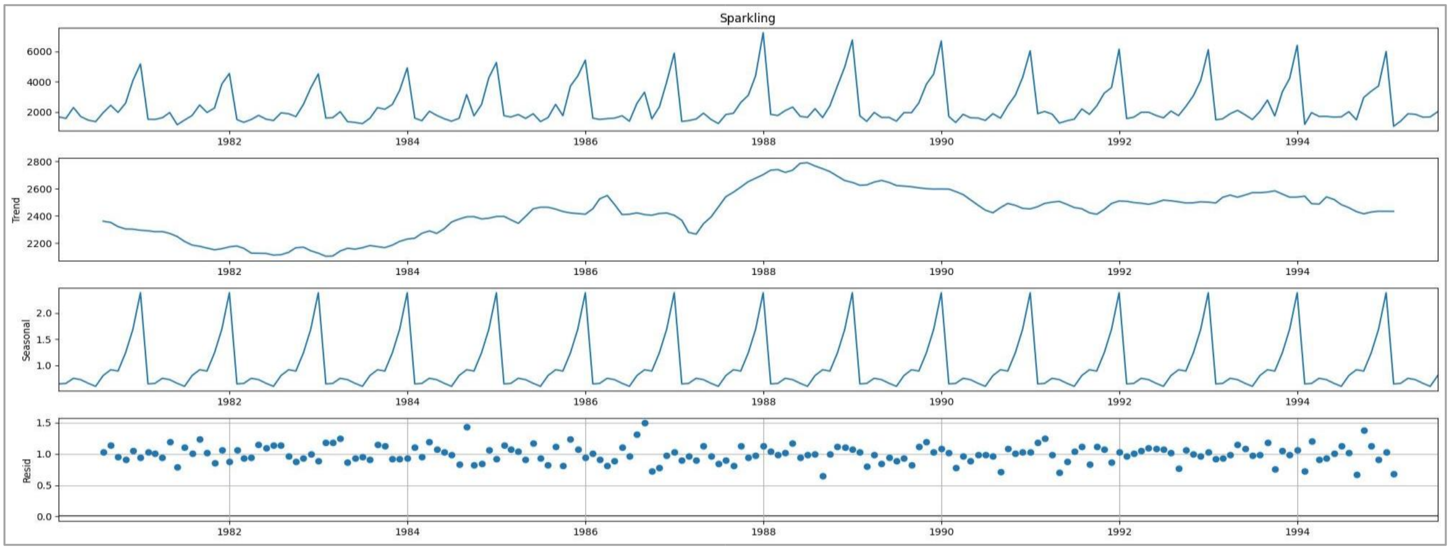



**Fig1.11 Decomposed Time Series Multiplicative**

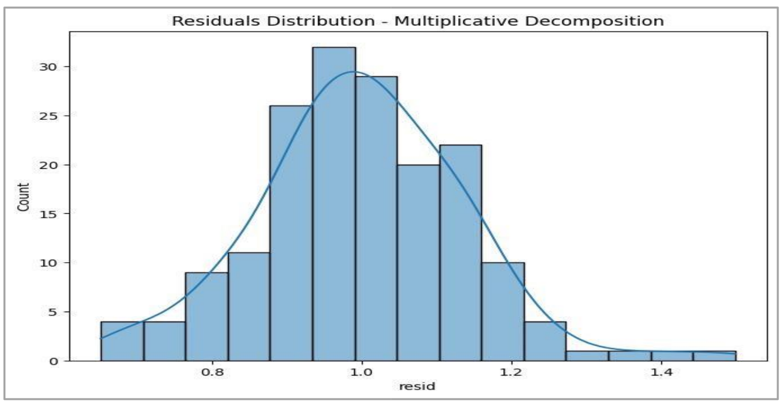



**Fig1.12 Residuals Histogram Multiplicative Decomposition**

- ## Test for Normality

We will use the Shapiro Wilk Test for Normality. Let's define the Null & alternate hypothesis: -

Ho: The residuals are normally distributed

Ha: The residuals are not normally distributed
p–value of the Shapiro-Wilk Test on the residuals = 0.08

Since the p-value > 0.05: we fail to reject the Null hypothesis

Residuals are normally distributed at 95% confidence level. The time series is a multiplicative time series.

➢ **Time series components for Multiplicative:**

| Trend | |
|---|---|
| Year_Month | |
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 2360.666667 |
| 1980-08-31 | 2351.333333 |
| 1980-09-30 | 2320.541667 |
| 1980-10-31 | 2303.583333 |
| 1980-11-30 | 2302.041667 |
| 1980-12-31 | 2293.791667 |
| Name: trend, dtype: float64 | |

| Seasonality | |
|---|---|
| Year_Month | |
| 1980-01-31 | 0.649843 |
| 1980-02-29 | 0.659214 |
| 1980-03-31 | 0.757440 |
| 1980-04-30 | 0.730351 |
| 1980-05-31 | 0.660609 |
| 1980-06-30 | 0.603468 |
| 1980-07-31 | 0.809164 |
| 1980-08-31 | 0.918822 |
| 1980-09-30 | 0.894367 |
| 1980-10-31 | 1.241789 |
| 1980-11-30 | 1.690158 |
| 1980-12-31 | 2.384776 |
| Name: seasonal, dtype: float64 | |

| Residual | |
|---|---|
| Year_Month | |
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 1.029230 |
| 1980-08-31 | 1.135407 |
| 1980-09-30 | 0.955954 |
| 1980-10-31 | 0.907513 |
| 1980-11-30 | 1.050423 |
| 1980-12-31 | 0.946770 |
| Name: resid, dtype: float64 | |

**Table1.7 Decomposed Time Series Components**

Based on the decomposed data provided for both additive and multiplicative decomposition, we can see that the

**Time Series is Multiplicative**

- The Trend appears to be increasing over time with some fluctuations, indicating a positive growth pattern.
- the seasonality shows both positive and negative values, suggesting regular cycles or seasonal effects.
- the residual component includes random fluctuations and unexplained variance in the time series.

## 1.3 Split the data into training and test. The test data should start in 1991.

- the data was split into a train and test set.
- the splitting was done chronologically, with data from the year 1991 forming the test set.
- the train set contains 132 records.
- the test set contains 55 records.

```
Dimentions of Original Dataset: (187, 1)
Dimentions of Training data: (132, 1)
Dimentions of Training data: (55, 1)
```

**Table 1.8 Dimensions of Original, Train & Test Data**

➢ **Training data sample**

First few rows of Train

| | Sparkling |
|---|---|
| Year_Month | |
| 1980-01-31 | 1686 |
| 1980-02-29 | 1591 |
| 1980-03-31 | 2304 |
| 1980-04-30 | 1712 |
| 1980-05-31 | 1471 |

Last few rows of Train

| | Sparkling |
|---|---|
| Year_Month | |
| 1990-08-31 | 1605 |
| 1990-09-30 | 2424 |
| 1990-10-31 | 3116 |
| 1990-11-30 | 4286 |
| 1990-12-31 | 6047 |

**Table1.9 Sample of Training Data**

- **Test data sample**

First few rows of Test

| Year_Month | Sparkling |
|---|---|
| 1991-01-31 | 1902 |
| 1991-02-28 | 2049 |
| 1991-03-31 | 1874 |
| 1991-04-30 | 1279 |
| 1991-05-31 | 1432 |

Last few rows of Test I

| Year_Month | Sparkling |
|---|---|
| 1995-03-31 | 1897 |
| 1995-04-30 | 1862 |
| 1995-05-31 | 1670 |
| 1995-06-30 | 1688 |
| 1995-07-31 | 2031 |

**Table 1.10 Sample of Test Data**

- **Train Test Split Plot**

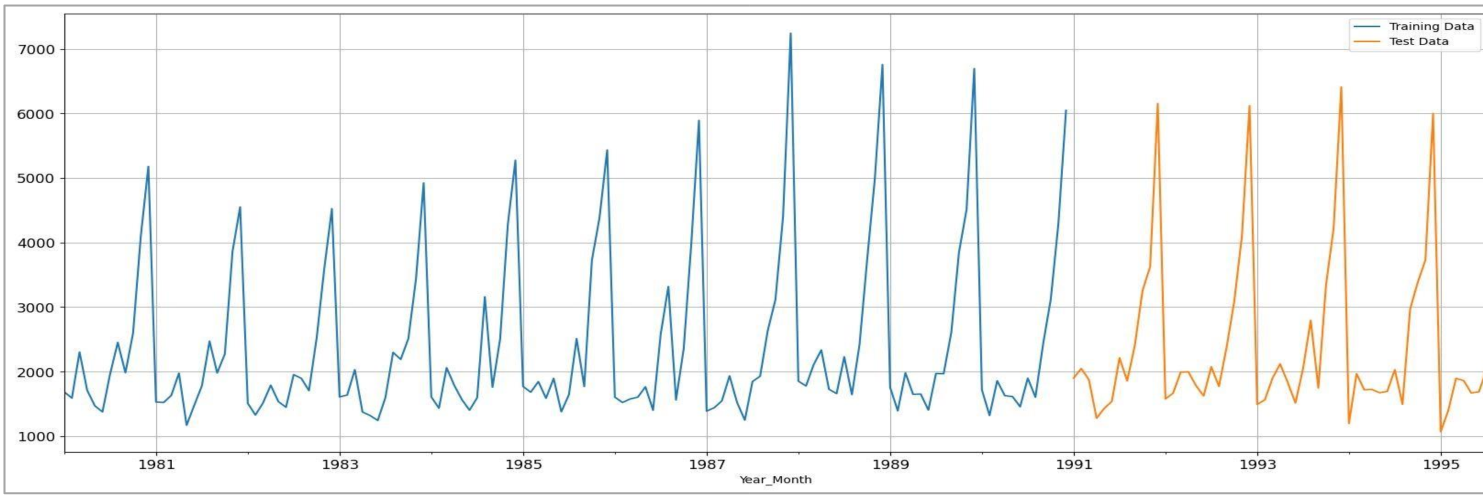



**Fig1.13 Train & Test Split Time Series**

## 1.4 Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

## Building different models and comparing the accuracy metrics.

- **Linear Regression Model**

For this linear regression, we are going to regress the 'Sparkling' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression. We generated the numerical time instance order for both the training and test set. Sample of the Train & Test data.

| YearMonth | Sparkling | time | RegOnTime |
|---|---|---|---|
| 1980-01-01 | 1686 | 1 | 2021.741171 |
| 1980-02-01 | 1591 | 2 | 2027.573830 |
| 1980-03-01 | 2304 | 3 | 2033.406488 |
| 1980-04-01 | 1712 | 4 | 2039.239147 |
| 1980-05-01 | 1471 | 5 | 2045.071805 |

| YearMonth | Sparkling | time | RegOnTime |
|---|---|---|---|
| 1991-01-01 | 1902 | 133 | 2791.652093 |
| 1991-02-01 | 2049 | 134 | 2797.484752 |
| 1991-03-01 | 1874 | 135 | 2803.317410 |
| 1991-04-01 | 1279 | 136 | 2809.150069 |
| 1991-05-01 | 1432 | 137 | 2814.982727 |

Table 1.11 Sample data of Linear Regression Train & Test Data

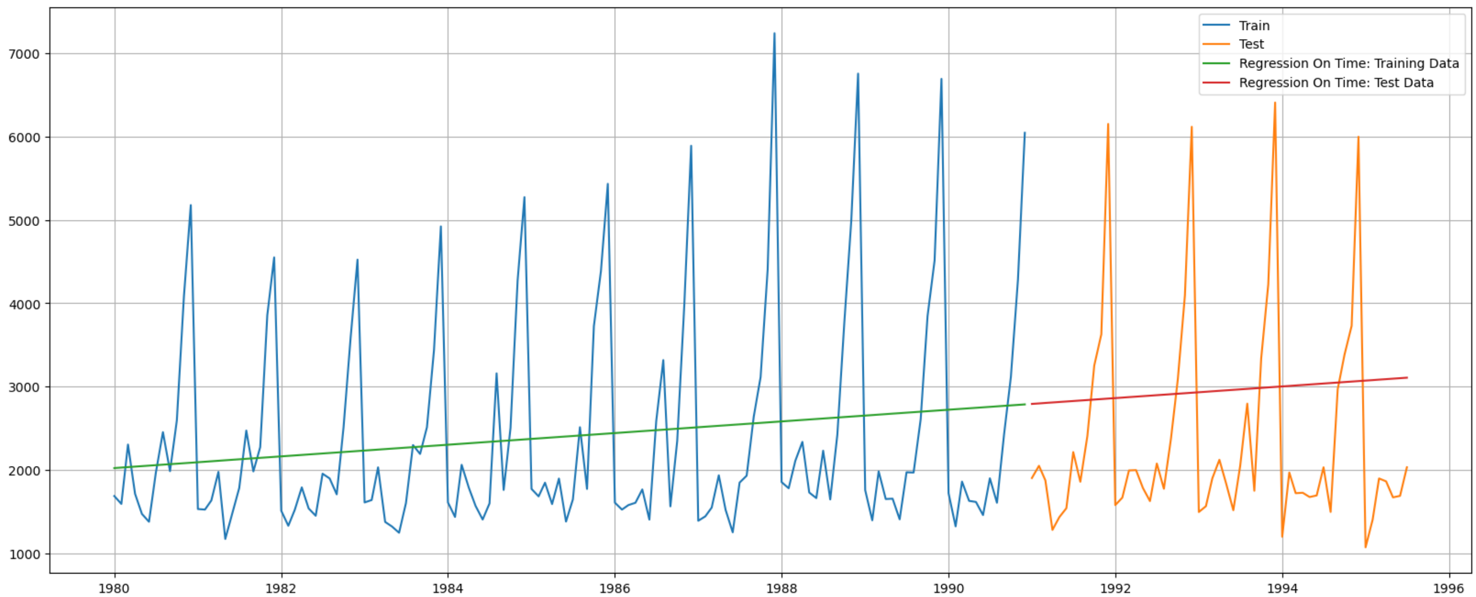



Fig 1.14 Time Series plot of Linear Regression

| Model Performance | | | |
|---|---|---|---|
| **Model** | **Train RMSE** | **Test RMSE** | **Test MAPE** |
| **Linear Regression** | **1279.32** | **1389.14** | **50.15** |

Table 1.12 Model Performance Summary – Linear Regression

- Linear regression captures the trend but not the seasonality.
- Test RMSE is 1389.14, MAPE is 59.35 for Linear Regression, indicating difficulty in handling seasonality.

**Model 2: Naïve Forecast Model**

Model Output Visualized:

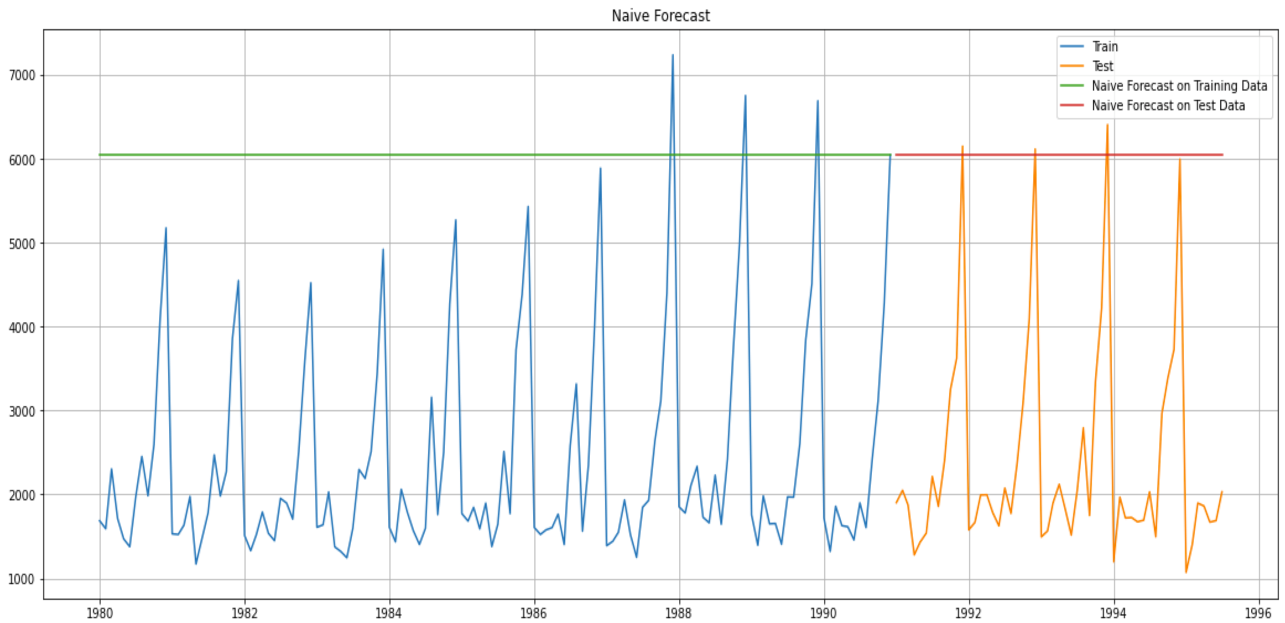



Fig 1.15: Time Series Plot for Naïve Model

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **3867.701** | **153.17** |
| **Test Data** | **3864.279** | **152.87** |

Observation:

The Naive model is dependent on the last observed value, which in our training data is the month of December. We know that the month of December records peak sales every year. So this value is clearly not representative of the dataset at large. Hence the expected very high RMSE scores.

**Model 3: Simple Average:**

For the simple average method, we will forecast by using the average of the training values.

Model Output Visualised:

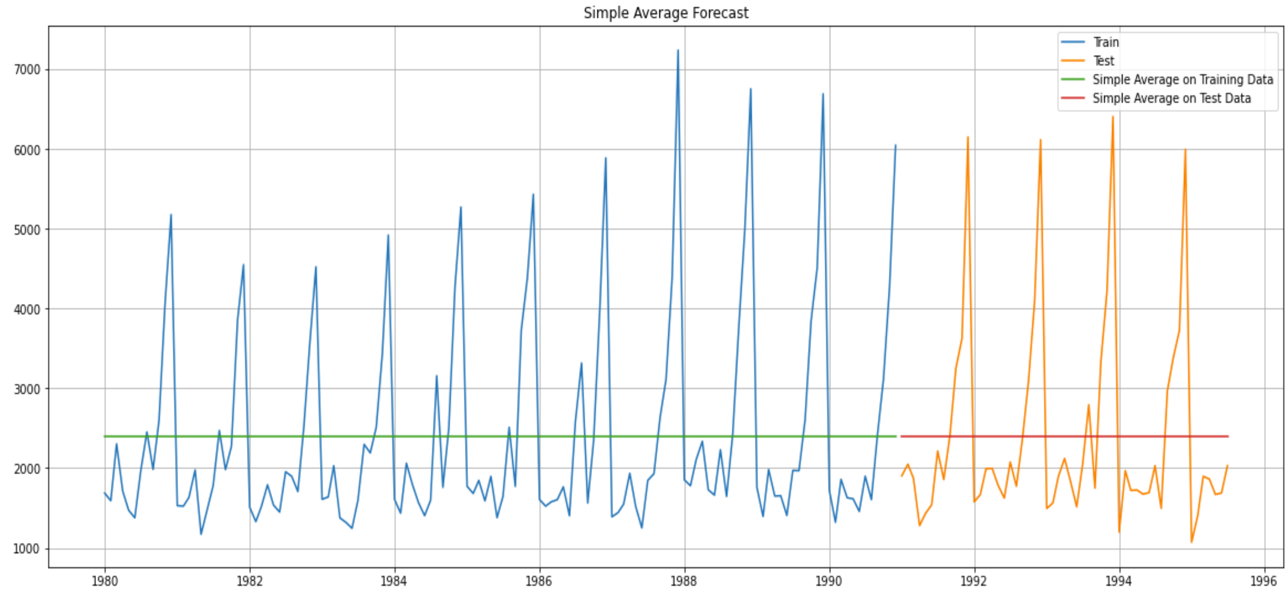



Fig 1.15: Time Series Plot For Simple Average

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **1298.484** | **40.36** |
| **Test Data** | **1275.073** | **38.81** |

Table 1.15: model perfromance

Observation:

A simple average model is not a great fit for the data, since seasonality is predominant in this time series. It therefore misses out on much of the variation, resulting in high RMSE scores.

- Performs better than linear regression in this case as test data has a constant trend different from the training data's underlying trend.
- Simple Average Model forecasts the mean of the training data. It ignores both the trend and seasonality.

**Model 4: Moving Average/s**

Model Output Visualized:

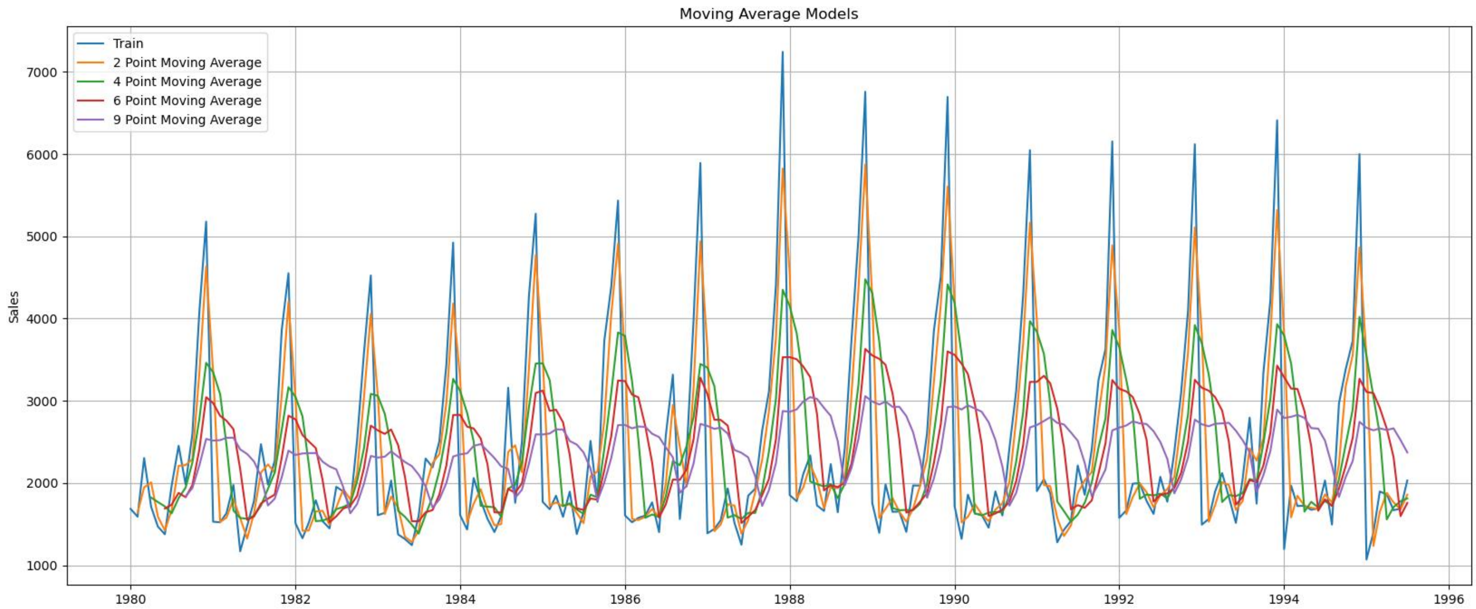



Fig 1.16: Time series plot of moving averages

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **2 point Moving Average** | **813.401** | **19.70** |
| **4 point Moving Average** | **1156.590** | **35.96** |
| **6 point Moving Average** | **1283.927** | **43.86** |
| **9 point Moving Average** | **1346.278** | **46.86** |

Table 1.16: model performance

Observation:

Moving Average Models are able to better track the variation of the time series, especially the 3 point MA in our case. To forecast for a year, we will need to use at least a 12-point MA, which however fails to capture the seasonal variation that is predominant in the data.   Model Comparison

The following is a visual depiction of how the models we've built thus far, perform.

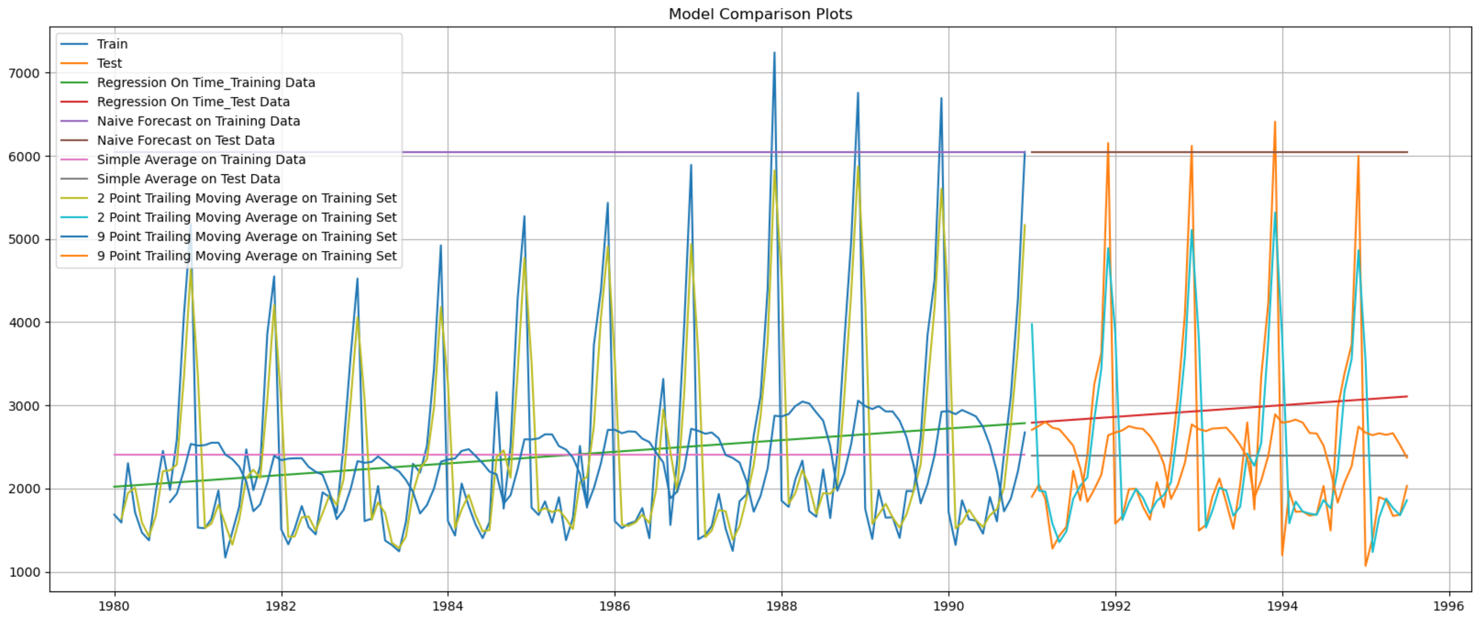



Fig: 1.17: Time Series plot of Moving Average on Whole Data

We will now proceed with the Exponential Smoothing Models.

