Time Series Forecasting – ROSE Wine
Coded Project

DSBA – Course

Business Report

Created by – Rishabh Gupta

Foreword

Context –

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties. Our objective is to delve into the data, analyze trends, patterns, and factors influencing wine sales over the course of the century. By leveraging data analytics and forecasting techniques, we aim to gain actionable insights that can inform strategic decision-making and optimize sales strategies for the future.

Objective -

The primary objective of this project is to analyze and forecast wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

Contents

List of Tables

List of Figures

Objective

The primary objective of this project is to analyse and forecast wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

For this assignment, we will analyze data on different types of wine sales from the 20th century. Both datasets come from the same company but represent different wine varieties. As an analyst at ABC Estate Wines, our task is to analyze and forecast wine sales during this period.

We are going to individually perform the following tasks on each of the two datasets.

Data Analysis of problem statement –

The dataset contains data corresponding to polished and unpolished stones.

Sheet name –

1. *Sparkling.csv*
2. *Rose.csv*

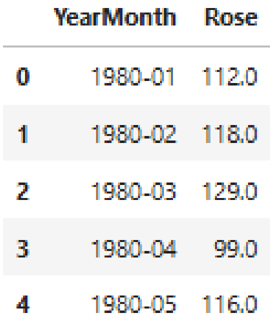
1.Rose Wines Sales

Data Dictionary –

1. **YearMonth:** displays time for sales
2. **Rose:** No of wines sales of this type

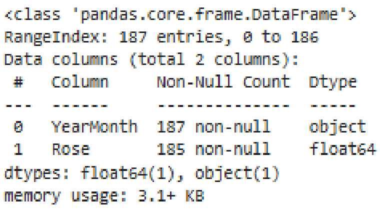
Data Overview –

To start the data analysis, we need to import the necessary libraries, specify the working directory, and load the dataset. Then, we will view the first five rows using head () function to get a preliminary understanding of the data. The Dataset has 187 number of rows with 2 columns.



