Time Series Forecasting PROJECT

Contents

EXECUTIVE SUMMARY

This report presents a comprehensive analysis of ABC Estate Wines' historical sales data spanning the 20th century, focusing on sparkling and rose wine varieties. By examining key sales trends, patterns, and influencing factors, we aim to provide actionable insights to guide strategic decision-making and optimize future sales strategies.

Through advanced data analytics and forecasting techniques, our findings highlight how consumer preferences, economic conditions, and market dynamics have shaped wine sales over the decades. Notably, seasonal patterns and external factors such as global events and technological advancements have significantly influenced the demand for sparkling and rose wines.

Leveraging these insights, we propose targeted strategies to capitalize on emerging market opportunities, enhance product positioning, and improve forecasting accuracy. By adopting data-driven approaches, ABC Estate Wines can sustain its competitive edge and position itself for growth in an ever-evolving wine industry.

INTRODUCTION

The wine industry has been a dynamic and competitive sector, influenced by shifting consumer preferences, economic trends, and cultural factors. As a prominent player, ABC Estate Wines has maintained a rich history of crafting quality wines that appeal to a diverse clientele. To remain competitive in the modern market, understanding historical sales trends and identifying future opportunities are imperative.

This project focuses on analyzing historical sales data from ABC Estate Wines for sparkling and rosé wines. Covering the entirety of the 20th century, the datasets provide a unique opportunity to uncover trends and patterns across decades, highlighting factors that have influenced wine consumption.

DATA DESCRIPTION

The CSV files "Rose.CSV" contains Rose wines sold from time period of (1980-01 to 1995-07).

The files contain two columns YearMonth and Rose wines sold.

YearMonth	Represents the year and month in which the sales were recorded
Rose	Number of wine units sold

Data Description

We read the data and checked for the data types present in the CSV file.

```
YearMonth     object
Rose         float64
dtype: object
```

Table 1: Data Types (Rose)

YearMonth column is not seen as a date object. So we converted it into index column after converting it into date-time format.

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Table 2: First 5 rows(Rose)

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Table 3: Bottom 5 rows(Rose)

The shape of the dataframe is (187, 1), 187 rows and 1 column as we have changed YearMonth as Index.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
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 4 : Basic Information of Data set(Rose)

Out of 187 total rows the basic information is showing 185 non null values which means two rows have null values or missing values.

Statistical Summary

	Rose
count	185.000
mean	90.395
std	39.175
min	28.000
25%	63.000
50%	86.000
75%	112.000
max	267.000

Table 5 : Statistical Summary of Data set(Rose)

- The variability in sales suggests that demand for rosé wine is influenced by factors such as seasonality, promotions, or external market conditions.
- Rosé wine sales ranged from 28 units (minimum) to 267 units (maximum) across the analyzed months, reflecting variability in demand.
- The average monthly sales of rosé wine were approximately 90.4 units, indicating a moderate level of overall demand during the analyzed period.
- The maximum sales of 267 units are significantly higher than the 75th percentile (112 units), suggesting potential outliers or periods of exceptional demand, such as holidays or special promotions.

Exploratory Data Analysis

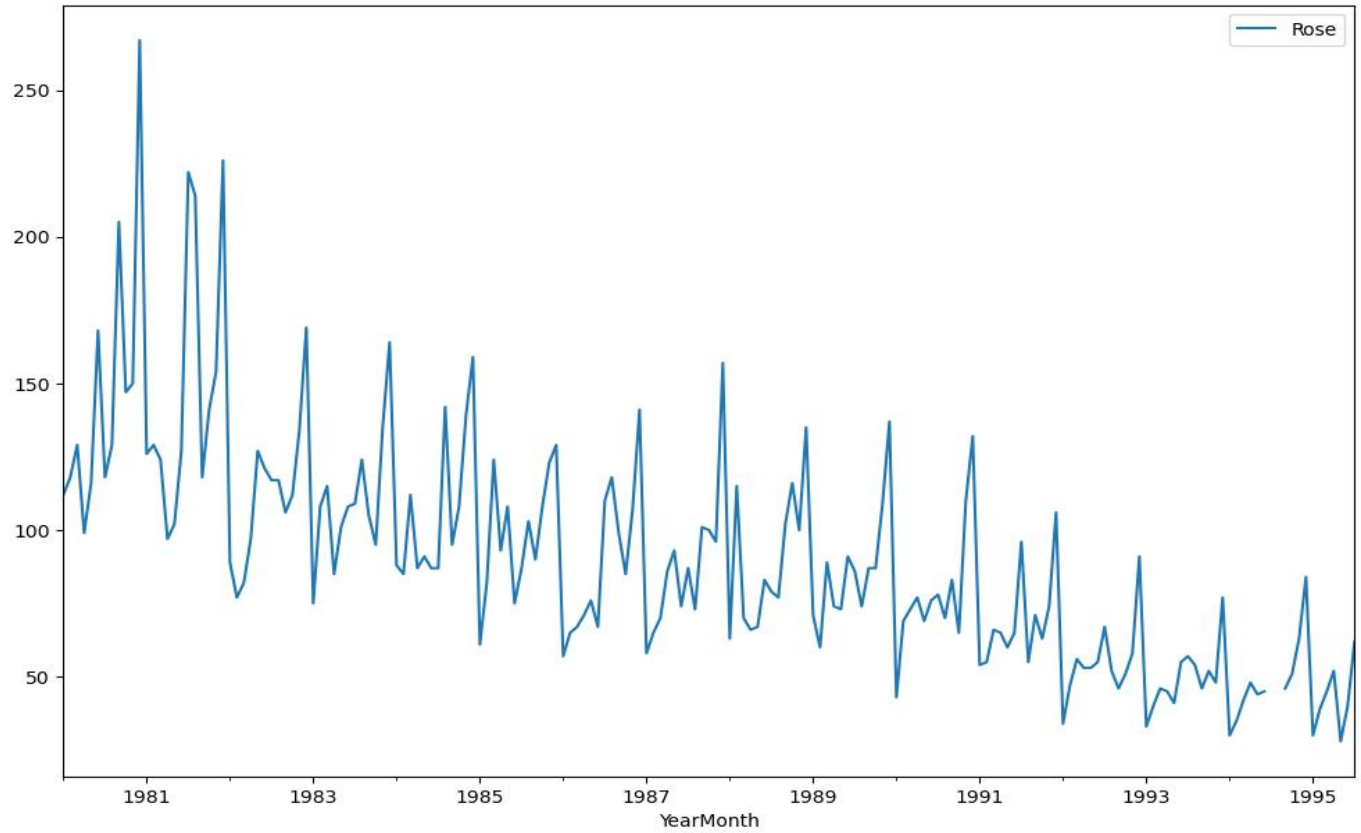

Figure 1 :Time Series(Rose)

We see an decreasing trend and seasonality which is not constant in nature.

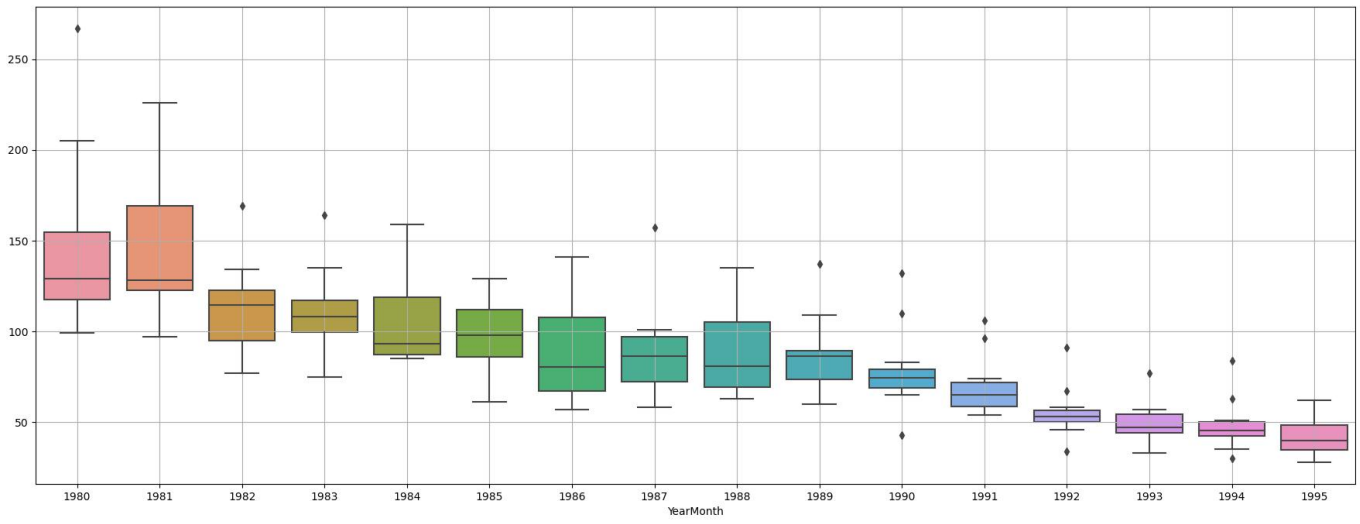

Figure 2 :Yearly Boxplot(Rose)

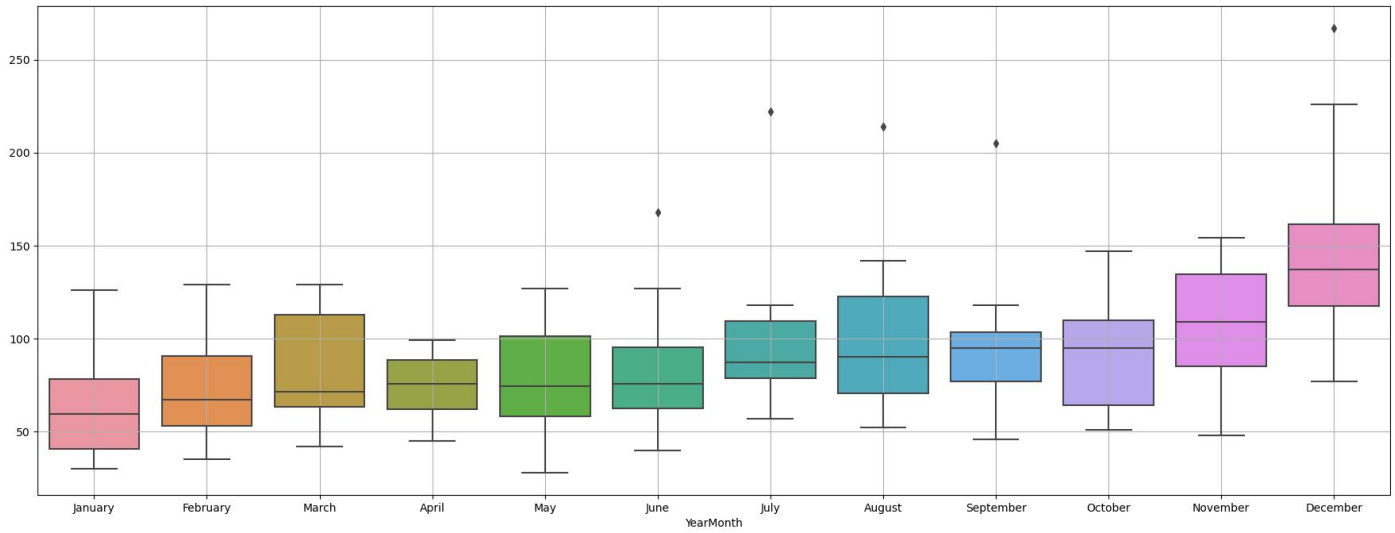

Figure 3 :Monthly Boxplot(Rose)

- Maximum sales in the month of December.
- The sales seems to usually pick in the last 4 months.
- Average order is greater in December and lowest in January.

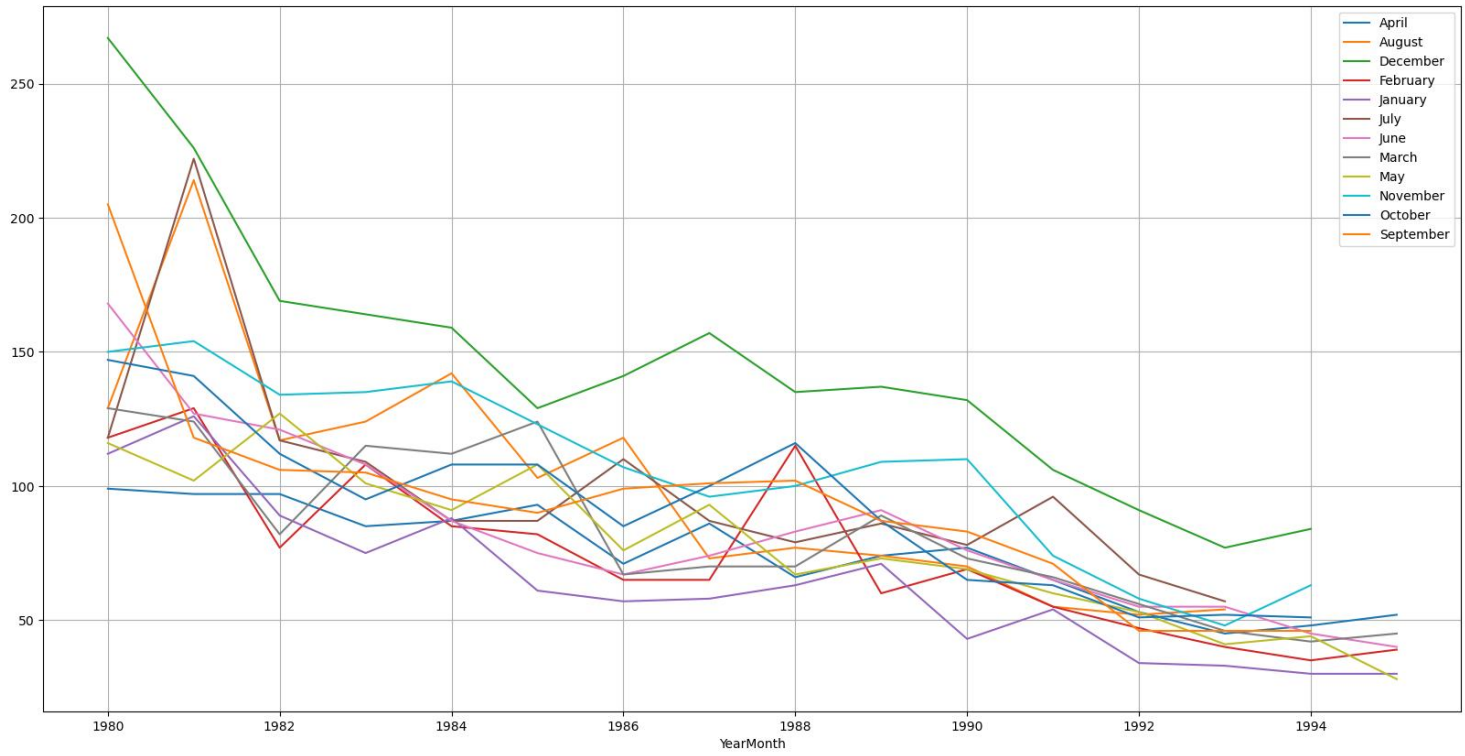

Figure 4 :Monthly Sales across Years(Rose)

- December consistently shows high sales in most years, indicating it may be a peak season due to holidays or festivities
- February consistently shows lower sales, likely due to seasonal or market factors.
- December consistently outperforms other months, suggesting it is a key period for marketing efforts.
- Months like January, February, and March exhibit relatively low sales across all years.

- The decline in sales is uniform across all months, suggesting no single month has drastically diverging trends.
- The sharp drop in sales from the 1980s to the 1990s suggests underlying challenges, such as reduced demand, competition, or economic factors. Investigating external market conditions or changes in company operations during this period might help.

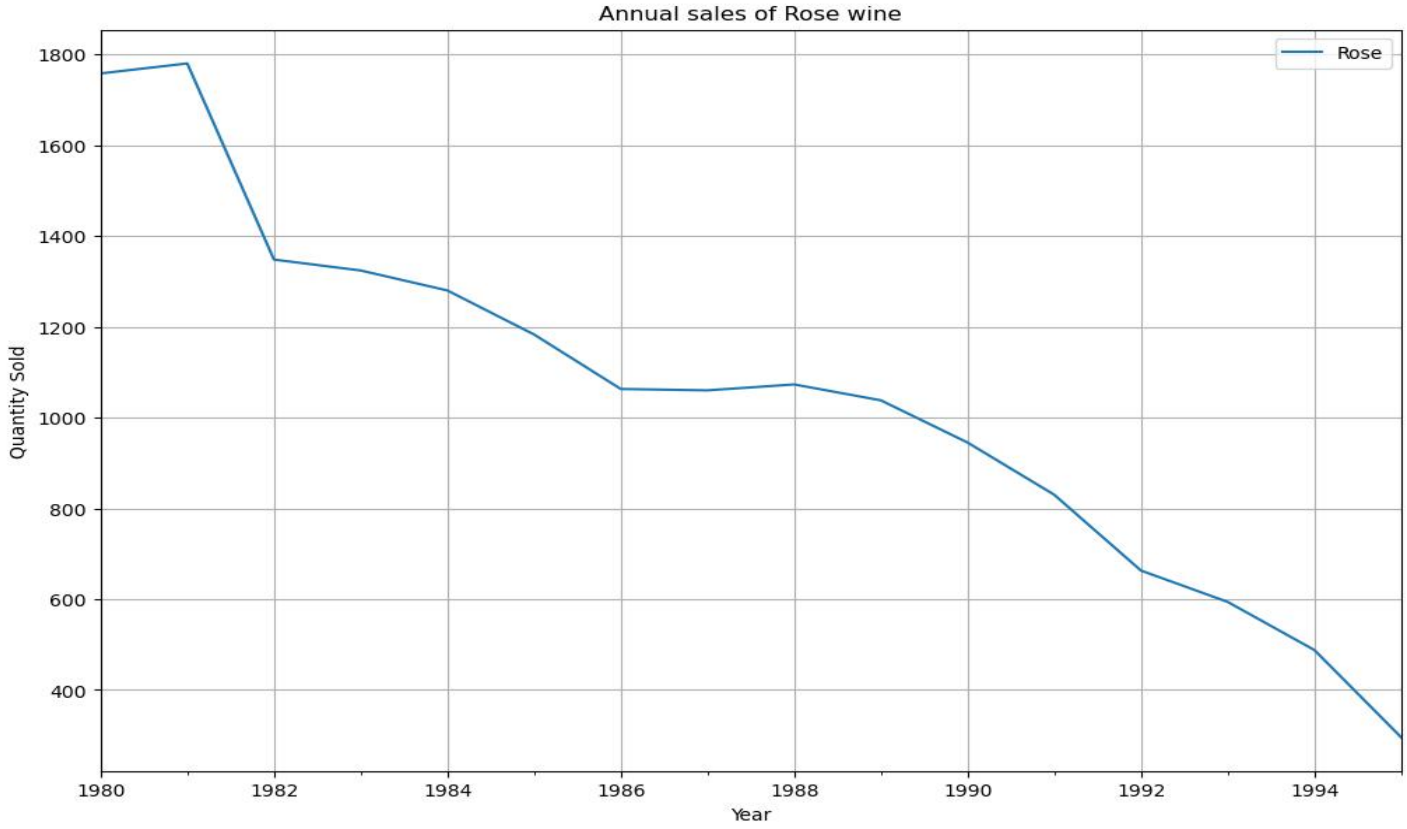

Figure 5 :Annual Sales (Rose)

- There is a consistent decline in annual sales from 1980 to 1995.
- Sales dropped from 1758 units in 1980 to just 296 units in 1995, representing an 83% decline over 15 years.
- The sharpest declines occurred in the early 1990s,this indicates that the early 1990s might have been particularly challenging for Rose wine sales, possibly due to market shifts or economic factors.
- Increased competition from other wine varieties or beverages could have impacted sales.Economic recessions or shifts in consumer spending habits during the early 1990s may have influenced the drop.

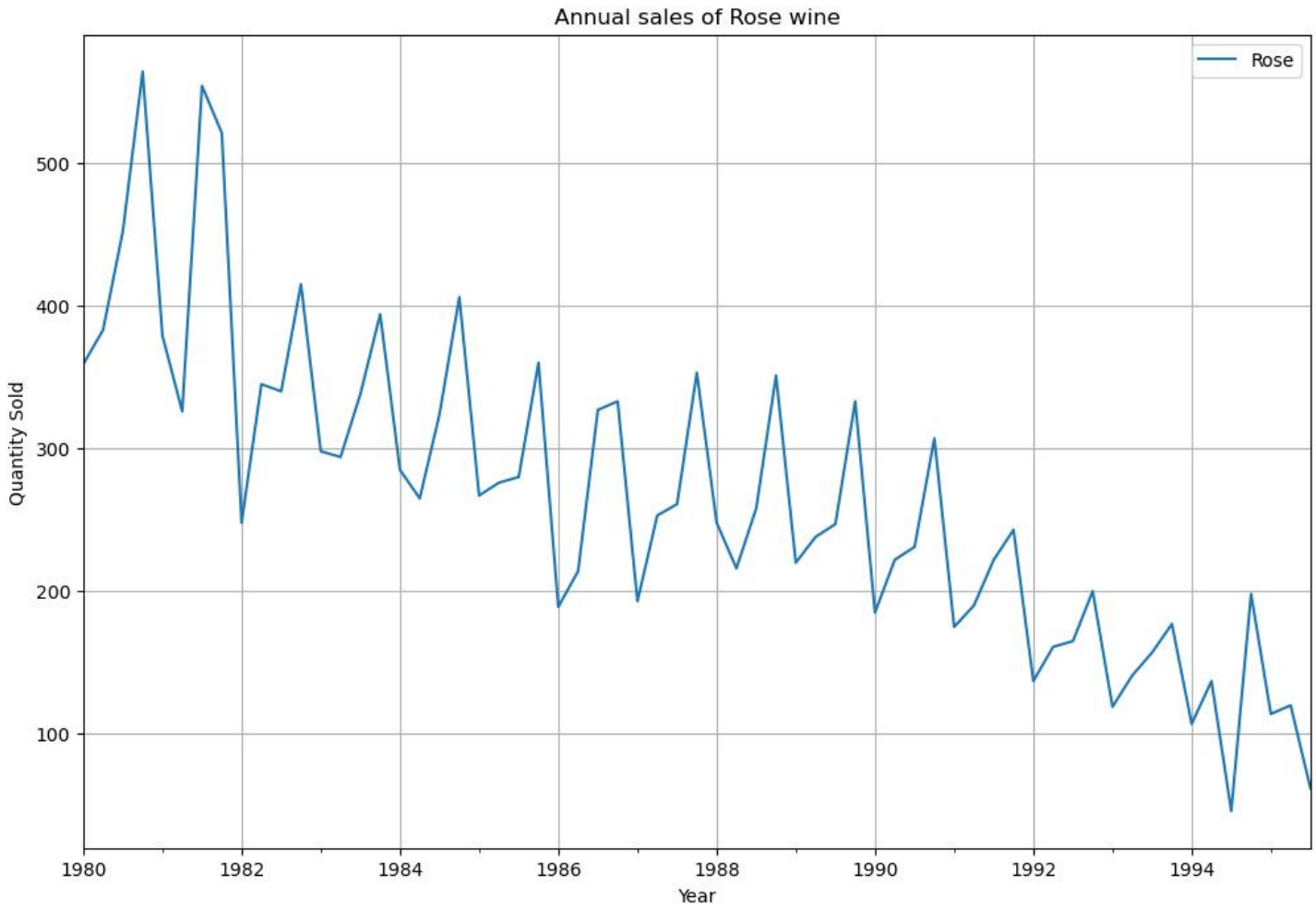

Figure 6 :Quarter Sales across Years (Rose)

- Q4 (October-December) consistently shows the highest sales in the year.
- Q2 (April-June) and Q1 (January-March) show comparatively lower sales, indicating these are weaker periods for Rose wine sales.
- Q2 (April-June) and Q1 (January-March) show comparatively lower sales, indicating these are weaker periods for Rose wine sales.

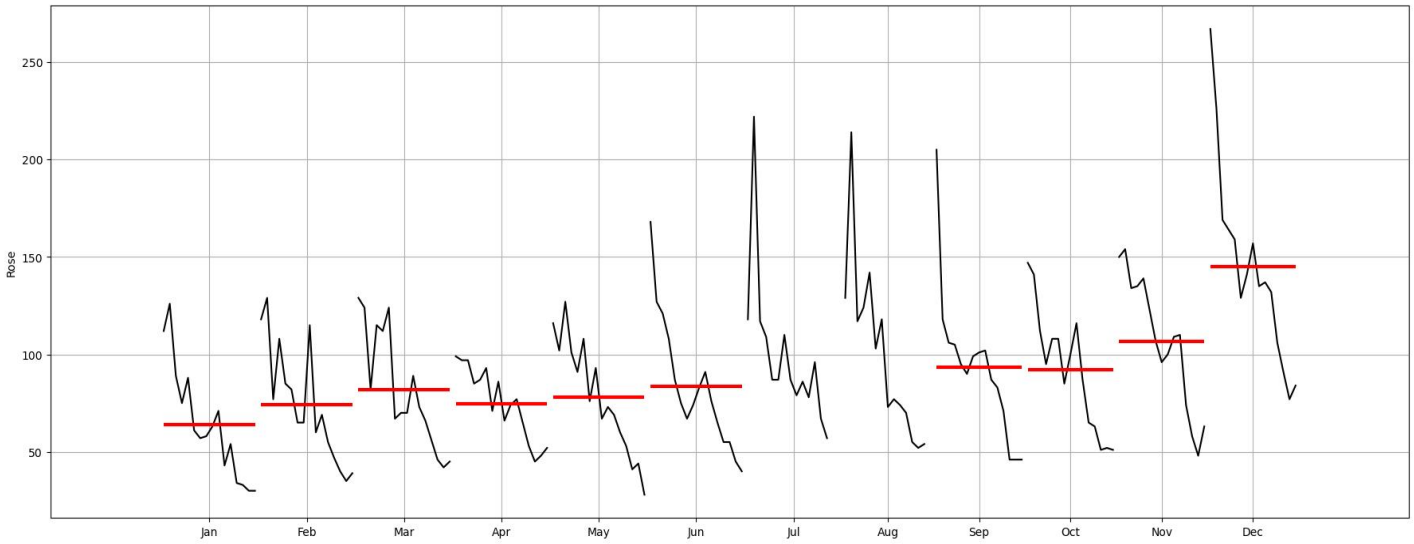

Figure 7 :Seasonal Patterns (Rose)

Missing values

Two rows have null value for Rose wine sold, the data is missing and we have to treat this.

YearMonth	Rose
1994-07-01	NaN
1994-08-01	NaN

Table 6 : Missing Values (Rose)

- The missing values for the months of July and August have been imputed with linear interpolation
- By interpolating missing values, the dataset becomes complete and ready for further analysis or forecasting.
- By interpolating missing values, the dataset becomes complete and ready for further analysis or forecasting.

Time Series Decomposition

Additive Model

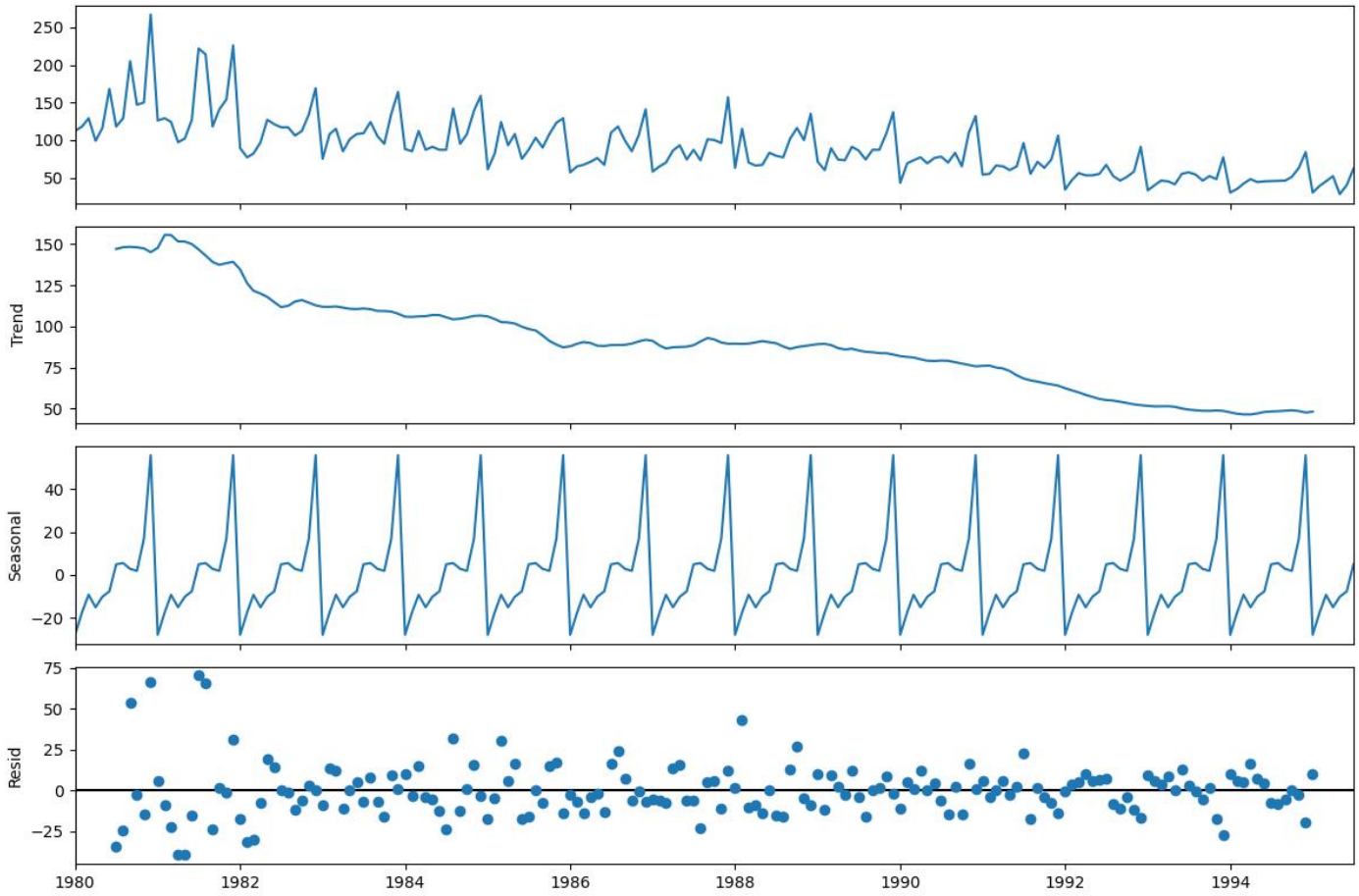

Figure 8 :Additive Model Decomposition(Rose)

- The trend represents the long-term movement in sales over time
  1)The trend captures sustained changes in consumer behavior or market conditions over time.
  2)This indicates a slight downward trend in overall sales.

- The seasonality captures the recurring patterns or fluctuations in sales due to seasonal effects
  1) The Rose sales data exhibits clear seasonality

- The residual represents the random noise or deviations not explained by the trend or seasonality.
  1)If we decompose a multiplicative series using an additive method, the error will continue o bear elements of seasonality.
  2)Residuals help identify anomalies or events that may require investigation.

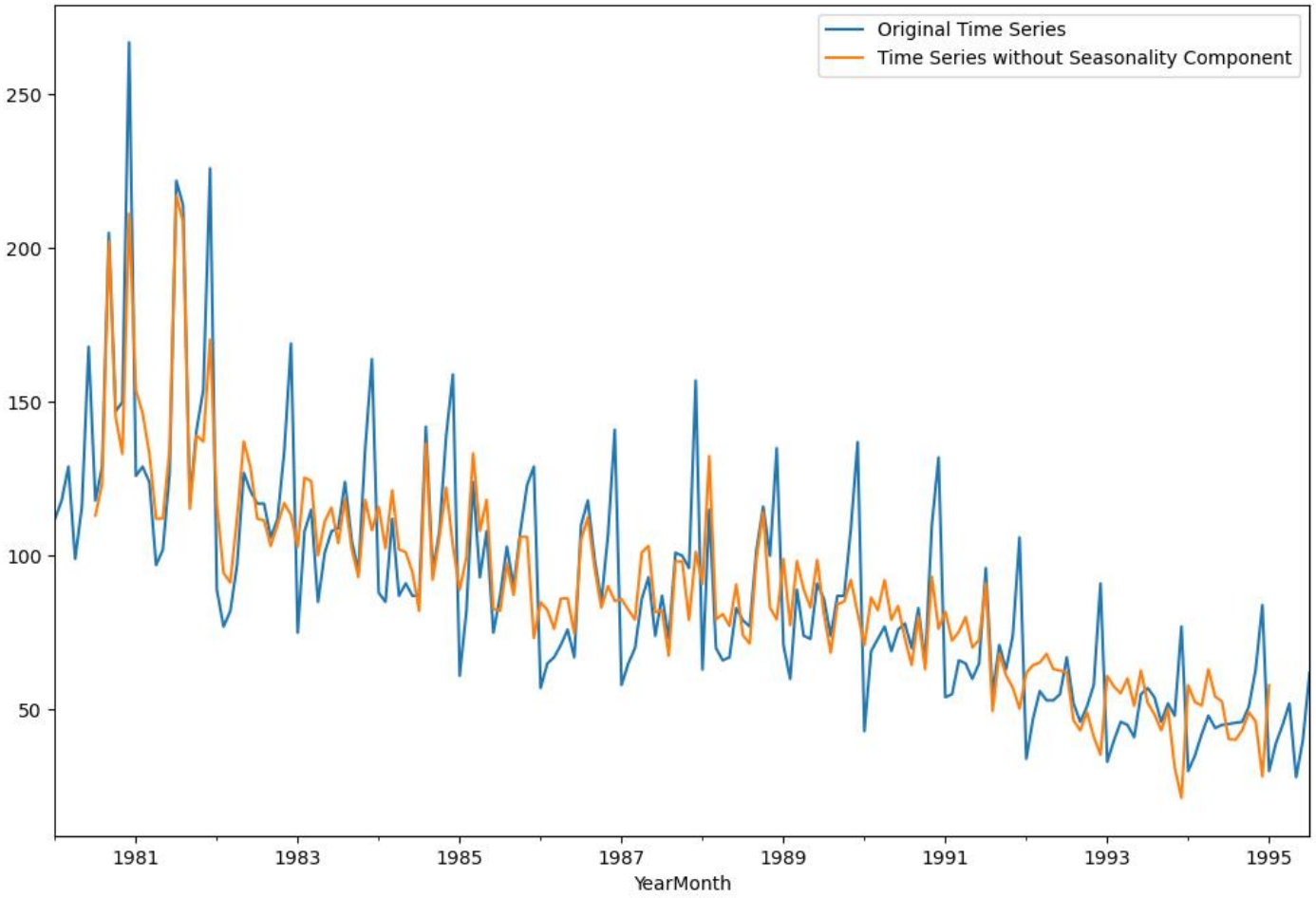

Figure 9 :Comparison of Original and Deseasonalized Rosé Wine Sales Time Series

Multiplicative Model

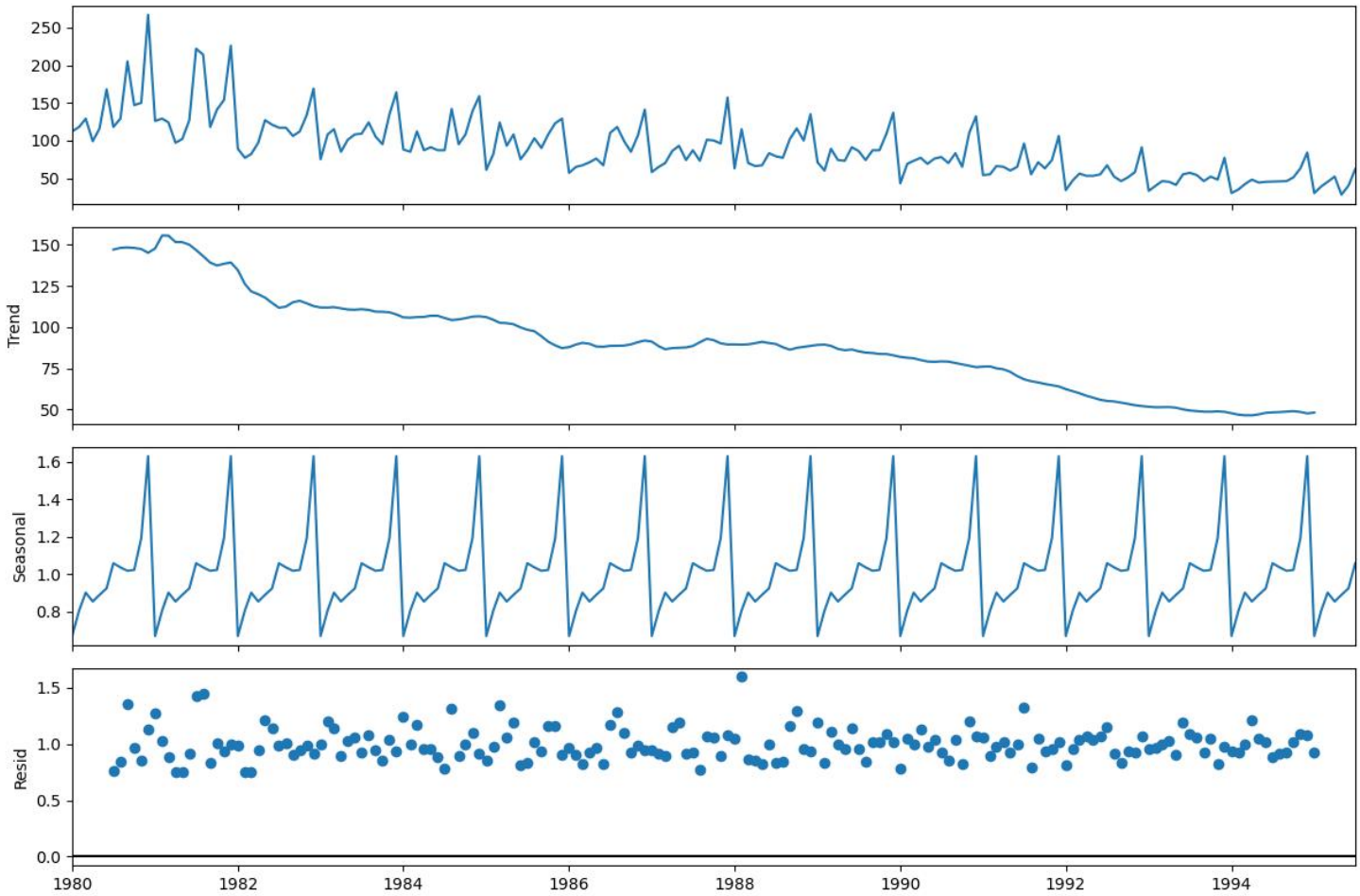

Figure 10 :Multiplicative Model Decomposition(Rose)

- Trend

  The trend is relatively stable with slight fluctuations, but there isn't a sharp exponential growth or decline. Still, the varying seasonal factors (multiplicative nature) seem more aligned with a multiplicative decomposition method.

- Seasonality

  The values for seasonality fluctuate around a base value of 1, which suggests that the seasonal variations are proportional to the trend (i.e., when the value of the trend increases, the seasonal effect also increases).

- Residual

  The residuals show fluctuations but they are relatively small and seem to be random noise. In multiplicative models, the residuals represent the variation that is left after accounting for trend and seasonality, and their behavior seems typical for a multiplicative decomposition.

As the trend appears to have relatively steady values with slight fluctuations, the seasonality is better represented as a proportional change to the trend. Multiplicative decomposition is appropriate for this data set.

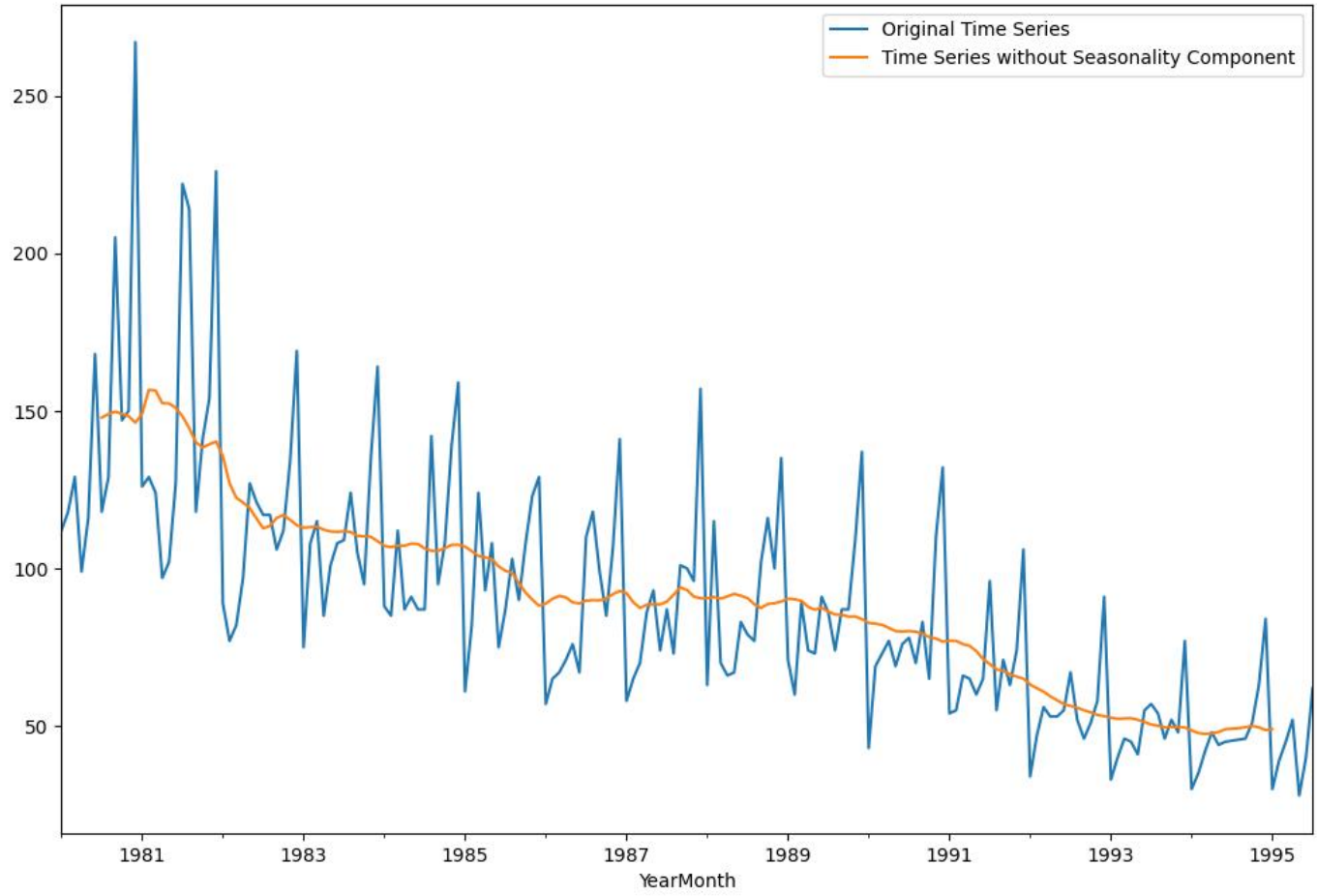

Figure 11 :Comparison of Original and Deseasonalized Rosé Wine Sales Time Series(Multiplicative)

Train-test split

- The data is split into two distinct subsets: training and test datasets. This division is essential for developing and evaluating predictive models.

