# TIME SERIES FORECASTING FINAL PROJECT


**Balaji M P**
**PGP DSBA Online -March' 22**
**Date: 23.10.2022**

greatlearning
*Power Ahead*

# TABLE OF CONTENTS





# LIST OF FIGURES

# LIST OF TABLES

# Rose Wine Analysis

## Executive Summary

Data on wine sales from the 20th century are available from ABC Estate Wines, a wine producing firm, and should be examined. With the provided information, an estimate of wine sales in the 20th century must be forecasted.


**Fig.1 Rose Wine Analysis**

## Introduction

The purpose of this **report** is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of **sales of Rose wine from 20th century.**

## Data Dictionary

| Variable Name | Description |
|---|---|
| **YearMonth** | Represents the year and month in which the sales were recorded |
| **Rose** | Denotes the number of wine units sold |

## Data Description

1. **YearMonth:** Datatime variable from 1980-01 to 1995-07
2. **Rose:** Continuous from 89 to 267

## Sample of the dataset

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

**Table 1. Sample of first 5 rows of the dataset**

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

**Table 2. Sample of last 5 rows of the dataset**

Dataset has 2 columns which captures the Year and Month of recorded data and the number of units sold on corresponding Year-Month respectively.

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

**Let us check the types of variables in the data frame and check for missing values in the dataset**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   YearMonth  187 non-null    object 
 1   Rose       185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.0+ KB
```

**Fig.2 Details of the dataset columns**

**The dataset has 2 variables and 187 rows in total**. The **"YearMonth"** column can be **deleted after creating a suitable time stamp column because it is not necessary** for our modelling. The column **Rose** is of **float type**. Additionally, we can observe from the data above that **Rose column has some missing values** which needs to be imputed further as it's a time series.

### Time Stamp created from 'YearMonth' column

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

**Fig.3 Details of the dataset columns**

## Resulting dataset after removing the "Year-Month" column and appending Time_Stamp column

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

**Fig.4 Details of the dataset columns**

**Time_Stamp** column has been **set as index** of the dataset and **column Rose** has **been renamed as Rose_Wine_Sales.**

## Renaming the columns of the data frame

The below mentioned columns of the data frame have been renamed as shown.

| Original Column Name | Renamed Column Name |
|---|---|
| **Rose** | **Rose_Wine_Sales** |

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

**Fig.5 Details of the dataset columns after renaming**

## Checking null values in the dataset

| Time_Stamp | Rose_Wine_Sales |
|---|---|
| 1994-07-31 | NaN |
| 1994-08-31 | NaN |

**Fig.6 Null values in the dataset**

**As can be seen from the above figure, there are 2 null values present in the dataset. Since it's a time series we cannot remove it and hence must be imputed.**


**Fig.7 Graph plot of the Rose wine sales dataset**

**Observation:**

- The data set provided contains sales information from January 1980 to July 1995.
- We can see from the plot that there has been a decline in sales over time. Over the years, the sales have gradually decreased. The data also exhibit some seasonality, as may be shown.
- There are 2 missing values which must be imputed.

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

### Handling Missing Values

```
Time_Stamp
1994-01-31    30.00
1994-02-28    35.00
1994-03-31    42.00
1994-04-30    48.00
1994-05-31    44.00
1994-06-30    45.00
1994-07-31    45.34
1994-08-31    45.67
1994-09-30    46.00
1994-10-31    51.00
1994-11-30    63.00
1994-12-31    84.00
Name: Rose_Wine_Sales, dtype: float64
```

**Fig.8 Imputed values of the dataset**

As can be seen from Fig.6, **values are missing for July and August month of 1994**. Since it's a time series, the missing values cannot be removed. We **have imputed them using linear interpolation**.