	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

Table 1 - TOP 5 ROWS OF DATASET



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   YearMonth  187 non-null    object 
 1   Rose       185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.1+ KB
```

Table 2 - Basic Info. of dataset

Regarding the datatype of the columns in the dataset, we can confirm that-

• We can observe that there around 1 numerical datatype and 1 object

Data Pre-processing-

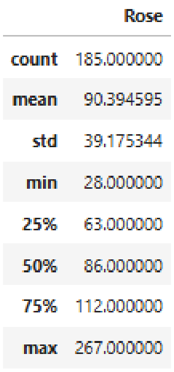
Missing value treatment and Analysis-

- On analysis, we can observe there are missing values. As this is time series, we need to impute them with linear interpolation.

Please refer code.

Statistical Summary –

Using Describe () function, we can analyses the summary statistics of the dataset –



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

Table 3 - Statistical summary of dataset

Observations-

- **Count (185.000000):** There are 185 data points in the "Rose" column. This indicates the sample size.
- **Mean (90.394595):** The average value of the "Rose" data is approximately 90.4. This represents the central tendency of the data.

- **Standard Deviation (std) (39.175344):** The standard deviation is about 39.18. This measures the spread or dispersion of the data around the mean. A higher standard deviation indicates greater variability.
- **Minimum (min) (28.000000):** The smallest value in the "Rose" column is 28.
- **25th Percentile (25%) (63.000000):** 25% of the data points are less than or equal to 63.
- **50th Percentile (50%) (86.000000):** 50% of the data points are less than or equal to 86. This is also the median of the data.
- **75th Percentile (75%) (112.000000):** 75% of the data points are less than or equal to 112.
- **Maximum (max) (267.000000):** The largest value in the "Rose" column is 267.

Overall Observations:

- **Distribution:** The data appears to be somewhat skewed to the right, as the mean (90.4) is higher than the median (86), and the maximum value (267) is significantly larger than the 75th percentile (112).
- **Range:** The data has a considerable range, from 28 to 267, indicating substantial variability.
- **Potential Outliers:** The maximum value of 267 is quite far from the 75th percentile (112), suggesting a possible outlier.

Exploratory Data Analysis

First lets plot the Time Series to understand the behaviour of the data.

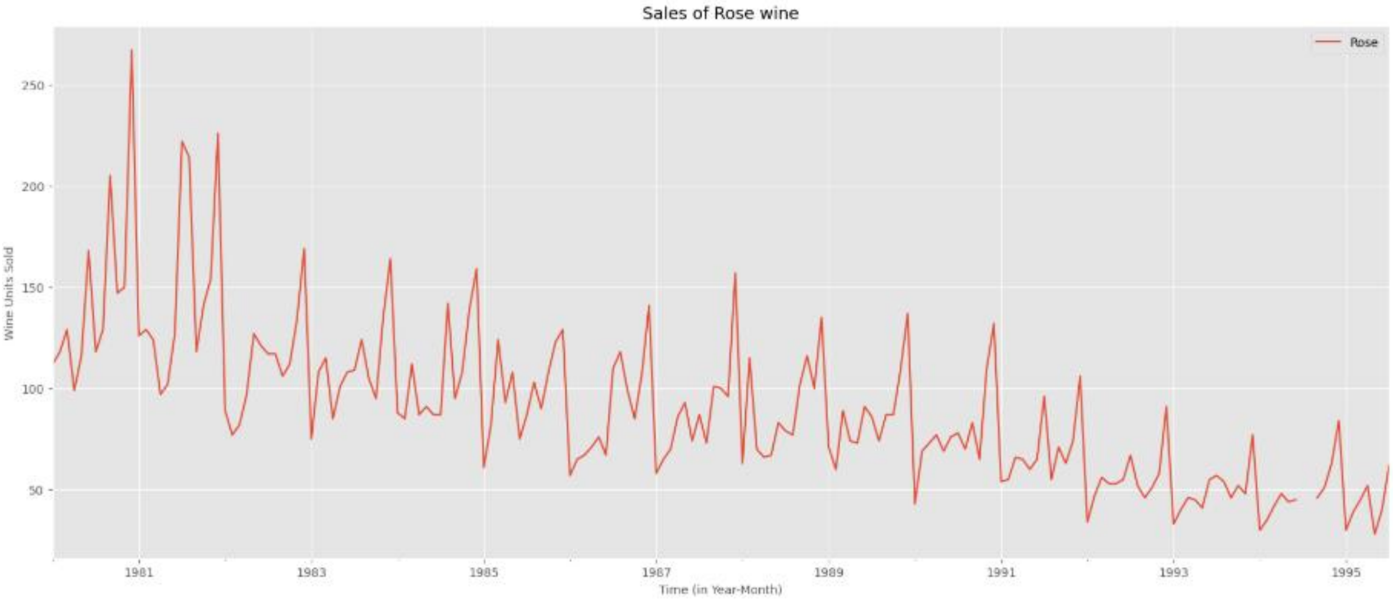


FIGURE 1 - TIMESERIES PLOT FOR ROSE WINES

- **Downward Trend:**
  - The most prominent feature is a clear downward trend in Rose wine sales over the observed period. Sales were generally higher in the early 1980s and gradually declined towards the mid-1990s.
- **Volatility:**
  - The sales data exhibits significant volatility, with numerous fluctuations and peaks and troughs. This suggests that sales are subject to various short-term factors.
- **Potential Seasonality:**
  - While not as pronounced as a clear, regular seasonal pattern, there appear to be some repeating fluctuations that might indicate some level of seasonality. Further analysis (e.g., using ACF/PACF plots or decomposition) would be needed to confirm this.
- **Peaks and Troughs:**

  - There are several notable peaks and troughs in the data. The highest peak occurs around 1981, and the lowest troughs occur in the later years, particularly around 1993 and 1995.

- **Time Range:**
  - The data spans from approximately 1981 to 1995, providing a long-term view of Rose wine sales.

Plotting a boxplot to understand the spread of sales across different years and within different months across years.

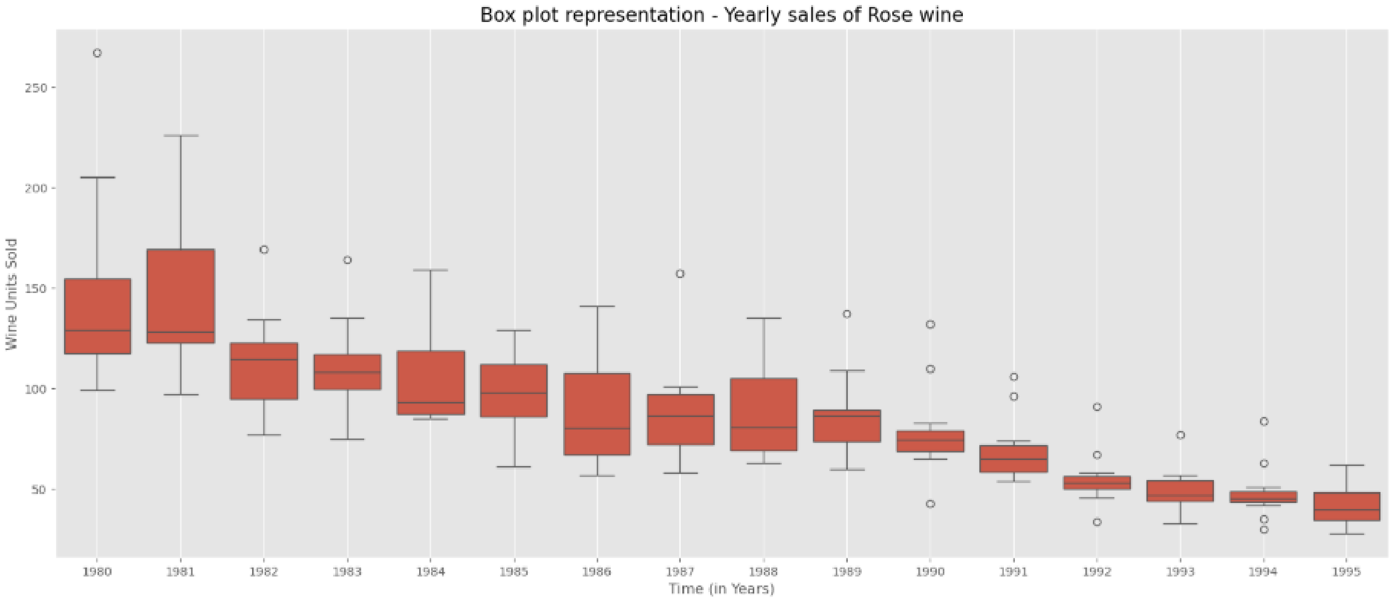


FIGURE 2 - BOXPLOT FOR ROSE WINE SALES OVER YEARS

- **Downward Trend:** The median sales (line inside each box) show a clear decline over the years, aligning with the time-series trend.
- **Decreasing Variability:** The interquartile range (IQR) shrinks over time, indicating reduced fluctuations in Rose wine sales.
- **Outliers:** Several data points fall outside the boxes, especially in the early (1980-1982) and later years.
- **Median Shift:** The median consistently moves downward, reinforcing the decline in sales.
- **Range Decline:** The overall spread of sales (distance between whiskers) is decreasing.

- **Steady Decline:** Sales have consistently dropped over the 16-year period.
- **Market Stability:** Reduced variability suggests a more predictable market, though at lower sales levels.
- **Outlier Investigation:** Outliers need further analysis to confirm if they are anomalies or valid data points.
- **Yearly Distribution:** The box plots effectively illustrate sales distribution, highlighting trends, variations, and outliers.

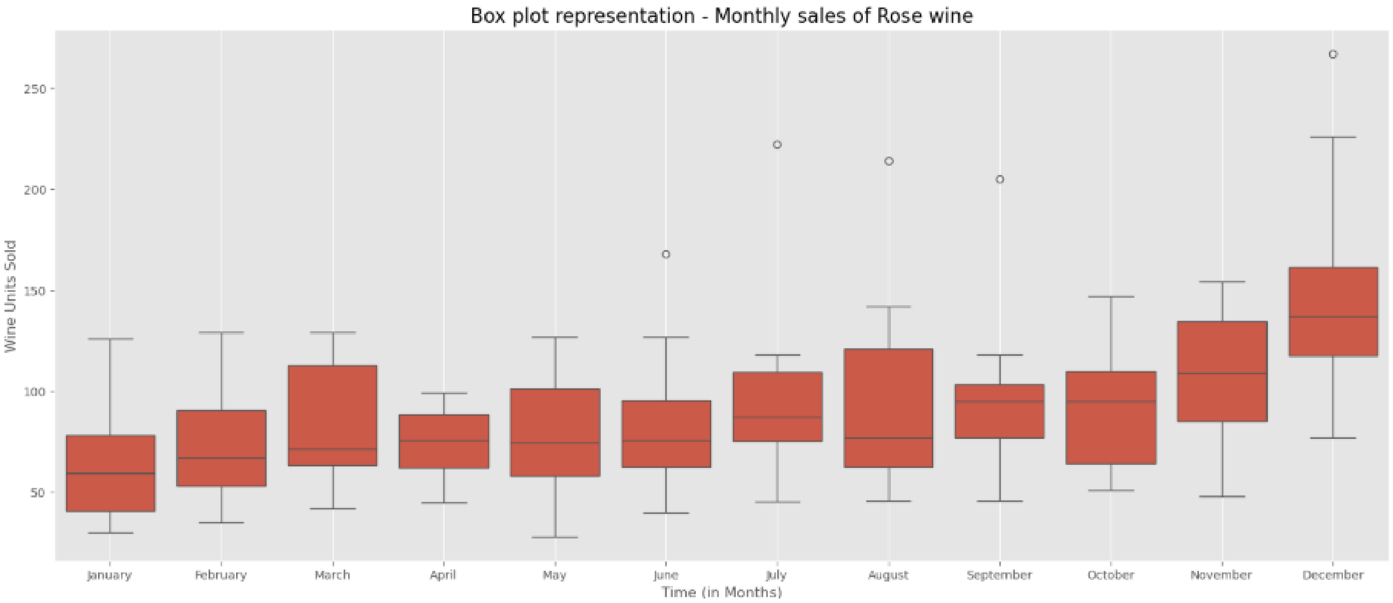


FIGURE 3 - BOXPLOT FOR ROSE WINE MONTHLY SALES

- **Seasonal Trend:** Sales are lower from **January to June** and rise from **July to December**.
- **December Peak:** Highest median sales and largest range, making it the strongest sales month.
- **January & February Slump:** Lowest median sales and smallest range, marking the weakest months.
- **Gradual Build-up:** Sales steadily increase from early year, accelerating from **September to December**.
- **Higher Variability:** Sales fluctuations increase in later months, especially in December.
- **Outliers:** Some months show unusually high or low sales, requiring further analysis.

- **Holiday Boost:** December sales surge suggests a strong holiday season influence.
- **Seasonal Demand:** Consistent patterns point to season-driven consumer behavior.
- **Predictability:** The trend repeats yearly, making demand forecasting more reliable.
- **Marketing Opportunity:** Clear seasonality allows for strategic promotions and inventory planning.

Now let's observe and plot the empirical cumulative Distribution function

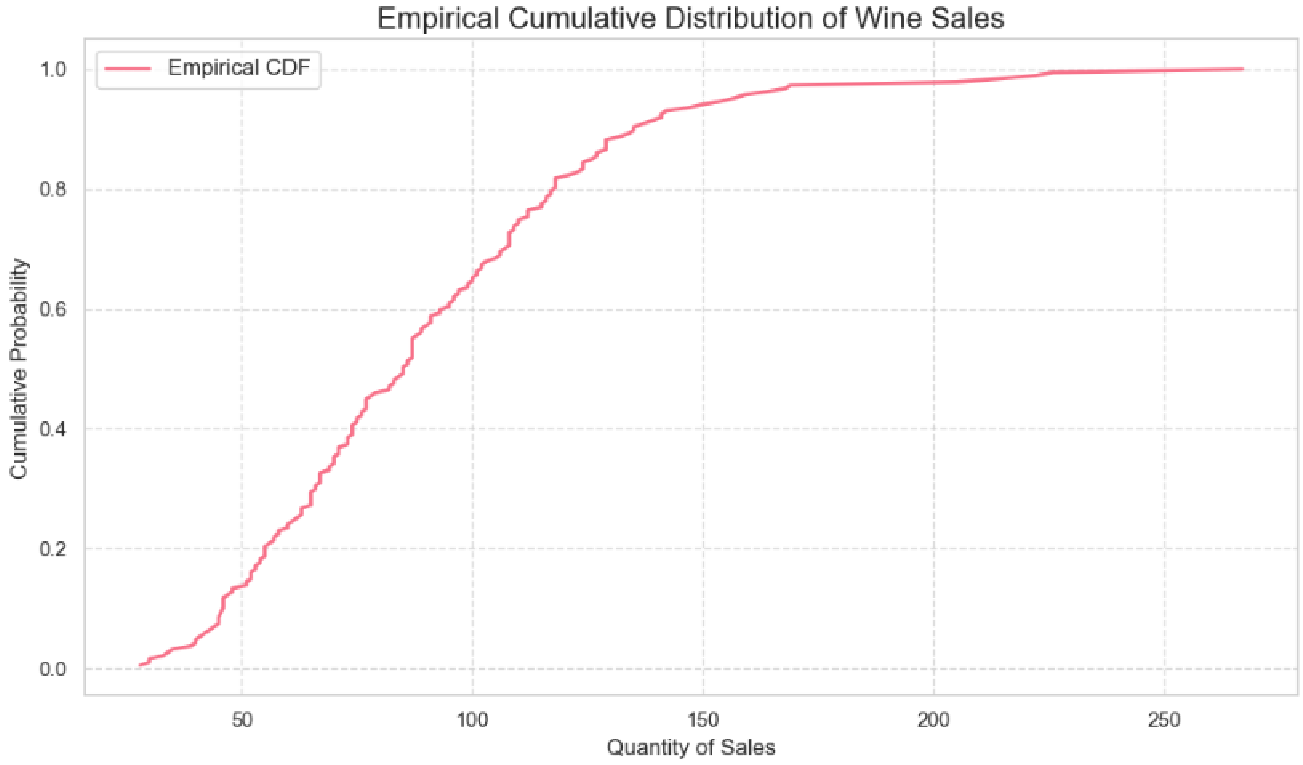


FIGURE 4 - ECDF FOR ROSE WINE MONTHLY SALES

- **Step-like ECDF Curve:** The plot rises from 0 to 1, showing cumulative sales distribution.
- **X-Axis (Sales Quantity):** Ranges from **0 to 250** units.
- **Y-Axis (Cumulative Probability):** Ranges from **0 to 1**, showing the proportion of sales ≤ a given quantity.
- **Visual Style:** Pinkish-red curve with a clean, light gray grid and white background.
- **Legend:** Labeled **"Empirical CDF"** in the top-left corner.

- **Sales Distribution:** ~20% of sales are **below 50** units, and ~80% are **below 150** units.
- **Steepness:** The sharp rise in the **50-100** range indicates high sales concentration.
- **Outliers:** The flat tail above **200 units** suggests rare high-sales instances.
- **Median Sales:** Estimated around **80-90** units (where cumulative probability = 0.5).

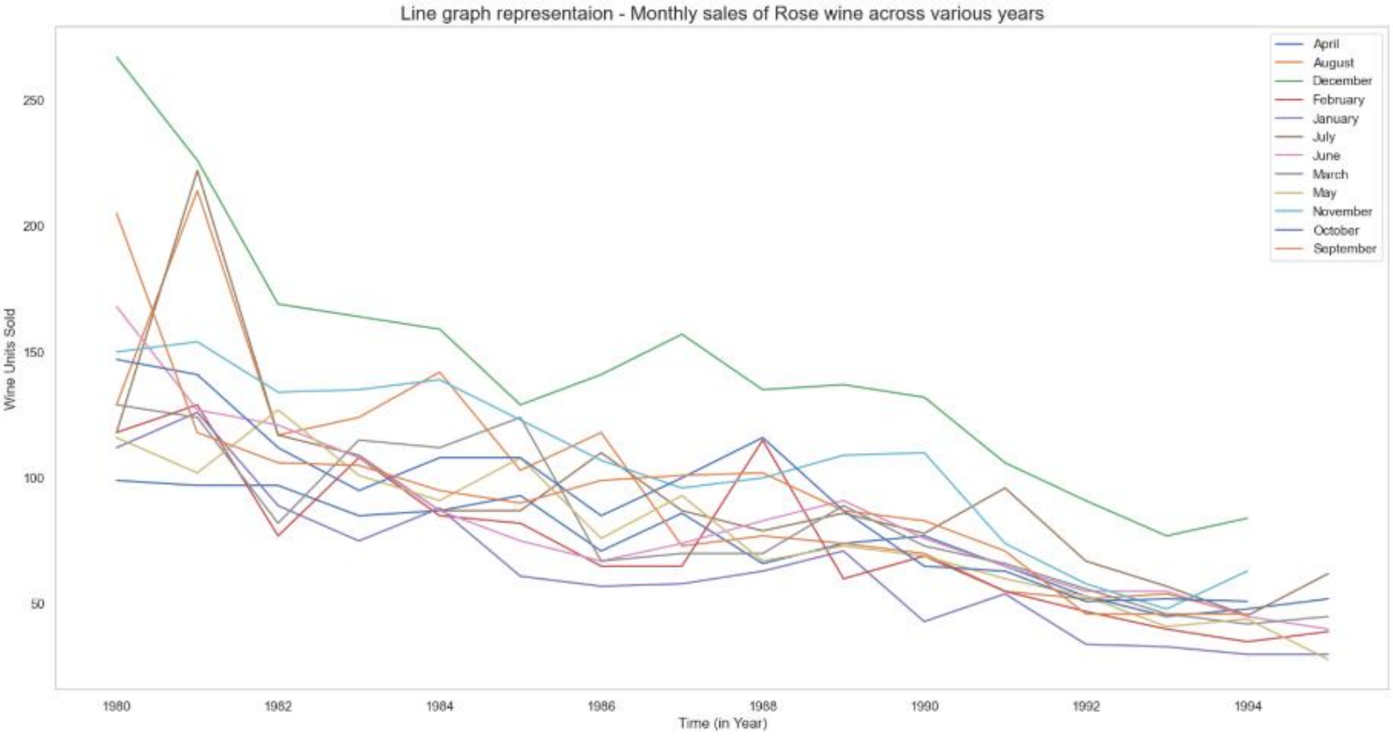


FIGURE 5 - PLOT FOR MONTHLY SALES OVER YEARS

- **Monthly Trends:** Each line represents a specific month's sales trend over the years, with a clear legend for identification.
- **Declining Sales:** Most monthly trends show a consistent downward trajectory, reinforcing the overall decline in Rose wine sales.
- **Early-Year Variability:** Sales were more volatile between **1980-1985**, with wider fluctuations.
- **Stabilization in Later Years:** Monthly sales converge over time, showing less variation and lower values.
- **Seasonal Patterns:** Some months, like **December**, consistently had higher sales, especially in earlier years.

- **Color Differentiation:** Distinct colors help separate trends, though some may appear similar.
- **Steady Market Decline:** The downward trend is persistent across all months.
- **Market Stabilization:** Reduced fluctuations in later years suggest a more predictable market at lower sales levels.
- **Seasonality Clues:** While the focus is yearly trends, seasonal patterns are evident and warrant deeper analysis.

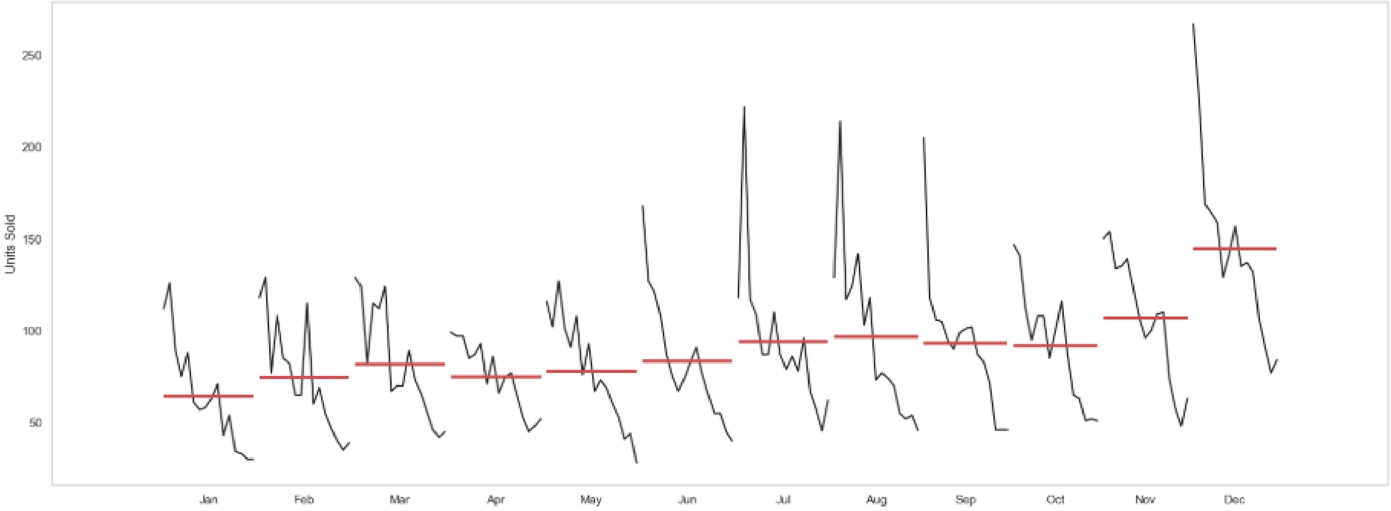


FIGURE 6 - TIMESERIES PLOT FOR MONTHLY SALES

- **Monthly Time Series:** Black lines show sales fluctuations over time for each month.
- **Monthly Averages:** Red horizontal lines highlight the average sales for each month, allowing quick comparisons.
- **Seasonal Pattern:** Sales are lower from **January to June** and peak from **July to December**, with **December** having the highest average.
- **Data Variability:** Sales fluctuate significantly within each month, with some months showing sharper spikes.
- **Visual Clarity:** The contrast between black time series lines and red averages ensures easy interpretation.
- **Seasonal Sales Trend:** Sales follow a clear seasonal pattern, peaking in later months.

- **Benchmarking Performance:** Red average lines help assess each month's typical sales level.
- **Strategic Opportunities:** Seasonal trends offer insights for **targeted marketing** and **inventory planning**.

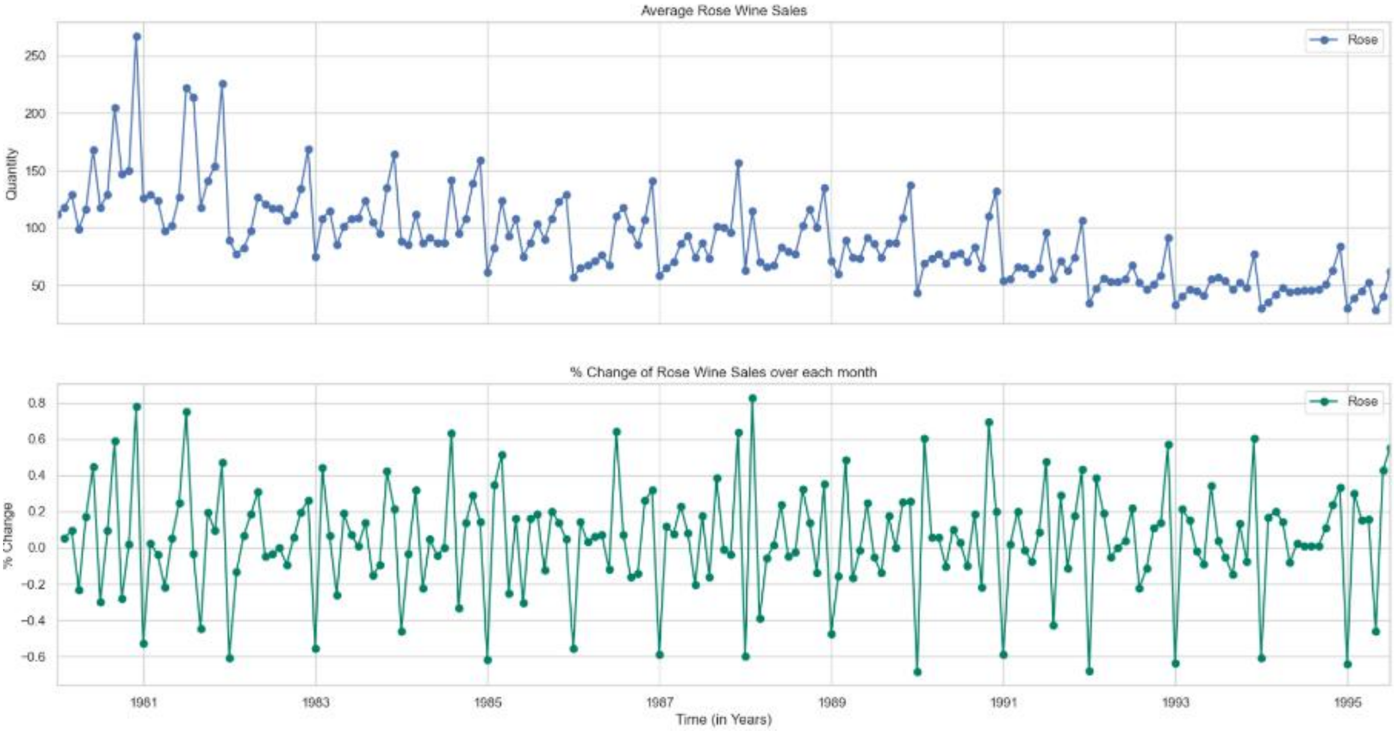


FIGURE 7 - PLOT FOR AVG SALES AND CHANGE IN % EVERY MONTH

Top Plot: Average Rose Wine Sales

- **Trend:** Sales show a **clear downward trajectory** from **1980 to 1995**.
- **Volatility:** Noticeable fluctuations with peaks and troughs.
- **Markers:** Circular markers highlight individual data points for better visibility.
- **Style:** Clean white grid background enhances readability.

Bottom Plot: % Change in Monthly Sales

- **Volatility:** Higher fluctuations than the average sales plot.
- **Pattern:** Some repeating trends suggest possible **seasonality** or cyclical influences.
- **Color:** Teal/green distinguishes it from the top plot.

- **Markers:** Circular markers emphasize individual data points.

- **Steady Decline:** The top plot confirms a **long-term decrease** in Rose wine sales.
- **Short-Term Instability:** The bottom plot highlights **significant monthly fluctuations**.
- **Potential Seasonality:** Patterns in percentage change hint at **seasonal trends** worth further analysis.
- **Market Shifts:** The combination of long-term decline and short-term volatility suggests **changing market dynamics**.

Decomposition of the Time Series

Additive decomposition

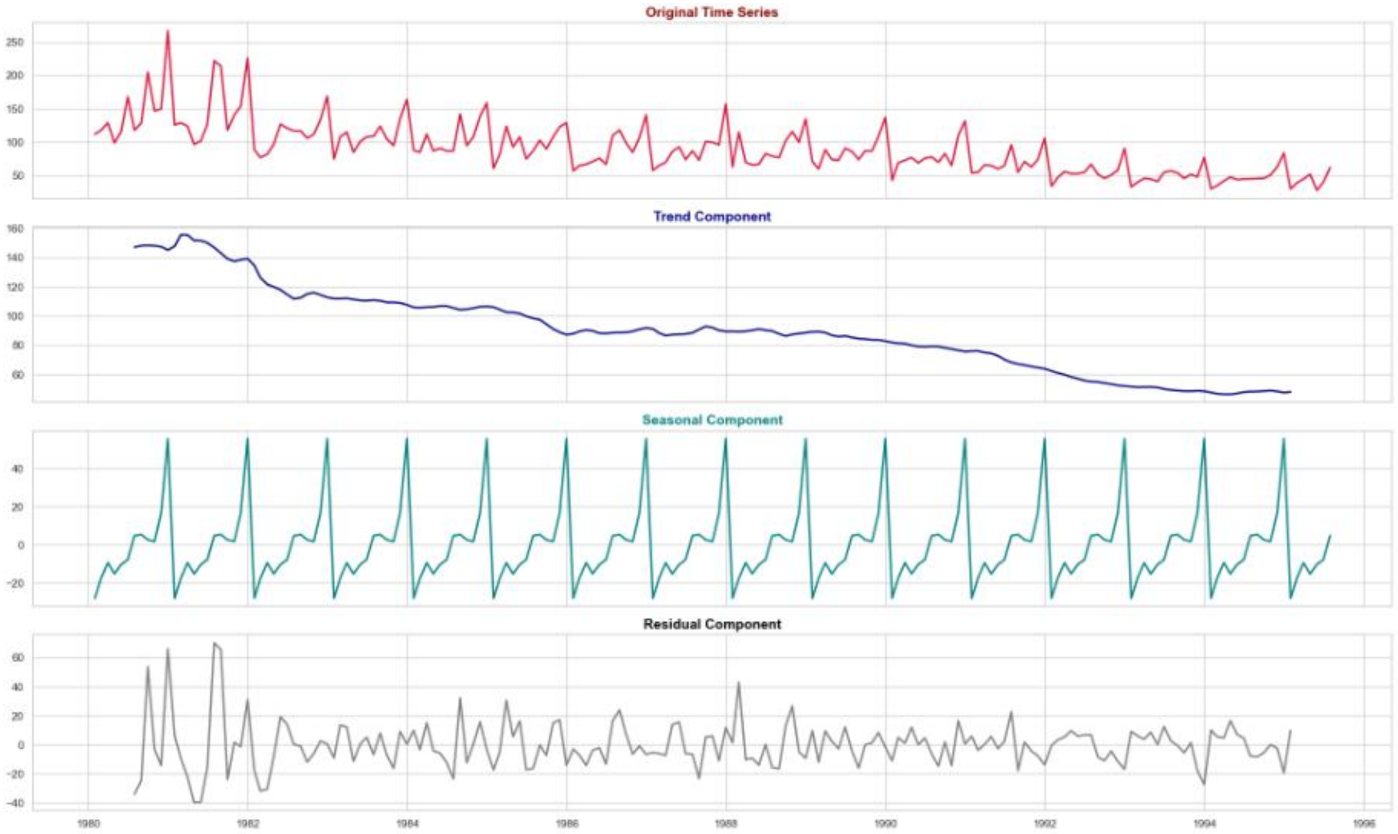


FIGURE 8 - ADDICTIVE DECOMPOSITION

1. Original Time Series (Top Panel)

- Trend: A clear downward trend over the years (1980–1996).
- Volatility: Frequent ups and downs indicate significant fluctuations.
- Potential Seasonality: Some recurring patterns suggest possible seasonal effects.

2. Trend Component (Second Panel)

- Smooth Long-Term Trend: Confirms the gradual decline seen in the original data.
- Noise Reduction: Short-term fluctuations are removed for a clearer direction.

3. Seasonal Component (Third Panel)

- Consistent Seasonality: Repeating peaks and troughs at regular intervals.
- Stable Amplitude: Seasonal variations remain relatively constant over time.
- Predictability: Strong and stable seasonal influence suggests recurring patterns.

4. Residual Component (Bottom Panel)

- Random Variations: Mostly unstructured noise, with no clear patterns.
- Volatility: Some spikes suggest possible outliers or external influences.

- Clear Trend & Seasonality: The decomposition effectively separates these key patterns.
- Stable Seasonal Influence: Sales fluctuations follow a predictable seasonal cycle.
- Residual Analysis Needed: Outliers or unexplained variations might require further investigation.
- This breakdown helps in forecasting, trend analysis, and strategic planning!

Multiplicative decomposition

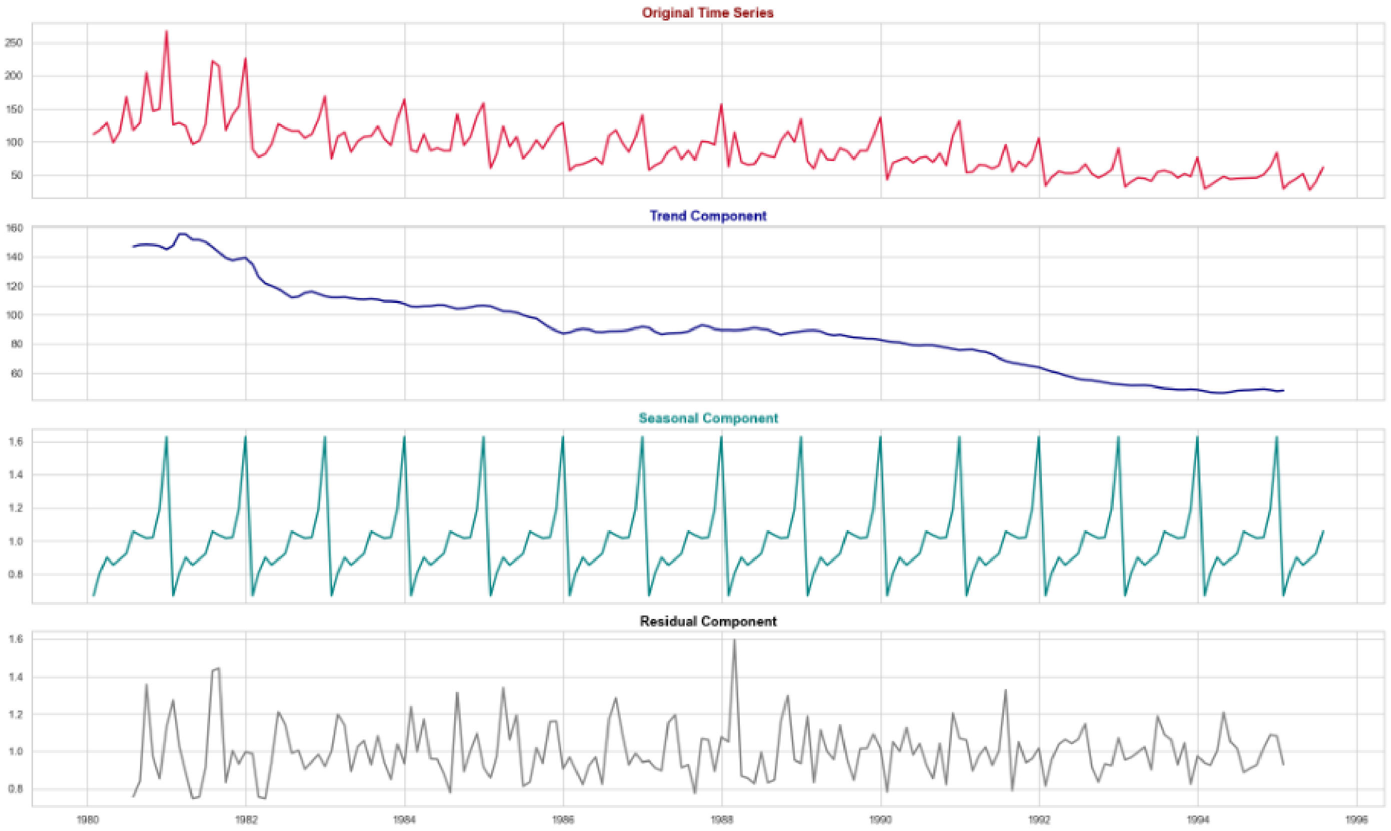


Figure 9 - Multiplicative decomposition

1. Original Time Series (Top Panel)

Overview: Displays the raw data, showing actual fluctuations from **1980 to 1996**.

- **Downward Trend:** Sales steadily decline over time.
- **High Volatility:** Frequent ups and downs indicate fluctuations.
- **Possible Seasonality:** Recurring patterns hint at seasonal effects.

2. Trend Component (Second Panel)

Overview: Extracts the **long-term trend**, smoothing out short-term fluctuations.

- **Confirms Downward Trend:** Reinforces the sales decline seen in raw data.
- **Smooth Representation:** Captures overall direction without noise.

3. Seasonal Component (Third Panel)

Overview: Isolates repeating seasonal patterns.

- **Consistent Peaks & Troughs:** Indicates a **stable seasonal effect**.
- **Fixed Amplitude:** Suggests **additive seasonality** (same magnitude over time).
- **Predictable Cycles:** Periodicity remains constant across years.

4. Residual Component (Bottom Panel)

Overview: Represents random fluctuations after removing trend & seasonality.

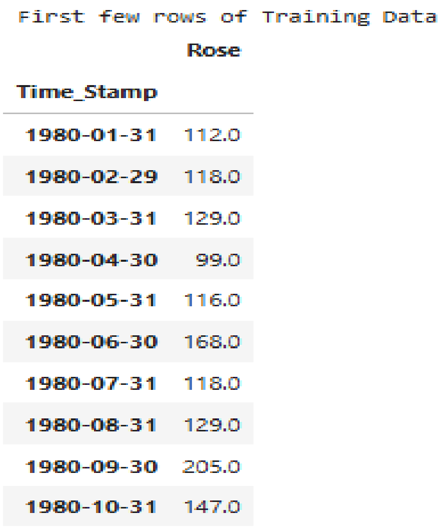
- **Mostly Random Noise:** No clear pattern remains.
- **Some Volatility:** Indicates additional influences beyond trend & seasonality.
- **Potential Outliers:** Sudden spikes suggest **unusual events** or anomalies.

- ➢ **Trend & Seasonality Dominate:** Sales decline consistently, with strong seasonal effects.
- ➢ **Residuals Are Random:** Suggests no major hidden patterns, but some outliers exist.
- ➢ **Predictable Sales Behavior:** Seasonal trends offer insights for **forecasting & strategy**.

Test Train Split

We have split the data into training and testing sets.

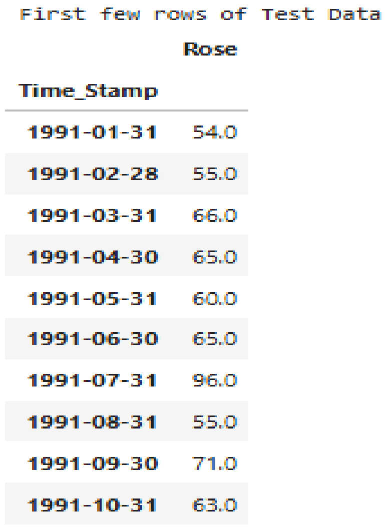
Lets have a view to top 10 rows from training set and test set –



First few rows of Training Data

	Rose
Time_Stamp	
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
1980-06-30	168.0
1980-07-31	118.0
1980-08-31	129.0
1980-09-30	205.0
1980-10-31	147.0

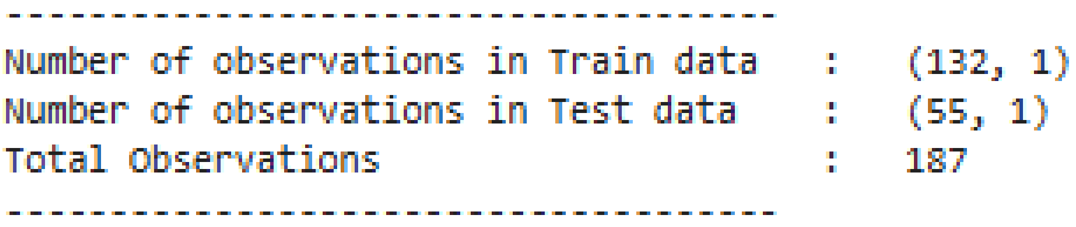
Table 4 – Train set



First few rows of Test Data

	Rose
Time_Stamp	
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
1991-06-30	65.0
1991-07-31	96.0
1991-08-31	55.0
1991-09-30	71.0
1991-10-31	63.0

Table 5 – Test set