- For the moving average models, here are the insights based on their RMSE and MAPE values:
- 2-Point Moving Average: RMSE = 813.4, MAPE = 19.70. It performs relatively well in capturing the trend and seasonality but still has room for improvement.
- 4-Point Moving Average: RMSE = 1156.59, MAPE = 35.96. It performs slightly worse than the 2-Point model, but it still shows better accuracy than simpler models like the Simple Average.
- 6-Point Moving Average: RMSE = 1283.82, MAPE = 43.86. It performs reasonably well but is not as accurate as the 2-Point or 4-Point models.
- 9-Point Moving Average: RMSE = 1346.28, MAPE = 46.86. It shows reasonable performance but is not as good as the 2-Point model.
- Overall, the 2-Point Moving Average model stands out as the best-performing model with the lowest RMSE and MAPE values. However, there is still room for improvement in all the models to better capture the trend and seasonality and reduce errors in the forecasts.

**Model 5: Single Exponential Smoothing (Auto-fit, alpha = 0) :**

**This method is suitable for forecasting data with no clear trend or seasonal pattern. It gives more weight to recent observations, which means that recent data points have a stronger influence on the forecast than older ones. This approach allows the model to capture short-term trends and adapt quickly to changes in the data.**

Parameter Alpha ($\alpha$) is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing

SES, Alpha = 0

The autofit model finds the most optimal parameters according to python while fitting on the train data.

Simple Exponential Smoothing optimal parameters:

Smoothing Level (alpha) = 0.07 Initial Level 1763.93.

```
{'smoothing_level': 0.039534883720930 23,
 '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 1.17 Autofill Simple Exponential Smoothing Optimal Parameters**

Model Output Visualized:

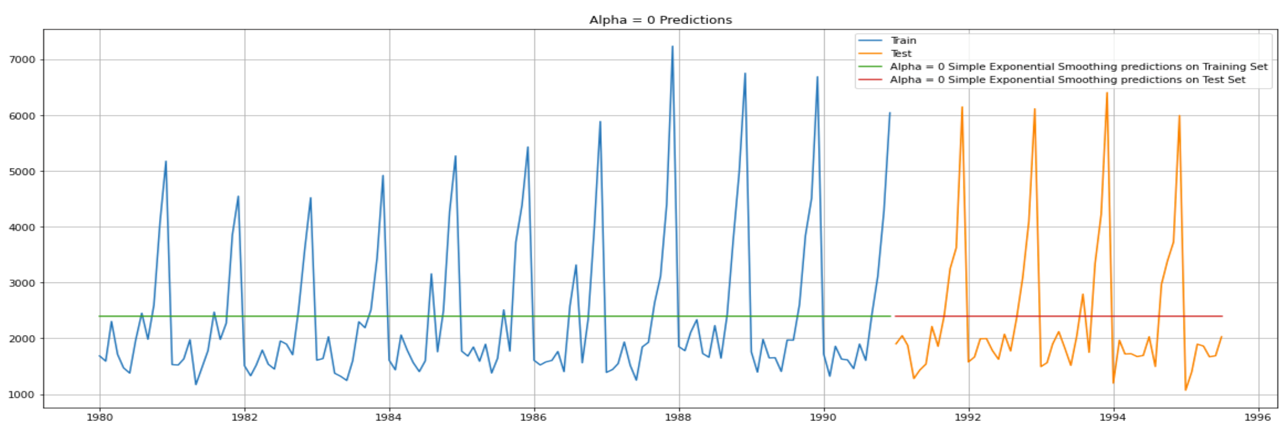



Fig:1.18 Time series plot for SES

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **1317.31** | **39.05** |
| **Test Data** | **1304.927** | **44.48** |

Table 1.17 Model performance for SES

Observation:

While this particular time series has no overall upward or downward trend, the level extrapolated by the SES model will be a useful component. However, since this time series has a predominant seasonal characteristic, it will not be able to capture the variation in the data.

- Simple exponential smoothing model provides one-step-ahead forecast. It ignores both the trend and seasonality in the data.
- RMSE is 1304.927 and MAPE is 44.48, indicating poor performance in capturing underlying patterns
- The low smoothing parameter (0) implies a heavy reliance on past averages. This makes it less accurate compared to more sophisticated methods.

**<u>Model 5a: Single Exponential Smoothing (using a Range of alpha values)</u>**

SES, Alpha ranging from 0.1 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0, 0.1 and 0.2)

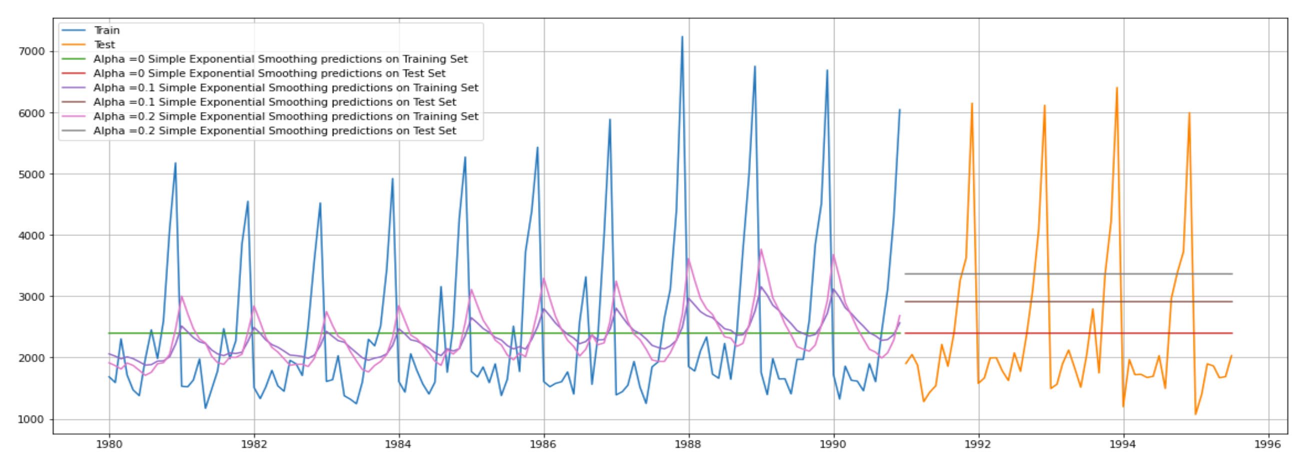



Fig:1.19 time Series plot on SES Alpha =0.01 – 0.2

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **Alpha = 0.1** | **1375.394** | **49.53** |
| **Alpha = 0.2** | **1595.207** | **60.46** |

Table: 1.19 Model Performance summary

Observation:

While this particular time series has no overall upward or downward trend, the level extrapolated by the SES model will be a useful component. However, since this time series has a predominant seasonal characteristic, it will not be able to capture the variation in the data.

- The Brute Force simple exponential smoothing model generates one-step-ahead forecasts, overlooking both trend and seasonality in the time series.
- The model exhibits an RMSE of 1278.50 and MAPE of 42.41, signifying its inability to effectively capture the underlying trend and seasonal patterns.
- The small smoothing parameter (0.02) indicates that the model relies heavily on past data averages rather than recent observations, making its accuracy comparable to the simple average model.

**Model 6: Double Exponential Smoothing (Auto-fit, alpha = 0.65, beta = 0)**

DES (Holt's Model), Alpha = 0.65, Beta = 0

- **DES: Auto Fill Method**

The autofit model finds the most optimal parameters according to python while fitting on the train data.

Double Exponential Smoothing optimal parameters: -

Smoothing Level (Alpha) = 0.65

Smoothing Trend (beta) = 0


```
{'smoothing_level': 0.6649999999999999,
 'smoothing_trend': 0.0001,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1502.1999999999991,
 'initial_trend': 74.87272727272739,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```


Table 1.20 parameters for DES

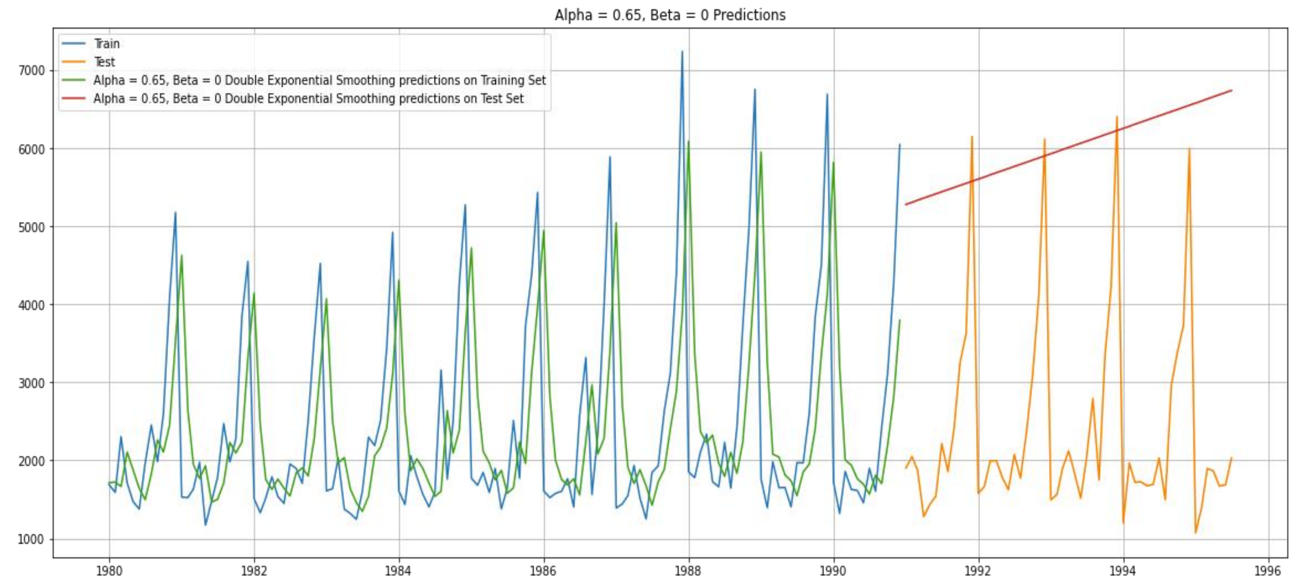



Fig:1.20 Time Series plot for DES

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **1339.501** | **38.82** |
| **Test Data** | **5291.880** | **208.74** |

Table:1.21 Model performance for DES

Observation:

The Holt’s model isn’t a very good fit for this time series as the data has no trend, and strong seasonality, as evidenced by the high RMSE.

**Model 6a: Double Exponential Smoothing (using a Range of alpha, beta values)**

DES, Alpha ranging from 0.1 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and beta = 0.1)

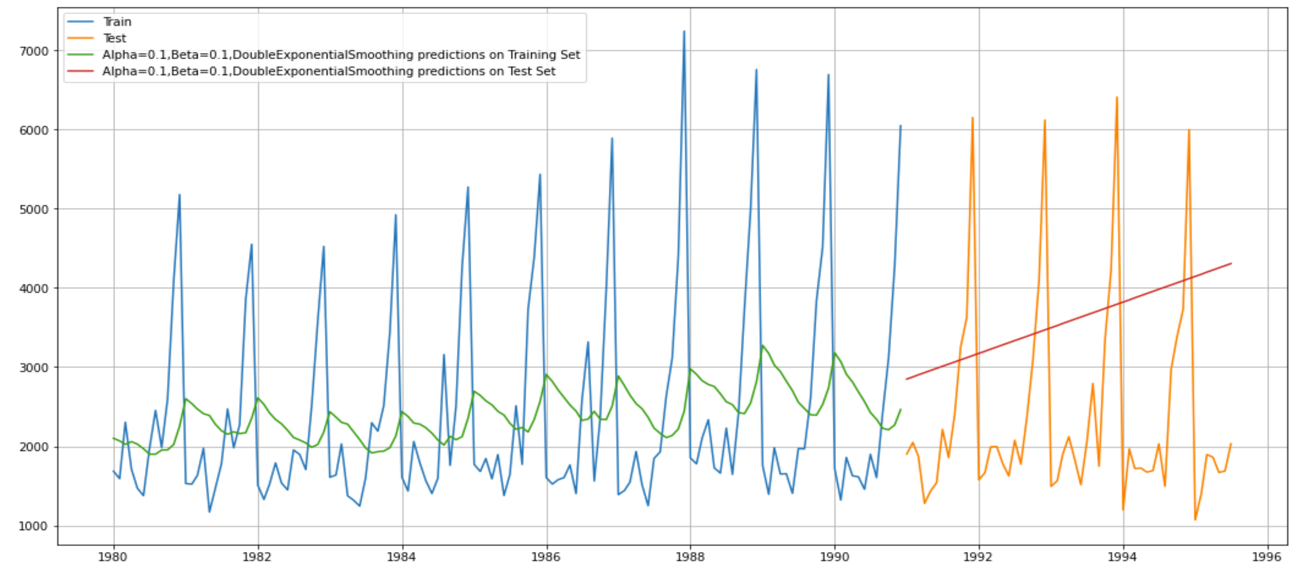



Fig:1.21 Time series plot for DES Alpha & Beta 0.1

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **1363.474** | **44.26** |
| **Test Data** | **1779.425** | **67.23** |

Table: 1.22 Model performance for DES

Observation:

The Holt's model isn't a very good fit for this time series as the data has no trend, and strong seasonality, as evidenced by the high RMSE.

- The Brute Force double exponential smoothing model effectively captures the trend but overlooks the seasonality of the time series.
- This performs better than auto fill method (Double Exponential Smoothing alpha= 0.65, Beta = 0.1)
- The model exhibits a high RMSE of 1363.474 and MAPE of 44.26, indicating its failure to capture the seasonality in the data.
- The level smoothing parameter (alpha) is close to 0 (0.1), suggesting that the level forecast is primarily based on the past data rather than recent observations.
- The trend smoothing parameter (beta) is 0.1, which lies between 0 and 1, indicating a balanced approach that considers both historical data and recent observations to forecast the trend.

## Model 7: Triple Exponential Smoothing (Auto-fit: Alpha=0.15, Beta=0, Gamma=0.37)

**The triple exponential smoothing model is suitable for time series with both trend and seasonality.**

**Since the trend variation is linear and the seasonal decomposition suggests a multiplicative time series, we use a triple exponential smoothing model with additive trend and multiplicative seasonality.**

TES (Holt-Winters' Model), Alpha=0.15, Beta=0, Gamma=0.37

- **TES: Auto Fill Method**

The autofit model finds the most optimal parameters according to python while fitting on the train data.

Triple Exponential Smoothing optimal parameters: -

Smoothing Level (Alpha) = 0.111

Smoothing Trend (Beta) = 0.049

Smoothing Seasonal (Gamma) = 0.362

```
{'smoothing_level': 0.11101471561088701,
 'smoothing_trend': 0.0493145907614654,
 'smoothing_seasonal': 0.36244934537370843,
 'damping_trend': nan,
 'initial_level': 2356.496908624238,
 'initial_trend': -9.809526161838415,
 'initial_seasons': array([0.713711  , 0.68278724, 0.90458411, 0.8053878 , 0.65571739,
       0.65388935, 0.88616088, 1.13350811, 0.91894498, 1.21186447,
       1.87099202, 2.37505867]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

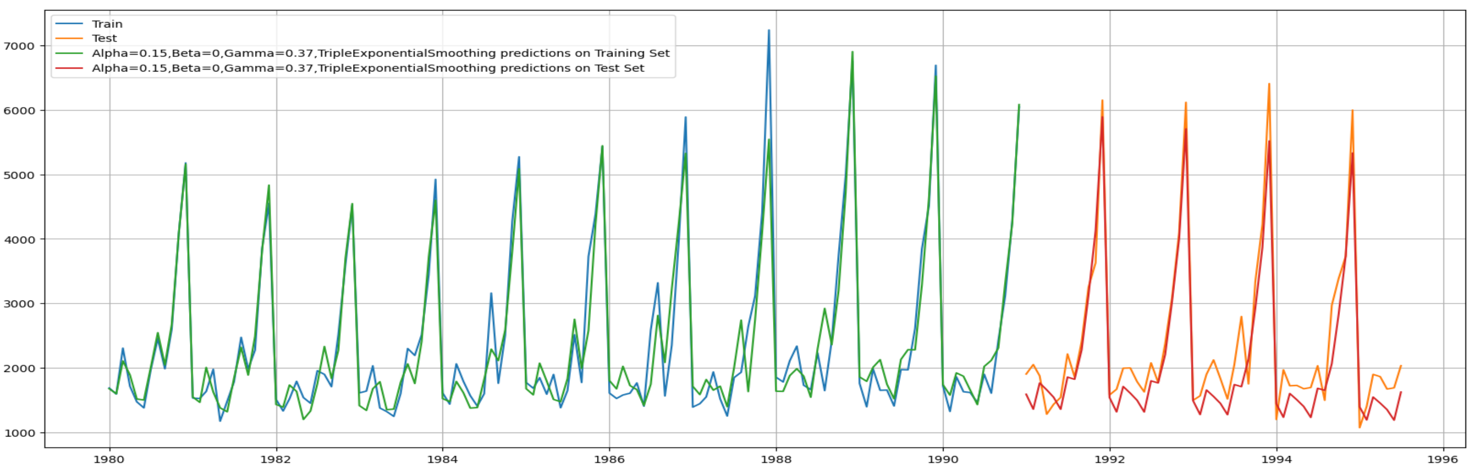



Fig: 1.22 Time series plot for TES Auto fill method

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **355.776** | **10.19** |
| **Test Data** | **402.936** | **13.88** |

Table:1.23 Performance metrics

Observation:

The Triple Exponential Smoothing model is a good fit for this time series, owing to the strong Seasonal component. This is corroborated by a low RMSE.

Exponential Smoothing model captures both the trend and seasonality of the time series.

- The test RMSE is 355.776, and the MAPE is 10.19. This model has the lowest error among all the evaluated models.
- The values of alpha and beta, being close to 0.15, indicate that the forecast relies heavily on historical data to build the model.

On the other hand, the gamma value has a moderate value, indicating that it considers more than just the recent past data but less than the entire historical data

**Model 7a: Triple Exponential Smoothing BRUTE Force (using a Range of alpha, beta, gamma values)**

TES, Alpha, Beta, Gamma ranging from 0 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and beta = 0.0, gamma = 0.3)

The brute force model tests various smoothing parameter values to find the best ones for accurate test data forecasting. Below is the table for various parameters, sorted with least Test RMSE on top.

| | Alpha | Beta | Gamma | Train RMSE | Test RMSE |
|---|---|---|---|---|---|
| 41 | 0.01 | 0.04 | 0.25 | 470.837600 | 302.728640 |
| 40 | 0.01 | 0.04 | 0.22 | 481.966174 | 303.298599 |
| 42 | 0.01 | 0.04 | 0.28 | 461.026879 | 303.899551 |
| 1163 | 0.04 | 0.07 | 0.25 | 366.400702 | 305.639893 |
| 1164 | 0.04 | 0.07 | 0.28 | 365.138125 | 305.990053 |

**Table 1.24 Brute Force Triple Exponential Smoothing Parameters**

Since alpha = 0.01, beta = 0.04, gamma = 0.25 yield the least test RMSE, indicating the best fit for our test data, we select them to build our Triple Exponential Smoothing model.

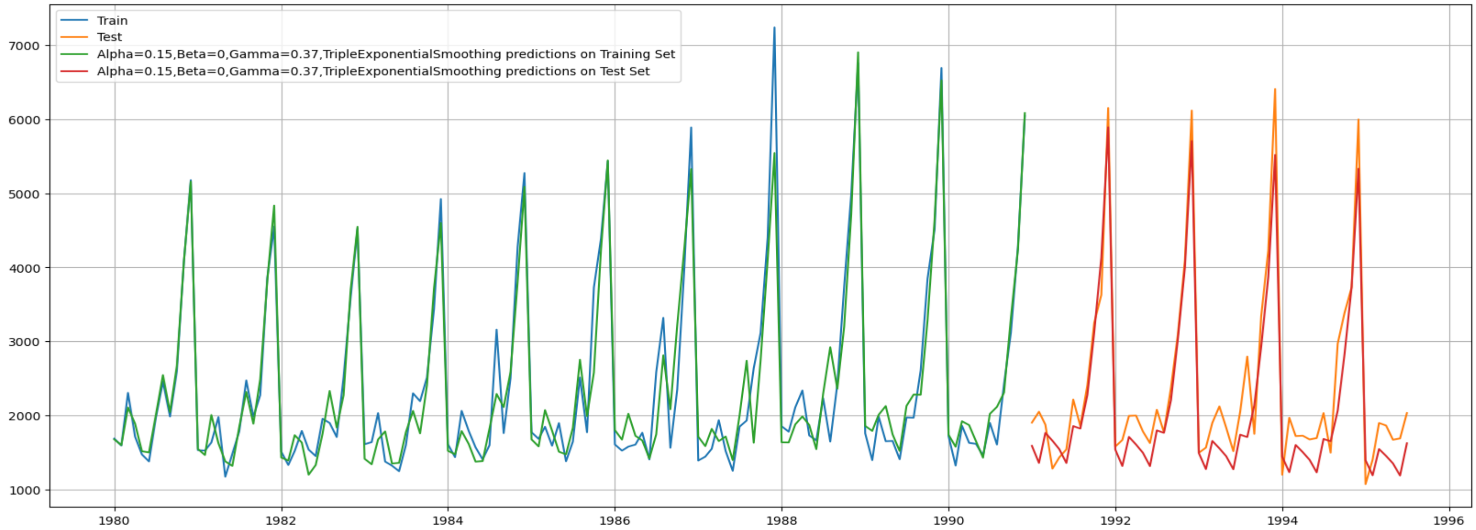



**Fig 1.23 Time Series Plot: Triple Exponential Smoothing Alpha = 0.15, Beta= 0, Gamma= 0.37**

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **370.698** | **9.58** |
| **Test Data** | **314.939** | **10.1** |

Observation:

The Triple Exponential Smoothing model is a good fit for this time series, owing to the strong Seasonal component. This is corroborated by a low RMSE.

- The Brute Force Triple Exponential smoothing model captures both the trend and seasonality of the time series effectively.
- With RMSE of 314.939 and MAPE of 10.1, this model exhibits the best accuracy among all the evaluated models so far.
- The values of alpha and beta being close to 0.15 imply that the model heavily relies on historical data to make forecasts.
- On the other hand, the moderate value of gamma indicates that it strikes a balance between recent past data and the entire historical past to capture seasonality.

A Consolidated Plot of all the Exponential Models built:

- We see that the Triple Exponential Models are the best suited for this time series.

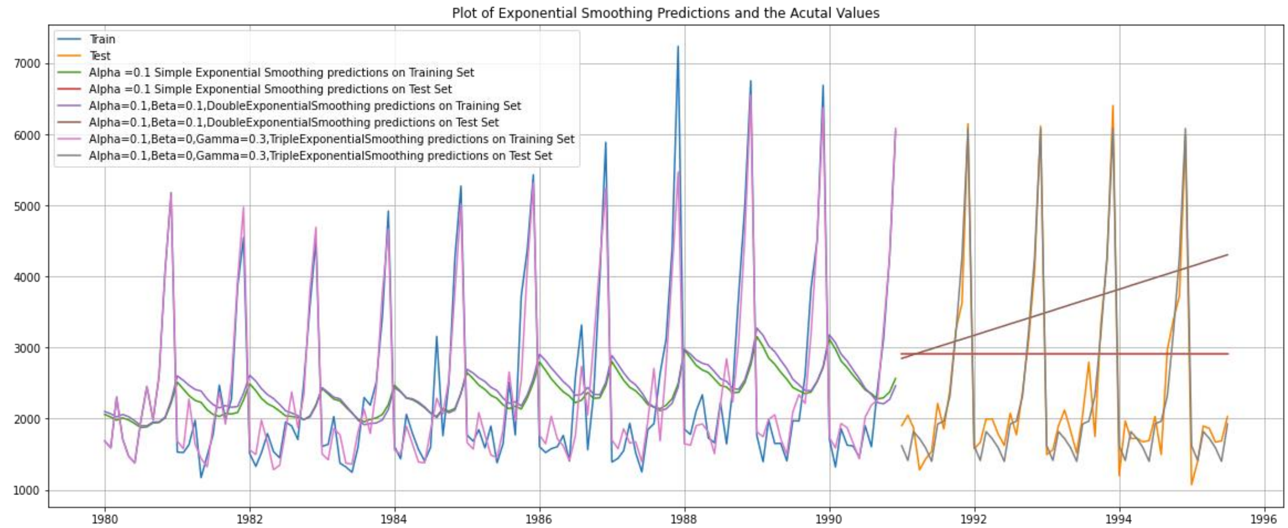



Fig: 1.24 Time series plot for all the DES , TES And SES

- The triple exponential smoothing model captures both the trend and seasonality of the time series effectively.
- With a Root Mean Square Error (RMSE) of 333.273 and a Mean Absolute Percentage Error (MAPE) of 10.16, this model exhibits the best accuracy among all the evaluated models.
- The values of alpha and beta being close to 0.15 imply that the model heavily relies on historical data to make forecasts. On the other hand, the mid-range value of gamma indicates that it strikes a balance between recent past data and the entire historical past to capture seasonality.

**1.5 Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.**

- **Note: Stationarity should be checked at alpha = 0.05.**

The Augmented Dickey Fuller (ADF) Test is a statistical test for affirming whether or not a time series is Stationary.

The Null Hypothesis H0 is: Time Series is non-stationary

The Alternative Hypothesis H1 is: Time Series is Stationary

**TEST 1:** We administer the ADF test on the Original Time Series.

Results of Dickey-Fuller Test: Test Statistic: -
1.208926 p-value: 0.669744

With the resultant ADF test p-value at 0.67, we cannot reject the Null Hypothesis (at alpha 0.05).

<u>We hence conclude that the Time Series is non-stationary</u>.
In order to make a Time Series Stationary, we need to transform the original series by taking a Difference of the original values. Usually, a 1 period difference suffices to transform a non-stationary series into a Stationary one.

**TEST 2**: We administer the ADF test on the new series - derived by taking a 1 period Difference of the original series.

Results of Dickey-Fuller Test: Test Statistic: -8.005007e+00 p-value: 2.280104e-12

The resultant ADF p-value (0.000000000002) is significantly less than 0.05 (alpha). We can hence reject the null hypothesis in the case of the new series, which is derived by differencing the original series over 1 period.

<u>We conclude that at Difference 1, the time series is Stationary.</u>

**6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**

- ### Automated version of an ARIMA model

ARIMA: - Auto Regressive Integrated Moving Average is a way of modeling time series data for forecasting or predicting future data points. Improving AR Models by making Time Series stationary through Moving Average Forecasts

ARIMA models consist of 3 components: -

AR model: The data is modelled based on past observations.

Integrated component: Whether the data needs to be differenced/transformed.

MA model: Previous forecast errors are incorporated into the model.

The best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

- ARIMA Model building to estimate best 'p', 'd', 'q' parameters (Lowest AIC Approach)

**Model 8: ARIMA (Lowest AIC parameters: p=2, d=1, q=2)**

Of the ARIMA models generated using various combinations of parameters p and q, the model with the lowest AIC score was: ARIMA (2, 1, 2) with an AIC score of 2213.617708

Model Summary:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sat, 10 Aug 2024   AIC                           2213.509
Time:                        14:22:22   BIC                           2227.885
Sample:                    01-01-1980   HQIC                          2219.351
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.781      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.740      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.216      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                14.46
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.79e+28. Standard errors may be unstable.
```