```
Rose_Wine_Sales    0
dtype: int64
```

**Fig.9 Null values after imputation**

## Descriptive Summary of the Dataset

| | Rose_Wine_Sales |
|---|---|
| count | 187.000000 |
| mean | 89.914492 |
| std | 39.238264 |
| min | 28.000000 |
| 25% | 62.500000 |
| 50% | 85.000000 |
| 75% | 111.000000 |
| max | 267.000000 |

**Fig.10 Descriptive Summary of Rose_Wine_Sales column**

**Observation:**

- 90 bottles of rose wine are typically sold each month.
- Between 62 and 111 units make up more than 50% of the sold rose wine units.
- The lowest unit sold is 28 units, while the highest unit sold is 267 units.

## Exploratory Analysis

Let us analyze the wine sales across different years and months using boxplots

### Yearly Plot

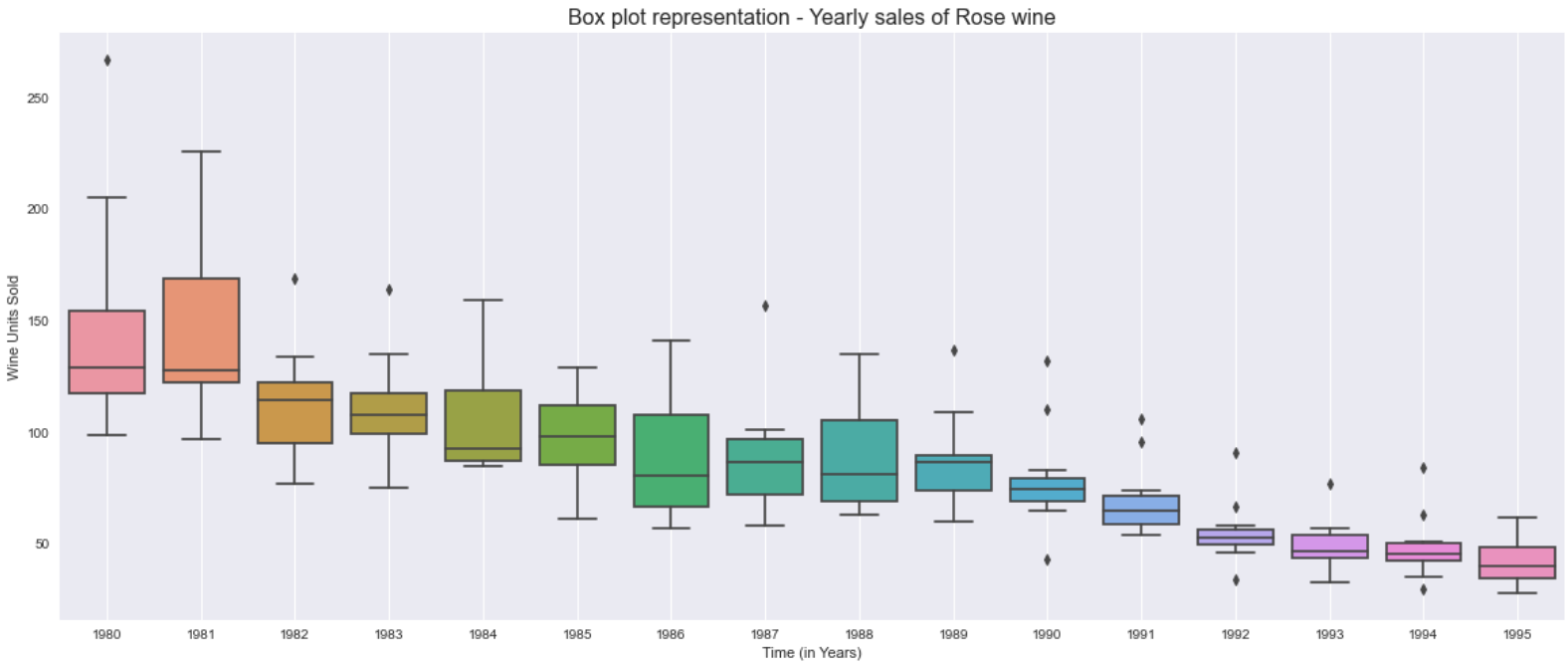


**Fig.11 Yearly plot of Rose wine sales**

**Observation:**

- We can see from the figure above that sales of rose wine have been declining over time.
- After 1992, the median sales have been at their lowest levels, having peaked in 1980 and 1981.
- Additionally, we can see that there are outliers in the box plots.

## Monthly Plot

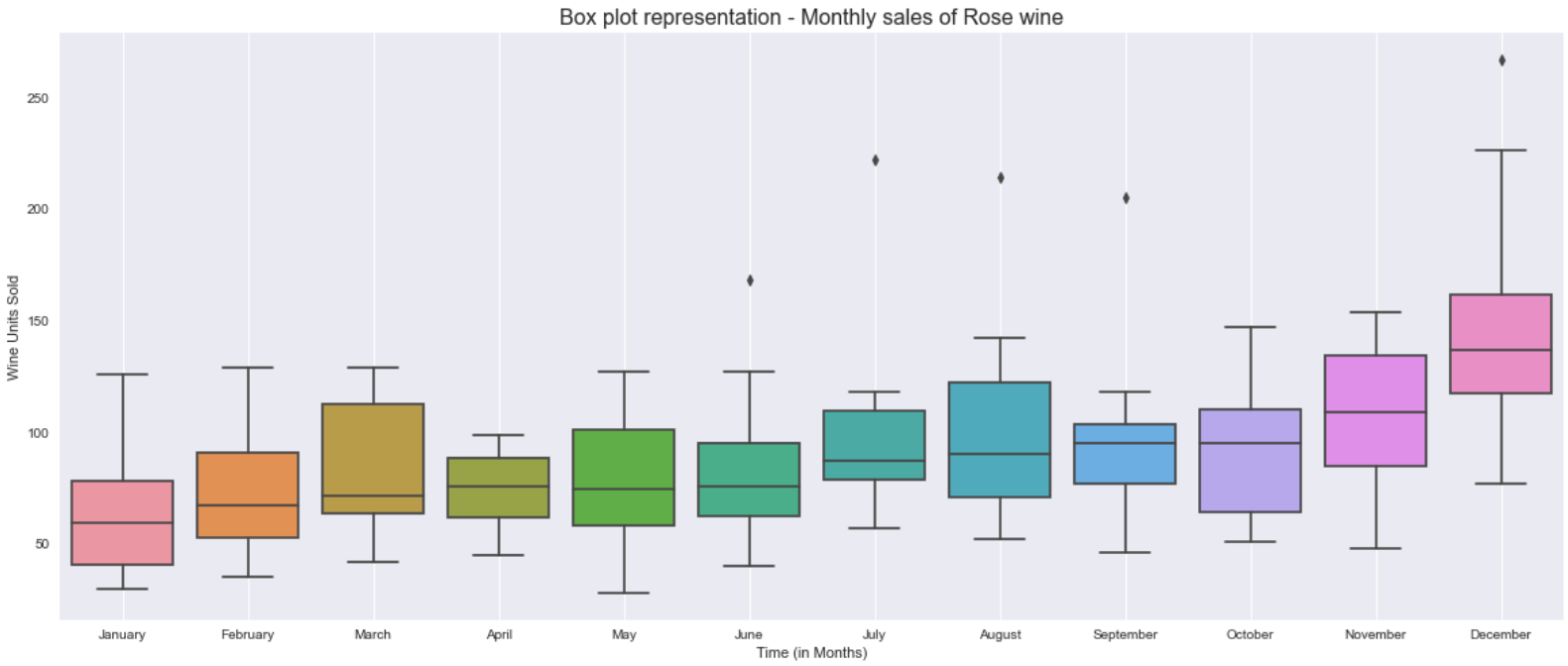


**Fig.11 Monthly plot of Rose wine sales**

**Observation:**

- The sales trajectory appears to be precisely the reverse of that seen in the yearly plot, increasing near the end of each year.
- January has the lowest wine sales while December sees the greatest. The sales modestly grow from January to August and then sharply climb after that.
- Additionally, we can see that there are outliers in the box plots.

## Annual Sales

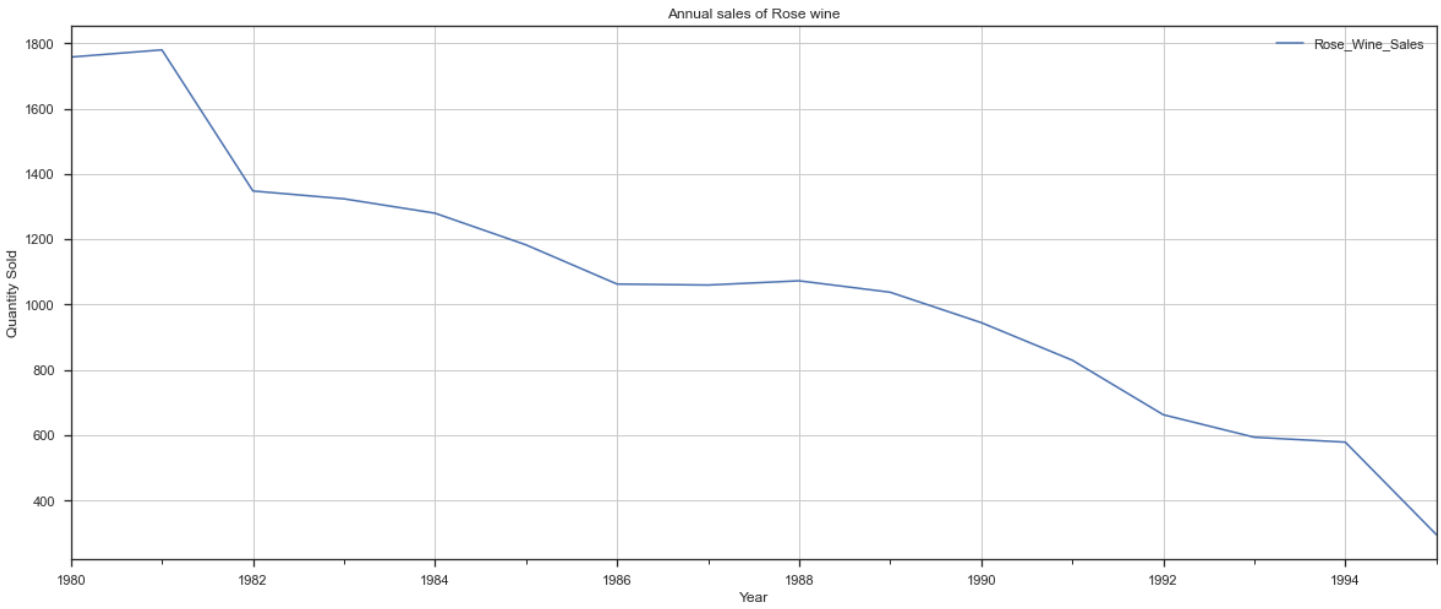


**Fig.12 Line plot – Annual sales**

## Quarterly Sales

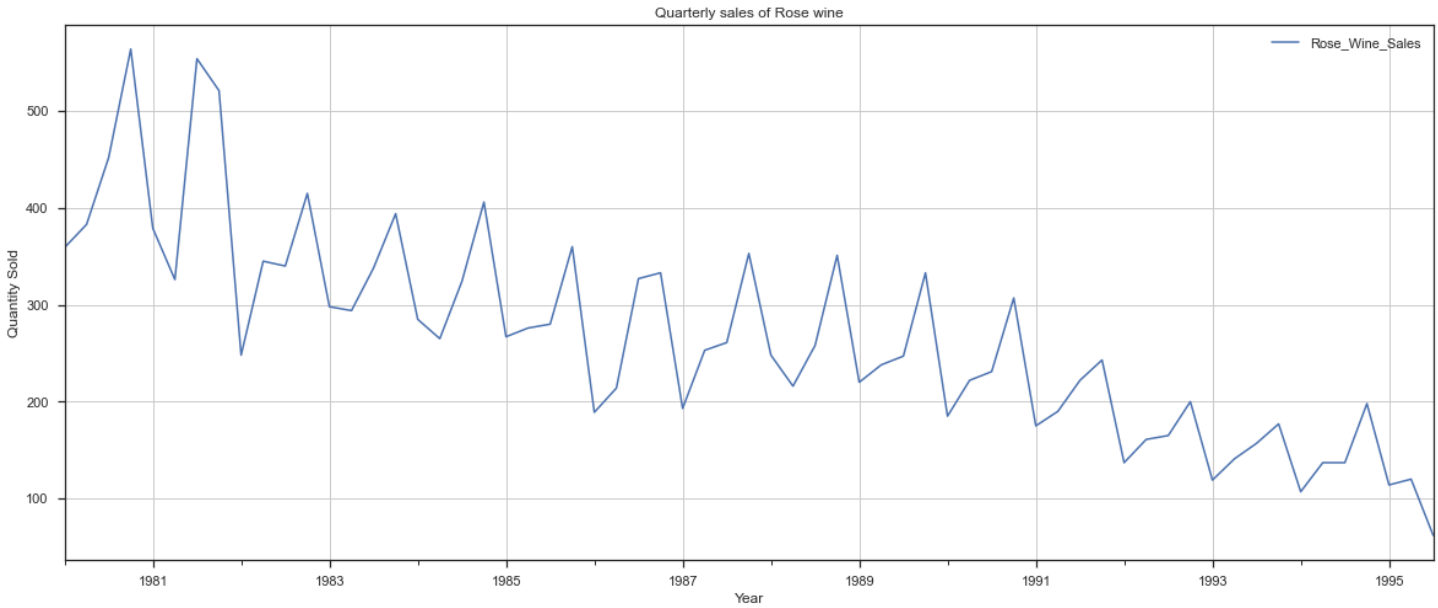


**Fig.13 Line plot – Quarterly sales**

## Monthly Sales across Different Years

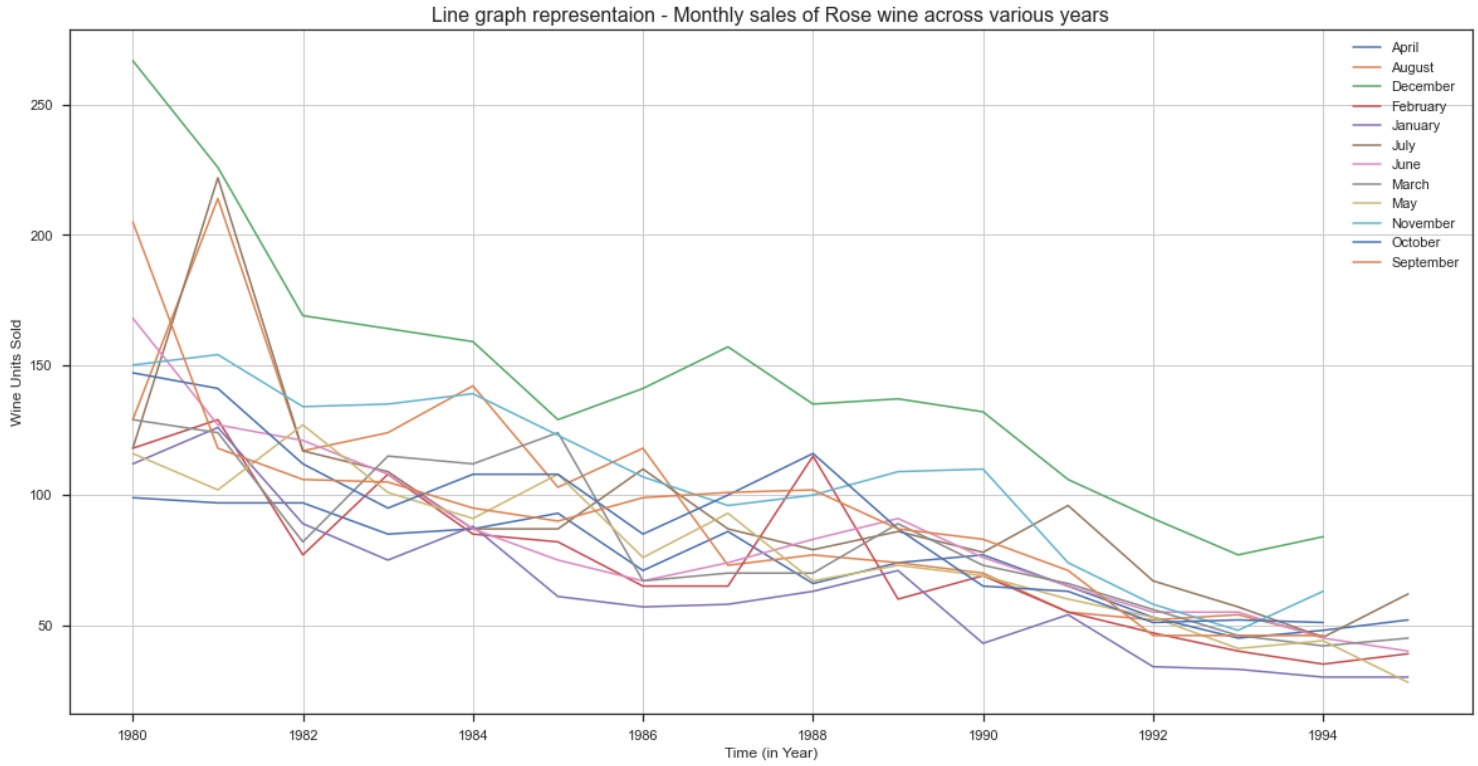


**Fig.14 Line plot – Monthly sales across different years**

## Empirical Cumulative Distribution Plot

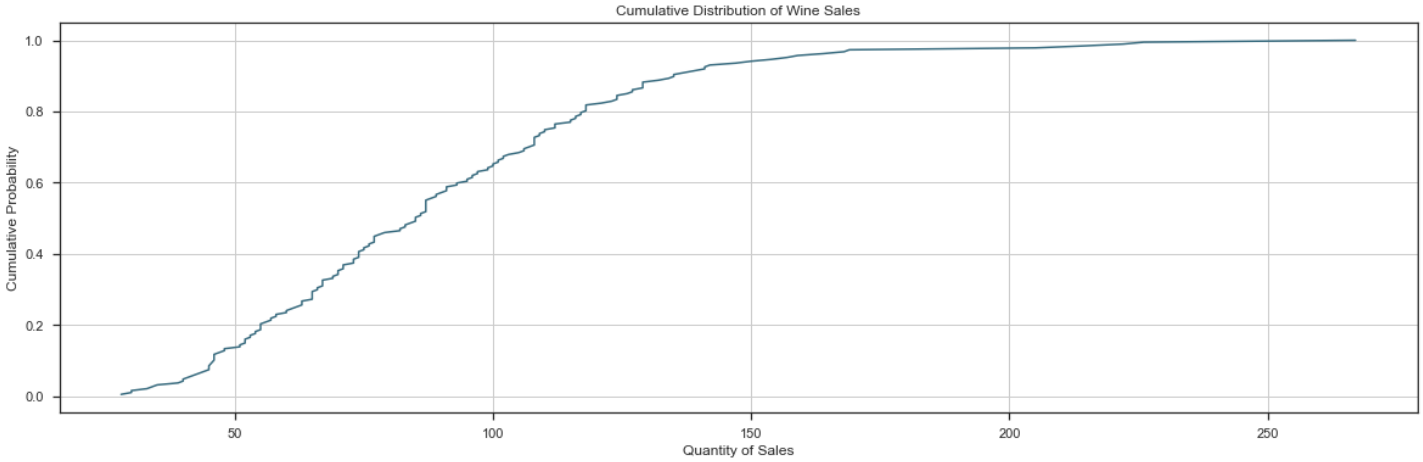


**Fig.15 Line plot – Empirical cumulative distribution function**

## Monthly Time Series Plot

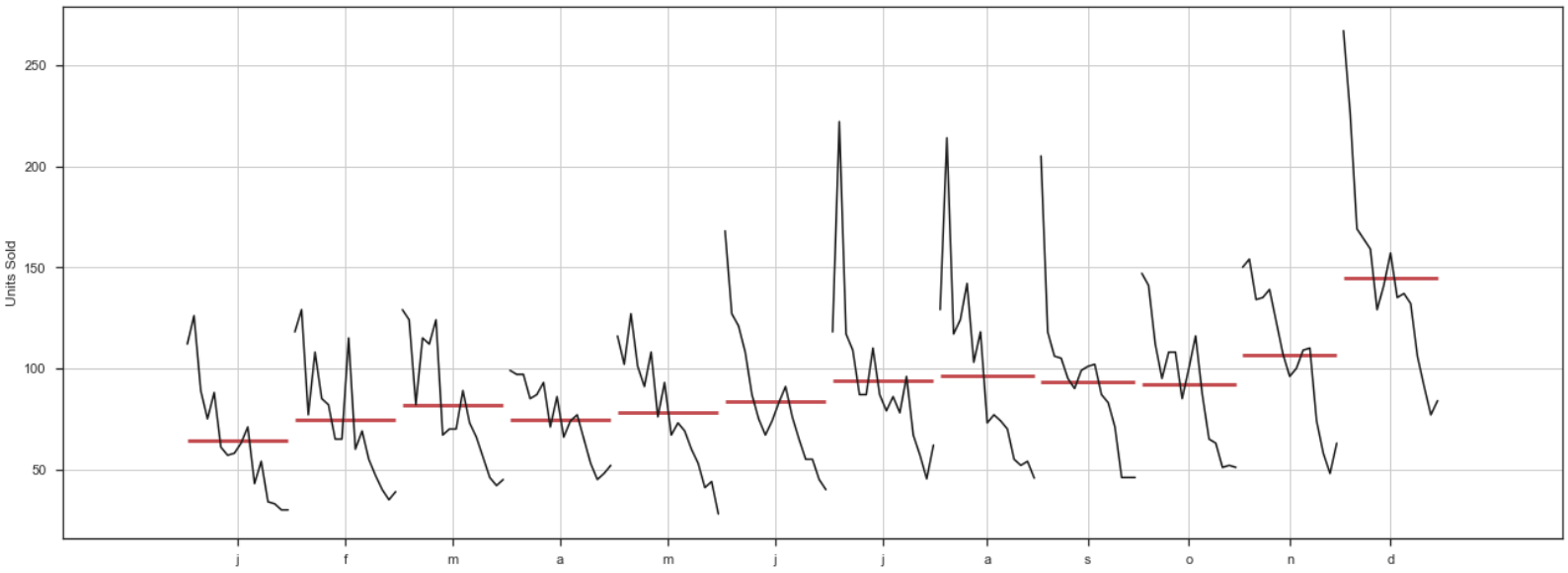


**Fig.16 Line plot – Monthly time series**

**Observation:**

- After 1981, the sales fell drastically. Sales are typically lowest in the first quarter and highest in the fourth quarter.
- Every year, December has the highest sales, followed by November and October. January had the lowest sales.
- From the cumulative distribution graph, we can observe that around 70 to 75 percent of the units sold are fewer than 100, and 90% of the units sold are less than 150. Only 15% of sales involved less than 50 items. Therefore, it is clear that the bulk of sales were in the range of 50 to 100 units.

## Average Wine sales per month & change percentage over each month

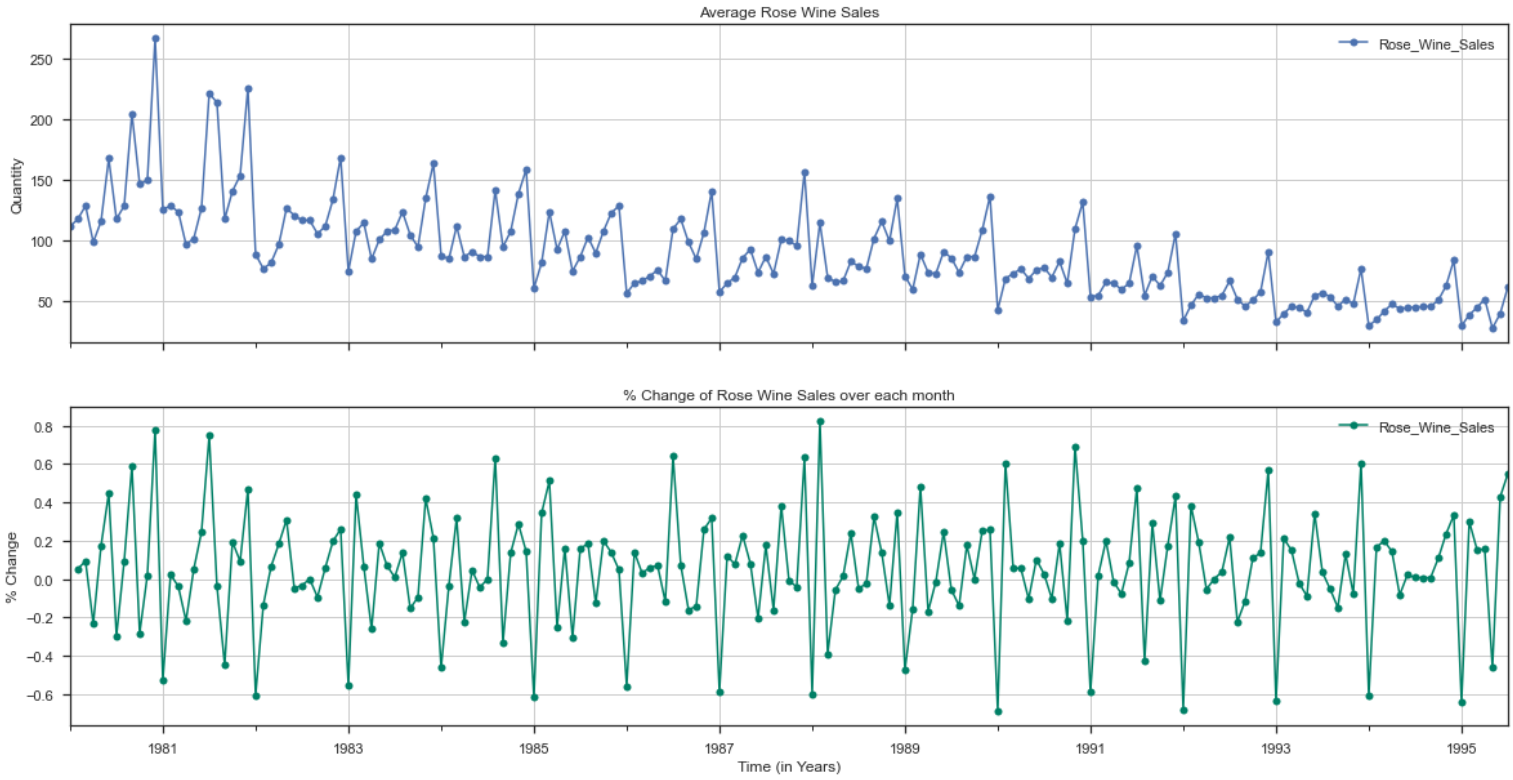


**Fig.17 Line plot – Average and % Change over each month**

**Observation:**

- We can see that there is a declining trend and seasonality from the average sales and % change plots. Additionally, the seasonality in the percentage change appears to be consistent throughout all the years.

## Decomposition of Time Series

### Additive Decomposition

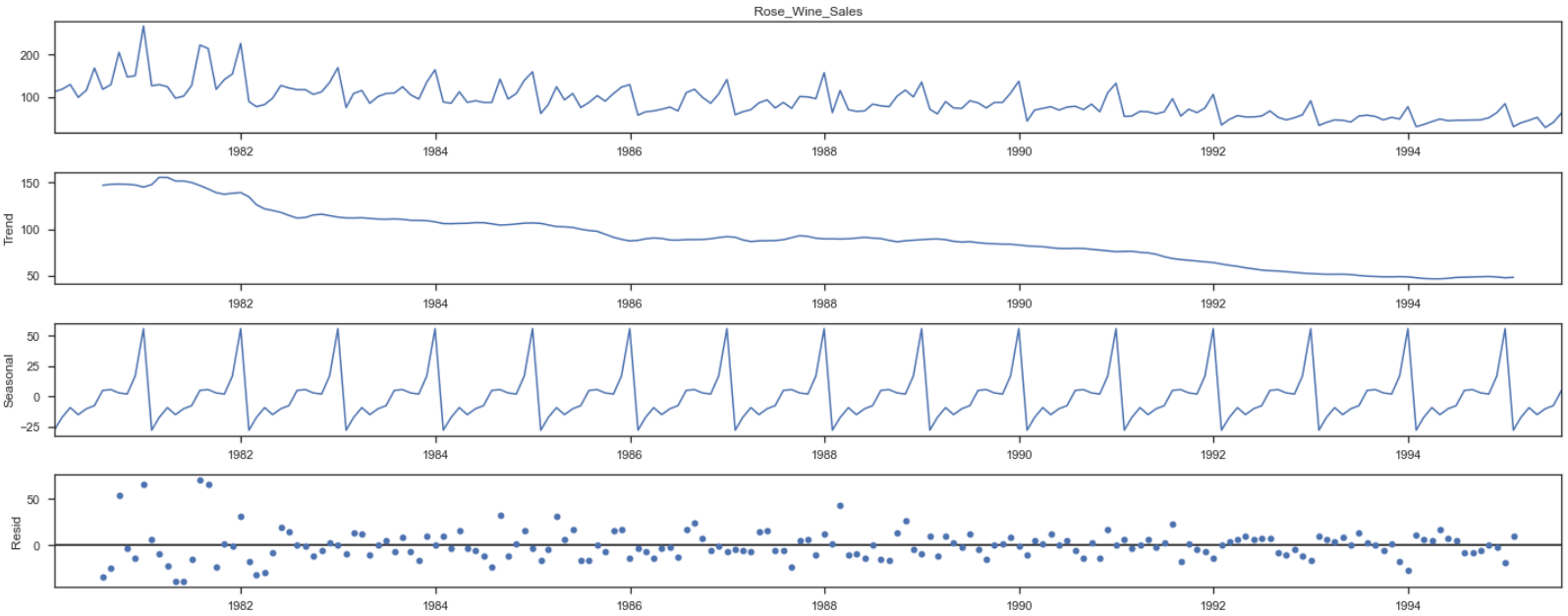


**Fig.18 Additive decomposition of time series**

**Trend**

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

Name: trend, dtype: float64

**Seasonality**

| Time_Stamp | |
|---|---|
| 1980-01-31 | -27.908702 |
| 1980-02-29 | -17.435680 |
| 1980-03-31 | -9.285889 |
| 1980-04-30 | -15.098389 |
| 1980-05-31 | -10.196603 |
| 1980-06-30 | -7.678746 |
| 1980-07-31 | 4.897298 |
| 1980-08-31 | 5.499853 |
| 1980-09-30 | 2.774631 |
| 1980-10-31 | 1.871853 |
| 1980-11-30 | 16.846853 |
| 1980-12-31 | 55.713520 |

Name: seasonal, dtype: float64

**Residual**

| Time_Stamp | |
|---|---|
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | -33.980631 |
| 1980-08-31 | -24.624853 |
| 1980-09-30 | 53.850369 |
| 1980-10-31 | -2.955187 |
| 1980-11-30 | -14.263520 |
| 1980-12-31 | 66.161480 |

Name: resid, dtype: float64

**Fig.19 Additive Decomposition - Sample of Trend, Seasonality & Residual values**

## Multiplicative Decomposition

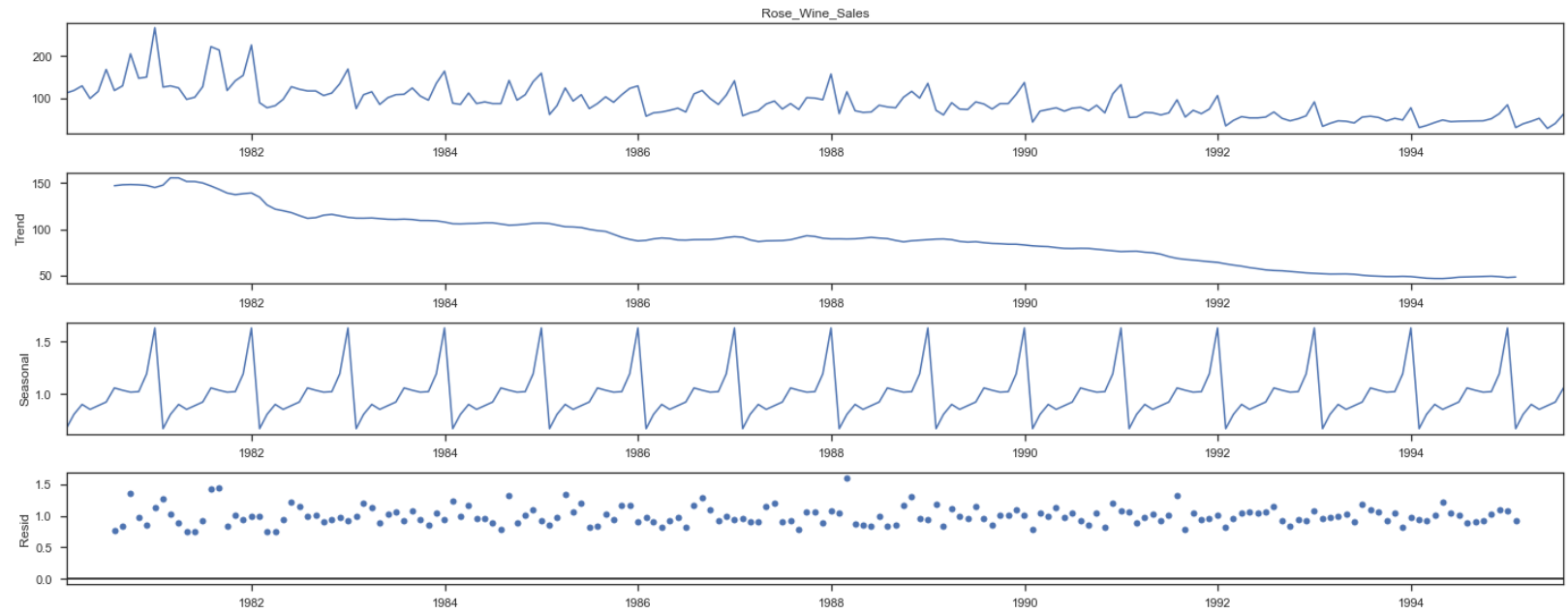


**Fig.20.1 Multiplicative decomposition of time series**

**Trend**

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

Name: trend, dtype: float64

**Seasonality**

| Time_Stamp | |
|---|---|
| 1980-01-31 | 0.670111 |
| 1980-02-29 | 0.806163 |
| 1980-03-31 | 0.901163 |
| 1980-04-30 | 0.854023 |
| 1980-05-31 | 0.889414 |
| 1980-06-30 | 0.923984 |
| 1980-07-31 | 1.058046 |
| 1980-08-31 | 1.035885 |
| 1980-09-30 | 1.017647 |
| 1980-10-31 | 1.022572 |
| 1980-11-30 | 1.192347 |
| 1980-12-31 | 1.628644 |

Name: seasonal, dtype: float64

**Residual**

| Time_Stamp | |
|---|---|
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 0.758253 |
| 1980-08-31 | 0.840717 |
| 1980-09-30 | 1.357675 |
| 1980-10-31 | 0.970772 |
| 1980-11-30 | 0.853379 |
| 1980-12-31 | 1.129647 |

Name: resid, dtype: float64

**Fig.20.2 Multiplicative Decomposition - Sample of Trend, Seasonality & Residual values**

**Observation:**

- We can see from the graphs above that the time series has a falling trend and is seasonal.
- The residual patterns after additive decomposition of the time series appear to represent the seasonal element and exhibit substantial variation.
- In the multiplicative decomposition of the time series, it has been observed that the seasonal fluctuation of residuals is under control.
- The size of the seasonal variations doesn't change on comparison, but the residuals are tightly controlled by the multiplicative decomposition. In addition to this, the residuals are not independent of seasonality thus we may assume that it is **multiplicative**.

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

Train and test data are separated from the provided dataset. Sales data up to 1991 is included in the training data, while data from 1991 through 1995 is used for testing.

First few rows of Training Data

| Time_Stamp | Rose_Wine_Sales |
|---|---|
| 1980-01-31 | 112.0 |
| 1980-02-29 | 118.0 |
| 1980-03-31 | 129.0 |
| 1980-04-30 | 99.0 |
| 1980-05-31 | 116.0 |
| 1980-06-30 | 168.0 |
| 1980-07-31 | 118.0 |
| 1980-08-31 | 129.0 |
| 1980-09-30 | 205.0 |
| 1980-10-31 | 147.0 |

Last few rows of Training Data

| Time_Stamp | Rose_Wine_Sales |
|---|---|
| 1990-03-31 | 73.0 |
| 1990-04-30 | 77.0 |
| 1990-05-31 | 69.0 |
| 1990-06-30 | 76.0 |
| 1990-07-31 | 78.0 |
| 1990-08-31 | 70.0 |
| 1990-09-30 | 83.0 |
| 1990-10-31 | 65.0 |
| 1990-11-30 | 110.0 |
| 1990-12-31 | 132.0 |

**Fig.21.1 First and Last few rows of Train data**

First few rows of Test Data

| Time_Stamp | Rose_Wine_Sales |
|---|---|
| 1991-01-31 | 54.0 |
| 1991-02-28 | 55.0 |
| 1991-03-31 | 66.0 |
| 1991-04-30 | 65.0 |
| 1991-05-31 | 60.0 |
| 1991-06-30 | 65.0 |
| 1991-07-31 | 96.0 |
| 1991-08-31 | 55.0 |
| 1991-09-30 | 71.0 |
| 1991-10-31 | 63.0 |

Last few rows of Test Data

| Time_Stamp | Rose_Wine_Sales |
|---|---|
| 1994-10-31 | 51.0 |
| 1994-11-30 | 63.0 |
| 1994-12-31 | 84.0 |
| 1995-01-31 | 30.0 |
| 1995-02-28 | 39.0 |
| 1995-03-31 | 45.0 |
| 1995-04-30 | 52.0 |
| 1995-05-31 | 28.0 |
| 1995-06-30 | 40.0 |
| 1995-07-31 | 62.0 |

**Fig.21.2 First and Last few rows of Test data**

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

**Fig.22 Count summary on train and test data**

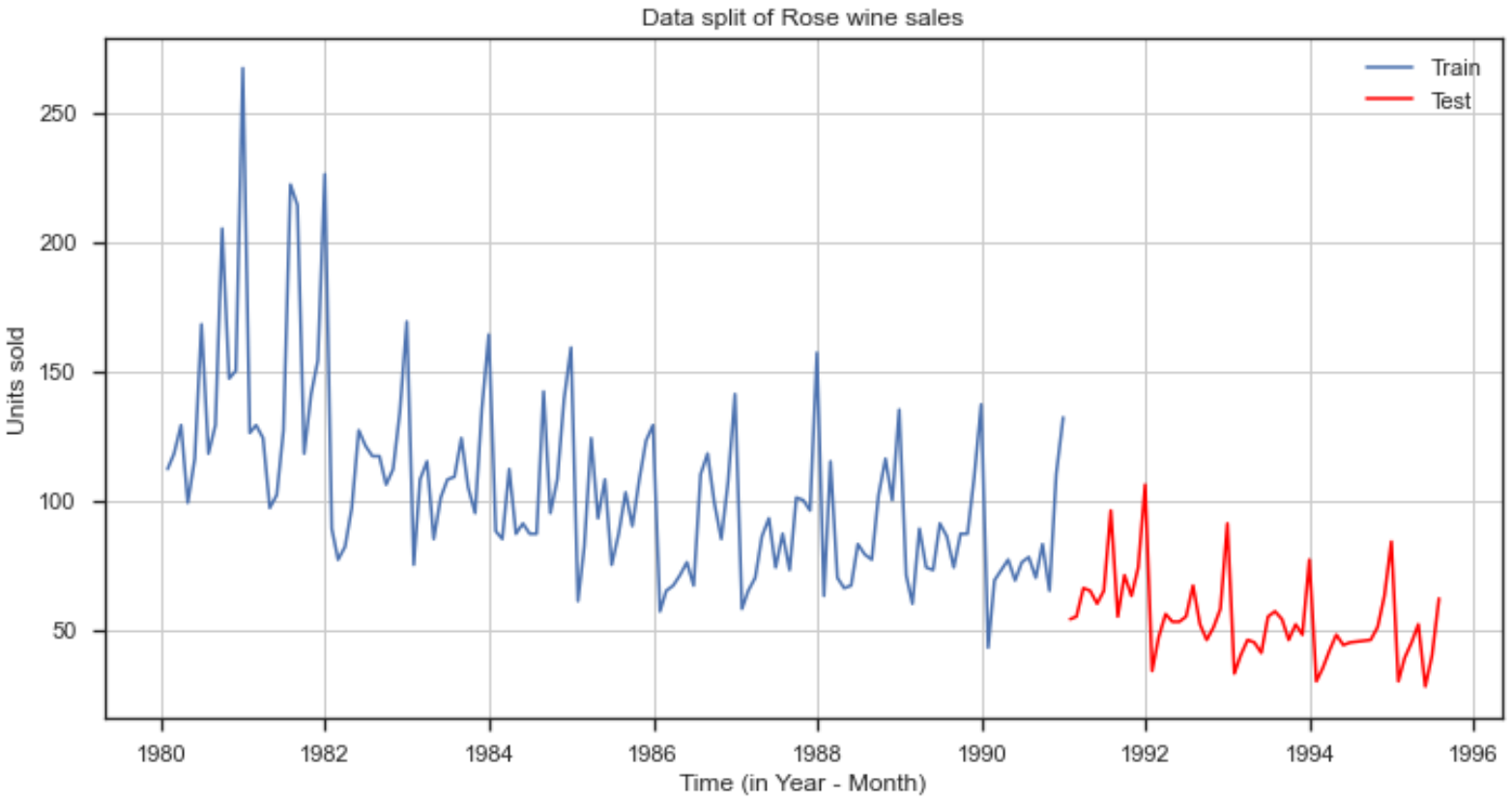


**Fig.23 Line Plot – Splitting of time series into Train & Test data**

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

### Model 1 – Linear Regression

**For this particular linear regression, we are going to regress the 'Rose_Wine_Sales' variable against the order of the occurrence.**

**For the selection criteria, the below Linear Regression model is built by using default parameters.**