```
----------------------------------------
Number of observations in Train data    :    (132, 1)
Number of observations in Test data     :    (55, 1)
Total Observations                      :    187
----------------------------------------
```

Model Building

Model 1: Linear Regression

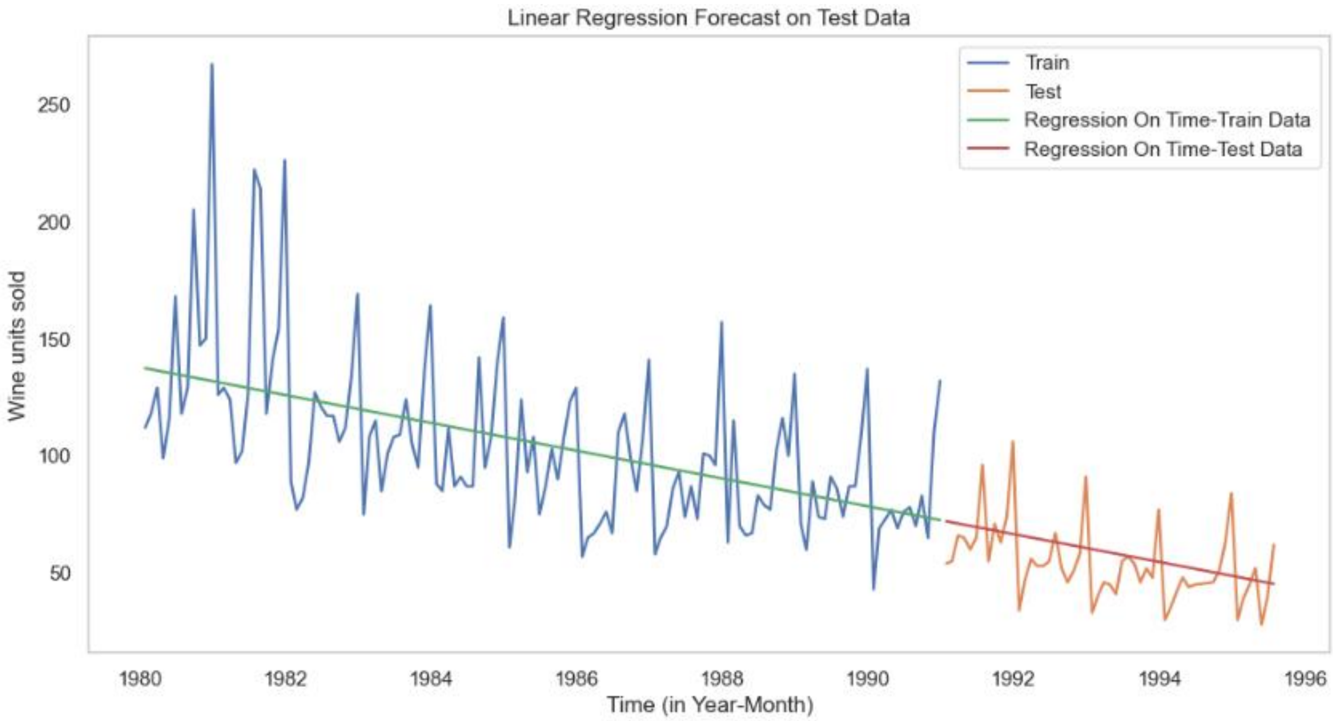


Figure 10 - Linear regression forecast

1. Train-Test Split

- Dataset is split into training (blue) and test (orange) sets.
- Clear separation ensures proper model validation.
- Test data follows the training trend but with some fluctuations.

2. Linear Regression Fit

- Green line represents the linear regression fit on training data.
- Red line represents the linear regression fit on test data.
- Both lines show a downward trend in sales.
- The model fails to capture seasonality and short-term fluctuations.

3. Performance & Limitations

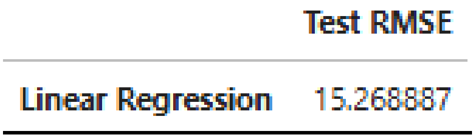
- The model correctly identifies the overall decline in sales.
- It does not capture seasonal patterns or short-term variations.
- The linear approach is too simplistic for time-series forecasting.

Final Insights

- Linear regression works well for detecting long-term trends.
- It is not suitable for capturing non-linear variations.
- More advanced models like ARIMA, Exponential Smoothing, or LSTMs may improve accuracy

Model evaluation - Linear Regression

So a per code calculations, For RegressionOnTime forecast on the Test Data, RMSE is 15.269



	Test RMSE
Linear Regression	15.268887

TABLE 6 – LINEAR REGRESSION RMSE

Model 2: Simple Average

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

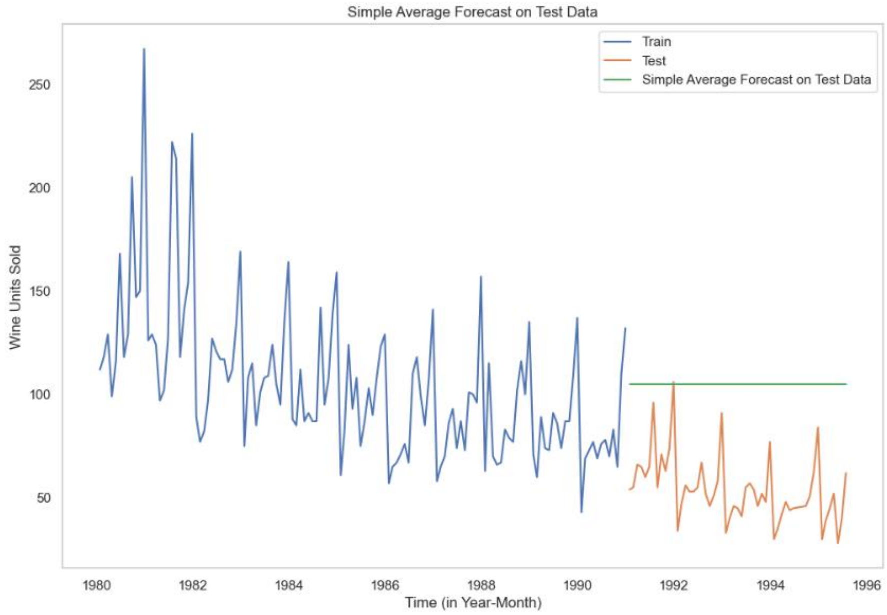


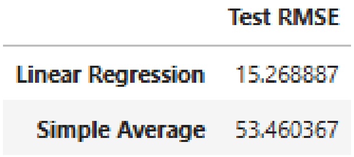
FIGURE 11 - SIMPLE AVG FORECAST

- **Train-Test Split:**
    - Blue line represents training data, orange line represents test data.
- **Simple Average Forecast**:
    - Green line represents the simple average forecast for test data.
    - Forecast is a flat, horizontal line around 105 units.
- **Lack of Fit:**
    - The forecast fails to capture trend, seasonality, and volatility.
    - Actual test data shows fluctuations and a downward trend, which the model misses.

- Simple Model Limitations: The simple average method is ineffective for time-series forecasting with trends and seasonality.
- Poor Predictive Power: It does not reflect underlying patterns, making it an unreliable predictor.
- Baseline Model: Can be used as a basic benchmark for evaluating more advanced forecasting models.

Model evaluation

So a per code calculations For Simple Average forecast on the Test Data, RMSE is 53.460



	Test RMSE
Linear Regression	15.268887
Simple Average	53.460367

TABLE 7 – SIMPLE AVG RMSE

Model 3: Moving Average

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.

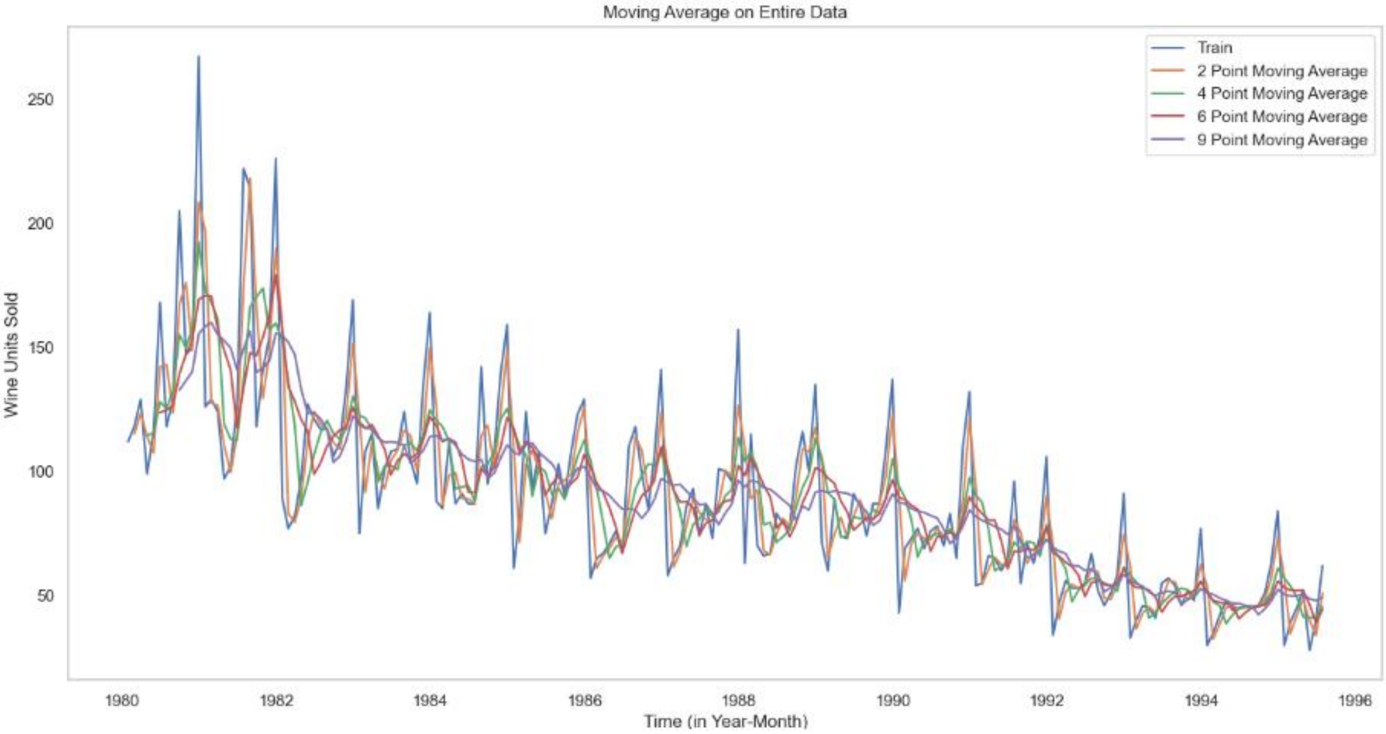


FIGURE 12 - MOVING AVG FORECAST

- **Original Data (Train - Blue Line):**
    - Shows significant fluctuations and a general downward trend.
- **Moving Averages:**
    - Different window sizes (2, 4, 6, and 9 points) applied to smooth the data.
    - **2-Point MA (Orange):** Follows data closely but smooths minor noise.
    - **4-Point MA (Green):** Reduces sharp peaks and troughs.
    - **6-Point MA (Red):** Further smooths, capturing a broader trend.
    - **9-Point MA (Purple):** Smoothest representation, minimizing short-term fluctuations.
- **Smoothing & Trend Capture:**
    - Larger window sizes smooth data more but introduce a lag.

- All moving averages confirm the overall downward trend.
- Higher window sizes reduce seasonality effects.

- **Noise Reduction:** Moving averages help filter out short-term fluctuations.
- **Trade-off:** Smaller windows respond quickly to changes, while larger windows provide a clearer trend but with more lag.
- **Trend Identification:** The 9-point moving average highlights the long-term trend effectively.

Now let's see for first 3 years-

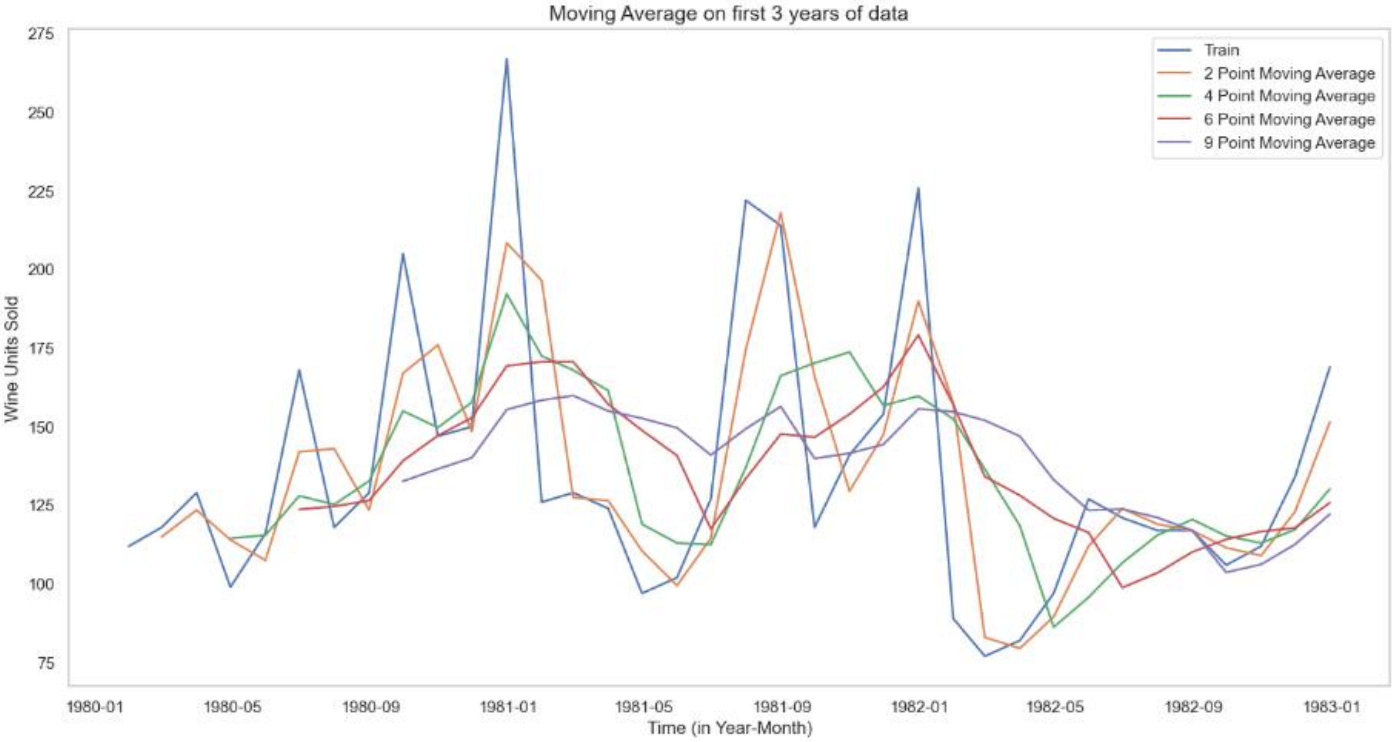


Figure 13 - Moving Avg for first 3 years forecast

- **Original Data (Train - Blue Line):**
  - The blue line represents the original time series data for the first three years, showing significant fluctuations.
- **Moving Averages:**
  - The plot displays moving averages calculated with window sizes of 2, 4, 6, and 9 points.

  - **2 Point Moving Average (Orange Line):**
    - Closely follows the original data but smooths out some of the highest frequency noise.
  - **4 Point Moving Average (Green Line):**
    - Smooths the data more significantly, reducing the sharp peaks and troughs.
  - **6 Point Moving Average (Red Line):**
    - Further smooths the data, creating a more generalized trend line.
  - **9 Point Moving Average (Purple Line):**
    - Provides the smoothest representation, capturing the overall trend while minimizing short-term fluctuations.
- **Smoothing Effect:**
  - As the window size increases, the moving average lines become smoother, indicating a stronger smoothing effect.
- **Lagging Effect:**
  - Moving averages introduce a lagging effect, where the smoothed line trails behind the original data. The larger the window size, the greater the lag.
- **Seasonal Pattern Observation:**
  - The original data shows a potential seasonal pattern with peaks and troughs, which is gradually smoothed out by the moving averages.

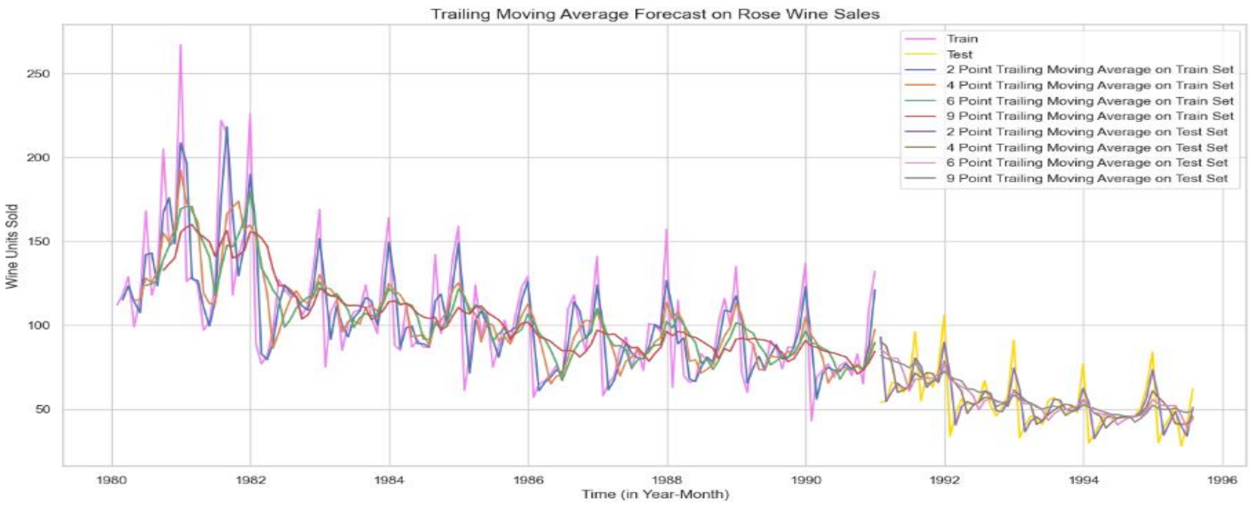


Figure 14 - Trailing Moving Avg forecast on Train vs Test

- **Train-Test Split:**
  - The plot clearly distinguishes between the training data (pink line) and the test data (yellow line), allowing for a clear evaluation of the model's performance on unseen data.
- **Trailing Moving Averages on Training Data:**
  - The blue, green, red, and purple lines represent the trailing moving averages with window sizes of 2, 4, 6, and 9 points, respectively, calculated on the training dataset.
  - These lines progressively smooth the training data as the window size increases, demonstrating the smoothing effect of larger window sizes.
- **Trailing Moving Averages on Test Data:**
  - The darker-shaded lines (corresponding to the same window sizes) represent the trailing moving averages calculated on the test dataset.
  - The same smoothing effect that is shown on the training data, is also shown on the testing data.
- **Smoothing and Lagging Effect:**
  - As the window size increases, the moving average lines become smoother, effectively reducing noise and highlighting the underlying trend.
  - However, this smoothing introduces a lagging effect, where the smoothed lines trail behind the actual data points. The larger the window size, the greater the lag.

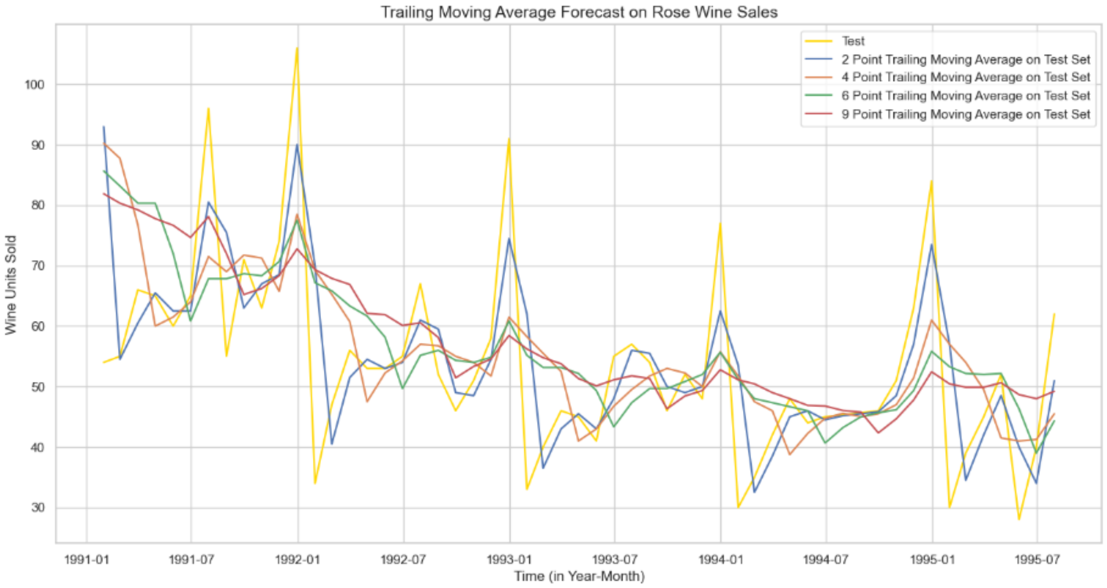
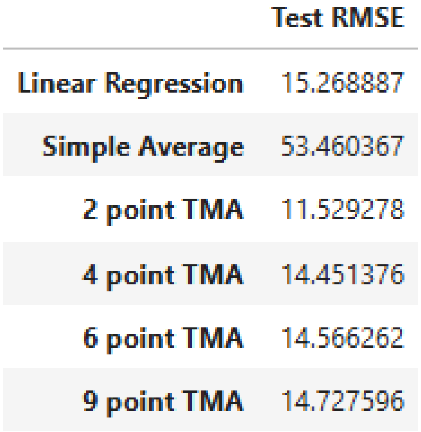


Figure 15 - Trailing Moving Avg forecast on Test data

Model evaluation

So a per code calculations,



	Test RMSE
Linear Regression	15.268887
Simple Average	53.460367
2 point TMA	11.529278
4 point TMA	14.451376
6 point TMA	14.566262
9 point TMA	14.727596

Table 8 – Moving avg RMSE

Model 4: Simple Exponential Smoothing

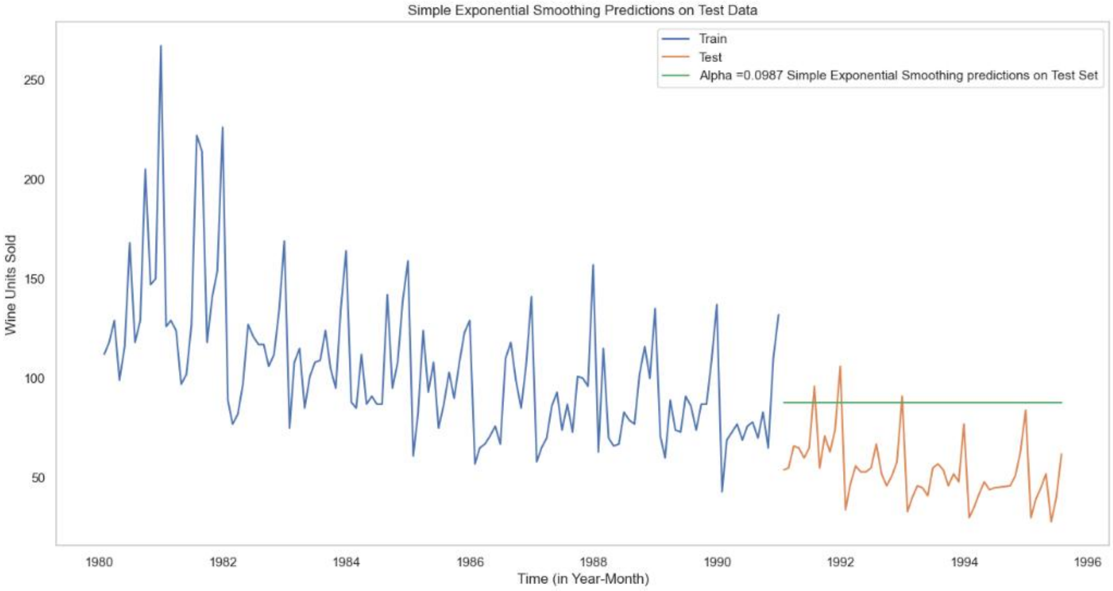
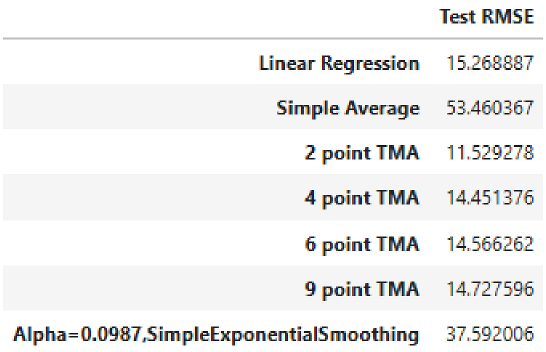


Figure 16 - SES predictions

Model evaluation

So a per code calculations,



	Test RMSE
Linear Regression	15.268887
Simple Average	53.460367
2 point TMA	11.529278
4 point TMA	14.451376
6 point TMA	14.566262
9 point TMA	14.727596
Alpha=0.0987,SimpleExponentialSmoothing	37.592006

Table 9 – SES RMSE

Setting different alpha values, the higher the alpha value more weightage is given to the more recent observation.

We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

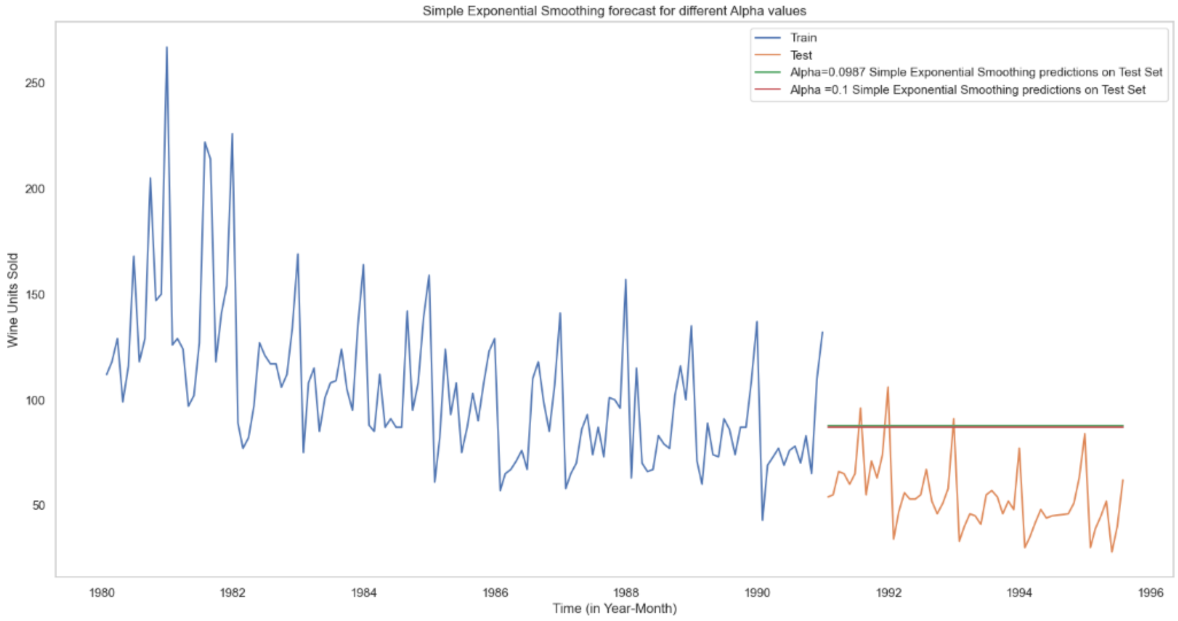


Figure 17 - SES predictions with different alpha

- **Train-Test Split:**
  - The blue line represents the training data, and the orange line represents the test data.