Fig: 1.25 Sarima results summary

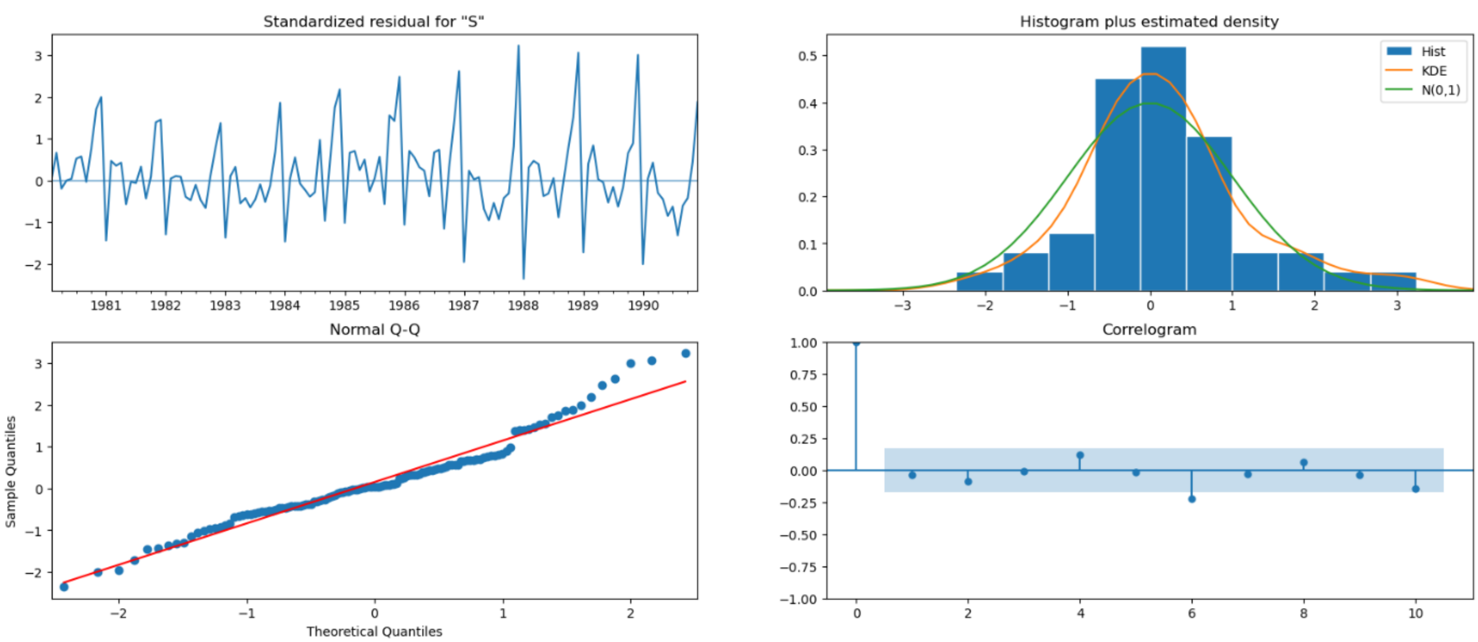



Fig: 1.26 Diagnostic plot of ARIMA (2,1,2)

| Model Performance | | | |
|---|---|---|---|
| Model | Train RMSE | Test RMSE | Test MAPE |
| Automated ARIMA (2, 1, 2) | 2213.509 | 1230.53 | 65.74 |

**Table 1.26 Model Performance Summary – Automated ARIMA (2, 1, 2)**

- **The Auto ARIMA model aims to capture the underlying trend in the data but does not consider the seasonality component.**
- **The model's performance is evaluated with a Root Mean Square Error of 1230.53 and a Mean Absolute Percentage Error of 65.74. The model performed poorer on Test data as compared to Training**

## ➢ Automated version of a SARIMA model

The ARIMA models can be extended/improved to handle seasonal components of a data series.

The seasonal autoregressive moving average model is given by SARIMA (p, d, q) (P, D, Q)F

The above model consists of:

- Autoregressive and moving average components (p, q)
- Seasonal autoregressive and moving average components (P, Q)
- The ordinary and seasonal difference components of order 'd' and 'D'
- Seasonal frequency 'F'

The value for the parameters (p,d,q) and (P, D, Q) can be decided by comparing different values for each and taking the lowest AIC value for the model build. The value for F can be consolidated by ACF plot.

**the values: p = 1, d =1, q = 2, P = 1, D = 0, Q = 2, S = 12**

| | param | seasonal | AIC |
|---|---|---|---|
| **50** | (1, 1, 2) | (1, 0, 2, 12) | 1555.584247 |
| **53** | (1, 1, 2) | (2, 0, 2, 12) | 1555.934563 |
| **26** | (0, 1, 2) | (2, 0, 2, 12) | 1557.121563 |
| **23** | (0, 1, 2) | (1, 0, 2, 12) | 1557.160507 |
| **77** | (2, 1, 2) | (1, 0, 2, 12) | 1557.340404 |

**Table 1.27 SARIMA AIC Parameters without Seasoning**

**Model 9: SARIMA (Lowest AIC parameters: p=2, d=1, q=2, P=0, D=1, Q=2)**

The SARIMA models generated using various combinations of parameters p, q, P, Q and D, the model with the lowest AIC score was: SARIMA (1, 1, 2)x(0, 1, 2, 12) with an AIC score of 1382.347780

Model Summary:

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(0, 1, 2, 12)   Log Likelihood                -685.174
Date:                            Sat, 10 Aug 2024   AIC                           1382.348
Time:                                    14:45:33   BIC                           1397.479
Sample:                                         0   HQIC                          1388.455
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5507      0.287     -1.922      0.055      -1.112       0.011
ma.L1         -0.1612      0.235     -0.687      0.492      -0.621       0.299
ma.L2         -0.7218      0.175     -4.132      0.000      -1.064      -0.379
ma.S.L12      -0.4062      0.092     -4.401      0.000      -0.587      -0.225
ma.S.L24      -0.0274      0.138     -0.198      0.843      -0.298       0.243
sigma2      1.705e+05   2.45e+04      6.956      0.000    1.22e+05    2.19e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                13.48
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               0.89   Skew:                             0.60
Prob(H) (two-sided):                  0.75   Kurtosis:                         4.44
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Fig: 1.27 Automated SARIMA without differencing model summary

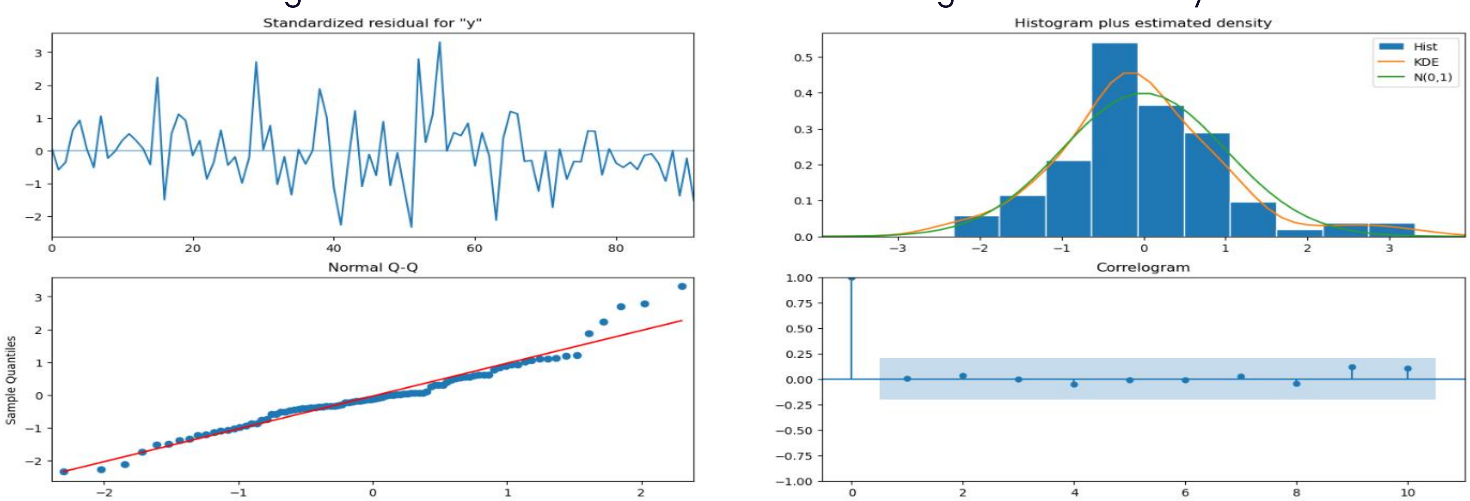



Fig: 1.28 Diagnostic plot of Automatred SARIMA

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **SARIMA (1,1,2) (0,1,2,12)** | **382.577** | **12.87** |

Table: 1.28 SARIMA mpodel

Observation:

A SARIMA model is better equipped for a time series with a very strong seasonal component. And this model matches up in performance to the TES model which is thus far the strongest model for this time series.

- The SARIMA model successfully captures both the trend and seasonality in the data.
- The Root Mean Square Error is 382.56, and the Mean Absolute Percentage Error is 12.87 for the automated SARIMA model with seasonal differencing.
- This model performs better than the model without seasonal differencing, indicating that incorporating seasonal differencing improves the accuracy of the forecast.

**7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**

From the ADF test conducted in an earlier section, we know that the original time series is non-stationary.

So the first step would be to transform the time series by differencing the original series over 1 period, and make it Stationary.

We then need to plot the ACF and PACF on the transformed stationary Time Series.

The following are the ACF and PACF plots:

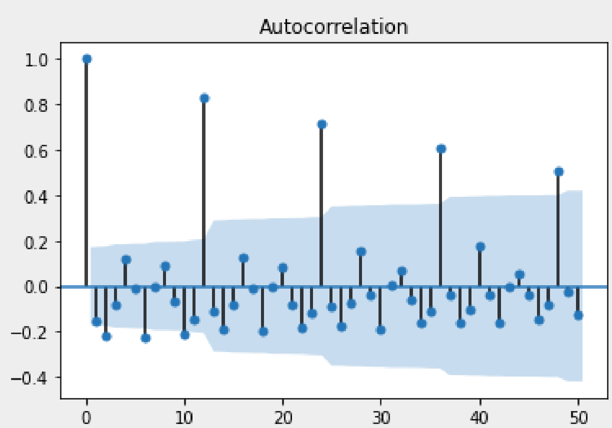
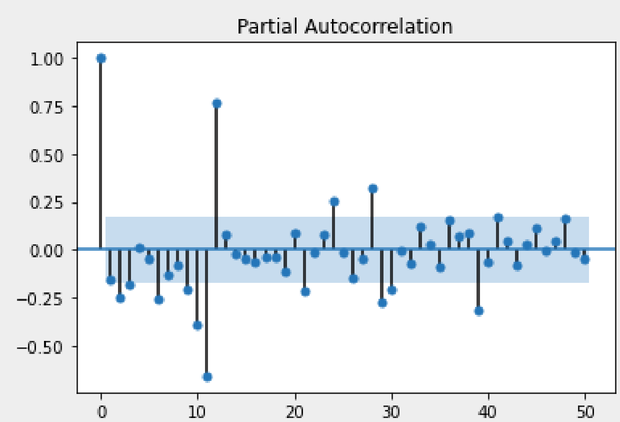
| **ACF Plot** | **PACF Plot** |
|---|---|
|  |  |

Fig: 1.29

The ACF plot:

- The cut-off appears right after lag 0: Lag 1 appears to be insignificant, though only marginally.
- The first significant point is at lag 2, after which the subsequent lags are again insignificant.
- Hence based on the ACF plot, we can assign 2 as the order of the MA component (q) of the ARIMA/SARIMA model.

The PACF plot:

- The cut-off here too appears right after lag 0: Lag 1 appears to be insignificant, though only marginally.
- The first significant point is at Lag 2.
- Lag 3 too appears to be significant, although it is right on the border.
- Hence based on the PACF plot, we can assign 3 as the order of the AR component (p) of the ARIMA/SARIMA model.
- We know that the Seasonal component is very apparent in the series. The ACF plot also clearly shows a pattern repeating every year - indicating strong seasonal behaviour.
- So a SARIMA model would be appropriate for modelling such a series. For employing the seasonal component, we will accord value to a 12-period lag, which will map to a value of 1 to both P and Q.
- The value of d will be 1, as the series has been Differenced by 1 period, in order to make it stationary.
- **Hence based on the ACF and PACF plots, we can develop a SARIMA (3,1,2)x(1,1,1) model.**

- Model 9a: SARIMA (ACF, PACF plot parameters: p=3, d=1, q=2, P=1, D=1, Q=1)

- **Model Summary:**

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                 132
Model:             SARIMAX(1, 1, 2)x(0, 1, 2, 12)   Log Likelihood                -685.174
Date:                             Wed, 12 Aug 2020   AIC                           1382.348
Time:                                     15:02:14   BIC                           1397.479
Sample:                                          0   HQIC                          1388.455
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5507      0.287     -1.922      0.055      -1.112       0.011
ma.L1         -0.1612      0.235     -0.687      0.492      -0.621       0.299
ma.L2         -0.7218      0.175     -4.132      0.000      -1.064      -0.379
ma.S.L12      -0.4062      0.092     -4.401      0.000      -0.587      -0.225
ma.S.L24      -0.0274      0.138     -0.198      0.843      -0.298       0.243
sigma2      1.705e+05   2.45e+04      6.956      0.000    1.22e+05    2.19e+05
===================================================================================
Ljung-Box (Q):                       20.51   Jarque-Bera (JB):                13.48
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.89   Skew:                             0.60
Prob(H) (two-sided):                  0.75   Kurtosis:                         4.44
===================================================================================
```

**Fig: 1.30 Model Summary**

**Performance Metrics on Test Data:**

| | RMSE | MAPE |
|---|---|---|
| SARIMA (1,1,2) (0,1,2,12) | **393.01** | **13.27** |

**Table : 1.29 Performance metrics**

**Observations:**

As seen in the previous instance, A SARIMA model is better equipped for a time series with very strong seasonal component. And this model, too, comes close to matching up in performance to the TES model.

## 1.8 Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

All models, ranked according to their performance. The lowest Test RMSE score is the highest ranked or the best performing model.

| | Test RMSE | Test MAPE |
|---|---|---|
| **Alpha=0.1,Beta=0,Gamma=0.3,TripleExponentialSmoothing** | 333.275019 | 10.16 |
| **SARIMA(1,1,2)(0,1,2,12)** | 382.576759 | 12.87 |
| **SARIMA(3,1,2)(1,1,1,12)** | 393.115110 | 13.27 |
| **Alpha=0.15,Beta=0,Gamma=0.37,TripleExponentialSmoothing** | 402.936179 | 13.88 |
| **2pointTrailingMovingAverage** | 813.400684 | 19.70 |
| **4pointTrailingMovingAverage** | 1156.589694 | 35.96 |
| **ARIMA(2,1,2)** | 1230.532934 | 65.74 |
| **SimpleAverageModel** | 1275.073380 | 38.81 |
| **6pointTrailingMovingAverage** | 1283.927428 | 43.86 |
| **Alpha=0,SimpleExponentialSmoothing** | 1304.927405 | 44.48 |
| **9pointTrailingMovingAverage** | 1346.278315 | 46.86 |
| **Alpha=0.1,SimpleExponentialSmoothing** | 1375.393398 | 49.53 |
| **RegressionOnTime** | 1389.135175 | 50.15 |
| **Alpha=0.2,SimpleExponentialSmoothing** | 1595.206839 | 60.46 |
| **Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing** | 1779.424896 | 67.23 |
| **NaiveModel** | 3864.279352 | 152.87 |
| **Alpha=0.65,Beta=0,DoubleExponentialSmoothing** | 5291.879833 | 208.74 |

**Table 1.30. Best Performing Models**

- Tuned Triple Exponential Model with Alpha = 0.1, Beta = 0, and Gamma = 0.3 having test RMSE 333.275 and MAPE 10.16
- Automated SARIMA with seasonal differencing - SARIMA(1, 1, 2)(0, 1, 2, 12) having test RMSE 393.11 and MAPE 12.87
- Automated SARIMA with seasonal differencing - SARIMA(3, 1, 2)(1, 1, 1, 12) having test RMSE 382.57 and MAPE 13.27

**We'll forecast on Top 2 Models**

➢ Forecasting on Tuned Triple Exponential Model with Alpha = 0.01, Beta = 0.04, Gamma = 0.25

To forecast 12 months into the future, we build the model on the full data first before forecasting

RMSE Full Model = 416.50 ○ Assumption: Forecast distribution's standard deviation ≈ Residual standard deviation.

- Purpose: Helps estimate uncertainty in the forecast.

Use: Construct confidence intervals with a specified level of confidence.

From our comparison between models, **the top 2 models** are:

1. Triple Exponential Smoothing Model (alpha=0.1, beta = 0, gamma = 0.3)
2. SARIMAX (1,1,2)x(0,1,2,12).

We will re-build these 2 models on the complete data, and make 12 month forecasts.

**Part A: TES model on complete data**

Building a Triple Exponential Smoothing (alpha = 0.1, beta = 0, gamma = 0.3) model on the complete data, and forecasting for the next 12 months

Model Summary:

```
                       ExponentialSmoothing Model Results
================================================================================
Dep. Variable:                    endog   No. Observations:                  187
Model:             ExponentialSmoothing   SSE                     23177578.963
Optimized:                         True   AIC                         2225.059
Trend:                         Additive   BIC                         2276.757
Seasonal:                Multiplicative   AICC                        2229.130
Seasonal Periods:                    12   Date:                Fri, 14 Aug 2020
Box-Cox:                          False   Time:                        20:47:45
Box-Cox Coeff.:                    None
=================================================================================
                          coeff                 code              optimized
---------------------------------------------------------------------------------
smoothing_level               0.1000000                alpha                False
smoothing_slope                0.000000                 beta                False
smoothing_seasonal            0.3000000                gamma                False
initial_level                 1580.0000                  l.0                 True
initial_slope                 0.0100000                  b.0                 True
initial_seasons.0             1.0670886                  s.0                 True
initial_seasons.1             1.0069620                  s.1                 True
initial_seasons.2             1.4582278                  s.2                 True
initial_seasons.3             1.0835443                  s.3                 True
initial_seasons.4             0.9310127                  s.4                 True
initial_seasons.5             0.8715190                  s.5                 True
initial_seasons.6             1.2443038                  s.6                 True
initial_seasons.7             1.5525316                  s.7                 True
initial_seasons.8             1.2556962                  s.8                 True
initial_seasons.9             1.6430380                  s.9                 True
initial_seasons.10            2.5867089                 s.10                 True
initial_seasons.11            3.2778481                 s.11                 True
---------------------------------------------------------------------------------
```

Fig: 1.31 Exponential smoothing models result

TES (alpha=0.1, beta = 0, gamma = 0.3) Model 12-month Forecast:

| Timeline | Sparkling Wine Sales Forecast |
|---|---|
| 1995-08-01 | 1926.181 |
| 1995-09-01 | 2392.240 |
| 1995-10-01 | 3214.95643 |
| 1995-11-01 | 3940.48285 |
| 1995-12-01 | 6063.414727 |
| 1996-01-01 | 1354.0173521 |
| 1996-02-01 | 1628.346929 |
| 1996-03-01 | 1864.353429 |
| 1996-04-01 | 1828.938459 |
| 1996-05-01 | 1675.33698 |
| 1996-06-01 | 1594.829942 |
| 1996-07-01 | 2004.61669 |

Table:1.31 12- mponth forecast

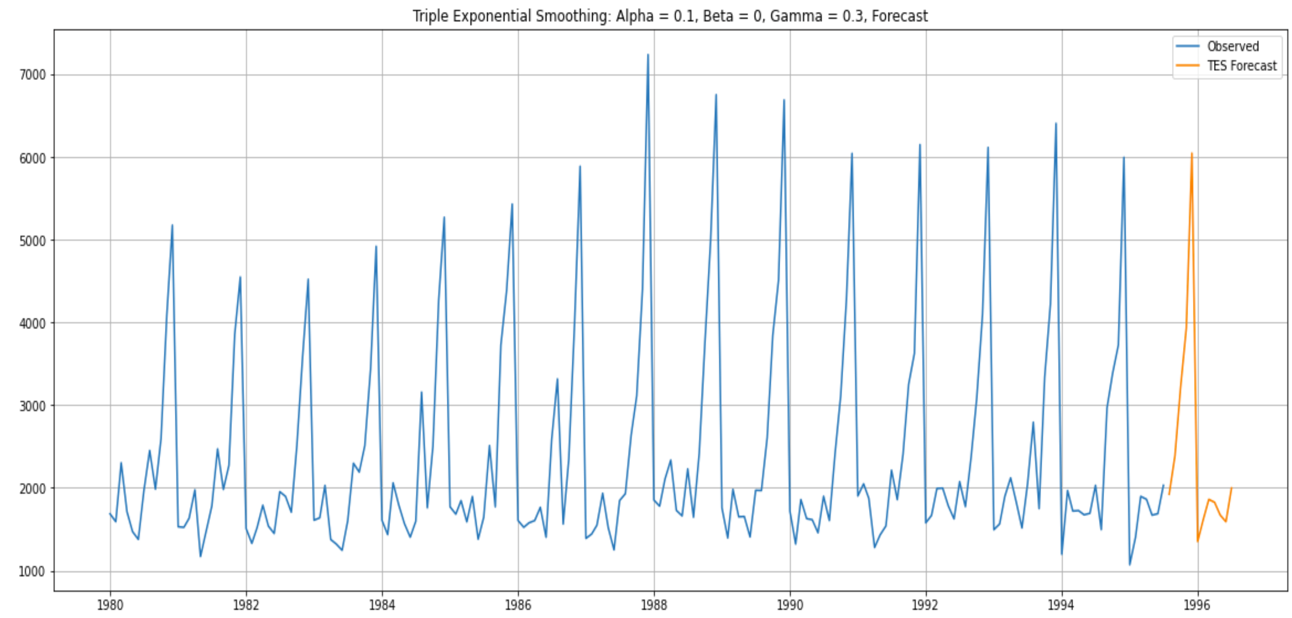



Fig: 1.32 Time series forecast for TES

**Part B: SARIMA model on complete data**

Building a SARIMAX (1,1,2)x(0,1,2,12) model on the complete data, and forecasting for the next 12 months

Model Summary:

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  187
Model:             SARIMAX(1, 1, 2)x(0, 1, 2, 12)  Log Likelihood               -1086.537
Date:                            Fri, 14 Aug 2020  AIC                           2185.074
Time:                                    21:04:10  BIC                           2203.017
Sample:                                        0   HQIC                          2192.364
                                           - 187
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5604      0.367     -1.528      0.127      -1.279       0.159
ma.L1         -0.2809      0.339     -0.830      0.407      -0.945       0.383
ma.L2         -0.6547      0.311     -2.102      0.036      -1.265      -0.044
ma.S.L12      -0.5443      0.068     -8.052      0.000      -0.677      -0.412
ma.S.L24      -0.0177      0.085     -0.207      0.836      -0.185       0.150
sigma2      1.515e+05   1.54e+04      9.820      0.000    1.21e+05    1.82e+05
===================================================================================
Ljung-Box (Q):                       18.57   Jarque-Bera (JB):                36.50
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.81   Skew:                             0.67
Prob(H) (two-sided):                  0.46   Kurtosis:                         5.04
===================================================================================
```