```
▾ LinearRegression
LinearRegression()
```

**Fig.24 Rose Wine – Linear regression model**

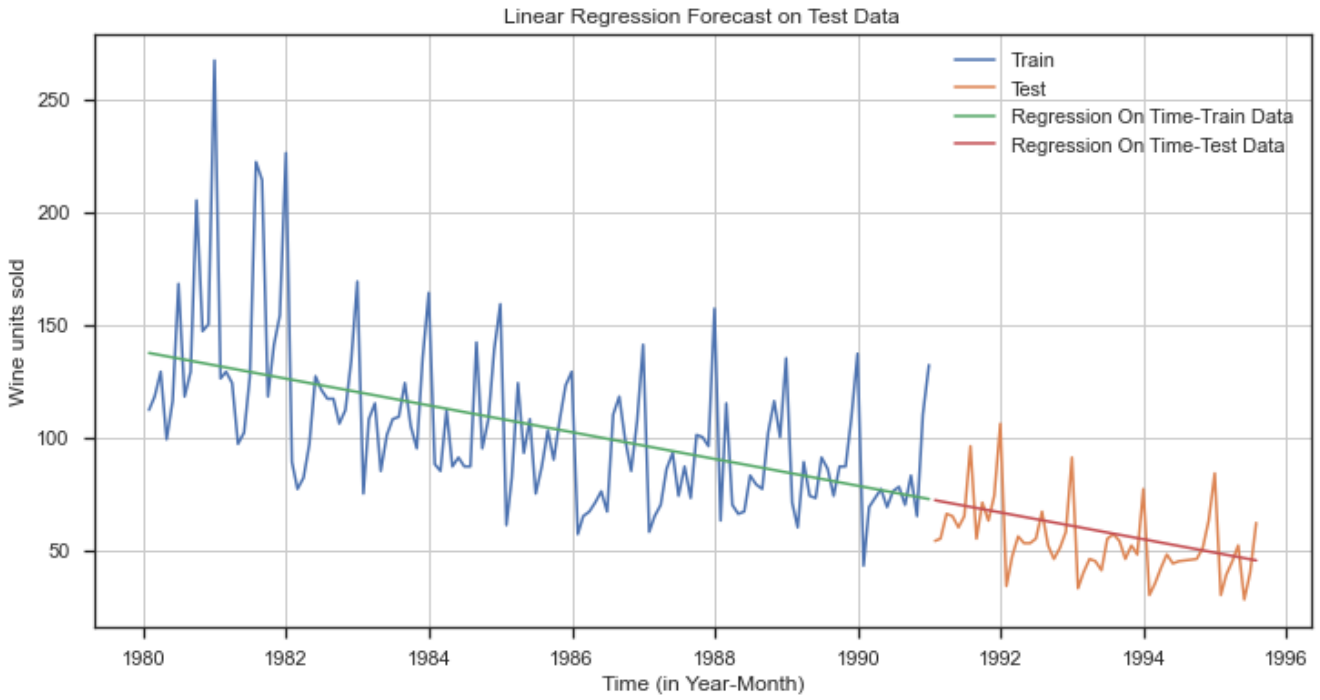


**Fig.25 Linear regression on Test data**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- The train and test data **trends have been caught** by the linear regression model however, it is **unable to account for seasonality**
- The root means squared error **(RMSE)** for the linear regression model is **15.268**. The size of the seasonal

## Linear Regression: Model Evaluation

| Performance Metric | |
|---|---|
| **Test RMSE** | **15.268887** |

## Model 2 – Naïve Forecast

**For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.**

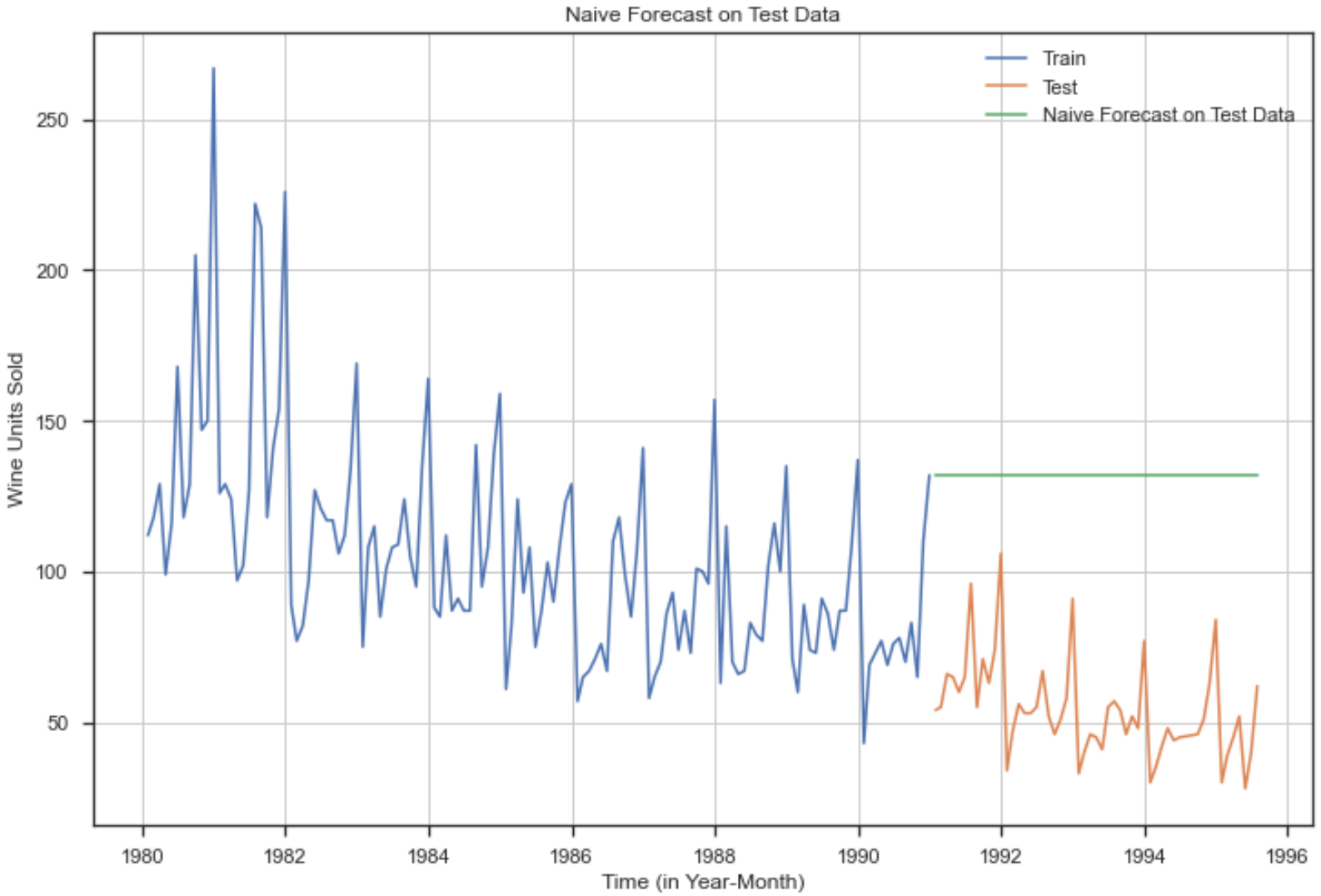


**Fig.26 Naïve forecast on Test data**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- The **seasonality and trend** of the time series data **cannot be captured** by the simple forecast model.
- The root means squared error **(RMSE)** for the naïve forecast model is **79.719** which is significantly higher than the regression model.

## Naïve Forecast: Model Evaluation

| Performance Metric | |
|---|---|
| **Test RMSE** | **79.718576** |

## Model 3 – Simple Average

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

| Time_Stamp | Rose_Wine_Sales | mean_forecast |
|---|---|---|
| 1991-01-31 | 54.0 | 104.939394 |
| 1991-02-28 | 55.0 | 104.939394 |
| 1991-03-31 | 66.0 | 104.939394 |
| 1991-04-30 | 65.0 | 104.939394 |
| 1991-05-31 | 60.0 | 104.939394 |

**Fig.27 Rose Wine – Simple Average model**

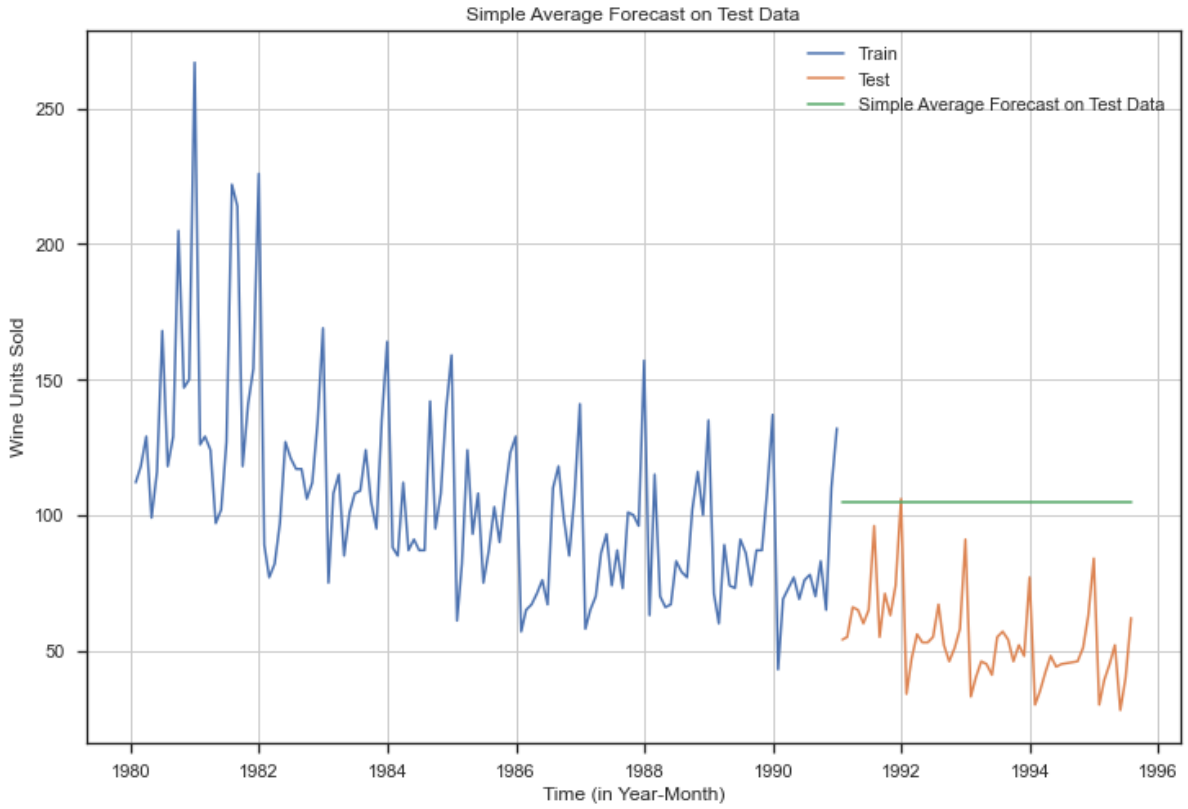


**Fig.28 Simple Average model predictions on Test data**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- The **seasonality and trend** of the time series data **cannot be captured** by the simple average model.
- The root means squared error **(RMSE)** for the simple average model is **53.46** which is significantly higher than the regression model but lower than naïve forecast model.

## Simple Average: Model Evaluation

| Performance Metric | |
|---|---|
| **Test RMSE** | **53.460367** |

## Model 4 – 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.**

| Time_Stamp | Rose_Wine_Sales | Trailing_2 | Trailing_4 | Trailing_6 | Trailing_9 |
|---|---|---|---|---|---|
| 1980-01-31 | 112.0 | NaN | NaN | NaN | NaN |
| 1980-02-29 | 118.0 | 115.0 | NaN | NaN | NaN |
| 1980-03-31 | 129.0 | 123.5 | NaN | NaN | NaN |
| 1980-04-30 | 99.0 | 114.0 | 114.50 | NaN | NaN |
| 1980-05-31 | 116.0 | 107.5 | 115.50 | NaN | NaN |
| 1980-06-30 | 168.0 | 142.0 | 128.00 | 123.666667 | NaN |
| 1980-07-31 | 118.0 | 143.0 | 125.25 | 124.666667 | NaN |
| 1980-08-31 | 129.0 | 123.5 | 132.75 | 126.500000 | NaN |
| 1980-09-30 | 205.0 | 167.0 | 155.00 | 139.166667 | 132.666667 |

**Fig.29 Rose Wine – Sample of Trailing Moving Averages**

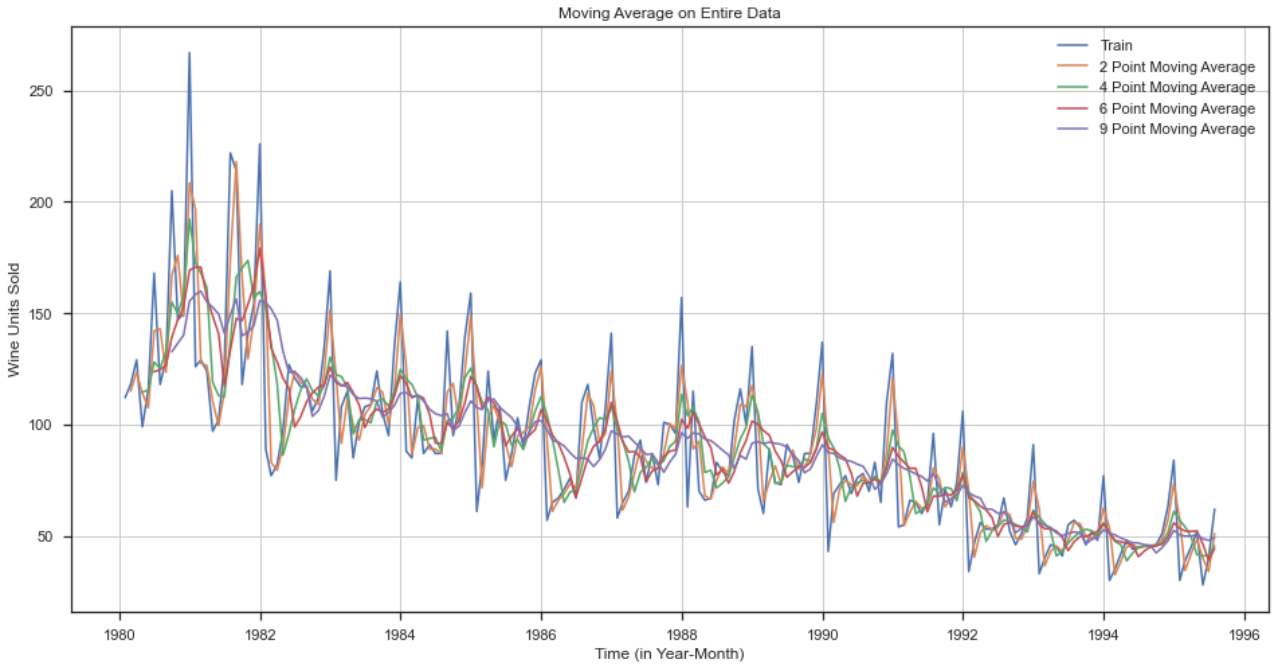


**Fig.30 Moving Average on Entire data**

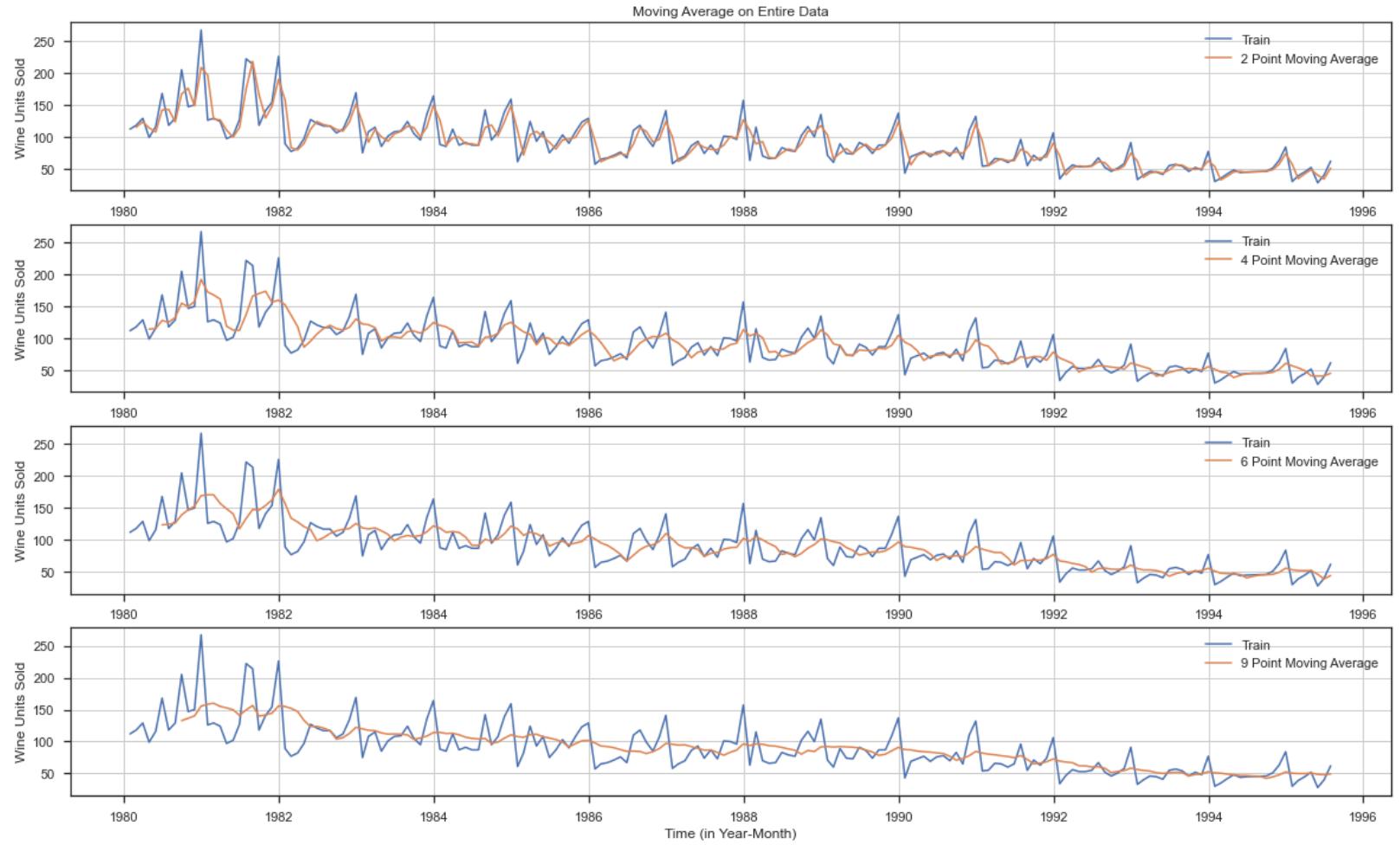


**Fig.31 Individual visualization of moving averages on entire data**

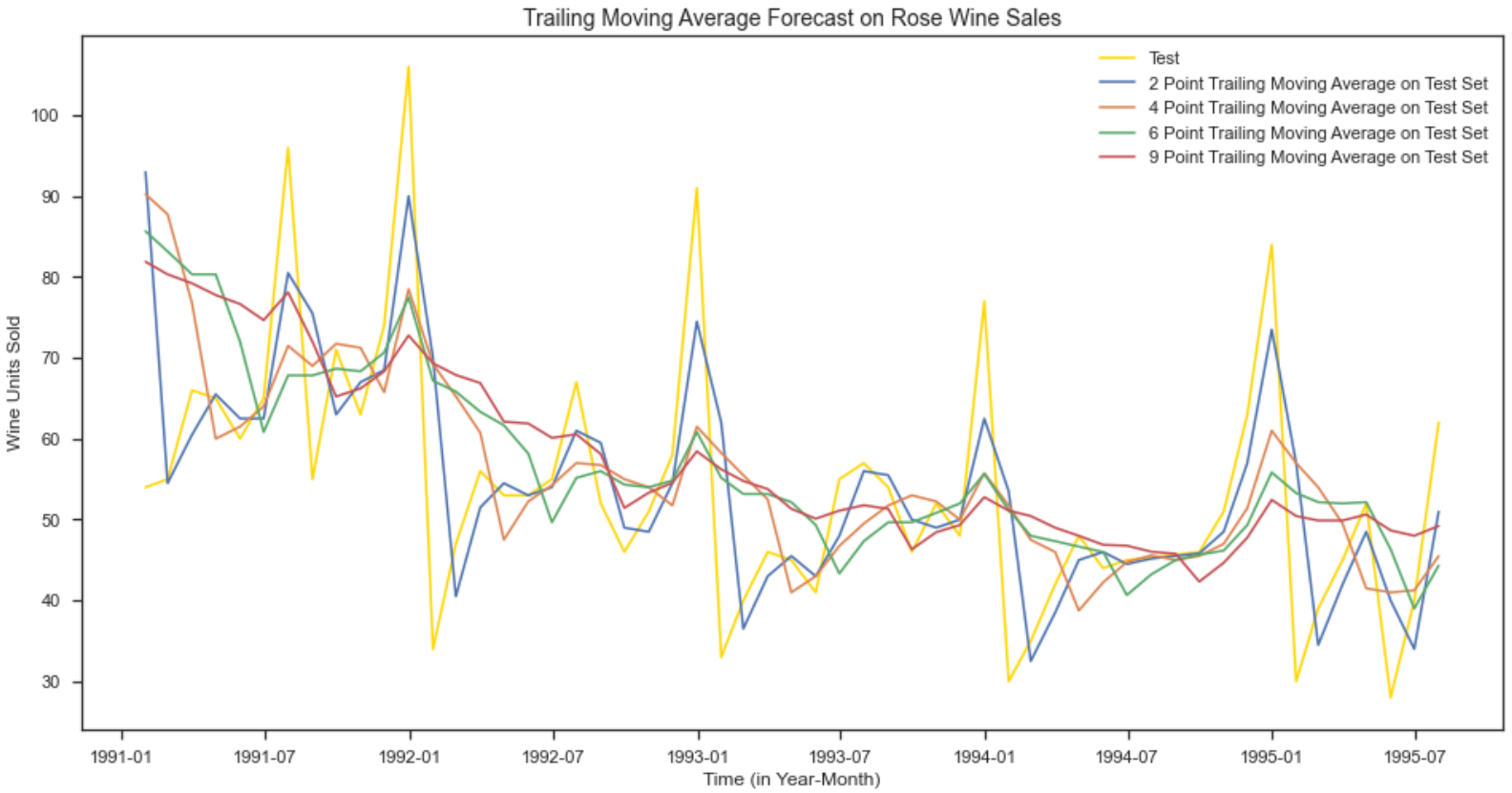


**Fig.32 Moving averages forecast on test data**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- The **seasonality and trend** of the time series data **may both be predicted** using moving average models.
- We can see how the data smooth out as the number of observation points taken increases. The **2-point TMA has characteristics that are more similar to test results** than the 9-point TMA.
- The root means squared error **(RMSE)** for the **2-point trailing average model is 11.529,** which is lowest than all models build so far.

## Moving Average: Model Evaluation

| Model | Test RMSE |
|---|---|
| 2 Point Trailing Moving Average | 11.529278 |
| 4 Point Trailing Moving Average | 14.451376 |
| 6 Point Trailing Moving Average | 14.566262 |
| 9 Point Trailing Moving Average | 14.727596 |

Let's compare the visualization of each model's predictions that we have constructed so far before investigating exponential smoothing methods.

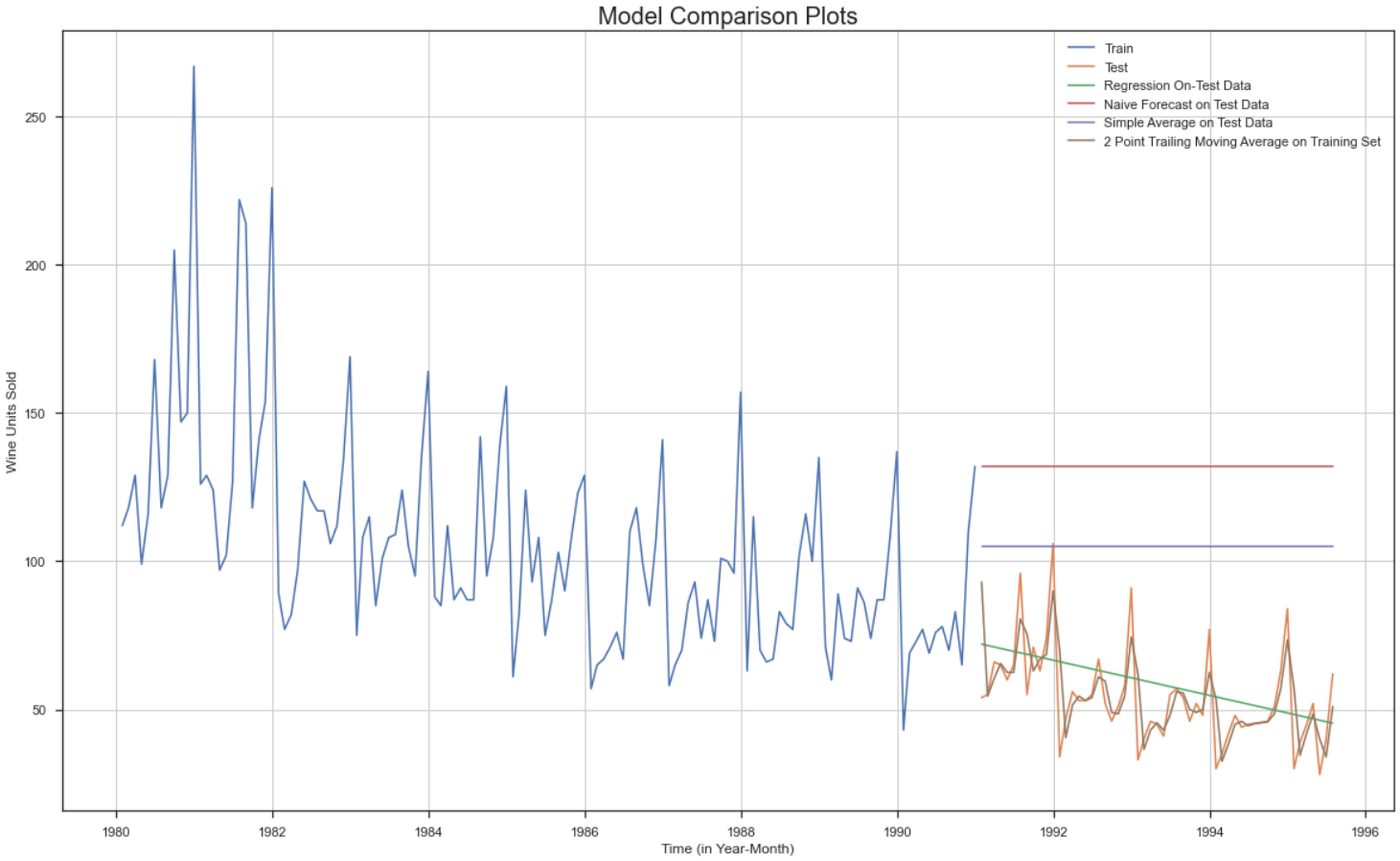


**Fig.33 Comparison of different models on test data (Regression, Naïve, Simple and Moving Average)**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- We can see from the graph above that simple average and naive forecast models fail to adequately describe the characteristics of the test data.
- The trend portion of the series has been caught using linear regression, however the seasonality has been missed
- Both trend and seasonality may be accounted for using moving average models

## Model 5 – Simple Exponential Smoothing

The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES). This method is suitable for forecasting data with no clear trend or seasonal pattern.

**In Single ES, the forecast at time (t + 1) is given by Winters,1960**

$$F_{t+1}=\alpha Y_t + (1-\alpha)F_t$$

**Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.**

**For the selection criteria, the below Simple Exponential Smoothing is built by using optimized parameters.**