- **Simple Exponential Smoothing Forecasts:**
  - The green line represents the forecast with Alpha = 0.0987.
  - The red line represents the forecast with Alpha = 0.1.
- **Flat Forecasts:**
  - Both forecasts are flat, horizontal lines, indicating that Simple Exponential Smoothing produces a constant forecast for the test period.
- **Similar Forecast Values:**
  - The forecasts with Alpha = 0.0987 and Alpha = 0.1 are very close in value, suggesting that small changes in Alpha have little impact in this case.
- **Lack of Trend and Seasonality Capture:**
  - Neither forecast captures the trend, seasonality, or volatility present in the actual test data (orange line).

- **Simple Exponential Smoothing Limitations:** This plot highlights the limitations of Simple Exponential Smoothing for forecasting time series data with trends and seasonality.
- **Constant Forecast:** Simple Exponential Smoothing produces a constant forecast, which is not suitable for data with changing patterns.
- **Insensitivity to Small Alpha Changes:** The forecasts are insensitive to small changes in Alpha, indicating that fine-tuning Alpha might not significantly improve the forecast.
- **Poor Predictive Power:** The forecasts have poor predictive power, as they don't reflect the underlying patterns in the data.

Model 5: Double Exponential Smoothing

Two parameters α and β are estimated in this model. Level and Trend are accounted for in this model.

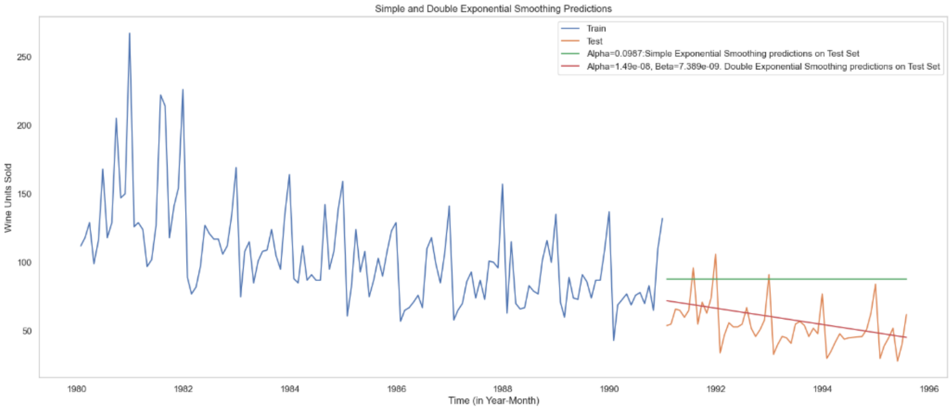
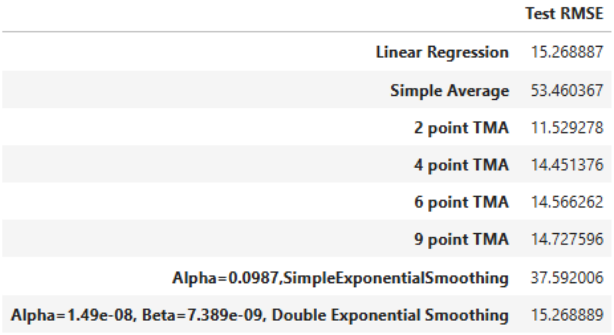


Figure 18 - DES predictions



	Test RMSE
Linear Regression	15.268887
Simple Average	53.460367
2 point TMA	11.529278
4 point TMA	14.451376
6 point TMA	14.566262
9 point TMA	14.727596
Alpha=0.0987,SimpleExponentialSmoothing	37.592006
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.268889

Table 10 – DES RMSE

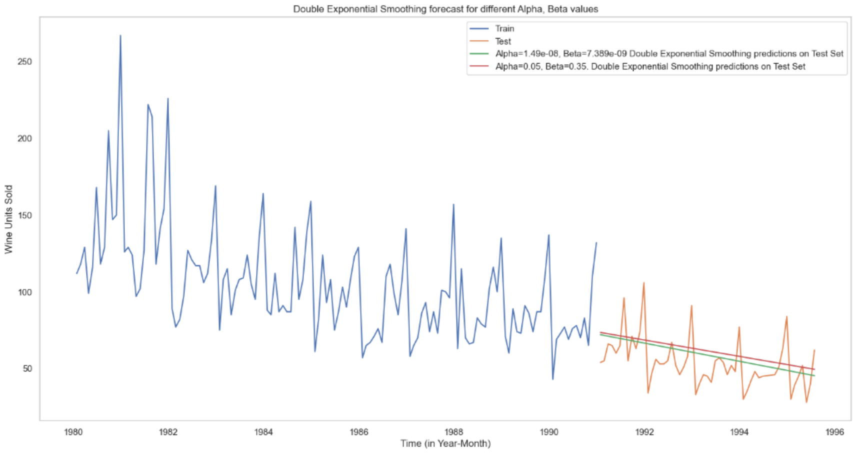


Figure 19 - DES predictions with alpha

- **Train-Test Split:**
  - The blue line represents the training data, and the orange line represents the test data.
- **Double Exponential Smoothing Forecasts:**
  - The green line represents the forecast with Alpha = 1.49e-08 and Beta = 7.389e-09 (very small values).
  - The red line represents the forecast with Alpha = 0.05 and Beta = 0.35.
- **Trend Capture:**
  - Both forecasts capture a downward trend, which aligns with the general trend in the test data.
  - The red line (Alpha = 0.05, Beta = 0.35) seems to capture the trend slightly better.
- **Lack of Seasonality Capture:**
  - Neither forecast captures the seasonality or short-term fluctuations present in the actual test data (orange line).
- **Forecast Behavior:**
  - The green line (very small Alpha and Beta) is almost a straight line, indicating very little smoothing or responsiveness to recent data.
  - The red line (Alpha = 0.05, Beta = 0.35) shows a slightly more pronounced downward trend, indicating more responsiveness to recent trend changes.

- **Double Exponential Smoothing Trend Capture:** Double Exponential Smoothing can capture the trend component in the data, as seen in both forecasts.
- **Impact of Alpha and Beta:** The choice of Alpha and Beta values significantly affects the forecast. Smaller values lead to smoother forecasts and less responsiveness, while larger values lead to more responsiveness.
- **Limitations:** Double Exponential Smoothing fails to capture the seasonality and short-term fluctuations, limiting its overall predictive power in this case.
- **Better Fit with Higher Beta:** The red line (Alpha = 0.05, Beta = 0.35) seems to provide a better fit to the trend in the test data, suggesting that a higher Beta value might be more appropriate for this dataset.

Model 6: Triple Exponential Smoothing

Three parameters α, β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Also the fit of the model is by the best parameters that Python thinks for the model. It uses a brute force method to choose the parameters.

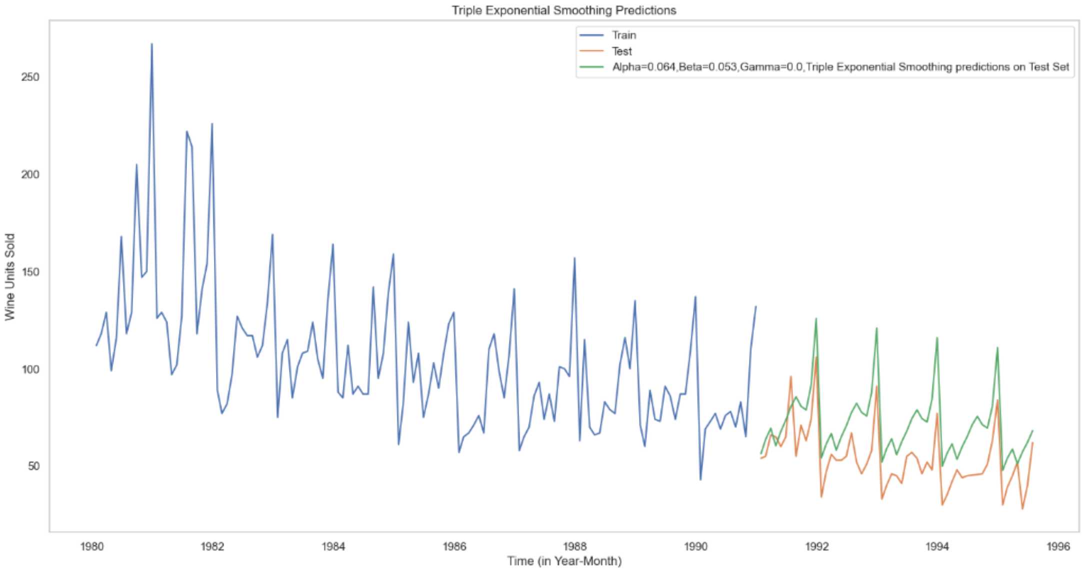
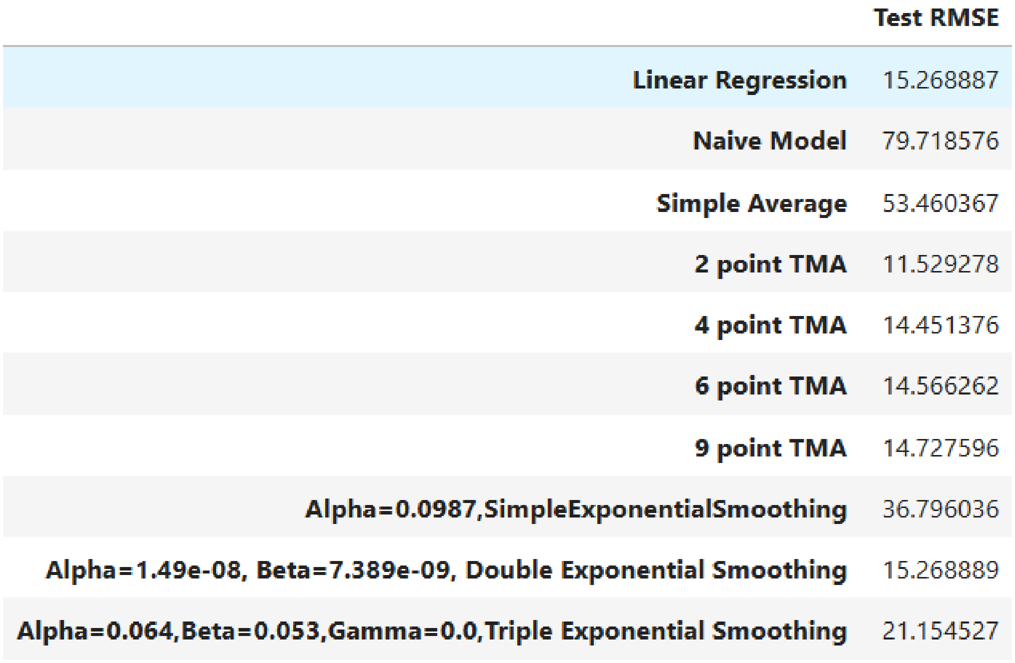


Figure 20 - TES predictions



	Test RMSE
Linear Regression	15.268887
Naive Model	79.718576
Simple Average	53.460367
2 point TMA	11.529278
4 point TMA	14.451376
6 point TMA	14.566262
9 point TMA	14.727596
Alpha=0.0987,SimpleExponentialSmoothing	36.796036
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.268889
Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing	21.154527

Table 11– TES RMSE

Now, calculating the performance metrics for different values of alpha, beta and gamma

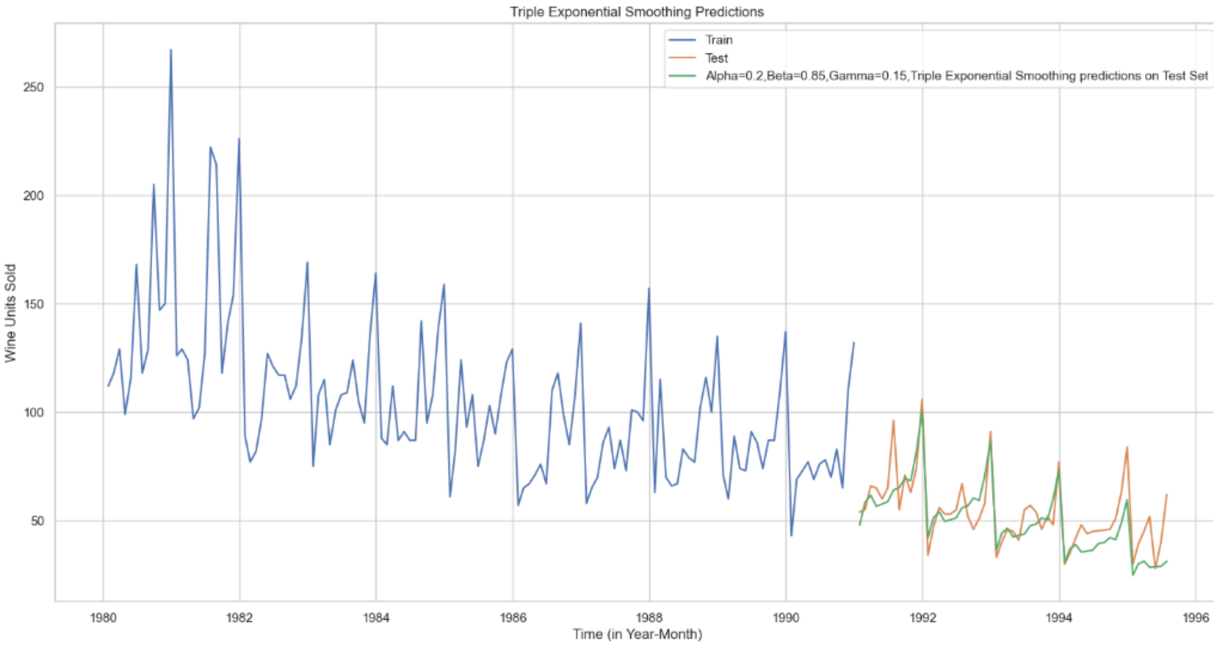


Figure 21 - TES predictions with alpha, beta & gamma

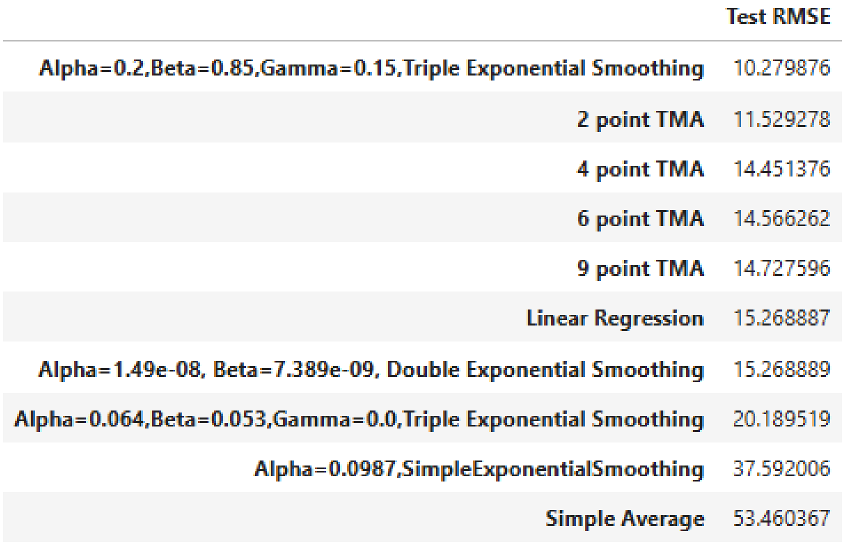
- **Train-Test Split:**
  - The blue line represents the training data, and the orange line represents the test data.
- **TES Forecast:**
  - The green line represents the Triple Exponential Smoothing forecast on the test data, using Alpha = 0.2, Beta = 0.85, and Gamma = 0.15.
- **Trend and Seasonality Capture:**
  - The TES forecast appears to capture both the downward trend and the seasonality present in the test data.
  - The green line (forecast) follows the general pattern of the orange line (test data).
- **Magnitude Discrepancies:**
  - While the forecast captures the overall pattern, there are some discrepancies in the magnitude of the predicted values, especially around peaks and troughs.
  - The green line does not have the same highs and lows as the orange line.
- **Model Fit:**
  - The model demonstrates a reasonable fit to the test data, indicating that Triple Exponential Smoothing is a suitable method for this time series.

- **Effectiveness of TES:** The plot demonstrates the effectiveness of Triple Exponential Smoothing in capturing both trend and seasonality in time series data.
- **Parameter Impact:** The specific Alpha, Beta, and Gamma values used in the forecast have a significant impact on the model's performance.
- **Potential for Improvement:** While the model performs well, there is potential for further improvement by fine-tuning the parameters or exploring other model specifications.

Model evaluation –

RMSE for Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing -10.279876

Overall comparison for all models–



	Test RMSE
Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing	10.279876
2 point TMA	11.529278
4 point TMA	14.451376
6 point TMA	14.566262
9 point TMA	14.727596
Linear Regression	15.268887
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.268889
Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing	20.189519
Alpha=0.0987,SimpleExponentialSmoothing	37.592006
Simple Average	53.460367

Table 12– Overall RMSE comparison for all models

Since this dataset exhibits both trend and seasonality, **Triple Exponential Smoothing** is theoretically expected to perform better than Simple and Double Exponential Smoothing. However, as this was a model-building exercise, we explored multiple models and compared their performance based on the best RMSE value on the test data.

Stationarity

Stationarity should be tested at α = 0.05.

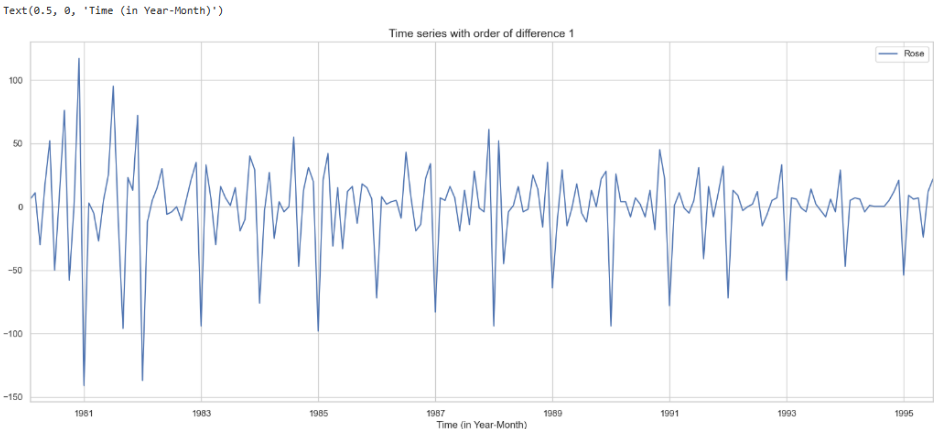
- Perform a stationarity check on the entire time series data.
- The Augmented Dickey-Fuller (ADF) test is a unit root test that determines whether the series has a unit root, indicating non-stationarity.

ADF Test Hypotheses:

- Null Hypothesis (H_0): The time series has a unit root (i.e., it is non-stationary).
- Alternative Hypothesis (H_1): The time series does not have a unit root (i.e., it is stationary).

Since ARIMA models require stationarity, the goal is to obtain a p-value lower than the chosen significance level (α = 0.05).

As per code, we observe that at 5% significant level the Time Series is non-stationary. Let us take one level of differencing to see whether the series becomes stationary.