- Training Set:This subset includes the first 70% of the available data. It is used to fit and train the model, allowing it to learn the underlying patterns and trends from the historical sales data.

- Test Set : The remaining 30% of the data is set aside as the test set. This portion is used to assess the model's performance and generalization ability. It simulates how the model will perform when applied to future, unseen data.

- By splitting the data in this way, we ensure that the model is trained on a large enough dataset, while still maintaining an independent portion for testing and validating its predictions. This process helps in assessing how well the model is likely to perform on real-world data, improving its reliability and robustness.

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Last few rows of Training Data

YearMonth	Rose
1990-06-01	76.0
1990-07-01	78.0
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0

First few rows of Test Data

YearMonth	Rose
1990-11-01	110.0
1990-12-01	132.0
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0

Last few rows of Test Data

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Table 7 : Training and Test Data (Rose)

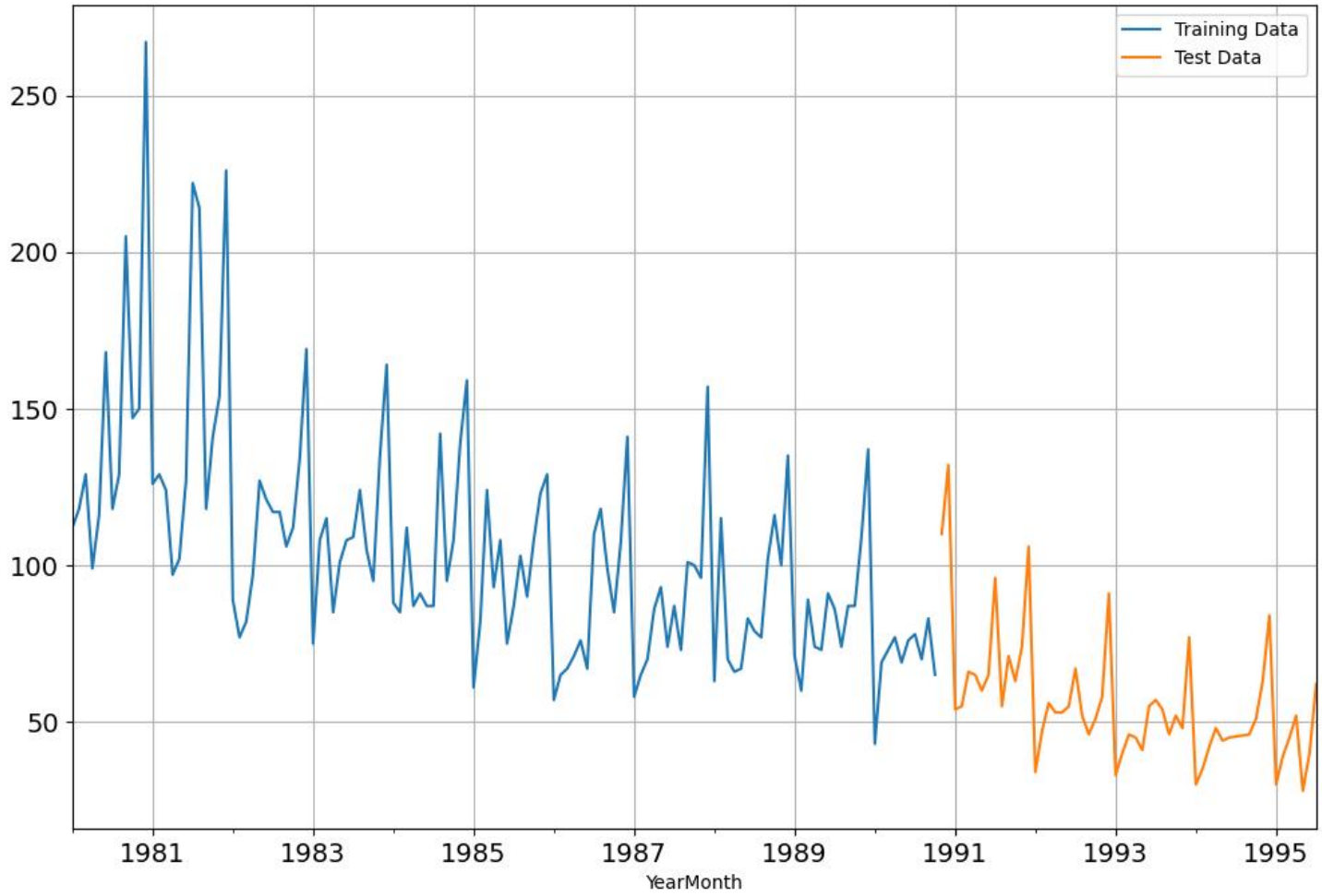

Figure 12 :Training and Test Split of Rose Wine Sales Data

Model Building - Original Data

1. Model 1: Linear Regression

Training Time instance

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130]

Test Time instance
[131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]

We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

First few rows of Training Data

YearMonth	Rose	time
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

Last few rows of Training Data

YearMonth	Rose	time
1990-06-01	76.0	126
1990-07-01	78.0	127
1990-08-01	70.0	128
1990-09-01	83.0	129
1990-10-01	65.0	130

First few rows of Test Data

YearMonth	Rose	time
1990-11-01	110.0	131
1990-12-01	132.0	132
1991-01-01	54.0	133
1991-02-01	55.0	134
1991-03-01	66.0	135

Last few rows of Test Data

YearMonth	Rose	time
1995-03-01	45.0	183
1995-04-01	52.0	184
1995-05-01	28.0	185
1995-06-01	40.0	186
1995-07-01	62.0	187

Table 8 : Training and Test Data Linear Regression (Rose)

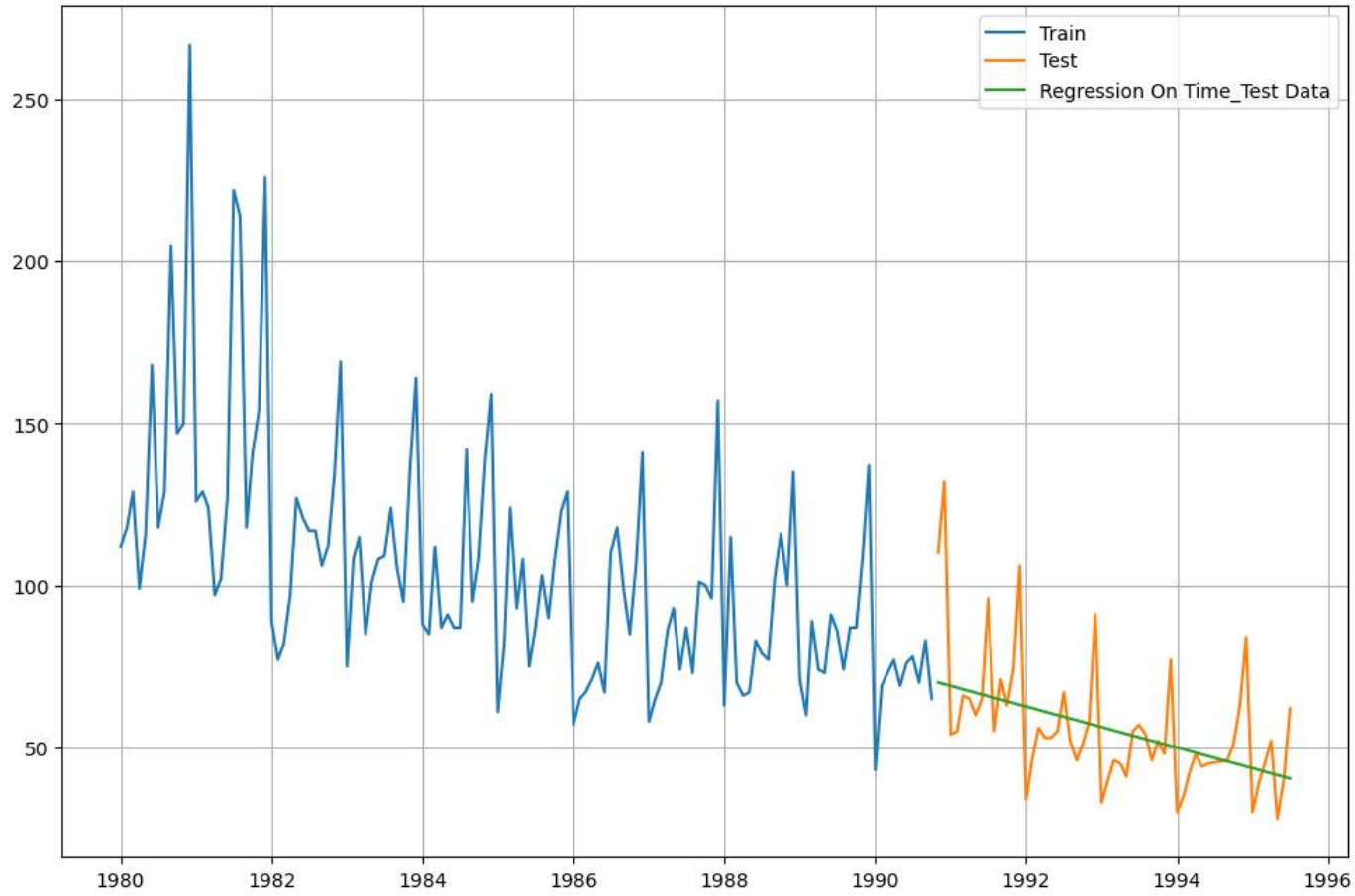

Figure 13 :Linear Regression Model Predictions for Rosé Wine Sales on Training and Test Data

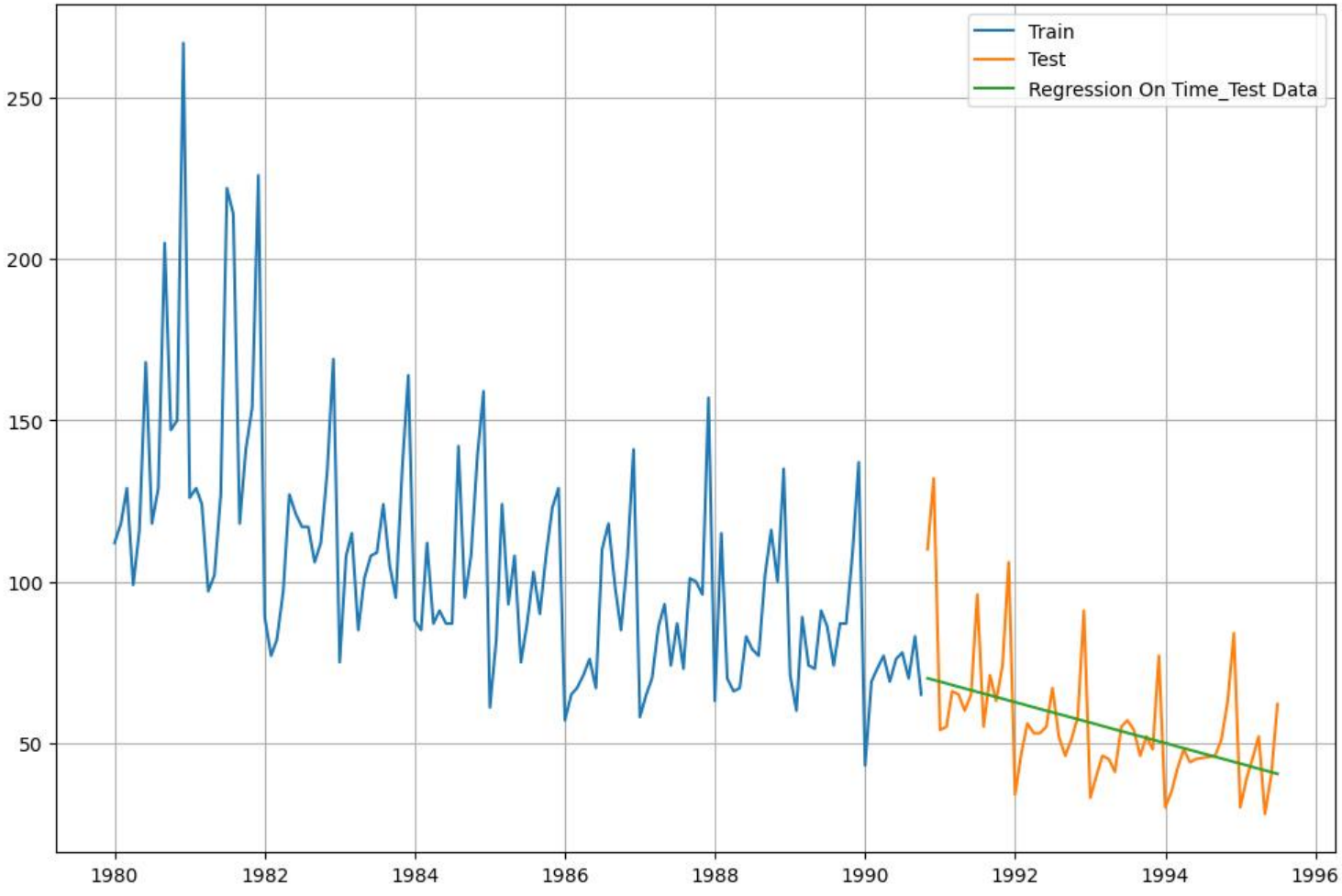

Figure 14 :Linear Regression Predictions and Accuracy Evaluation for Rosé Wine Sales

For Logistic regression forecast on the Test Data, RMSE is 17.356.

The RMSE for the RegressionOnTime forecast on the test data is 17.36. This means that, on average, the model's predictions are off by 17.36 units of wine sales (in the same scale as the sales data).

Model 2: Simple Average

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

YearMonth	Rose	mean_forecast
1990-11-01	110.0	104.692308
1990-12-01	132.0	104.692308
1991-01-01	54.0	104.692308
1991-02-01	55.0	104.692308
1991-03-01	66.0	104.692308

Table 9 : Mean forecast Rose

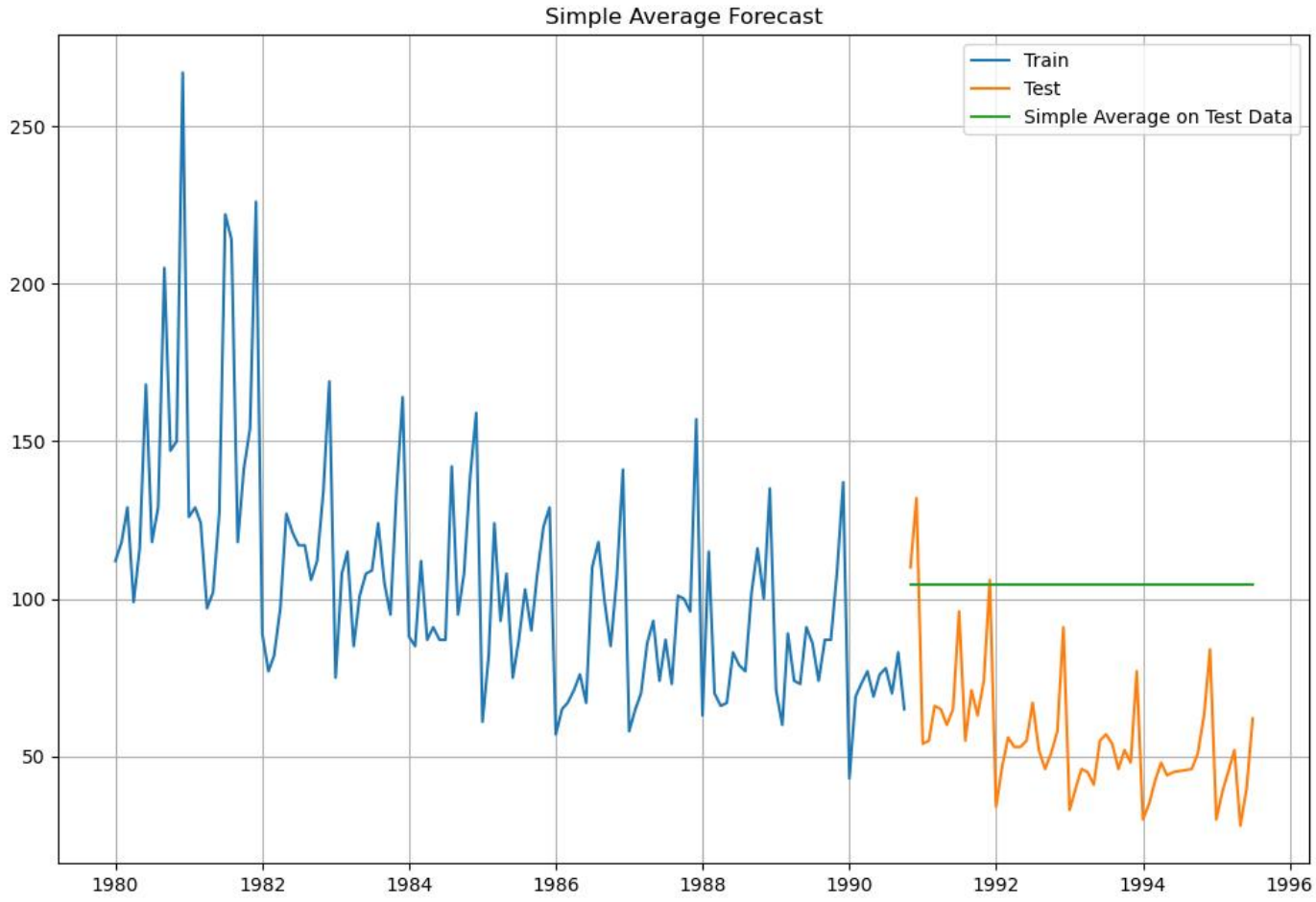

Figure 15 :Comparison of Actual and Simple Average Forecast for Rosé Wine Sales

For Simple Average forecast on the Test Data, RMSE is 52.412.

The RMSE for the Simple Average forecast on the test data is 52.4. This means that, on average, the model's predictions are off by 52.4 units of wine sales (in the same scale as the sales data).

Model 3: Moving Average(MA)

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.

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	114.5	NaN	NaN
1980-05-01	116.0	107.5	115.5	NaN	NaN

Table 10 : Trailing moving averages

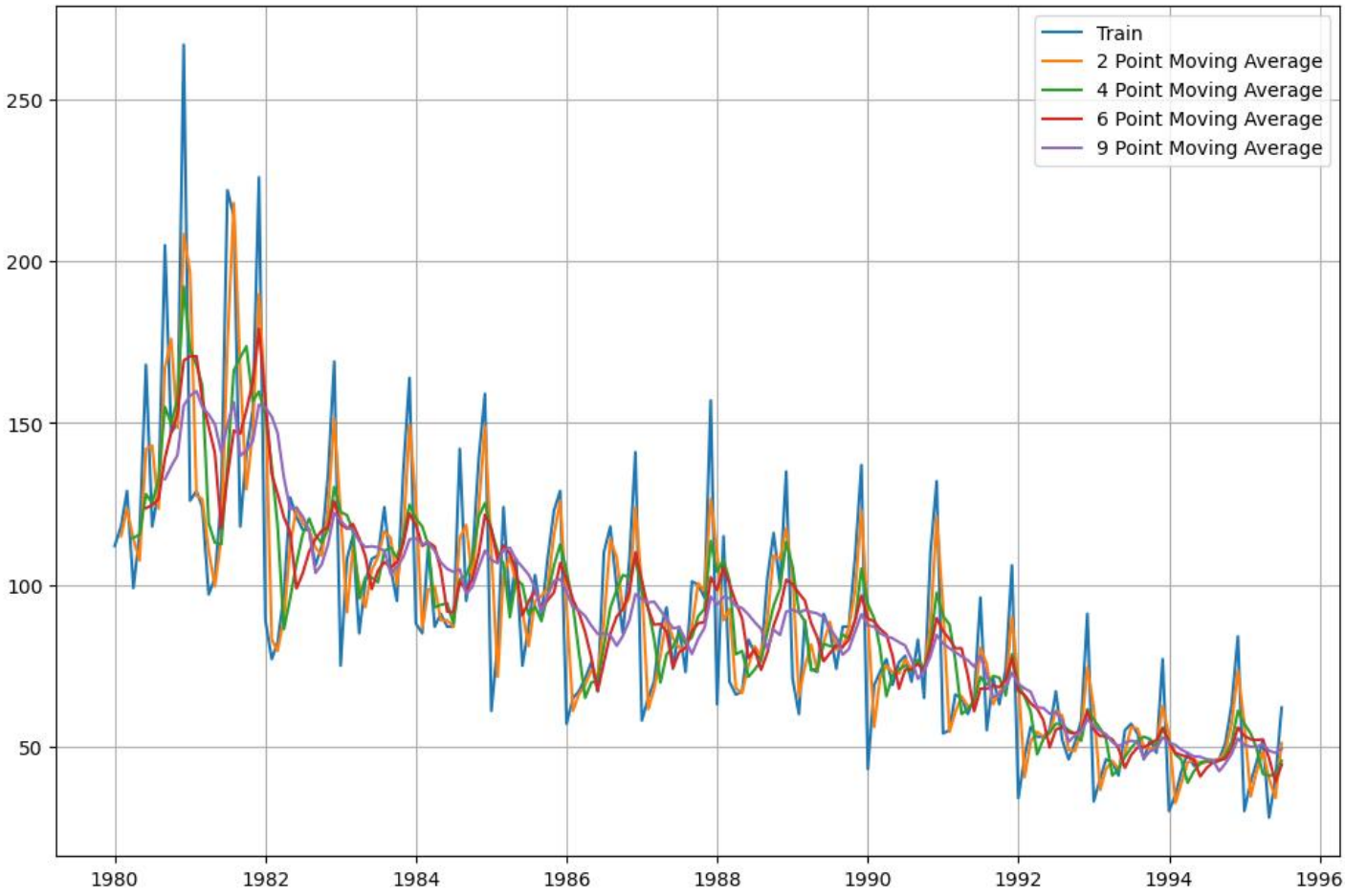

Figure 16 :Comparison of Rosé Wine Sales with Different Moving Averages

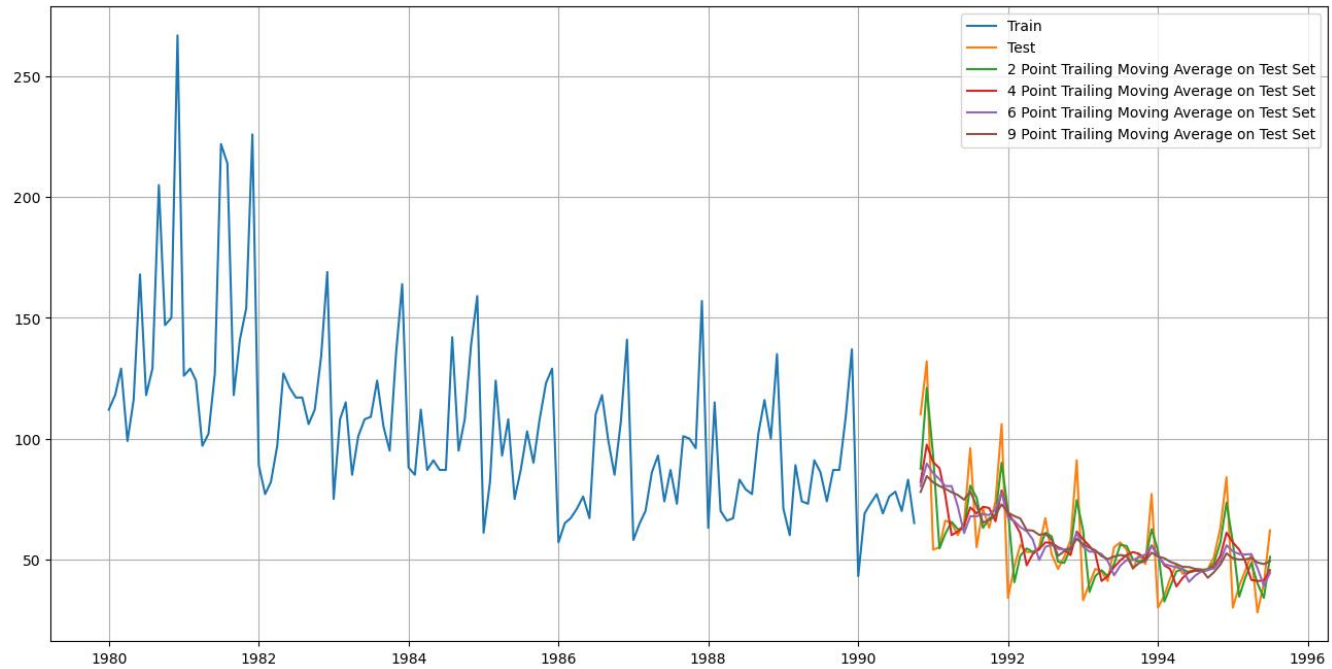

Figure 17 :Comparison of Actual Sales and Trailing Moving Averages on Training and Test Data (Rose)

For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.801
For 4 point Moving Average Model forecast on the Training Data, RMSE is 15.367
For 6 point Moving Average Model forecast on the Training Data, RMSE is 15.862

For 9 point Moving Average Model forecast on the Training Data, RMSE is 16.342

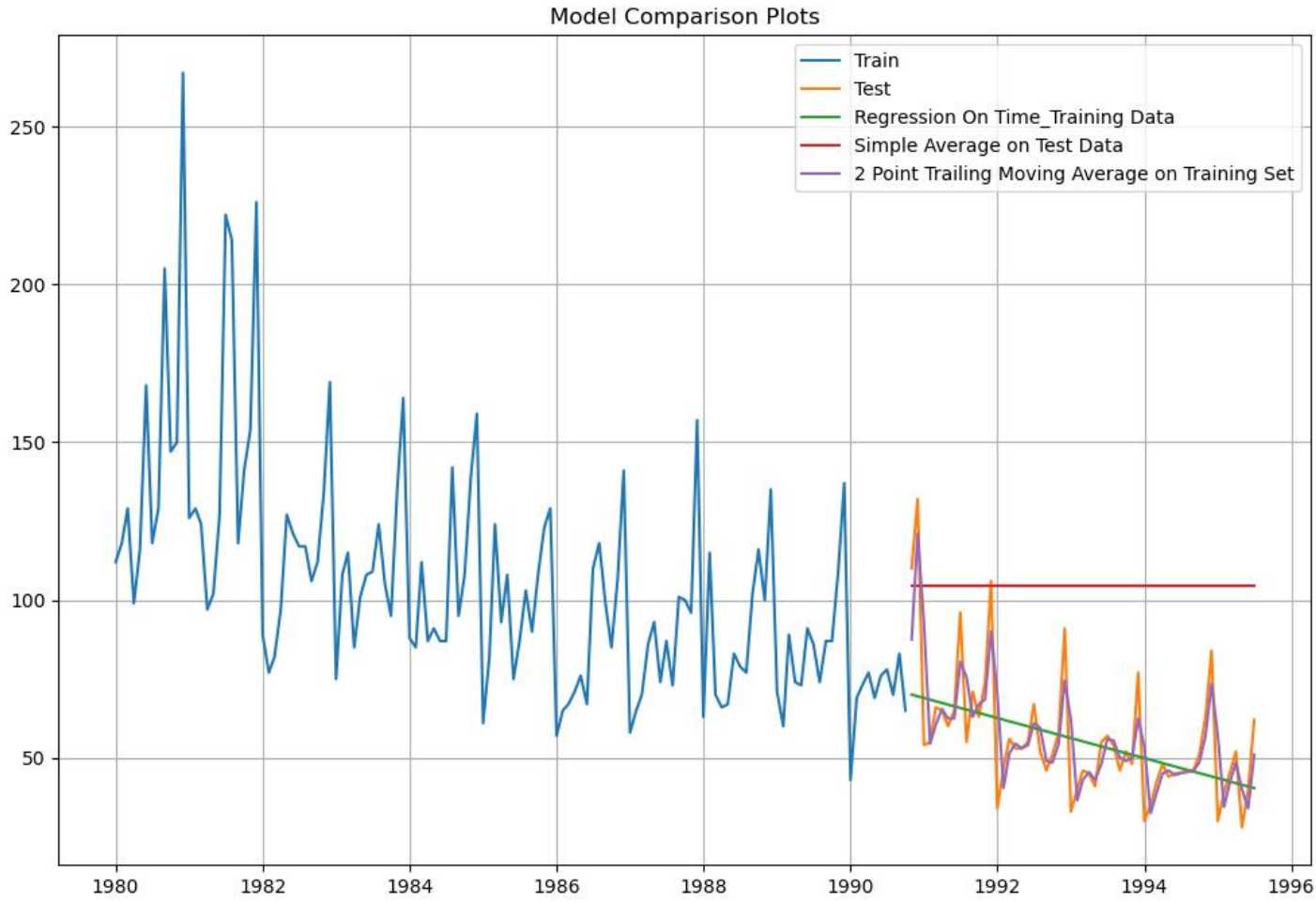

Figure 18 :Comparison of Model Predictions and Actual Sales for Rosé Wine (Training and Test Data)

Model 4: Simple Exponential Smoothing

```
{'smoothing_level': 0.1277774077775358,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 112.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table 11 : Parameters of the Simple Exponential Smoothing (SES) Model

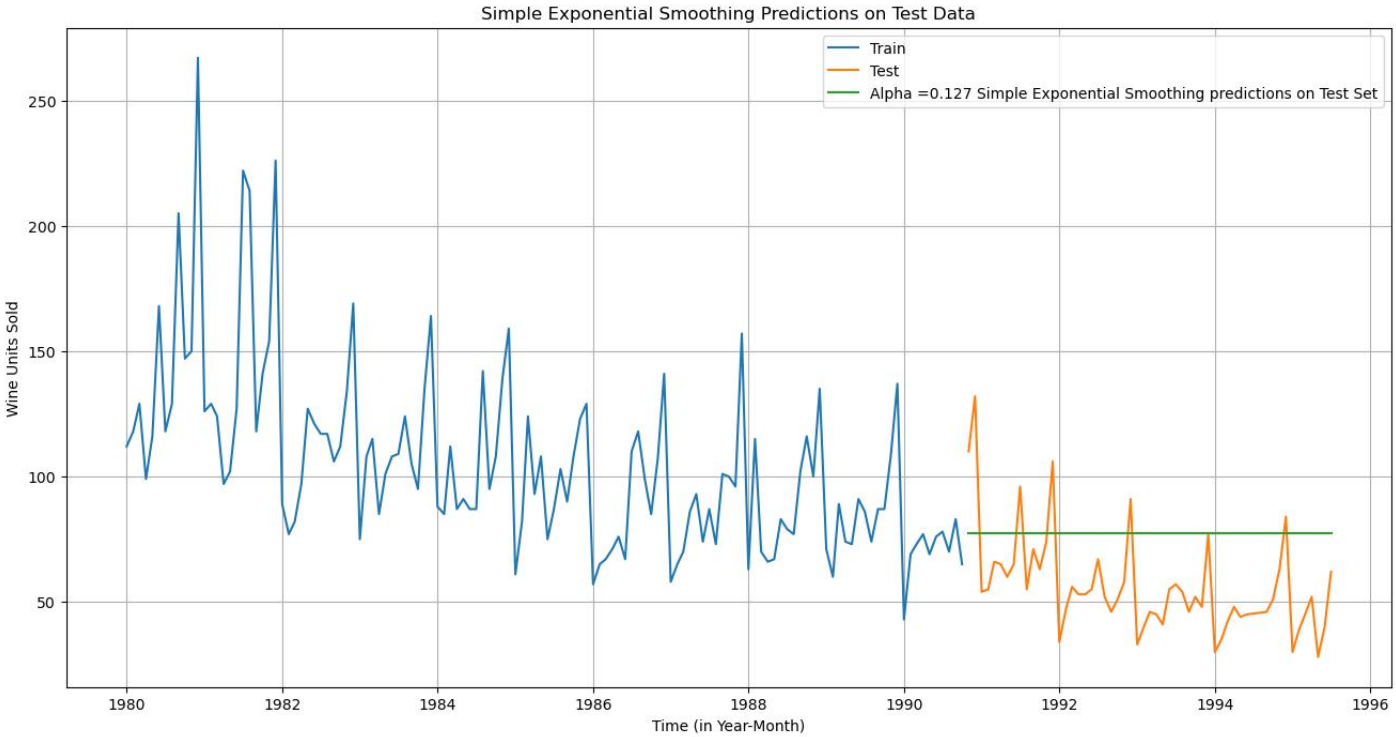

Figure 19 :Simple Exponential Smoothing Predictions vs Actual Sales for Rosé Wine (Training and Test Data)

For Alpha =0.127 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 29.224.

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

	Alpha Values	Train RMSE	Test RMSE
17	0.95	38.150218	22.132051
16	0.90	37.507371	22.496984
15	0.85	36.901698	22.858190
14	0.80	36.330954	23.212947
13	0.75	35.793559	23.559167
12	0.70	35.288467	23.895104
11	0.65	34.815003	24.219182
10	0.60	34.372651	24.529990
9	0.55	33.960778	24.826649
8	0.50	33.578304	25.109786
7	0.45	33.223346	25.383412
6	0.40	32.893017	25.657948
5	0.35	32.583675	25.954512
4	0.30	32.292266	26.310533
3	0.25	32.019860	26.787297
2	0.20	31.779467	27.482320
1	0.15	31.613462	28.556111
0	0.10	31.643829	30.310782

Table 12 : Sorted Results of Test RMSE for Model Comparison

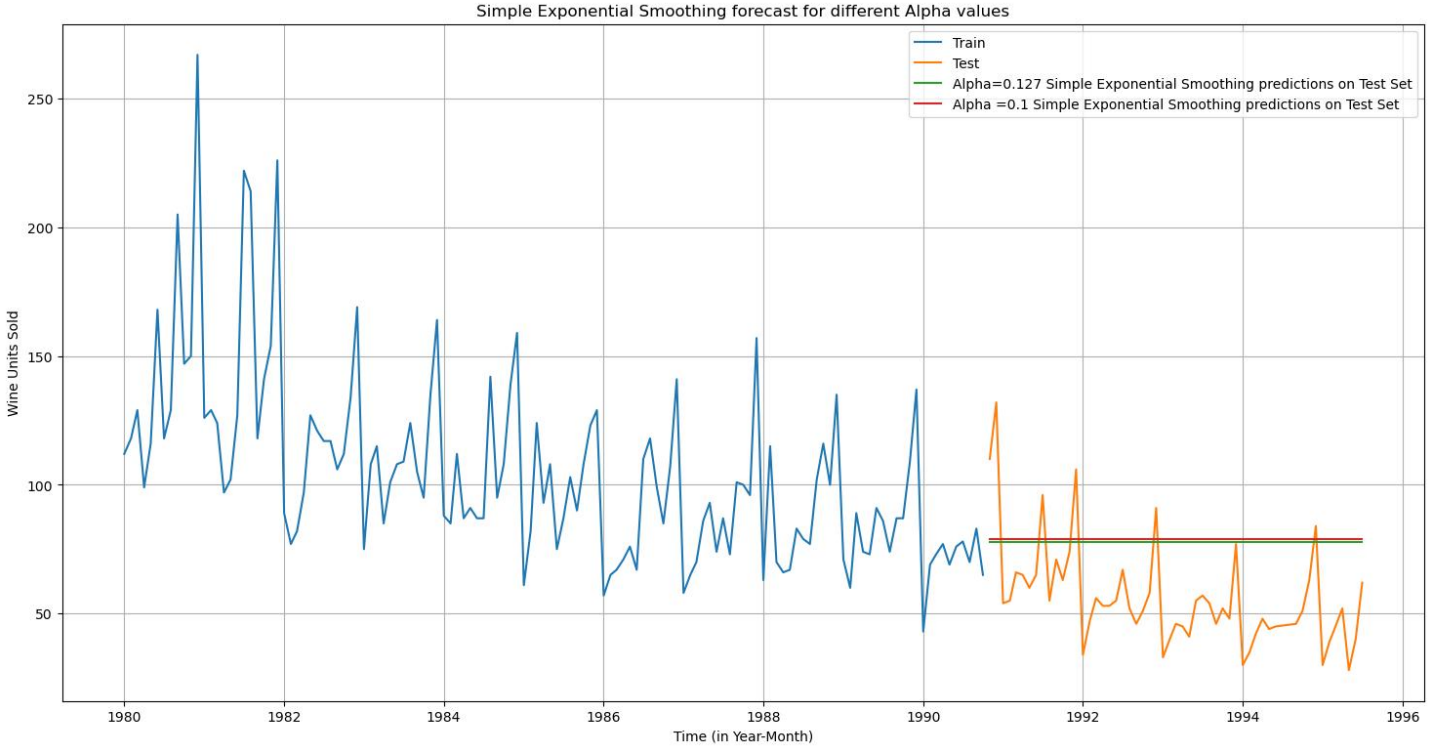

Figure 20 :Simple Exponential Smoothing forecast for different Alpha values

Model 5: Double Exponential Smoothing (Holt's Model)

```
{'smoothing_level': 1.4901161193847656e-08,
 'smoothing_trend': 7.755984441513712e-11,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 139.3527819489728,
 'initial_trend': -0.5291705700335453,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table 13 : Parameters of the Double Exponential Smoothing (DES) Model

Forecasting using this model for the test set

YearMonth	
1990-11-01	70.031437
1990-12-01	69.502267
1991-01-01	68.973096
1991-02-01	68.443926
1991-03-01	67.914755
1991-04-01	67.385584
1991-05-01	66.856414
1991-06-01	66.327243
1991-07-01	65.798073
1991-08-01	65.268902
1991-09-01	64.739732
1991-10-01	64.210561
1991-11-01	63.681390
1991-12-01	63.152220
1992-01-01	62.623049
1992-02-01	62.093879
1992-03-01	61.564708
1992-04-01	61.035538
1992-05-01	60.506367
1992-06-01	59.977196
1992-07-01	59.448026
1992-08-01	58.918855
1992-09-01	58.389685
1992-10-01	57.860514
1992-11-01	57.331344
1992-12-01	56.802173
1993-01-01	56.273002
1993-02-01	55.743832
1993-03-01	55.214661
1993-04-01	54.685491
1993-05-01	54.156320

1993-06-01	53.627150
1993-07-01	53.097979
1993-08-01	52.568808
1993-09-01	52.039638
1993-10-01	51.510467
1993-11-01	50.981297
1993-12-01	50.452126
1994-01-01	49.922956
1994-02-01	49.393785
1994-03-01	48.864614
1994-04-01	48.335444
1994-05-01	47.806273
1994-06-01	47.277103
1994-07-01	46.747932
1994-08-01	46.218762
1994-09-01	45.689591
1994-10-01	45.160420
1994-11-01	44.631250
1994-12-01	44.102079
1995-01-01	43.572909
1995-02-01	43.043738
1995-03-01	42.514568
1995-04-01	41.985397
1995-05-01	41.456226
1995-06-01	40.927056
1995-07-01	40.397885

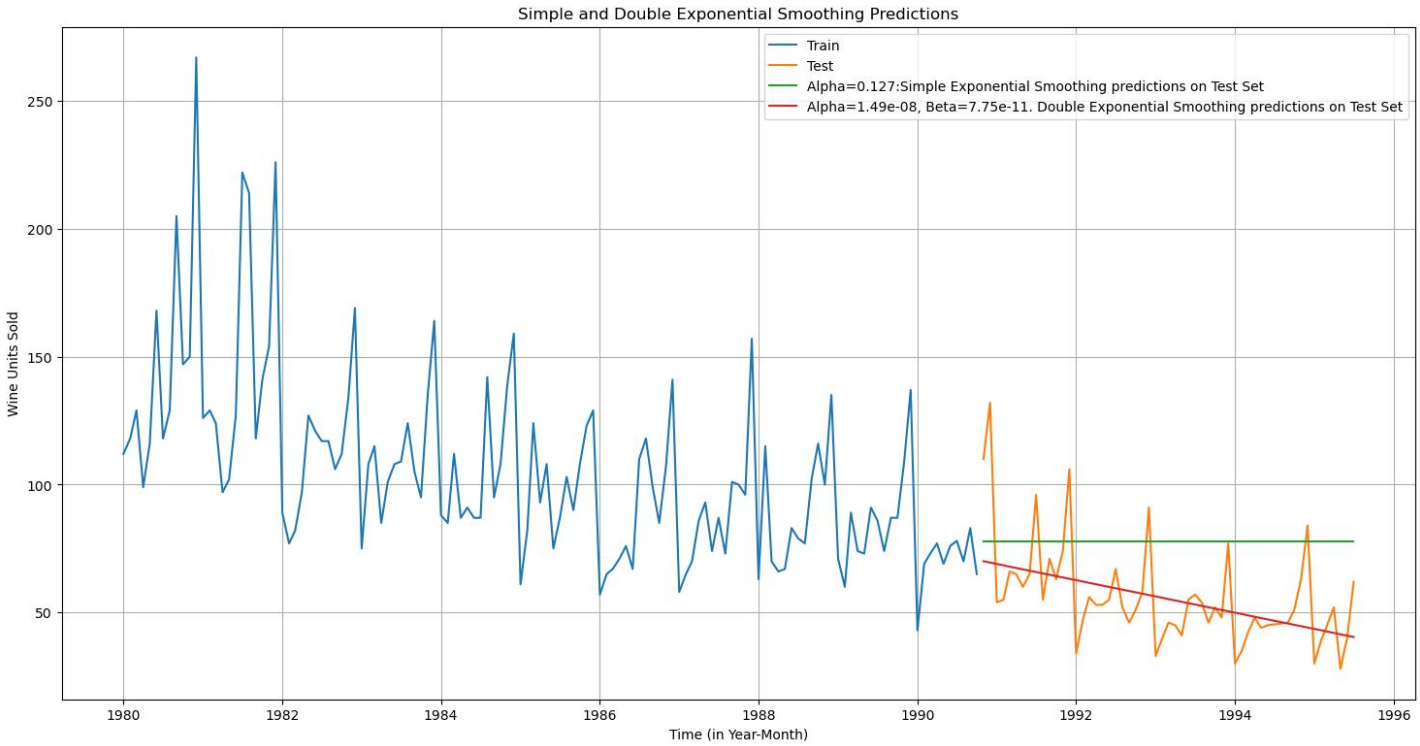

Figure 21 :Comparison of Simple and Double Exponential Smoothing Predictions for Rosé Wine Sales

For DES forecast on the Rose Testing Data: RMSE is 17.356.