```
{'smoothing_level': 0.09874989825614361,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38702255613862,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

**Fig.34 Rose Wine – Simple Exponential Smoothing Model**

| Time_Stamp | Rose_Wine_Sales | predict |
|---|---|---|
| 1991-01-31 | 54.0 | 87.104999 |
| 1991-02-28 | 55.0 | 87.104999 |
| 1991-03-31 | 66.0 | 87.104999 |
| 1991-04-30 | 65.0 | 87.104999 |
| 1991-05-31 | 60.0 | 87.104999 |

**Fig.35 Sample of SES predictions**

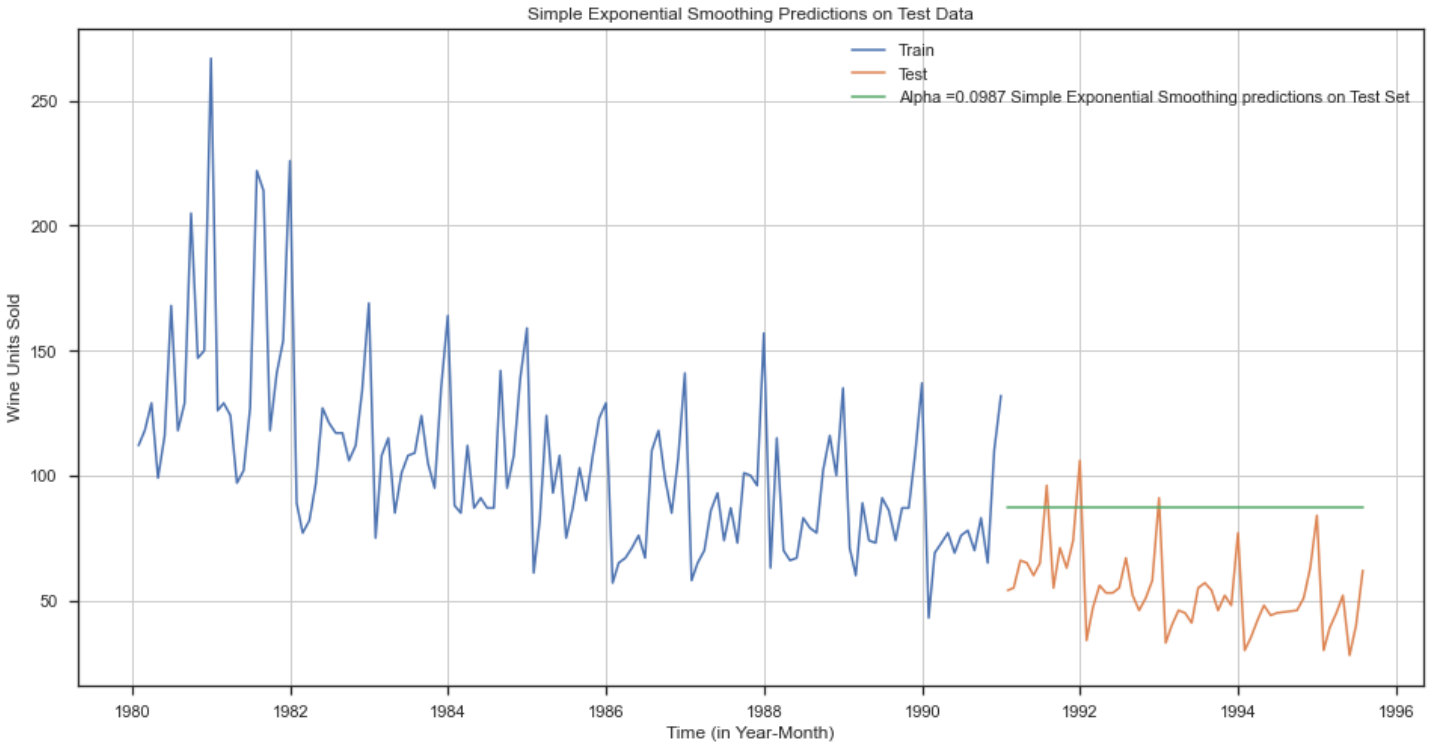


**Fig.36 Rose Wine - SES predictions on Test data**

**The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.**

**The range of alpha value is from 0.1 to 0.95 and the respective RMSE for train and test data are calculated for analyzing the performance metrics.**

| | Alpha Values | Train RMSE | Test RMSE |
|---|---|---|---|
| **0** | 0.10 | 31.815610 | 36.827827 |
| **1** | 0.15 | 31.809845 | 38.721920 |
| **2** | 0.20 | 31.979391 | 41.361671 |
| **3** | 0.25 | 32.211871 | 44.360591 |
| **4** | 0.30 | 32.470164 | 47.504617 |
| **5** | 0.35 | 32.744341 | 50.665469 |
| **6** | 0.40 | 33.035130 | 53.767204 |
| **7** | 0.45 | 33.346578 | 56.766932 |
| **8** | 0.50 | 33.682839 | 59.641585 |
| **9** | 0.55 | 34.047042 | 62.378789 |
| **10** | 0.60 | 34.441171 | 64.971088 |
| **11** | 0.65 | 34.866356 | 67.412703 |
| **12** | 0.70 | 35.323261 | 69.697963 |
| **13** | 0.75 | 35.812435 | 71.820654 |
| **14** | 0.80 | 36.334596 | 73.773794 |
| **15** | 0.85 | 36.890835 | 75.549538 |
| **16** | 0.90 | 37.482782 | 77.139078 |
| **17** | 0.95 | 38.112735 | 78.532498 |

**Fig.37 SES prediction metrics for different alpha values**

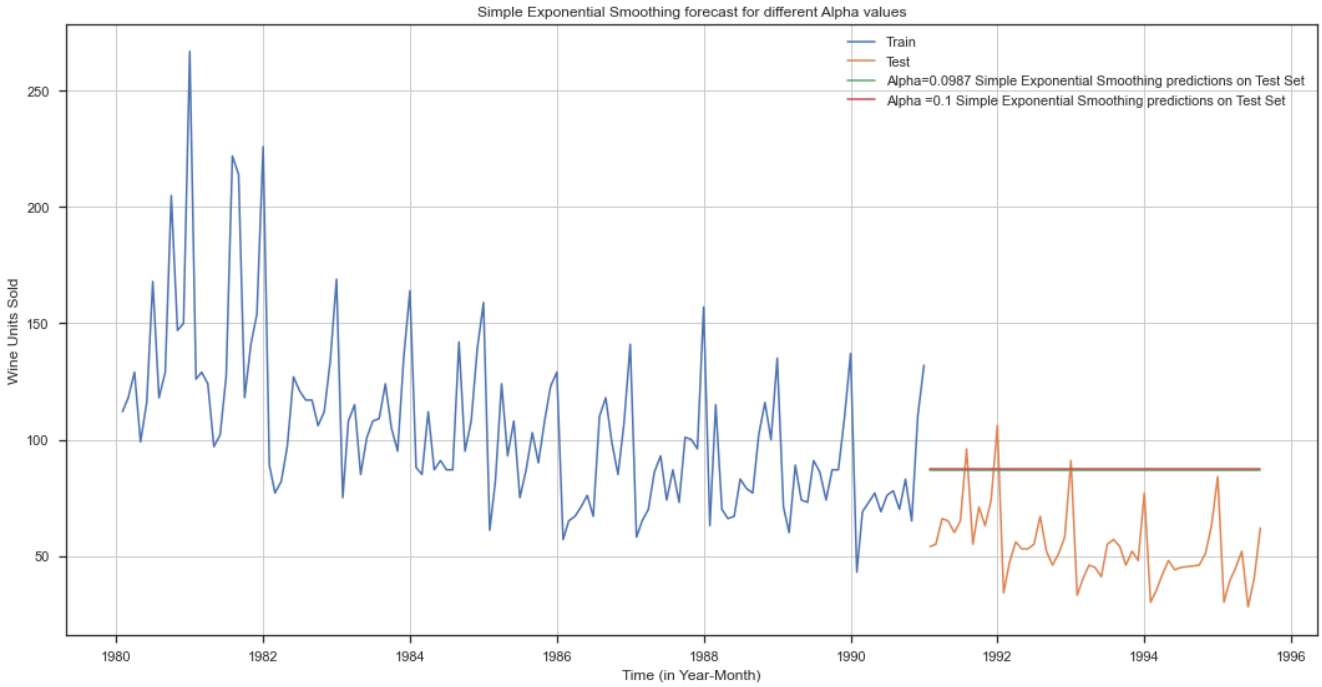


**Fig.38 SES forecast for different Alpha values**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- When there is **neither a trend nor a seasonal component to the time series, simple exponential smoothing is typically used**. It is due to this reason, it unable to capture the characteristics of the time series data.
- The root means squared error **(RMSE)** for the simple exponential smoothing model with **Alpha=0.0987** is **36.796** and for **Alpha=0.1, RMSE is 36.827.**
- **The Simple Exponential Smoothing with alpha=0.0987 is taken as the best model among two as it has the lowest test RMSE.**

## Simple Exponential Smoothing: Model Evaluation

| Model | Test RMSE |
|---|---|
| SES (Alpha = 0.0987) | 36.796036 |
| SES (Alpha = 0.1) | 36.827827 |

## Model 6 – Double Exponential Smoothing (Holt's Model)

**This model is an extension of SES known as Double Exponential model which estimates two smoothing parameters. Applicable when data has Trend but no seasonality. Two separate components are considered: Level and Trend. Level is the local mean. One smoothing parameter α corresponds to the level series. A second smoothing parameter β corresponds to the trend series.**

**Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short-term average value or level and the other for capturing the trend.**