```
Text(0.5, 0, 'Time (in Year-Month)')
```

Figure 22 - Stationary series plot

- The plot shows the first-order differenced time series, where each point represents the change from the previous one, helping to remove trends.
- The data fluctuates around zero, indicating that differencing has reduced non-stationarity, though volatility remains.
- No clear upward or downward trend is visible, but subtle patterns suggest potential seasonality requiring further analysis.
- While stationarity is improved, additional modeling (e.g., seasonal differencing or ACF/PACF analysis) may be needed.

ACF plot

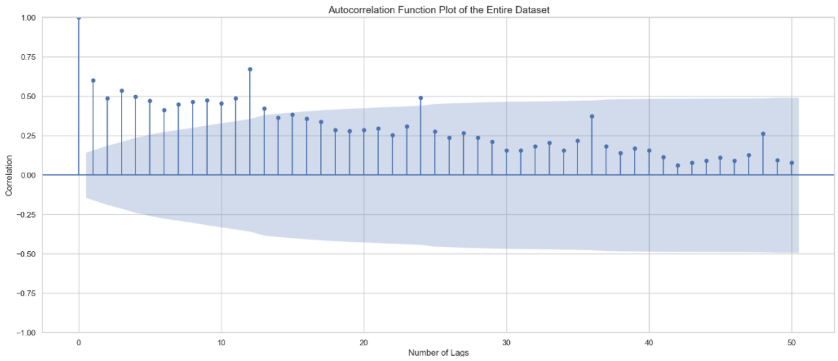


Figure 23 - ACF Plot

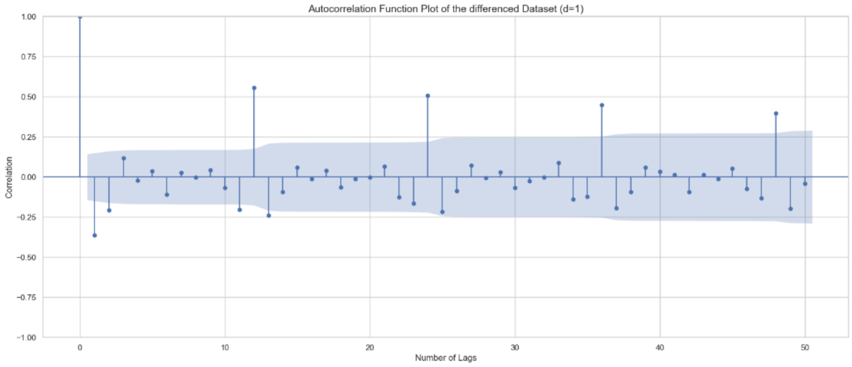


Figure 24 - ACF Plot with d=1

- **ACF Plot Overview:** Displays autocorrelation values for lags ranging from 0 to 50.
- **Axes:** Y-axis represents correlation (-1 to 1), and X-axis represents lag numbers.
- **Significant Autocorrelation:** High positive autocorrelation at lag 0 and multiple subsequent lags.
- **Gradual Decay:** Autocorrelation values decrease slowly, indicating long-term dependence.
- **Significance Threshold:** The blue shaded area marks the confidence interval; values outside are statistically significant.
- **Non-Stationarity Indication:** Gradual decay suggests the time series is non-stationary.
- **Trend/Seasonality Presence:** Strong autocorrelations at multiple lags suggest trend or seasonal patterns.
- Also the other plot shows the Autocorrelation Function (ACF) of a time series after applying a first-order difference (d=1). The data now fluctuates around zero, indicating stationarity, with most lags falling within the significance threshold. However, a few significant spikes suggest potential remaining autocorrelation that might need further investigation.

PACF plot

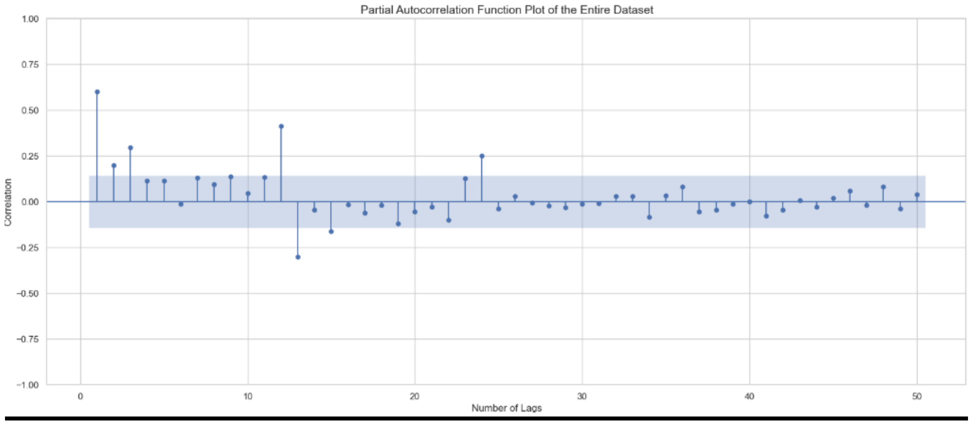


Figure 25 - PACF Plot

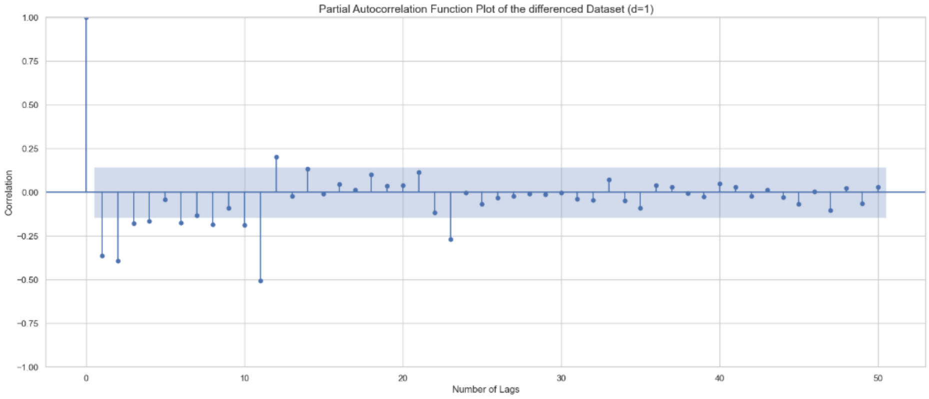
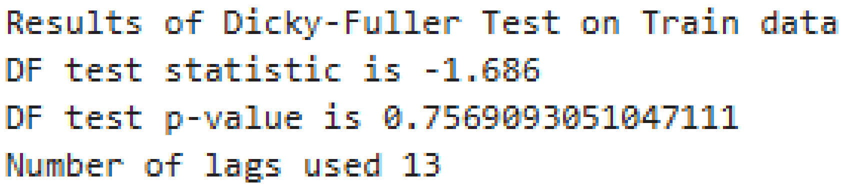


Figure 26 - PACF Plot with d=1

- This plot displays the Partial Autocorrelation Function (PACF) of the entire dataset.
- Significant spikes are observed at lags 1 and 11, indicating potential direct correlations at those lags.
- Most other lags fall within the confidence interval, suggesting they are not significantly correlated.
- The other plot shows the Partial Autocorrelation Function (PACF) of the dataset after a first-order difference (d=1).
- A strong spike is observed at lag 1, indicating a significant direct correlation with the immediately preceding value.
- Several other lags, notably lag 11, also show significant partial autocorrelation, suggesting potential direct dependencies at those specific time intervals.

Now, lets check for stationarity of the Training Data Time Series.