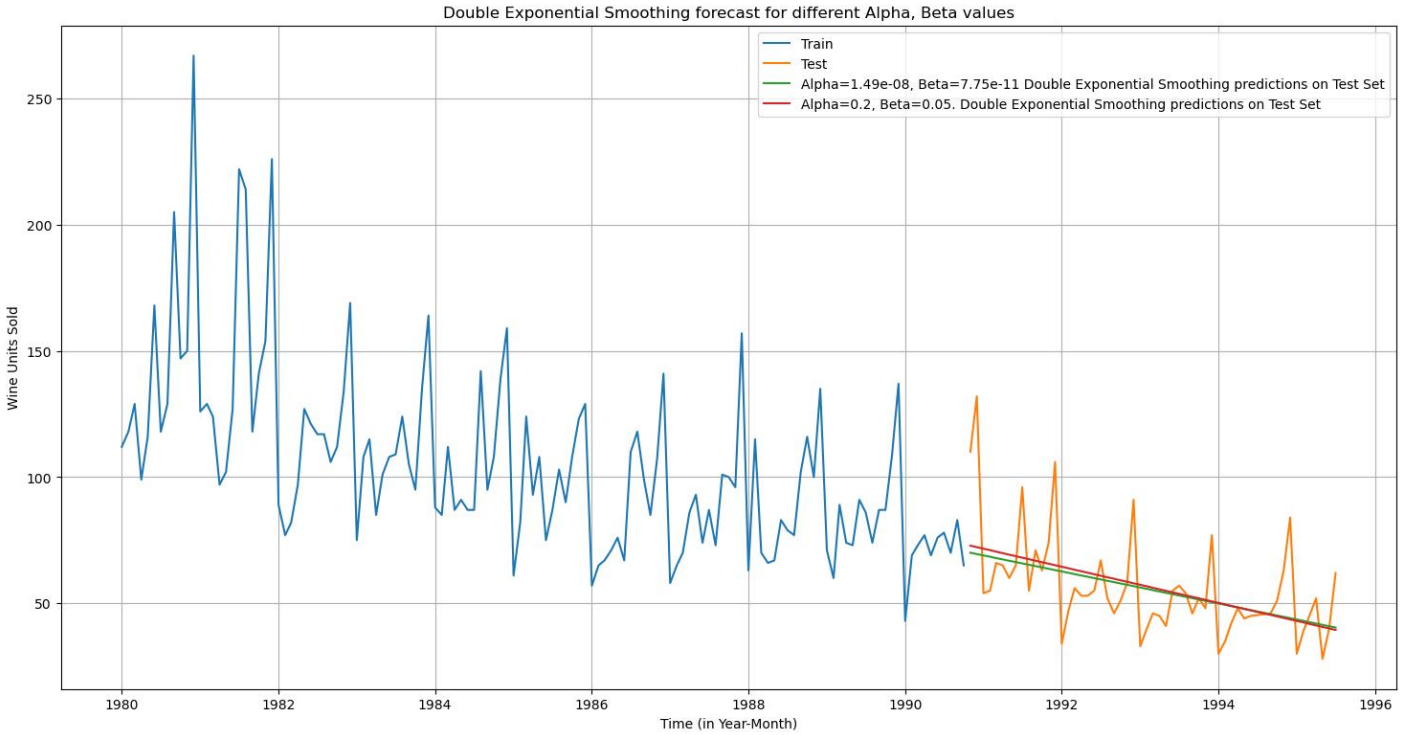

Figure 22 :Double Exponential Smoothing forecast for different Alpha, Beta values (Rose)