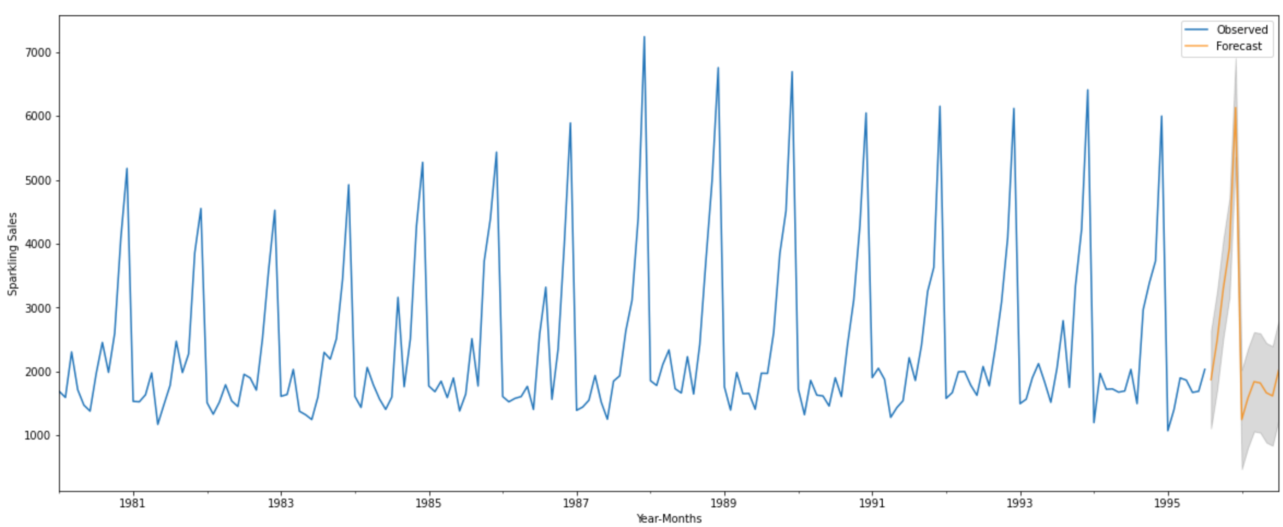
Fig:1.33 SARIMAX result summary

SARIMA model 12 month Forecast:

| timeline | mean | mean_se | mean_ci_lower | mean_ci_upper |
|---|---|---|---|---|
| 1995-08-01 | 1869.770067 | 389.241605 | 1106.870539 | 2632.669594 |
| 1995-09-01 | 2484.479875 | 394.108109 | 1712.042176 | 3256.917575 |
| 1995-10-01 | 3294.052682 | 394.223309 | 2521.389194 | 4066.716169 |
| 1995-11-01 | 3932.866644 | 395.394534 | 3157.907597 | 4707.825691 |
| 1995-12-01 | 6131.680558 | 395.475843 | 5356.562149 | 6906.798966 |
| 1996-01-01 | 1245.188368 | 396.010406 | 469.022235 | 2021.354502 |
| 1996-02-01 | 1580.094096 | 396.241732 | 803.474572 | 2356.713619 |
| 1996-03-01 | 1837.50603 | 396.62706 | 1060.131278 | 2614.880783 |
| 1996-04-01 | 1818.652204 | 396.920827 | 1040.701677 | 2596.60273 |
| 1996-05-01 | 1664.131274 | 397.263937 | 885.508265 | 2442.754284 |
| 1996-06-01 | 1615.681123 | 397.578529 | 836.441526 | 2394.920721 |
| 1996-07-01 | 2016.79348 | 397.908527 | 1236.907098 | 2796.679862 |

Table 1.32 12-month forecast

SARIMA (1,1,2)x(0,1,2,12) model 12-month Forecast visualised:




12-month forecast Time series

## 1.9 Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

Forecasting Insights:

- **Data Analysis:**
    - The sales of Sparkling products have shown a steady increasing trend over the years, indicating a growth pattern in customer demand.
    - Seasonal patterns are evident in the data, with sales spiking during November and December, which can be attributed to the holiday season when customers tend to buy more Sparkling products for celebrations.
- **Time Series Characteristics:**
    - Based on the decomposed data provided, the time series is observed to be of a multiplicative nature.
    - the trend appears to be increasing over time with some fluctuations, suggesting a growth pattern.
    - Seasonality shows both positive and negative values, indicating regular cycles or seasonal effects.
    - the residual component includes random fluctuations and unexplained variance in the time series.
- **Model Performance**
    - After evaluating various forecasting models, the top-performing ones are:
    - Tuned Triple Exponential Model (Alpha = 0.1, Beta = 0, Gamma = 0.3): It shows the best accuracy with a Test RMSE of 333.4673 and MAPE of 10.16.
        - Automated SARIMA with Seasonal Differencing (SARIMA(2, 1, 2)(0, 1, 2, 12)): It also performs well, with a Test RMSE of 386.06 and MAPE of 12.87.
        - Triple Exponential Smoothing (Alpha = 0.111, Beta = 0.049, Gamma = 0.362): This model is the forth best with a Test RMSE of 402.931 and MAPE of 13.88.
        - On the other hand, the Alpha=0.665, Beta = 0.0001 Double Exponential Smoothing model performs poorly with a high Test RMSE of 5291.88 and MAPE of 208.74.
- **Predictions Model**
    - For comprehensive forecasting and predicting 12 months ahead, we built the following models:
        - Tuned Triple Exponential Model with Alpha = 0.1, Beta = 0, and Gamma = 0.3, the RMSE of the Full Model is 416.50.

- Automated SARIMA with Seasonal Differencing (SARIMA(0, 0, 2)(0, 1, 2, 12)): It also performs well, with a full model RMSE of 539.99

**Measures for Future Sales**

- Based on the analysis, the company can take the following measures to improve future sales:
  - Capitalize on Seasonal Trends: With observed seasonal patterns during November and December, the company should plan production and marketing efforts to meet increased demand during holiday seasons.
  - Inventory Management: Implement effective inventory management to avoid stockouts during peak periods and minimize excess inventory during slower periods.
  - Pricing Strategy: Utilize dynamic pricing to adjust prices during peak and off-peak periods, attracting more customers and optimizing revenue.

Customer Engagement: Strengthen customer relationships through personalized offers, loyalty programs, and active engagement to foster repeat purchases.

- The Sparkling Wine Sales have steadied over the last four years, and the levels don't show any significant growth or decline.
- The seasonal patterns recur consistently, especially the last quarter which sees the maximum sales every year.
- The models appear to forecast a similar range and pattern for the next 12 months. And given the low RMSE score, along with a certain consistency of past behavior, the forecast looks dependable.
- The consumption pattern points to the fact that Sparkling Wines are most in demand in holiday and festive seasons, and is therefore positioned as a premium product, and meant for special occasions.
- To increase sales, we could look to capitalize on the last quarter demand, and create campaigns to push more consumption during the period. This would maintain the special-ness of the product.
- We could alternatively promote Sparkling Wine consumption during other festive periods and other seasons conducive for it. There are smaller spikes in consumption during March and June, which can be explored.
- Also, other geographies can be explored to tap into other festive seasons globally.

# Problem 2: Rose Wine Sales

## Problem Statement:

**As an analyst in the ABC Estate Wines, your task is to analyze and forecast Wine Sales in the 20th century. Data set for the Problem: Rose.csv**

## Data Dictionary:

**Year Month: Month & Year of the sale**

**Rose: Total Number of Rose Wine sales in particular Month-Year**

## 2.1 Read the data as an appropriate Time Series data and plot the data. Read the data as an appropriate Time Series data and plot the data.

## Basic Information about the dataset

- **Sample of the dataset: First & last 5 values of the dataset:**

| | 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 |

| | 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 |

**Table2.1 First 5 and Last 5 Samplesofthe Dataset**

- **Converting the Year Month Column to Date time Index & dropping default index. Sample:**

| Year_Month | Rose |
|---|---|
| 1980-01-31 | 112.0 |
| 1980-02-29 | 118.0 |
| 1980-03-31 | 129.0 |
| 1980-04-30 | 99.0 |
| 1980-05-31 | 116.0 |

**Table 2.2 First 5 Samples of the Converted Dataset**

- **Information about the dataset:**

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

**Table 2.3 Info of the Dataset**

- The DataFrame has 187 entries with a Date time Index ranging from January 1980 to July 1995.
- The 'Rose' column is of integer type (int64), and it has 187 non-null values.

- **Time Series Plot:**

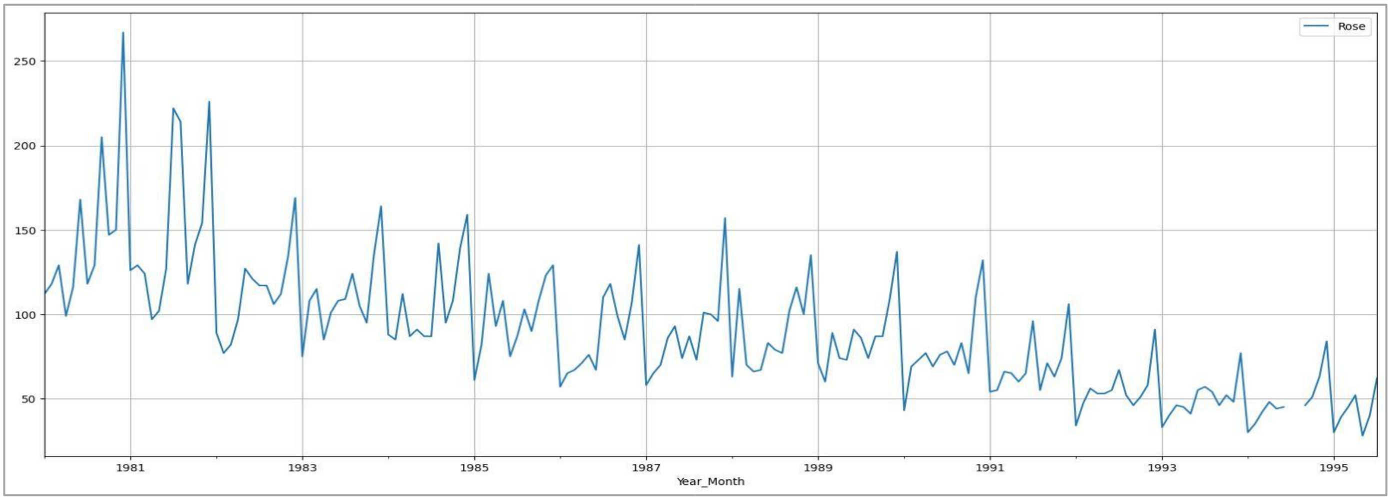



**Fig 2.1 Time Series Plot**

- Decreasing Trend: The data shows a consistent decrease in Rose wine sales over the years, suggesting a declining pattern in customer demand for this type of wine.
- Seasonal Patterns: Despite the overall decreasing trend, there are still seasonal variations in the data, with sales peaking in certain months and dropping in others.
- Missing Data: There are missing values in the year 1994 and which will be handled accordingly

## 2.2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

- **Missing Values:**

- There are 2 Null values in the dataset, for 07-1994 & 08-1994, which need to be corrected. Let's check the Null values
- Interpolate the missing values using cubic interpolation.

| Year_Month | Rose |
|---|---|
| 1994-07-31 | NaN |
| 1994-08-31 | NaN |

| Year_Month | Rose |
|---|---|
| 1994-07-31 | 45.270014 |
| 1994-08-31 | 44.504928 |

**Table 2.4 Missing Values Before & After Treatment**

➢ **Time Series Plot Post Null Treatment:**

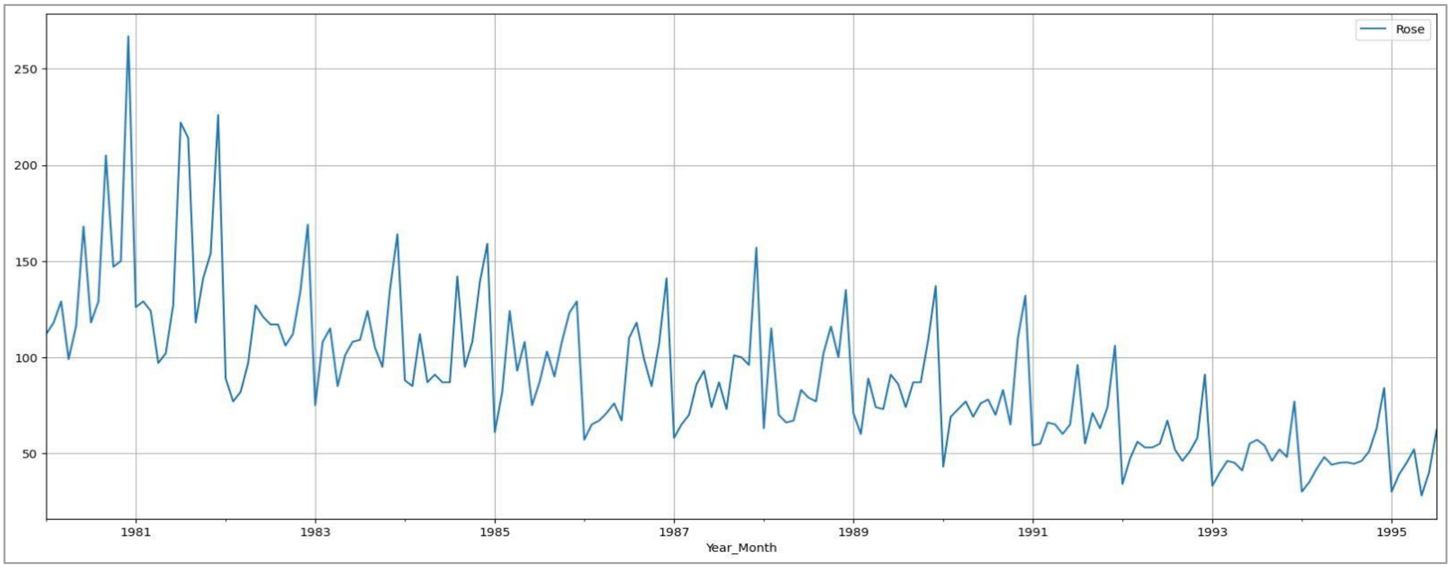



**Fig 2.1 b Time Series Plot Post Null Treatment**

➢ **Duplicate Values:**

o The dataset shows 90 Duplicates rows for Rose wine sales, however, when checked further, these were the same no of sales at different year. Hence, we conclude that there are no duplicate values

➢ **DescriptiveStatistics:**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Rose** | 187.0 | 89.907887 | 39.245847 | 28.0 | 62.5 | 85.0 | 111.0 | 267.0 |

**Table 2.5 Descriptive Statistic**

- The dataset contains 187 observations of Rose product sales.
- It shows significant variability, with sales ranging from a minimum of 28 units to a maximum of 267 units.
- On average, the monthly sales of Rose wine are approximately 89.91 units ○ The data is right-skewed, with the mean 89.91 being higher than the median 85

- **Histogram & Boxplot:**

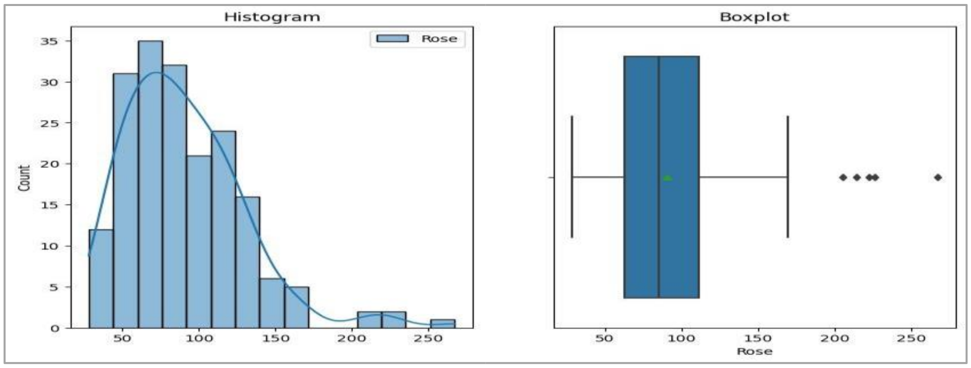



**Fig 2.2 Histogram & Boxplot**

- The dataset is right skewed with the presence of outliers on the right tail

- **Spread of Sales: Year-on-Year Boxplot**

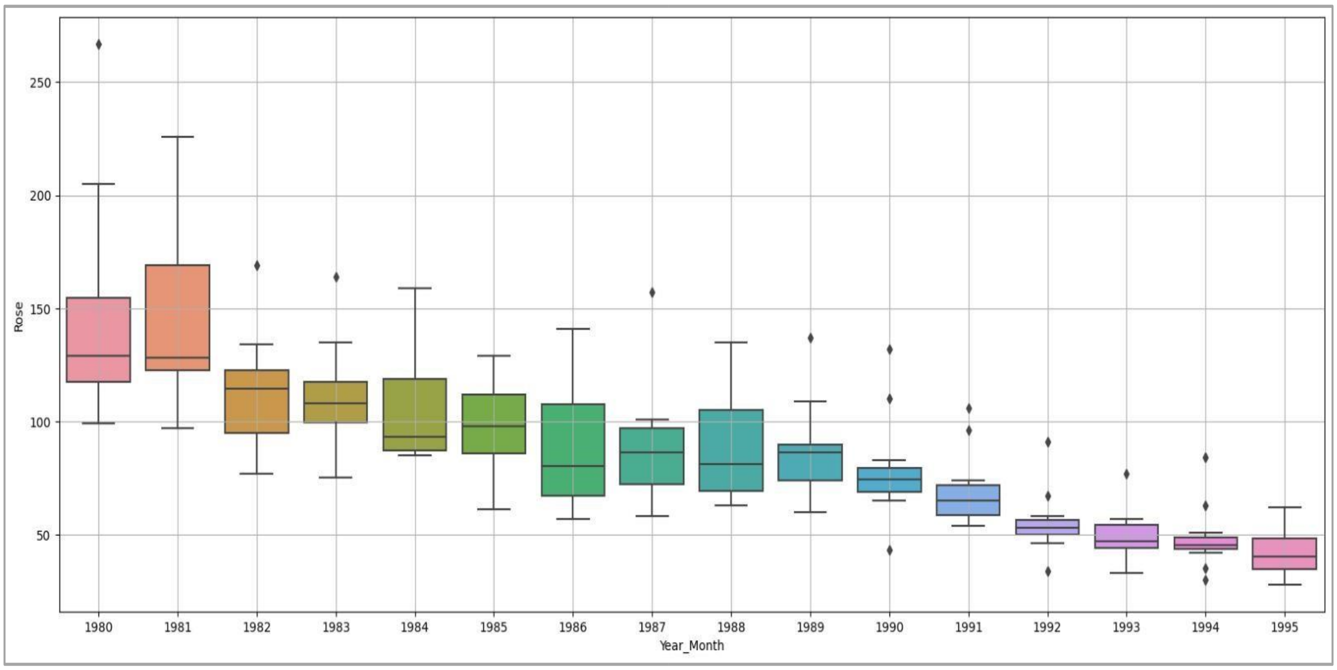



**Fig2.3 Spread of Sale Across Different Years**

- Sales of Rose products have been declining over time.
- the dataset is skewed with the presence of outliers on the right tail.
- There is significant variability in sales from year to year There are a few outliers in the

data, which could be due to special promotions, holidays or other occasions

➢ **Spread of Sales: Month-on- Month**

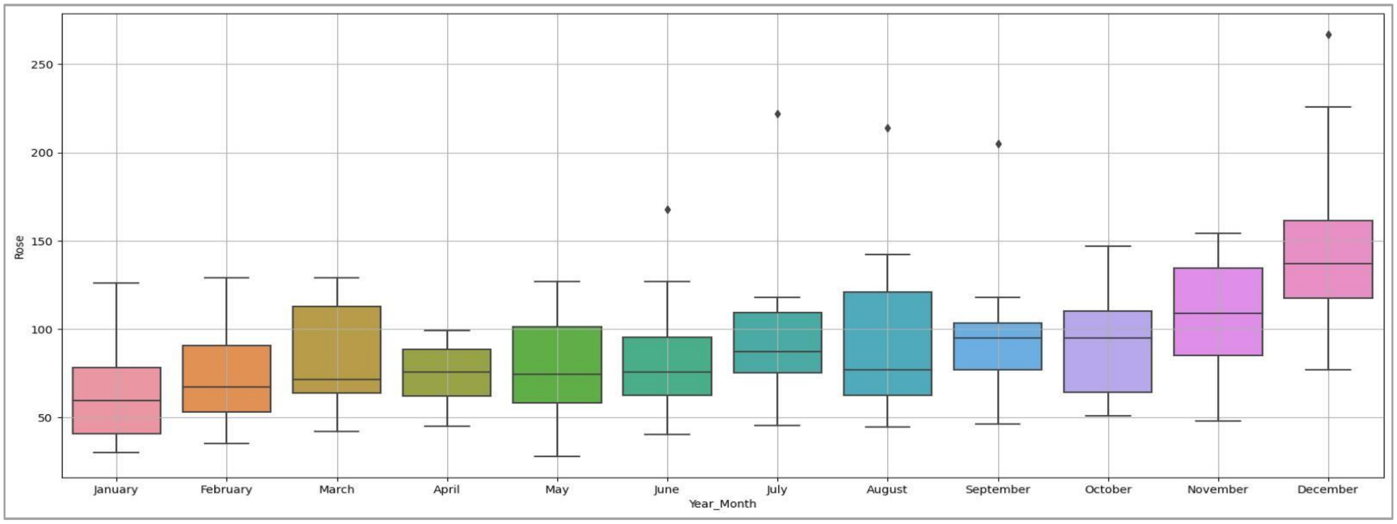



**Fig 2.4 Spread of Sales Across Different Months**

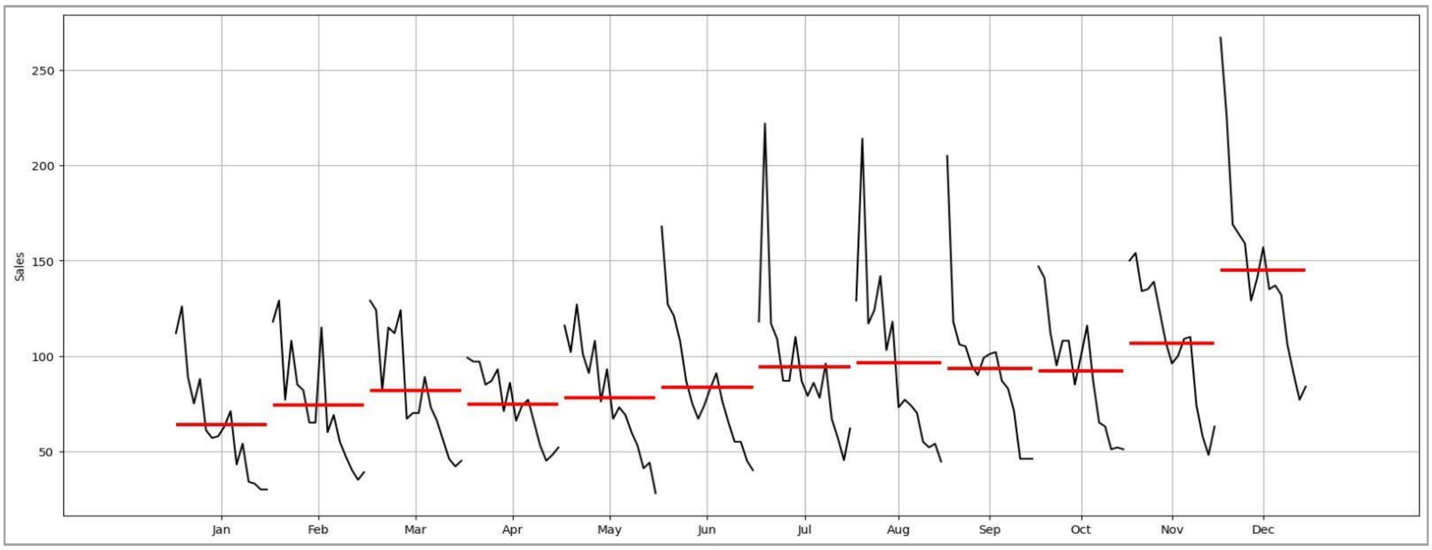



**Fig2.5 Distribution of time series across different months**

- The Month Plot demonstrates a downward trend across all years for each month, indicating consistent seasonality throughout the months.
- Sales of Rose wine exhibit seasonality, with a significant increase from August to December, peaking in December due to the winter season and Christmas Holidays.
- There is a sharp drop in sales in January (post the holiday season), and stability can be seen in the mid-months (March to August). o Outliers on the maximum side likely correspond to December sales.