```
Results of Dicky-Fuller Test on Train data
DF test statistic is -1.686
DF test p-value is 0.7569093051047111
Number of lags used 13
```

Table 13– Stationarity on Training data

The training data is non-stationary at 95% confidence level. Let us take a first level of differencing to stationarize the Time Series.

Plot the differenced training data

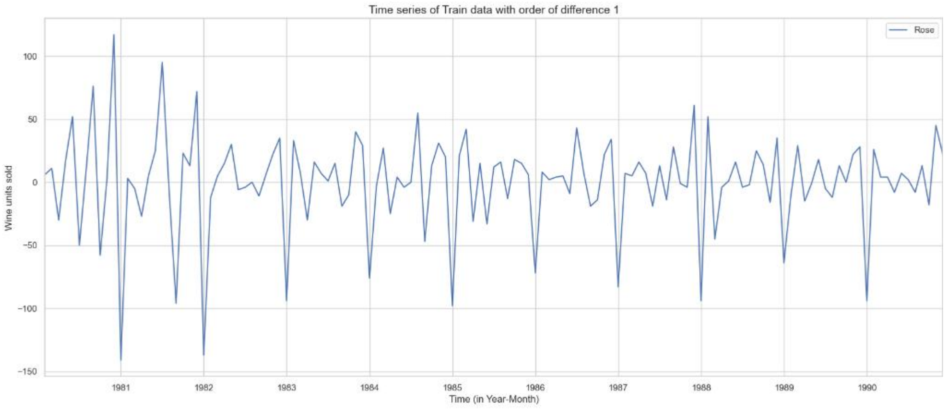
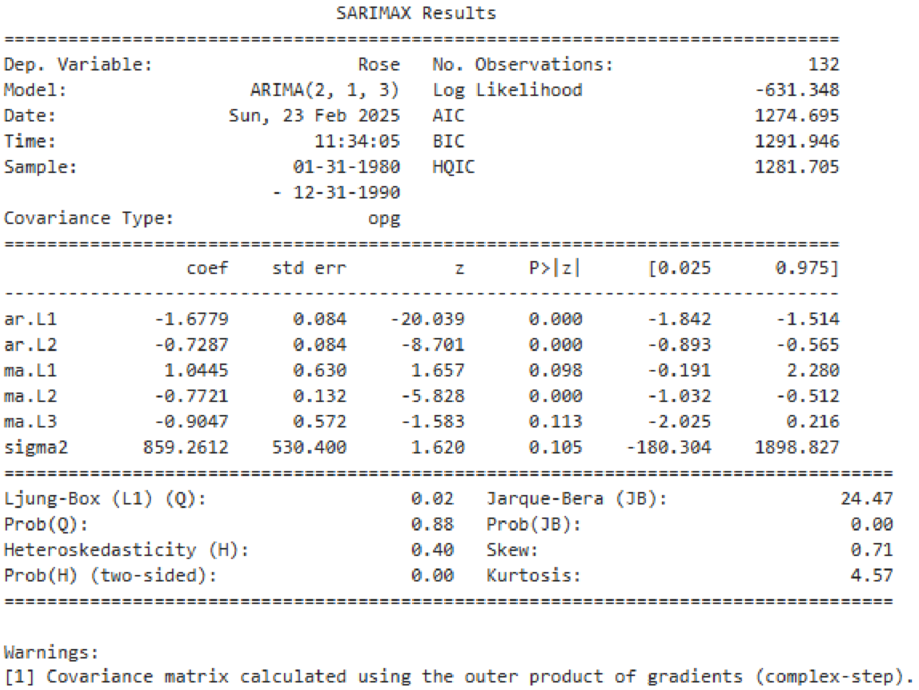


Figure 27 - Difference Training data Plot

- **Differenced Data:** The plot represents the change between consecutive "Rose" sales values in the training dataset.
- **Stationarity Goal:** This transformation helps eliminate trends, making the data more suitable for time series modeling.
- **Zero-Centered Fluctuations:** The differenced values hover around zero, suggesting that the trend has been largely removed.
- **Persistent Volatility:** Despite trend removal, the data still shows significant fluctuations, indicating other variations remain.
- **Y-Axis Clarification:** The Y-axis should reflect "Change in Wine Units Sold" instead of raw sales to avoid misinterpretation.

ARIMA Model