Model 6: Triple Exponential Smoothing (Holt - Winter's Model)

```
{'smoothing_level': 0.06208669130988934,
 'smoothing_trend': 0.018324026350319066,
 'smoothing_seasonal': 0.000890407920947476,
 'damping_trend': nan,
 'initial_level': 175.13549001567995,
 'initial_trend': 0.9929819792289998,
 'initial_seasons': array([0.64356811, 0.72961666, 0.79892089, 0.69823075, 0.7853838 ,
        0.85201631, 0.93662832, 0.99971478, 0.94480076, 0.92473909,
        1.06761581, 1.47942515]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table 14 : Parameters of the Triple Exponential Smoothing

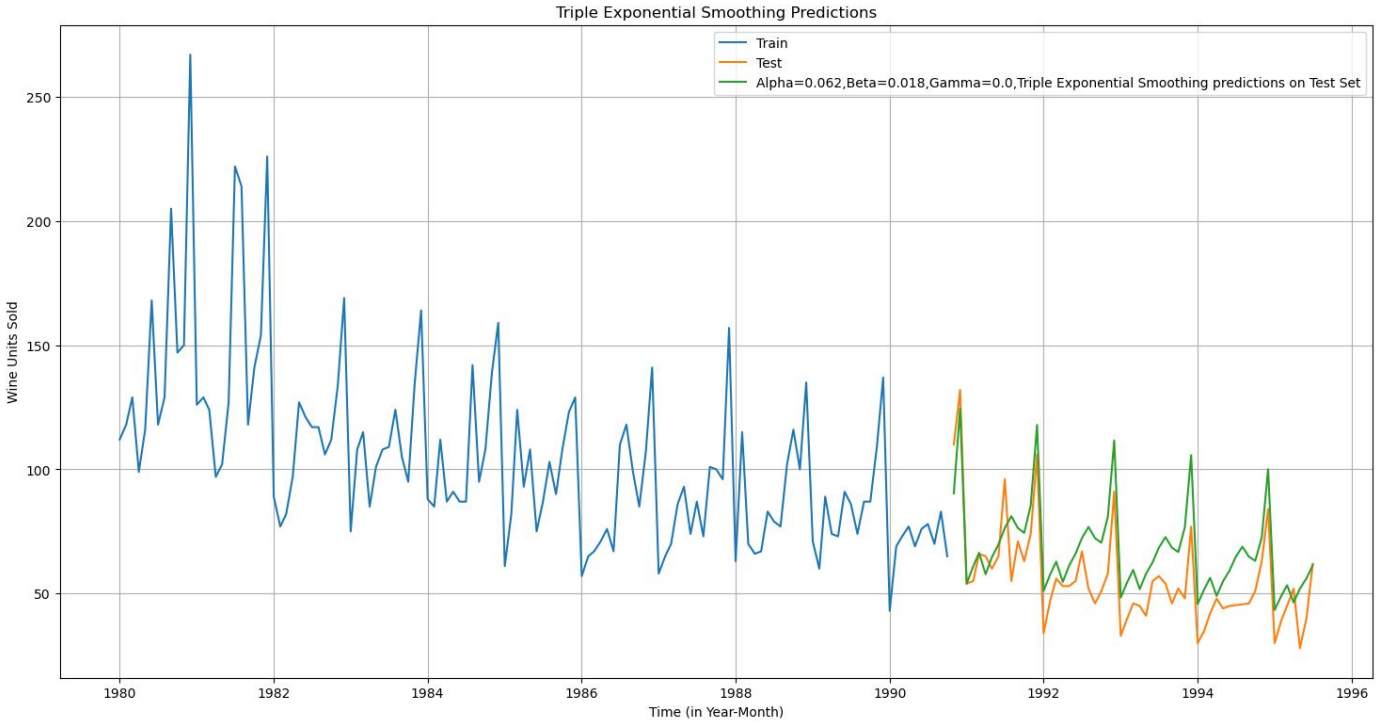

Figure 23 :Triple Exponential Smoothing Predictions

For Alpha=0.062,Beta=0.018,Gamma=0.0, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 15.273

	Test RMSE
Alpha=0.1,Beta=0.25,Gamma=0.85,Triple Exponential Smoothing	8.377119
2pointTrailingMovingAverage	11.801043
Alpha=0.062,Beta=0.018,Gamma=0.0,Triple Exponential Smoothing	15.273169
4pointTrailingMovingAverage	15.367244
6pointTrailingMovingAverage	15.862398
9pointTrailingMovingAverage	16.341947
Alpha=1.49e-08, Beta=7.75e-11, Double Exponential Smoothing	17.355736
Linear Regression	17.355804
Alpha=0.127,SimpleExponentialSmoothing	29.223870
SimpleAverageModel	52.412291

Table 15 : Sorted by RMSE values on the Test Data

We see that the best model is the Triple Exponential Smoothing with multiplicative seasonality with the parameters $\alpha = 0.1$, $\beta = 0.25$ and $\gamma = 0.85$.

- Triple Exponential Smoothing (TES):
  With Alpha=0.1, Beta=0.25, Gamma=0.85, TES achieved the lowest RMSE (8.377).
  TES effectively captures seasonality and trends, making it the most suitable model for this dataset.
- Moving Average Models:
  Performance decreases as the window size increases.
  Suitable for basic trend smoothing but not ideal for datasets with significant seasonality or trend complexity.
- Double Exponential Smoothing (DES):
  While DES captures trends better than SES, it does not handle seasonality, leading to a higher RMSE than TES.
- Linear Regression:
  Comparable to DES but not competitive with TES, suggesting limited utility in this context.
- Simple Exponential Smoothing (SES):
  Focuses only on the level component and is not effective for data with trends or seasonality.
- Simple Average Model:
  The poorest performer, as expected, due to its simplistic nature and inability to adapt to changes in the data.

With its ability to model level, trend, and seasonality, TES provides the most accurate forecasts.

Parameter tuning is critical; the optimal configuration (Alpha=0.1, Beta=0.25, Gamma=0.85) should be used.

Check for Stationarity

A Time Series is considered to be stationary when statistical properties such as the variance and (auto) correlation are constant over time.

Stationary Time Series allows us to think of the statistical properties of the time series as not changing in time, which enables us to build appropriate statistical models for forecasting based on past data.

Dickey-Fuller Test - Dicky Fuller Test on the timeseries is run to check for stationarity of data.

Null Hypothesis H0 : Time Series is non-stationary.
Alternate Hypothesis Ha : Time Series is stationary.
So Ideally if p-value < 0.05 then null hypothesis: TS is non-stationary is rejected else the TS is non-stationary is failed to be rejected .

Results of Dicky-Fuller Test
DF test statistic is -2.240
DF test p-value is 0.46713522477982006
Number of lags used 13

We see 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.

- Differencing 'd' is done on a non-stationary time series data one or more times to convert it into stationary.

- (d=1) 1st order differencing is done where the difference between the current and previous (1 lag before) series is taken and then checked for stationarity using the ADF(Augmented Dicky Fueller) test. If differenced time series is stationary, we proceed with AR modeling. Else we do (d=2) 2nd order differencing, and this process repeats till we get a stationary time series

- The variance of a time series may also not be the same over time. To remove this kind of non-stationarity, we can transform the data. If the variance is increasing over time, then a log transformation can stabilize the variance.

Results of Dicky-Fuller Test with differencing
DF test statistic is -8.162
DF test p-value is 3.016095098834424e-11
Number of lags used 12

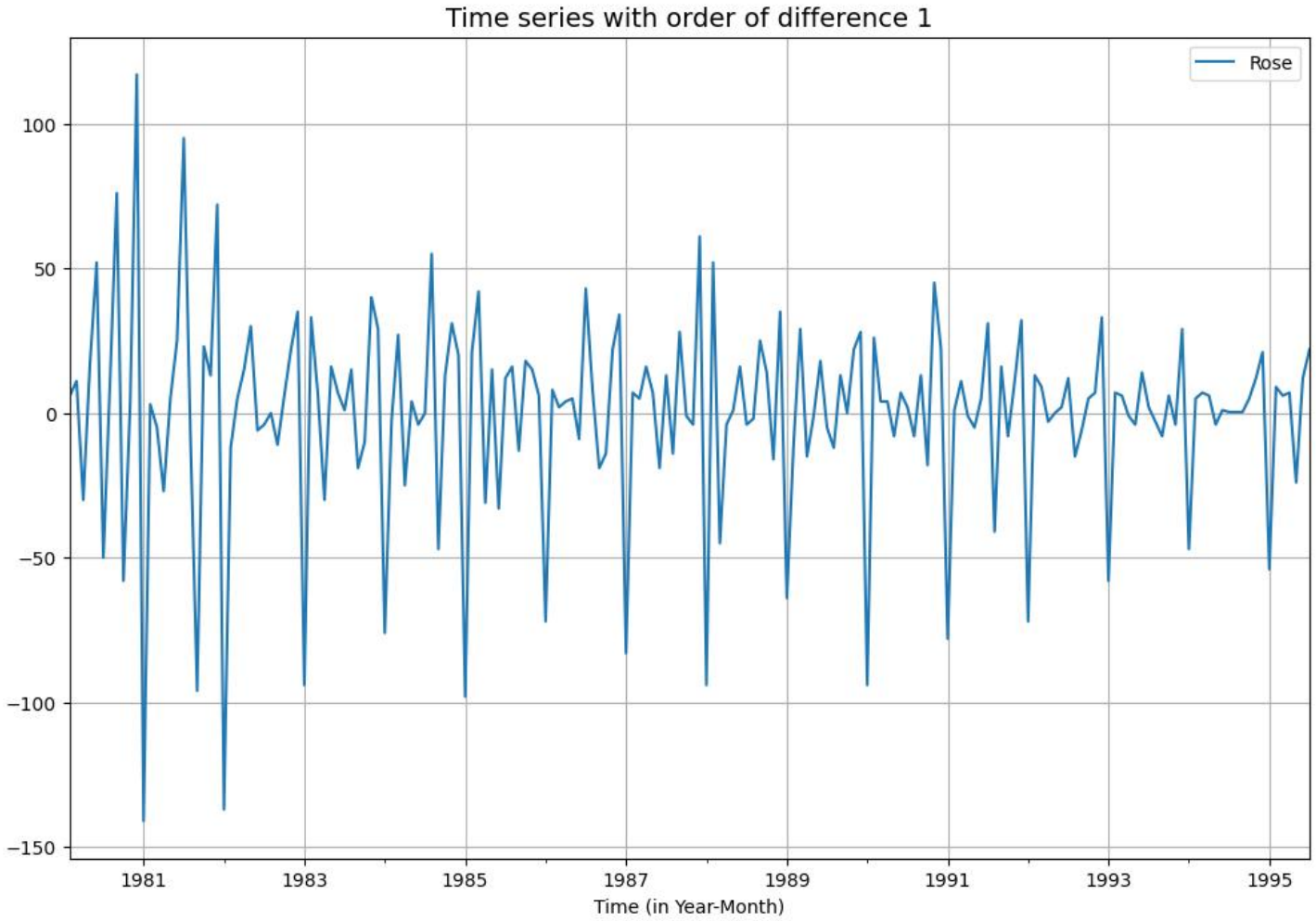

Figure 24 :Time series with order of difference 1 (Rose)

Model Building - Stationary Data

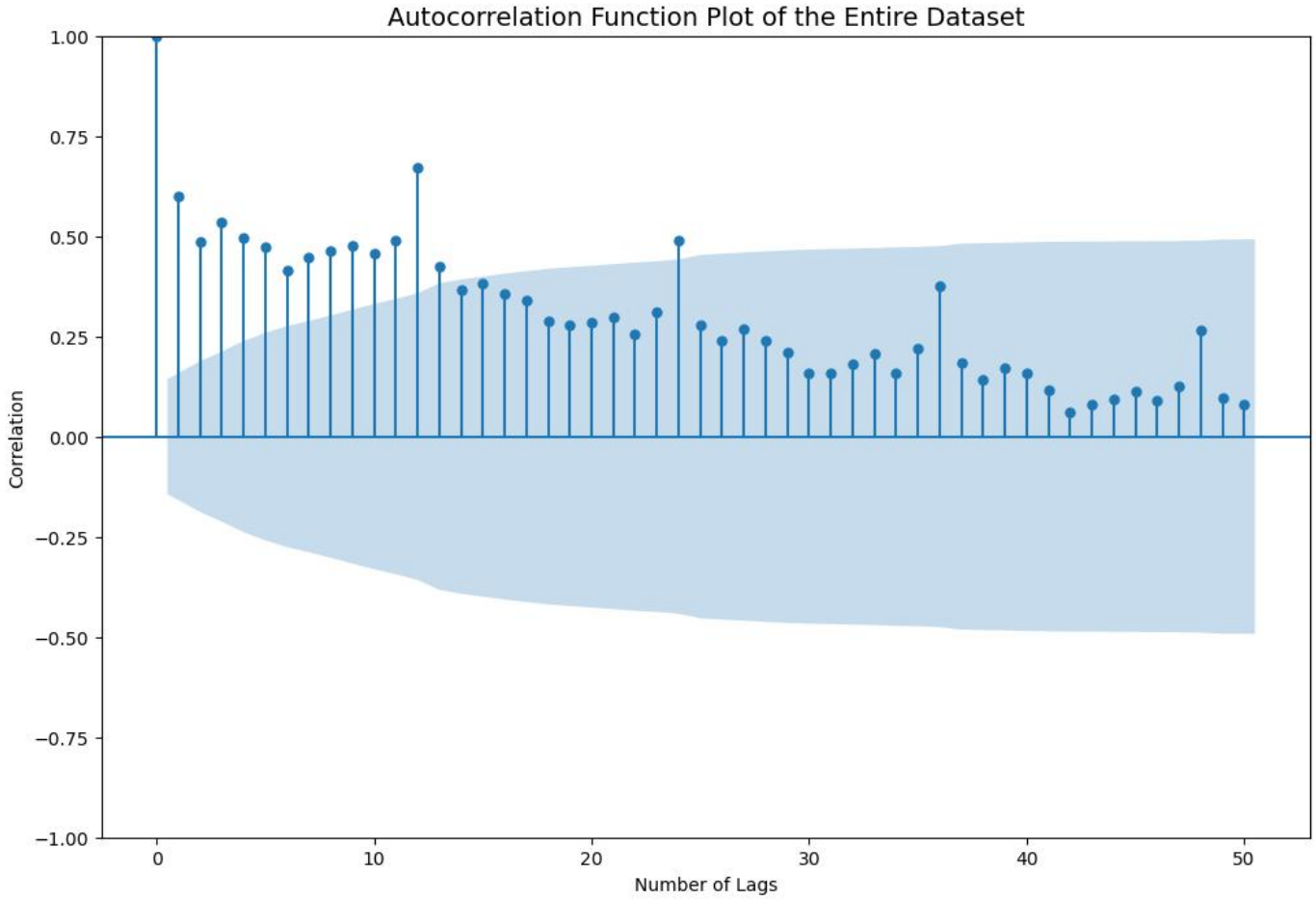

Figure 24 :ACF Plot (Rose)

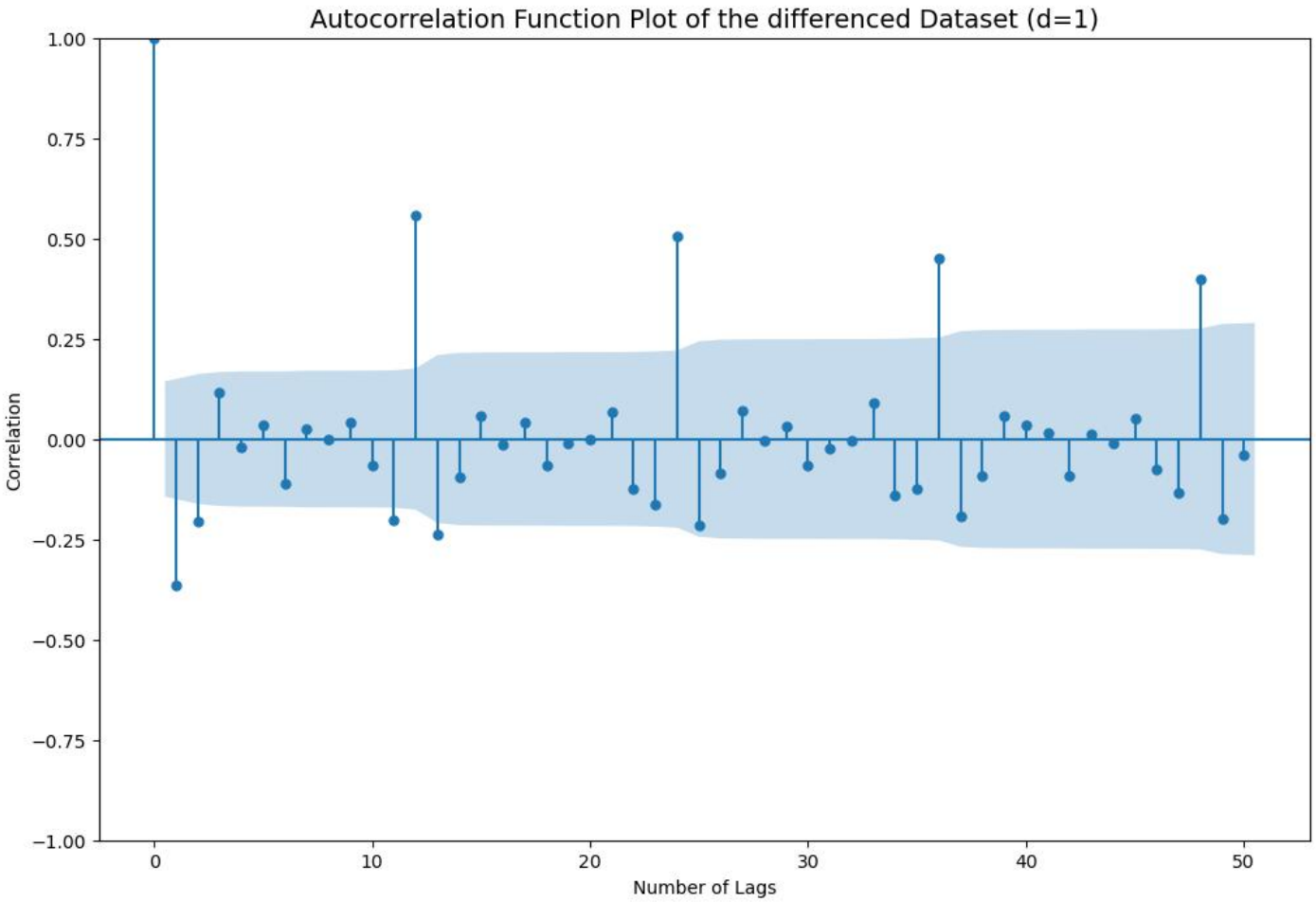

Figure 25 :Autocorrelation Function Plot of the differenced Dataset (d=1) (Rose)

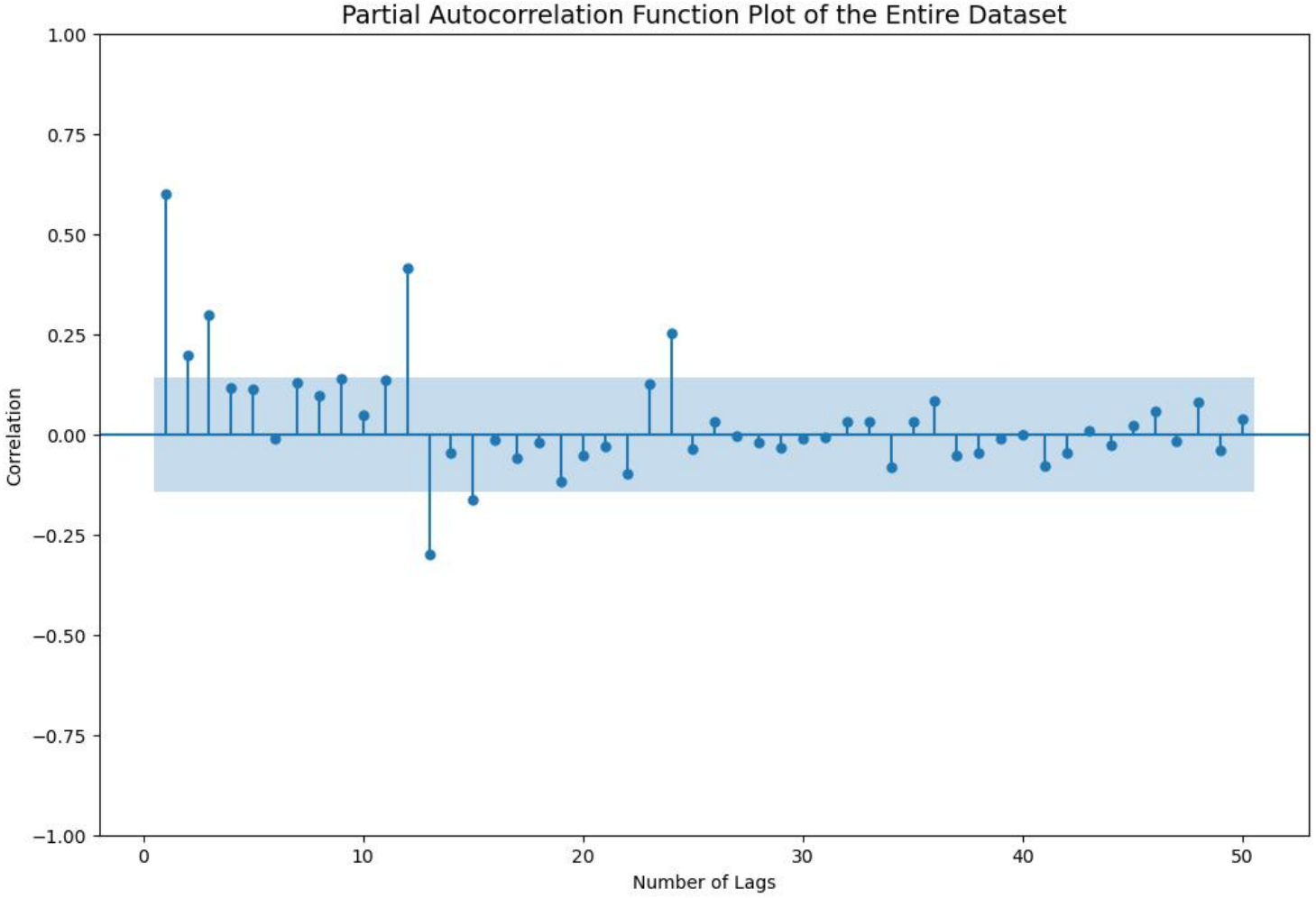

Figure 26 :Partial Autocorrelation Function Plot of the Entire Dataset (Rose)

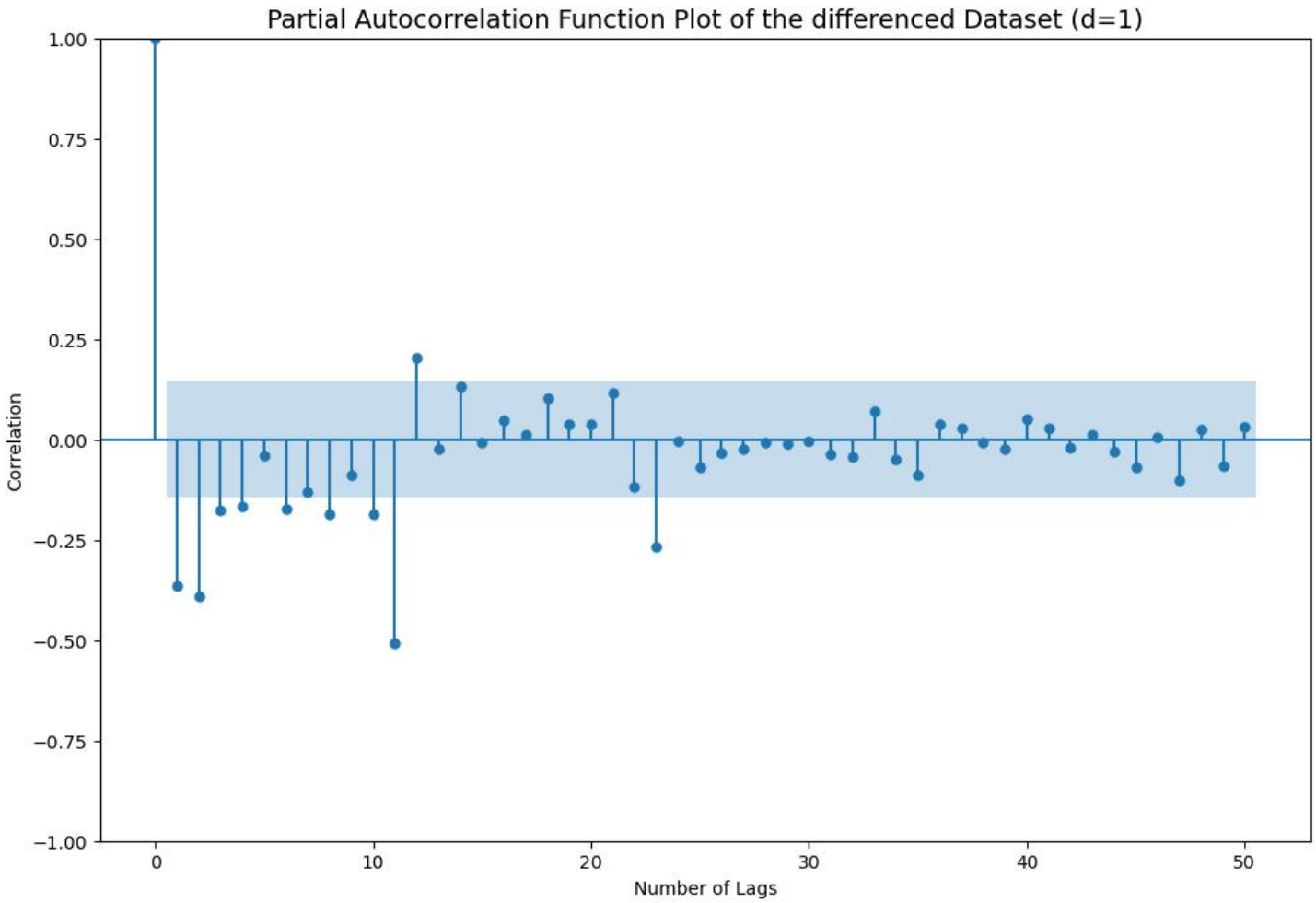

Figure 27 :Partial Autocorrelation Function Plot of the differenced Dataset (d=1)(Rose)

- Presence of seasonality found from the above plot.

Model 1: ARIMA Model

ARIMA(0, 1, 0) - AIC:1313.1758613526429
ARIMA(0, 1, 1) - AIC:1261.3274438405808
ARIMA(0, 1, 2) - AIC:1259.2477803151237
ARIMA(0, 1, 3) - AIC:1260.1328188315786
ARIMA(0, 1, 4) - AIC:1261.2882323654271
ARIMA(1, 1, 0) - AIC:1297.0772943848615
ARIMA(1, 1, 1) - AIC:1260.0367627036055
ARIMA(1, 1, 2) - AIC:1259.4732049501204
ARIMA(1, 1, 3) - AIC:1261.4721912366276
ARIMA(1, 1, 4) - AIC:1259.187773565361
ARIMA(2, 1, 0) - AIC:1278.1352807484318
ARIMA(2, 1, 1) - AIC:1261.0140762916922
ARIMA(2, 1, 2) - AIC:1261.472000656906
ARIMA(2, 1, 3) - AIC:1258.119722760092

ARIMA(2, 1, 4) - AIC:1256.9235351115808
ARIMA(3, 1, 0) - AIC:1276.8427173909395
ARIMA(3, 1, 1) - AIC:1261.969097760944
ARIMA(3, 1, 2) - AIC:1263.33176710444
ARIMA(3, 1, 3) - AIC:1258.4353016064729
ARIMA(3, 1, 4) - AIC:1263.791167401926
ARIMA(4, 1, 0) - AIC:1275.6764376351036
ARIMA(4, 1, 1) - AIC:1263.386892538002
ARIMA(4, 1, 2) - AIC:1265.330538097643
ARIMA(4, 1, 3) - AIC:1258.4359772931566
ARIMA(4, 1, 4) - AIC:1261.8801587631947

	param	AIC
14	(2, 1, 4)	1256.923535
13	(2, 1, 3)	1258.119723
18	(3, 1, 3)	1258.435302
23	(4, 1, 3)	1258.435977
9	(1, 1, 4)	1259.187774

Table 16 : AIC values in the ascending order

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	130
Model:	ARIMA(2, 1, 4)	Log Likelihood	-621.462
Date:	Sun, 17 Nov 2024	AIC	1256.924
Time:	12:36:15	BIC	1276.942
Sample:	01-01-1980	HQIC	1265.058
	- 10-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9906	0.038	-26.141	0.000	-1.065	-0.916
ar.L2	-0.9873	0.036	-27.339	0.000	-1.058	-0.917
ma.L1	0.3248	0.144	2.249	0.025	0.042	0.608
ma.L2	0.1227	0.176	0.698	0.485	-0.222	0.467
ma.L3	-0.8700	0.158	-5.505	0.000	-1.180	-0.560
ma.L4	-0.2640	0.106	-2.492	0.013	-0.472	-0.056
sigma2	917.2663	167.781	5.467	0.000	588.422	1246.110

Ljung-Box (L1) (Q):	0.08	Jarque-Bera (JB):	49.64
Prob(Q):	0.78	Prob(JB):	0.00
Heteroskedasticity (H):	0.32	Skew:	0.93
Prob(H) (two-sided):	0.00	Kurtosis:	5.40

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

Table 17 : Auto ARIMA Model Summary for Rosé Wine Sales Forecasting

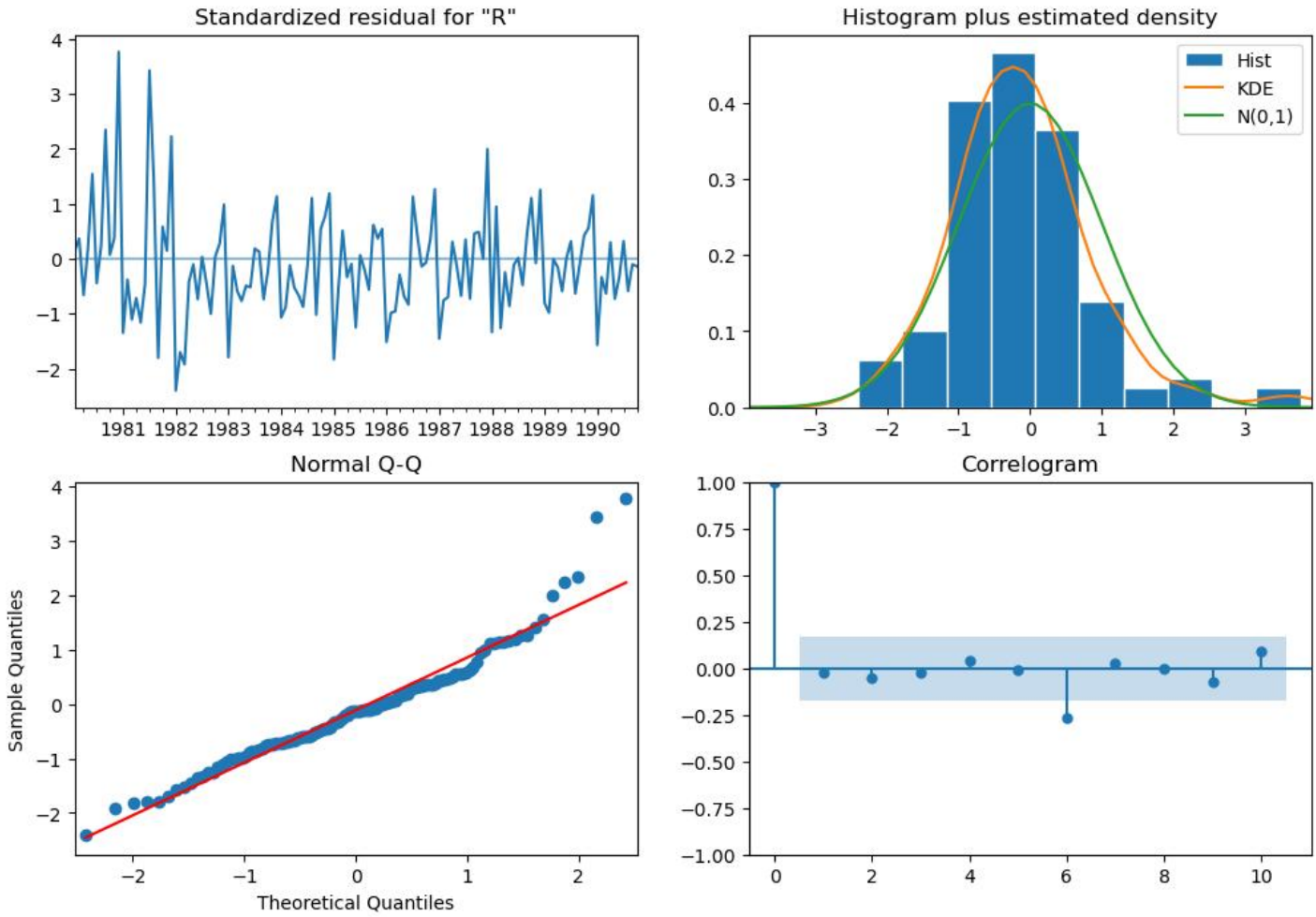

Figure 28 :Diagnostics plot Auto ARIMA (Rose)

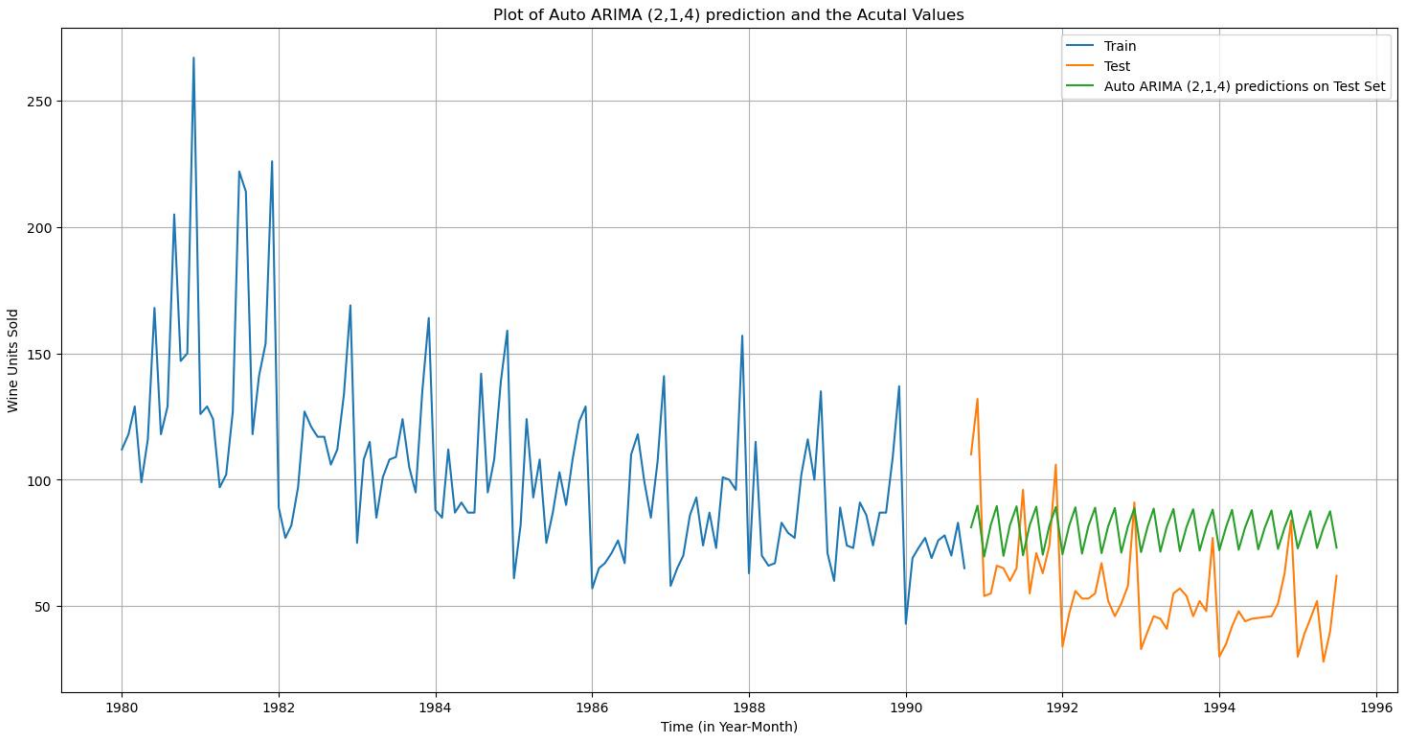

Figure 29 :Plot of Auto ARIMA (2,1,4) prediction and the Actual Values (Rose)

- **RMSE: 31.15452801990538**
- **MAPE: 60.92411545222197**

Model 2: Manual ARIMA Model

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	130
Model:	ARIMA(2, 1, 2)	Log Likelihood	-625.736
Date:	Sun, 17 Nov 2024	AIC	1261.472
Time:	12:36:17	BIC	1275.771
Sample:	01-01-1980 - 10-01-1990	HQIC	1267.282
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4622	0.483	-0.957	0.338	-1.409	0.484
ar.L2	-0.0039	0.169	-0.023	0.981	-0.335	0.327
ma.L1	-0.2523	0.473	-0.534	0.594	-1.179	0.674
ma.L2	-0.5931	0.442	-1.341	0.180	-1.460	0.274
sigma2	945.0138	89.890	10.513	0.000	768.832	1121.196

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	39.93
Prob(Q):	0.87	Prob(JB):	0.00
Heteroskedasticity (H):	0.33	Skew:	0.85
Prob(H) (two-sided):	0.00	Kurtosis:	5.14

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

Table 18 : Manual ARIMA Model Summary for Rose Wine Sales Forecasting

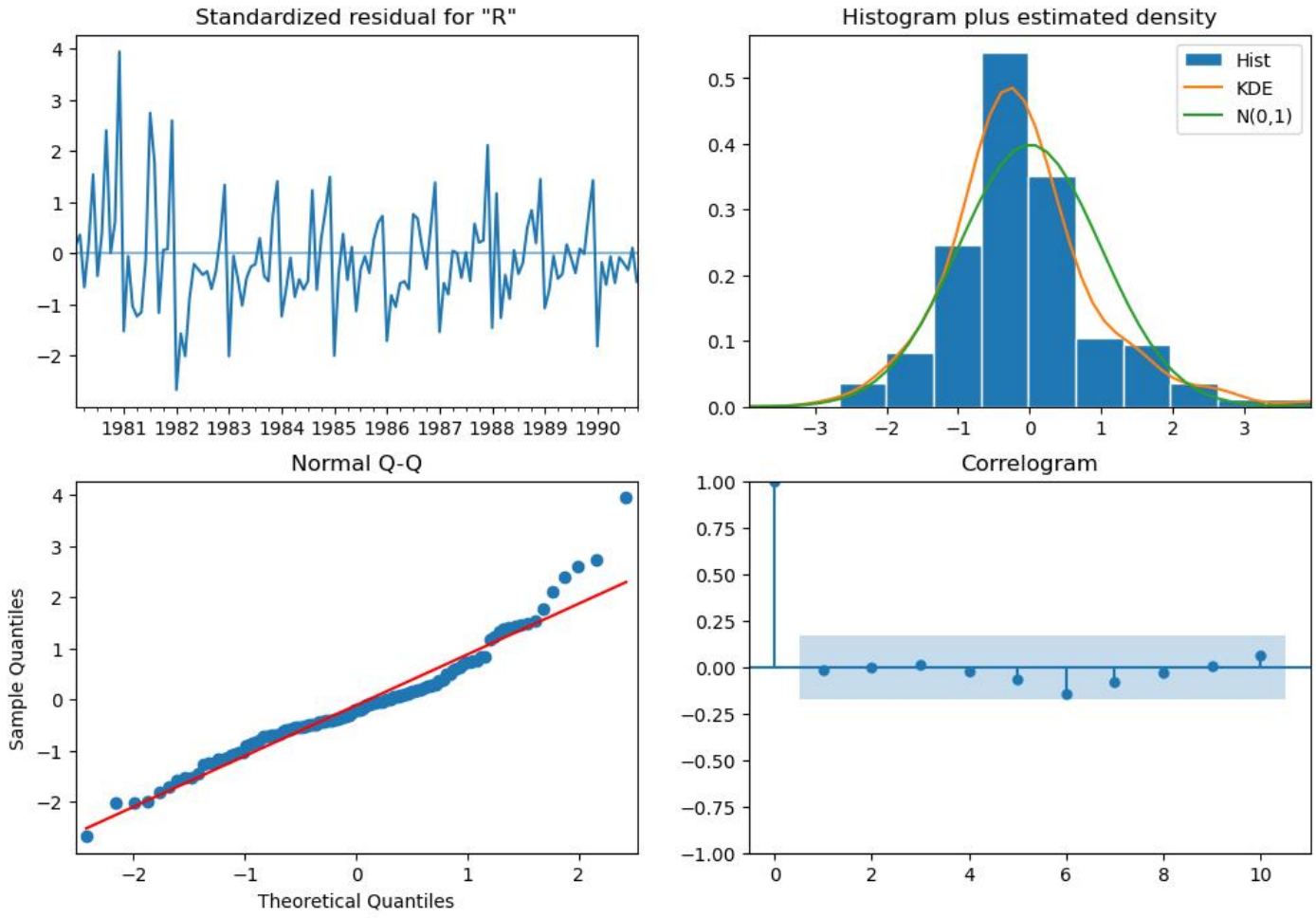

Figure 30 :Diagnostic Plot Manual ARIMA (Rose)

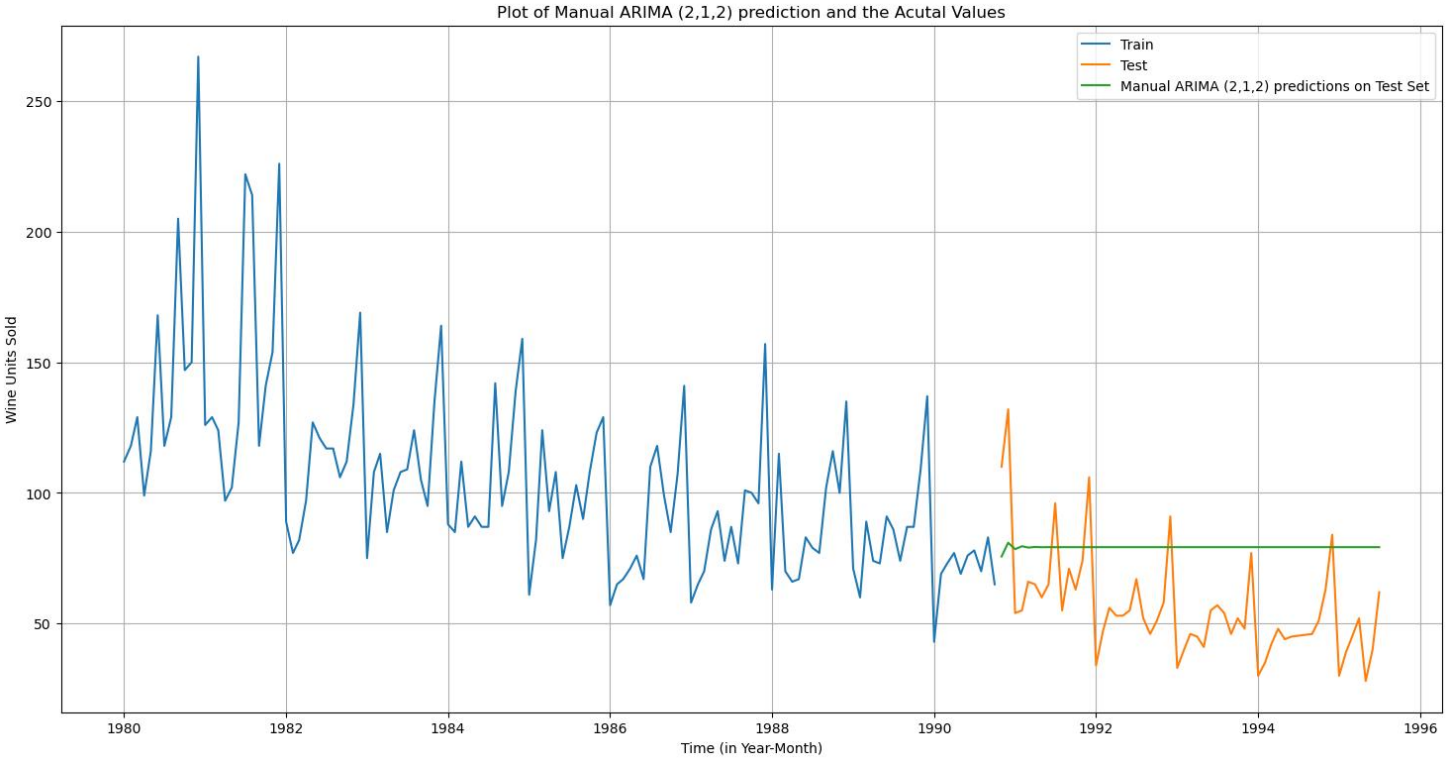

Figure 31 :Plot of Manual ARIMA (2,1,2) prediction and the Acutal Values (Rose)

RMSE: 30.448609840075147
MAPE: 60.26583908102812

Model 3: Auto SARIMA Model

27	(0, 1, 1)	NaN	381.030735
59	(0, 1, 3)	NaN	665.165121
222	(3, 1, 1)	NaN	758.399542
238	(3, 1, 2)	NaN	759.007034
252	(3, 1, 3)	NaN	759.060266
220	(3, 1, 1)	NaN	759.394980
221	(3, 1, 1)	NaN	759.587878
253	(3, 1, 3)	NaN	759.955600
237	(3, 1, 2)	NaN	760.555695
236	(3, 1, 2)	NaN	761.246193
254	(3, 1, 3)	NaN	761.852918
189	(2, 1, 3)	NaN	763.457230
188	(2, 1, 3)	NaN	765.735947
190	(2, 1, 3)	NaN	768.673639
172	(2, 1, 2)	NaN	769.143806

Table 19 : SARIMA Models Sorted by AIC for Rosé Wine Sales Forecasting

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	130
Model:	SARIMAX(0, 1, 1)x(3, 0, [1, 2], 12)	Log Likelihood	-385.304
Date:	Sun, 17 Nov 2024	AIC	784.607
Time:	12:39:22	BIC	802.336
Sample:	01-01-1980	HQIC	791.766
	- 10-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-1.0243	0.083	-12.338	0.000	-1.187	-0.862
ar.S.L12	0.6942	0.149	4.647	0.000	0.401	0.987
ar.S.L24	0.1212	0.126	0.959	0.338	-0.127	0.369
ar.S.L36	0.0742	0.065	1.142	0.253	-0.053	0.201
ma.S.L12	-0.4587	0.222	-2.069	0.039	-0.893	-0.024
ma.S.L24	-0.1888	0.180	-1.048	0.295	-0.542	0.164
sigma2	199.6067	41.993	4.753	0.000	117.302	281.911

Ljung-Box (L1) (Q):	0.67	Jarque-Bera (JB):	3.06
Prob(Q):	0.41	Prob(JB):	0.22
Heteroskedasticity (H):	0.66	Skew:	0.44
Prob(H) (two-sided):	0.25	Kurtosis:	3.06

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

Table 20 : Auto SARIMA Model Summary for Rose Wine Sales Forecasting

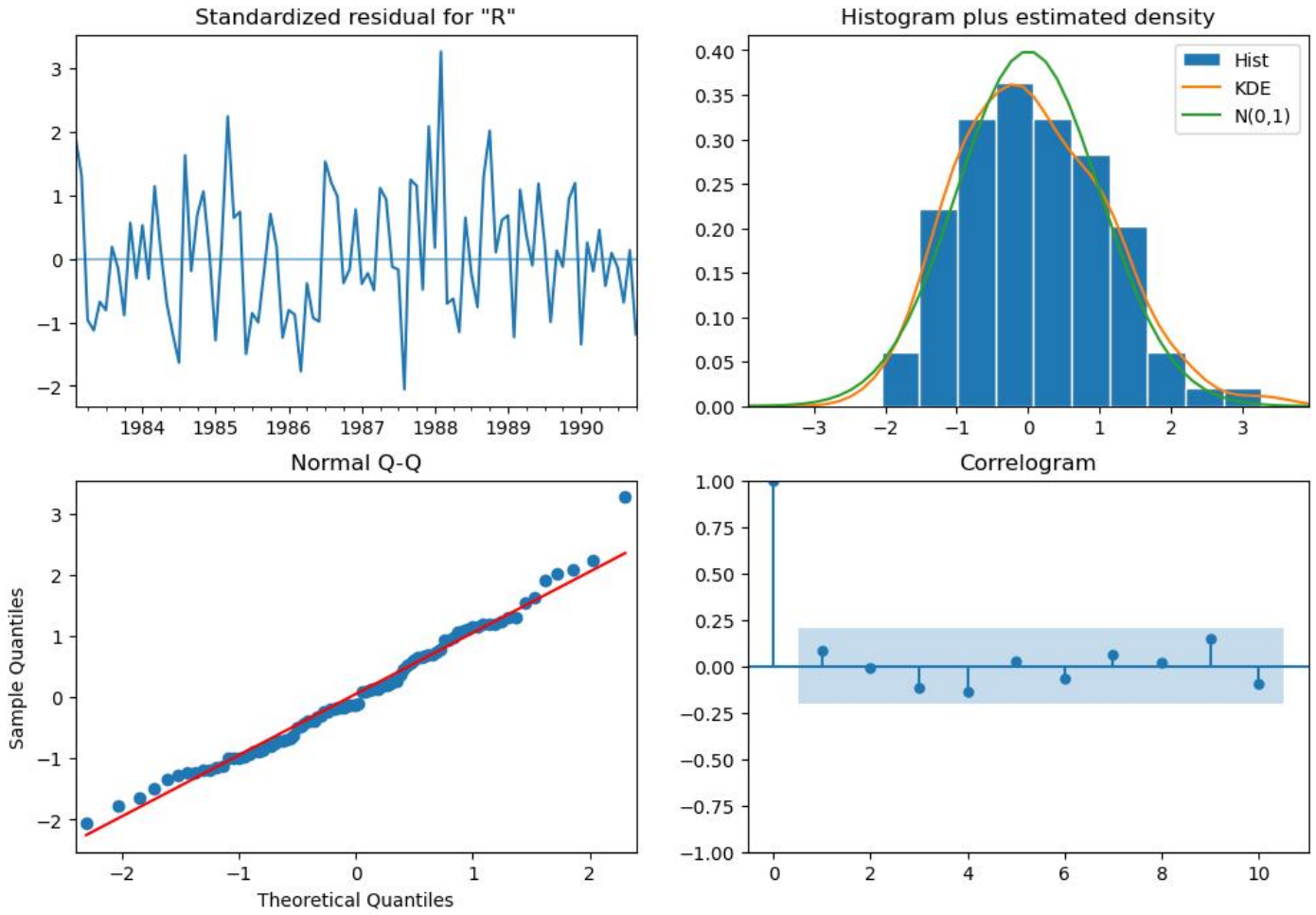

Figure 32 :Diagnostic Plot of Auto SARIMA (Rose)

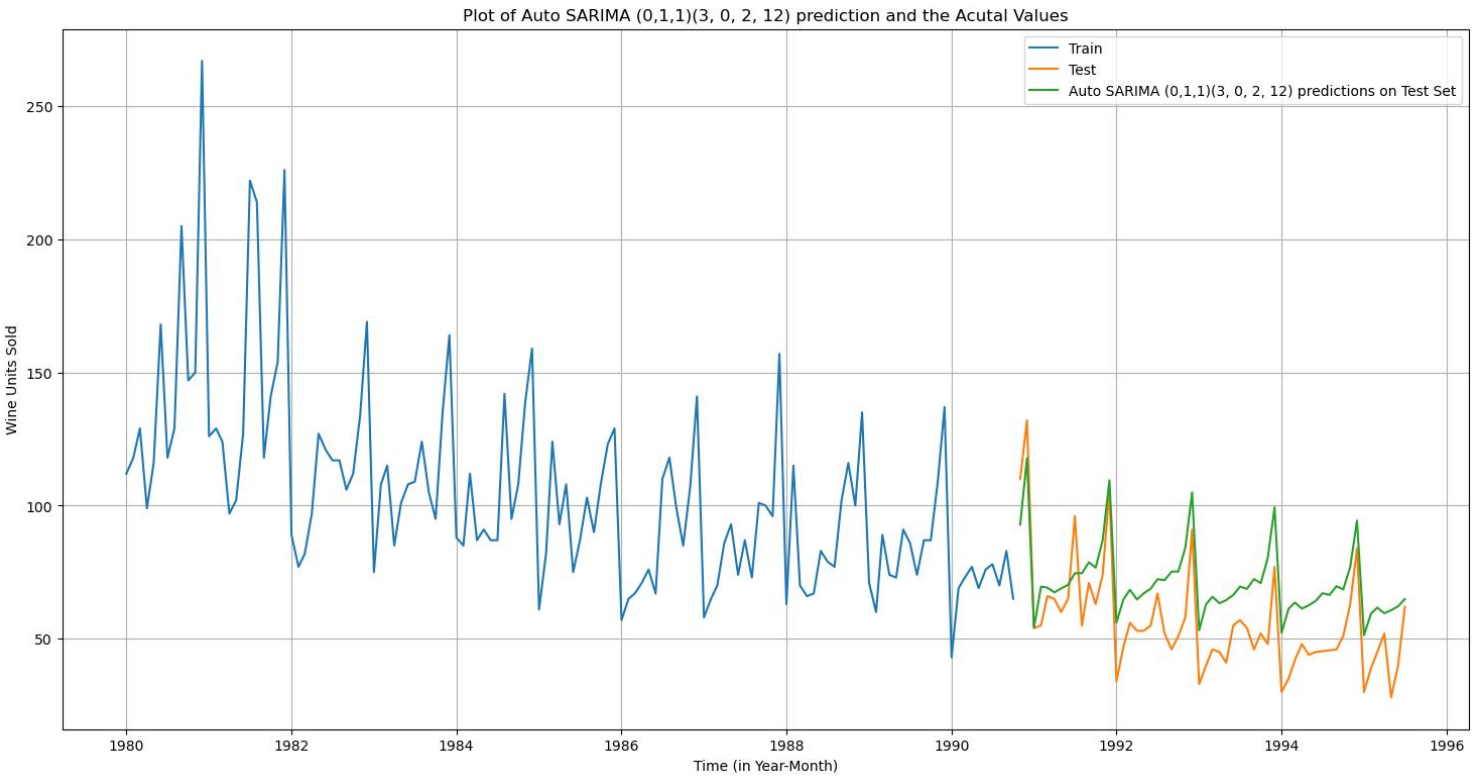

Figure 33 :Plot of Auto SARIMA (0,1,1)(3, 0, 2, 12) prediction and the Acutal Values (Rose)

RMSE: 18.287570169793984
MAPE: 35.30270091722601

Model 4: Manual SARIMA Model

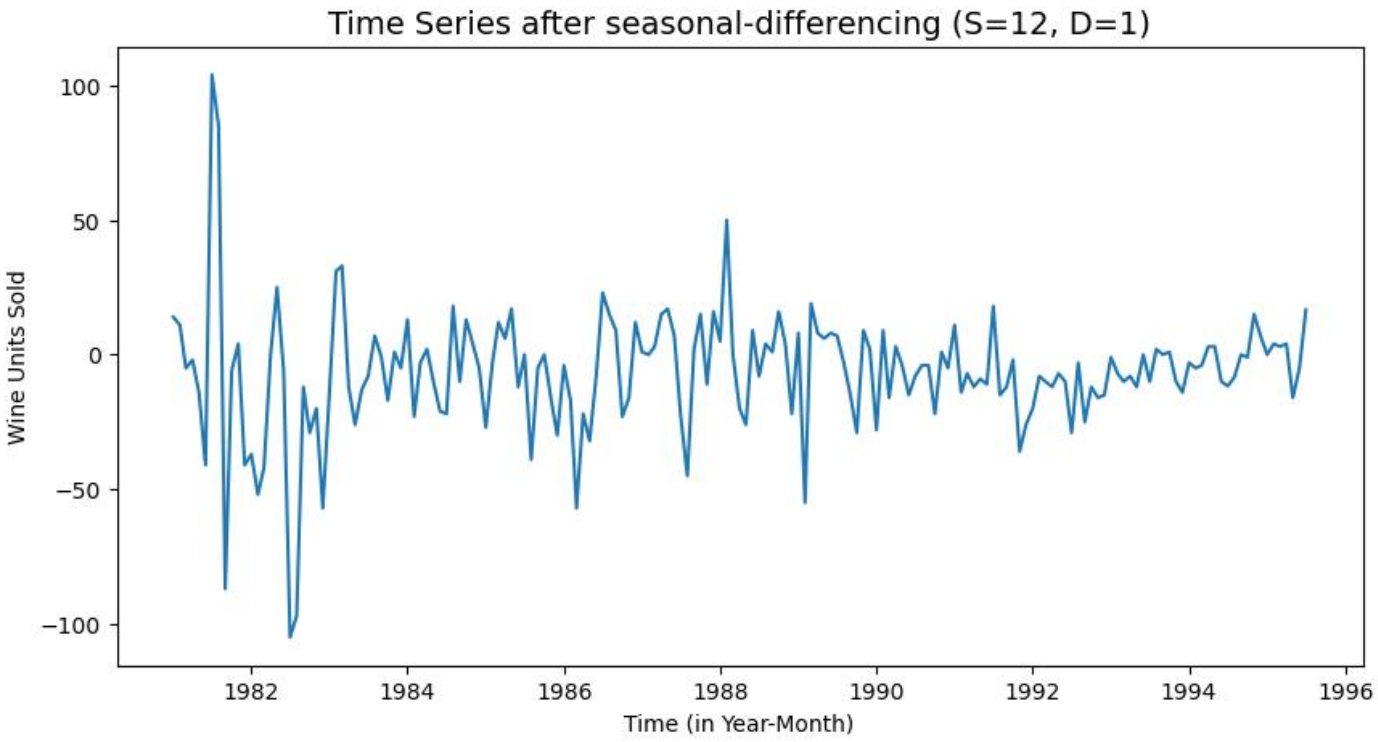

Figure 34 :Time Series after seasonal-differencing (S=12, D=1)

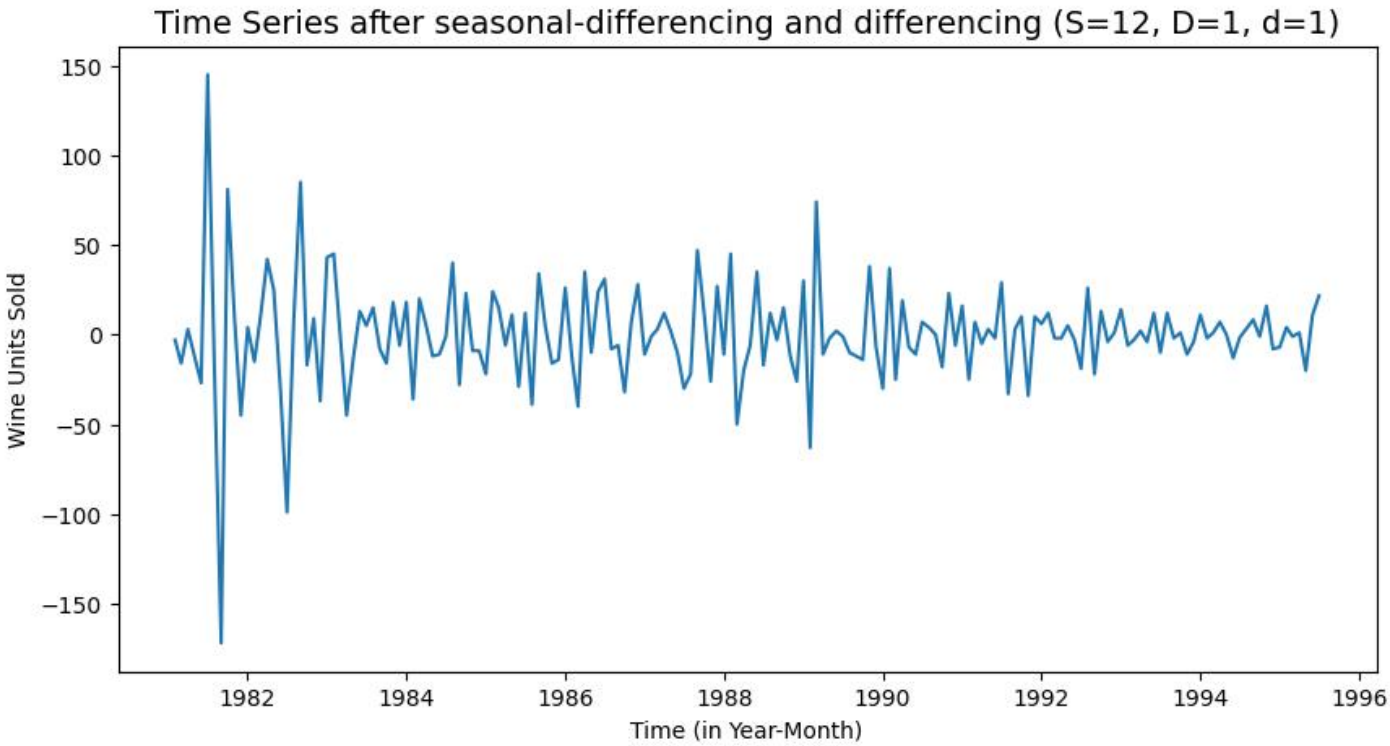

Figure 35 :Time Series after seasonal-differencing and differencing (S=12, D=1, d=1)

DF test statistic is -4.551
DF test p-value is 0.0012457501217352787
Number of lags used 11

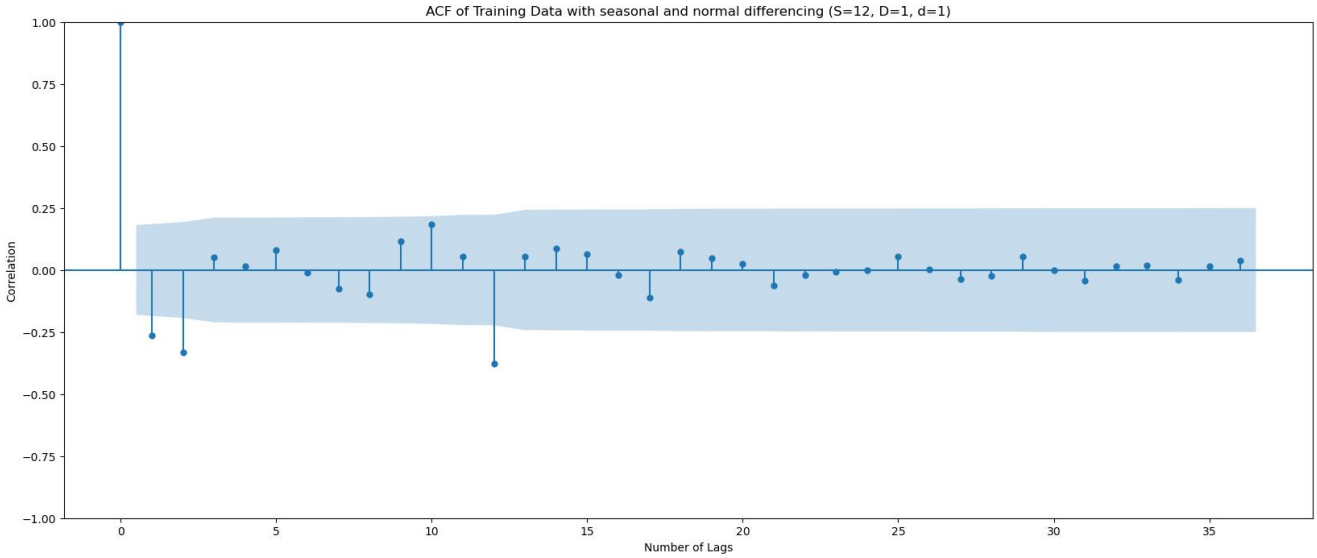

Figure 36 :ACF of Training Data with seasonal and normal differencing (S=12, D=1, d=1)

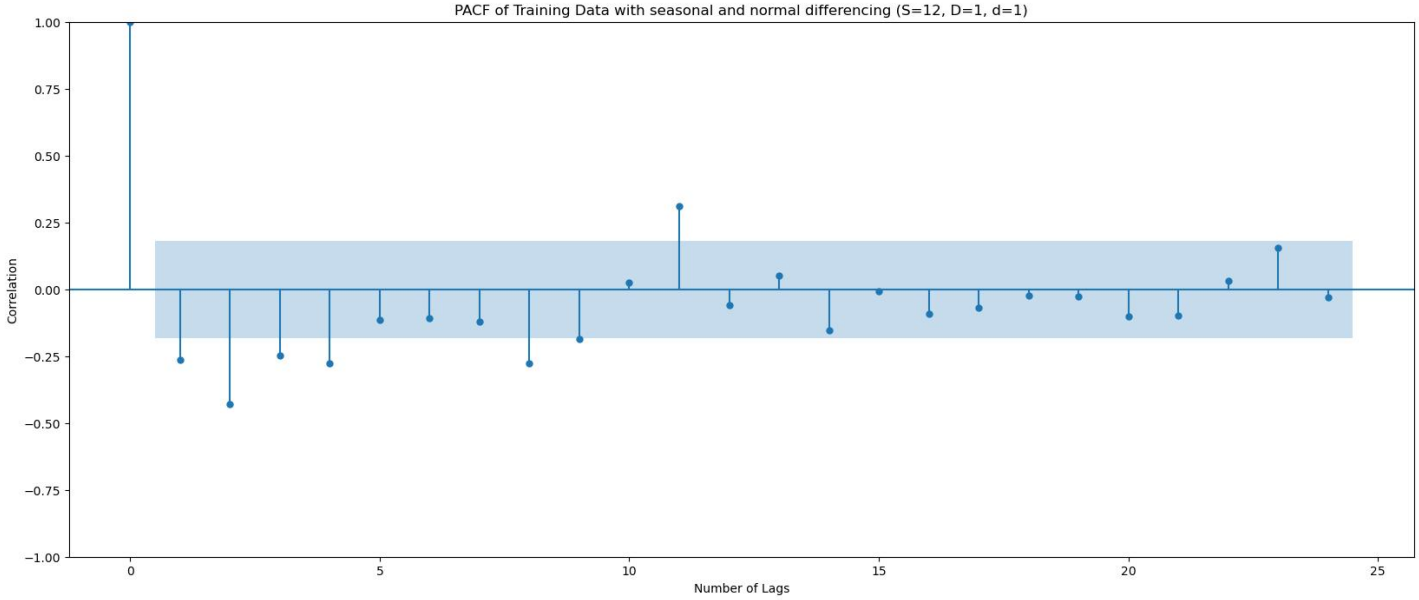

Figure 37 :ACF of Training Data with seasonal and normal differencing (S=12, D=1, d=1)

- 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 seasonal 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.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	130
Model:	SARIMAX(4, 1, 2)x(0, 1, [1], 12)	Log Likelihood	-438.007
Date:	Sun, 17 Nov 2024	AIC	892.013
Time:	12:39:25	BIC	913.013
Sample:	01-01-1980 - 10-01-1990	HQIC	900.517
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7924	0.121	-6.563	0.000	-1.029	-0.556
ar.L2	0.0454	0.142	0.319	0.750	-0.234	0.325
ar.L3	-0.2352	0.149	-1.580	0.114	-0.527	0.057
ar.L4	-0.1870	0.109	-1.719	0.086	-0.400	0.026
ma.L1	0.1467	456.900	0.000	1.000	-895.362	895.655
ma.L2	-0.8533	389.846	-0.002	0.998	-764.937	763.231
ma.S.L12	-0.5378	0.085	-6.292	0.000	-0.705	-0.370
sigma2	299.4903	1.37e+05	0.002	0.998	-2.68e+05	2.69e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	0.01
Prob(Q):	0.94	Prob(JB):	0.99
Heteroskedasticity (H):	0.58	Skew:	-0.00
Prob(H) (two-sided):	0.11	Kurtosis:	3.05

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

Table 21 : Manual SARIMA Model Summary for Rose Wine Sales Forecasting

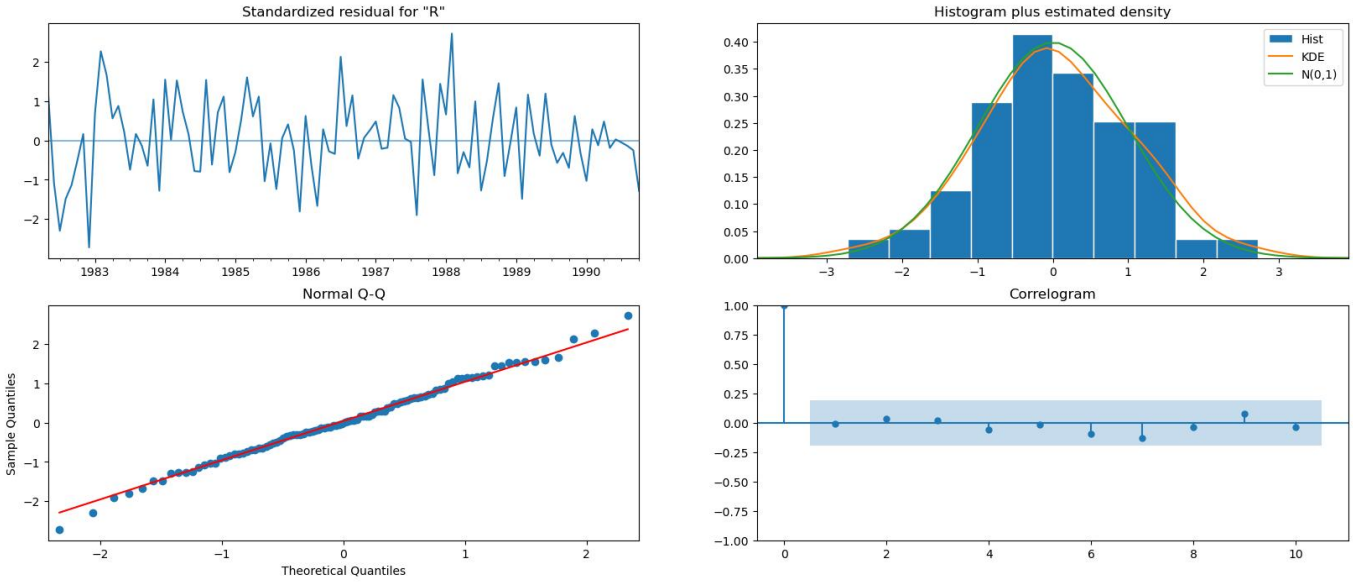

Figure 38 :diagnostics plot Manual SARIMA

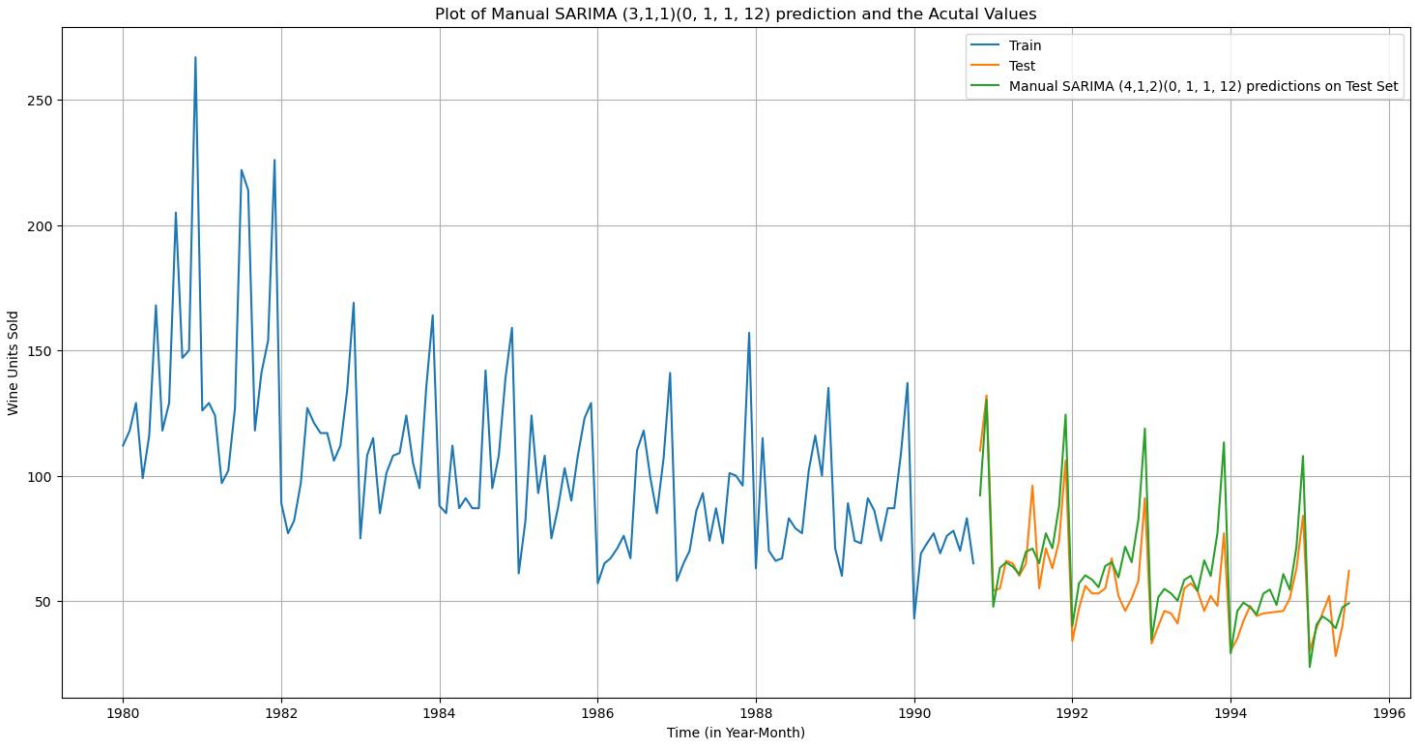

Figure 39 :Plot of Manual SARIMA (3,1,1)(0, 1, 1, 12) prediction and the Actual Values

RMSE: 12.625446085422606
MAPE: 17.26837893363303

	Test RMSE	MAPE
Auto ARIMA (2,1,4)	31.154528	60.924115
Manual ARIMA(2,1,2)	30.448610	60.265839
Auto SARIMA (0,1,1)(3, 0, 2, 12)	18.287570	35.302701
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	12.625446	17.268379

Table 22 : Performance of ARIMA models (Rose)

Performance of the models built

	Test RMSE	MAPE
Alpha=0.1,Beta=0.25,Gamma=0.85,Triple Exponential Smoothing	8.377119	NaN
2pointTrailingMovingAverage	11.801043	NaN
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	12.625446	17.268379
Alpha=0.062,Beta=0.018,Gamma=0.0,Triple Exponential Smoothing	15.273169	NaN
4pointTrailingMovingAverage	15.367244	NaN
6pointTrailingMovingAverage	15.862398	NaN
9pointTrailingMovingAverage	16.341947	NaN
Alpha=1.49e-08, Beta=7.75e-11, Double Exponential Smoothing	17.355736	NaN
Linear Regression	17.355804	NaN
Auto SARIMA (0,1,1)(3, 0, 2, 12)	18.287570	35.302701
Alpha=0.127,SimpleExponentialSmoothing	29.223870	NaN
Manual ARIMA(2,1,2)	30.448610	60.265839
Auto ARIMA (2,1,4)	31.154528	60.924115
SimpleAverageModel	52.412291	NaN

Compare the performance of the models

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

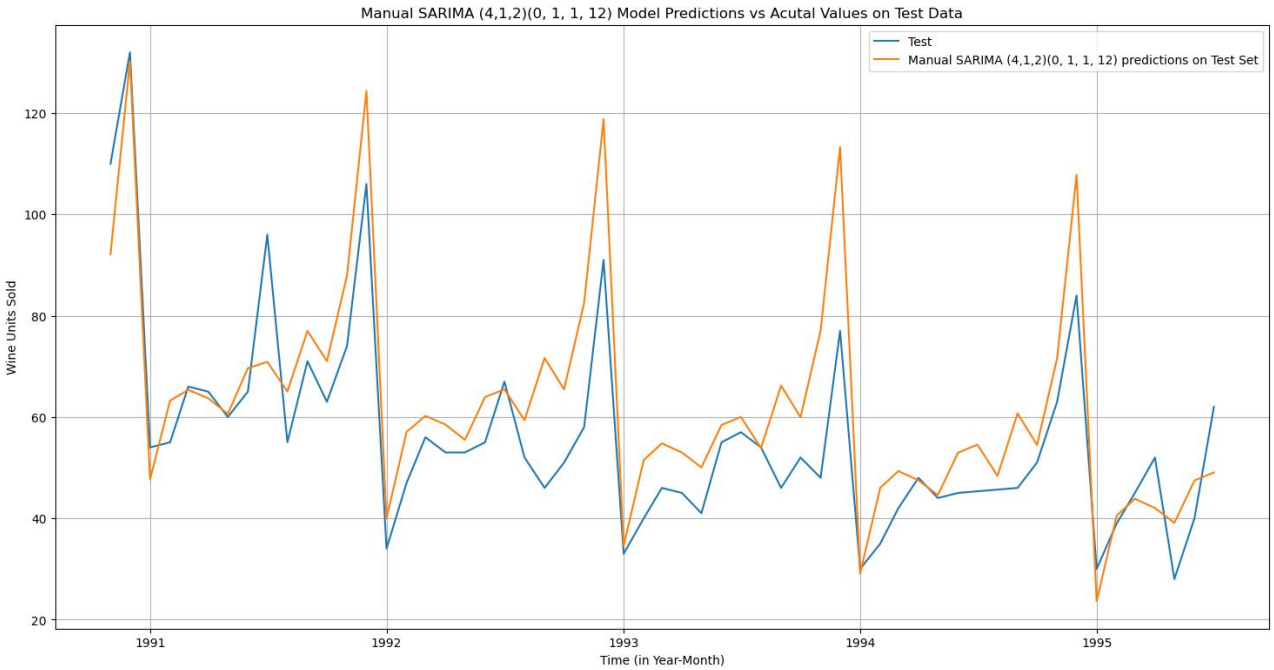

Figure 40 :Manual SARIMA (4,1,2)(0, 1, 1, 12) Model Predictions vs Actual Values on Test Data

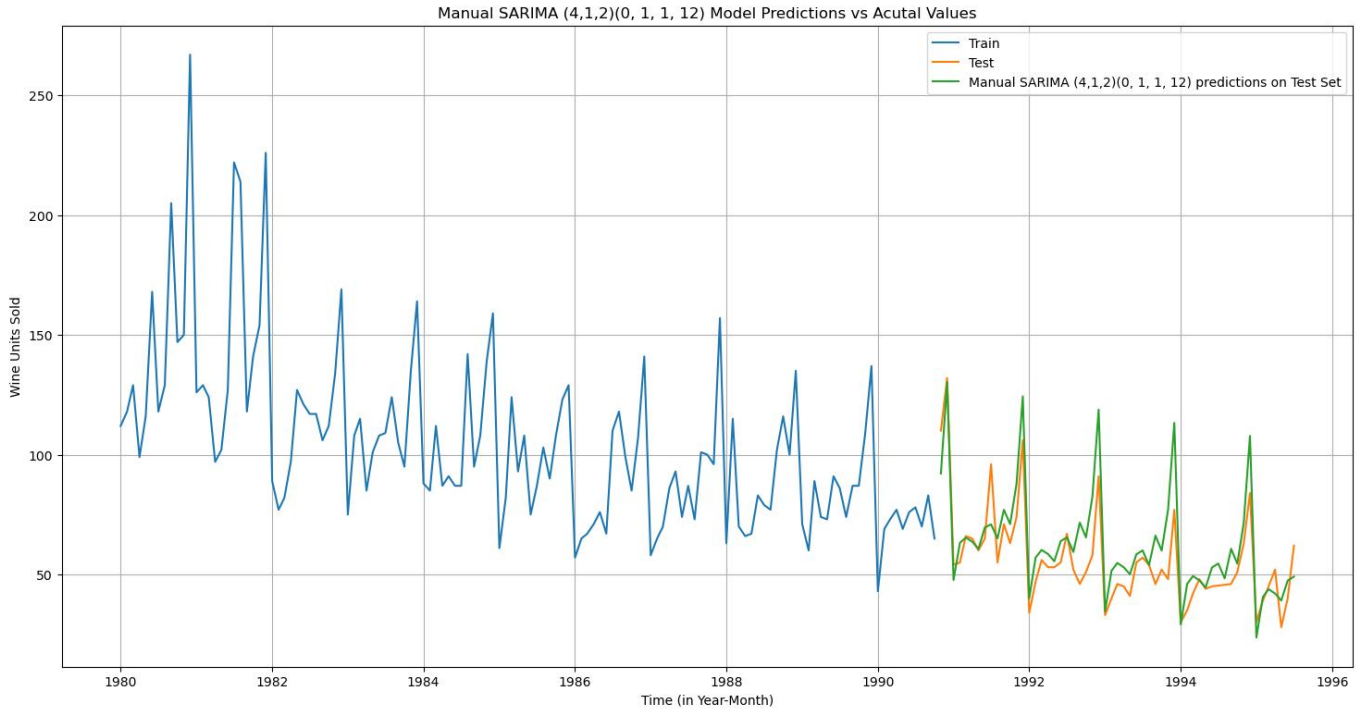

Figure 41 :Manual SARIMA (4,1,2)(0, 1, 1, 12) Model Predictions vs Actual Values

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(4, 1, 2)x(0, 1, [1], 12)	Log Likelihood	-658.936
Date:	Sun, 17 Nov 2024	AIC	1333.871
Time:	12:39:29	BIC	1358.423
Sample:	01-01-1980	HQIC	1343.841
	- 07-01-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8240	0.083	-9.941	0.000	-0.986	-0.662
ar.L2	0.0469	0.107	0.437	0.662	-0.164	0.257
ar.L3	-0.2148	0.110	-1.956	0.051	-0.430	0.000
ar.L4	-0.1693	0.078	-2.182	0.029	-0.321	-0.017
ma.L1	0.1565	82.334	0.002	0.998	-161.215	161.528
ma.L2	-0.8435	69.448	-0.012	0.990	-136.959	135.272
ma.S.L12	-0.5418	0.061	-8.898	0.000	-0.661	-0.422
sigma2	225.0308	1.85e+04	0.012	0.990	-3.61e+04	3.65e+04

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	3.11
Prob(Q):	0.88	Prob(JB):	0.21
Heteroskedasticity (H):	0.24	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.68

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

Table 23 : Full model manual SARIMA summary(Rose)

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	48.269778	15.053216	18.766018	77.773539
1995-09-01	44.985412	15.830585	13.958036	76.012787
1995-10-01	45.474816	15.888941	14.333064	76.616568
1995-11-01	54.808063	15.898941	23.646712	85.969414
1995-12-01	81.906511	15.913690	50.716251	113.096772

Table 24 : Predicted Manual SARIMA full data summary(Rose)

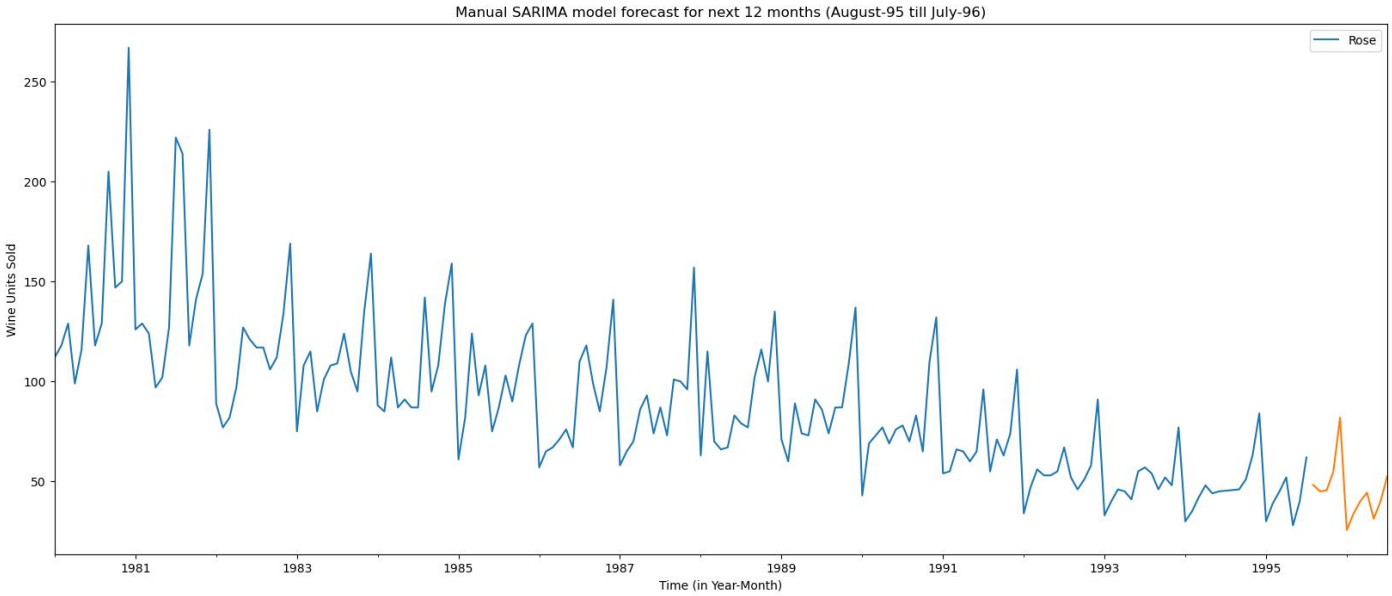

Figure 42 :Manual SARIMA model forecast for next 12 months (August-95 till July-96)

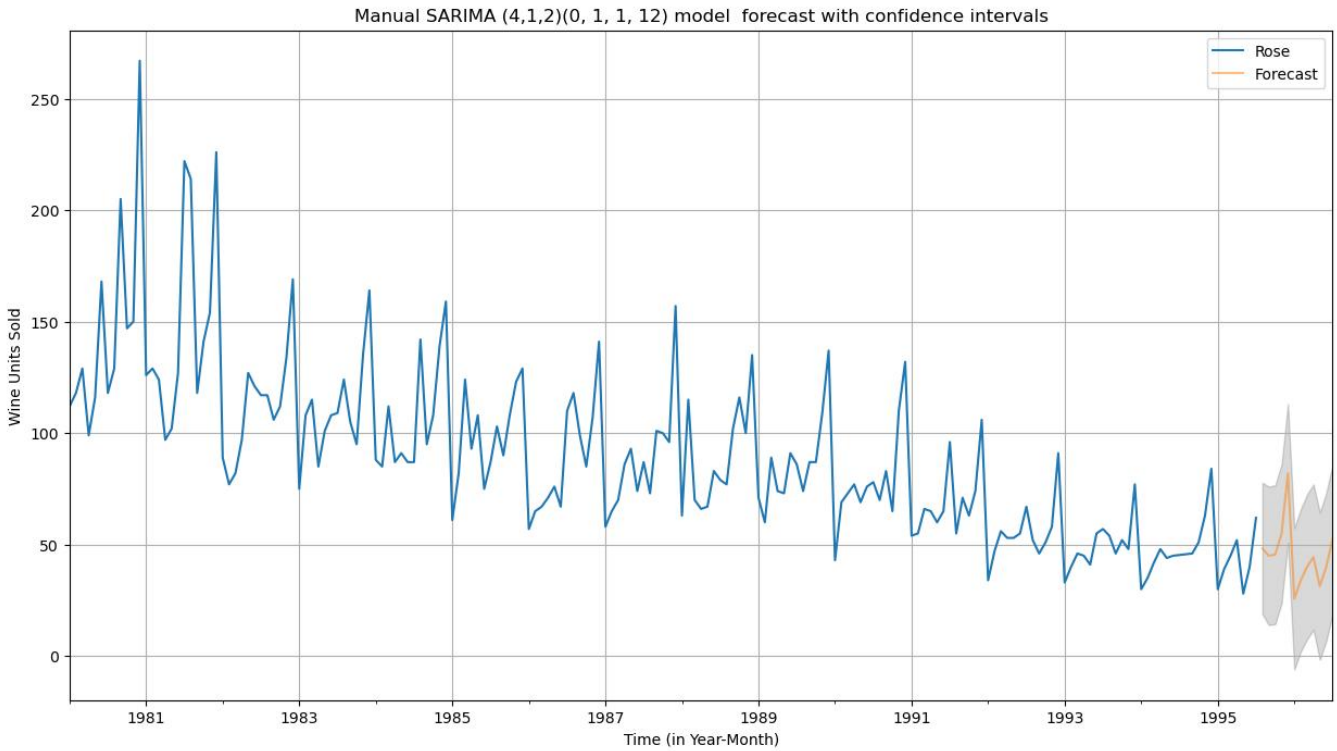

Figure 43 :Manual SARIMA (4,1,2)(0, 1, 1, 12) model forecast with confidence intervals

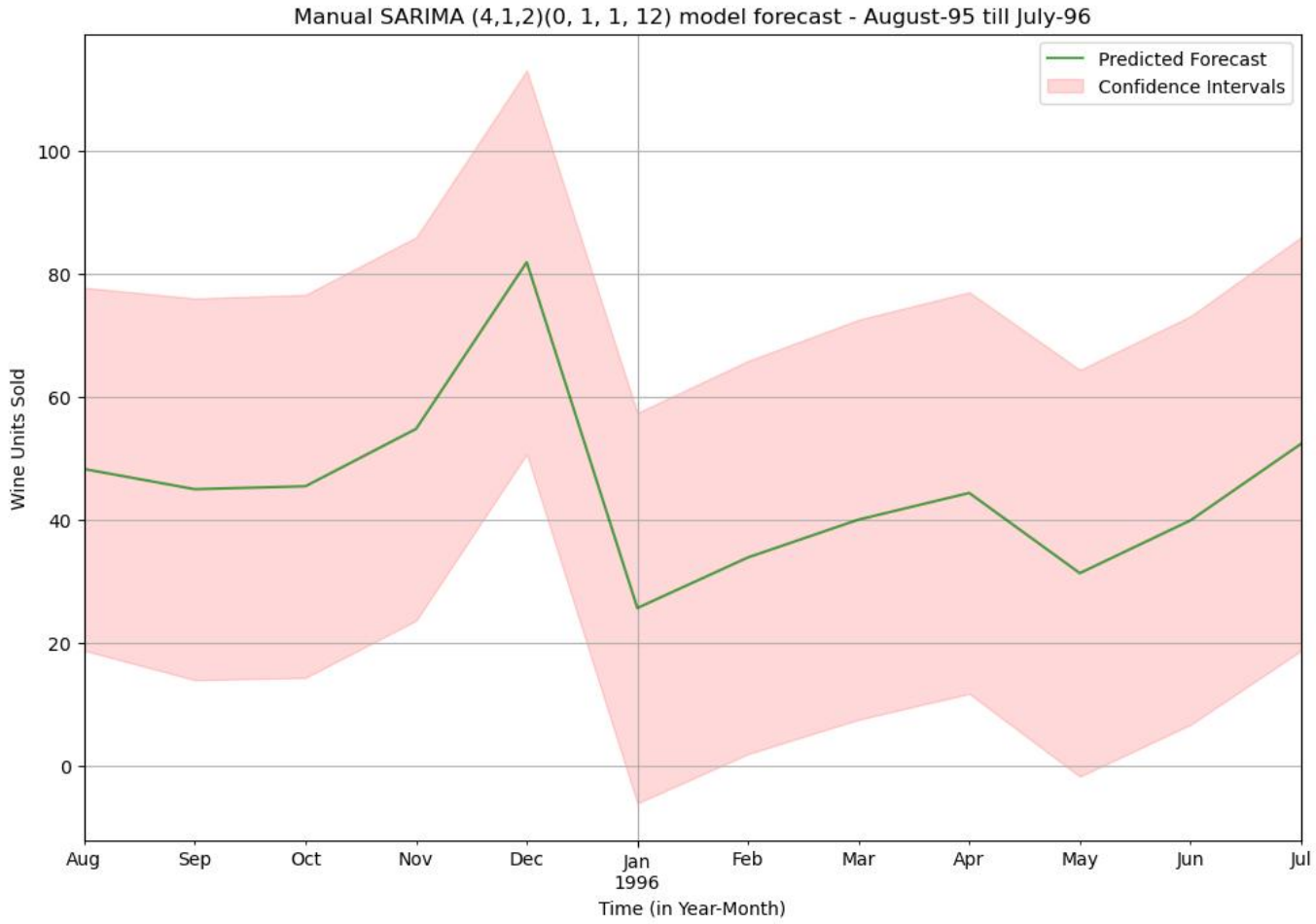

Figure 44 :Manual SARIMA (4,1,2)(0, 1, 1, 12) model forecast - August-95 till July-96

Actionable Insights & Recommendations

- Allocate significant marketing budgets for October to December to capitalize on peak demand.
- Introduce promotions or new products to drive demand in Q1 (January–March).
- Introduce seasonal wine varieties to maintain engagement year-round.
- Manual SARIMA (4,1,2)(0,1,1,12): Achieves the lowest RMSE (12.63) and a significantly lower MAPE (17.27%), making it the most accurate model for forecasting.
- Triple Exponential Smoothing (Alpha=0.1, Beta=0.25, Gamma=0.85): Offers excellent performance (RMSE: 8.38) and is computationally simpler. Suitable if SARIMA implementation is constrained.

- Adopt Manual SARIMA for long-term forecasts to drive business decisions regarding inventory, marketing, and sales strategies.
- Use Triple Exponential Smoothing for operational tasks requiring quick, reliable forecasts.
- Discontinue simpler models like Simple Average and Linear Regression due to poor predictive performance.
- The Manual SARIMA (4,1,2)(0,1,1,12) model effectively captures the trends and seasonality of the time series, as evident from its performance metrics on the test data.RMSE for the model on test data is 12.63, which indicates relatively accurate predictions compared to other models in the analysis.
- While the full dataset RMSE is slightly higher at 32.22, this includes noise and long-term residual variability, suggesting the model performs well overall.
- The forecasted values for the 12 months (August 1995 – July 1996) align closely with observed trends and expected seasonal fluctuations in the wine sales data.
- Confidence intervals for the forecast provide a reasonable range for uncertainty, ensuring robustness in prediction and planning.
- Diagnostic statistics (Jarque-Bera test, Ljung-Box test) suggest that residuals are approximately normally distributed with no significant autocorrelation, validating the model's assumptions.
- Heteroskedasticity test indicates some variation in residual variance, which might slightly affect confidence in predictions.
- The model forecasts a steady increase in wine sales towards the end of 1995 followed by a decline in early 1996, consistent with seasonal trends.

Recommendations

- Continue refining the model by monitoring performance as new data becomes available.
- Consider incorporating external variables (e.g., pricing, promotions, economic indicators) to improve predictive accuracy.

- Use the forecasted values and confidence intervals to guide business decisions and mitigate risks associated with demand variability.

EXECUTIVE SUMMARY

This report presents a comprehensive analysis of ABC Estate Wines' historical sales data spanning the 20th century, focusing on sparkling and rose wine varieties. By examining key sales trends, patterns, and influencing factors, we aim to provide actionable insights to guide strategic decision-making and optimize future sales strategies.

Through advanced data analytics and forecasting techniques, our findings highlight how consumer preferences, economic conditions, and market dynamics have shaped wine sales over the decades. Notably, seasonal patterns and external factors such as global events and technological advancements have significantly influenced the demand for sparkling and rose wines.

Leveraging these insights, we propose targeted strategies to capitalize on emerging market opportunities, enhance product positioning, and improve forecasting accuracy. By adopting data-driven approaches, ABC Estate Wines can sustain its competitive edge and position itself for growth in an ever-evolving wine industry.

INTRODUCTION

The wine industry has been a dynamic and competitive sector, influenced by shifting consumer preferences, economic trends, and cultural factors. As a prominent player, ABC Estate Wines has maintained a rich history of crafting quality wines that appeal to a diverse clientele. To remain competitive in the modern market, understanding historical sales trends and identifying future opportunities are imperative.

This project focuses on analyzing historical sales data from ABC Estate Wines for sparkling and rosé wines. Covering the entirety of the 20th century, the datasets provide a unique opportunity to uncover trends and patterns across decades, highlighting factors that have influenced wine consumption.

DATA DESCRIPTION

The CSV files "Sparkling.CSV" contains Sparkling wines sold from time period of (1980-01 to 1995-07).

The files contain two columns YearMonth and Sparkling wines sold.

YearMonth	Represents the year and month in which the sales were recorded
Sparkling	Number of wine units sold

Data Description

We read the data and checked for the data types present in the CSV file.