**Intercept or Level equation, $L_t$ is given by: $L_t = \alpha Y_t + (1-\alpha)F_t$**

**Trend equation is given by $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$**

**Here, αα and ββ are the smoothing constants for level and trend, respectively,**

**$0 < \alpha < 1$ and $0 < \beta < 1$.**

**The forecast at time t + 1 is given by**

$$F_{t+1} = L_t + T_t$$

$$F_{t+n} = L_t + nT_t$$

**For the selection criteria, the below Double Exponential Smoothing is built by using optimized parameters.**

```
{'smoothing_level': 1.4901247095597348e-08,
 'smoothing_trend': 7.3896641488640725e-09,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 137.81551313502814,
 'initial_trend': -0.4943777717865305,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

**Fig.39 Rose Wine – Double Exponential Smoothing Model**

| Time_Stamp | |
|---|---|
| 1991-01-31 | 72.063269 |
| 1991-02-28 | 71.568892 |
| 1991-03-31 | 71.074514 |
| 1991-04-30 | 70.580136 |
| 1991-05-31 | 70.085758 |
| 1991-06-30 | 69.591381 |
| 1991-07-31 | 69.097003 |
| 1991-08-31 | 68.602625 |
| 1991-09-30 | 68.108247 |
| 1991-10-31 | 67.613870 |
| 1991-11-30 | 67.119492 |
| 1991-12-31 | 66.625114 |
| 1992-01-31 | 66.130736 |
| 1992-02-29 | 65.636358 |
| 1992-03-31 | 65.141981 |
| 1992-04-30 | 64.647603 |

**Fig.40 Sample of DES predictions**

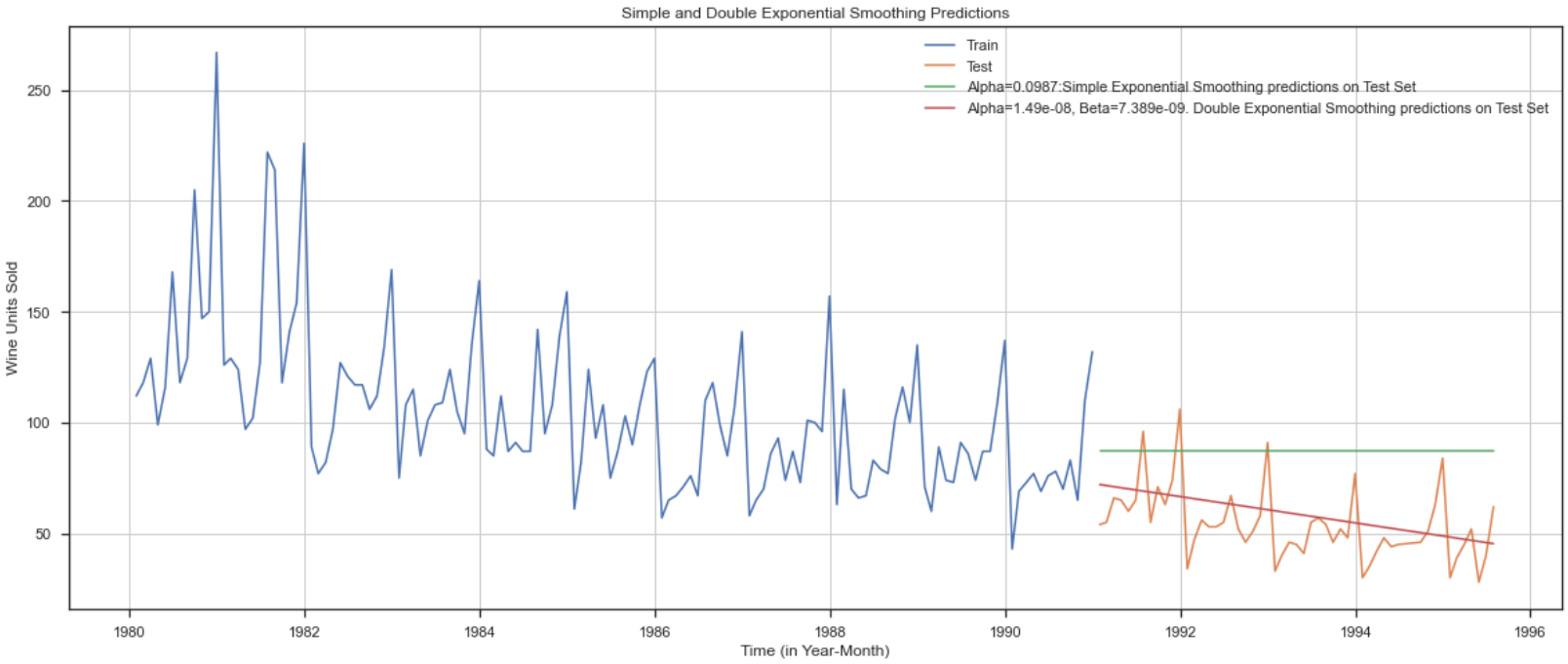


**Fig.41 Rose Wine - DES predictions on Test data**

**The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.**

**The range of alpha value is from 0.05 to 1.0 and the respective RMSE for train and test data are calculated for analyzing the performance metrics.**

| | Alpha | Beta | Train RMSE | Test RMSE |
|---|---|---|---|---|
| 6 | 0.05 | 0.35 | 36.233997 | 16.328994 |
| 5 | 0.05 | 0.30 | 36.616877 | 18.624520 |
| 2 | 0.05 | 0.15 | 39.106563 | 23.716787 |
| 0 | 0.05 | 0.05 | 49.734056 | 31.526698 |
| 7 | 0.05 | 0.40 | 35.783737 | 31.577953 |

**Fig.42 DES prediction metrics for different alpha, beta values**

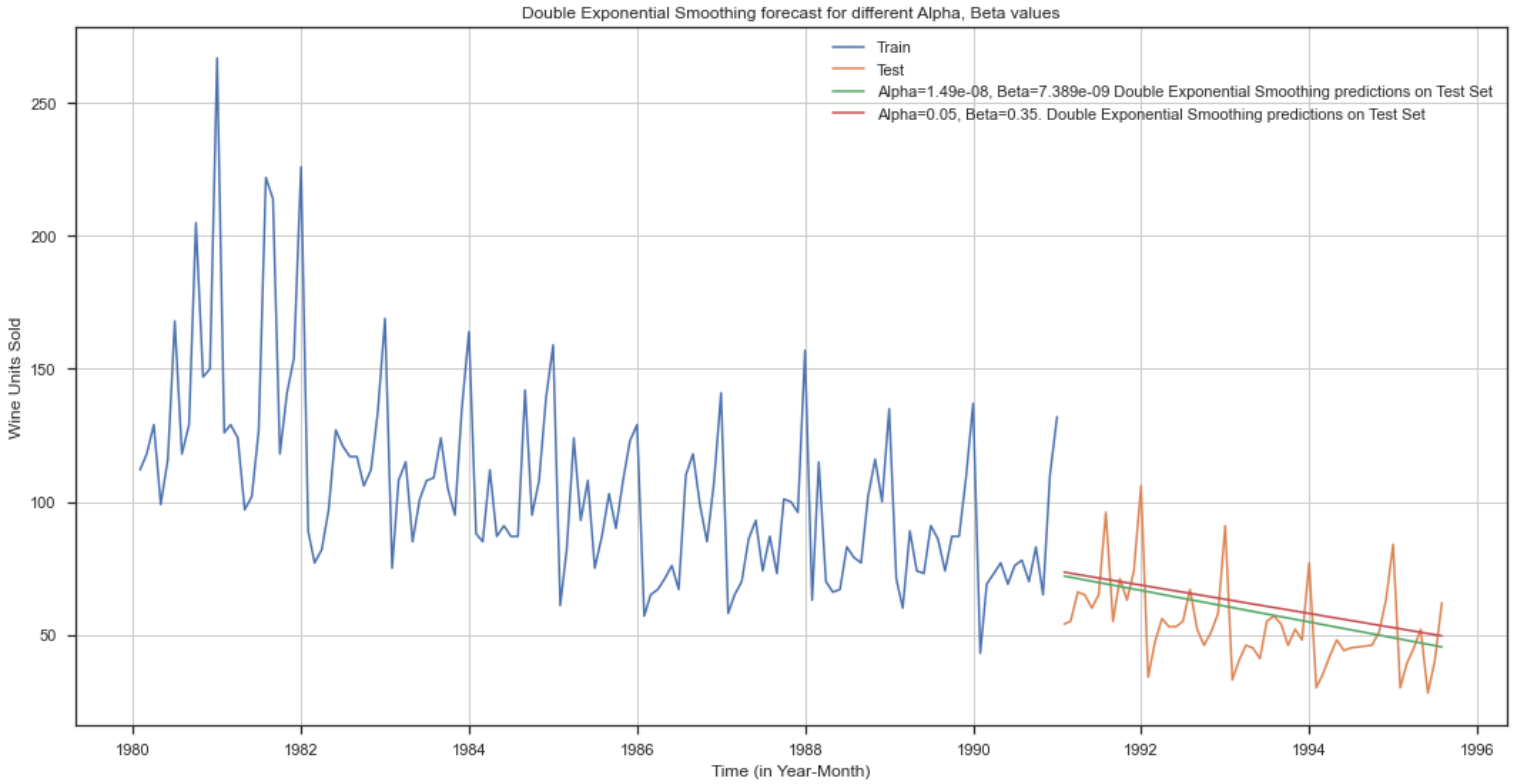


**Fig.43 DES forecast for different Alpha, Beta values**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- When there is simply trend and no seasonality in the time series data, the double exponential smoothing model performs well. It is due to this reason it is only able to capture the trend characteristics of the data and seasonality is not accounted for.
- The root means squared error **(RMSE)** for the double exponential smoothing model with **Alpha=1.49e-08, Beta=7.389e-09** is **15.268** and for **Alpha=0.05, Beta=0.35 (Auto tuned model)**, **RMSE** is **16.328994.**
- **The Double Exponential Smoothing with Alpha=1.49e-08, Beta=7.389e-09 is taken as the best model among two as it has the lowest test RMSE.**
- Additionally, it should be highlighted that compared to the simple exponential smoothing model, the double exponential smoothing model has almost halved the RMSE values.

## Double Exponential Smoothing: Model Evaluation

| Model | Test RMSE |
|---|---|
| DES (Alpha=1.49e-08, Beta=7.389e-09) | 15.268889 |
| DES (Alpha=0.05, Beta=0.35) | 16.328994 |

## Model 7 – Triple Exponential Smoothing (Holt-Winter's Model)

**This model is an extension of DES known as Triple Exponential Smoothing model which estimates three smoothing parameters. Applicable when data has both Trend and seasonality. Three separate components are considered: Level, Trend and Seasonality.**

**One smoothing parameter α corresponds to the level series.**

**A second smoothing parameter β corresponds to the trend series.**

**A third smoothing parameter γ corresponds to the seasonality series**

**where,**

**0 < α <1,**

**0 < β <1,**

**0 < γ <1**

**For the selection criteria, the below Triple Exponential Smoothing is built by using optimized parameters.**

```
{'smoothing_level': 0.06467234615091698,
 'smoothing_trend': 0.05315920636255018,
 'smoothing_seasonal': 0.0,
 'damping_trend': nan,
 'initial_level': 50.880912909225756,
 'initial_trend': -0.31656840824205823,
 'initial_seasons': array([2.21583703, 2.51439498, 2.74693025, 2.40118428, 2.69
936273,
        2.94338111, 3.2353888 , 3.44052906, 3.26420741, 3.19365239,
        3.72269442, 5.13435788]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

**Fig.44 Rose Wine – Triple Exponential Smoothing Model**

| Time_Stamp | Rose_Wine_Sales | auto_predict |
|---|---|---|
| 1991-01-31 | 54.0 | 56.755640 |
| 1991-02-28 | 55.0 | 64.211013 |
| 1991-03-31 | 66.0 | 69.939833 |
| 1991-04-30 | 65.0 | 60.953618 |
| 1991-05-31 | 60.0 | 68.316934 |

**Fig.45 Sample of TES predictions**

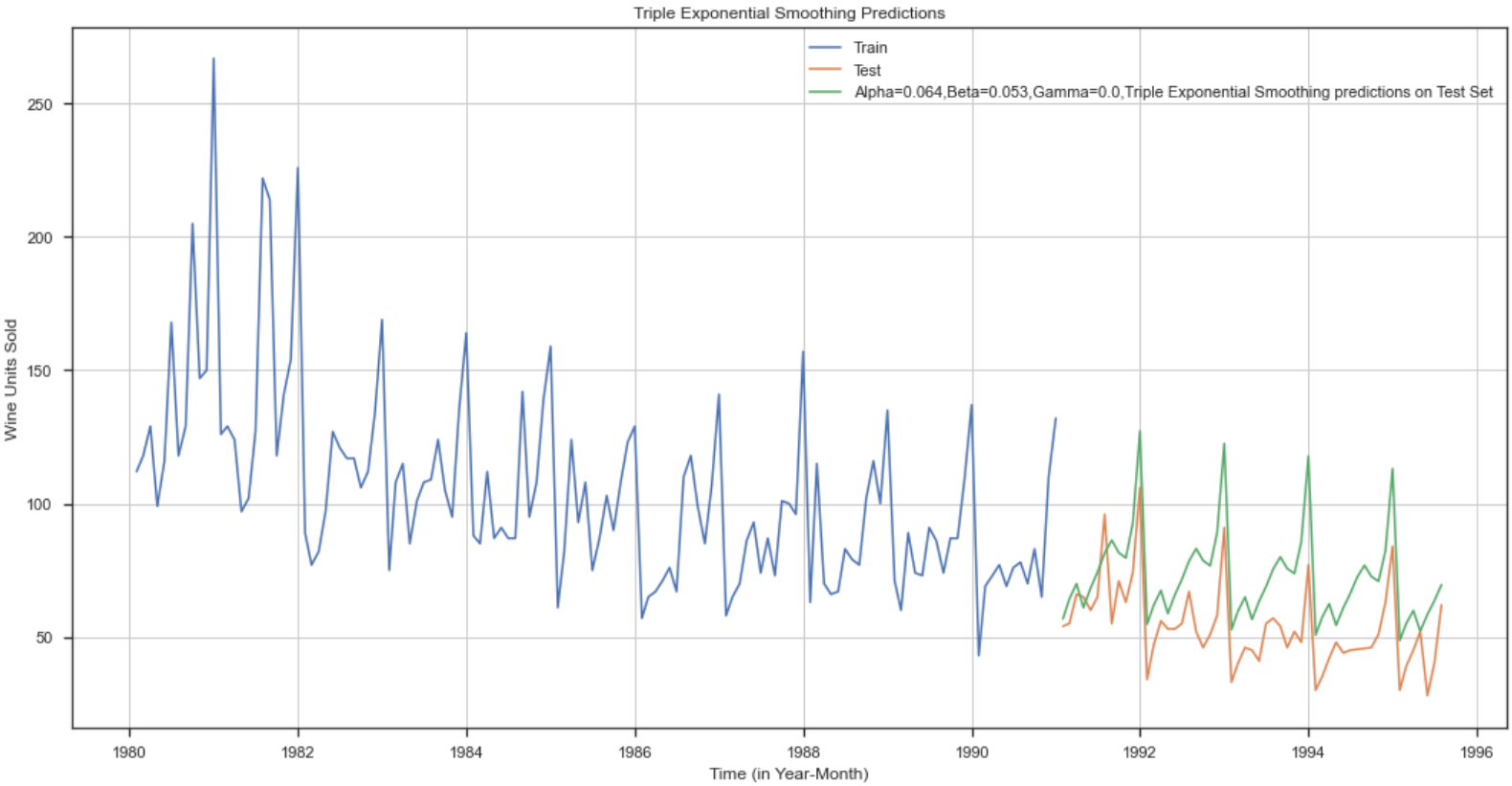


**Fig.46 Rose Wine - TES predictions on Test data**

**The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.**

**The range of alpha value is from 0.1 to 1.0 and the respective RMSE for train and test data are calculated for analyzing the performance metrics.**

| | Alpha | Beta | Gamma | Train RMSE | Test RMSE |
|---|---|---|---|---|---|
| **1008** | 0.2 | 0.85 | 0.15 | 30.302008 | 9.121757 |
| **951** | 0.2 | 0.70 | 0.15 | 29.220767 | 9.447912 |
| **21** | 0.1 | 0.15 | 0.20 | 24.197704 | 9.620332 |
| **39** | 0.1 | 0.20 | 0.15 | 24.798001 | 9.626348 |
| **40** | 0.1 | 0.20 | 0.20 | 24.365597 | 9.640614 |

**Fig.47 TES prediction metrics for different alpha, beta and gamma values**

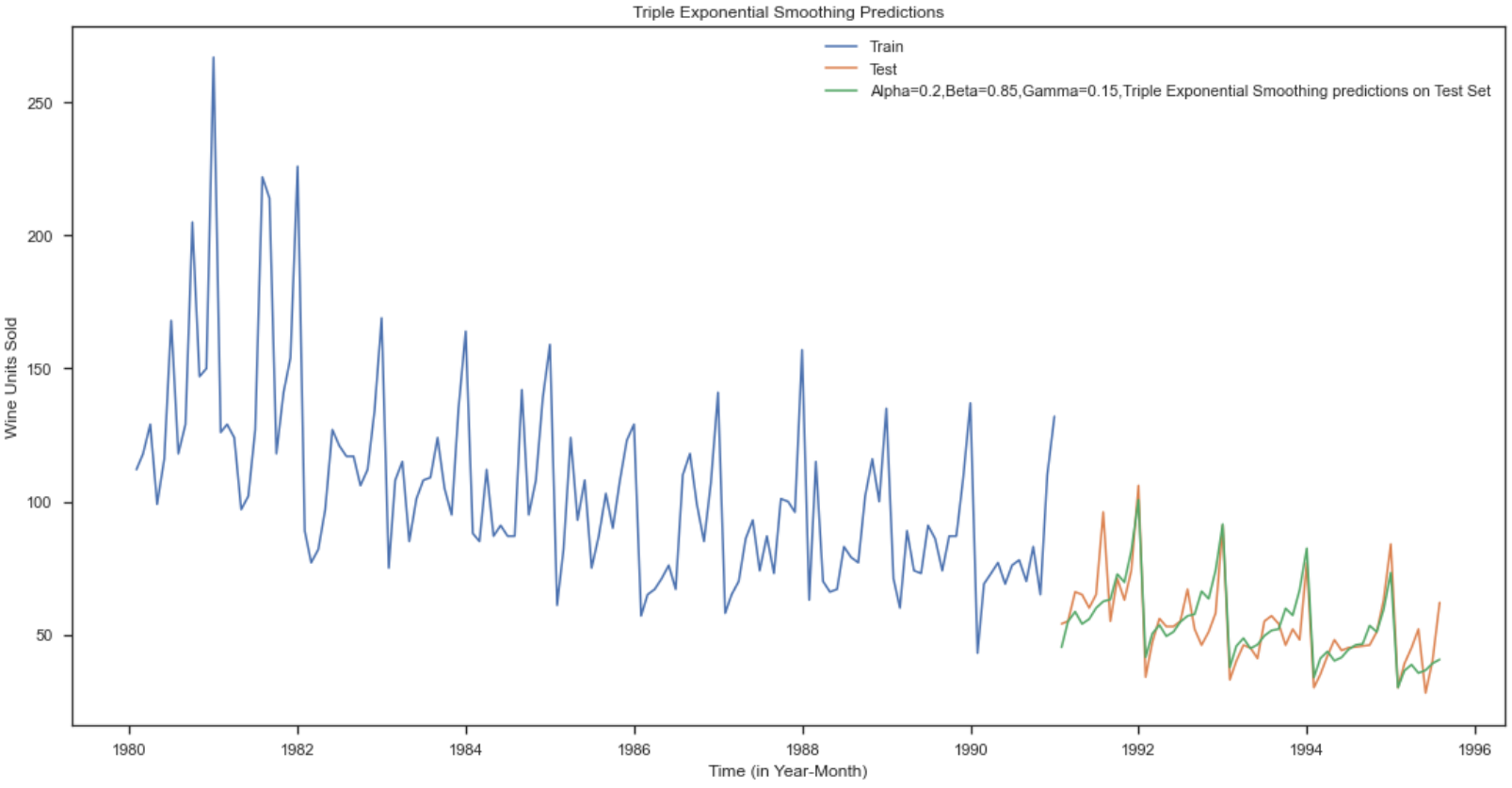


**Fig.48 TES forecast for automated model parameters**

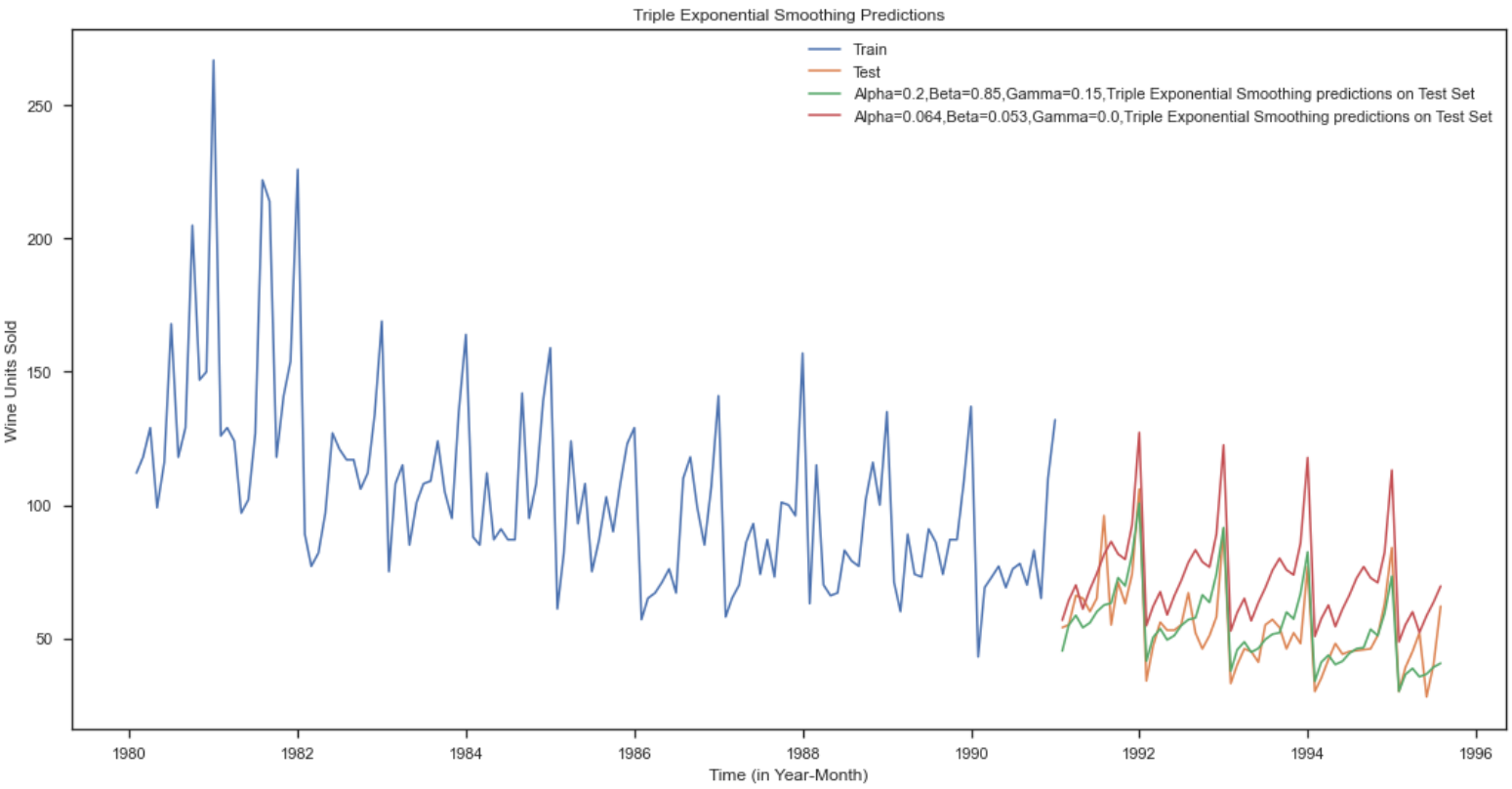


**Fig.49 TES forecast for different model parameters**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- When there is both trend and seasonality in the time series data, the triple exponential model works well. It is due to this reason it able to capture both the trend and seasonal characteristics and nearly match the actual test data plot.
- The root means squared error **(RMSE)** for the double exponential smoothing model with **Alpha=0.064, Beta=0.053, Gamma=0.0** is **21.154** and for **Alpha=0.2, Beta=0.85, Gamma=0.15 (Auto tuned model)**, **RMSE** is **9.121.**
- **The Triple Exponential Smoothing with Alpha=0.2, Beta=0.85, Gamma=0.15 is taken as the best model among two as it has the lowest test RMSE.**
- Additionally, it should be highlighted that compared to the double exponential smoothing model, the **triple exponential smoothing model has almost reduced the RMSE value by 40%.**