```
                                SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 23 Feb 2025   AIC                           1274.695
Time:                        11:34:05   BIC                           1291.946
Sample:                    01-31-1980   HQIC                          1281.705
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6779      0.084    -20.039      0.000      -1.842      -1.514
ar.L2         -0.7287      0.084     -8.701      0.000      -0.893      -0.565
ma.L1          1.0445      0.630      1.657      0.098      -0.191       2.280
ma.L2         -0.7721      0.132     -5.828      0.000      -1.032      -0.512
ma.L3         -0.9047      0.572     -1.583      0.113      -2.025       0.216
sigma2       859.2612    530.400      1.620      0.105    -180.304    1898.827
===================================================================================
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 14– Auto Arima results

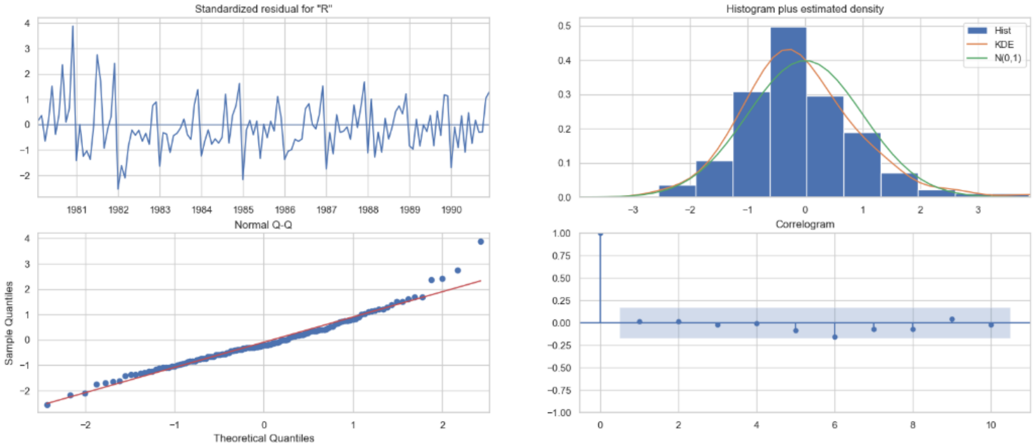


Figure 28 - Auto Arima diagnostic results

- **Model Fit:** The model is an ARIMA(2, 1, 3), indicating 2 autoregressive terms, 1 difference, and 3 moving average terms. The log-likelihood is -631.348, and AIC, BIC, and HQIC are reported.
- **Coefficient Significance:** The coefficients for ar.L1, ar.L2, and ma.L2 are statistically significant (P>|z| < 0.05). ma.L1 and ma.L3 are not statistically significant.

- **Residual Diagnostics:** The Ljung-Box test shows no significant autocorrelation in the residuals (Prob(Q) = 0.88). The Jarque-Bera test indicates non-normal residuals (Prob(JB) = 0.00).
- **Heteroskedasticity:** The Heteroskedasticity test indicates significant heteroskedasticity (Prob(H) = 0.00), meaning the variance of the residuals is not constant.
- **Top Left: Standardized Residuals:** The residuals fluctuate around zero, suggesting no obvious trend or pattern, but there are some periods of higher volatility, especially in the early years.
- **Top Right: Histogram and Density:** The histogram of the residuals shows a roughly bell-shaped distribution, but with slightly heavier tails than a normal distribution, as indicated by the KDE and N(0,1) lines.
- **Bottom Left: Normal Q-Q Plot:** The Q-Q plot shows some deviations from the red line, particularly at the tails, indicating that the residuals are not perfectly normally distributed.
- **Bottom Right: Correlogram:** The correlogram shows no significant autocorrelation in the residuals, as all lags fall within the confidence interval, suggesting the model has captured the temporal dependencies in the data.

Now lets plot auto Arima results on Training vs Test data –

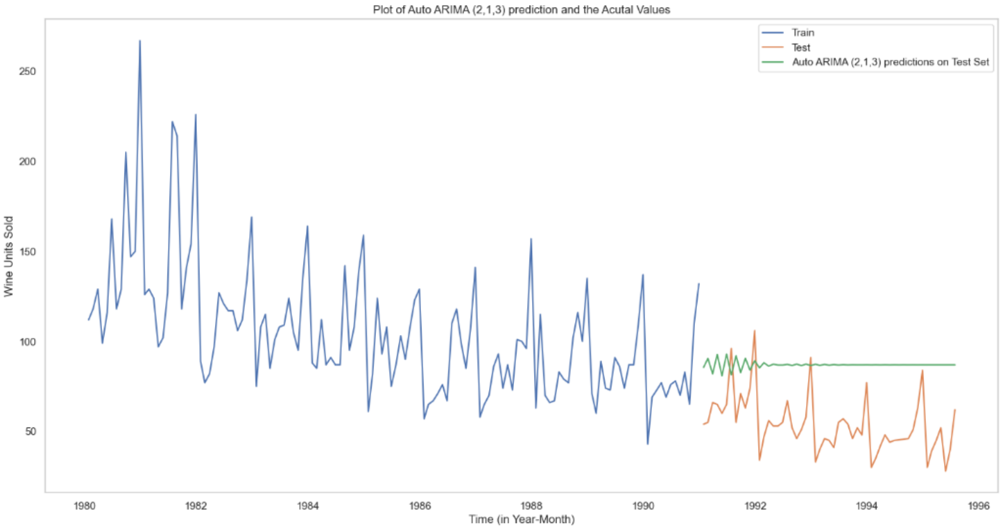


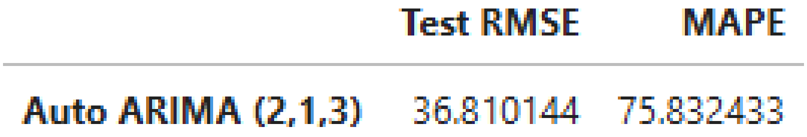
Figure 29 - Auto Arima on training vs Test data

- **Train-Test Split:** The plot clearly delineates the training data (blue line) from the test data (orange line), allowing for visual assessment of the model's predictive capabilities on unseen data.

- **Auto ARIMA Forecast:** The green line represents the Auto ARIMA(2,1,3) predictions on the test set, demonstrating a flat, constant forecast despite the actual test data's fluctuations.
- **Model Limitations:** The flat forecast indicates that the Auto ARIMA(2,1,3) model, in this instance, fails to capture the inherent volatility and potential seasonality present in the test data.
- **Poor Predictive Accuracy:** The significant divergence between the forecasted (green) and actual (orange) values in the test set highlights the model's poor predictive accuracy for this particular dataset.