```
YearMonth    object
Sparkling     int64
dtype: object
```

Table 25: Data Types (Sparkling)

YearMonth column is not seen as a date object. So we converted it into index column after converting it into date-time format.

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

Table 26: First 5 rows(Sparkling)

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Table 27: Bottom 5 rows(Sparkling)

The shape of the dataframe is (187, 1), 187 rows and 1 column as we have changed YearMonth as Index.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
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 28 : Basic Information of Data set(Sparkling)

Statistical Summary

	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000
50%	1874.000
75%	2549.000
max	7242.000

Table 29 : Statistical Summary of Data set(Sparkling)

- The mean sales of sparkling wine are 2402.42 units, indicating that on average, the sales volume is fairly high.
- Sales fluctuate between a minimum of 1070 units and a maximum of 7242 units. The wide range indicates variability in demand, which could be tied to seasonality, promotional events, or other external factors.

Exploratory Data Analysis

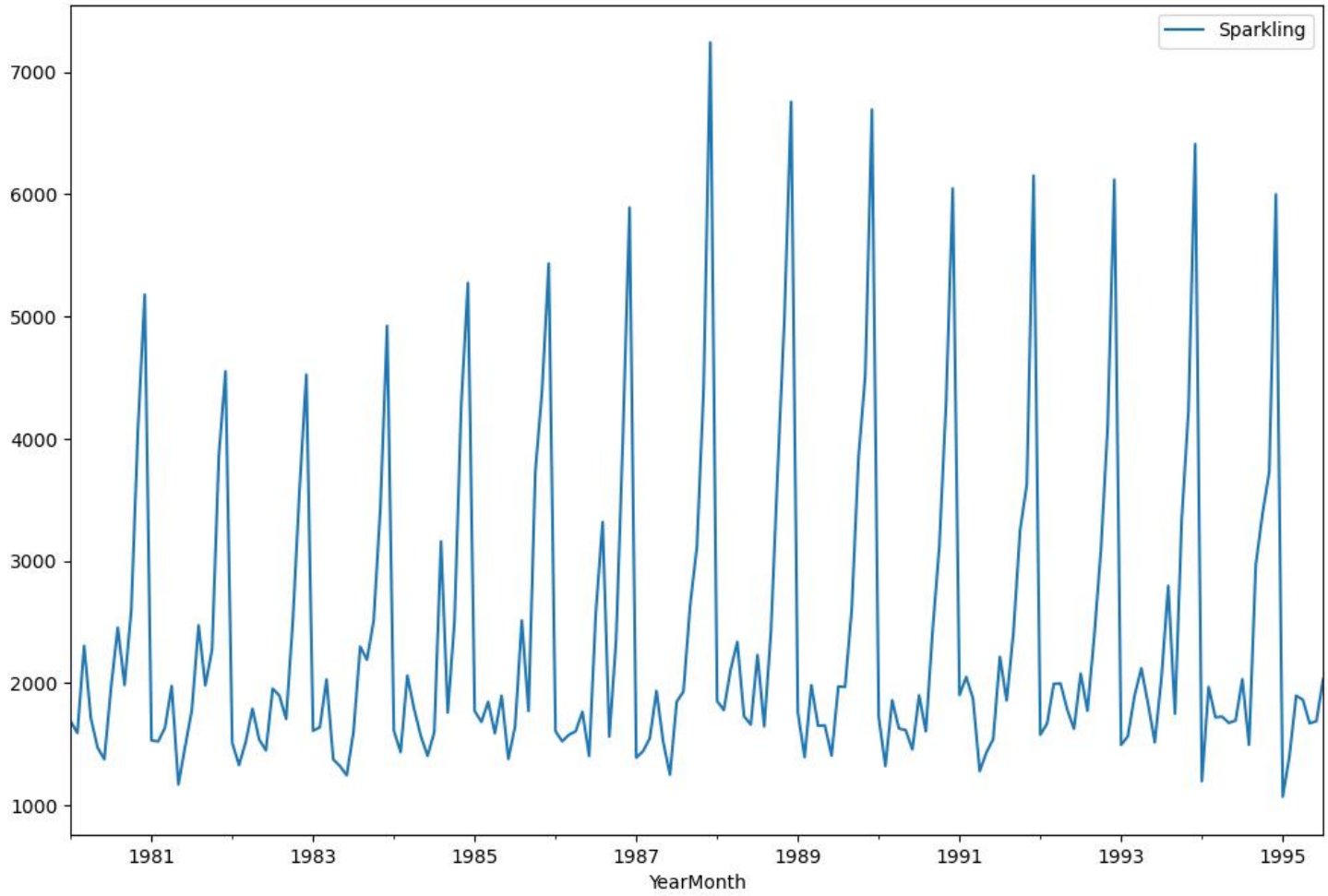

Figure 45 :Time Series Plot

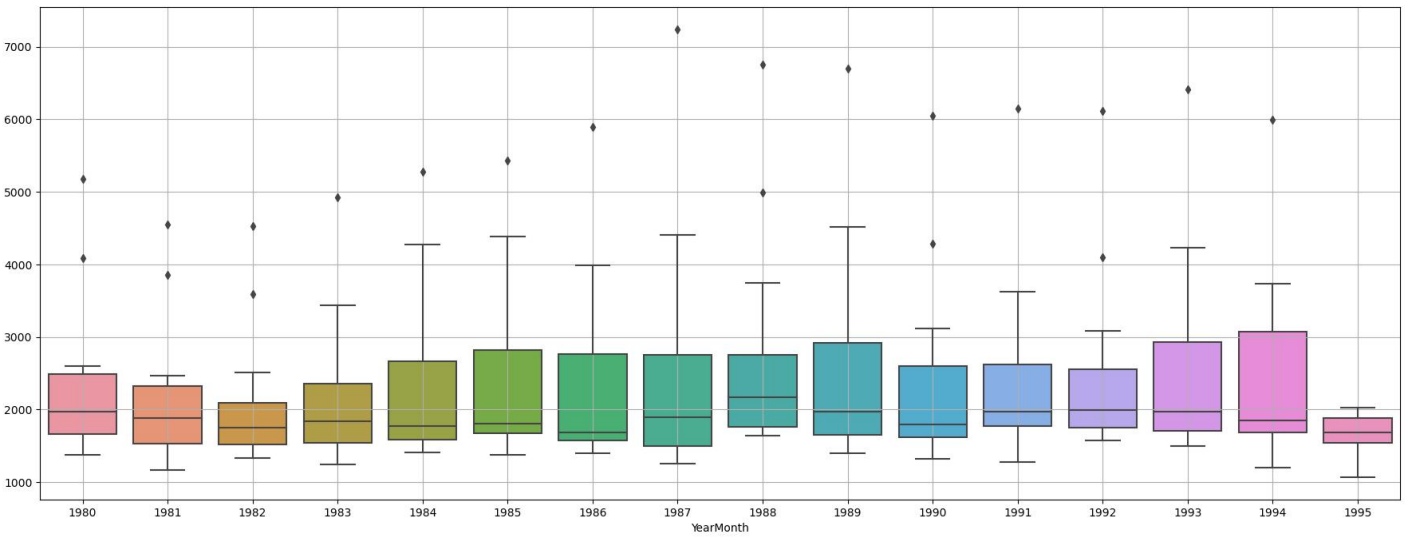

Figure 46 :Yearly Box Plot

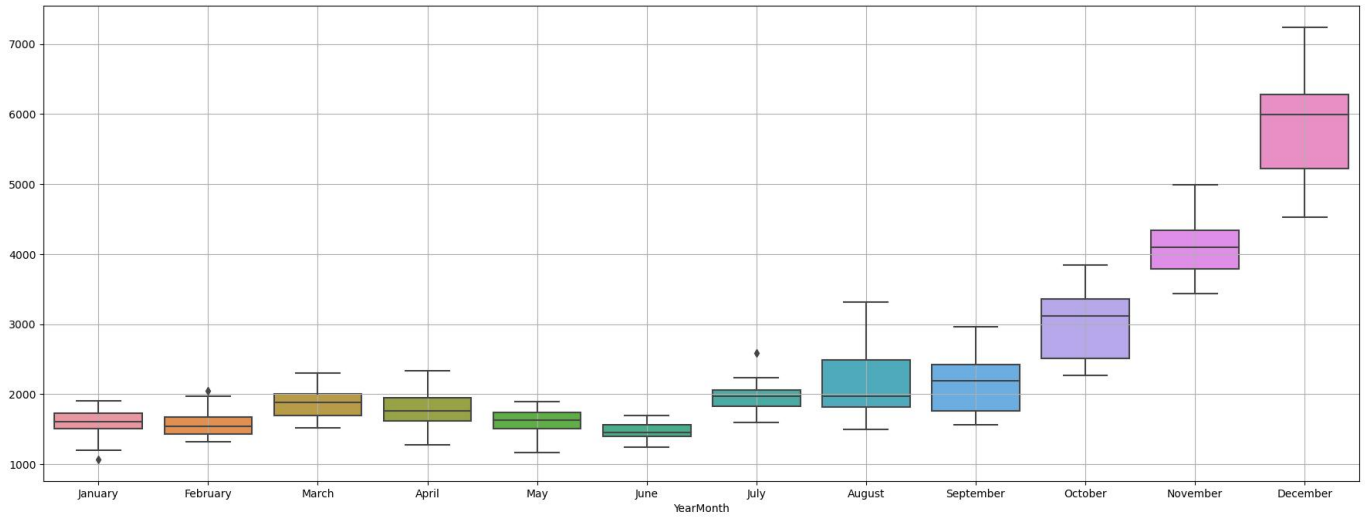

Figure 47 :Monthly Box Plot

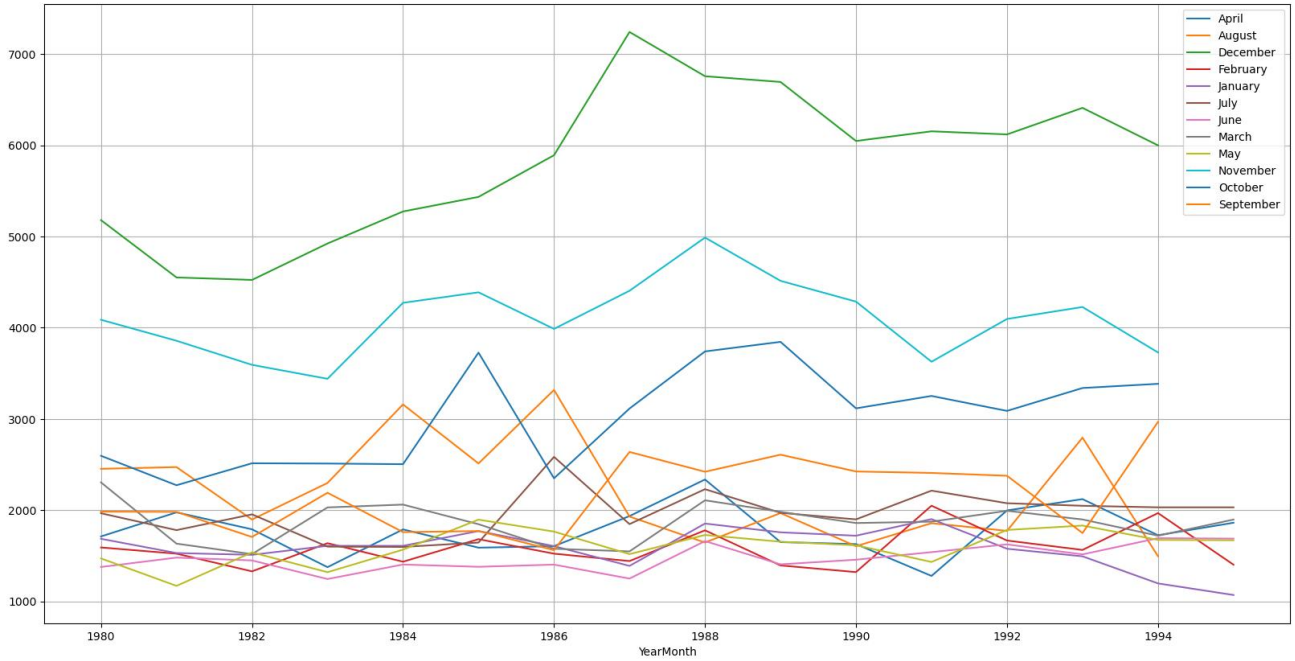

Figure 47: Monthly Sales across Years

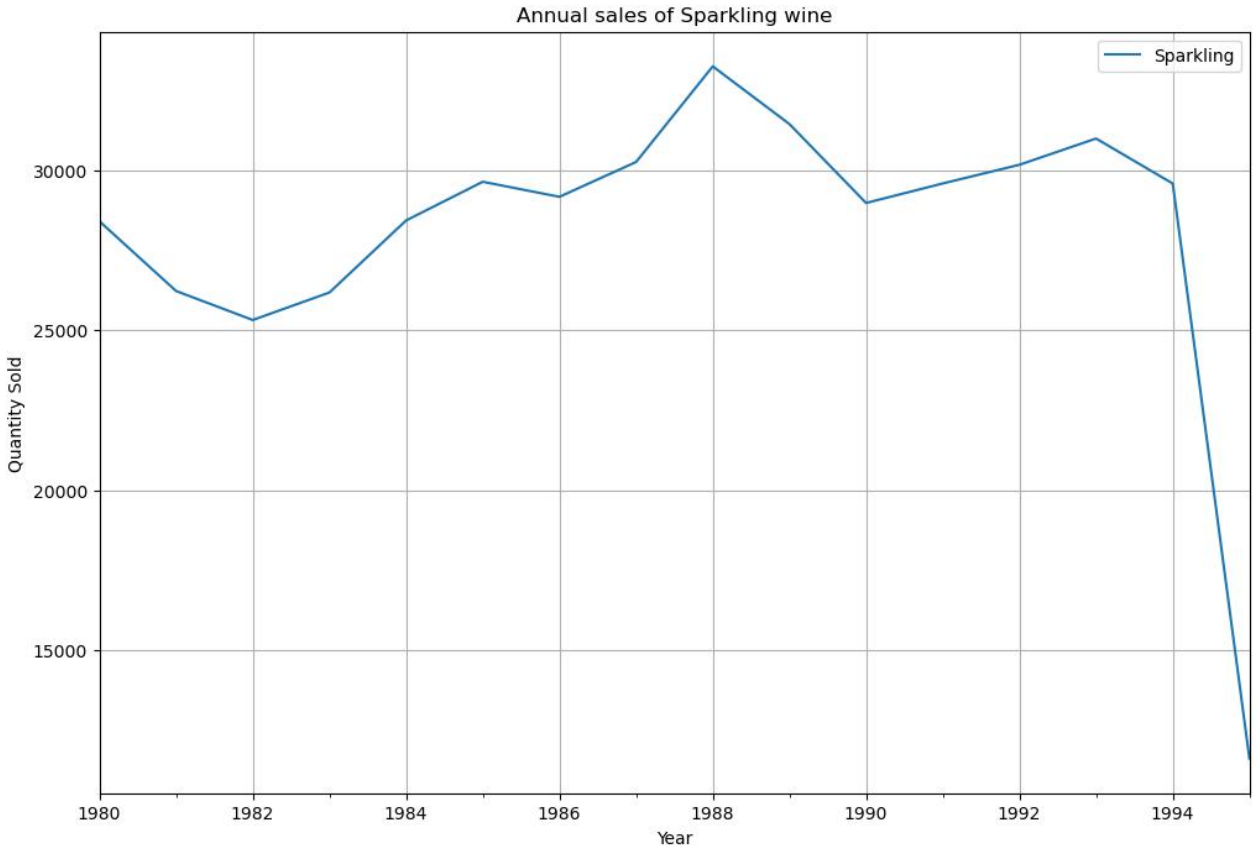

Figure 48: Annual sales of Sparkling wine

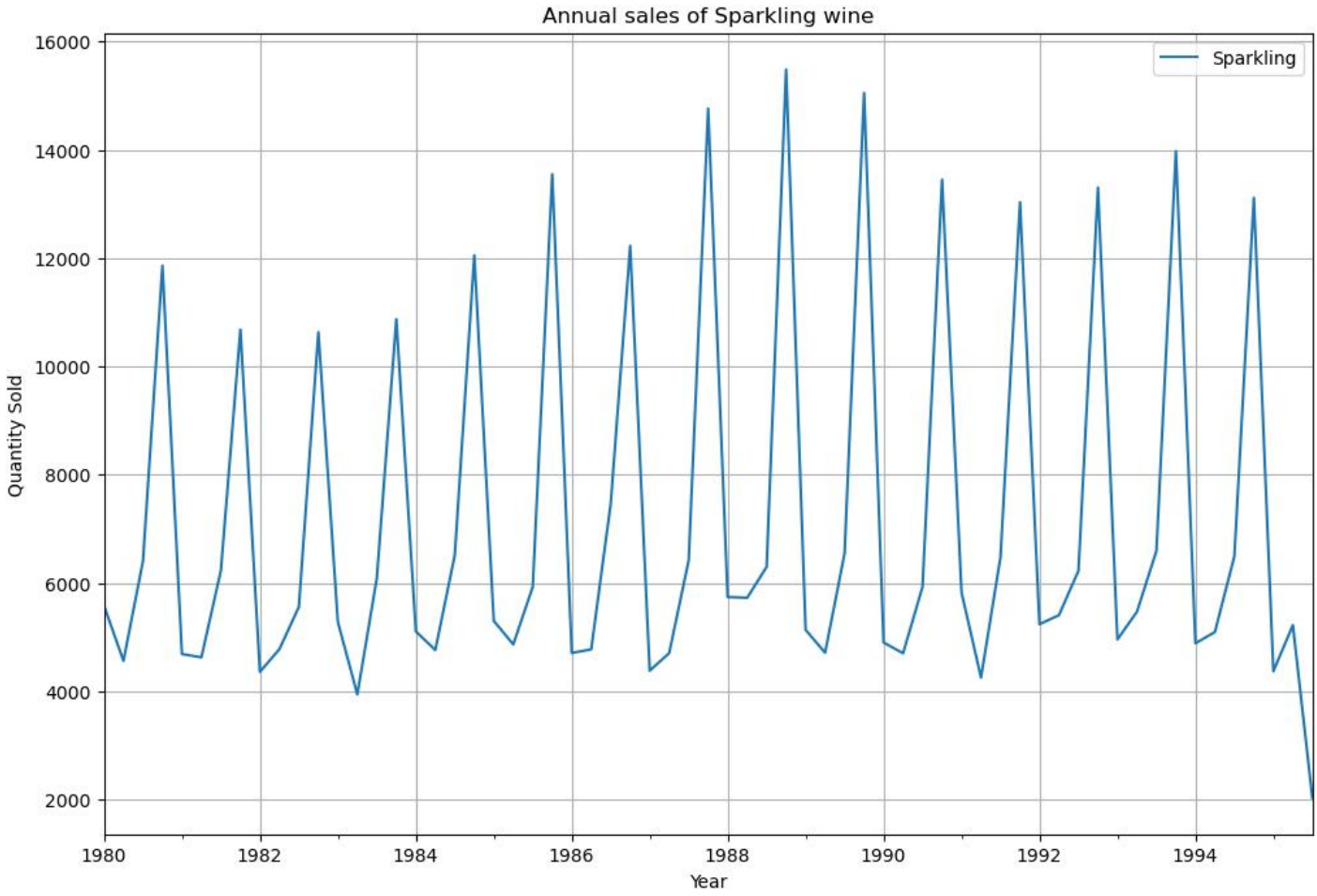

Figure 49: Quarter Sales across Years(Sparkling)

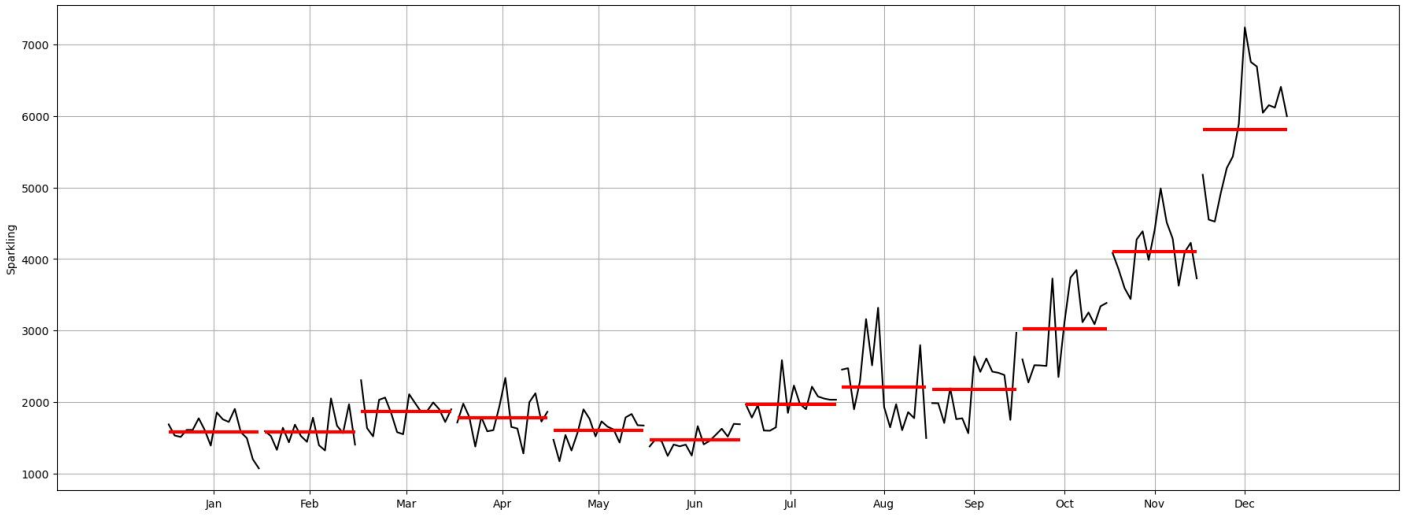

Figure 50: Seasonal Patterns (Sparkling)

Annual Trends:

- The annual sales for sparkling wine show fluctuations from 1980 to 1995. The sales for 1980 to 1984 hover between 25,000 to 28,000 units annually, with 1981 and 1982 being on the lower end and 1983 and 1984 reaching towards the higher end.
- There is a noticeable growth trend from 1985 to 1989, with sales peaking at 33,246 units in 1988, which represents a significant rise from 1980-1984.
- However, after 1989, there is a decline in sales, with 1990-1994 showing figures ranging from 28,977 to 29,584 units annually.
- 1995 shows a sharp decline in sales, dropping to 11,620 units. This could indicate a significant market shift or external factors affecting demand.

Seasonal Trends:

- Monthly sales fluctuate throughout the year. The data suggests that the end of year months (November-December) tend to have higher sales figures for sparkling wine, peaking notably in December:
- In 1980, sales reach 11,862 in December, marking the highest monthly figure for the year. This trend is also observed in other years, where December tends to outperform other months, suggesting a seasonal spike in demand, likely due to holiday sales.
- March and June months tend to show lower sales in most years, with March 1980 reaching 5,581 and June 1980 showing 4,560. These months might be off-peak periods with fewer promotional activities or less seasonal demand.

Fluctuations in Specific Years:

- The year 1984 is a standout with 28,431 units in December, a peak that could be linked to a successful promotional push or strong holiday sales.
- The year 1987 also saw a high in December with 30,258 units, reinforcing the seasonal trends.

- 1989 saw a slight dip in December sales compared to 1988, but still performed well with 31,443 units, showing consistency in holiday demand.

Time Series Decomposition

Additive Model

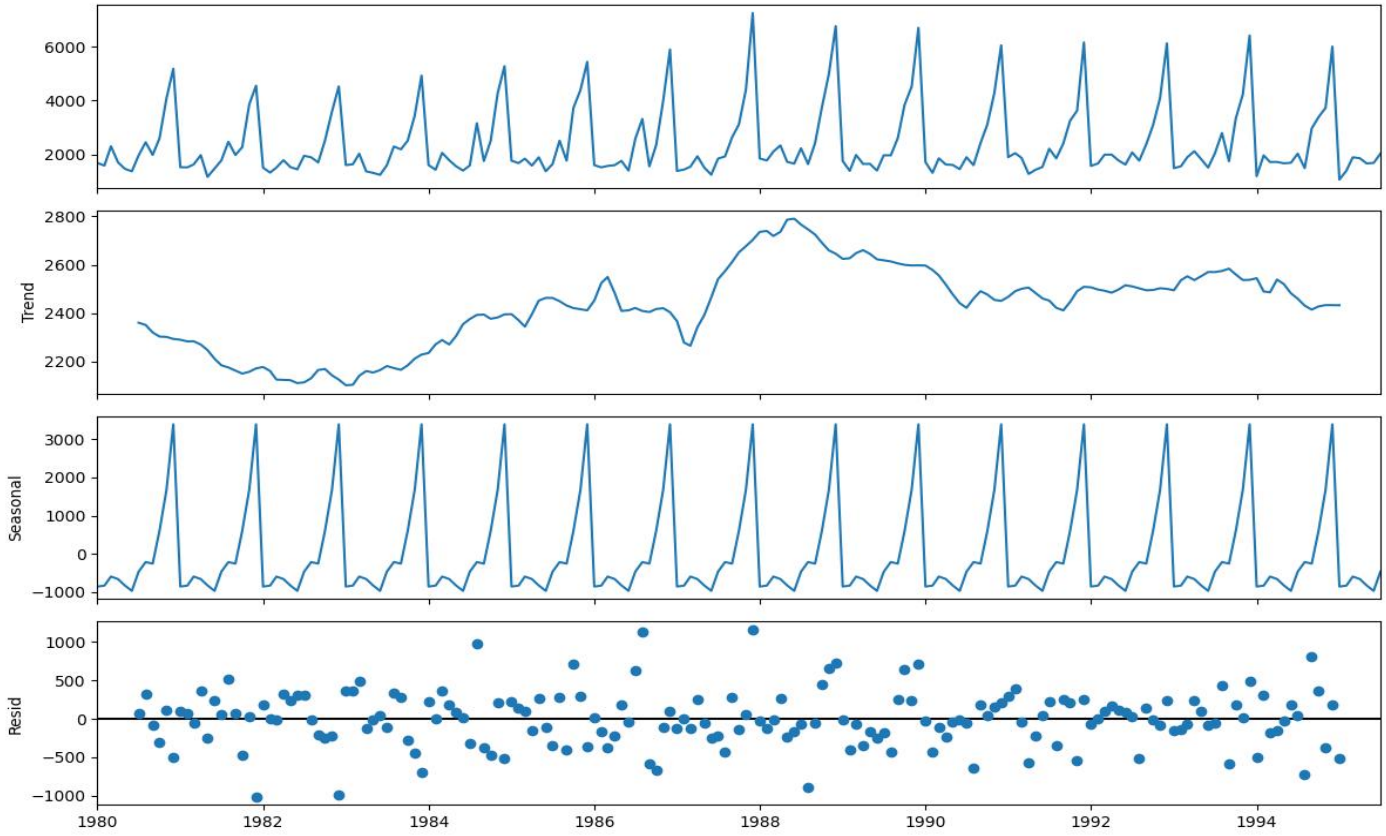

Figure 51: Additive Model Decomposition (Sparkling)

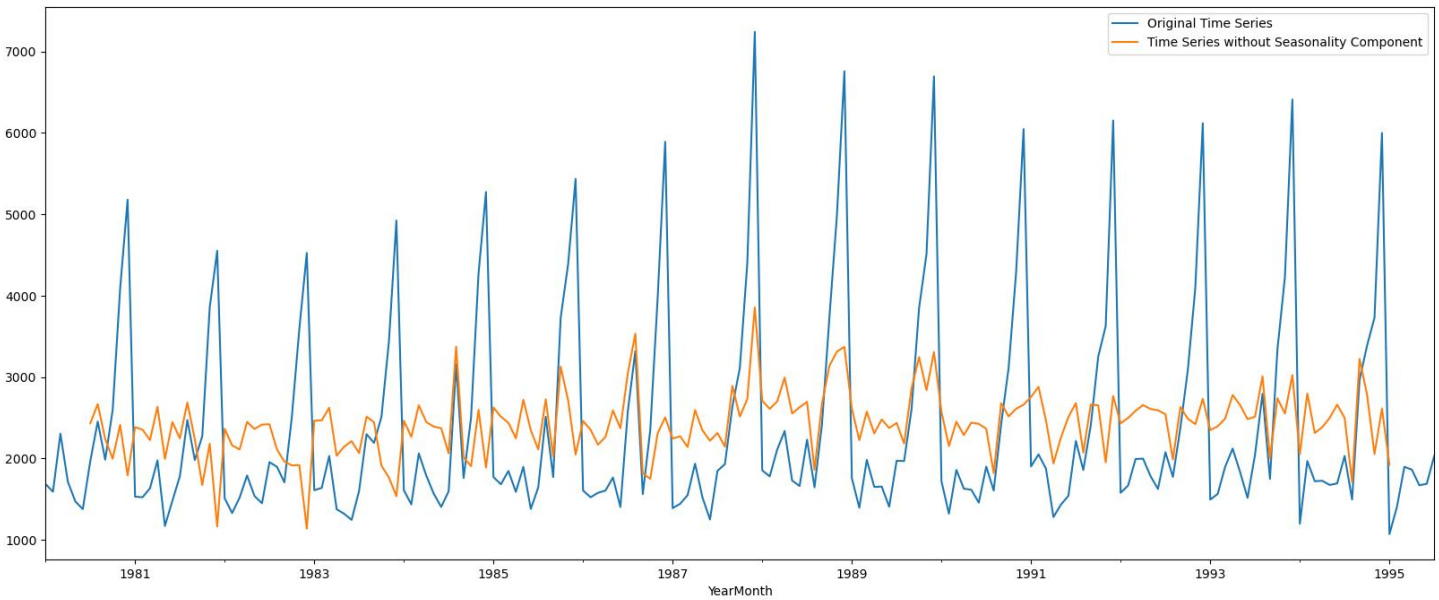

Figure 52: Comparison of Original and Deseasonalized Time Series (Sparkling)

Multiplicative Model

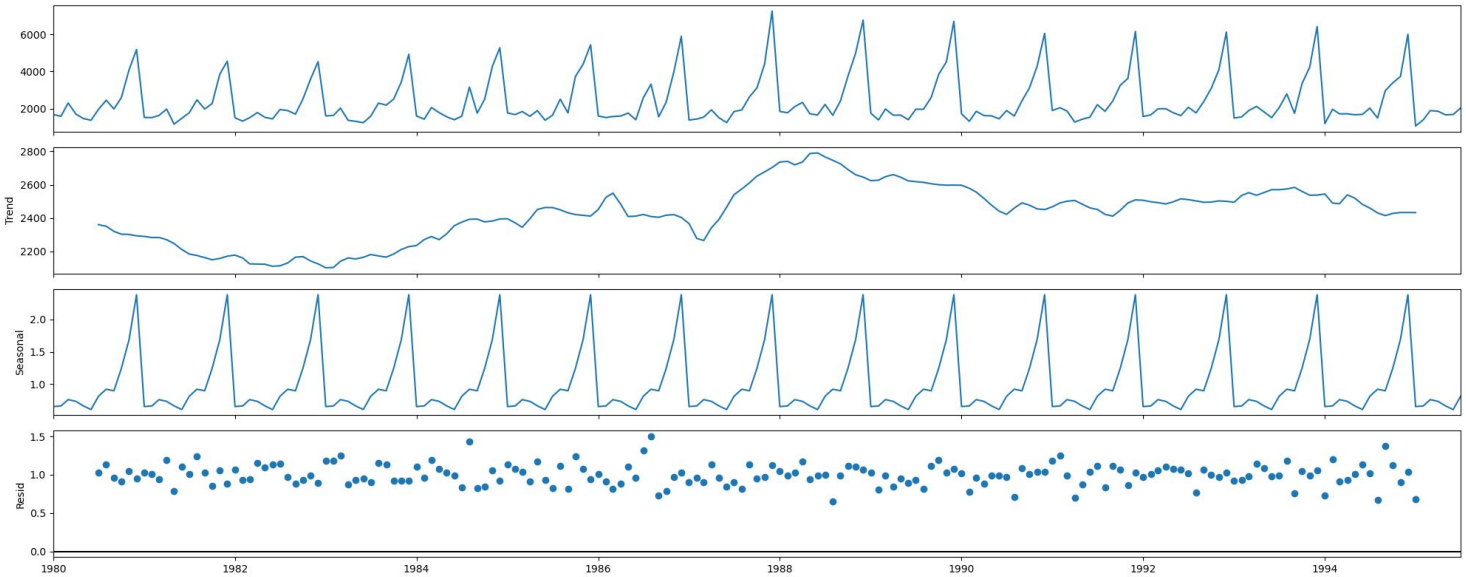

Figure 53 :Multiplicative Model Decomposition(Sparkling)

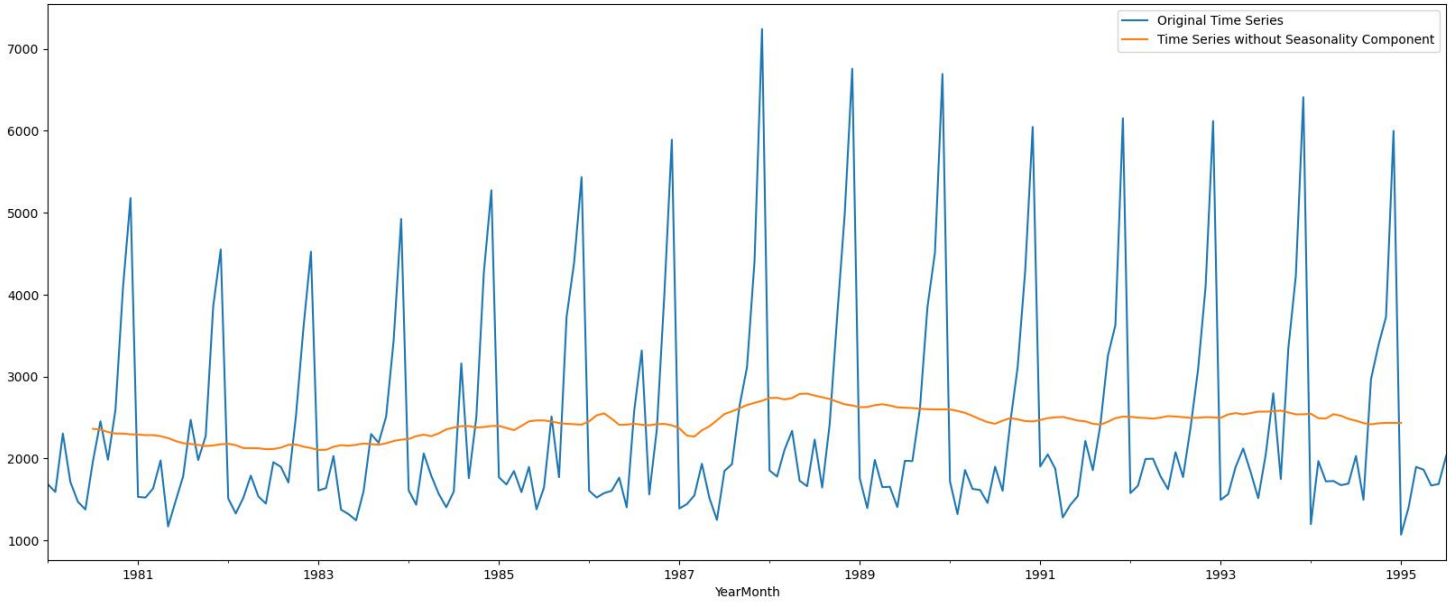

Figure 54 :Comparison of Original and Deseasonalized Time Series(Sparkling)

Train-test split

- The data is split into two distinct subsets: training and test datasets. This division is essential for developing and evaluating predictive models.

- Training Set:This subset includes the first 70% of the available data. It is used to fit and train the model, allowing it to learn the underlying patterns and trends from the historical sales data.

- Test Set : The remaining 30% of the data is set aside as the test set. This portion is used to assess the model's performance and generalization ability. It simulates how the model will perform when applied to future, unseen data.

By splitting the data in this way, we ensure that the model is trained on a large enough dataset, while still maintaining an independent portion for testing and validating its predictions. This process helps in assessing how well the model is likely to perform on real-world data, improving its reliability and robustness.

First few rows of Training Data

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

Last few rows of Training Data

YearMonth	Sparkling
1990-06-01	1457
1990-07-01	1899
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116

First few rows of Test Data

YearMonth	Sparkling
1990-11-01	4286
1990-12-01	6047
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874

Last few rows of Test Data

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Table 30 : Training and Test Data (Sparkling)

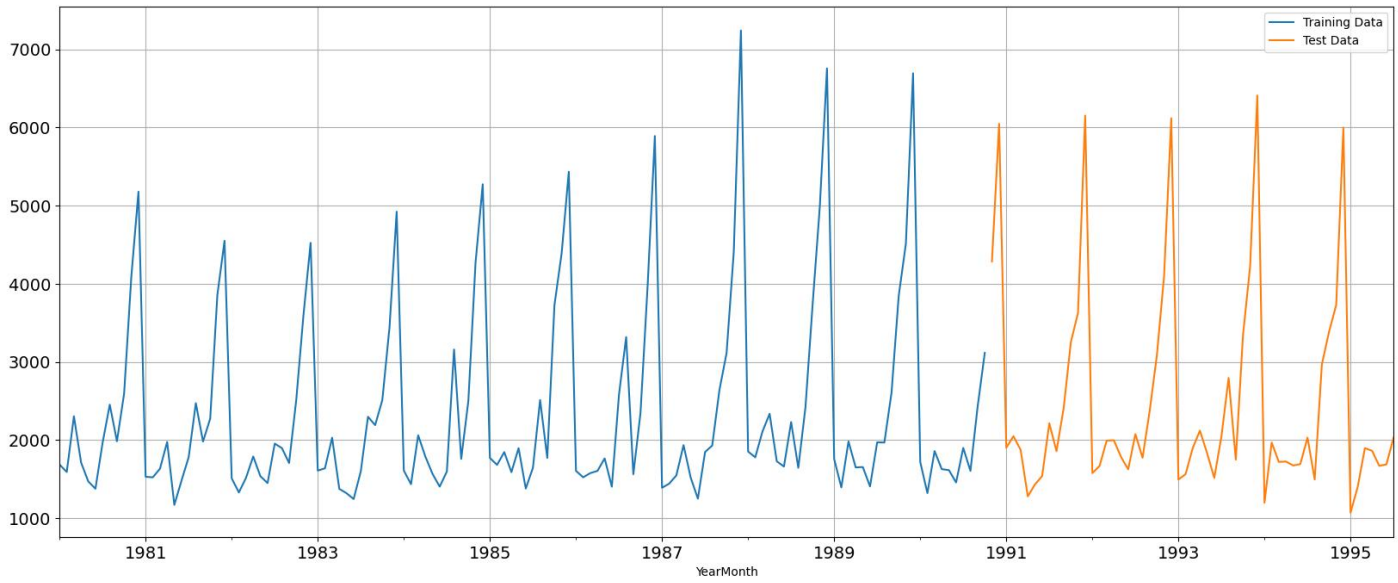

Figure 55:Training and Test Split of Sparkling Wine Sales Data

Model Building - Original Data

2. Model 1: Linear Regression

Training Time instance

Training Time instance
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130]

Test Time instance
[131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]

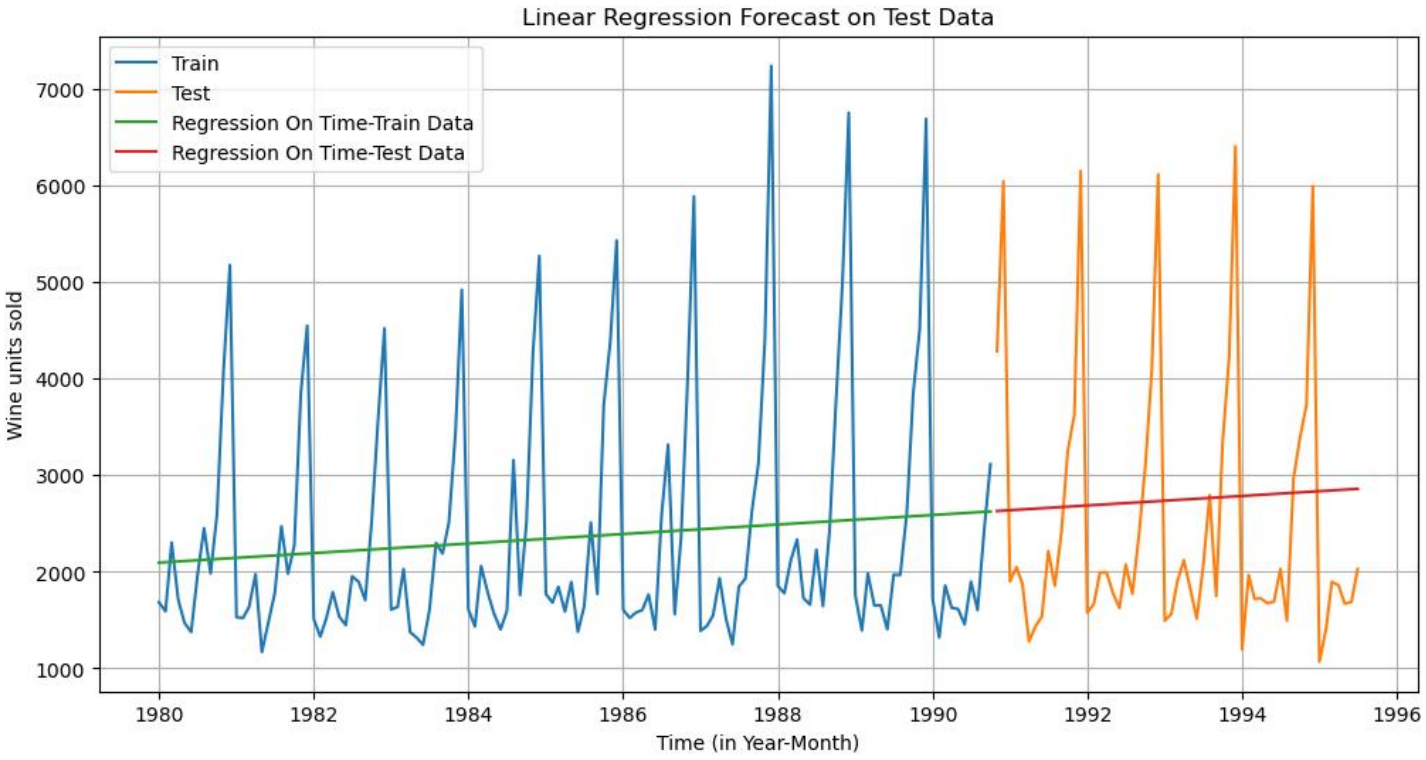

Figure 56: Linear Regression Forecast on Test Data(Sparkling)

For RegressionOnTime forecast on the Test Data, RMSE is 1392.438.

Model 2: Simple Average

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

YearMonth	Sparkling	mean_forecast
1990-11-01	4286	2361.276923
1990-12-01	6047	2361.276923
1991-01-01	1902	2361.276923
1991-02-01	2049	2361.276923
1991-03-01	1874	2361.276923

Table 31 : Mean forecast Sparkling

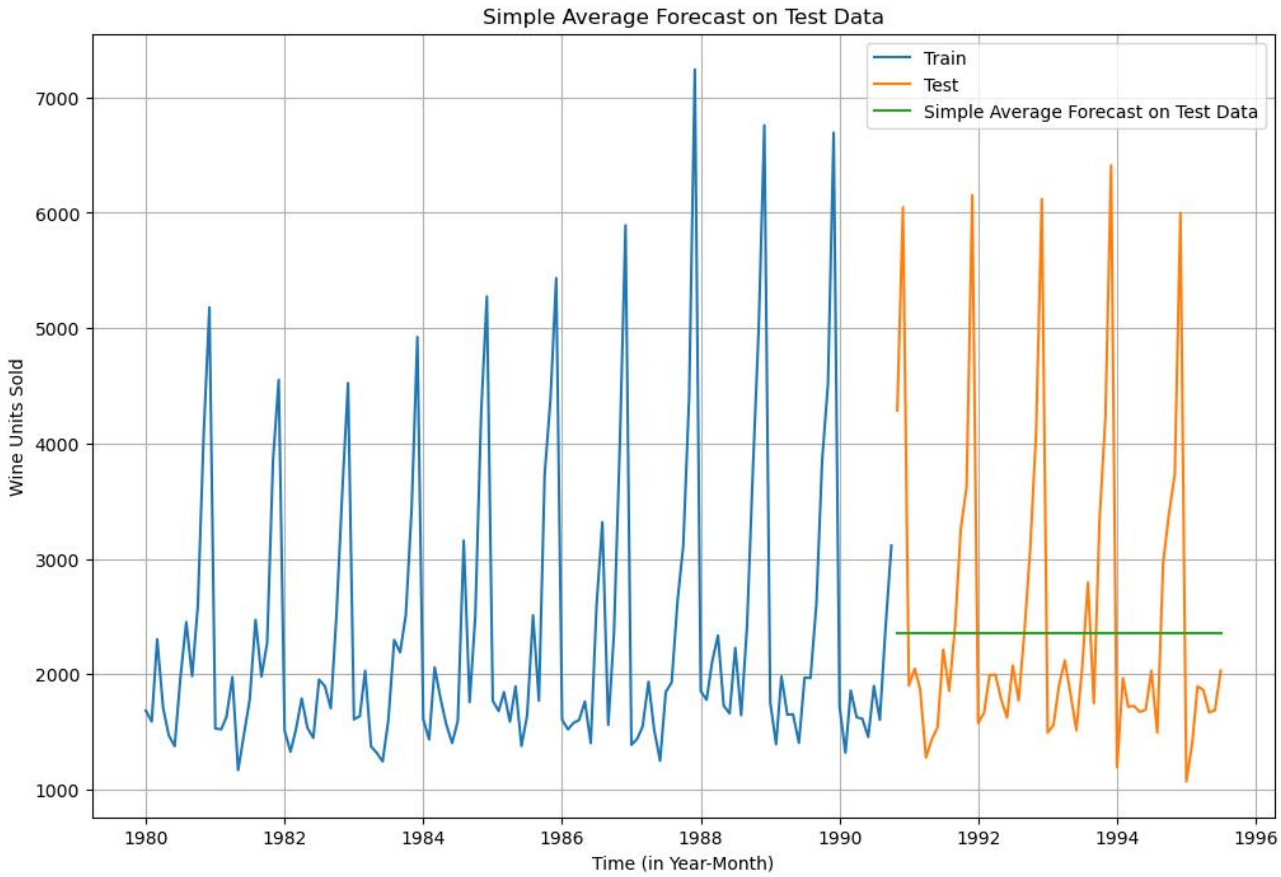

Figure 57: Comparison of Actual and Simple Average Forecast for Sparkling Wine Sales

For Simple Average forecast on the Test Data, RMSE is 1368.747.

Model 3: Moving Average(MA)

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.

YearMonth	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN

Table 32 : Trailing moving averages(Sparkling)

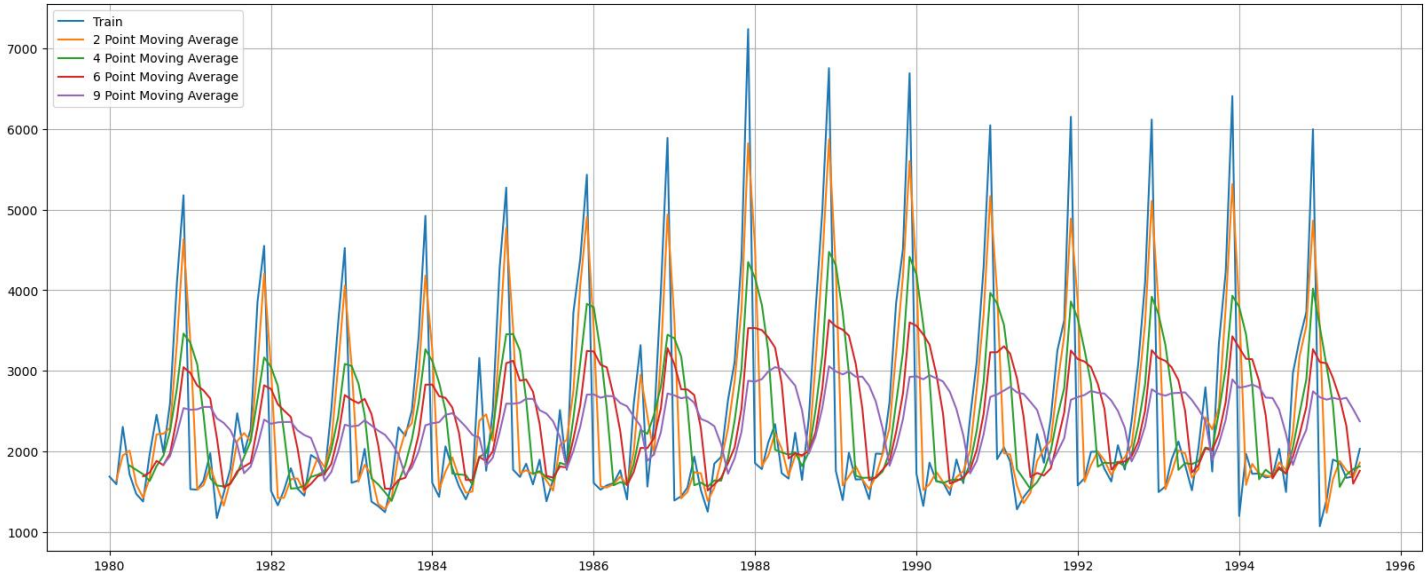

Figure 58: Comparison of Sparkling Wine Sales with Different Moving Averages

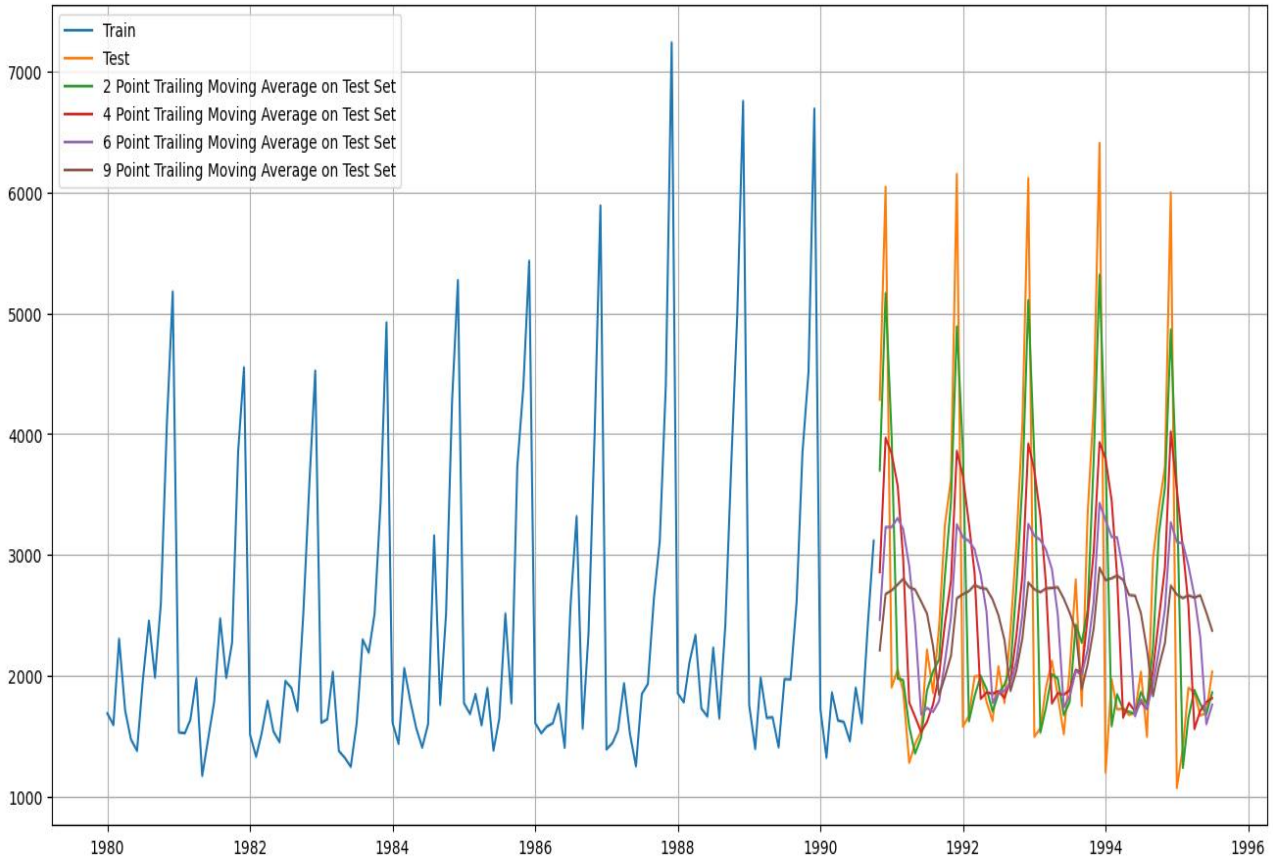

Figure 59: Comparison of Actual Sales and Trailing Moving Averages on Training and Test Data (Sparkling)

- **For 2 point Moving Average Model forecast on the Training Data, RMSE is 811.179**
- **For 4 point Moving Average Model forecast on the Training Data, RMSE is 1184.213**
- **For 6 point Moving Average Model forecast on the Training Data, RMSE is 1337.201**
- **For 9 point Moving Average Model forecast on the Training Data, RMSE is 1422.653**

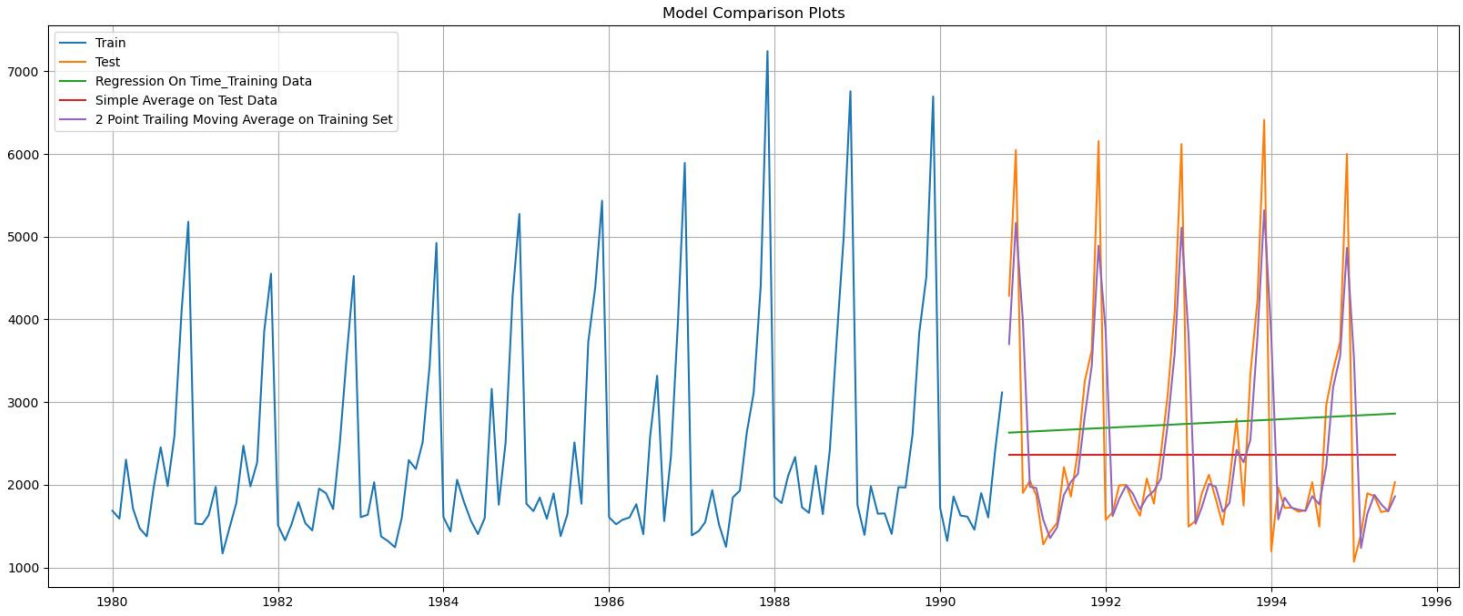

Figure 60: Comparison of Model Predictions and Actual Sales for Sparkling Wine (Training and Test Data)

Model 4: Simple Exponential Smoothing