➢ **Spread of Sales: Year-on-Year & Monthly Comparison:**

| Year_Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Year_Month** | | | | | | | | | | | | |
| **1980** | 112.0 | 118.0 | 129.0 | 99.0 | 116.0 | 168.0 | 118.000000 | 129.000000 | 205.0 | 147.0 | 150.0 | 267.0 |
| **1981** | 126.0 | 129.0 | 124.0 | 97.0 | 102.0 | 127.0 | 222.000000 | 214.000000 | 118.0 | 141.0 | 154.0 | 226.0 |
| **1982** | 89.0 | 77.0 | 82.0 | 97.0 | 127.0 | 121.0 | 117.000000 | 117.000000 | 106.0 | 112.0 | 134.0 | 169.0 |
| **1983** | 75.0 | 108.0 | 115.0 | 85.0 | 101.0 | 108.0 | 109.000000 | 124.000000 | 105.0 | 95.0 | 135.0 | 164.0 |
| **1984** | 88.0 | 85.0 | 112.0 | 87.0 | 91.0 | 87.0 | 87.000000 | 142.000000 | 95.0 | 108.0 | 139.0 | 159.0 |
| **1985** | 61.0 | 82.0 | 124.0 | 93.0 | 108.0 | 75.0 | 87.000000 | 103.000000 | 90.0 | 108.0 | 123.0 | 129.0 |
| **1986** | 57.0 | 65.0 | 67.0 | 71.0 | 76.0 | 67.0 | 110.000000 | 118.000000 | 99.0 | 85.0 | 107.0 | 141.0 |
| **1987** | 58.0 | 65.0 | 70.0 | 86.0 | 93.0 | 74.0 | 87.000000 | 73.000000 | 101.0 | 100.0 | 96.0 | 157.0 |
| **1988** | 63.0 | 115.0 | 70.0 | 66.0 | 67.0 | 83.0 | 79.000000 | 77.000000 | 102.0 | 116.0 | 100.0 | 135.0 |
| **1989** | 71.0 | 60.0 | 89.0 | 74.0 | 73.0 | 91.0 | 86.000000 | 74.000000 | 87.0 | 87.0 | 109.0 | 137.0 |
| **1990** | 43.0 | 69.0 | 73.0 | 77.0 | 69.0 | 76.0 | 78.000000 | 70.000000 | 83.0 | 65.0 | 110.0 | 132.0 |
| **1991** | 54.0 | 55.0 | 66.0 | 65.0 | 60.0 | 65.0 | 96.000000 | 55.000000 | 71.0 | 63.0 | 74.0 | 106.0 |
| **1992** | 34.0 | 47.0 | 56.0 | 53.0 | 53.0 | 55.0 | 67.000000 | 52.000000 | 46.0 | 51.0 | 58.0 | 91.0 |
| **1993** | 33.0 | 40.0 | 46.0 | 45.0 | 41.0 | 55.0 | 57.000000 | 54.000000 | 46.0 | 52.0 | 48.0 | 77.0 |
| **1994** | 30.0 | 35.0 | 42.0 | 48.0 | 44.0 | 45.0 | 45.270014 | 44.504928 | 46.0 | 51.0 | 63.0 | 84.0 |
| **1995** | 30.0 | 39.0 | 45.0 | 52.0 | 28.0 | 40.0 | 62.000000 | NaN | NaN | NaN | NaN | NaN |

**Table 2.6 Year-on-Year Monthly Sales**

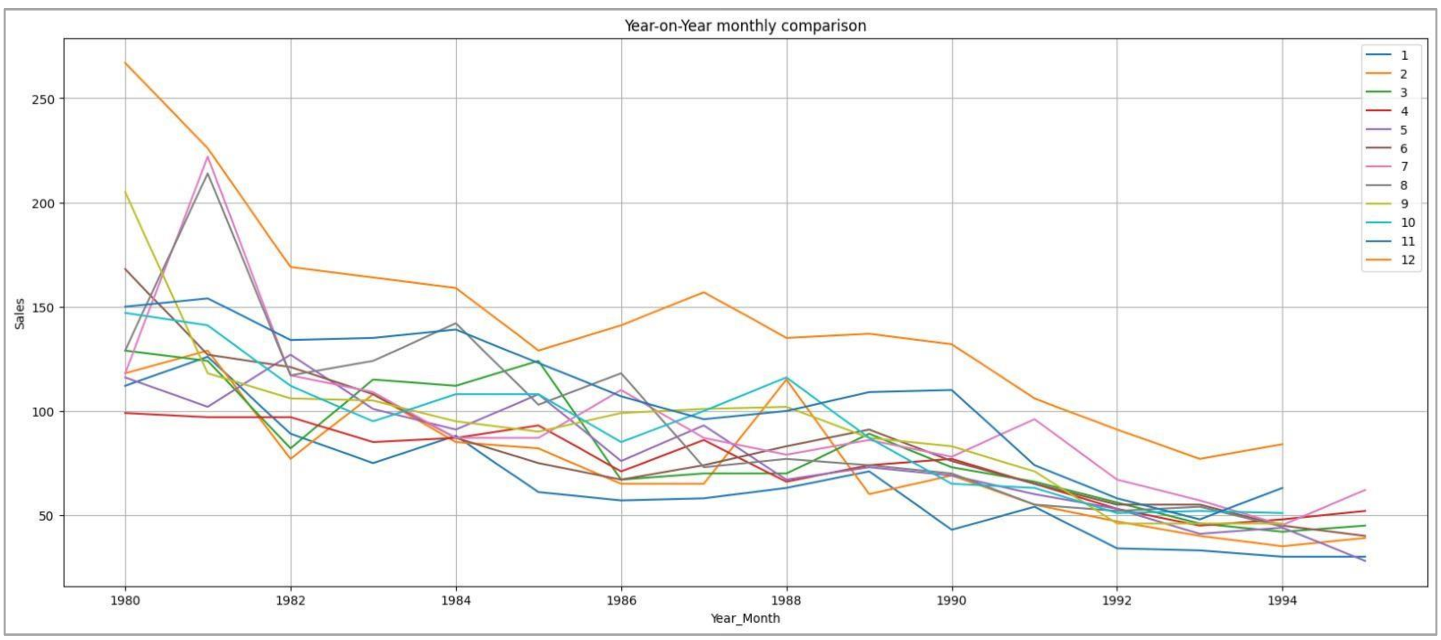



**Fig 2.6 Year-on-Monthly Comparison**

- December consistently shows the highest sales across all the years, while January has the lowest sales for most years.
- A noticeable downward trend can be observed across all months over the years.
- The outliers in the monthly boxplot are likely from the years 1980 or 1981.

- **Empirical Cumulative Distribution Plot**

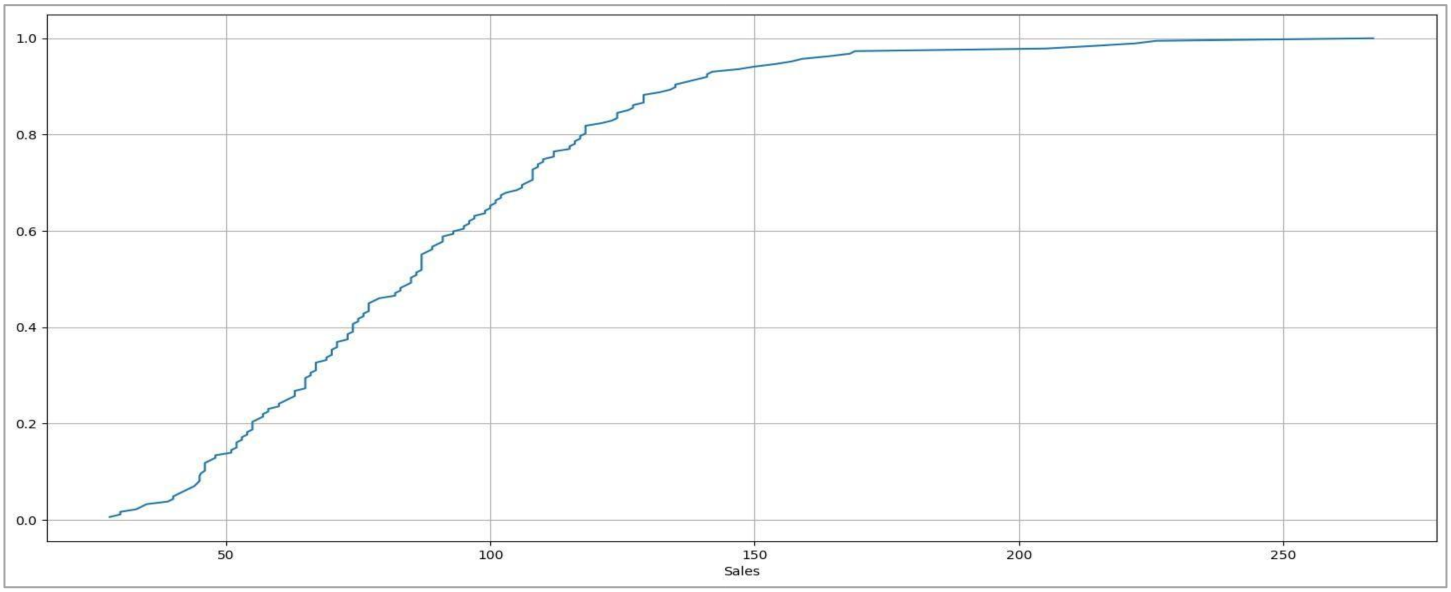



**Fig 2.7 Empirical Cumulative Distribution Plot**

- This graph tells us what percentage of data points refer to what number of Sales.

- **Average Rose Sales**

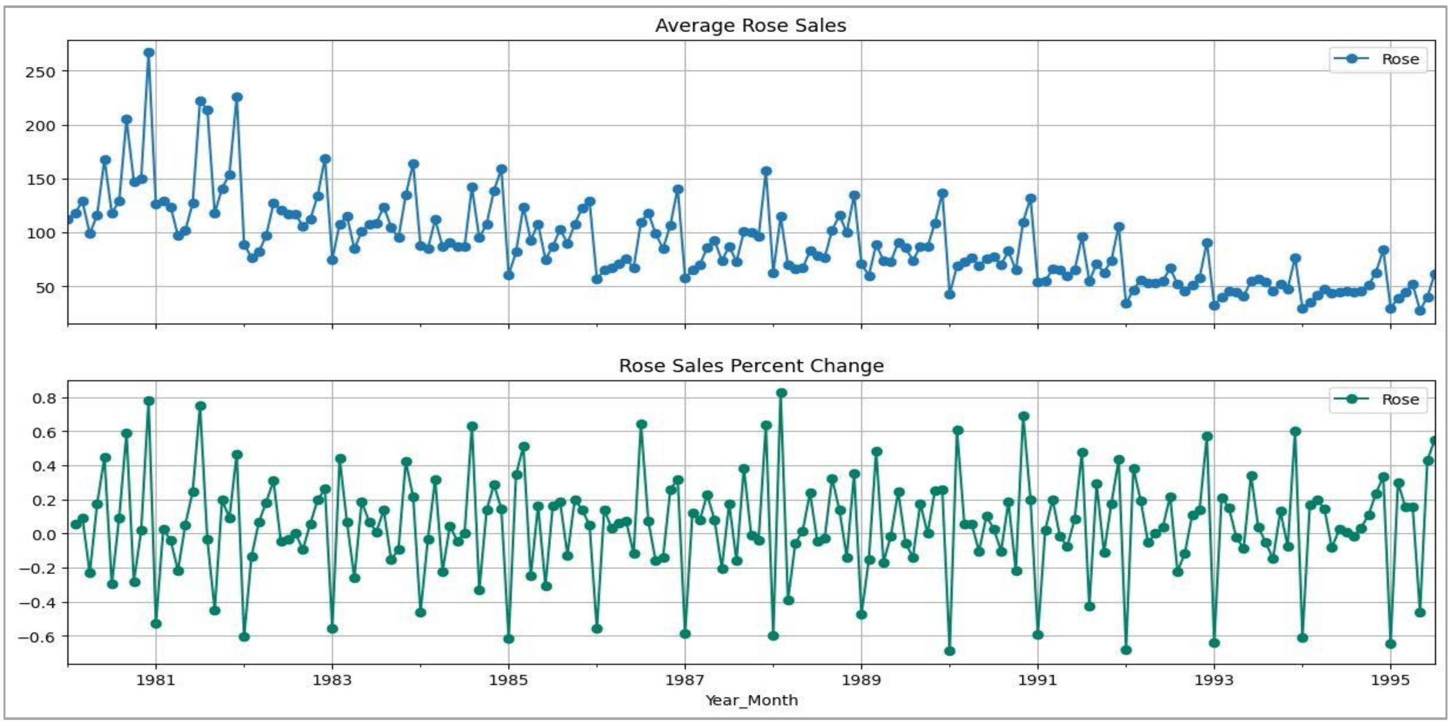



**Fig2.8 Year-on-Average Sale**

- the average sales are declining year-on-year. This is evident from the fact that the line graph is generally decreasing.
- There is a seasonal pattern in sales, with sales being highest in December.

## Decomposition:

➢ **Additive Decomposition:**

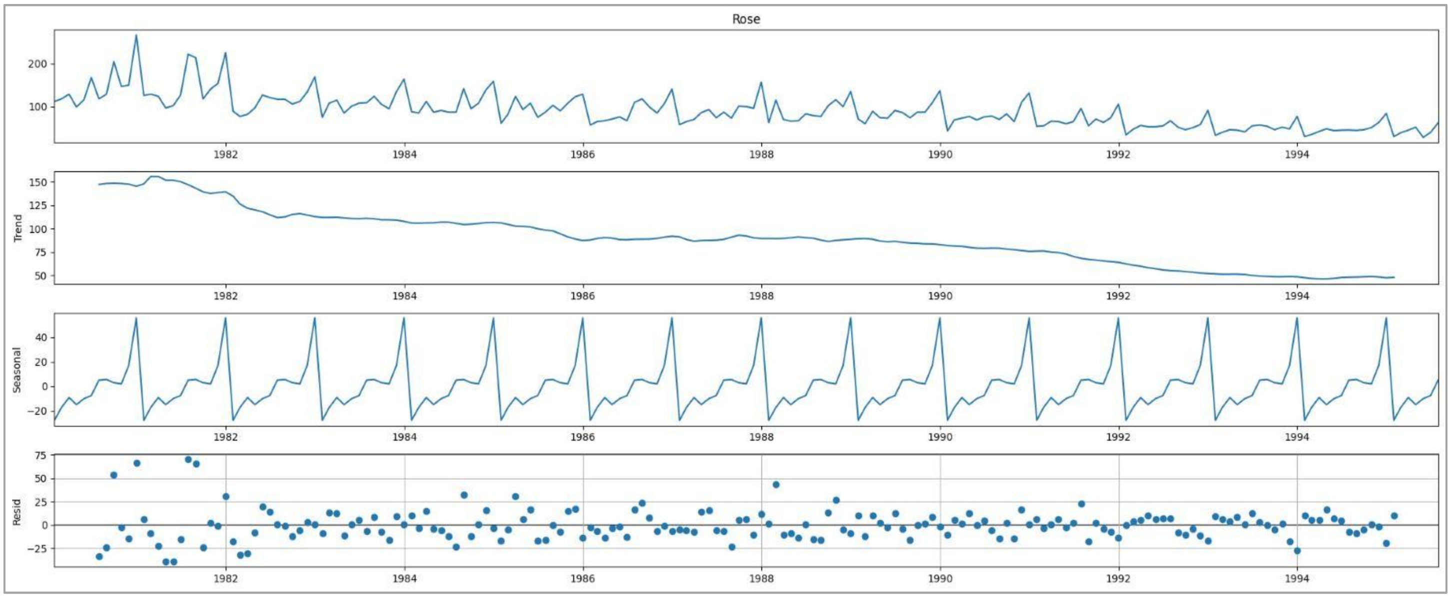



**Fig 2.9 Decomposed Time Series Additive**

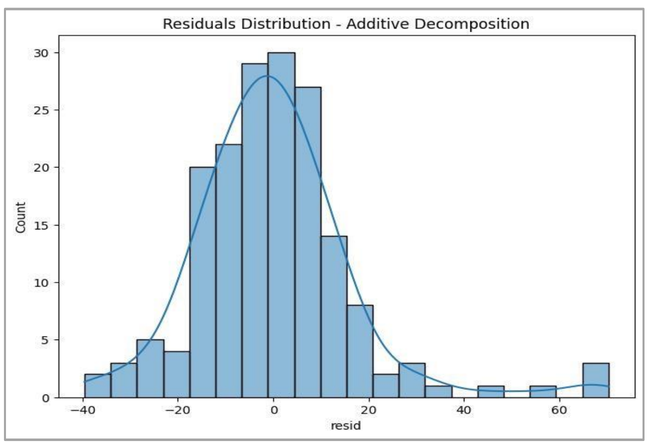



**Fig 2.10 Residuals Histogram Additive Decomposition**

- **Test for Normality**

We will use the Shapiro Wilk Test for Normality. Let's define the Null & alternate hypothesis: -

Ho: The residuals are normally distributed

Ha: The residuals are not normally distributed

p–value of the Shapiro-Wilk Test on the residuals = ~7.984797711912961e-09

Since the p-value < 0.05 - We Reject the null hypothesis.

Hence Residuals are not normally distributed at 95% confidence level. The time series is not an additive time series.

## Multiplicative Decomposition:

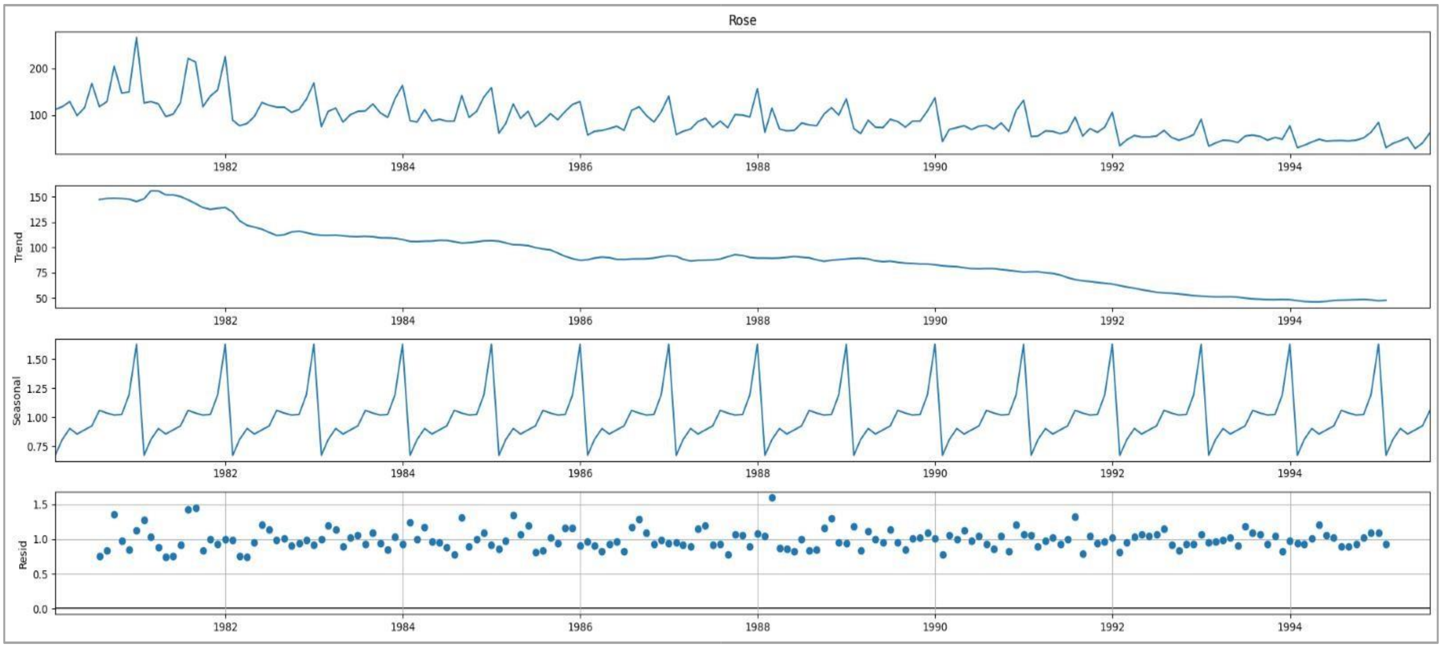



**Fig 2.11 Decomposed Time Series Multiplicative**

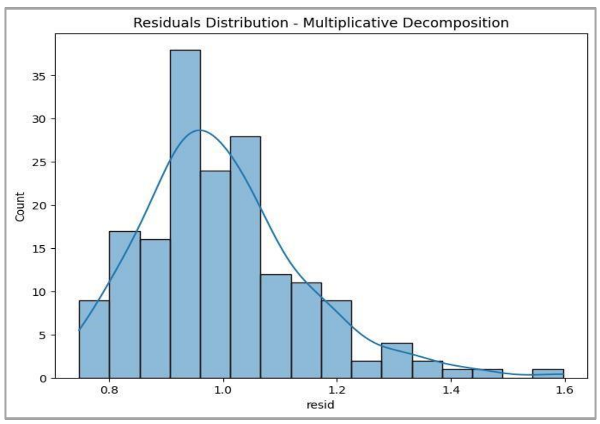



**Fig 2.12 Residuals Histogram Multiplicative Decomposition**

- **Test for Normality**

We will use the Shapiro Wilk Test for Normality. Let's define the Null & alternate hypothesis: -

Ho: The residuals are normally distributed Ha:
The residuals are not normally distributed

p–value of the Shapiro-Wilk Test on the residuals = 6.26639530310058e-06

Since the p-value < 0.05: we fail to reject the Null hypothesis

Residuals are not normally distributed at 95% confidence level. It cannot be determined from this that the time series is a pure multiplicative time series either.

➢ **Time series components for Multiplicative:**

Trend

| Year_Month | |
|---|---|
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 147.083333 |
| 1980-08-31 | 148.125000 |
| 1980-09-30 | 148.375000 |
| 1980-10-31 | 148.083333 |
| 1980-11-30 | 147.416667 |
| 1980-12-31 | 145.125000 |

Name: trend, dtype: float64

Seasonality

| Year_Month | |
|---|---|
| 1980-01-31 | 0.670198 |
| 1980-02-29 | 0.806223 |
| 1980-03-31 | 0.901304 |
| 1980-04-30 | 0.854185 |
| 1980-05-31 | 0.889558 |
| 1980-06-30 | 0.924126 |
| 1980-07-31 | 1.058081 |
| 1980-08-31 | 1.034408 |
| 1980-09-30 | 1.017778 |
| 1980-10-31 | 1.022716 |
| 1980-11-30 | 1.192528 |
| 1980-12-31 | 1.628895 |

Name: seasonal, dtype: float64

Residual

| Year_Month | |
|---|---|
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 0.758228 |
| 1980-08-31 | 0.841918 |
| 1980-09-30 | 1.357501 |
| 1980-10-31 | 0.970636 |
| 1980-11-30 | 0.853249 |
| 1980-12-31 | 1.129473 |

Name: resid, dtype: float64

**Table2.7 Decomposed TimeSeriesComponents**

- The time series shows a clear downward trend across the years, with a sharper dip observed after 1991 compared to before 1991.
- Seasonality is present in the data, as sales pick up in the ending months of the year.
- While the time series exhibits characteristics closer to a multiplicative nature, it cannot be definitively classified as either an additive or multiplicative time series

## 2.3 Split the data into training and test. The test data should start in 1991.

- The data was split into a train and test set.
- The splitting was done chronologically, with data from the year 1991 forming the test set.
- The train set contains 132 records, while the test set contains 55 records.

```
Dimentions of Original Dataset: (187, 1)
Dimentions of Training data: (132, 1)
Dimentions of Training data: (55, 1)
```

**Table 2.8 Dimensions of Original, Train & Test Data**

➢ **Training data sample**

First few rows of Train

| Year_Month | Rose |
|---|---|
| 1980-01-31 | 112.0 |
| 1980-02-29 | 118.0 |
| 1980-03-31 | 129.0 |
| 1980-04-30 | 99.0 |
| 1980-05-31 | 116.0 |

Last few rows of Train

| Year_Month | Rose |
|---|---|
| 1990-08-31 | 70.0 |
| 1990-09-30 | 83.0 |
| 1990-10-31 | 65.0 |
| 1990-11-30 | 110.0 |
| 1990-12-31 | 132.0 |

**Table2.9 Sampleof Training Data**

- **Test data sample**

First few rows of Test

| Year_Month | Rose |
|---|---|
| 1991-01-31 | 54.0 |
| 1991-02-28 | 55.0 |
| 1991-03-31 | 66.0 |
| 1991-04-30 | 65.0 |
| 1991-05-31 | 60.0 |

Last few rows of Test

| Year_Month | Rose |
|---|---|
| 1995-03-31 | 45.0 |
| 1995-04-30 | 52.0 |
| 1995-05-31 | 28.0 |
| 1995-06-30 | 40.0 |
| 1995-07-31 | 62.0 |

**Table 2.10 Sample of Test Data**

- **Train Test Split Plot**

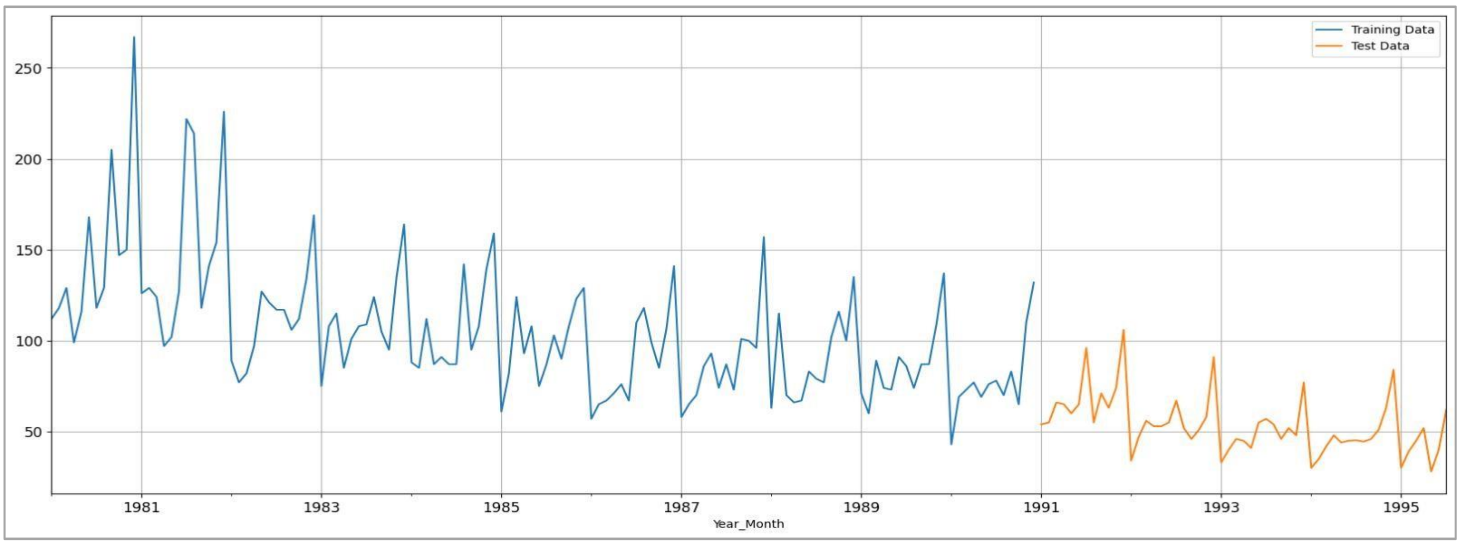



**Fig 2.13 Train & Test Split Time Series**

## 2.4 Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

**<u>Building different models and comparing the accuracy metrics.</u>**

- **Linear Regression Model**

For this linear regression, we are going to regress the 'Rose' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

We generated the numerical time instance order for both the training and test set. Sample of the Train & Test data

**Model 1: Linear Regression**

Model Output Visualised:

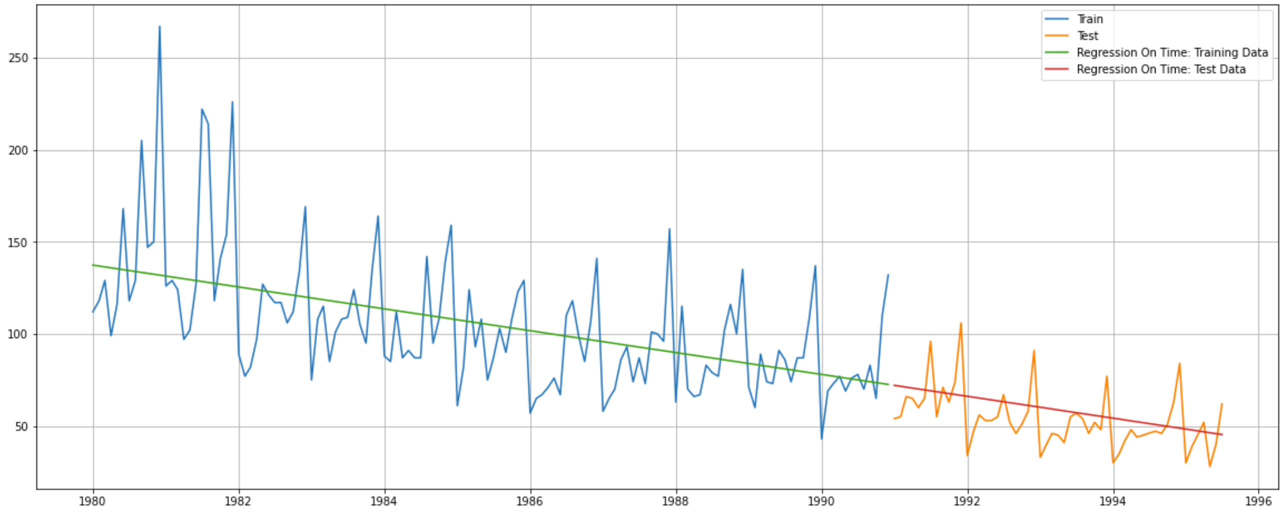



Fig: 2.14 time series plot – Linear Regression

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **30.718** | **21.22** |
| **Test Data** | **15.279** | **22.88** |

Table: 2.11 LR Model Performance

Observation:

A linear regression model captures the declining trend. It however is unable to capture the seasonal variation that is a characteristic feature of the time series.