Model evaluation –

As per code, we get –

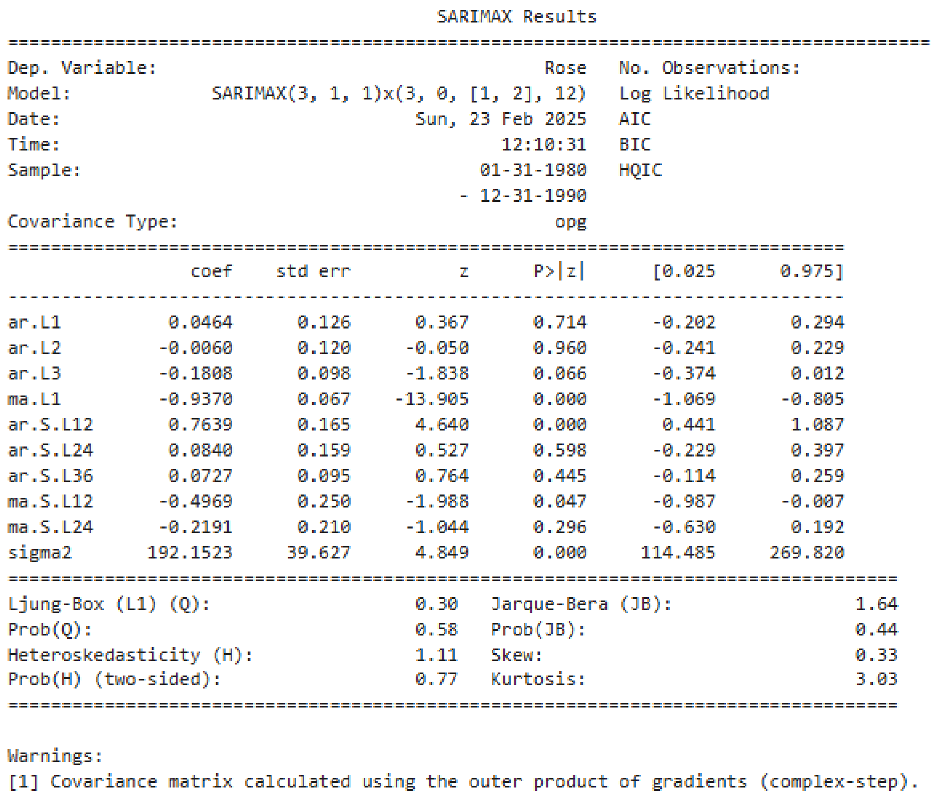


	Test RMSE	MAPE
Auto ARIMA (2,1,3)	36.810144	75.832433

Table 15– Auto Arima RMSE and MAPE

SARIMA Model

As we can observe seasonality in ACF plot.



```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  Rose   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)   Log Likelihood                -377.200
Date:                               Sun, 23 Feb 2025   AIC                            774.400
Time:                                       12:10:31   BIC                            799.618
Sample:                                   01-31-1980   HQIC                           784.578
                                        - 12-31-1990
Covariance Type:                                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0464      0.126      0.367      0.714      -0.202       0.294
ar.L2         -0.0060      0.120     -0.050      0.960      -0.241       0.229
ar.L3         -0.1808      0.098     -1.838      0.066      -0.374       0.012
ma.L1         -0.9370      0.067    -13.905      0.000      -1.069      -0.805
ar.S.L12       0.7639      0.165      4.640      0.000       0.441       1.087
ar.S.L24       0.0840      0.159      0.527      0.598      -0.229       0.397
ar.S.L36       0.0727      0.095      0.764      0.445      -0.114       0.259
ma.S.L12      -0.4969      0.250     -1.988      0.047      -0.987      -0.007
ma.S.L24      -0.2191      0.210     -1.044      0.296      -0.630       0.192
sigma2       192.1523     39.627      4.849      0.000     114.485     269.820
===================================================================================
Ljung-Box (L1) (Q):                   0.30   Jarque-Bera (JB):                 1.64
Prob(Q):                              0.58   Prob(JB):                         0.44
Heteroskedasticity (H):               1.11   Skew:                             0.33
Prob(H) (two-sided):                  0.77   Kurtosis:                         3.03
===================================================================================

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

Table 16– Auto Sarima results

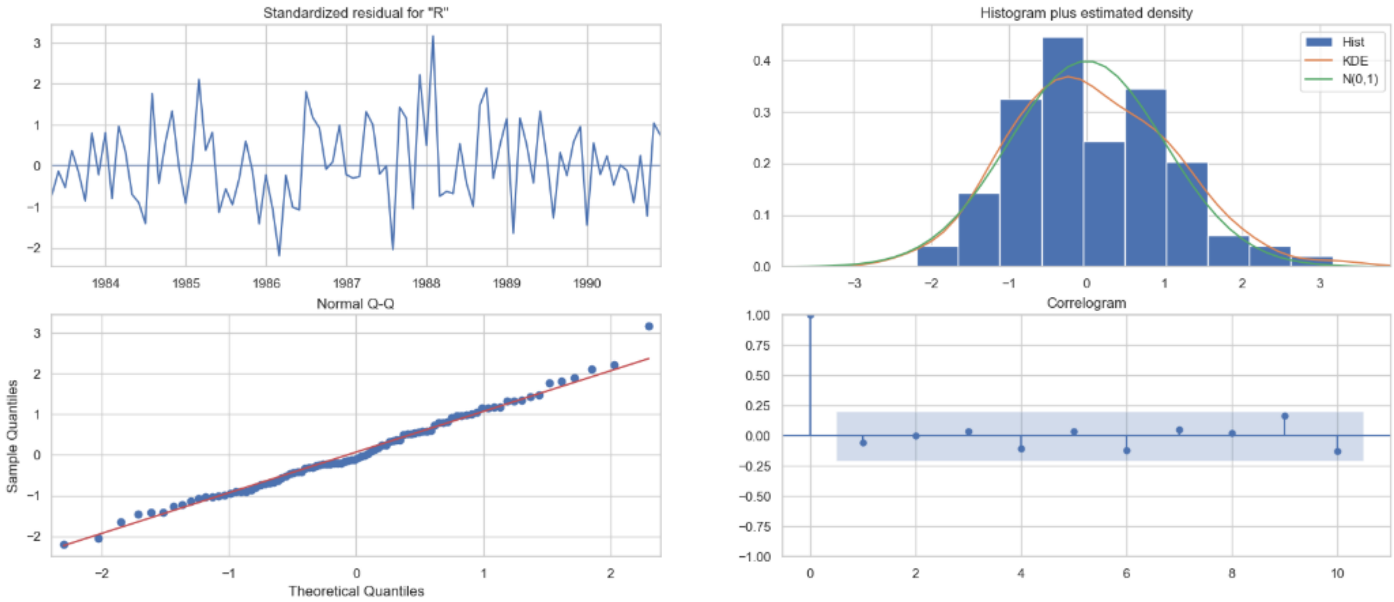


Figure 30 - Auto Sarima results diagnostic

- **Top Left: Standardized Residuals**
  - **Observation:** Residuals fluctuate around zero, showing no clear trend. However, periods of higher volatility are noticeable, especially in 1984-1985.
  - **Interpretation:** The model captures most of the systematic variation, but some unexplained volatility or heteroscedasticity remains.
- **Top Right: Histogram & Density Estimate**
  - **Observation:** The residual histogram is roughly bell-shaped but has slightly heavier tails compared to a normal distribution, as shown by the KDE and N(0,1) lines.
  - **Interpretation:** Residuals are approximately normal but deviate slightly, particularly in the tails.
- **Bottom Left: Normal Q-Q Plot**
  - **Observation:** Some deviations from the expected normal distribution occur, especially at the tails.
  - **Interpretation:** The residuals are not perfectly normal, indicating potential outliers or extreme values.
- **Bottom Right: Correlogram**
  - **Observation:** No significant autocorrelation is present, as all lags fall within the confidence interval.
  - **Interpretation:** The model effectively captures temporal dependencies, with no remaining autocorrelation in residuals.

Overall Interpretation

The model successfully explains most variations and dependencies in the data. However, residuals show slight deviations from normality, particularly in the tails. While this may not significantly impact performance, further analysis of the causes behind these heavier tails could be beneficial.

Now, lets predict on the Test Set using this model and evaluate the model.

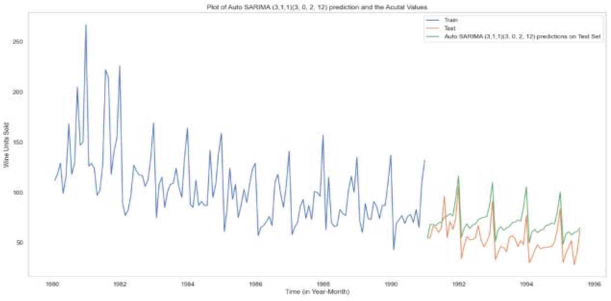
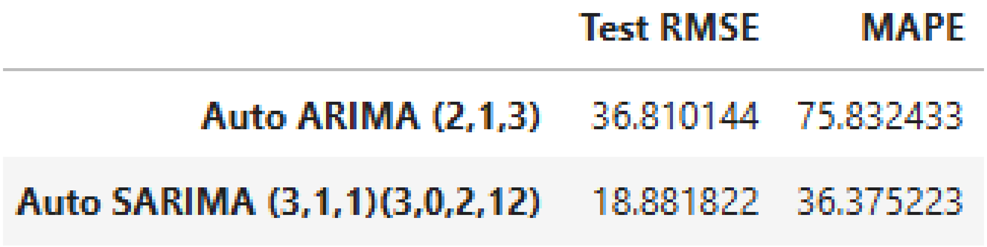


Figure 31 - Auto Sarima on Test set

- This plot shows the performance of an **Auto SARIMA(3,1,1)(3,0,2,12)** model in forecasting "Wine Units Sold":
- **Train-Test Split:** The plot clearly distinguishes between the training data (blue line) and the test data (orange line), allowing for visual assessment of the model's predictive capabilities on unseen data.
- **SARIMA Forecast:** The green line represents the Auto SARIMA(3,1,1)(3,0,2,12) predictions on the test set, demonstrating a model that captures both trend and seasonal patterns.
- **Model Fit:** The SARIMA model's predictions closely follow the actual test data, suggesting a good fit and indicating that the model has successfully captured the underlying time series dynamics.
- **Seasonal Pattern Capture:** The model adeptly captures the seasonal fluctuations present in the test data, showcasing the effectiveness of SARIMA models for time series with seasonality.
- **Improved Accuracy:** Compared to simpler models, the SARIMA model demonstrates improved accuracy in forecasting the "Wine Units Sold", especially in capturing the seasonal and trend components.

Model evaluation -



	Test RMSE	MAPE
Auto ARIMA (2,1,3)	36.810144	75.832433
Auto SARIMA (3,1,1)(3,0,2,12)	18.881822	36.375223

Table 17– Auto Sarima RMSE and MAPE

Manual ARIMA & SARIMA Model

Here, we have taken alpha=0.05.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag after which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0.

By looking at the above plots, we will take the value of p and q to be 2 and 2 respectively.

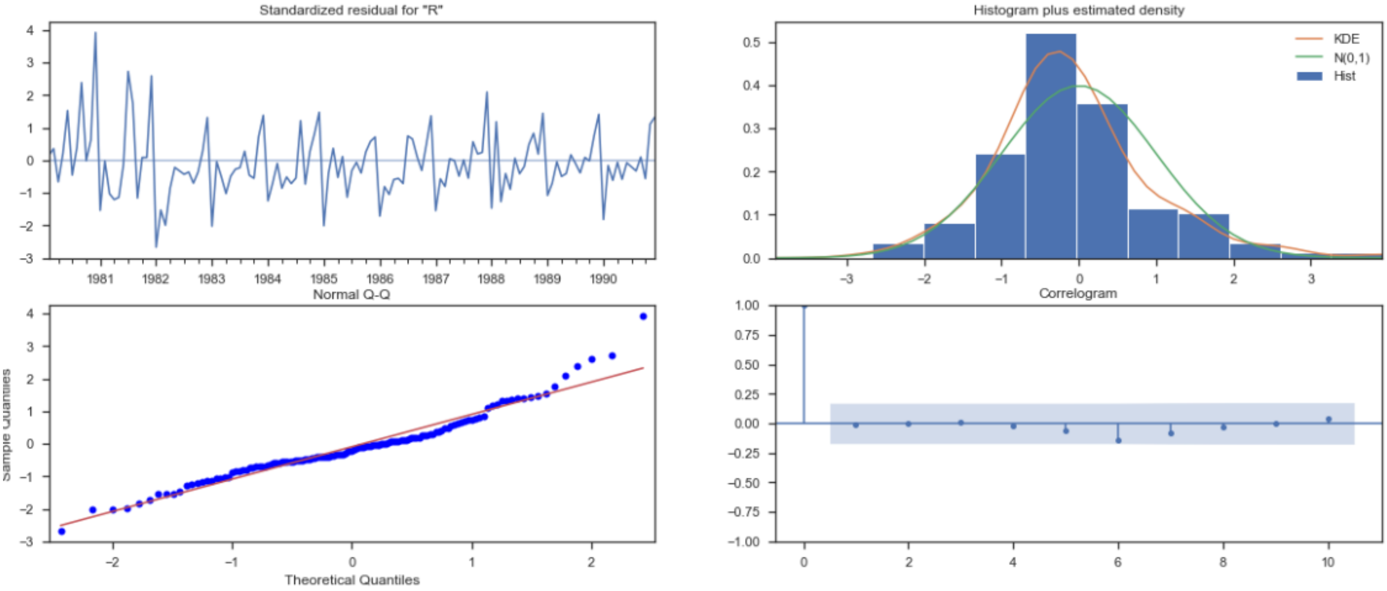


Figure 32 - Manual Arima results and Diagnostics

- **Standardized Residuals (Top Left):** The residuals fluctuate around zero, suggesting the model captures the mean well. However, volatility seems unevenly distributed, potentially indicating heteroscedasticity.
- **Histogram and Density (Top Right):** The histogram approximates a normal distribution, but with slightly heavier tails, suggesting some deviation from perfect normality.
- **Normal Q-Q Plot (Bottom Left):** The Q-Q plot shows deviations from the straight line, particularly in the tails, reinforcing the observation of non-normality in the residuals.
- **Correlogram (Bottom Right):** The correlogram shows no significant autocorrelation, indicating that the model has effectively captured the temporal dependencies in the data.

- **Overall:** While the model seems to have captured the time series structure well (no autocorrelation), the residuals exhibit some non-normality and potential heteroscedasticity, suggesting room for model improvement or further investigation.

Predict on the Test Set using this model and evaluate the model.

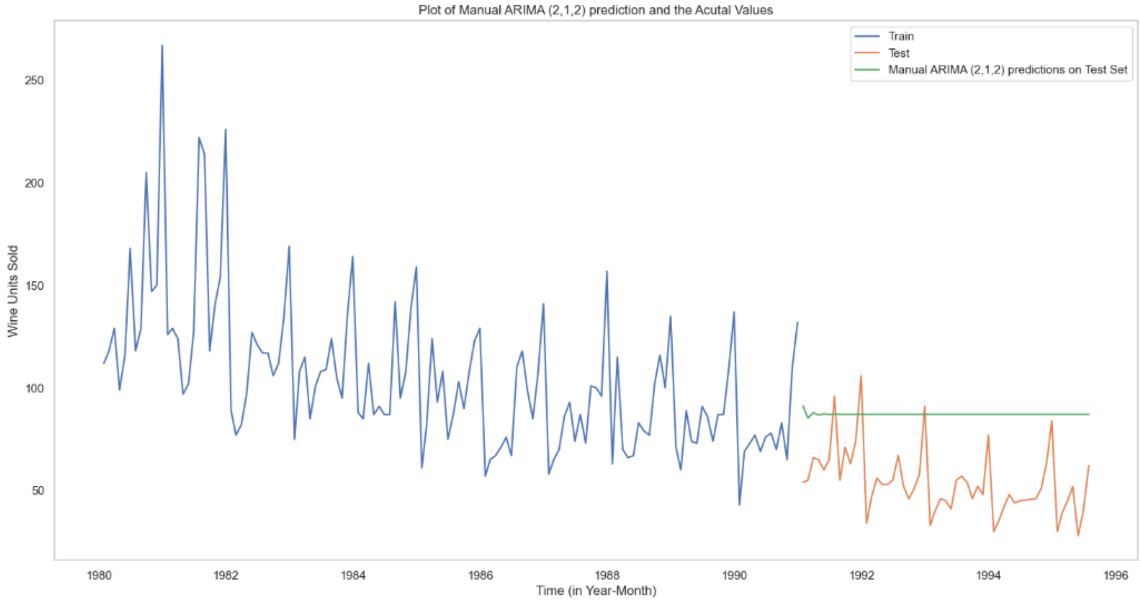
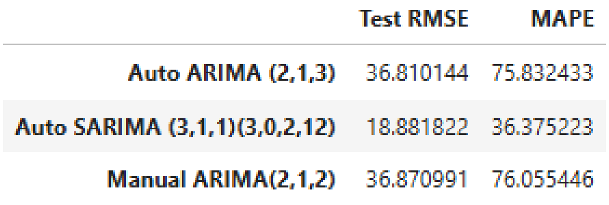


Figure 33 - Manual Arima model results on Train and Test data

- **Visualizes Train/Test Split:** The plot clearly distinguishes between the training (blue) and test (orange) data, allowing for a visual assessment of model performance.
- **Flat Forecast:** The green line, representing the Manual ARIMA(2,1,2) forecast on the test set, is nearly flat, indicating constant predictions.
- **Poor Performance:** The flat forecast significantly deviates from the actual test data's volatility, highlighting the model's poor ability to capture the underlying dynamics.
- **Model Limitations:** The ARIMA(2,1,2) model, with these parameters, fails to capture the trend, seasonality, and fluctuations observed in the test data.

Model evaluation Manual ARIMA–



	Test RMSE	MAPE
Auto ARIMA (2,1,3)	36.810144	75.832433
Auto SARIMA (3,1,1)(3,0,2,12)	18.881822	36.375223
Manual ARIMA(2,1,2)	36.870991	76.055446

Table 18– Manual arima RMSE and MAPE

Now lets see Manual SARIMA –

We see that our ACF plot at the seasonal interval (12) does not taper off quickly. So, we go ahead and take a seasonal differencing of the original series.

- Also, Here we have taken alpha = 0.05 and seasonal period as 12.
- From the PACF plot it can be seen that till lag 4 is significant before cut-off, so AR term ‘p = 4’ is chosen. At seasonal lag of 12, it cuts off, so keep seasonal AR ‘P = 0’.
- From ACF plot, lag 1 and 2 are significant before it cuts off, so lets keep MA term ‘q = 2’ and at seasonal lag of 12, a significant lag is apparent and no seaonal lags are apparent at lags 24, 36 or afterwards, so lets keep ‘Q = 1'.
- The final selected terms for SARIMA model is (4, 1, 2)x(0, 1, 1, 12), as inferred from the ACF and PACF plots.

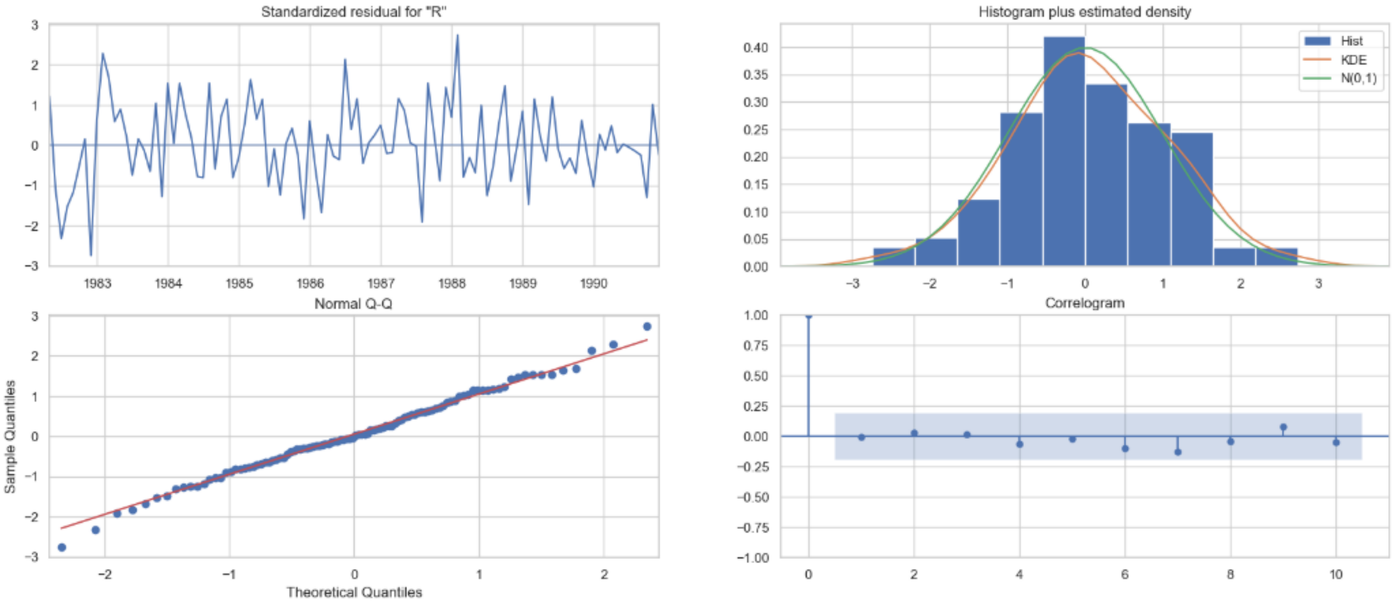


Figure 34 - Manual Sarima model results and Diagnostics

- **Standardized Residuals (Top Left):** The residuals fluctuate around zero, indicating the model captures the mean well, but show uneven volatility, suggesting potential heteroscedasticity.
- **Histogram and Density (Top Right):** The histogram approximates a normal distribution, but with slightly heavier tails, indicating some deviation from perfect normality.
- **Normal Q-Q Plot (Bottom Left):** The Q-Q plot shows deviations from the straight line, particularly in the tails, reinforcing the observation of non-normality in the residuals.
- **Correlogram (Bottom Right):** The correlogram shows no significant autocorrelation, indicating that the model has effectively captured the temporal dependencies in the data.

- **Overall:** While the model captures the time series structure well (no autocorrelation), the residuals exhibit some non-normality and potential heteroscedasticity, suggesting room for model improvement or further investigation.

Predict on the Test Set using this model and evaluate the model.

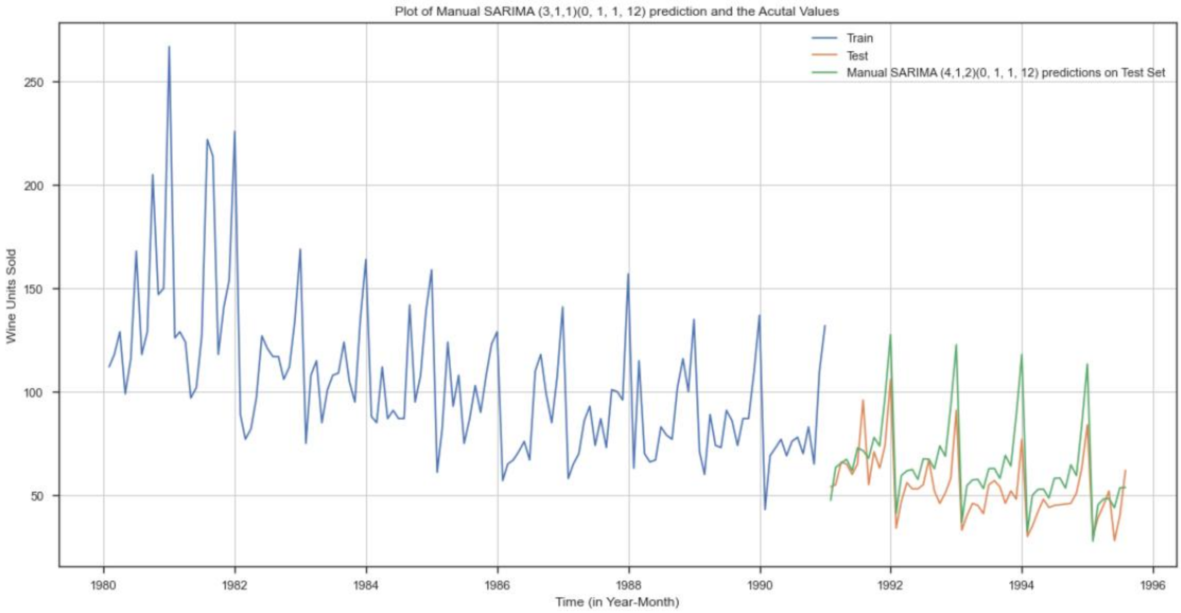
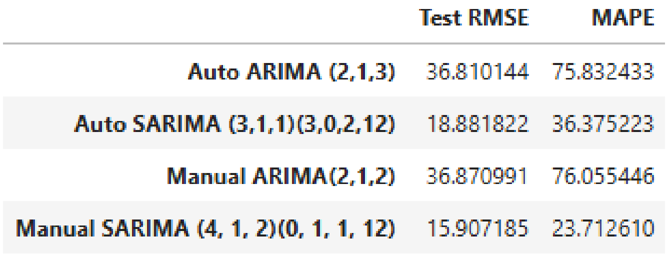


Figure 35 - Manual Sarima model results Test and Train datasets

Model evaluation Manual Sarima results –

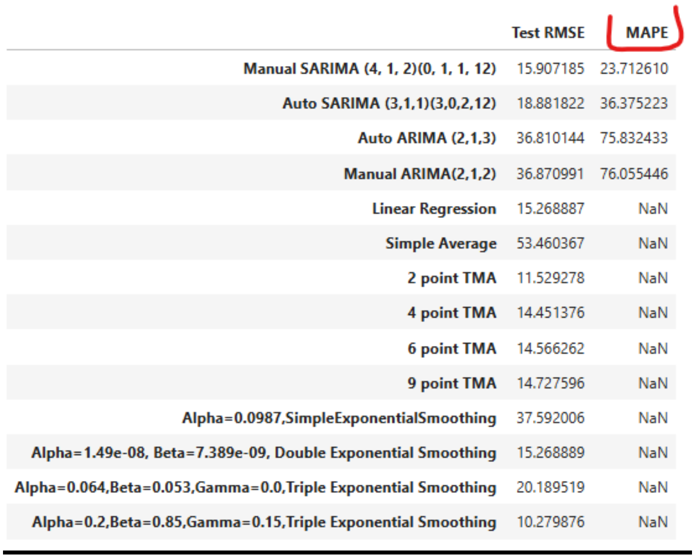


	Test RMSE	MAPE
Auto ARIMA (2,1,3)	36.810144	75.832433
Auto SARIMA (3,1,1)(3,0,2,12)	18.881822	36.375223
Manual ARIMA(2,1,2)	36.870991	76.055446
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	15.907185	23.712610

Table 19– Manual Sarima RMSE and MAPE

Final Comparison and Analysis –

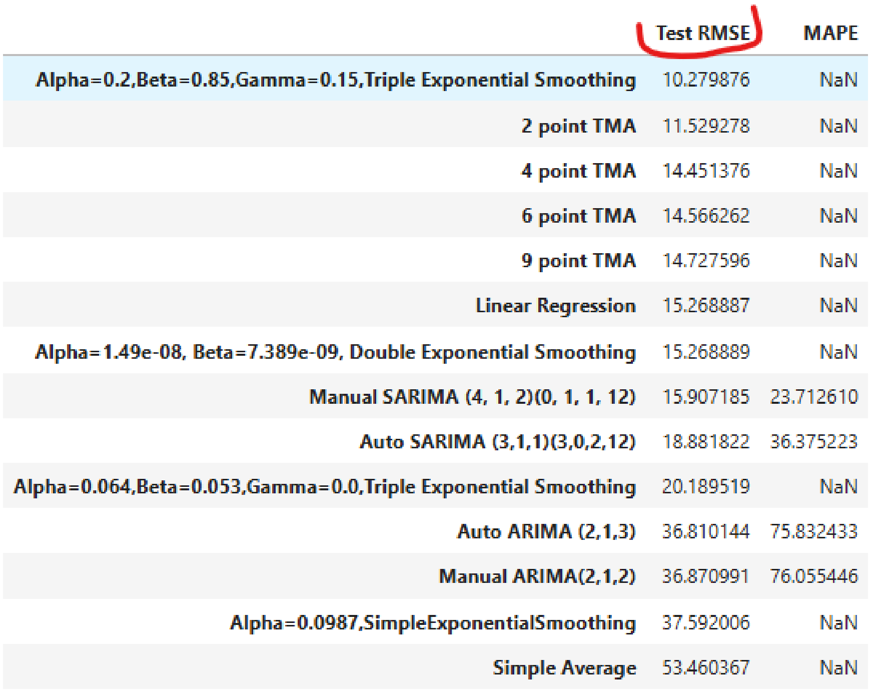
Mape values –



	Test RMSE	MAPE
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	15.907185	23.712610
Auto SARIMA (3,1,1)(3,0,2,12)	18.881822	36.375223
Auto ARIMA (2,1,3)	36.810144	75.832433
Manual ARIMA(2,1,2)	36.870991	76.055446
Linear Regression	15.268887	NaN
Simple Average	53.460367	NaN
2 point TMA	11.529278	NaN
4 point TMA	14.451376	NaN
6 point TMA	14.566262	NaN
9 point TMA	14.727596	NaN
Alpha=0.0987,SimpleExponentialSmoothing	37.592006	NaN
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.268889	NaN
Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing	20.189519	NaN
Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing	10.279876	NaN

Table 20– Sorted Mape values

RMSE values –



	Test RMSE	MAPE
Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing	10.279876	NaN
2 point TMA	11.529278	NaN
4 point TMA	14.451376	NaN
6 point TMA	14.566262	NaN
9 point TMA	14.727596	NaN
Linear Regression	15.268887	NaN
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.268889	NaN
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	15.907185	23.712610
Auto SARIMA (3,1,1)(3,0,2,12)	18.881822	36.375223
Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing	20.189519	NaN
Auto ARIMA (2,1,3)	36.810144	75.832433
Manual ARIMA(2,1,2)	36.870991	76.055446
Alpha=0.0987,SimpleExponentialSmoothing	37.592006	NaN
Simple Average	53.460367	NaN

Table 21– Sorted RMSE values

Building optimum model and predicting 12-month data-

From the above results we can see that Triple exponential model is the optimum model followed by Trailing moving average models. However lets take TES and Manual SARIMA and predict for the future.

Optimum Model - Triple Exponential Smoothing Model (Alpha=0.2, Beta=0.85, Gamma=0.15)

Please refer code.

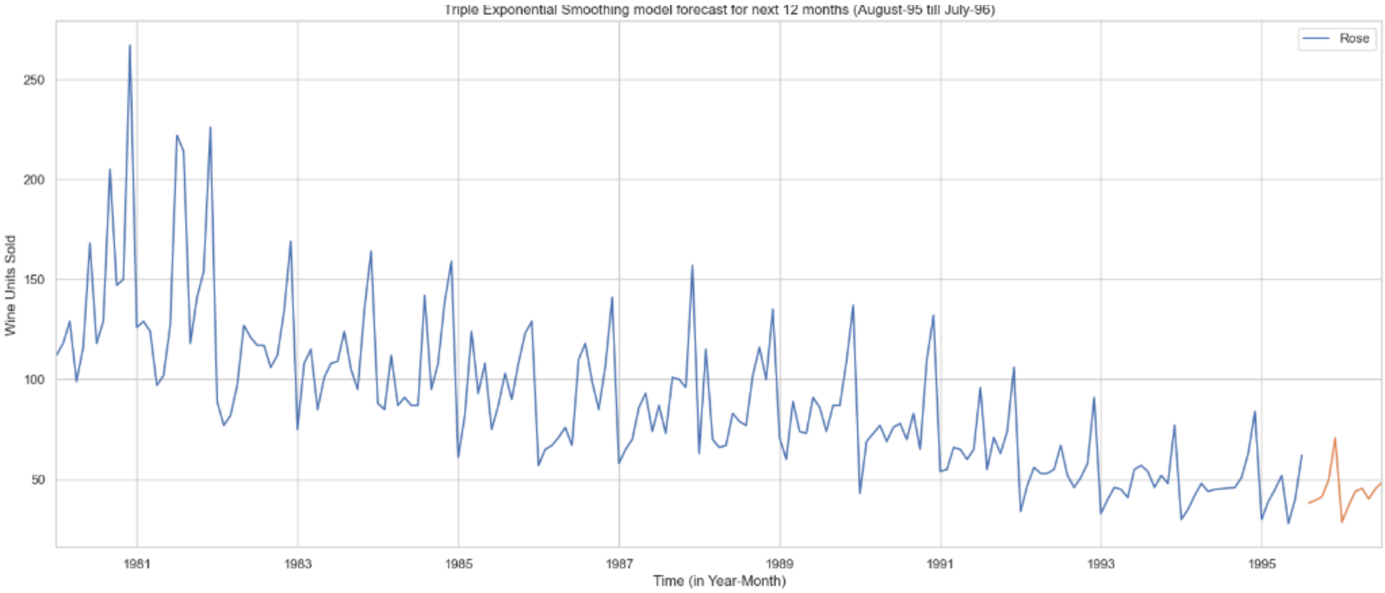


Figure 36 - TES for next 12 months

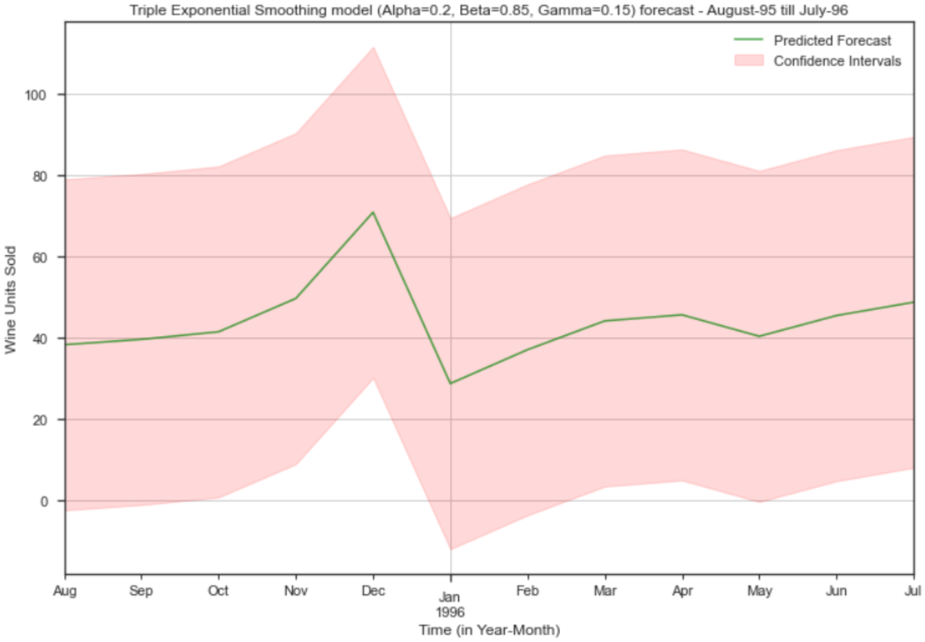


Figure 37 - TES model for next 12 months

This plot shows a **Triple Exponential Smoothing (TES) model forecast for the next 12 months (August 1995 to July 1996)**, including **confidence intervals**. Here's a breakdown:

- **Forecast Period:** The plot specifically focuses on the 12-month forecast period from August 1995 to July 1996.
- **Predicted Forecast (Green Line):** The green line represents the point forecasts generated by the TES model, using Alpha = 0.2, Beta = 0.85, and Gamma = 0.15.
- **Confidence Intervals (Pink Shaded Area):** The pink shaded area represents the confidence intervals around the point forecasts. This area indicates the range within which the actual values are likely to fall with a certain level of confidence (typically 95%).
- **Seasonal Pattern Capture:** The forecast shows a clear seasonal pattern, with peaks and troughs occurring at specific times of the year. This indicates that the TES model has successfully captured the seasonality in the data.
- **Confidence Interval Width:** The confidence intervals are wider at certain points, particularly around the peaks and troughs of the seasonal pattern. This suggests that the model is less certain about the exact values during these periods of higher variability.
- **Overall Trend:** The forecast suggests a relatively stable trend with seasonal fluctuations. There is a noticeable drop in sales predicted for January 1996.

Now,

Optimum Model - Manual SARIMA Model (4, 1, 2)(0, 1, 1, 12)

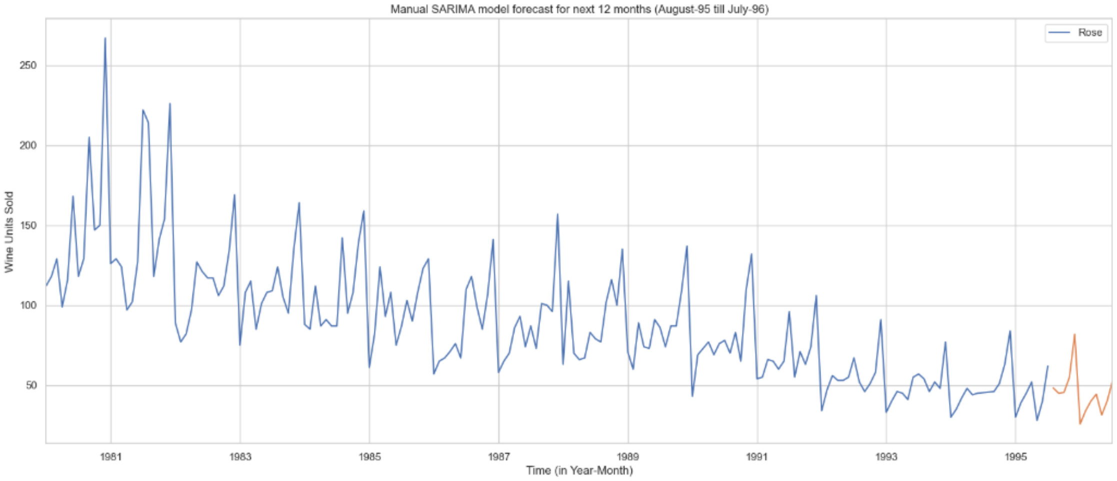


Figure 38 - Manual Sarima results for next 12 months

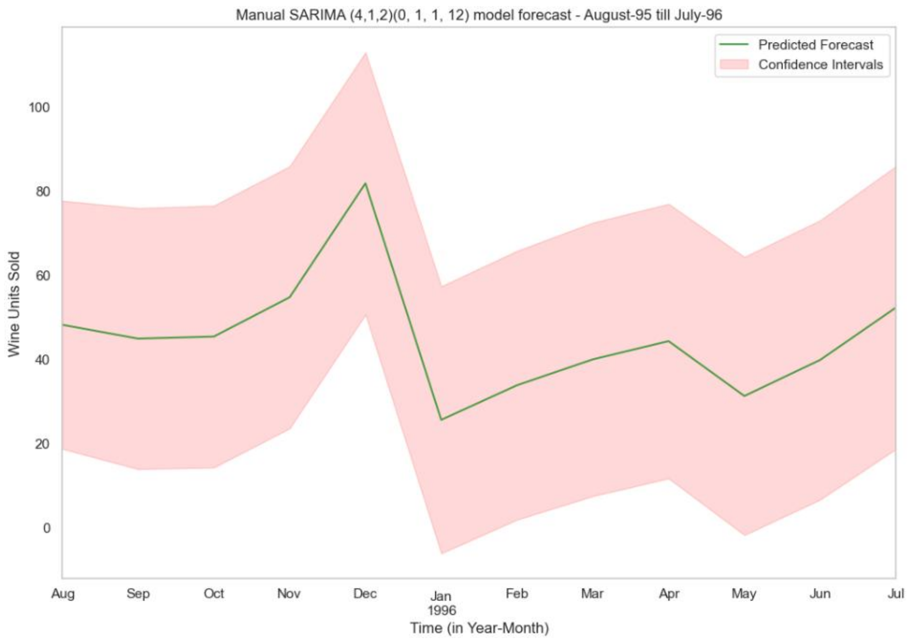


Figure 39 - Manual Sarima Plot for next 12 months

This plot displays a **Manual SARIMA(4,1,2)(0, 1, 1, 12) model forecast** for "Wine Units Sold" from August 1995 to July 1996, including confidence intervals:

- **Forecast Period:** The plot specifically focuses on the 12-month forecast period from August 1995 to July 1996.
- **Predicted Forecast (Green Line):** The green line represents the point forecasts generated by the SARIMA model.
- **Confidence Intervals (Pink Shaded Area):** The pink shaded area represents the confidence intervals around the point forecasts, indicating the range of likely values.
- **Seasonal Pattern:** The forecast shows a clear seasonal pattern, with a significant peak in December and a trough in January, reflecting the model's ability to capture seasonality with a period of 12.
- **Confidence Interval Width:** The confidence intervals vary in width, being wider around the peak and trough, indicating higher uncertainty during periods of greater fluctuation.

Hence it is evident that TES is the best model.

Actionable Insights and Recommendations:

Model Insights:

- The time series shows a declining trend with stable seasonality. Triple Exponential Smoothing and SARIMA models consistently perform best for forecasting, given their ability to handle both trend and seasonality.
- The Root Mean Squared Error (RMSE) is used to evaluate model performance. The model with the lowest RMSE is preferred.
- Triple Exponential Smoothing produced the lowest RMSE, making it the most suitable model for forecasting.

Historical Sales Insights:

- Rosé wine sales peaked in 1980-1981 but have since declined, reaching a low in 1995 (data available for only the first seven months).
- Seasonal trends show that sales increase toward the end of each year, with December having the highest sales and January the lowest.
- The average monthly sales are 90 units, with more than 50% of sales between 62 and 111 units.
- 70-75% of sales are below 100 units, while only 20% exceed 120 units. Large-volume purchases are rare, as 90% of sales are below 150 units.

Forecast Insights:

- The model predicts average monthly sales of 44 units, marking a 50% decline from historical averages.
- Minimum sales are projected at 28 units, unchanged from past trends.
- Maximum sales are expected to be 70 units, a 73% drop from the previous high of 267 units.
- Forecasted sales volatility is lower, with the standard deviation dropping from 62 to 10 units (83% decrease).

- October to December show peak sales, while January sees a sharp decline before a gradual recovery leading up to October.

Recommendations to Improve Sales

1. Capitalizing on Seasonal Demand

- 40% of sales occur from September to December, coinciding with festivals and holiday shopping.
- Promotional campaigns should focus on Thanksgiving, Christmas, and New Year's, when people stock up on wine for celebrations and gifting.
- Bulk purchase discounts or free shipping incentives can encourage larger orders.

2. Enhancing Marketing and Promotions

- Target holiday shoppers with special deals and bundles, making Rosé wine an attractive gifting option.
- Offer free gifts (e.g., branded accessories, small wine samples) for orders above a threshold to enhance customer experience.
- E-commerce campaigns (flash sales, giveaways, and competitions) can increase engagement and reach a wider audience.
- Tailored marketing strategies for different customer segments can boost conversions.

3. Addressing the Off-Season Sales Decline (January-June)

- Conduct market research to identify factors impacting sales during this period.
- Introduce a more affordable version of Rosé wine to attract a broader customer base and sustain sales.
- Offer discounts or seasonal pairings (e.g., pairing Rosé with light summer dishes) to drive demand during off-peak months.

4. Improving Sales Forecasting & Market Research

- While the current model effectively tracks historical trends, additional factors could influence sales.

- Conduct in-depth research on external factors (economic conditions, consumer preferences, competitor pricing).
- Integrate market trends and customer behaviour data into forecasting models to improve prediction accuracy.
-

By leveraging these strategies, businesses can mitigate declining sales, enhance customer engagement, and maximize revenue throughout the year.