```
{'smoothing_level': 0.037534298998536025,
 '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 33 : Parameters of the Simple Exponential Smoothing (SES) Mode

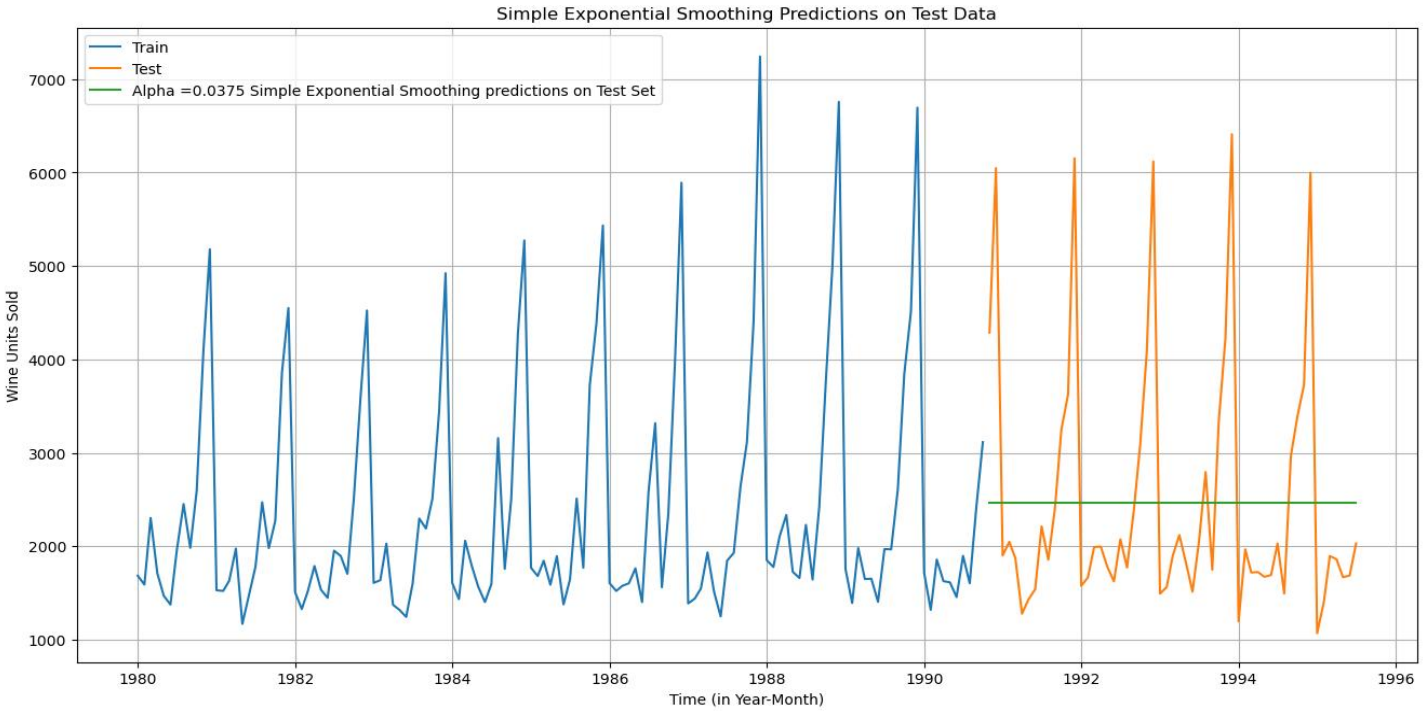

Figure 61: Simple Exponential Smoothing Predictions vs Actual Sales for Sparkling Wine (Training and Test Data)

- For Alpha =0.0375 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1362.429

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

	Alpha Values	Train RMSE	Test RMSE
7	0.45	1327.999697	1362.187063
6	0.40	1329.814823	1363.037803
8	0.50	1326.403864	1364.863549
5	0.35	1331.152869	1366.893767
0	0.10	1298.211536	1367.395642
9	0.55	1325.503170	1370.948695
4	0.30	1331.102204	1372.323705
1	0.15	1312.518677	1374.658019
3	0.25	1328.611553	1377.005722
2	0.20	1322.658289	1378.320562
10	0.60	1325.588422	1379.988733
11	0.65	1326.816031	1391.414538
12	0.70	1329.257530	1404.659104
13	0.75	1332.940498	1419.202408
14	0.80	1337.879425	1434.578214
15	0.85	1344.098031	1450.366060
16	0.90	1351.645478	1466.179706
17	0.95	1360.608670	1481.655953

Table 34 : Sorted Results of Test RMSE for Model Comparison

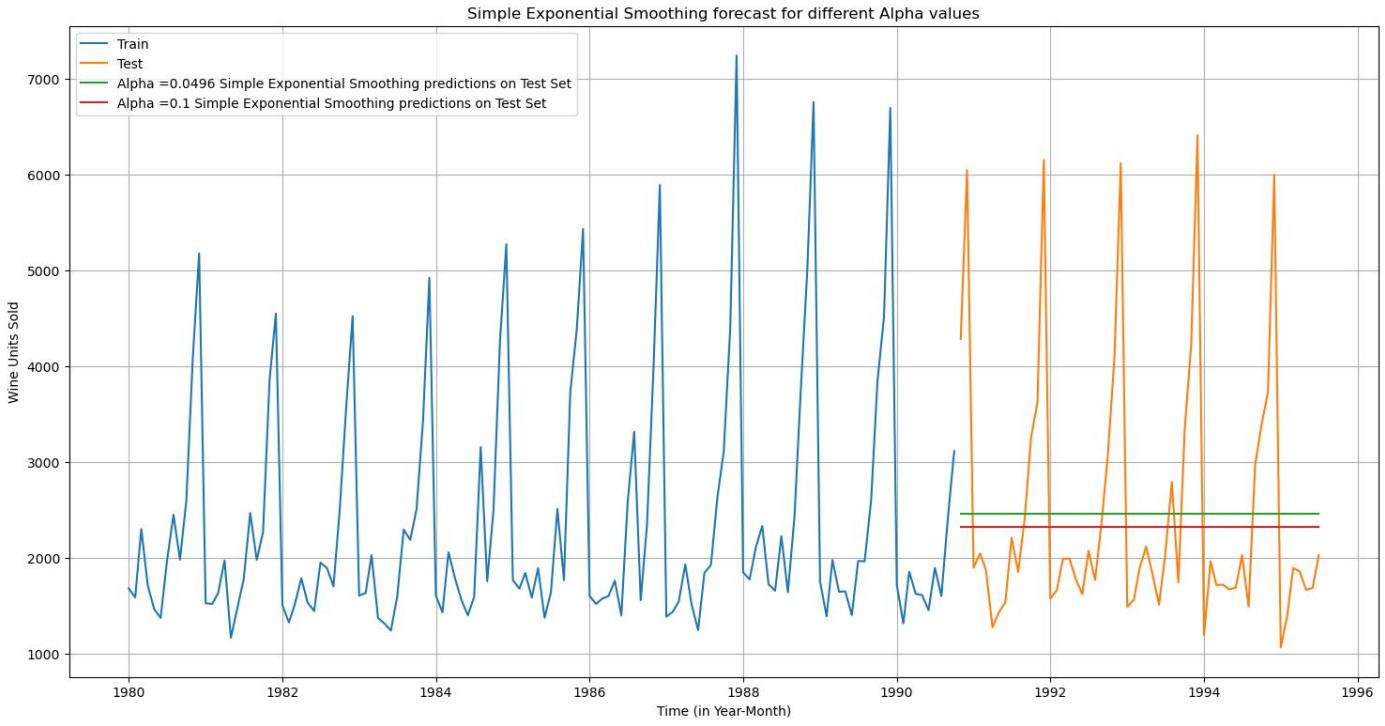

Figure 62: Simple Exponential Smoothing forecast for different Alpha values(Sparkling)

Model 5: Double Exponential Smoothing (Holt's Model)

```
{'smoothing_level': 0.6414285714285713,
 'smoothing_trend': 0.0001,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1686.0,
 'initial_trend': -95.0,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table 35 : Parameters of the Double Exponential Smoothing (DES) Model

YearMonth	
1990-11-01	2623.902015
1990-12-01	2530.231452
1991-01-01	2436.560888
1991-02-01	2342.890325
1991-03-01	2249.219761
1991-04-01	2155.549198
1991-05-01	2061.878634
1991-06-01	1968.208071
1991-07-01	1874.537507
1991-08-01	1780.866944
1991-09-01	1687.196380
1991-10-01	1593.525817
1991-11-01	1499.855253
1991-12-01	1406.184690
1992-01-01	1312.514126
1992-02-01	1218.843563
1992-03-01	1125.172999
1992-04-01	1031.502436
1992-05-01	937.831872
1992-06-01	844.161309
1992-07-01	750.490745
1992-08-01	656.820181
1992-09-01	563.149618
1992-10-01	469.479054
1992-11-01	375.808491
1992-12-01	282.137927
1993-01-01	188.467364

1993-02-01	94.796800
1993-03-01	1.126237
1993-04-01	-92.544327
1993-05-01	-186.214890
1993-06-01	-279.885454
1993-07-01	-373.556017
1993-08-01	-467.226581
1993-09-01	-560.897144
1993-10-01	-654.567708
1993-11-01	-748.238271
1993-12-01	-841.908835
1994-01-01	-935.579398
1994-02-01	-1029.249962
1994-03-01	-1122.920525
1994-04-01	-1216.591089
1994-05-01	-1310.261652
1994-06-01	-1403.932216
1994-07-01	-1497.602779
1994-08-01	-1591.273343
1994-09-01	-1684.943906
1994-10-01	-1778.614470
1994-11-01	-1872.285033
1994-12-01	-1965.955597
1995-01-01	-2059.626160
1995-02-01	-2153.296724
1995-03-01	-2246.967287
1995-04-01	-2340.637851
1995-05-01	-2434.308414
1995-06-01	-2527.978978
1995-07-01	-2621.649541

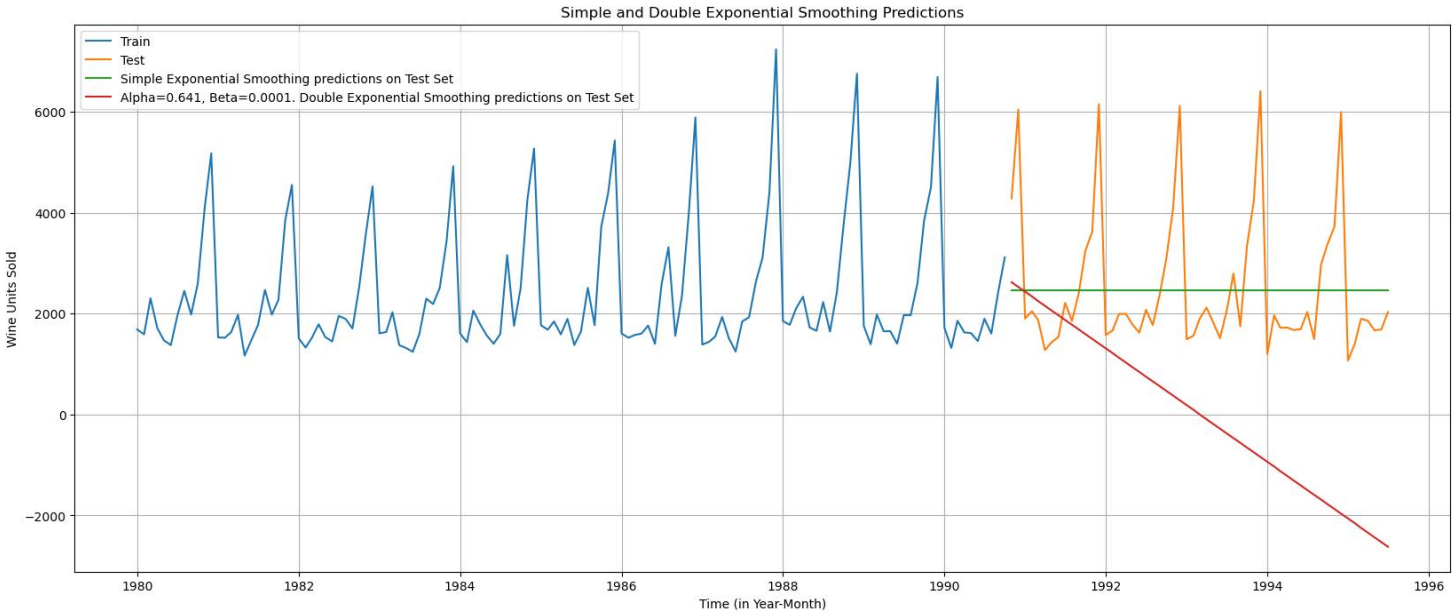

Figure 63: Comparison of Simple and Double Exponential Smoothing Predictions for Sparkling Wine Sales

- For DES forecast on the Rose Testing Data: RMSE is 3173.262

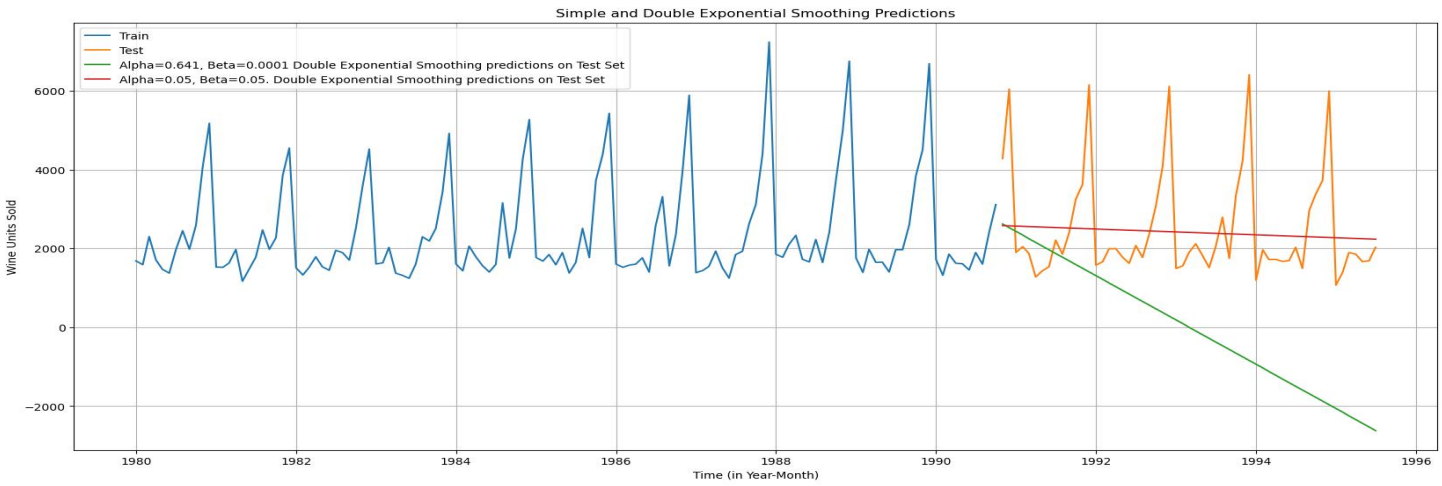

Figure 64: Double Exponential Smoothing forecast for different Alpha, Beta values (Sparkling)

Model 6: Triple Exponential Smoothing (Holt - Winter's Model)