## Triple Exponential Smoothing: Model Evaluation

| Model | Test RMSE |
|---|---|
| TES (Alpha=0.064, Beta=0.053, Gamma=0.0) | 21.154527 |
| TES (Alpha=0.2, Beta=0.85, Gamma=0.15) | 9.121757 |

Let's compare the RMSE values of the models we have constructed so far and visualize the plot of the best exponential smoothing models thus built.

| | Test RMSE |
|---|---|
| Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing | 9.121757 |
| 2 point TMA | 11.529278 |
| 4 point TMA | 14.451376 |
| 6 point TMA | 14.566262 |
| 9 point TMA | 14.727596 |
| Linear Regression | 15.268887 |
| Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing | 15.268889 |
| Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing | 21.154527 |
| Alpha=0.0987,SimpleExponentialSmoothing | 36.796036 |
| Simple Average | 53.460367 |
| Naive Model | 79.718576 |

**Fig.50 Comparison of Test RMSE values of different exponential smoothing models**

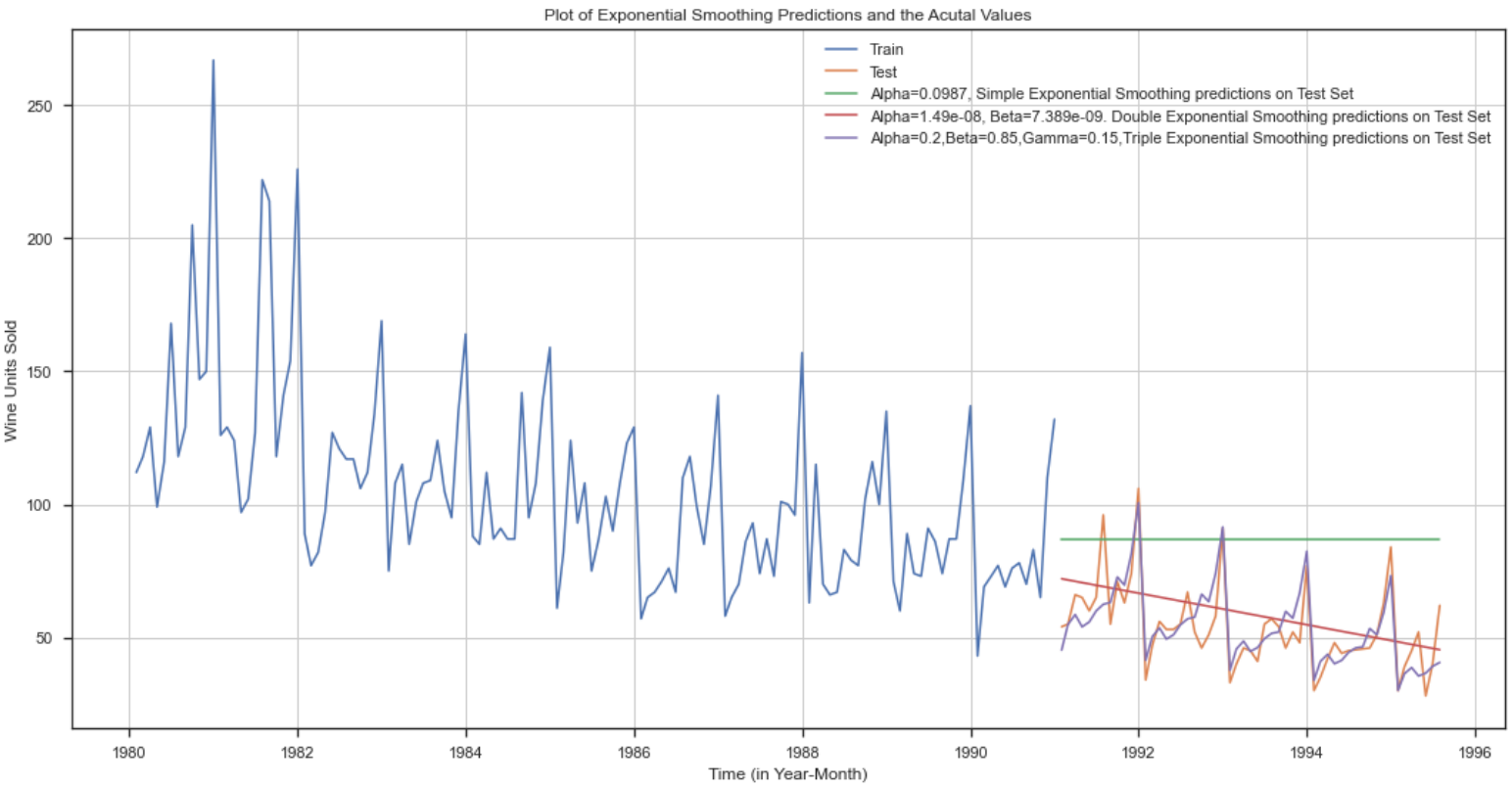


**Fig.51 Comparison of different models on test data (SES, DES and TES)**

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- **Simple exponential smoothing** is frequently employed when the time series doesn't include a trend or a seasonal component. This is the reason why it is unable to capture the time series data's features.
- The **double exponential smoothing** model works effectively when the time series data just contains trend and no seasonality. This explains why seasonality is not taken into consideration and just the trend features of the data are captured.
- The **triple exponential model** performs effectively when the time series data exhibit both trend and seasonality. This is the reason why it is essentially identical to the test data plot and is able to capture both the trend and seasonal aspects.
- The **Triple exponential model is the best model we have built so far as it has the lowest RMSE value**.

## 5) Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

### Checking for Stationarity of Entire Data

The **Augmented Dickey-Fuller test** is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

**Framing the hypothesis:**

$H_0$: The Time Series has a unit root and is thus non-stationary.

$H_1$: The Time Series does not have a unit root and is thus stationary.

**The series have to be stationary for building ARIMA/SARIMA models and thus we would want the p-value of this test to be less than the α value.**

```
Results of Dicky-Fuller Test
DF test statistic is -2.240
DF test p-value is 0.46713505298058916
Number of lags used 13
```

**Fig.52 Rose Wine – ADF summary**

**Inference:**

**We see that at 5% significant level the Time Series is non-stationary as p-value is 0.467 which is more than alpha value (0.05), therefore we fail to reject the null hypothesis. Let us take one level of differencing to see whether the series becomes stationary.**

```
Results of Dicky-Fuller Test with differencing
DF test statistic is -8.162
DF test p-value is 3.015892676563128e-11
Number of lags used 12
```

**Fig.53 Rose Wine – ADF summary with differencing**

**Inference:**

**We see that at 5% significant level the Time Series becomes stationary as p-value is 3.015e-11 which is less than alpha value (0.05), therefore we reject the null hypothesis. We can see that the provided time series becomes stationary with differencing.**

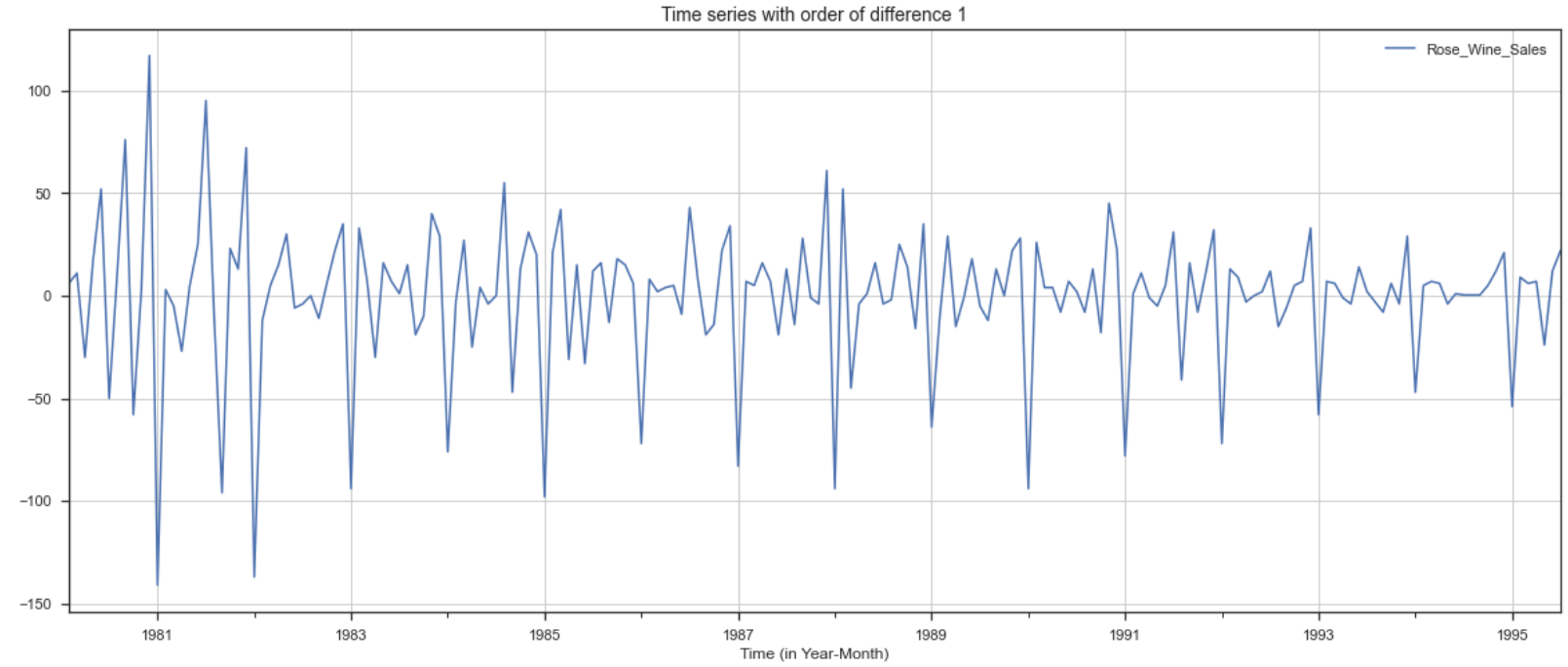


**Fig.54 Time Series Plot of Entire data – With differencing**

## Checking for Stationarity of Training Data

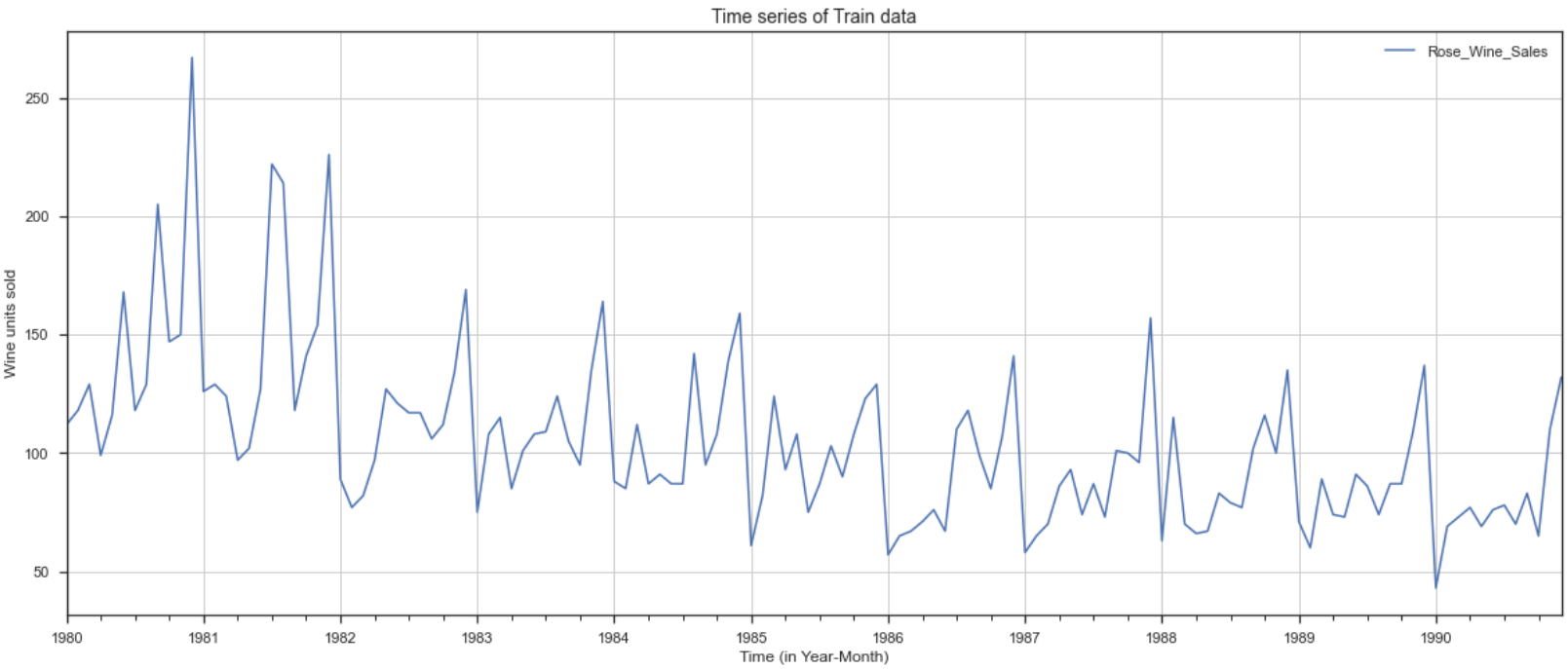


**Fig.54 Time Series Plot of Train data**

```
Results of Dicky-Fuller Test on Train data
DF test statistic is -1.686
DF test p-value is 0.7569093051047106
Number of lags used 13
```

**Fig.55 Rose Wine – ADF summary on train data**

**Inference:**

**We see that at 5% significant level the Time Series of training data is non-stationary as p-value is 0.756 which is more than alpha value (0.05), therefore we fail to reject the null hypothesis. Let us take one level of differencing to see whether the series becomes stationary.**

```
Results of Dicky-Fuller Test on Train data with differencing
DF test statistic is -6.804
DF test p-value is 3.894831356781761e-08
Number of lags used 12
```

**Fig.56 Rose Wine – ADF summary on train data with differencing**

**Inference:**

**We see that at 5% significant level the Time Series of training data is non-stationary as p-value is 3.894e-08 which is less than alpha value (0.05), therefore we reject the null hypothesis. We can see that the provided training time series becomes stationary with differencing.**

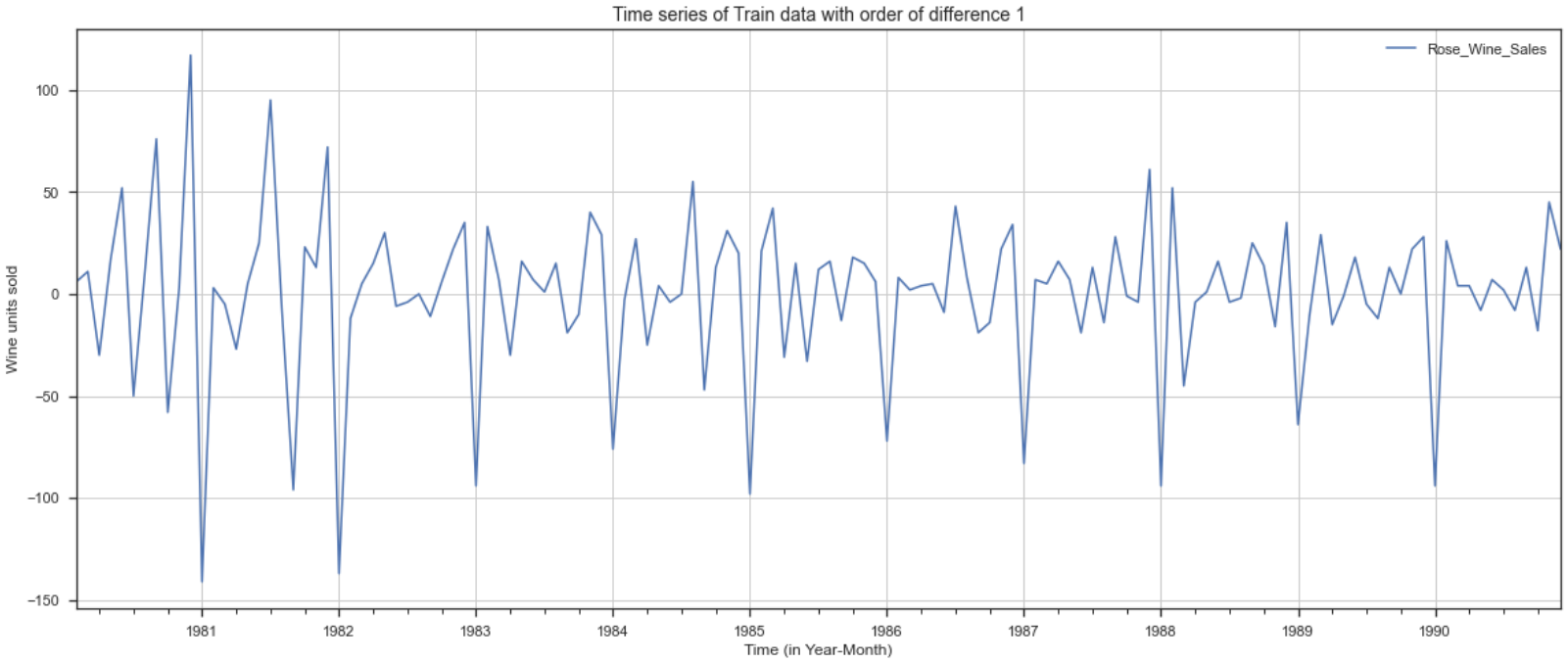


**Fig.57 Time Series Plot of Training data with differencing**

**Observation:**

- As per the Augmented Dicky-Fuller test, we observed that the time series data by itself is not stationary, however, it becomes stationary when differencing is done.
- The same thing is also observed with Training data. Therefore, for training the models, it can be built with order of difference d=1.

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

### Model 8 – Auto-Regressive Integrated Moving Average (ARIMA)

**Auto-regression means regression of a variable on itself. One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process. When the time series data is not stationary, then we have to convert the non-stationary time-series data to stationary time-series before applying AR.**

**ARIMA models may be used to represent any "non-seasonal" time series that has patterns and isn't just random noise.**

**An ARIMA model is characterized by 3 terms: p, d, q**

**where,**

**p is the order of the Auto Regressive (AR) term**

**q is the order of the Moving Average (MA) term**

**d is the number of differencing required to make the time series stationary**

**For the selection criteria of p,d,q the below ARIMA model is built by using automated model parameters with lowest Akaike Information Criteria.**

```
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)
```

**Fig.58 Parameter Combinations for ARIMA model**

```
ARIMA(0, 1, 0) - AIC:1333.1546729124348
ARIMA(0, 1, 1) - AIC:1282.3098319748299
ARIMA(0, 1, 2) - AIC:1279.6715288535765
ARIMA(0, 1, 3) - AIC:1280.5453761734668
ARIMA(0, 1, 4) - AIC:1281.6766982143947
ARIMA(1, 1, 0) - AIC:1317.3503105381546
ARIMA(1, 1, 1) - AIC:1280.5742295380073
ARIMA(1, 1, 2) - AIC:1279.8707234231897
ARIMA(1, 1, 3) - AIC:1281.8707223309998
ARIMA(1, 1, 4) - AIC:1279.6052633451109
ARIMA(2, 1, 0) - AIC:1298.6110341604983
ARIMA(2, 1, 1) - AIC:1281.5078621868424
ARIMA(2, 1, 2) - AIC:1281.8707222264402
ARIMA(2, 1, 3) - AIC:1274.6953561209548
ARIMA(2, 1, 4) - AIC:1278.7699014386199
ARIMA(3, 1, 0) - AIC:1297.4810917271725
ARIMA(3, 1, 1) - AIC:1282.4192776271946
ARIMA(3, 1, 2) - AIC:1283.720740597711
ARIMA(3, 1, 3) - AIC:1278.6619652725685
ARIMA(3, 1, 4) - AIC:1287.7190768737705
ARIMA(4, 1, 0) - AIC:1296.3266569004702
ARIMA(4, 1, 1) - AIC:1283.79317151231
ARIMA(4, 1, 2) - AIC:1285.718248563479
ARIMA(4, 1, 3) - AIC:1278.4514021021457
ARIMA(4, 1, 4) - AIC:1282.3372229344063
```

**Fig.59 AIC values for different parameter combinations**

| | param | AIC |
|---|---|---|
| **13** | (2, 1, 3) | 1274.695356 |
| **23** | (4, 1, 3) | 1278.451402 |
| **18** | (3, 1, 3) | 1278.661965 |
| **14** | (2, 1, 4) | 1278.769901 |
| **9** | (1, 1, 4) | 1279.605263 |