- Linear regression captures the downward trend but not the seasonality.
- Test RMSE is 15.2798, MAPE is 22.88 for Linear Regression, indicating difficulty in handling seasonality.

➢ **Naïve Forecast Model**

The Naive model predicts tomorrow's value based on today's observation, and the day after tomorrow's prediction is also the same as today's value.

Samples of Train & test data after we trained on Naive model:

```
Last 5 values of Train data:      Last 5 values of Test data:
              Rose                               Rose
YearMonth                         YearMonth
1990-08-01    70.0                1995-03-01   45.0
1990-09-01    83.0                1995-04-01   52.0
1990-10-01    65.0                1995-05-01   28.0
1990-11-01   110.0                1995-06-01   40.0
1990-12-01   132.0                1995-07-01   62.0
```

Model Output Visualized:

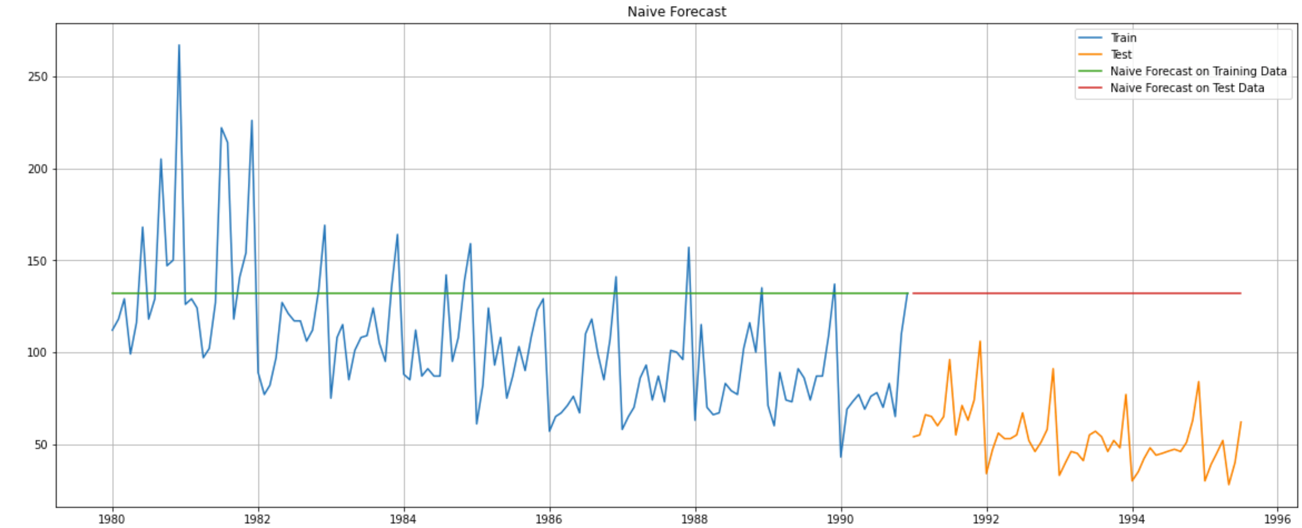



Fig: 2.15 Time Series plot for Naïve Model

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **45.064** | **36.38** |
| **Test Data** | **79.672** | **145.23** |

Table: 2.13 NM Performance model

Observation:

The Naive model is dependent on the last observed value, which in our training data is the month of December. We know that the month of December records peak sales every year. So this value is clearly not representative of the dataset at large, also when the overall sales are declining. Hence the expected very high RMSE scores on the Test data.

- Naive approach ignores trend and seasonality as it forecasts the last observed value
- Test RMSE is 79.74, MAPE is 145.23 for Naive forecast, showing significant errors due to the lack of trend and seasonality capture.

➢ **Simple Average Model:**

**For the simple average method, we will forecast by using the average of the training values.**

**Samples of Train & test data for Simple Average:**

| | Rose | naive |
|---|---|---|
| **YearMonth** | | |
| **1991-01-01** | 54.0 | 132.0 |
| **1991-02-01** | 55.0 | 132.0 |
| **1991-03-01** | 66.0 | 132.0 |
| **1991-04-01** | 65.0 | 132.0 |
| **1991-05-01** | 60.0 | 132.0 |

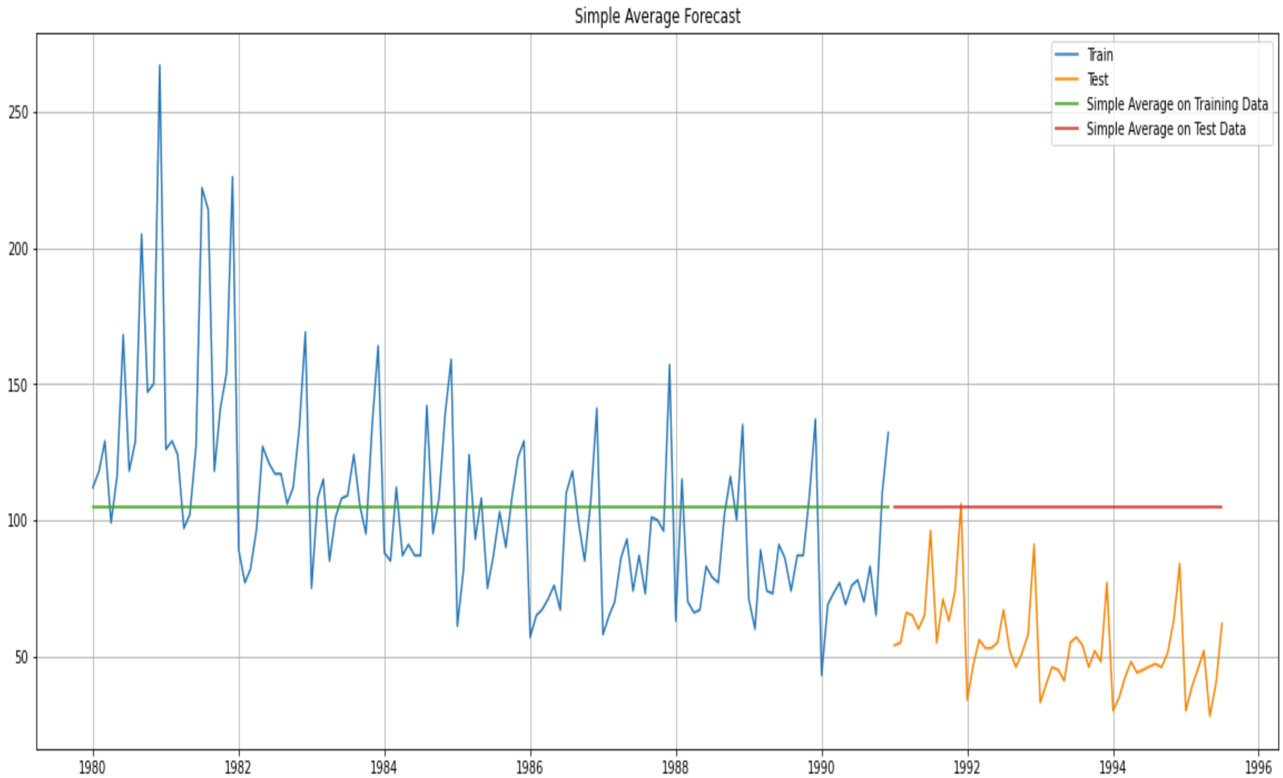



Fig: 2.16 Time Series plot for Average Model

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **36.034** | **25.39** |
| **Test Data** | **53.490** | **95.03** |

Table: 2.14 AM performance metrics summary

Observations:

A simple average model is not a great fit for the data, since it is unable to capture the trend, which in this case is one of decline. Also, seasonality is strong in this time series. It therefore misses out on much of the variation, resulting in high RMSE scores.

- Simple Average Model forecasts the mean of the training data. It ignores both the trend and seasonality.
- Test RMSE is 53.413, MAPE is 94.77. Errors are significant due to the lack of trend and seasonality capture.
- Performs better than Naïve, however, is not good enough for predictions

## ➢ Moving Average Model:

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals.

The best interval can be determined by the maximum accuracy (or the minimum error) over here.

For Moving Average, we are going to average over the entire data.

Moving Average Sample on Training data

| YearMonth | Rose | Trailing_2 | Trailing_4 | Trailing_6 | Trailing_9 |
|---|---|---|---|---|---|
| **1980-01-01** | 112.0 | NaN | NaN | NaN | NaN |
| **1980-02-01** | 118.0 | 115.0 | NaN | NaN | NaN |
| **1980-03-01** | 129.0 | 123.5 | NaN | NaN | NaN |
| **1980-04-01** | 99.0 | 114.0 | 114.50 | NaN | NaN |
| **1980-05-01** | 116.0 | 107.5 | 115.50 | NaN | NaN |
| **1980-06-01** | 168.0 | 142.0 | 128.00 | 123.67 | NaN |
| **1980-07-01** | 118.0 | 143.0 | 125.25 | 124.67 | NaN |
| **1980-08-01** | 129.0 | 123.5 | 132.75 | 126.50 | NaN |
| **1980-09-01** | 205.0 | 167.0 | 155.00 | 139.17 | 132.67 |
| **1980-10-01** | 147.0 | 176.0 | 149.75 | 147.17 | 136.56 |

Table: 2.15 Sample train Data of Moving Average

Model Output Visualised:

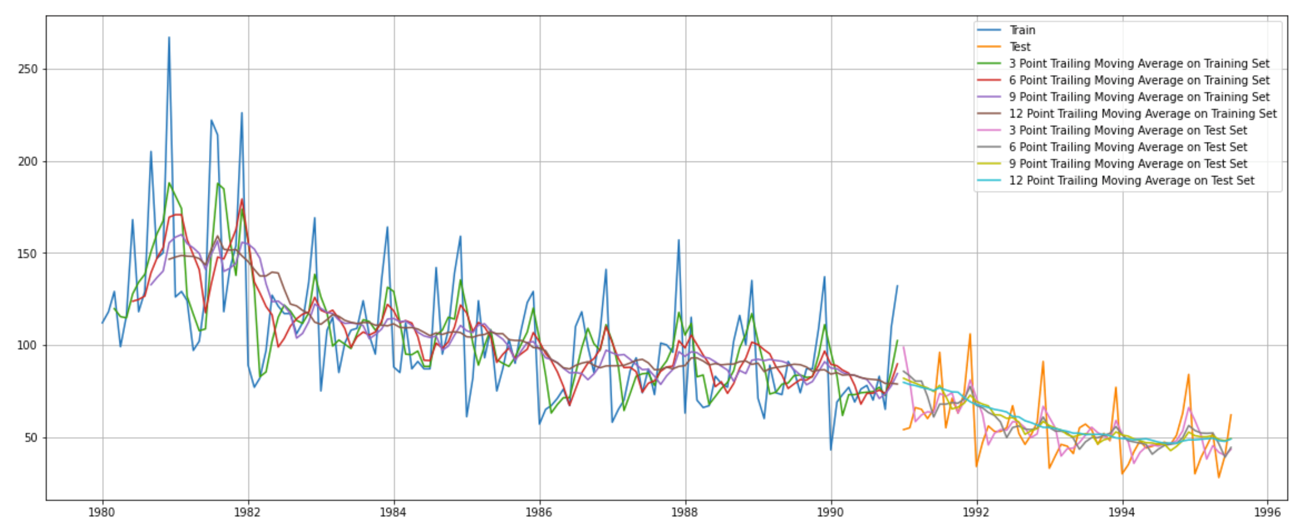



Fig: 2.17 Time Series Plot – Moving Average

Performance Metrics on Test Data:

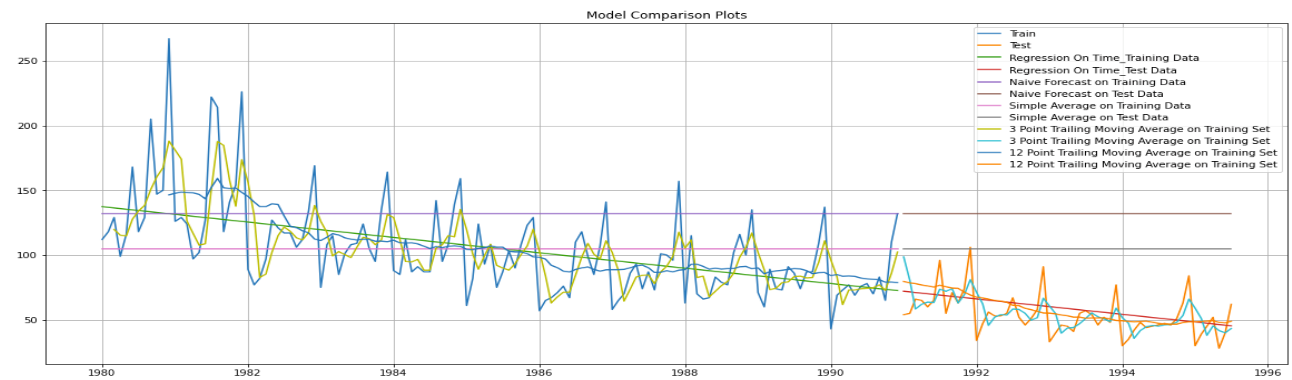
| | RMSE | MAPE |
|---|---|---|
| **2 point Moving Average** | **11.530** | **13.57** |
| **4 point Moving Average** | **14.458** | **19.52** |
| **6 point Moving Average** | **14.574** | **20.84** |
| **9 point Moving Average** | **14.733** | **21.07** |

Table: 2.16 Model Performance Summary – Moving Averages

Observation:

Moving Average Models are able to better track the variation of the time series. In this case it works especially as there is a clear trend. Plus the seasonal variation is also not very large. So one can see a relatively small difference in the performance of the 3-point and 12-point MA series.

Model Performance Comparison Visualised:




We will now proceed to the Exponential Models.

**For the moving average models, here are the insights based on their RMSE and MAPE values:**

- 2-Point Moving Average: RMSE = 19.67, Test RMSE = 11.53, Test MAPE = 13.57. It performs relatively well in capturing the trend and seasonality but still has room for improvement as it struggles to capture the sharp fluctuations in the sales pattern over time.
- 4-Point Moving Average: RMSE = 26.2, Test RMSE = 14.46, Test MAPE = 19.52. It performs slightly worse than the 2-Point model but still shows better accuracy than simpler models like the Simple Average.
- 6-Point Moving Average: RMSE = 28.52, Test RMSE = 14.57, Test MAPE = 20.84. It did not perform reasonably well and is not as accurate as the 2-Point or 4-Point models.
- 9-Point Moving Average: RMSE = 30.23, Test RMSE = 14.73, Test MAPE = 21.07. It is again not as good as the 2-Point model.
- Overall, the 2-Point Moving Average model stands out as the best-performing model so far among all with the lowest Test RMSE and Test MAPE values.
- However, there is still room for improvement in all the models to better capture the trend and seasonality and reduce errors in the forecasts.

## ➢ Simple Exponential Smoothing (SES)

**This method is suitable for forecasting data with no clear trend or seasonal pattern. It gives more weight to recent observations, which means that recent data points have a stronger influence on the forecast than older ones. This approach allows the model to capture short-term trends and adapt quickly to changes in the data.**

Parameter Alpha ($\alpha$) is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing

- **SES: Auto Fill Method**

The autofit model finds the most optimal parameters according to python while fitting on the train data.

Simple Exponential Smoothing optimal parameters: -

```
{'smoothing_level': 0.12362012817157356,
 '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}
```

Smoothing Level (alpha) = 0. 123

Initial Level = 112.0

**<u>Model 5: Single Exponential Smoothing (Auto-fit, alpha = 0.099)</u>**

SES, Alpha = 0.099

Model Output Visualised:

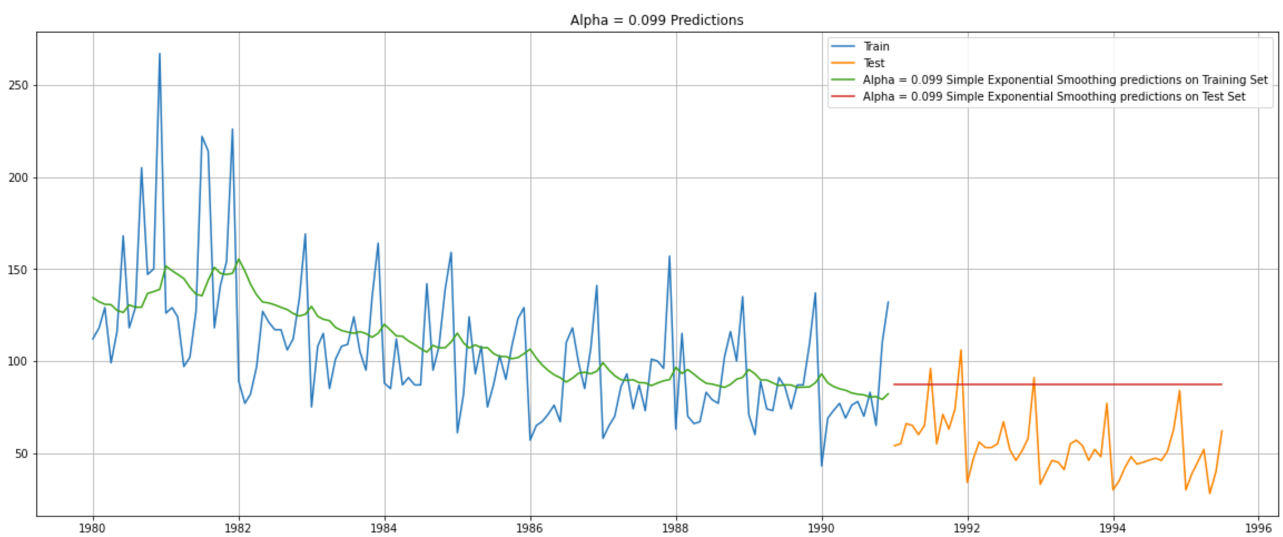



Fig: 2.18 SES Time Series Plot

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **31.501** | **22.73** |
| **Test Data** | **36.748** | **63.75** |

Table: 2.17 Model Performance Summary – Simple Exponential Smoothing Alpha = 0.099

Observation:

The level extrapolated by the SES model does not capture the trend (decline) or the seasonal variation of the time series.

- The Autofill Simple exponential smoothing model provides one-step-ahead forecast. It ignores both the trend and seasonality in the data.
- Test RMSE is 36.748 and MAPE is 63.75, indicating poor performance in capturing underlying patterns.
- The low smoothing parameter (0.099) implies a heavy reliance on past averages. This makes it less accurate compared to more sophisticated methods.

▪ **SES: Brute Force Method**

**The brute force model tests various smoothing parameter values to find the best ones for accurate test data forecasting. Below is the table for various parameters, sorted with least Test RMSE on top.**

**<u>Model 5a: Single Exponential Smoothing (using a Range of alpha values)</u>**

SES, Alpha ranging from 0.1 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and 0.2):

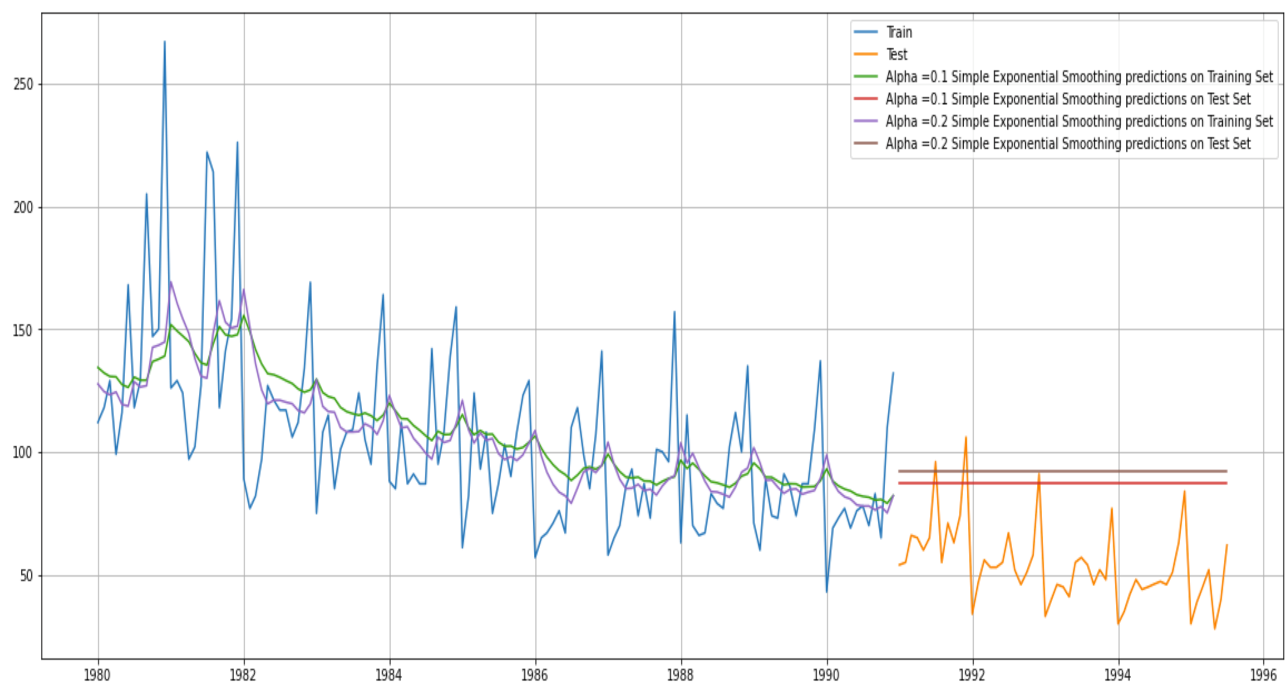



Fig: 2.19 Time Series Plot: Simple Exponential Smoothing Alpha = 0.1 & 0.2

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **Alpha = 0.1** | **36.780** | **63.81** |
| **Alpha = 0.2** | **41.314** | **72.07** |

Table: 2.18 Model Performance Summary – Simple Exponential Smoothing Alpha = 0.07

Observations:

The level extrapolated by the SES model does not capture the trend (decline) or the seasonal variation of the time series. Hence SES will not be a good fit for the dataset.

- The Brute Force simple exponential smoothing model generates one-step-ahead forecasts, overlooking both trend and seasonality in the time series.
- The model exhibits an RMSE of 36.46 and MAPE of 88.74, signifying its inability to effectively capture the underlying trend and seasonal patterns.
- The small smoothing parameter (0.07) indicates that the model relies heavily on past data averages rather than recent observations, making its accuracy comparable to the simple average model.

## ➢ Double Exponential Smoothing (DES)

**This method is applicable when data has Trend but no seasonality.**

- **DES: Auto Fill Method**

The autofit model finds the most optimal parameters according to python while fitting on the train data.