```
{'smoothing_level': 0.07571432471504627,
 'smoothing_trend': 0.06489794789923221,
 'smoothing_seasonal': 0.3765611795178487,
 'damping_trend': nan,
 'initial_level': 2356.5416847960546,
 'initial_trend': -9.182360270735833,
 'initial_seasons': array([0.71216394, 0.67829895, 0.89649052, 0.79723125, 0.64100433,
       0.63985644, 0.86674058, 1.1133546 , 0.89819179, 1.18511974,
       1.83459596, 2.32779881]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table 36: Parameters of the Triple Exponential Smoothing

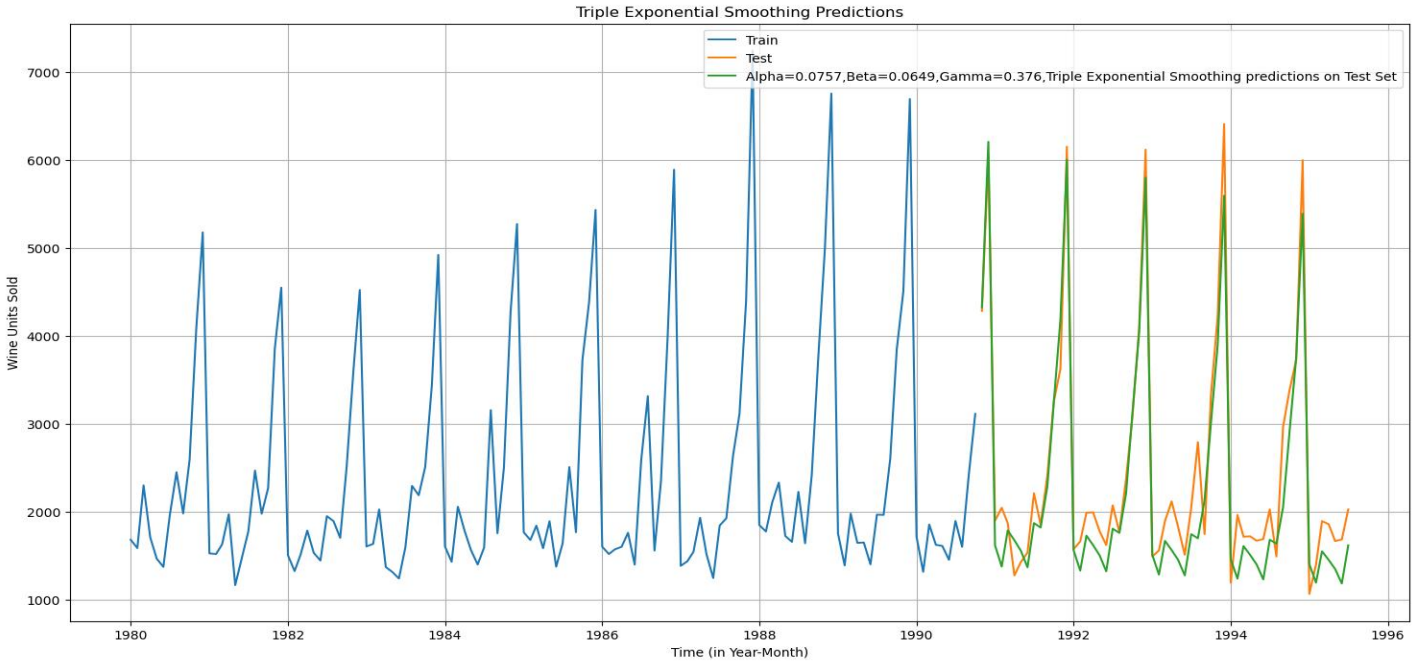

Figure 65 :Triple Exponential Smoothing Predictions

- For Alpha=0.0757,Beta=0.0649,Gamma=0.376, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 381.655.

	Test RMSE
Linear Regression	1392.438305
SimpleAverageModel	1368.746717
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281
Alpha =0.0375,SimpleExponentialSmoothing	1362.428949
Alpha=0.641, Beta=0.0001, Double Exponential Smoothing	3173.262078
Alpha=0.05, Beta=0.05, Double Exponential Smoothing	1359.261034
Alpha=0.0757,Beta=0.0649,Gamma=0.376,Triple Exponential Smoothing	381.655272
Alpha=0.20,Beta=0.35,Gamma=0.80,Triple Exponential Smoothing	326.867284

Table 37: Sorted by RMSE values on the Test Data

1. Best Performing Models:

- Triple Exponential Smoothing (Alpha=0.20, Beta=0.35, Gamma=0.80) has the lowest RMSE of 326.87, which indicates the most accurate model among the ones tested. This model is highly effective in capturing both trend and seasonality, likely due to its ability to adjust for level, trend, and seasonality simultaneously.
- Triple Exponential Smoothing (Alpha=0.0757, Beta=0.0649, Gamma=0.376) also performs well, with an RMSE of 381.66, which is close to the best-performing model.

This indicates that fine-tuning seasonality, trend, and level smoothing factors is a good approach for the given data.

3. Models with Moderate Performance:

- 2-point Trailing Moving Average (RMSE = 811.18) has a relatively higher error compared to the triple exponential smoothing models but still outperforms other simpler models like Linear Regression and Simple Exponential Smoothing.
- Linear Regression (RMSE = 1392.44) performs worse than most of the smoothing models. This suggests that Linear Regression might not capture the seasonal fluctuations and trends as effectively as the exponential smoothing or moving average models.

4. Models with Poor Performance:

- Double Exponential Smoothing (Alpha=0.641, Beta=0.0001) with an RMSE of 3173.26 has a very high error, indicating that the model is not well-suited for the data. The extremely low value for Beta (0.0001) suggests that the model might be too static, failing to account for the trend properly.
- Simple Average Model (RMSE = 1368.75) also performs poorly compared to the moving averages and smoothing models, indicating that it ' s a simple baseline without capturing the underlying trend or seasonality.

5. General Observations:

- Moving Averages (2-point, 4-point, 6-point, 9-point) tend to perform better than more complex models like Linear Regression and Simple Exponential Smoothing. This could be due to the smoothing effect that captures more recent data, which might be closer to actual sales behavior.
- Exponential Smoothing Models with well-tuned Alpha, Beta, and Gamma parameters (Triple Exponential Smoothing) deliver the most accurate forecasts, showing the importance of parameter optimization in forecasting methods.

Check for Stationarity

A Time Series is considered to be stationary when statistical properties such as the variance and (auto) correlation are constant over time.

Stationary Time Series allows us to think of the statistical properties of the time series as not changing in time, which enables us to build appropriate statistical models for forecasting based on past data.

Dickey-Fuller Test - Dicky Fuller Test on the timeseries is run to check for stationarity of data.

- Null Hypothesis H0 : Time Series is non-stationary.
- Alternate Hypothesis Ha : Time Series is stationary.
- So Ideally if p-value < 0.05 then null hypothesis: TS is non-stationary is rejected else the TS is non-stationary is failed to be rejected .

- **Results of Dicky-Fuller Test**
- **DF test statistic is -1.798**
- **DF test p-value is 0.7055958459932068**
- **Number of lags used 12**

We see 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.

- Differencing 'd' is done on a non-stationary time series data one or more times to convert it into stationary.

- (d=1) 1st order differencing is done where the difference between the current and previous (1 lag before) series is taken and then checked for stationarity using the ADF(Augmented Dicky Fueller) test. If differenced time series is stationary, we proceed with AR modeling. Else we do (d=2) 2nd order differencing, and this process repeats till we get a stationary time series

The variance of a time series may also not be the same over time. To remove this kind of non-stationarity, we can transform the data. If the variance is increasing over time, then a log transformation can stabilize the variance.

- **Results of Dicky-Fuller Test with differencing**
- **DF test statistic is -44.912**
- **DF test p-value is 0.0**
- **Number of lags used 10**

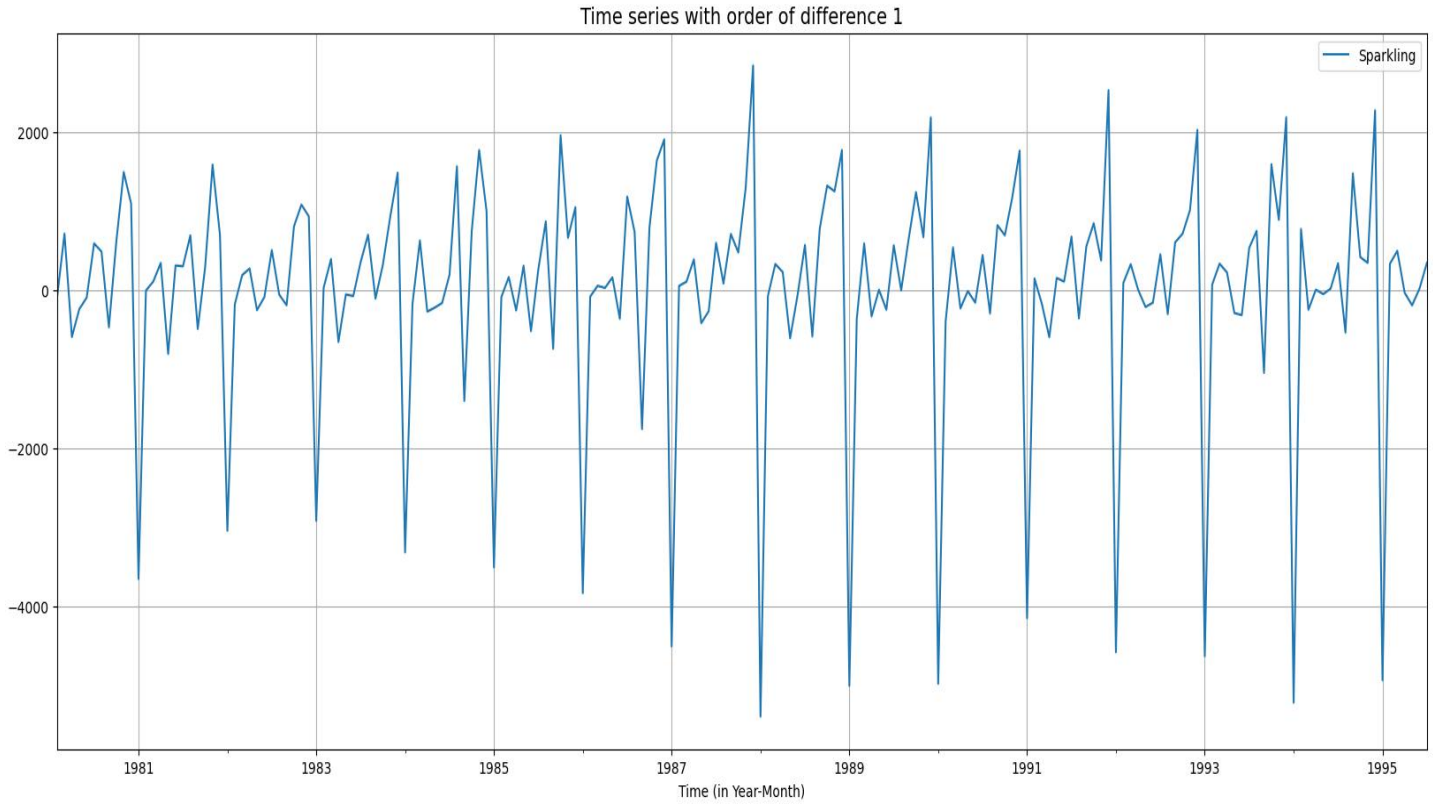

Figure 66 :Time series with order of difference 1 (Sparkling)

Model Building - Stationary Data

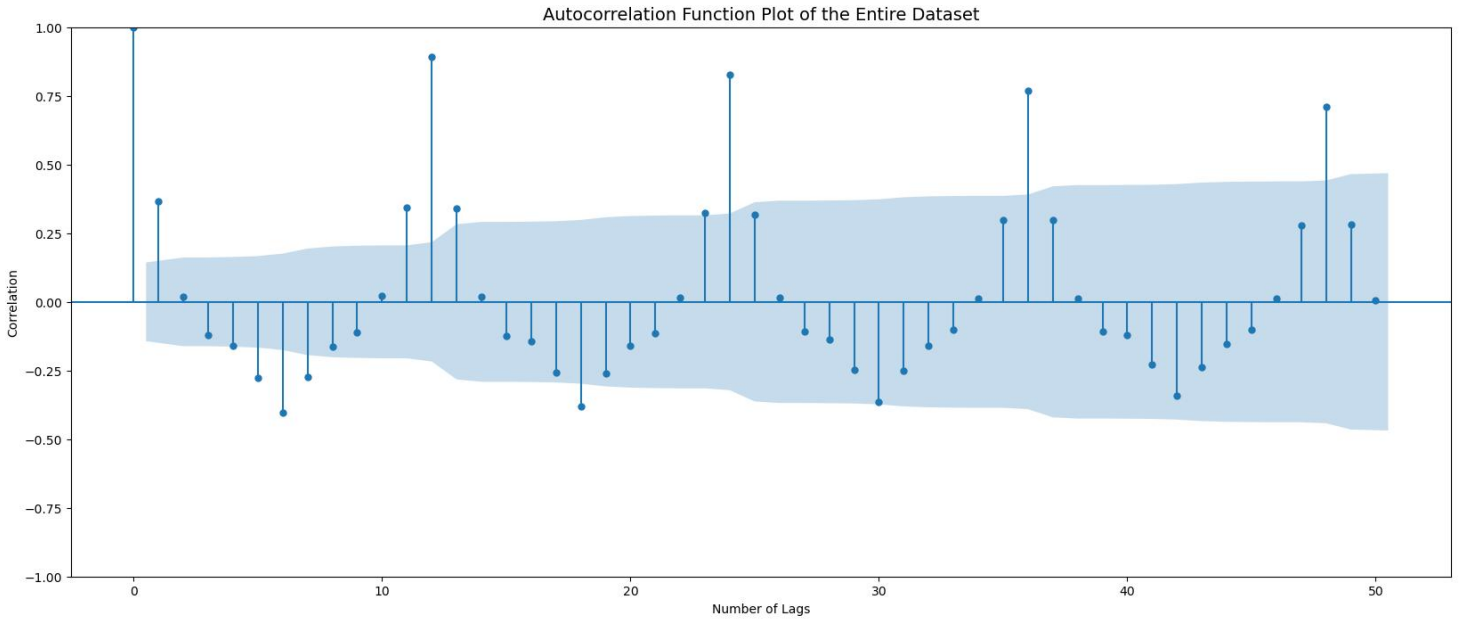

Figure 67 :Autocorrelation Function Plot of the Entire Dataset (Sparkling)

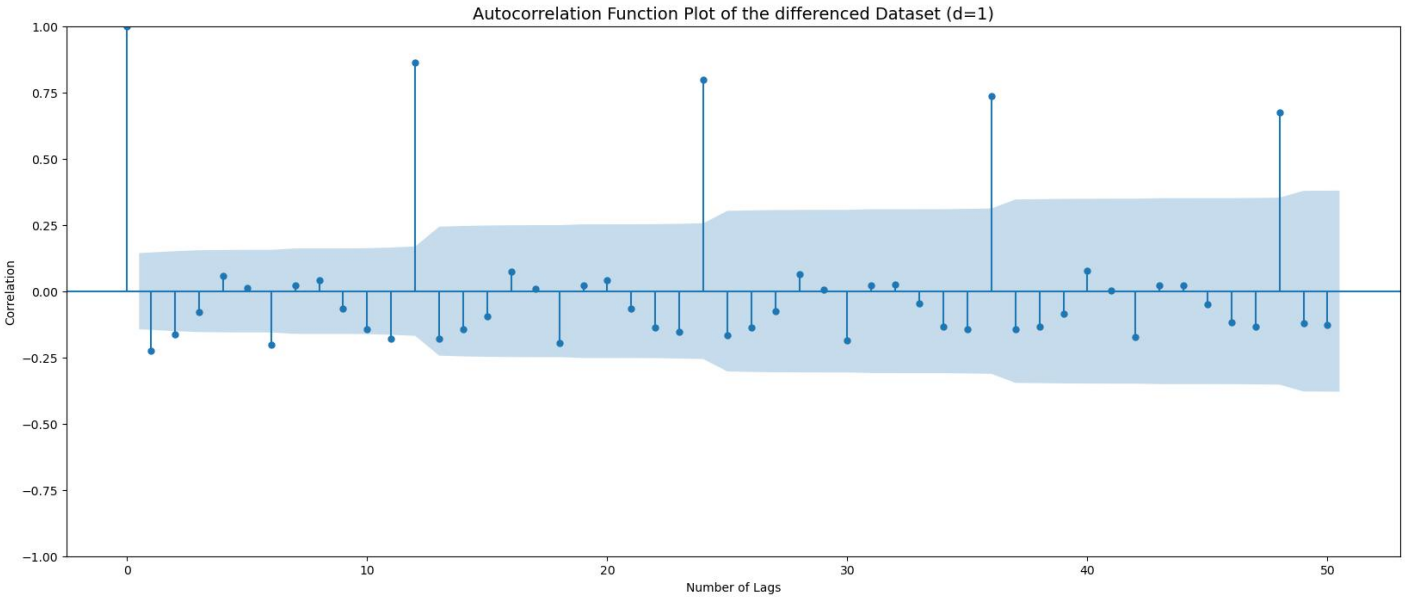

Figure 68 :Autocorrelation Function Plot of the differenced Dataset (d=1)(Sparkling)

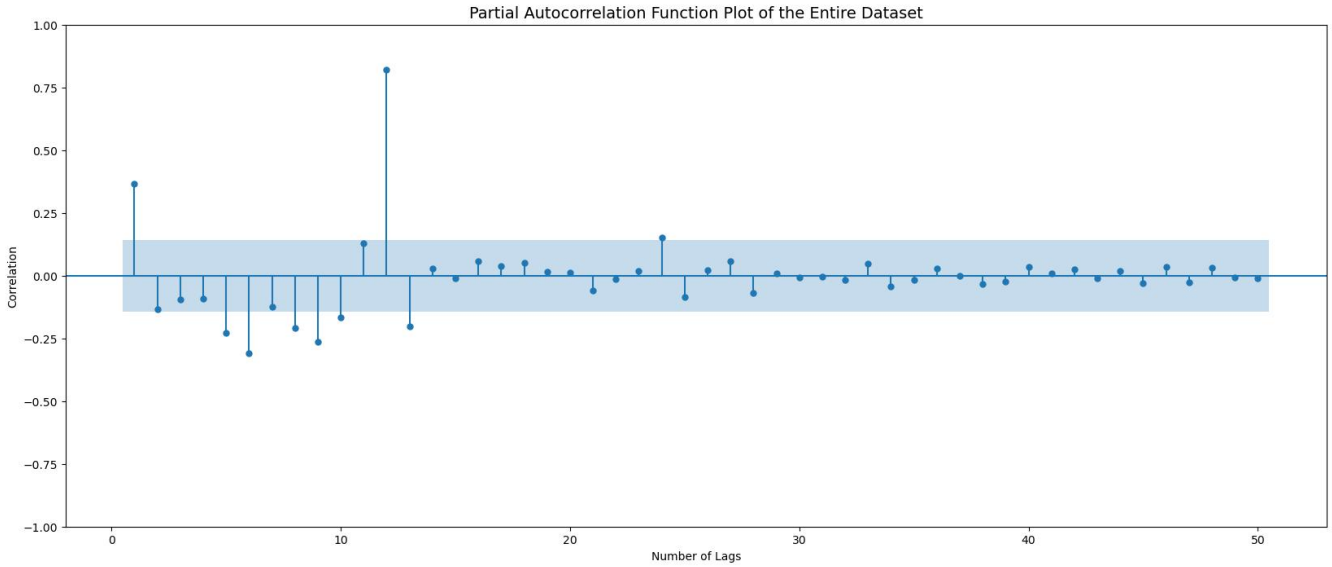

Figure 69 :Partial Autocorrelation Function Plot of the Entire Dataset(Sparkling)

Model 1: ARIMA Model

Examples of the parameter combinations for the Model

Model: (0, 1, 0)

Model: (0, 1, 1)

Model: (0, 1, 2)

Model: (0, 1, 3)

Model: (0, 1, 4)

Model: (1, 1, 0)

Model: (1, 1, 1)

Model: (1, 1, 2)

Model: (1, 1, 3)

Model: (1, 1, 4)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (2, 1, 4)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
Model: (3, 1, 4)
Model: (4, 1, 0)
Model: (4, 1, 1)
Model: (4, 1, 2)
Model: (4, 1, 3)
Model: (4, 1, 4)

	param	AIC
24	(4, 1, 4)	2176.435351
12	(2, 1, 2)	2178.109723
18	(3, 1, 3)	2181.617358
14	(2, 1, 4)	2183.063753
23	(4, 1, 3)	2184.812363

Table 38: AIC values in the ascending order

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	130
Model:	ARIMA(4, 1, 4)	Log Likelihood	-1079.218
Date:	Sun, 17 Nov 2024	AIC	2176.435
Time:	20:23:12	BIC	2202.174
Sample:	01-01-1980 - 10-01-1990	HQIC	2186.893
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4771	0.113	-4.231	0.000	-0.698	-0.256
ar.L2	-0.4781	0.072	-6.604	0.000	-0.620	-0.336
ar.L3	-0.4743	0.095	-4.983	0.000	-0.661	-0.288
ar.L4	0.5225	0.067	7.808	0.000	0.391	0.654
ma.L1	0.0010	8.264	0.000	1.000	-16.196	16.198
ma.L2	0.0119	16.676	0.001	0.999	-32.672	32.696
ma.L3	-0.0256	8.178	-0.003	0.998	-16.053	16.002
ma.L4	-0.9873	0.161	-6.126	0.000	-1.303	-0.671
sigma2	8.893e+05	3.33e-05	2.67e+10	0.000	8.89e+05	8.89e+05

Ljung-Box (L1) (Q):	0.16	Jarque-Bera (JB):	2.80
Prob(Q):	0.69	Prob(JB):	0.25
Heteroskedasticity (H):	2.66	Skew:	0.34
Prob(H) (two-sided):	0.00	Kurtosis:	3.22

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.43e+27. Standard errors may be unstable.

Table 39: Auto ARIMA Model Summary for Sparkling Wine Sales Forecasting

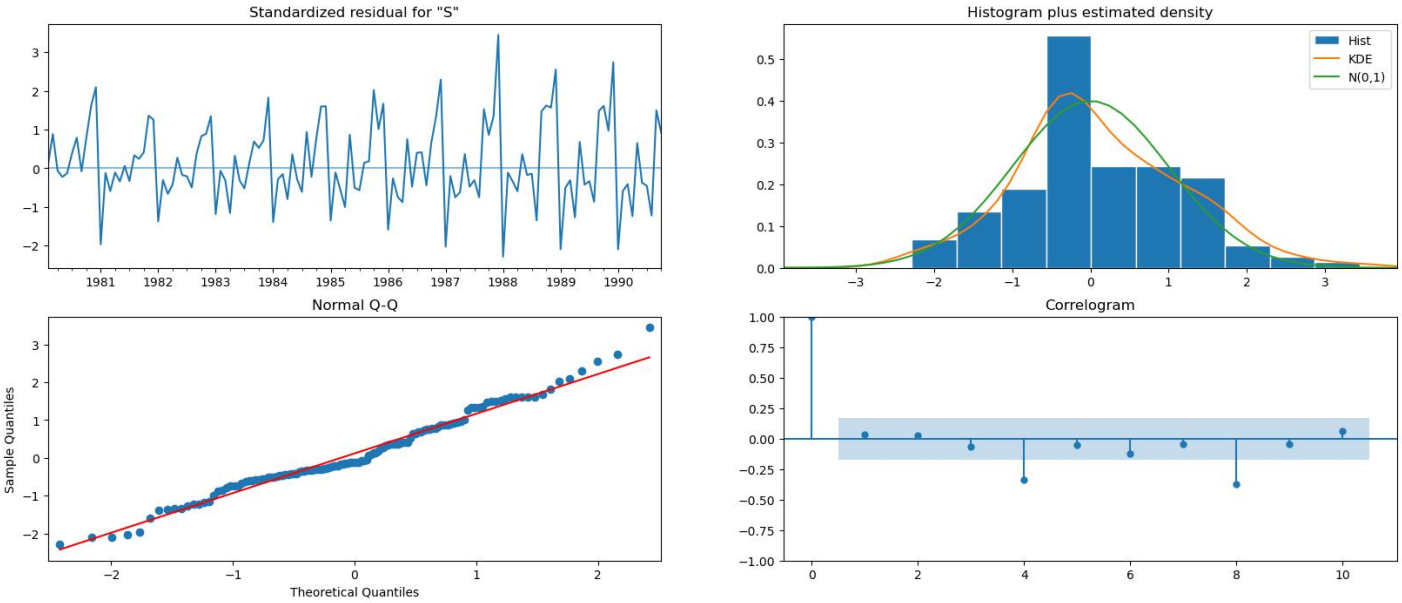

Figure 70:Diagnostics plot Auto ARIMA (Sparkling)

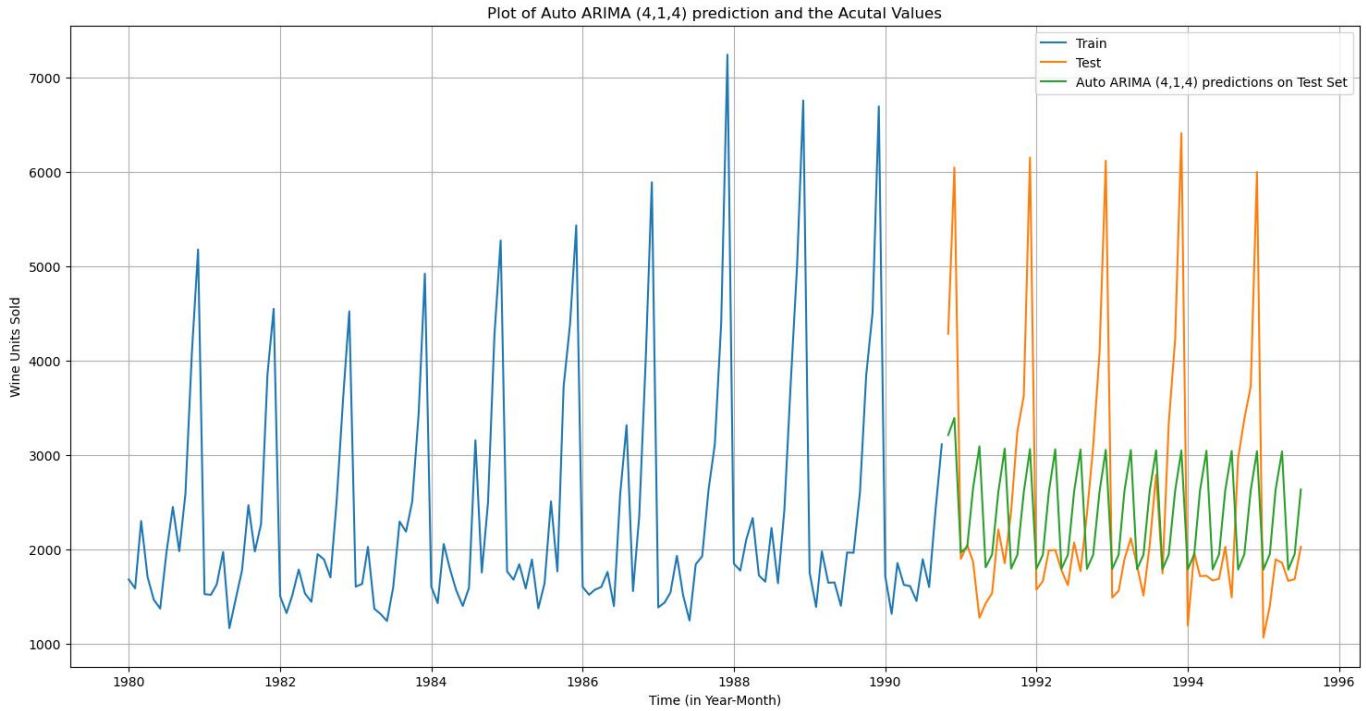

Figure 71:Plot of Auto ARIMA (4,1,4) prediction and the Actual Values (Sparkling)

- **RMSE: 1213.419548735715**
- **MAPE: 35.17936281350787**

Model 2: Manual ARIMA 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 can see that first lag cuts off in both plots and hence we start from lags after lag 0. Therefore we have taken the value of p and q to be 2 and 1 respectively.

We would also build a model with p=0,q=0

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	130
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1115.360
Date:	Sun, 17 Nov 2024	AIC	2232.719
Time:	20:23:14	BIC	2235.579
Sample:	01-01-1980 - 10-01-1990	HQIC	2233.881
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1.88e+06	1.28e+05	14.696	0.000	1.63e+06	2.13e+06

Ljung-Box (L1) (Q):	3.62	Jarque-Bera (JB):	196.64
Prob(Q):	0.06	Prob(JB):	0.00
Heteroskedasticity (H):	2.37	Skew:	-1.93
Prob(H) (two-sided):	0.01	Kurtosis:	7.66

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

Table 40: Manual ARIMA Model Summary for Sparkling Wine Sales Forecasting

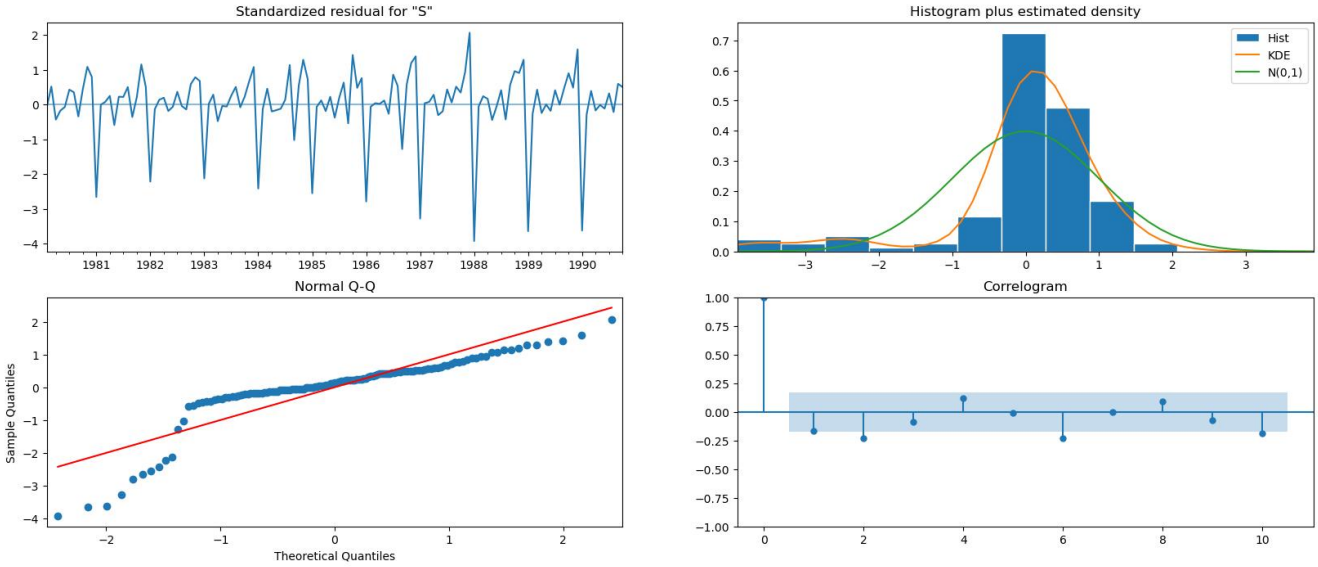

Figure 72:Diagnostic Plot Manual ARIMA (Sparkling)

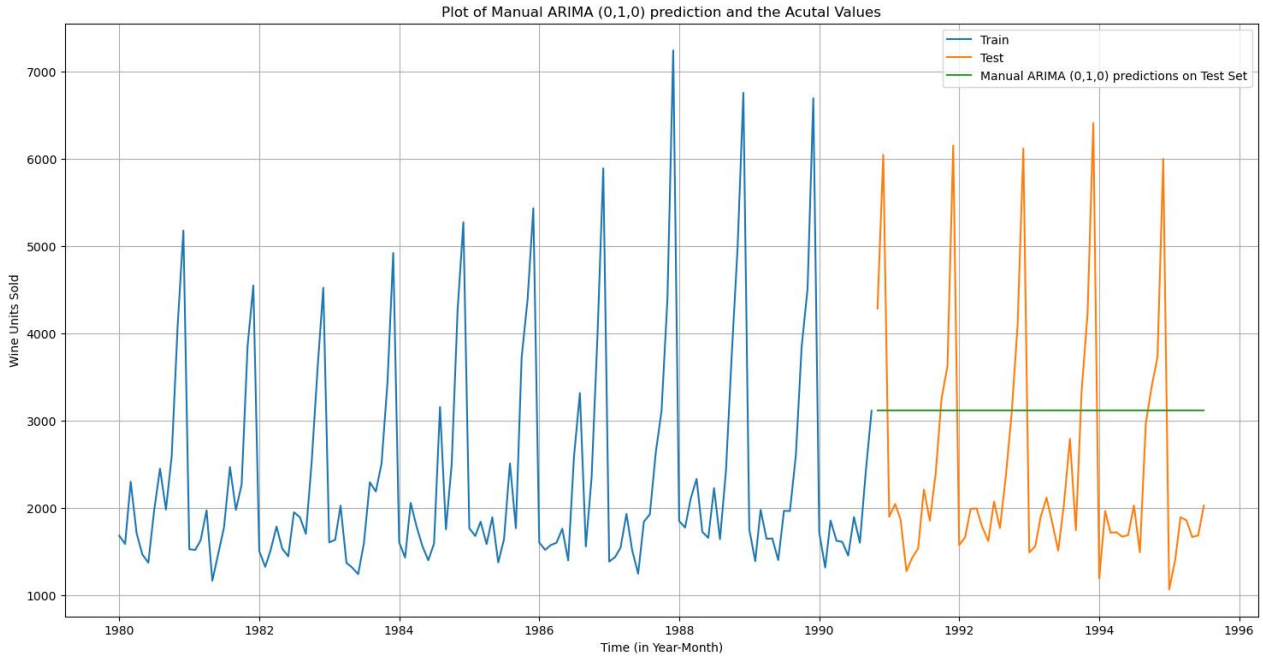

Figure 73:Plot of Manual ARIMA (0,1,0) prediction and the Actual Values (Sparkling)

- **RMSE: 1496.4446285498805**
- **MAPE: 64.94421166998026**

Model 3: Auto SARIMA Model

253	(3, 1, 3)	NaN	1349.705424
237	(3, 1, 2)	NaN	1352.349172
221	(3, 1, 1)	NaN	1352.506962
220	(3, 1, 1)	NaN	1352.668051
222	(3, 1, 1)	NaN	1353.713090
252	(3, 1, 3)	NaN	1353.922925
236	(3, 1, 2)	NaN	1354.677592
254	(3, 1, 3)	NaN	1360.949260
205	(3, 1, 0)	NaN	1365.058551
204	(3, 1, 0)	NaN	1365.264213
206	(3, 1, 0)	NaN	1365.983545
173	(2, 1, 2)	NaN	1365.999251
157	(2, 1, 1)	NaN	1367.087463
174	(2, 1, 2)	NaN	1367.432865
156	(2, 1, 1)	NaN	1367.448213

Table 41: SARIMA Models Sorted by AIC for Sparkling Wine Sales Forecasting

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	130
Model:	SARIMAX(3, 1, 3)x(3, 0, [1], 12)	Log Likelihood	-663.853
Date:	Sun, 17 Nov 2024	AIC	1349.705
Time:	20:29:00	BIC	1377.203
Sample:	01-01-1980	HQIC	1360.794
	- 10-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.8059	0.124	-14.589	0.000	-2.048	-1.563
ar.L2	-1.0083	0.229	-4.403	0.000	-1.457	-0.560
ar.L3	-0.1607	0.123	-1.308	0.191	-0.401	0.080
ma.L1	0.8919	0.260	3.433	0.001	0.383	1.401
ma.L2	-1.1008	0.145	-7.591	0.000	-1.385	-0.817
ma.L3	-1.0707	0.185	-5.776	0.000	-1.434	-0.707
ar.S.L12	0.8147	0.257	3.165	0.002	0.310	1.319
ar.S.L24	0.0342	0.204	0.167	0.867	-0.366	0.435
ar.S.L36	0.2536	0.138	1.841	0.066	-0.016	0.524
ma.S.L12	-0.4906	0.233	-2.108	0.035	-0.947	-0.034
sigma2	1.174e+05	3.6e-06	3.26e+10	0.000	1.17e+05	1.17e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	1.31
Prob(Q):	0.98	Prob(JB):	0.52
Heteroskedasticity (H):	1.53	Skew:	0.01
Prob(H) (two-sided):	0.25	Kurtosis:	3.59

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.27e+26. Standard errors may be unstable.