**Fig.60 Sorted AIC values for different parameter combinations**

**We can see that among all the possible given combinations, the AIC is lowest for the combination (2,1,3). Hence, the model is built with these parameters to determine the RMSE value of test data.**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Rose_Wine_Sales | No. Observations: | 132 |
| Model: | ARIMA(2, 1, 3) | Log Likelihood | -631.348 |
| Date: | Sat, 22 Oct 2022 | AIC | 1274.695 |
| Time: | 10:07:25 | BIC | 1291.947 |
| Sample: | 01-31-1980 | HQIC | 1281.705 |
| | - 12-31-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | -1.6781 | 0.084 | -19.992 | 0.000 | -1.843 | -1.514 |
| ar.L2 | -0.7289 | 0.084 | -8.684 | 0.000 | -0.893 | -0.564 |
| ma.L1 | 1.0446 | 0.628 | 1.665 | 0.096 | -0.185 | 2.275 |
| ma.L2 | -0.7720 | 0.133 | -5.824 | 0.000 | -1.032 | -0.512 |
| ma.L3 | -0.9046 | 0.569 | -1.590 | 0.112 | -2.020 | 0.210 |
| sigma2 | 860.6996 | 528.714 | 1.628 | 0.104 | -175.560 | 1896.959 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.02 | Jarque-Bera (JB): | 24.48 |
| Prob(Q): | 0.88 | Prob(JB): | 0.00 |
| Heteroskedasticity (H): | 0.40 | Skew: | 0.71 |
| Prob(H) (two-sided): | 0.00 | Kurtosis: | 4.57 |

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

**Fig.61 Rose Wine – Automated ARIMA model**

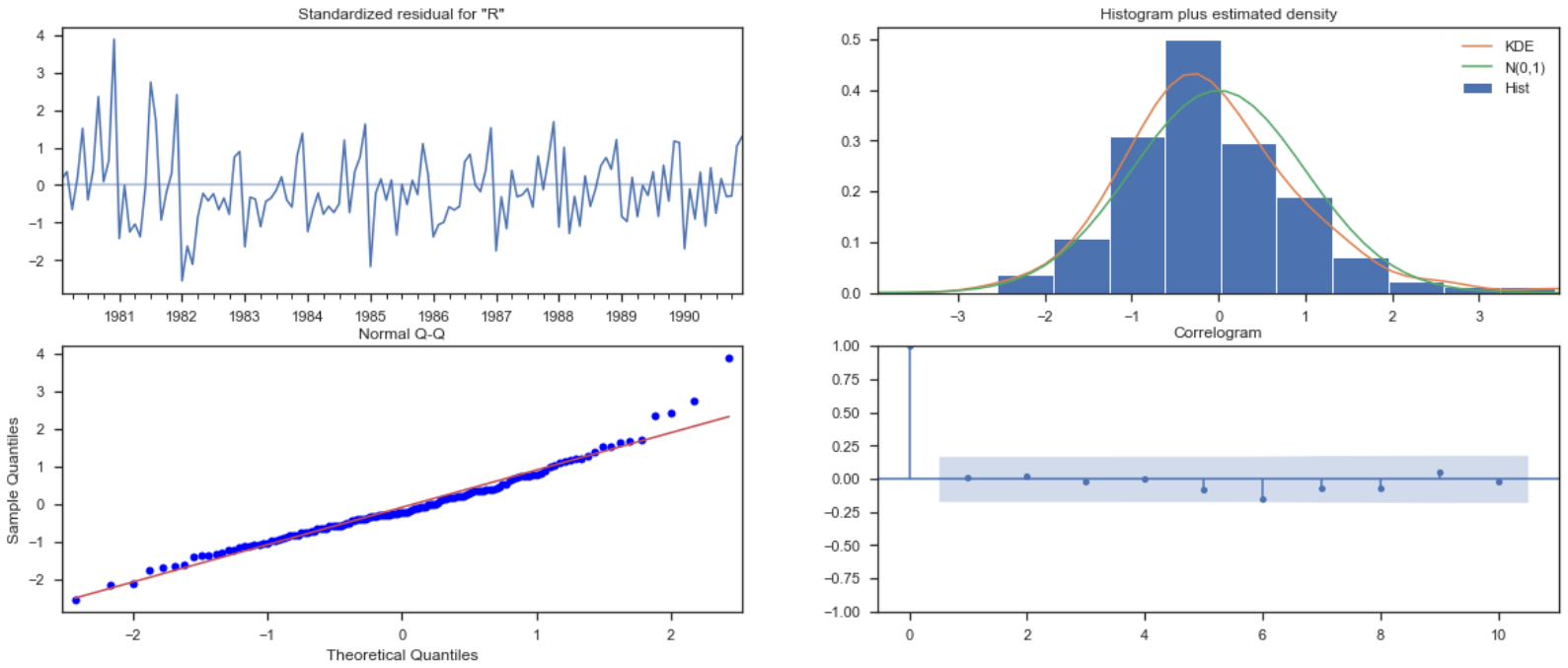


**Fig.62 Automated ARIMA – Diagnostics plot**

**Observation:**

- The optimal parameters are decided based on the lowest Akaike Information Criteria (AIC) values. **The AIC is lowest for the combination (2,1,3) as we see from the above results.**
- From the **Standardized residual plot** above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.
- The **histogram plus estimated density plot** suggests a slightly uniform distribution with mean zero and slightly skewed to the right.
- In **Normal Q-Q plot**, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.
- The **correlogram plot** of residuals shows that the residuals are not auto correlated.

| | |
|---|---|
| 1991-01-31 | 85.605078 |
| 1991-02-28 | 90.532552 |
| 1991-03-31 | 81.968167 |
| 1991-04-30 | 92.748215 |
| 1991-05-31 | 80.900983 |
| 1991-06-30 | 92.924087 |
| 1991-07-31 | 81.383696 |
| 1991-08-31 | 91.985881 |
| 1991-09-30 | 82.606225 |
| 1991-10-31 | 90.618235 |
| 1991-11-30 | 84.010151 |
| 1991-12-31 | 89.259200 |
| 1992-01-31 | 85.267413 |
| 1992-02-29 | 88.139999 |
| 1992-03-31 | 86.229119 |
| 1992-04-30 | 87.341950 |

**Fig.63 Sample of Automated ARIMA (2,1,3) predictions**

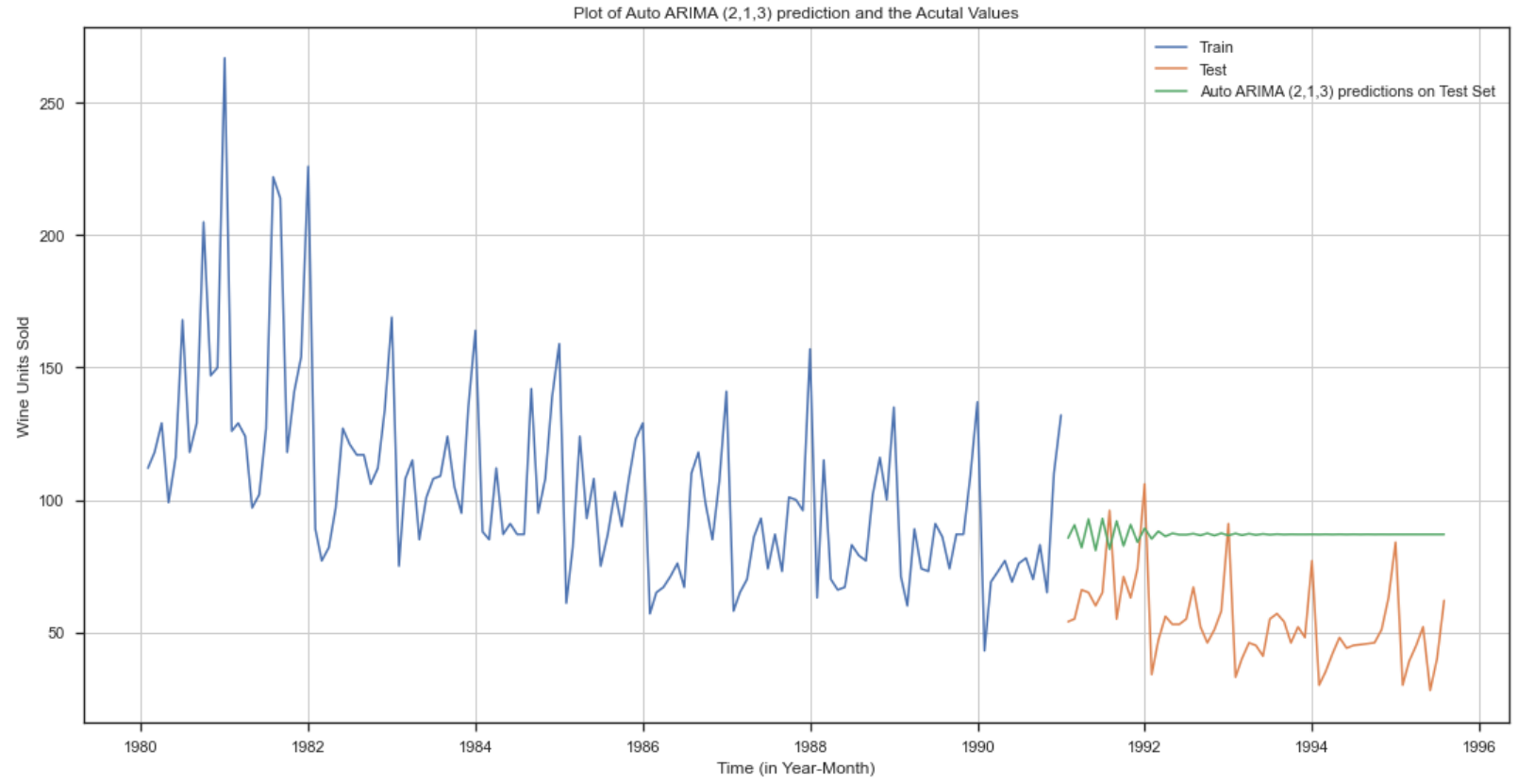


**Fig.64 Plot of Automated ARIMA (2,1,3) predictions on Test data**

## Automated ARIMA: Model Evaluation

**For evaluating the model's performance metrics, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)**

| Model | Test RMSE | Test MAPE |
|---|---|---|
| **ARIMA (p=2, d=1, q=3)** | **36.813** | **75.839** |

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- ARIMA models performs well on non-seasonal time series. It is due to this reason it is unable to capture the entire characteristics of the test data.
- The root means squared error **(RMSE)** of test data for the ARIMA model with **(p=2, d=1, q=3)** is **36.813.**
- Not surprisingly, the RMSE of the aforementioned ARIMA model is greater than the majority of previously constructed models.

## Model 9 – Seasonal Auto-Regressive Integrated Moving Average (SARIMA)

**SARIMA models or also known as Seasonal ARIMA is an extension of ARIMA for a time series data with defined seasonality. SARIMA models use seasonal differencing which is similar to regular differencing.**

**A SARIMA model is characterized by 7 terms: p, d, q, P, Q, D and F**

**where,**

**p is the order of the Auto Regressive (AR) term**

**q is the order of the Moving Average (MA) term**

**d is the number of differencing required to make the time series stationary**

**P is the order of the Seasonal Auto Regressive (AR) term**

**Q is the order of the Seasonal Moving Average (MA) term**

**D is the number of seasonal differencing required to make the time series stationary**

**F is the seasonal frequency of the time series**

**We must examine the PACF and ACF plots, respectively, at delays that are the multiple of "F" in order to determine the "P" and "Q" values, and determine where these cut-off values are (for appropriate confidence interval bands).**

**By examining the lowest AIC values, we can also estimate "p," "q," "P," and "Q" for the SARIMA models.**

**By examining the ACF plots, one may calculate the seasonal parameter 'F'. The existence of seasonality should be shown by a spike in the ACF plot at multiples of "F."**

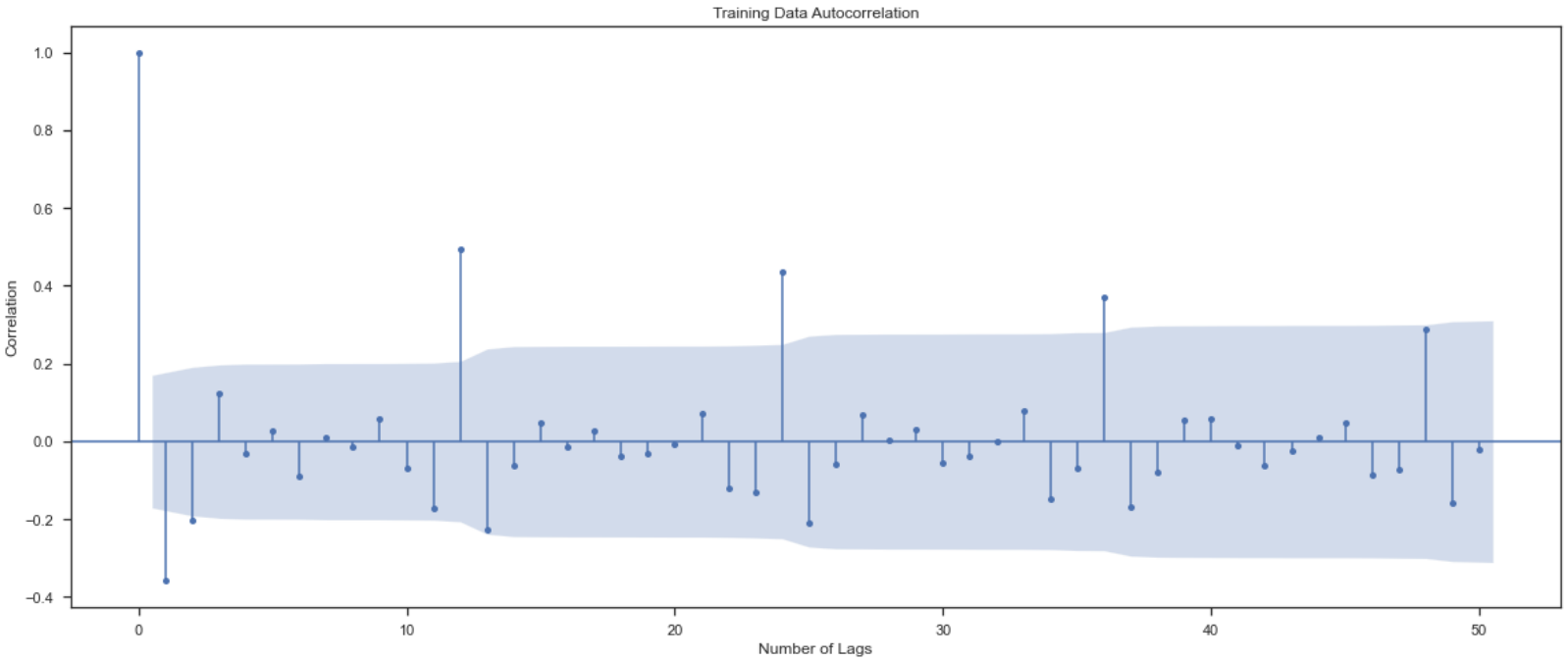


**Fig.65 ACF plot of Train data**

**From the above ACF plot we can observe that at every 12$^{th}$ lag is significant indicating the presence of seasonality. Hence for our model building we will consider the term F=12.**

**For the selection criteria of p, d, q, P, D, Q & F the below SARIMA model is built by using automated model parameters with lowest Akaike Information Criteria.**

```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (0, 1, 3)(0, 0, 3, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (1, 1, 3)(1, 0, 3, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
Model: (2, 1, 3)(2, 0, 3, 12)
Model: (3, 1, 0)(3, 0, 0, 12)
Model: (3, 1, 1)(3, 0, 1, 12)
Model: (3, 1, 2)(3, 0, 2, 12)
Model: (3, 1, 3)(3, 0, 3, 12)
```

**Fig.66 Parameter Combinations for SARIMA model**

```
SARIMA(0, 1, 0)x(0, 0, 0, 12) - AIC:1323.9657875279158
SARIMA(0, 1, 0)x(0, 0, 1, 12) - AIC:1145.4230827207386
SARIMA(0, 1, 0)x(0, 0, 2, 12) - AIC:976.4375296380903
SARIMA(0, 1, 0)x(0, 0, 3, 12) - AIC:3994.7201928422214
SARIMA(0, 1, 0)x(1, 0, 0, 12) - AIC:1139.921738995602
SARIMA(0, 1, 0)x(1, 0, 1, 12) - AIC:1116.0207869385767
SARIMA(0, 1, 0)x(1, 0, 2, 12) - AIC:969.6913635754913
SARIMA(0, 1, 0)x(1, 0, 3, 12) - AIC:3801.0056501147847
SARIMA(0, 1, 0)x(2, 0, 0, 12) - AIC:960.8812220353041
SARIMA(0, 1, 0)x(2, 0, 1, 12) - AIC:962.8794540697546
SARIMA(0, 1, 0)x(2, 0, 2, 12) - AIC:955.5735408945715
SARIMA(0, 1, 0)x(2, 0, 3, 12) - AIC:5644.471437825422
SARIMA(0, 1, 0)x(3, 0, 0, 12) - AIC:850.753540393109
SARIMA(0, 1, 0)x(3, 0, 1, 12) - AIC:851.7482702714087
SARIMA(0, 1, 0)x(3, 0, 2, 12) - AIC:850.5304136127778
SARIMA(0, 1, 0)x(3, 0, 3, 12) - AIC:3843.9094675774722
SARIMA(0, 1, 1)x(0, 0, 0, 12) - AIC:1263.5369097383968
SARIMA(0, 1, 1)x(0, 0, 1, 12) - AIC:1098.5554825918339
SARIMA(0, 1, 1)x(0, 0, 2, 12) - AIC:923.6314049383913
SARIMA(0, 1, 1)x(0, 0, 3, 12) - AIC:3812.2263068724546
SARIMA(0, 1, 1)x(1, 0, 0, 12) - AIC:1095.7936324918096
SARIMA(0, 1, 1)x(1, 0, 1, 12) - AIC:1054.7434330947592
SARIMA(0, 1, 1)x(1, 0, 2, 12) - AIC:918.8573483300271
SARIMA(0, 1, 1)x(1, 0, 3, 12) - AIC:3442.2711593212944
```

**Fig.67 AIC values for different parameter combinations**

| | param | seasonal | AIC |
|---|---|---|---|
| **222** | (3, 1, 1) | (3, 0, 2, 12) | 774.400286 |
| **238** | (3, 1, 2) | (3, 0, 2, 12) | 774.880935 |
| **220** | (3, 1, 1) | (3, 0, 0, 12) | 775.426699 |
| **221** | (3, 1, 1) | (3, 0, 1, 12) | 775.495330 |
| **252** | (3, 1, 3) | (3, 0, 0, 12) | 775.561019 |

**Fig.68 Sorted AIC values for different parameter combinations**

**We can see that among all the possible given combinations, the AIC is lowest for the combination (3,1,1) (3,0,2,12). Hence, the model is built with these parameters to determine the RMSE value of test data.**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Rose_Wine_Sales | No. Observations: | 132 |
| Model: | SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12) | Log Likelihood | -377.200 |
| Date: | Sat, 22 Oct 2022 | AIC | 774.400 |
| Time: | 10:10:53 | BIC | 799.618 |
| Sample: | 01-31-1980 | HQIC | 784.578 |
| | - 12-31-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | 0.0464 | 0.126 | 0.367 | 0.714 | -0.202 | 0.294 |
| ar.L2 | -0.0060 | 0.120 | -0.050 | 0.960 | -0.241 | 0.229 |
| ar.L3 | -0.1808 | 0.098 | -1.838 | 0.066 | -0.374 | 0.012 |
| ma.L1 | -0.9370 | 0.067 | -13.905 | 0.000 | -1.069 | -0.805 |
| ar.S.L12 | 0.7639 | 0.165 | 4.640 | 0.000 | 0.441 | 1.087 |
| ar.S.L24 | 0.0840 | 0.159 | 0.527 | 0.598 | -0.229 | 0.397 |
| ar.S.L36 | 0.0727 | 0.095 | 0.764 | 0.445 | -0.114 | 0.259 |
| ma.S.L12 | -0.4969 | 0.250 | -1.988 | 0.047 | -0.987 | -0.007 |
| ma.S.L24 | -0.2191 | 0.210 | -1.044 | 0.296 | -0.630 | 0.192 |
| sigma2 | 192.1518 | 39.627 | 4.849 | 0.000 | 114.484 | 269.819 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.30 | Jarque-Bera (JB): | 1.64 |
| Prob(Q): | 0.58 | Prob(JB): | 0.44 |
| Heteroskedasticity (H): | 1.11 | Skew: | 0.33 |
| Prob(H) (two-sided): | 0.77 | Kurtosis: | 3.03 |