Double Exponential Smoothing optimal parameters: -

Smoothing Level (Alpha) = 0.16

Smoothing Trend (beta) = 0.16

**Model 6: Double Exponential Smoothing (Auto-fit, alpha = 0.16, beta = 0.16)**

DES, Alpha = 0.16, Beta = 0.16

```
{'smoothing_level': 0.162133240249321,
 'smoothing_trend': 0.1315215020415439,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 112.0,
 'initial_trend': 6.0,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Model Output Visualized:

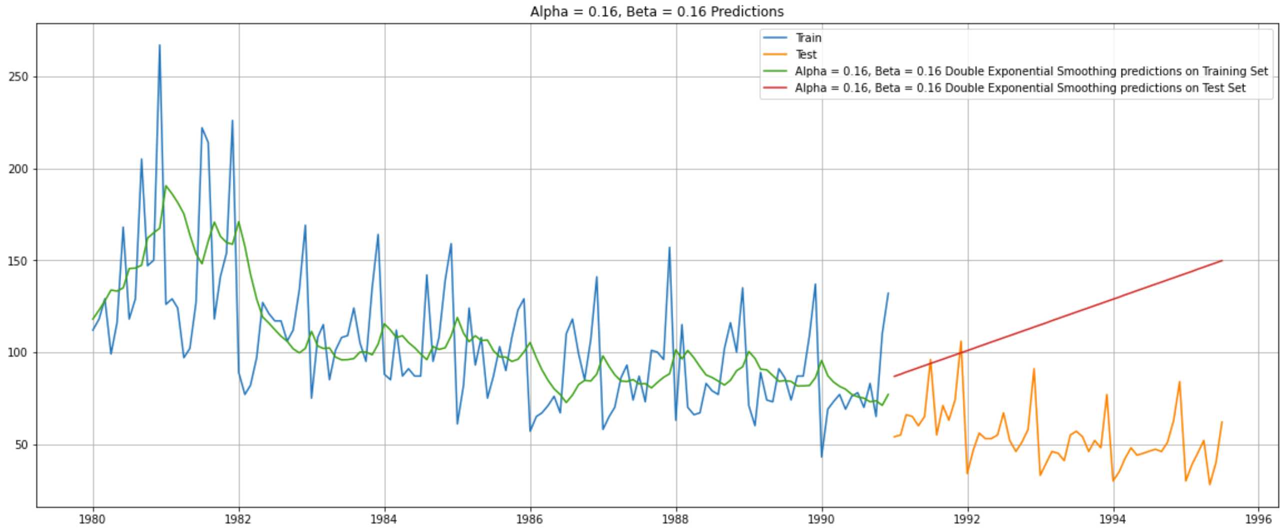



Fig: 2.20 Time Series Plot: Double Exponential Smoothing Alpha = 0.16, Beta = 0.16

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **35.059** | **23.85** |
| **Test Data** | **63.077** | **108.19** |

Table: 2.19 Model Performance Summary – Double Exponential Smoothing Alpha = 0.16, Beta = 0.16

Observation:

The Holt's model isn't a very good fit for this time series, as evidenced by the high RMSE. It is unable to capture the seasonality which is a strong component of the series. And in this case, it is unable to forecast the trend as well.

- The Autofill double exponential smoothing model forecasts the trend but disregards the seasonality of the time series.
- The RMSE is 63.077, and the MAPE is 108.19 for the Double exponential smoothing model. While the error is comparable to the Linear Regression Model
- Alpha & beta close to 0.16 means it has used the whole historical data to forecast the time series.

**DES: Brute Force Method**

The brute force model tests various smoothing parameter values to find the best ones for accurate test data forecasting. Below is the table for various parameters, sorted with least Test RMSE on top

**Model 6a: Double Exponential Smoothing (using a Range of alpha, beta values)**

DES, Alpha and Beta ranging from 0.1 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and beta = 0.1)

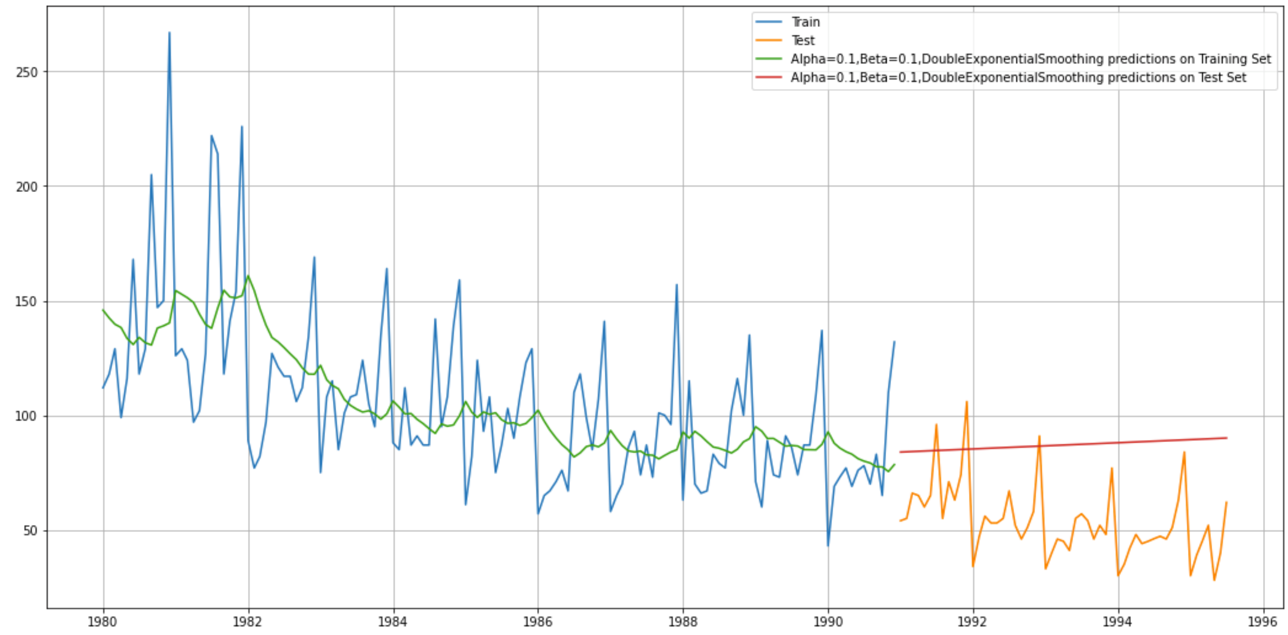



Fig: 2.21 Time Series Plot: Double Exponential Smoothing Alpha = 0.1 & Beta = 0.1

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **34.439** | **24.83** |
| **Test Data** | **36.96** | **63.86** |

**Table 2.2 Model Performance Summary – Double Exponential Smoothing Alpha = 0.1, Beta= 0.1**

Observation:

This model is an improvement over the earlier DES model. That said, it is unable to capture the seasonal variation which is a strong component of the series. And in this case, again, it is unable to capture the trend as well.

- **The Autofill double exponential smoothing model forecasts the trend but disregards the seasonality of the time series.**
- **The RMSE is 36.95, and the MAPE is 63.86 for the Double exponential smoothing model. While the error is comparable to the Linear Regression Model**
- **Alpha & beta to 0.1 means it has used the whole historical data to forecast the time series.**

## ➢ Triple Exponential Smoothing (TES)

- the triple exponential smoothing model is suitable for time series with both trend and seasonality.
- Since the trend variation is linear and the seasonal decomposition suggests a multiplicative time series, we use a triple exponential smoothing model with additive trend and multiplicative seasonality.

- **TES: Auto Fill Method**

The autofit model finds the most optimal parameters according to python while fitting on the train data.

Triple Exponential Smoothing optimal parameters: -

Smoothing Level (Alpha) = 0.071

Smoothing Trend (Beta) = 0.045

Smoothing Seasonal (Gamma) = 0.

```
{'smoothing_level': 0.07130285749243212,
 'smoothing_trend': 0.04550837652110988,
 'smoothing_seasonal': 8.385716703273524e-05,
 'damping_trend': nan,
 'initial_level': 163.60092654560762,
 'initial_trend': -0.9804841883026134,
 'initial_seasons': array([0.68714163, 0.77936108, 0.85184662, 0.74446365, 0.8372947 ,
        0.91182237, 1.00282327, 1.06745268, 1.01025249, 0.98957378,
        1.1535151 , 1.59037115]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

**<u>Model 7: Triple Exponential Smoothing (Auto-fit: Alpha=0.15, Beta=0, Gamma=0.37)</u>**

TES, Alpha=0.11, Beta=0.05, Gamma=0

Model Output:

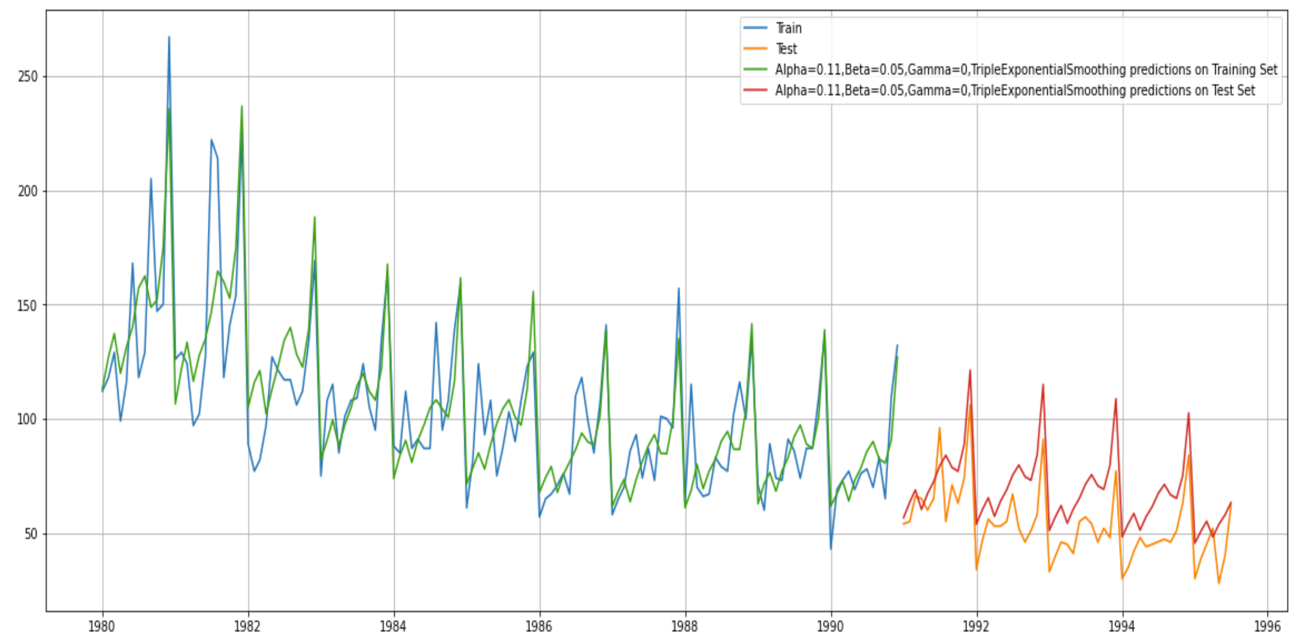



Fig: 2.22 Time Series Plot: Triple Exponential Smoothing Alpha = 0.15, Beta= 0.05, Gamma= 0.

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **18.406** | **12.58** |
| **Test Data** | **20.229** | **33.76** |

**Table 2.28 Model Performance Summary – Triple Exponential Smoothing Alpha = 0.15, Beta= 0.05, Gamma= 0**

Observation:

This Holt-Winters' model or the Triple Exponential Model is a better fit for this time series, as it has both a strong Trend and a strong Seasonal component. This is corroborated by a low RMSE.

- The Autofill Triple Exponential Smoothing model captures both the trend and seasonality of the time series.
- The test RMSE is 20.229, and the MAPE is 33.76. However, the error is higher than Double exponential Smoothening or Linear Regression model
- The values of Alpha, Beta and Gamma being close to 0.15 and 0.05 and 0, indicate that the forecast relies heavily on historical data to build the model giving a smoother prediction.

**TES: Brute Force Method**

The brute force model tests various smoothing parameter values to find the best ones for accurate test data forecasting. Below is the table for various parameters, sorted with least Test RMSE on top.

| | Alpha Values | Beta Values | Gamma Values | Train RMSE | Train MAPE | Test RMSE | Test MAPE |
|---|---|---|---|---|---|---|---|
| **2** | 0.1 | 0.0 | 0.3 | 20.143621 | 14.25 | 9.180148 | 12.80 |
| **1** | 0.1 | 0.0 | 0.2 | 19.562899 | 13.94 | 9.120683 | 12.93 |
| **3** | 0.1 | 0.0 | 0.4 | 20.779099 | 14.42 | 9.504917 | 12.96 |
| **273** | 0.3 | 0.5 | 0.4 | 25.873405 | 17.54 | 10.361217 | 13.21 |
| **20** | 0.1 | 0.2 | 0.1 | 19.651464 | 14.31 | 9.190042 | 13.25 |

Table: 2.29 Brute Force TES Parameters

Since Alpha = 0.1, Beta = 0.2, Gamma = 0.2 yield the least test RMSE, indicating the best fit for our test data, we select them to build our Triple Exponential Smoothing model.

**<u>Model 7a: Triple Exponential Smoothing (using a Range of alpha, beta, gamma values)</u>**

TES, Alpha, Beta, Gamma ranging from 0 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and beta = 0.2, gamma = 0.2)

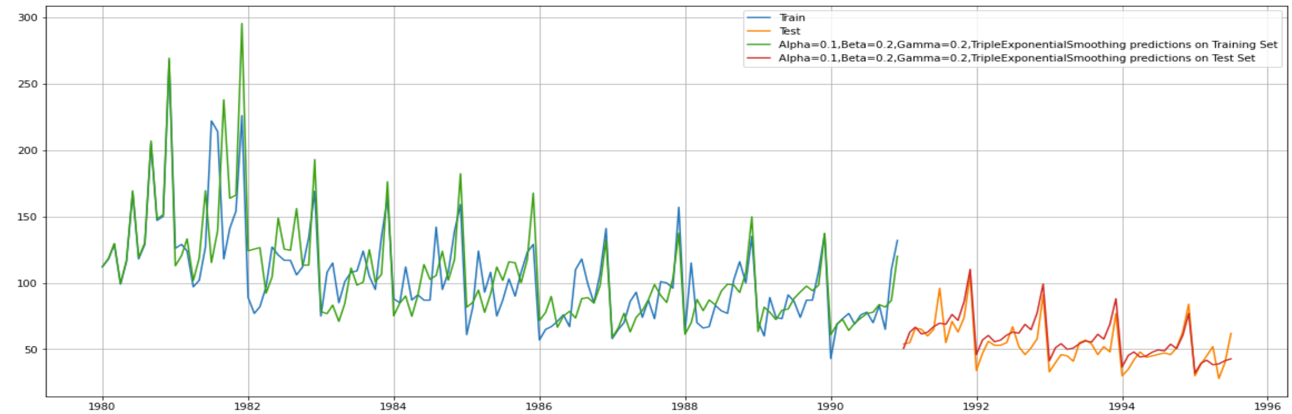



Fig: 2.23 Time Series Plot: Triple Exponential Smoothing Alpha = 0.1, Beta = 0.2, Gamma = 0.2

Performance Metrics:

| | RMSE | MAPE |
|---|---|---|
| **Training Data** | **19.562** | **13.94** |
| **Test Data** | **9.50** | **13.74** |

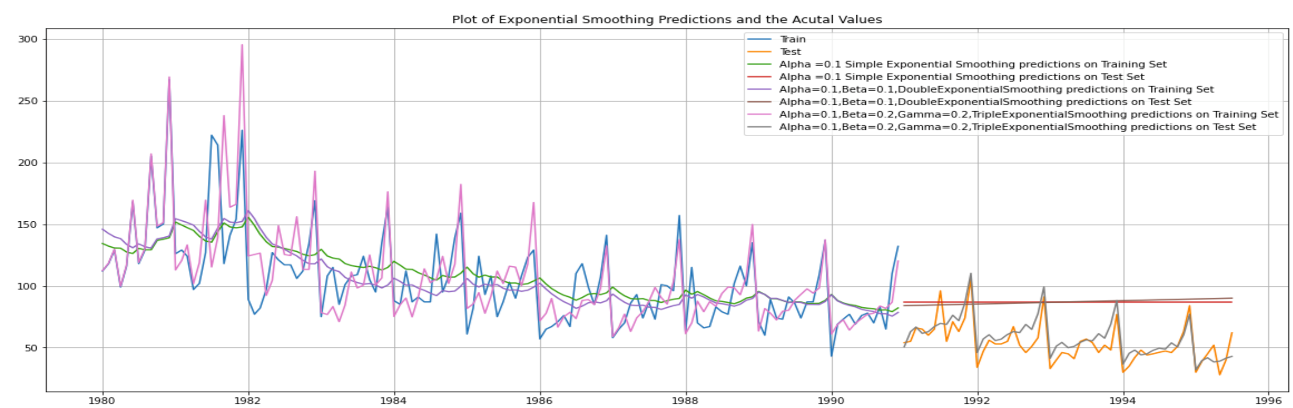
**Table 2.30 Model Performance Summary – Triple Exponential Smoothing Alpha = 0.1, Beta = 0.2, Gamma = 0.2**

Observation:

This Triple Exponential Smoothing model is an improvement on the earlier auto-fit TES model. This, in fact, is the best performing model thus far, as evidenced by the low RMSE. This Holt-Winters' model or the TES Model is a good fit for this time series, as it has both a strong Trend and a strong Seasonal component.

- The Brute Force triple exponential smoothing models show the best accuracy among all models evaluated, with the lowest Root Mean Square Error (RMSE) of 9.50 and Mean Absolute Percentage Error (MAPE) of 13.74
- This is followed by the 2-point moving average model, capturing both trend and seasonality in the time series.
- The rest of the other models are also not suitable for prediction as they do not capture both the Trend & Seasonality well required for the time series.

**A Consolidated Plot of all the Exponential Models built:**



**The Triple Exponential Models appear to be the strongest fit for this time series.**
**Force Triple Exponential smoothing model captures both the trend and seasonality.**

- With RMSE of 9.50 and MAPE of 13.74, this model exhibits the best accuracy among all the evaluated models so far.
- The values of alpha and beta being close to 0 imply that the model heavily relies on historical data to make forecasts.

**2.5 Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.**

**Note: Stationarity should be checked at alpha = 0.05.**

The **Augmented Dickey Fuller (ADF) Test** is a statistical test for affirming whether or not a time series is Stationary.
The Null Hypothesis H0 is: Time Series is non-stationary
The Alternative Hypothesis H1 is: Time Series is Stationary
**TEST 1**: We administer the ADF test on the Original Time Series.

Results of Dickey-Fuller Test: Test Statistic: -2.164250 p-value: 0.219476
With the resultant ADF test p-value at 0.21, we cannot reject the Null Hypothesis (at alpha 0.05).

We hence conclude that the Time Series is non-stationary.
In order to make a Time Series Stationary, we need to transform the original series by taking a Difference of the original values. Usually, a 1 period difference suffices to transform a non-Stationary series into a Stationary one.
**TEST 2:** We administer the ADF test on the new series - derived by taking a 1 period Difference of the original series.

Results of Dickey-Fuller Test: Test Statistic: -
6.592372e+00 p-value: 7.061944e-09

The resultant ADF p-value (0.000000007) is significantly less than 0.05 (alpha).
We can hence reject the null hypothesis in the case of the new series, which is derived by differencing the original series over 1 period.

We conclude that at Difference 1, the time series is Stationary.

## 2.6 Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

**ARIMA: -** Auto Regressive Integrated Moving Average is a way of modeling time series data for forecasting or predicting future data points. Improving AR Models by making Time Series stationary through Moving Average Forecasts

ARIMA models consist of 3 components: -

AR model: The data is modelled based on past observations.

Integrated component: Whether the data needs to be differenced/transformed.

MA model: Previous forecast errors are incorporated into the model.

**The best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).**

- ARIMA Model building to estimate best 'p', 'd', 'q' parameters (Lowest AIC Approach)

| | param | AIC |
|---|---|---|
| 13 | (2, 1, 3) | 1274.695121 |
| 23 | (4, 1, 3) | 1278.451407 |
| 18 | (3, 1, 3) | 1278.652189 |
| 14 | (2, 1, 4) | 1278.771124 |
| 9 | (1, 1, 4) | 1279.605263 |
| 2 | (0, 1, 2) | 1279.671529 |
| 7 | (1, 1, 2) | 1279.870723 |
| 3 | (0, 1, 3) | 1280.545376 |
| 6 | (1, 1, 1) | 1280.574230 |
| 11 | (2, 1, 1) | 1281.507862 |
| 4 | (0, 1, 4) | 1281.676698 |
| 12 | (2, 1, 2) | 1281.870722 |

**Table: 2.31 Arima AIC Parameters**

**Model 8: ARIMA (Lowest AIC model parameters: p=3, d=1, q=3)**

Of the ARIMA models generated using various combinations of parameters p and q, the model with the lowest AIC score was: ARIMA (2, 1, 3) with an AIC score of **1274.695**

Model Summary:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 11 Aug 2024   AIC                           1274.695
Time:                        12:58:13   BIC                           1291.946
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6778      0.084    -20.041      0.000      -1.842      -1.514
ar.L2         -0.7286      0.084     -8.702      0.000      -0.893      -0.565
ma.L1          1.0444      0.631      1.654      0.098      -0.193       2.282
ma.L2         -0.7722      0.133     -5.827      0.000      -1.032      -0.512
ma.L3         -0.9048      0.573     -1.580      0.114      -2.027       0.218
sigma2       859.2985    531.404      1.617      0.106    -182.234    1900.831
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.47
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Table: 2.32 Auto ARIMA Model SUMMARY

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **ARIMA (2,1,3)** | **36.88** | **73.89** |

Observation: ARIMA does not factor the seasonal component which is an important characteristic of this time series, and hence wouldn't be an ideal model for the series.

- The Auto ARIMA model aims to capture the underlying trend in the data but does not consider the seasonality component.
- The model's performance is evaluated with a Root Mean Square Error of 36.88 and a Mean Absolute Percentage Error of 73.89. The model performed poorer on Test data as compared to Training

## ➢ Automated version of a SARIMA model

The ARIMA models can be extended/improved to handle seasonal components of a data series.