Table 41: Auto SARIMA Model Summary for Sparkling Wine Sales Forecasting

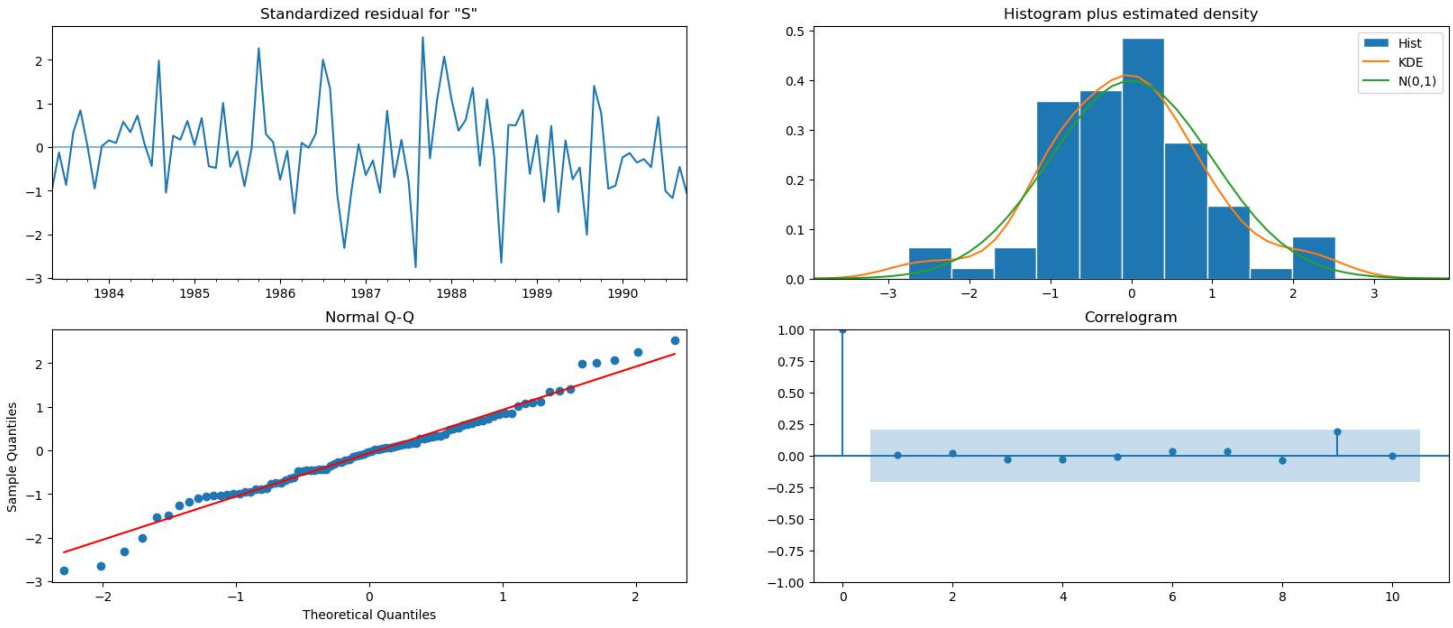

Figure 74:Diagnostic Plot of Auto SARIMA (Sparkling)

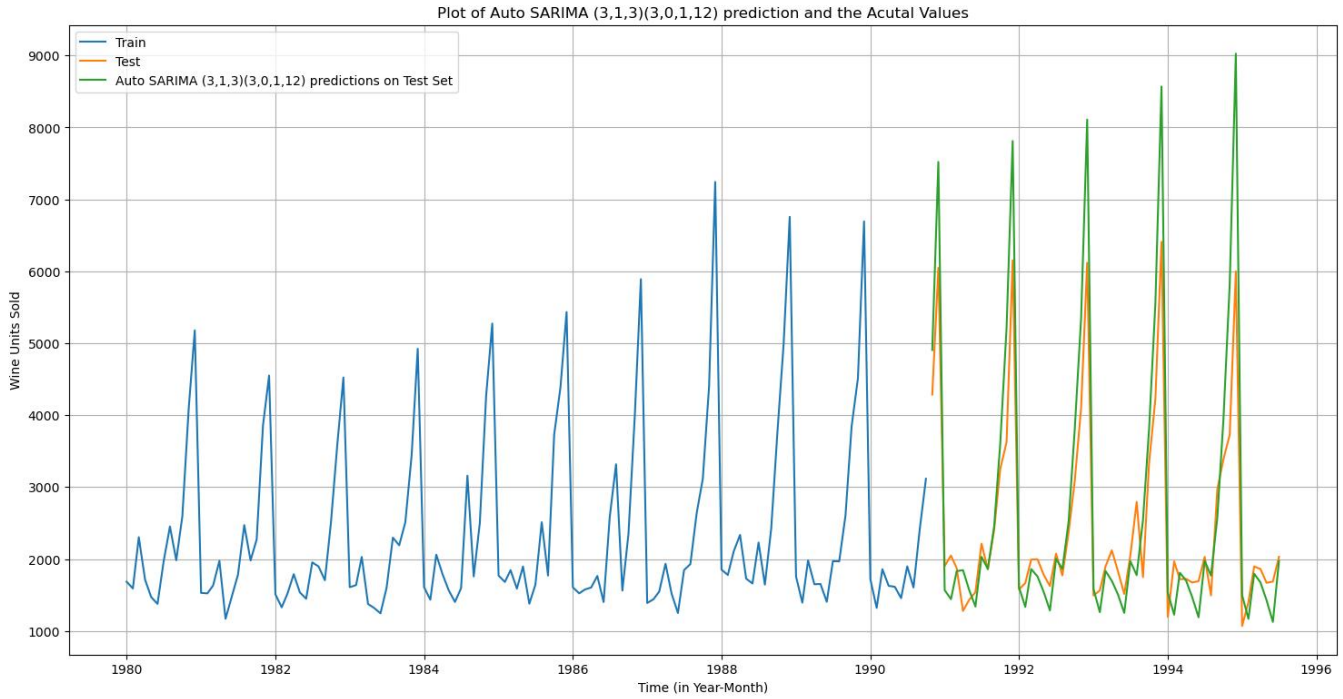

Figure 75:Plot of Auto SARIMA (3,1,3)(3,0,1,12) prediction and the Acutal Values (Sparkling)

- **RMSE: 835.4697748560455**
- **MAPE: 18.70777880972156**

Model 4: Manual SARIMA Model

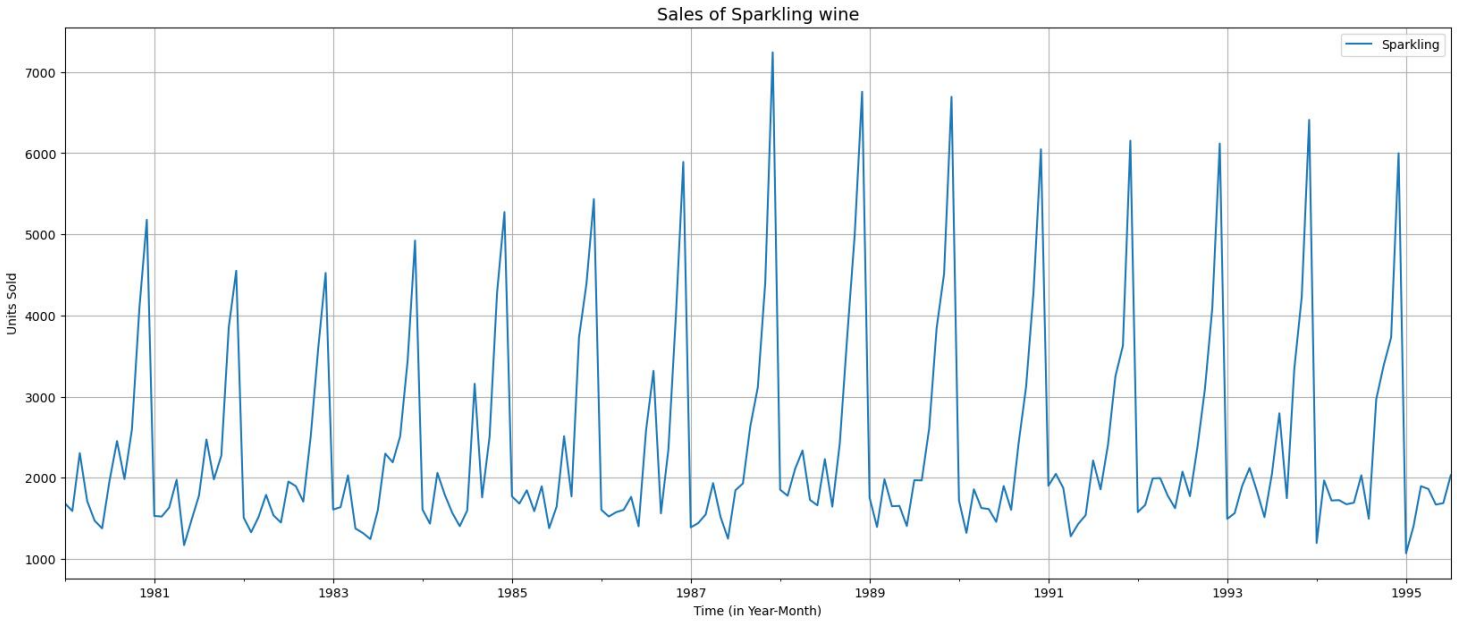

Figure 76:Sales of Sparkling wine (Sparkling)

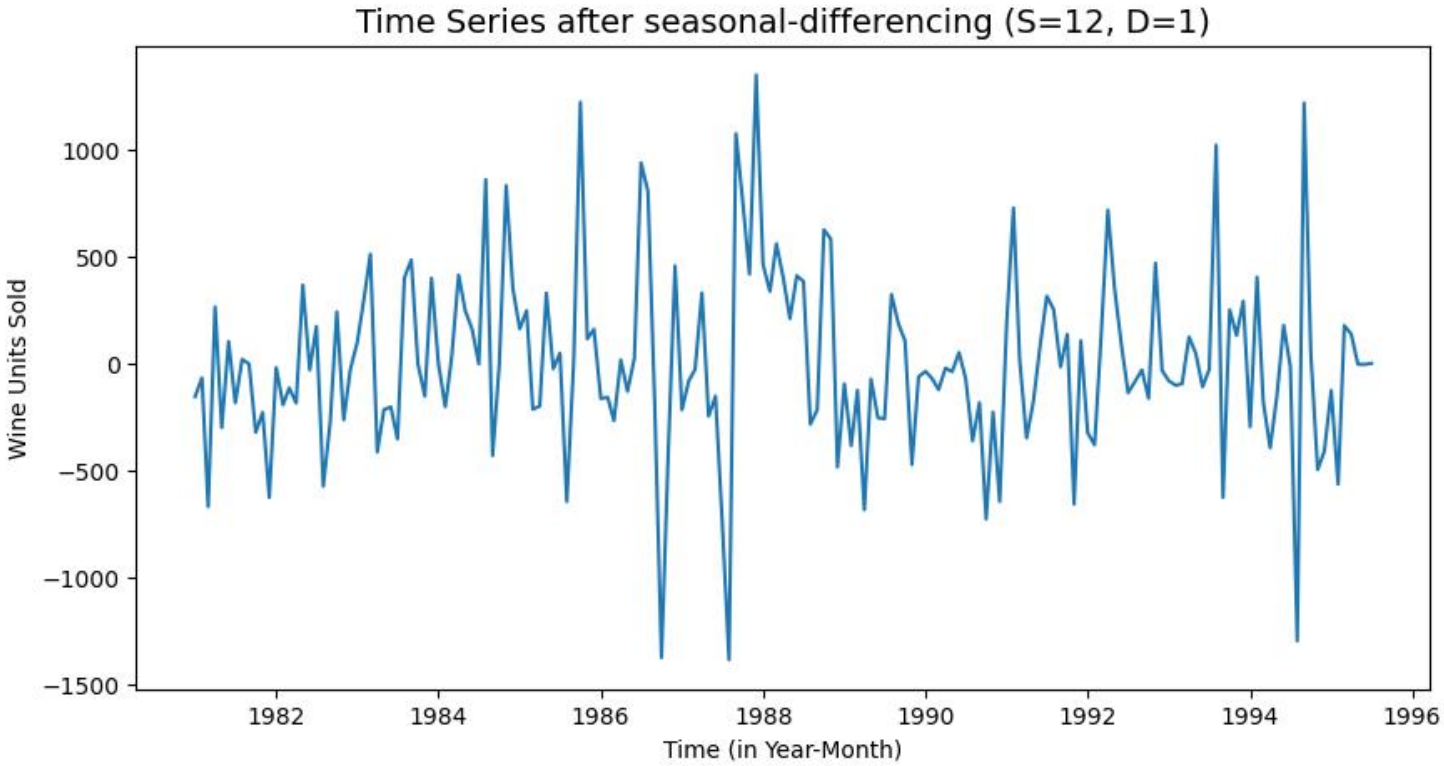

Figure 77:Time Series after seasonal-differencing (S=12, D=1) (Sparkling)

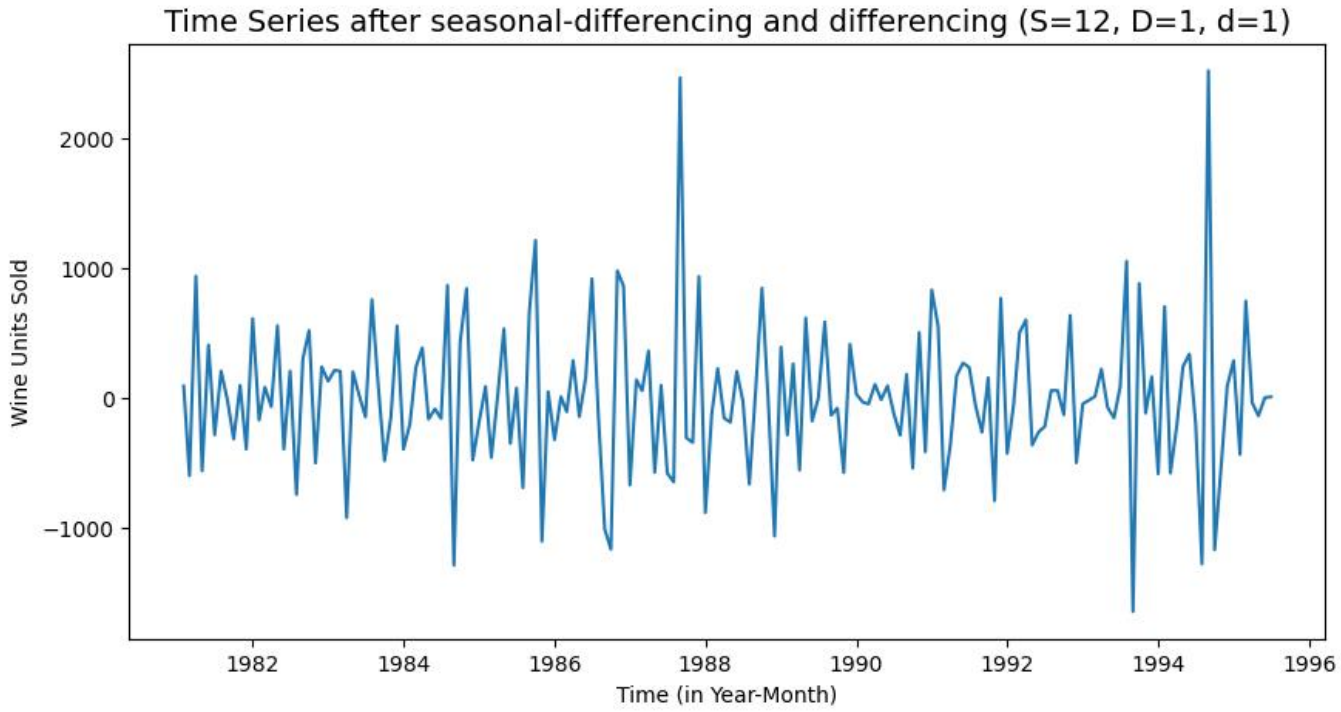

Figure 78:Time Series after seasonal-differencing and differencing (S=12, D=1, d=1) (Sparkling)

- **DF test statistic is -5.121**
- **DF test p-value is 0.00012316304915681252**
- **Number of lags used 11**

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	130
Model:	SARIMAX(4, 1, 2)x(0, 1, [1], 12)	Log Likelihood	-756.432
Date:	Sun, 17 Nov 2024	AIC	1528.864
Time:	20:29:03	BIC	1549.864
Sample:	01-01-1980 - 10-01-1990	HQIC	1537.367
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2161	0.527	-0.410	0.681	-1.248	0.816
ar.L2	-0.0770	0.181	-0.425	0.671	-0.432	0.278
ar.L3	0.0224	0.109	0.207	0.836	-0.190	0.235
ar.L4	-0.1624	0.144	-1.125	0.261	-0.445	0.121
ma.L1	-0.5070	0.556	-0.911	0.362	-1.598	0.584
ma.L2	-0.3625	0.532	-0.682	0.495	-1.404	0.679
ma.S.L12	-0.4715	0.085	-5.566	0.000	-0.638	-0.305
sigma2	1.611e+05	2.32e+04	6.945	0.000	1.16e+05	2.07e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	11.15
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	1.11	Skew:	0.50
Prob(H) (two-sided):	0.76	Kurtosis:	4.27

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

Table 42: Manual SARIMA Model Summary for Rose Wine Sales Forecasting

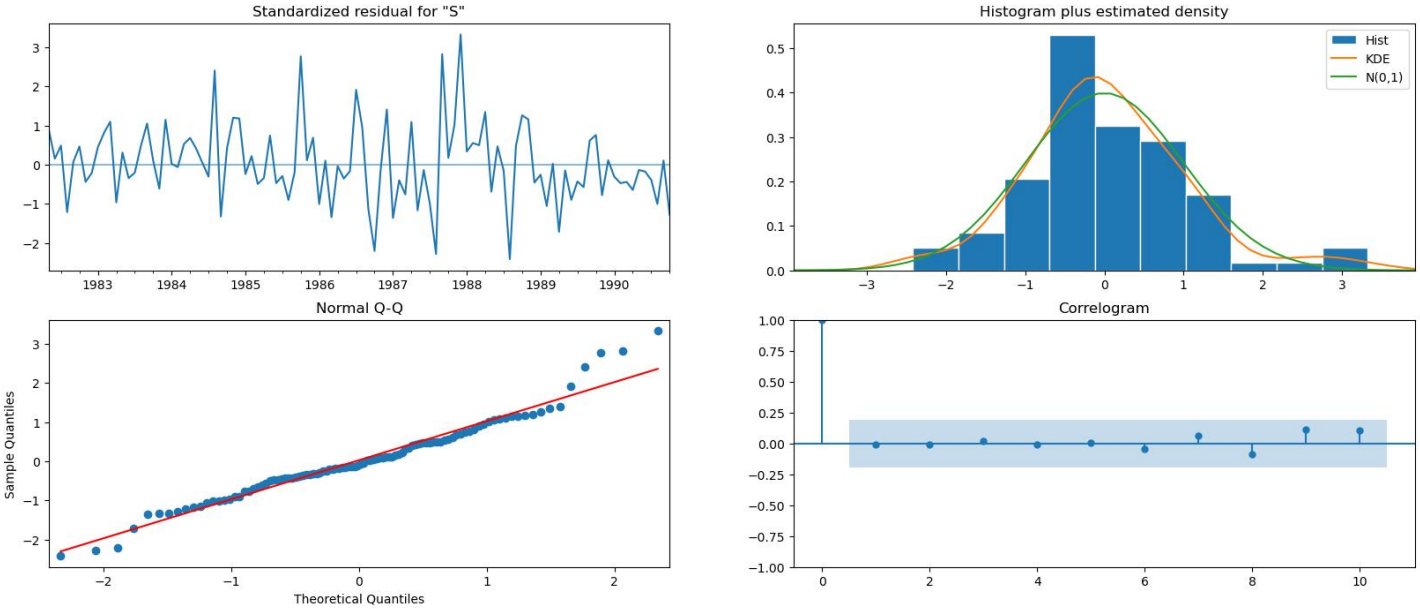

Figure 79: Diagnostics plot Manual SARIMA (Sparkling)

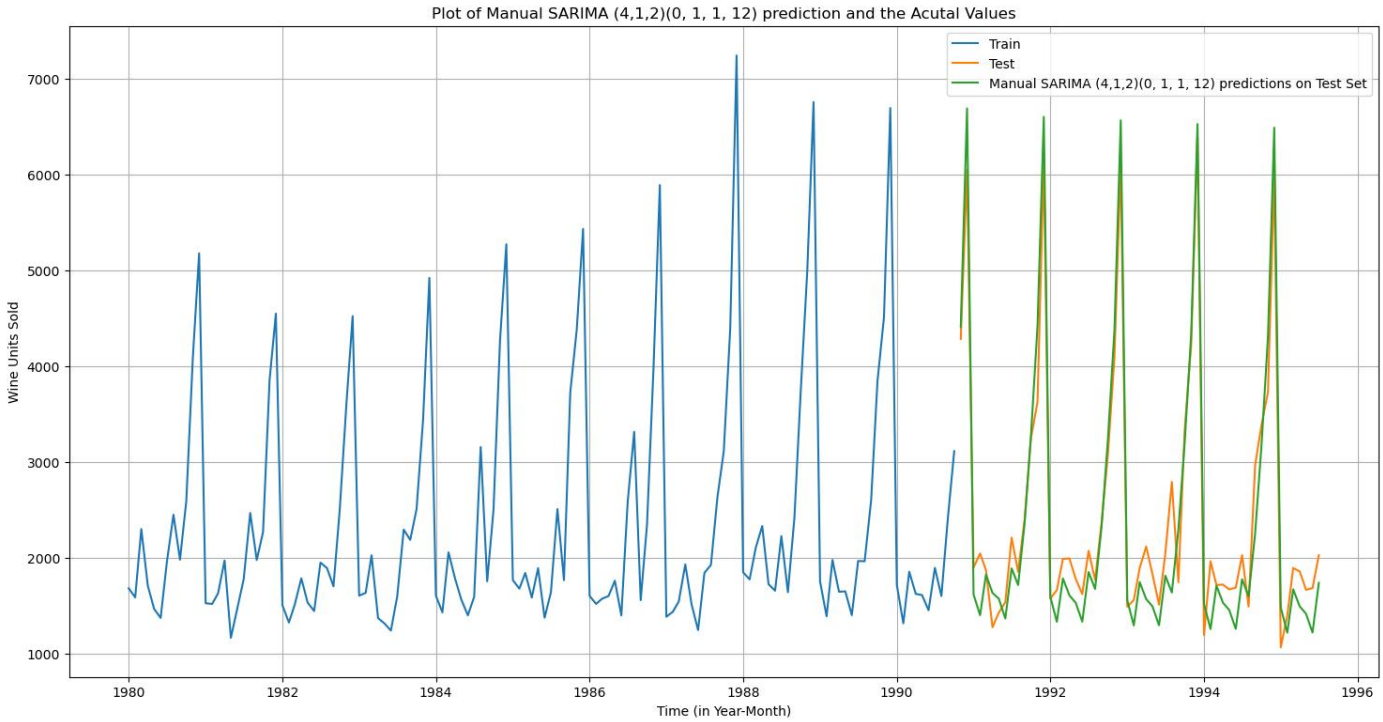

Figure 80: Plot of Manual SARIMA (4,1,2)(0, 1, 1, 12) prediction and the Actual Values (Sparkling)

- **RMSE: 373.57078182254025**
- **MAPE: 13.851089856067473**

	Test RMSE	MAPE
Auto ARIMA (4,1,4)	1213.419549	35.179363
Manual ARIMA (2,1,1)	1359.649838	37.619248
Auto SARIMA (3,1,2)(3,0,1,12)	835.469775	18.707779
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	373.570782	13.851090

Table 43: Performance of ARIMA models (Sparkling)

Performance of the models built

	Test RMSE	MAPE
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	373.570782	13.851090
Auto SARIMA (3,1,2)(3,0,1,12)	835.469775	18.707779
Auto ARIMA (4,1,4)	1213.419549	35.179363
Manual ARIMA (2,1,1)	1359.649838	37.619248
Linear Regression	1392.438305	NaN
SimpleAverageModel	1368.746717	NaN
2pointTrailingMovingAverage	811.178937	NaN
4pointTrailingMovingAverage	1184.213295	NaN
6pointTrailingMovingAverage	1337.200524	NaN
9pointTrailingMovingAverage	1422.653281	NaN
Alpha =0.0375,SimpleExponentialSmoothing	1362.428949	NaN
Alpha=0.641, Beta=0.0001, Double Exponential Smoothing	3173.262078	NaN
Alpha=0.05, Beta=0.05, Double Exponential Smoothing	1359.261034	NaN
Alpha=0.0757,Beta=0.0649,Gamma=0.376,Triple Exponential Smoothing	381.655272	NaN
Alpha=0.20,Beta=0.35,Gamma=0.80,Triple Exponential Smoothing	326.867284	NaN

Compare the performance of the models

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.

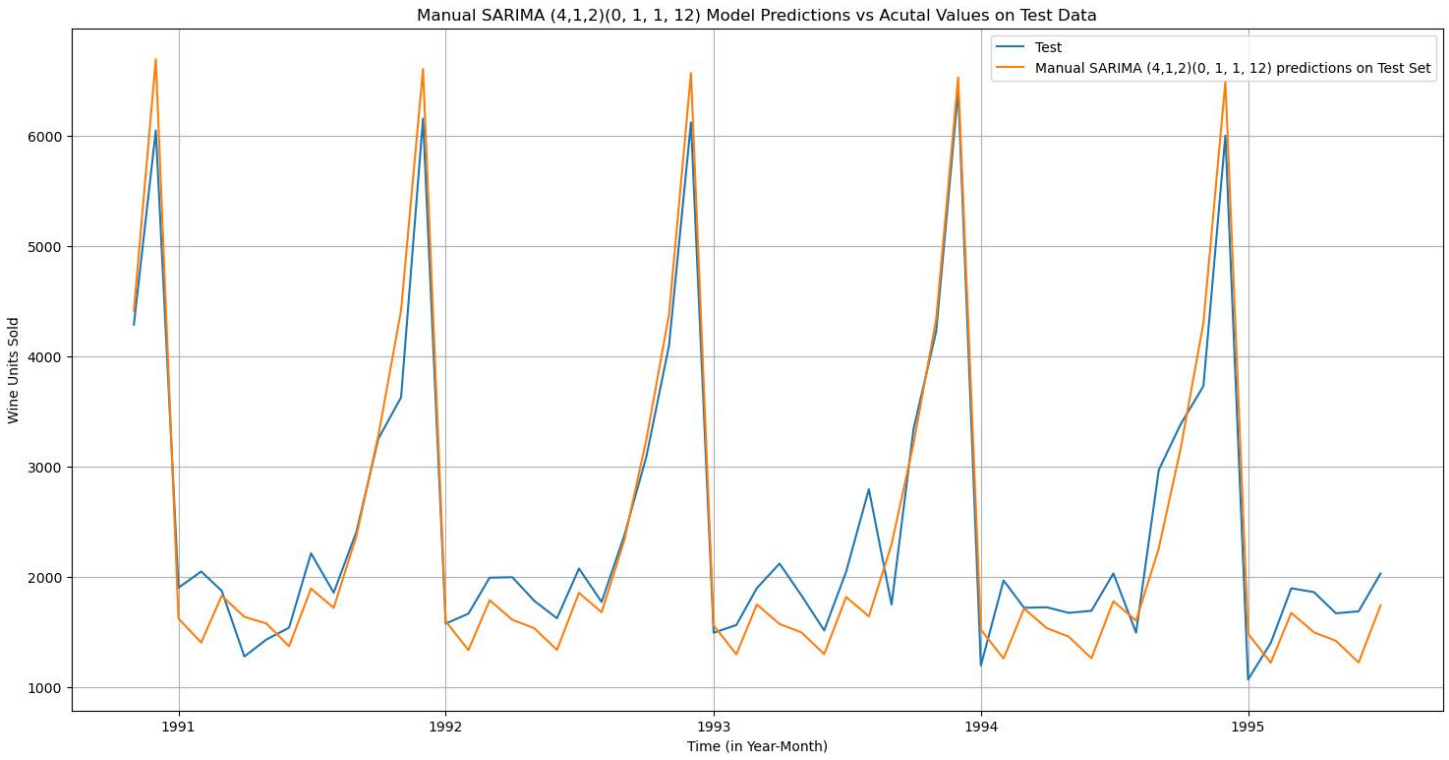

Figure 81: Manual SARIMA (4,1,2)(0, 1, 1, 12) Model Predictions vs Actual Values on Test Data (Sparkling)

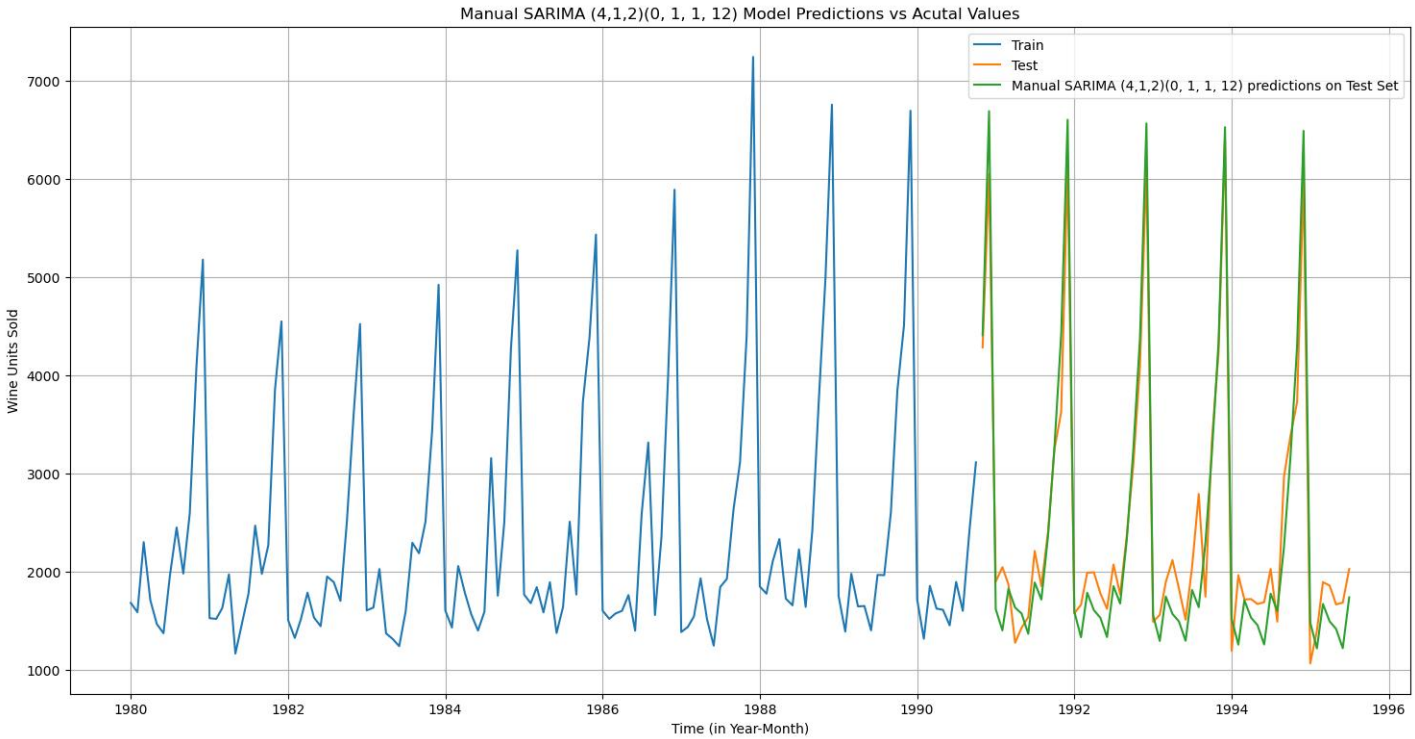

Figure 82: Manual SARIMA (4,1,2)(0, 1, 1, 12) Model Predictions vs Actual Values (Sparkling)

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(4, 1, 2)x(0, 1, [1], 12)	Log Likelihood	-1171.945
Date:	Sun, 17 Nov 2024	AIC	2359.889
Time:	20:29:08	BIC	2384.441
Sample:	01-01-1980 - 07-01-1995	HQIC	2369.859
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6275	0.517	1.213	0.225	-0.386	1.641
ar.L2	-0.1556	0.127	-1.222	0.222	-0.405	0.094
ar.L3	0.0695	0.123	0.563	0.573	-0.172	0.311
ar.L4	-0.1421	0.078	-1.826	0.068	-0.295	0.010
ma.L1	-1.4573	0.495	-2.944	0.003	-2.427	-0.487
ma.L2	0.5085	0.458	1.109	0.267	-0.390	1.407
ma.S.L12	-0.5978	0.059	-10.096	0.000	-0.714	-0.482
sigma2	1.458e+05	1.37e+04	10.681	0.000	1.19e+05	1.73e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	34.61
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.92	Skew:	0.58
Prob(H) (two-sided):	0.77	Kurtosis:	4.97

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

Table 44: Full model manual SARIMA summary(Sparkling)

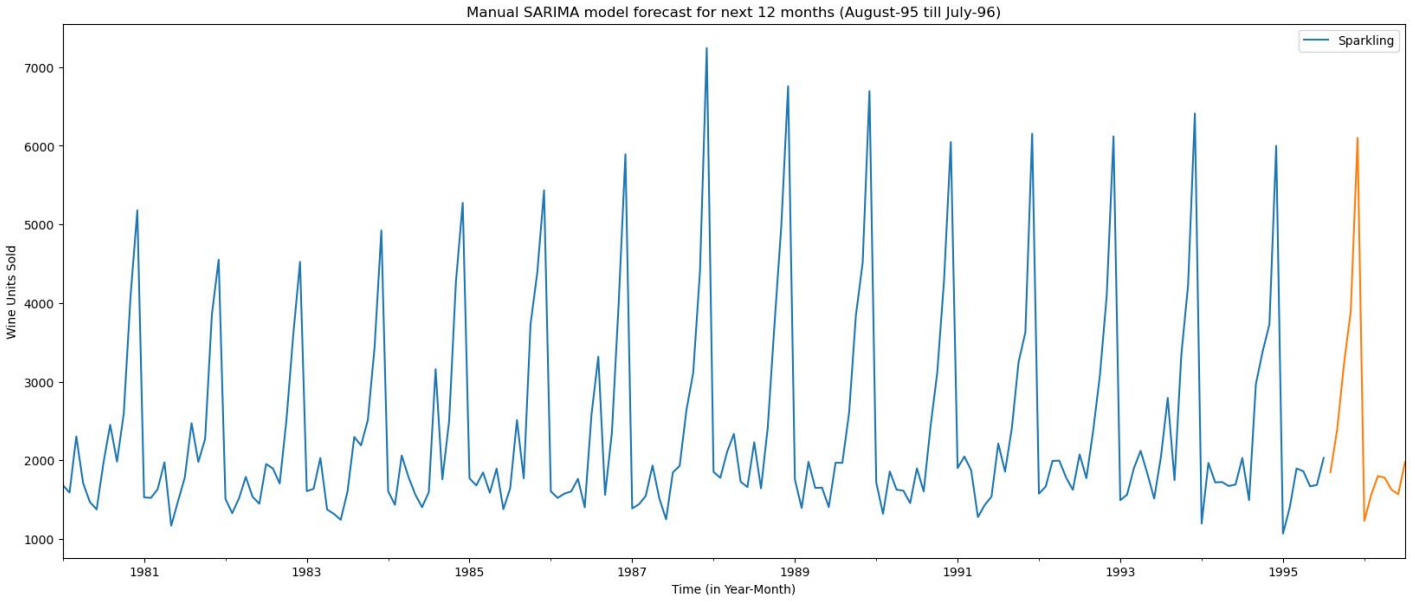

Figure 83: Manual SARIMA model forecast for next 12 months (August-95 till July-96) (Sparkling)

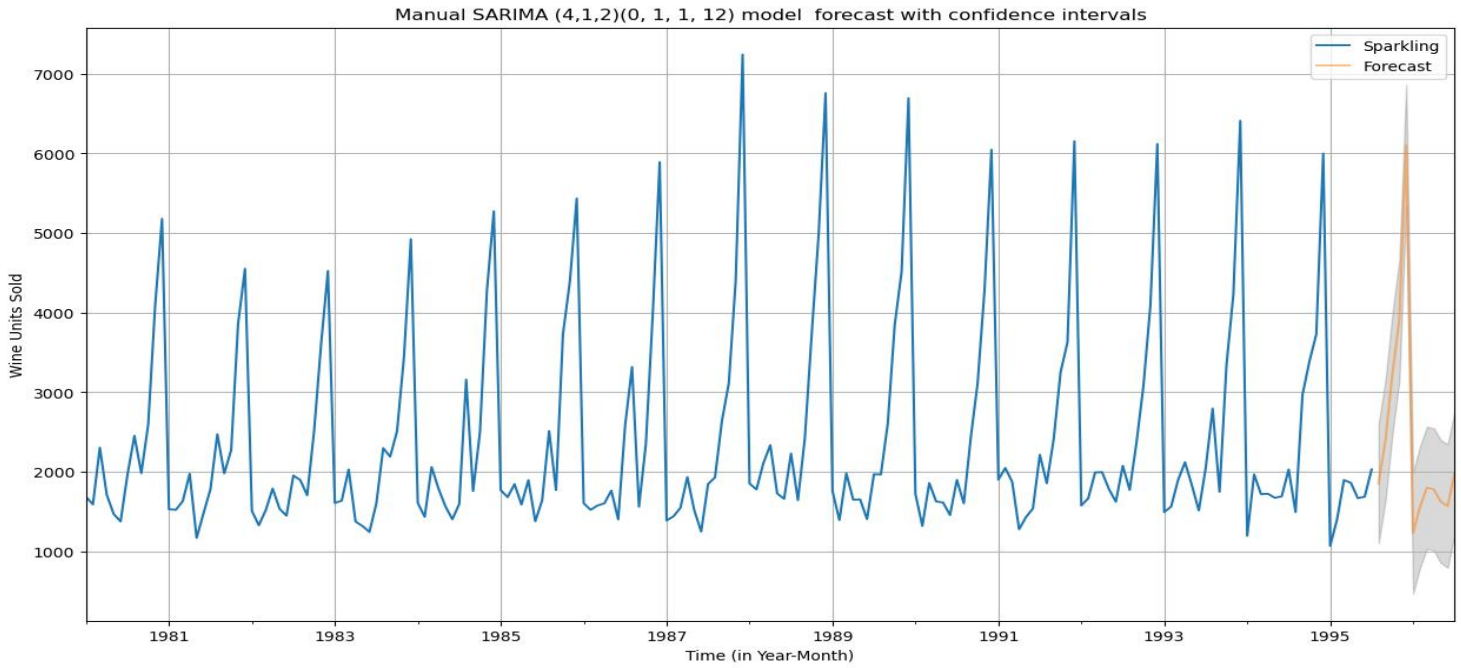

Figure 84:Manual SARIMA (4,1,2)(0, 1, 1, 12) model forecast with confidence intervals (Sparkling)

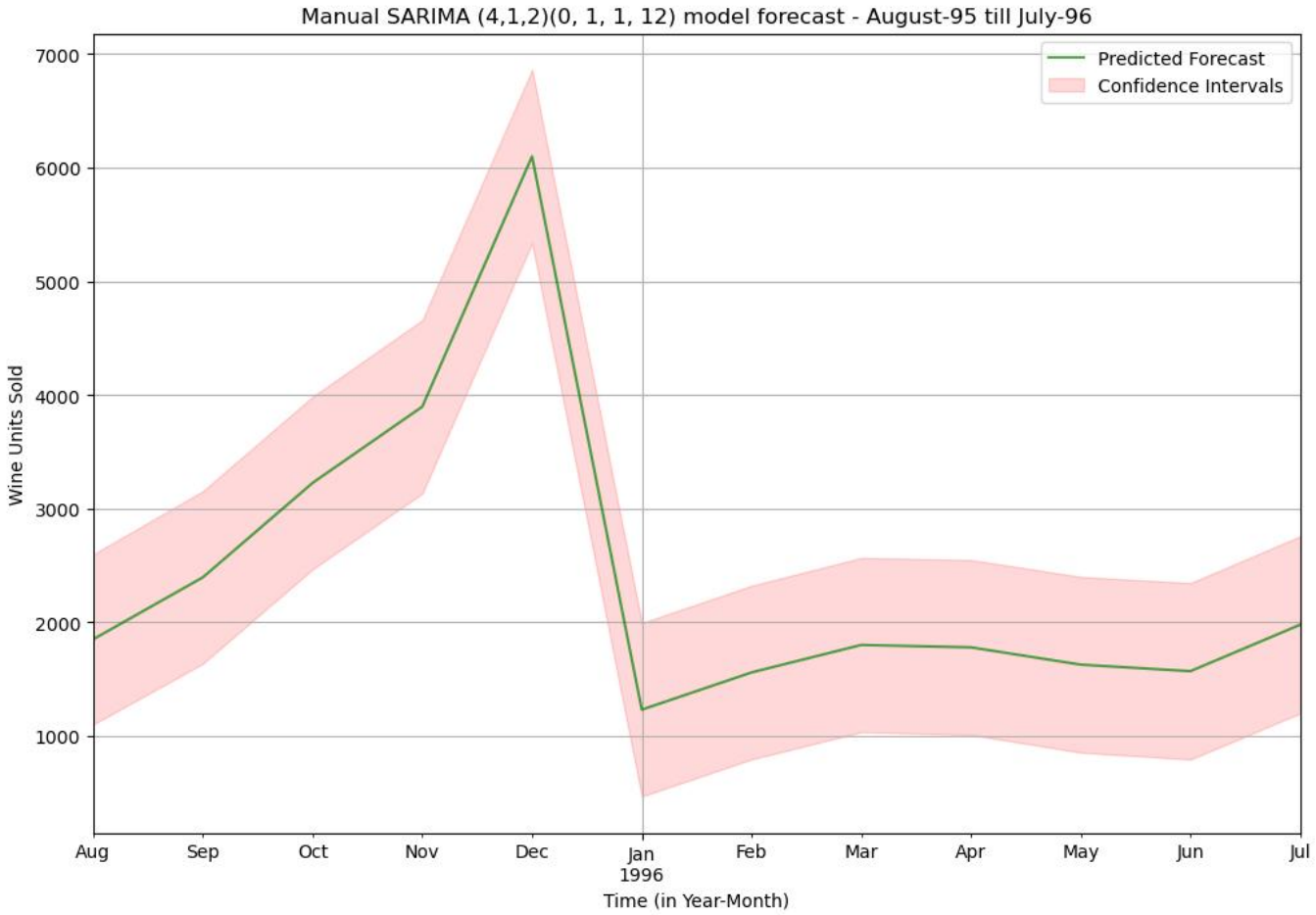

Figure 85 : Manual SARIMA (4,1,2)(0, 1, 1, 12) model forecast - August-95 till July-96(Sparkling)

```
1995-08-01    1870.288172
1995-09-01    3277.197950
1995-10-01    2960.054554
1995-11-01    3016.837313
1995-12-01    4497.760817
1996-01-01     773.278295
1996-02-01    1031.211281
1996-03-01    1299.619470
1996-04-01    1249.022045
1996-05-01    1132.057617
1996-06-01    1157.777524
1996-07-01    1604.041540
Freq: MS, dtype: float64
```

Actionable Insights & Recommendations

- SARIMA Models Excel: Seasonal adjustments in SARIMA models are critical for improving forecast accuracy in datasets with pronounced seasonality. The manual SARIMA model performs significantly better than the automated versions, emphasizing the importance of custom tuning.
- Exponential Smoothing Works Well: Triple exponential smoothing with optimal parameters is a close competitor to SARIMA models. It can be a preferred method when SARIMA complexity is unnecessary or parameter tuning is challenging.
- ARIMA Models Struggle: ARIMA models without seasonal components (e.g., Auto ARIMA and Manual ARIMA) perform poorly due to the dataset's strong seasonal nature.

- Use Manual SARIMA (4, 1, 2)(0, 1, 1, 12) or Triple Exponential Smoothing (Alpha=0.20, Beta=0.35, Gamma=0.80) for strategic decision-making, such as:
    1. Inventory Planning: Ensure optimal stock levels during high-demand months.
    2. Sales Targeting: Align sales and marketing efforts with predicted demand peaks.
    3. Budget Allocation: Allocate resources efficiently based on forecasted demand patterns.
- Consider Auto SARIMA for automated forecasting in scenarios requiring faster deployment with less manual intervention, especially for less critical forecasting tasks.
- Leverage the 2-point Trailing Moving Average for quick, short-term demand projections where computation simplicity is preferred over precision.
- Discontinue the use of Auto ARIMA, Manual ARIMA, and simpler methods like Linear Regression and Simple/Double Exponential Smoothing as they fail to capture seasonal and trend dynamics critical for accurate wine sales forecasting.
- Incorporate seasonal patterns into marketing and pricing strategies, such as Offering promotions or discounts during low-demand months and Stocking and advertising premium wines during historically high-demand periods (e.g., holidays).
- Regularly evaluate forecasting models using updated data to ensure continued accuracy and relevance.
- Fine-tune parameters of the best-performing models periodically, especially for manual SARIMA.