The seasonal autoregressive moving average model is given by SARIMA (p, d, q)(P, D, Q)F

The above model consists of:

- **Autoregressive** and **moving average** components (p, q)
- Seasonal autoregressive and moving average components (P, Q)
- The ordinary and seasonal difference components of order 'd' and 'D' ▫

  Seasonal frequency 'F'

The value for the parameters (p,d,q) and (P, D, Q) can be decided by comparing different values for each and taking the lowest AIC value for the model build. The value for F can be consolidated by ACF plot

- Without Seasonal Differencing (D = 0):

Let us look at the differenced ACF plot again to understand the seasonal parameter for the SARIMA model

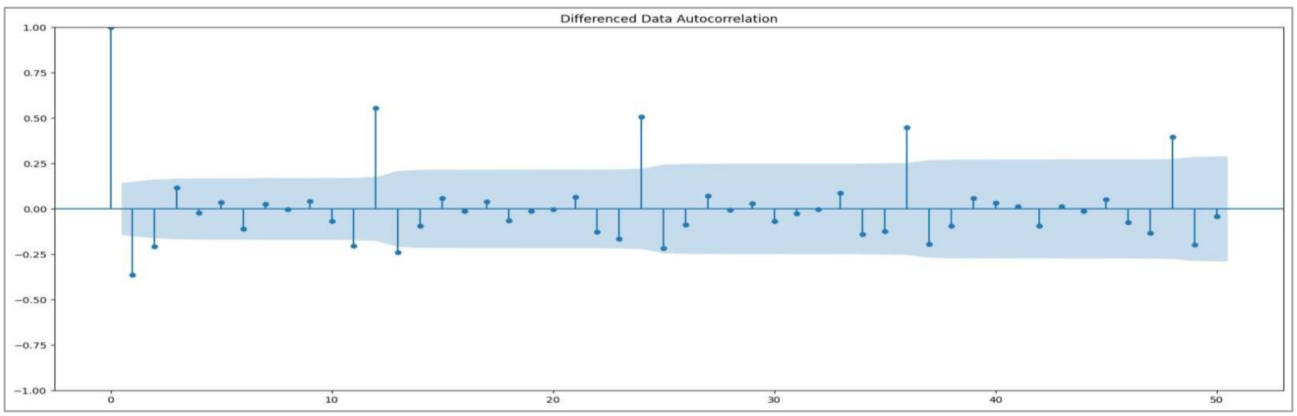



**Fig2.35 Differenced Autocorrelation Plot: SARIMA**

- S=12 is chosen for seasonal differencing as it is significant, and the ACF plot at S=12 does not taper off.
- This indicates the presence of seasonality, and applying seasonal differencing to the original series can improve the model's performance.
- d = 1 to make the time series stationary
- Seasonal differencing not yet applied to make the time series stationary D = 0

**<u>Model 9: SARIMA (Lowest AIC parameters: p=3, d=1, q=1, P=3, D=1, Q=1)</u>**

Of the SARIMA models generated using various combinations of parameters p, q, P, Q and D, the model with the lowest AIC score was: SARIMA (3, 1, 1) x (3, 1, 1, 12) with an AIC score of 681.362807

Model Summary:

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                    y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 1, 1, 12)   Log Likelihood                -331.681
Date:                            Thu, 13 Aug 2020   AIC                            681.363
Time:                                    00:05:15   BIC                            702.801
Sample:                                         0   HQIC                           689.958
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0173      0.151      0.114      0.909      -0.279       0.314
ar.L2         -0.0426      0.141     -0.302      0.763      -0.319       0.234
ar.L3         -0.0574      0.119     -0.484      0.628      -0.290       0.175
ma.L1         -0.9388      0.085    -11.104      0.000      -1.105      -0.773
ar.S.L12       0.0908      0.126      0.721      0.471      -0.156       0.337
ar.S.L24      -0.0437      0.108     -0.406      0.685      -0.254       0.167
ar.S.L36   -3.592e-05      0.053     -0.001      0.999      -0.104       0.103
ma.S.L12      -0.9998    237.117     -0.004      0.997    -465.741     463.741
sigma2       185.4128    4.4e+04      0.004      0.997    -8.6e+04    8.63e+04
===================================================================================
Ljung-Box (Q):                       42.97   Jarque-Bera (JB):                 2.56
Prob(Q):                              0.35   Prob(JB):                         0.28
Heteroskedasticity (H):               0.56   Skew:                             0.42
Prob(H) (two-sided):                  0.13   Kurtosis:                         3.22
===================================================================================
```

Table:2.33 SARIMA AIC Parameters

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **SARIMA (3,1,1) (3,1,1,12)** | **16.85** | **25.54** |

Observation:

A SARIMA model is generally better equipped for a time series with a strong seasonal component. However this model fails to improve on the ARIMA model score.

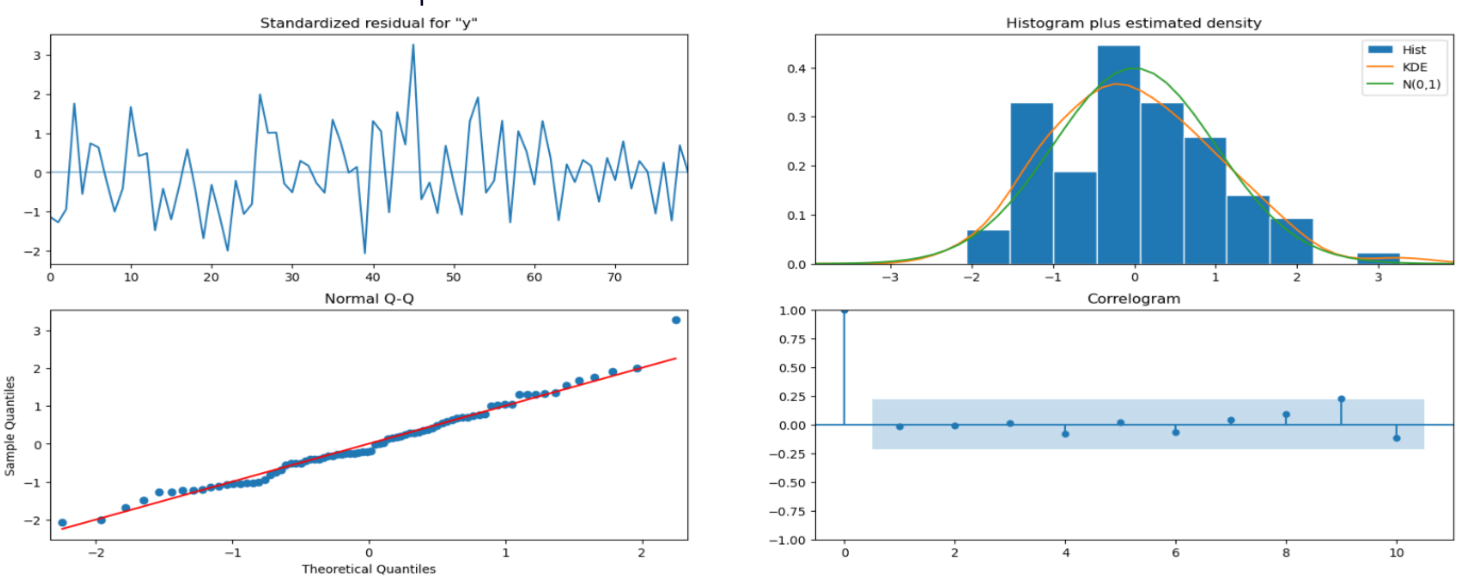



Fig: 2.25 Diagnostic plot Automated SARIMA

- The Automated SARIMA (3, 1, 1)(3, 1, 1, 12) model successfully captures both the trend and seasonality in the data.
- The Root Mean Square Error is 16.85, and the Mean Absolute Percentage Error is 25.54 for the automated SARIMA model with seasonal differencing.
- This model performs better than the model without seasonal differencing, indicating that incorporating seasonal differencing improves the accuracy of the forecast.

**2.7 Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**

From the ADF test conducted in an earlier section, we know that the original time series is Non-stationary.
So the first step would be to transform the time series by differencing the original series over 1 period, and make it Stationary.
We then need to plot the ACF and PACF on the transformed stationary Time Series.
The following are the ACF and PACF plots:

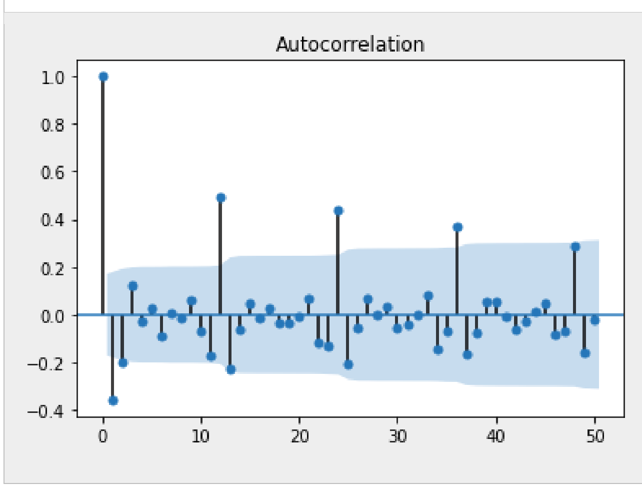
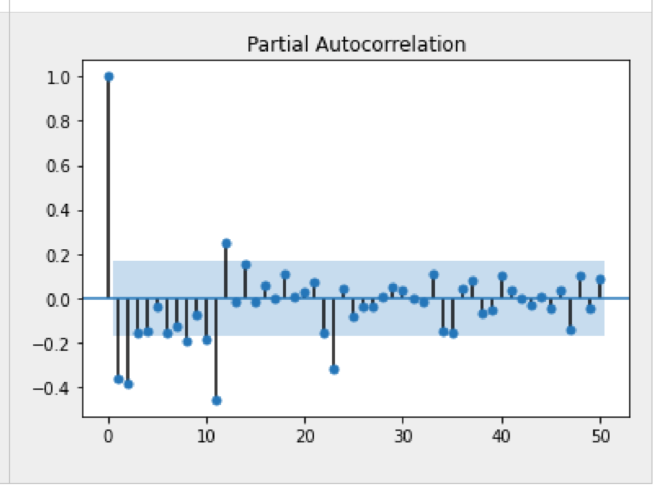
| **ACF Plot** | **PACF Plot** |
|---|---|
|  |  |



Fig: 2.26

The ACF plot:

- The cut-off is right after lag 2. It appears lag 2 is marginally above the border. Hence a possible value of q is 2.
- Hence from the ACF plot, we can assign 2 as the order of the MA component (q) of the ARIMA/SARIMA model.

The PACF plot:

- The cut-off here too appears right after lag 2
- Lag 3 appears to be insignificant, though only marginally.

- Hence from the PACF plot, we can assign 2 as the order of the AR component (p) of the ARIMA/SARIMA model.

We know that the Seasonal component is apparent in the series. The ACF plot also clearly shows a pattern repeating every year - indicating strong seasonal behavior.
So a SARIMA model would be appropriate for modelling such a series. For employing the seasonal component, we will accord value to a 12-period lag, which will map to a value of 1 to both P and Q. The value of d will be 1, as the series has been Differenced by 1 period, in order to make it stationary.
Hence based on the ACF and PACF plots, we can develop a SARIMA (2,1,2)x(1,1,1, 12) model.

**Model 9a: SARIMA (ACF, PACF plot parameters: p=2, d=1, q=2, P=1, D=1, Q=1)**

Model Summary:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(1, 1, [1], 12)   Log Likelihood                -450.847
Date:                            Fri, 14 Aug 2020   AIC                            915.693
Time:                                    23:26:33   BIC                            934.204
Sample:                                         0   HQIC                           923.192
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.1035      0.133      8.288      0.000       0.843       1.364
ar.L2         -0.3436      0.109     -3.150      0.002      -0.557      -0.130
ma.L1         -1.8152      0.105    -17.291      0.000      -2.021      -1.609
ma.L2          0.8668      0.093      9.276      0.000       0.684       1.050
ar.S.L12      -0.3877      0.069     -5.625      0.000      -0.523      -0.253
ma.S.L12      -0.0788      0.130     -0.606      0.545      -0.334       0.176
sigma2       338.2613     53.797      6.288      0.000     232.821     443.701
===================================================================================
Ljung-Box (Q):                       25.41   Jarque-Bera (JB):                 0.03
Prob(Q):                              0.96   Prob(JB):                         0.98
Heteroskedasticity (H):               0.66   Skew:                             0.04
Prob(H) (two-sided):                  0.22   Kurtosis:                         2.97
===================================================================================
```

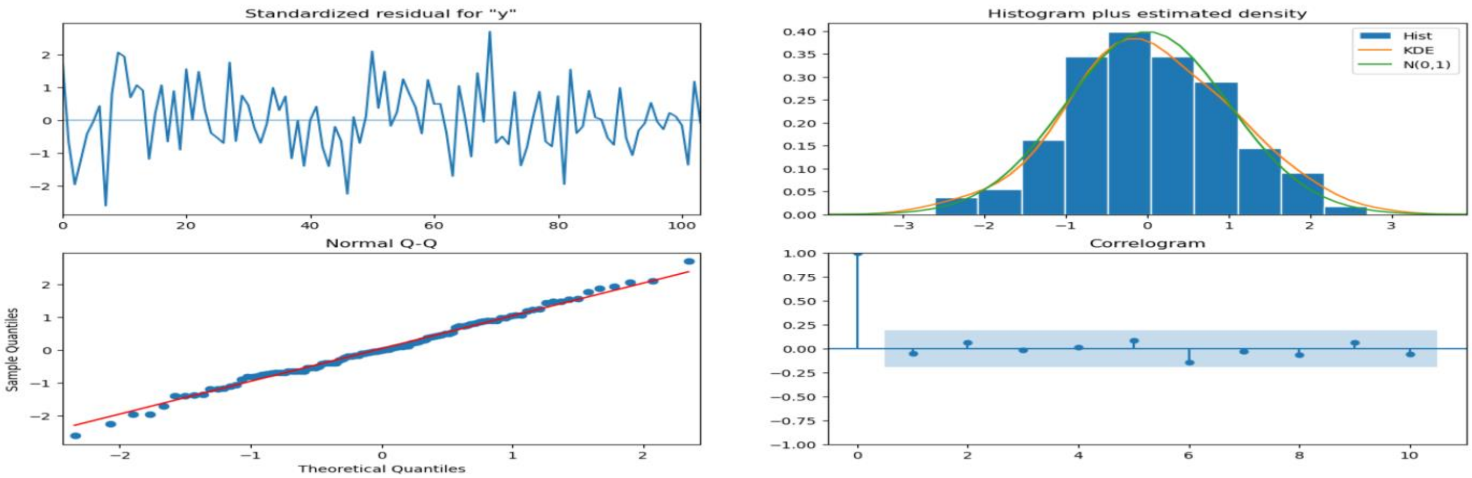
Table: 2.34 Sarimax Results

Performance Metrics on Test Data:

| | RMSE | MAPE |
|---|---|---|
| **SARIMA (2,1,2) (1,1,1,12)** | **13.275** | **17.53** |

Table: 2.35 Performance Matrix

Observation:
This SARIMA model appears to be an improvement over the last SARIMA model, going by the RMSE on the test data. The model is the second best model that we've developed.




| Model | Test RMSE | Test MAPE |
|---|---|---|
| **Automated SARIMA (2, 1, 2)(1, 1, 1, 12)** | **13.27** | **13.49** |

**Table 2.36 Model Performance Summary – Automated SARIMA (2, 1, 2)(1, 1, 1, 12)**

- **The Automated SARIMA (2, 1, 2)(1, 1, 1, 12) model successfully captures both the trend and seasonality in the data.**
- **The Root Mean Square Error is 13.27, and the Mean Absolute Percentage Error is 13.49 for the automated SARIMA model with seasonal differencing. o This model performs better than the model without seasonal differencing, indicating that incorporating seasonal differencing improves the accuracy of the forecast.**

| | Test RMSE | Test MAPE |
|---|---|---|
| Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing | 9.507991 | 13.74 |
| 2pointTrailingMovingAverage | 11.529901 | 13.57 |
| SARIMA(2,1,2)(1,1,1,12) | 13.270837 | 17.49 |
| 4pointTrailingMovingAverage | 14.458344 | 19.52 |
| 6pointTrailingMovingAverage | 14.573943 | 20.84 |
| 9pointTrailingMovingAverage | 14.733243 | 21.07 |
| RegressionOnTime | 15.279105 | 22.88 |
| SARIMA(3,1,1)(3,1,1,12) | 16.855496 | 25.54 |
| Alpha=0.15,Beta=0,Gamma=0.37,TripleExponentialSmoothing | 20.229428 | 33.76 |
| Alpha=0.1,SimpleExponentialSmoothing | 36.858463 | 64.02 |
| ARIMA(2,1,3) | 36.885626 | 73.89 |
| Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing | 36.954836 | 63.86 |
| Alpha=0.099,SimpleExponentialSmoothing | 37.622621 | 65.42 |
| SimpleAverageModel | 53.490397 | 95.03 |
| Alpha=0.16,Beta=0.16,DoubleExponentialSmoothing | 63.077144 | 108.19 |
| NaiveModel | 79.747814 | 145.23 |

Table: 2.36 Model Performance Summary – Consolidated

- **Model Performance: After evaluating various forecasting models, the top-performing ones are:**
  - Triple Exponential Smoothing Model (Alpha = 0.1, Beta = 0.2, Gamma = 0.2): This model exhibits the highest accuracy, with a Train RMSE of 21.69, Test RMSE of 9.50, and MAPE of 13.7.
  - 2-Point Trailing Moving Average: This model performs well, with a Train RMSE of 19.67, Test RMSE of 11.52, and MAPE of 13.57.
  - Automated SARIMA(2, 1, 2)(1, 1, 1, 12): This model shows reasonable accuracy, with a Train RMSE of 39.18, Test RMSE of 13.52, and MAPE of 17.49.
  - Triple Exponential Smoothing Model (Alpha = 0.111, Beta = 0.05, Gamma = 0.362): Another triple exponential model with a slightly lower performance, having a Train RMSE of 18.41, Test RMSE of 20.15, and MAPE of 33.37.

## 2.8 Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

From our comparison between models, the top 2 models are:

1. Triple Exponential Smoothing, Alpha=0.1,Beta=0.2,Gamma=0.2
2. SARIMA (2,1,2) x (1,1,1,12)

We will re-build these 2 models on the complete data, and make 12 month forecasts.

**Part A: TES model on complete data**

Building a Triple Exponential Smoothing (alpha = 0.1, beta = 0.2, gamma = 0.2) model on the complete data, and forecasting for the next 12 months

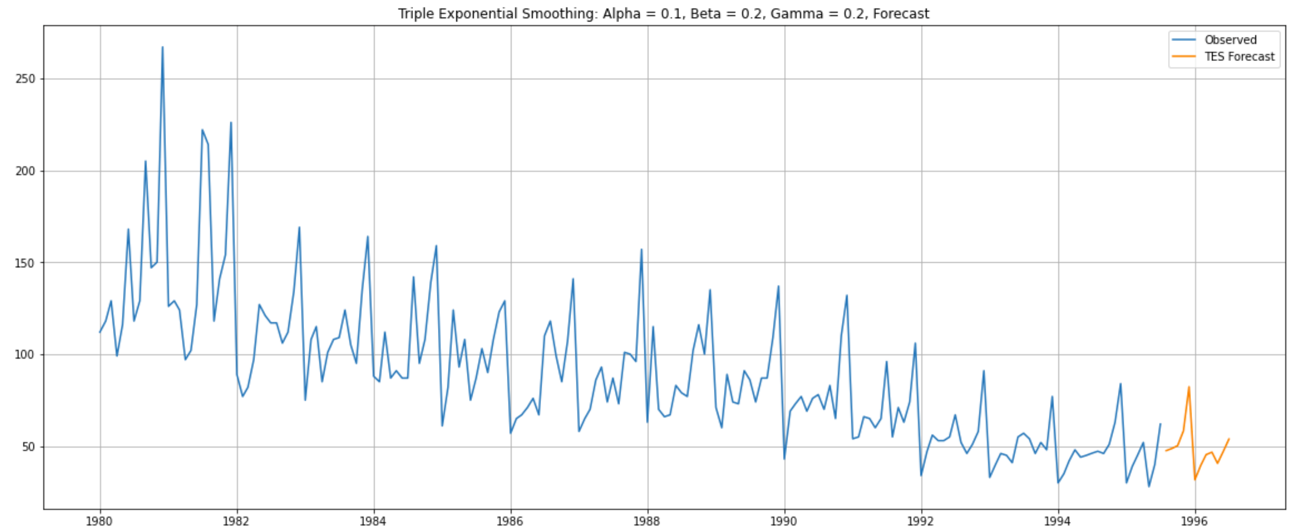
Model Summary:

```
                    ExponentialSmoothing Model Results
================================================================================
Dep. Variable:                     Rose   No. Observations:                  187
Model:             ExponentialSmoothing   SSE                          56655.762
Optimized:                         True   AIC                           1100.451
Trend:                         Additive   BIC                           1152.148
Seasonal:                Multiplicative   AICC                          1104.522
Seasonal Periods:                    12   Date:                 Sun, 11 Aug 2024
Box-Cox:                          False   Time:                         13:39:12
Box-Cox Coeff.:                    None
=================================================================================
                          coeff                 code              optimized
---------------------------------------------------------------------------------
smoothing_level               0.1000000                alpha                False
smoothing_trend               0.2000000                 beta                False
smoothing_seasonal            0.2000000                gamma                False
initial_level                 138.60653                  l.0                 True
initial_trend                 1.4125891                  b.0                 True
initial_seasons.0             0.8097786                  s.0                 True
initial_seasons.1             0.8568749                  s.1                 True
initial_seasons.2             0.9232260                  s.2                 True
initial_seasons.3             0.7906197                  s.3                 True
initial_seasons.4             0.8920215                  s.4                 True
initial_seasons.5             1.0334076                  s.5                 True
initial_seasons.6             1.1554022                  s.6                 True
initial_seasons.7             1.2275389                  s.7                 True
initial_seasons.8             1.1002290                  s.8                 True
initial_seasons.9             1.0723883                  s.9                 True
```

**TES (alpha = 0.1, beta = 0.2, gamma = 0.2) Model 12-month Forecast:**

| Timeline | Rose Sales Forecast |
|---|---|
| 1995-08-01 | 47.552982 |
| 1995-09-01 | 48.746129 |
| 1995-10-01 | 50.277107 |

| | |
|---|---|
| 1995-11-01 | 58.269327 |
| 1995-12-01 | 82.302948 |
| 1996-01-01 | 31.700169 |
| 1996-02-01 | 39.432145 |
| 1996-03-01 | 45.380415 |
| 1996-04-01 | 46.71215 |
| 1996-05-01 | 40.645794 |
| 1996-06-01 | 47.20148 |
| 1996-07-01 | 53.867965 |




## Part B: SARIMA model on complete data

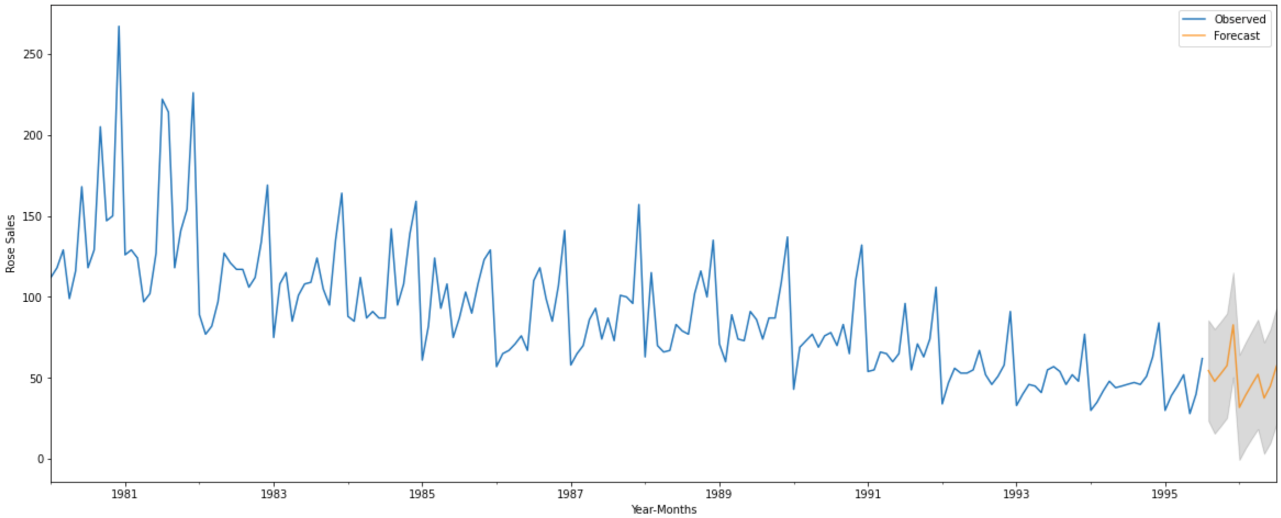
**Building a SARIMAX (2,1,2)x(1,1,1,12) model on the complete data, and forecasting for the next 12 months Model Summary:**

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                    y   No. Observations:               187
Model:             SARIMAX(2, 1, 2)x(1, 1, [1], 12)   Log Likelihood             -665.294
Date:                              Fri, 14 Aug 2020   AIC                        1344.588
Time:                                      23:45:34   BIC                        1366.070
Sample:                                           0   HQIC                       1353.312
                                              - 187
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.1081      0.093     11.951      0.000       0.926       1.290
ar.L2         -0.3257      0.080     -4.089      0.000      -0.482      -0.170
ma.L1         -1.8286      0.067    -27.351      0.000      -1.960      -1.698
ma.L2          0.8792      0.059     14.789      0.000       0.763       0.996
ar.S.L12      -0.3830      0.049     -7.757      0.000      -0.480      -0.286
ma.S.L12      -0.0828      0.093     -0.888      0.374      -0.265       0.100
sigma2       250.8891     26.925      9.318      0.000     198.118     303.660
===================================================================================
Ljung-Box (Q):                       35.53   Jarque-Bera (JB):                 3.01
Prob(Q):                              0.67   Prob(JB):                         0.22
Heteroskedasticity (H):               0.22   Skew:                             0.04
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.67
===================================================================================
```

**SARIMA (2,1,2)x(1,1,1,12) model 12-month Forecast:**

| timeline | mean | mean_se | mean_ci_lower | mean_ci_upper |
|---|---|---|---|---|
| 1995-08-01 | 54.545303 | 15.839478 | 23.500497 | 85.59011 |
| 1995-09-01 | 47.917716 | 16.446711 | 15.682754 | 80.152677 |
| 1995-10-01 | 52.623395 | 16.455915 | 20.370395 | 84.876396 |
| 1995-11-01 | 57.646815 | 16.455942 | 25.393762 | 89.899869 |
| 1995-12-01 | 82.776318 | 16.466521 | 50.50253 | 115.050107 |
| 1996-01-01 | 31.838413 | 16.531662 | -0.563049 | 64.239874 |
| 1996-02-01 | 39.28762 | 16.682039 | 6.591425 | 71.983816 |
| 1996-03-01 | 45.878564 | 16.915626 | 12.724547 | 79.032581 |
| 1996-04-01 | 52.236236 | 17.21173 | 18.501865 | 85.970606 |

| | | | | |
|---|---|---|---|---|
| **1996-05-01** | **37.623006** | **17.546784** | **3.231942** | **72.01407** |
| **1996-06-01** | **44.997111** | **17.902037** | **9.909764** | **80.084459** |
| **1996-07-01** | **57.374141** | **18.265042** | **21.575316** | **93.172966** |




## 2.10: Insights and Findings

- Looking at the pattern in the time series, one could infer that Rose wines have been going out of fashion for a while now. The time series shows a consistent decline in Sales of Rose.
- Perhaps there are other alternatives that are preferred in current times, in which case it may be worthwhile to focus more on the other festive wines in demand.
- But the last couple of years also shows a steadying of the decline. And in year 1995, the Sales in the first 2 quarters were surprisingly strong. This could mean that the interest in Rose might still be rekindled.
- The models appear to forecast a range and pattern for the next 12 months, which are similar to that of the past year. And given the low RMSE score, along with a certain consistency of past behaviour, the forecast looks dependable.

- The consumption pattern points to the fact that Rose Wines are most in demand in holiday and festive seasons. This brings them in competition with Sparkling Wines and other premium products.
- Perhaps Rose needs to be positioned differently, likely a level below premium. And then it can be promoted accordingly. The focus for Rose need not be the holiday season, because that is a very competitive space. A niche needs to be carved for Rose for sales and interest to pick up again.

➢ **Measures for Future Sales:**

**A critical decision must be taken to either discontinue the Rose wine or undertake product and process enhancements to boost sales. To improve future sales, ABC Estate Wines should consider the following strategies:**

**o Customer Engagement: Strengthen customer relationships through personalized offers, loyalty programs, and active engagement to foster repeat business.**

- **Marketing promotions, sponsoring small & large events having target audience**
- **Advertisements strategies such as launching non-alcoholic beverages with the same brand name to again popularity**
- **Capitalize on Seasonal Trends: Plan production and marketing efforts to meet increased demand during holiday seasons, especially in November and December.**
- **Inventory Management: Implement effective inventory management to avoid stockouts during peak periods and minimize excess inventory during slower periods.**
- **Pricing Strategy: Utilize dynamic pricing to adjust prices during peak and off-peak periods, attracting more customers and optimizing revenue.**
- **Competitor benchmarking: Figure out competitor businesses for similar wines to understand the demand of rose wine in the market and make necessary changes.**