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

**Fig.69 Rose Wine – Automated SARIMA model**

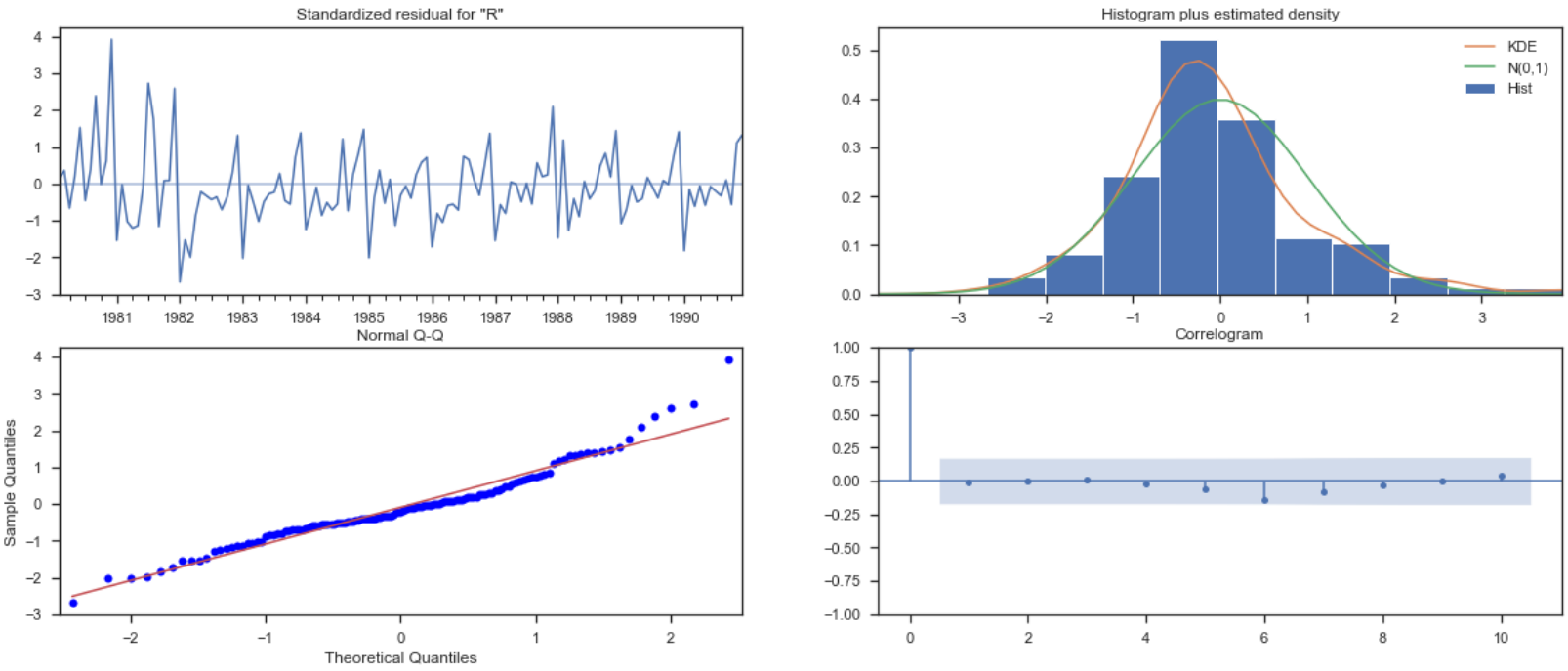


**Fig.70 Automated SARIMA – Diagnostics plot**

**Observation:**

- The optimal parameters are decided based on the lowest Akaike Information Criteria (AIC) values. **The AIC is lowest for the combination (3,1,1) (3,0,2,12) as we see from the above results.**
- From the **Standardized residual plot** above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.
- The **histogram plus estimated density plot** suggests a slightly uniform distribution with mean zero and slightly skewed to the right.
- In **Normal Q-Q plot**, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.
- The **correlogram plot** of residuals shows that the residuals are not auto correlated.

| | |
|---|---|
| 1991-01-31 | 55.235777 |
| 1991-02-28 | 68.122643 |
| 1991-03-31 | 67.908788 |
| 1991-04-30 | 66.786249 |
| 1991-05-31 | 69.760445 |
| 1991-06-30 | 70.329003 |
| 1991-07-31 | 75.359549 |
| 1991-08-31 | 76.492110 |
| 1991-09-30 | 78.971378 |
| 1991-10-31 | 76.538670 |
| 1991-11-30 | 93.249070 |
| 1991-12-31 | 116.283229 |
| 1992-01-31 | 55.202490 |
| 1992-02-29 | 64.444090 |
| 1992-03-31 | 68.547777 |
| 1992-04-30 | 63.872362 |
| 1992-05-31 | 67.700162 |
| 1992-06-30 | 68.443607 |
| 1992-07-31 | 72.972114 |

**Fig.71 Sample of Automated SARIMA (3,1,1) (3,0,2,12) predictions**

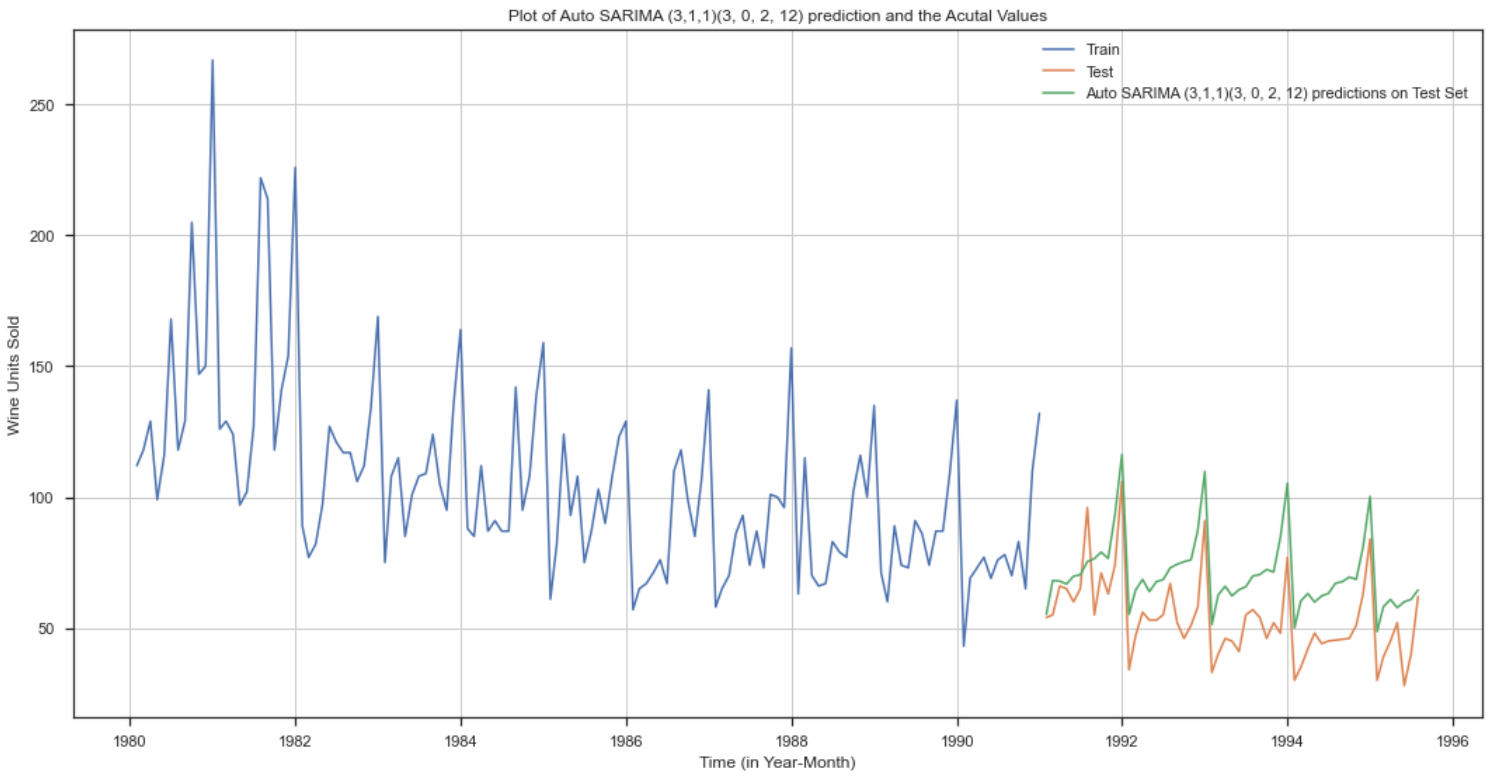


**Fig.72 Plot of Automated SARIMA (3,1,1) (3,0,2,12) predictions on Test data**

## Automated SARIMA: Model Evaluation

**For evaluating the model performance, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)**

| Model | Test RMSE | Test MAPE |
|---|---|---|
| **SARIMA (p=3, d=1, q=1) (P=3, D=0, Q=2, F=12)** | **18.881** | **36.375** |

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- SARIMA model performs well on seasonal time series. It is due to this reason it is able to capture the entire characteristics of the test data.
- The root means squared error **(RMSE)** of test data for the SARIMA model with **(p=3, d=1, q=1) (P=3, D=0, Q=2, F=12)** is **18.881.**
- Additionally, it should be highlighted that compared to the ARIMA model, the SARIMA model has almost halved the RMSE value.

## 7) Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

### Model 10 – Auto-Regressive Integrated Moving Average (ARIMA) - Manual

**An ARIMA model is characterized by 3 terms: p, d, q**

**where,**

**p is the order of the Auto Regressive (AR) term**

**q is the order of the Moving Average (MA) term**

**d is the number of differencing required to make the time series stationary**

Indicating which previous series values are most beneficial in forecasting future values, autocorrelation and partial autocorrelation are measures of relationship between present and past series values. You may identify the sequence of processes in an ARIMA model using this information.

**The parameters p & q can be determined by looking at the PACF & ACF plots respectively.**

**Autocorrelation function (ACF) -** At lag *k*, this is the correlation between series values that are *k* intervals apart.

**Partial autocorrelation function (PACF) -** At lag *k*, this is the correlation between series values that are *k* intervals apart, accounting for the values of the intervals between.

**In an ACF & PACF plots, each bar represents the size and direction of the connection. Bars that cross the red line are statistically significant.**

## ACF Plot – Training Data

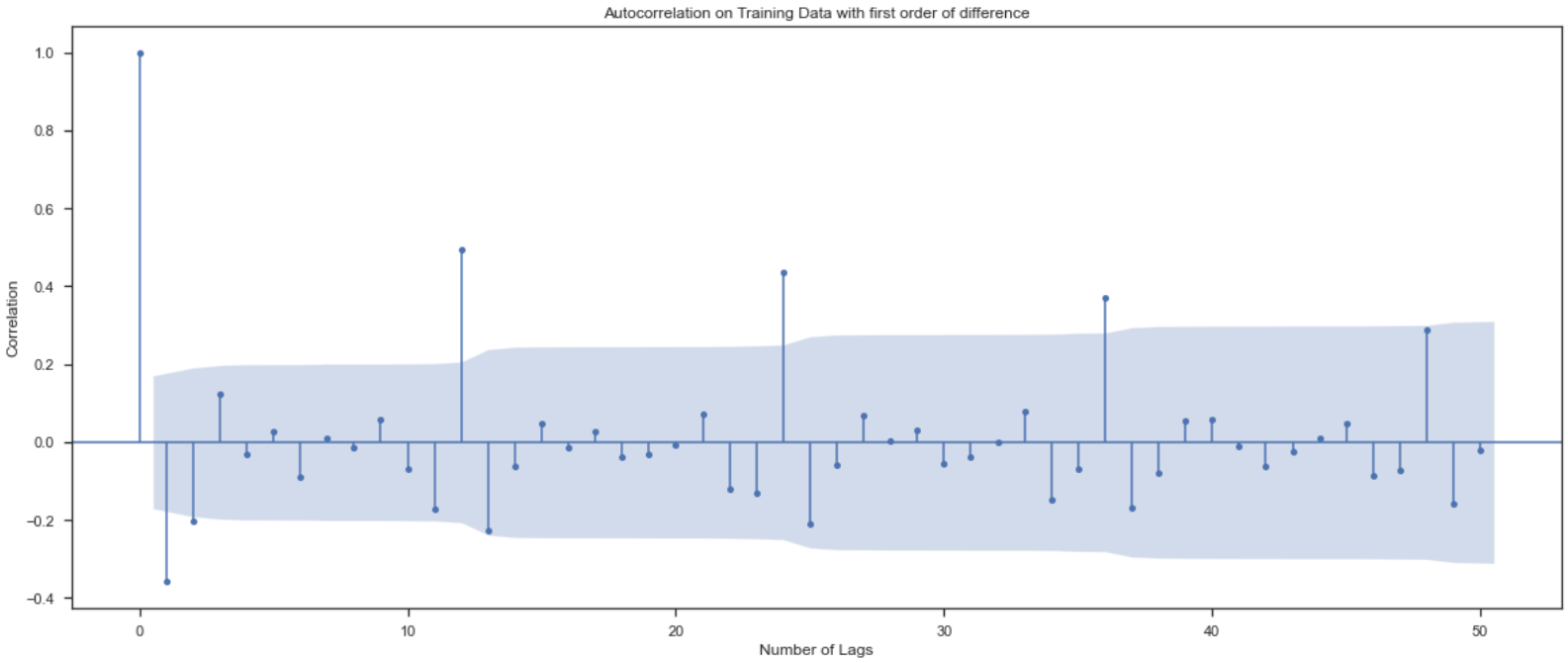


**Fig.73 ACF plot on differenced train data**

## PACF Plot – Training Data

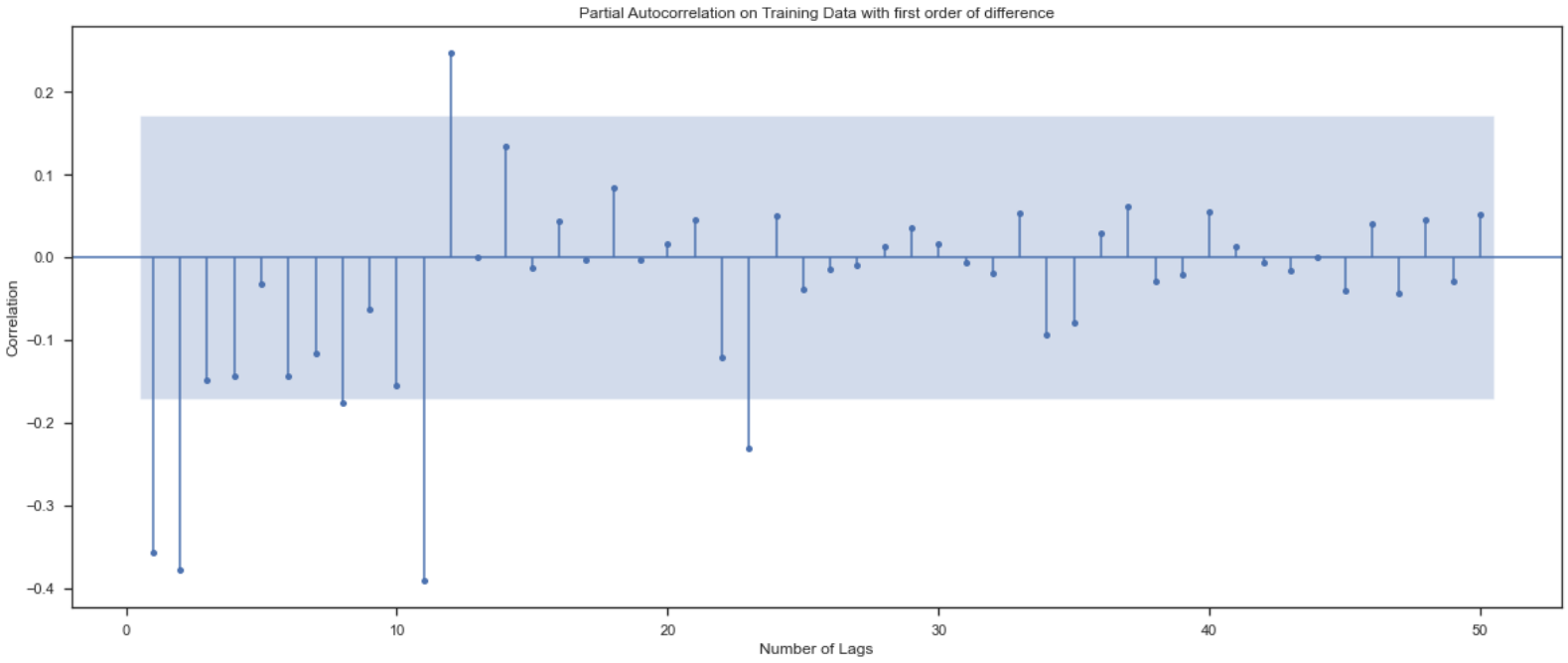


**Fig.74 PACF plot on differenced train data**

**Observation:**

- The **Auto-Regressive parameter** in an ARIMA model is **'p'** which comes from the significant lag after which the PACF plot cuts-off below the confidence interval.
- The **Moving-Average parameter** in an ARIMA model is **'q'** which comes from the significant lag after which the ACF plot cuts-off below the confidence interval.
- **By looking at the above plots, we will take the value of p=2 and q=2 respectively. The value of d=1, as with differencing the time series becomes stationary.**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Rose_Wine_Sales | No. Observations: | 132 |
| Model: | ARIMA(2, 1, 2) | Log Likelihood | -635.935 |
| Date: | Sat, 22 Oct 2022 | AIC | 1281.871 |
| Time: | 20:35:16 | BIC | 1296.247 |
| Sample: | 01-31-1980 | HQIC | 1287.712 |
| | - 12-31-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | -0.4540 | 0.469 | -0.969 | 0.333 | -1.372 | 0.464 |
| ar.L2 | 0.0001 | 0.170 | 0.001 | 0.999 | -0.334 | 0.334 |
| ma.L1 | -0.2541 | 0.459 | -0.554 | 0.580 | -1.154 | 0.646 |
| ma.L2 | -0.5984 | 0.430 | -1.390 | 0.164 | -1.442 | 0.245 |
| sigma2 | 952.1601 | 91.424 | 10.415 | 0.000 | 772.973 | 1131.347 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.02 | Jarque-Bera (JB): | 34.16 |
| Prob(Q): | 0.88 | Prob(JB): | 0.00 |
| Heteroskedasticity (H): | 0.37 | Skew: | 0.79 |
| Prob(H) (two-sided): | 0.00 | Kurtosis: | 4.94 |

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

**Fig.75 Rose Wine – Manual ARIMA model**

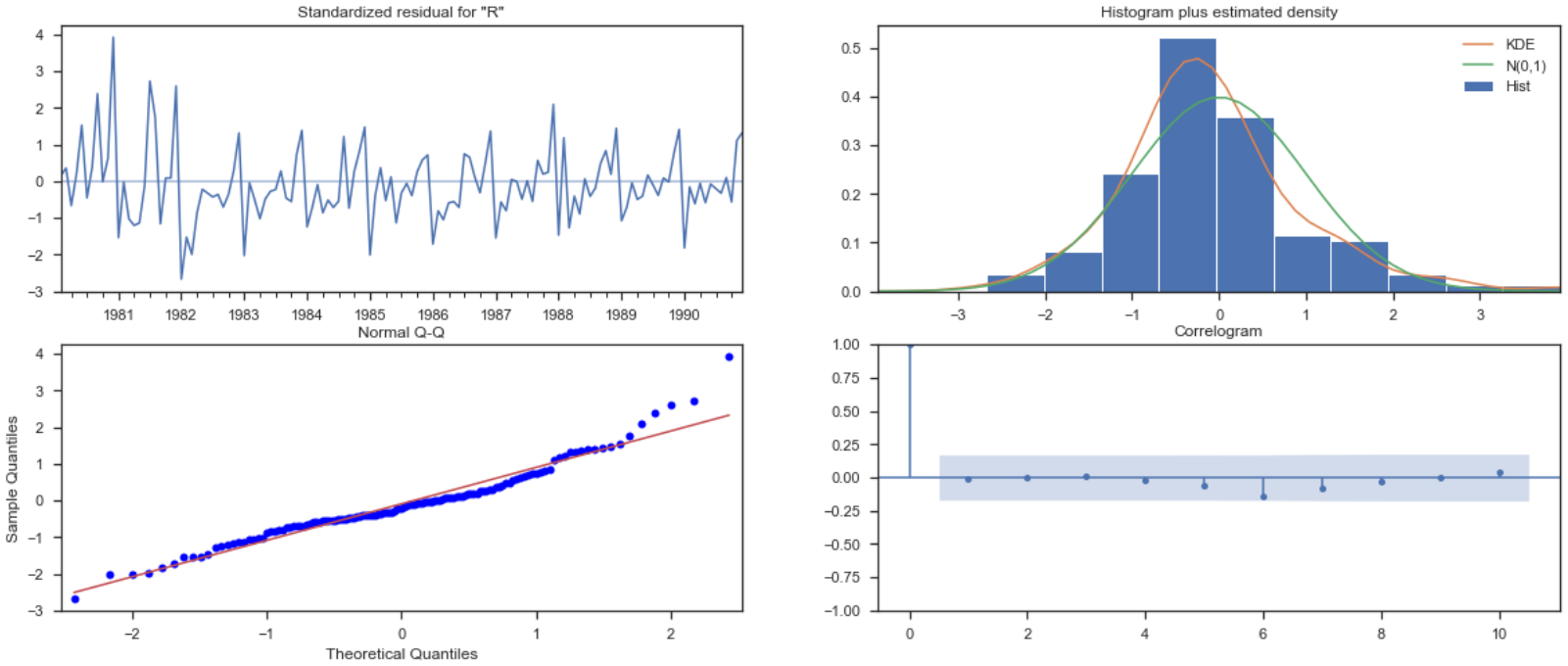


**Fig.76 Manual ARIMA – Diagnostics plot**

**Observation:**

- The model's parameters, p and q, were identified by examining the **ACF (q=2)** and **PACF (p=2)** graphs. Since we differenced the series to make it stationary, the parameter **d=1**.
- From the **Standardized residual plot** above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.
- The **histogram plus estimated density plot** suggests a slightly uniform distribution with mean zero and slightly skewed to the right.
- In **Normal Q-Q plot**, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.
- The **correlogram plot** of residuals shows that the residuals are not auto correlated.

| | |
|---|---|
| 1991-01-31 | 91.159512 |
| 1991-02-28 | 85.314628 |
| 1991-03-31 | 87.962468 |
| 1991-04-30 | 86.759563 |
| 1991-05-31 | 87.306038 |
| 1991-06-30 | 87.057777 |
| 1991-07-31 | 87.170561 |
| 1991-08-31 | 87.119324 |
| 1991-09-30 | 87.142601 |
| 1991-10-31 | 87.132026 |
| 1991-11-30 | 87.136830 |
| 1991-12-31 | 87.134648 |
| 1992-01-31 | 87.135639 |
| 1992-02-29 | 87.135189 |
| 1992-03-31 | 87.135393 |
| 1992-04-30 | 87.135300 |
| 1992-05-31 | 87.135342 |

**Fig.77 Sample of Manual ARIMA (2,1,2) predictions**

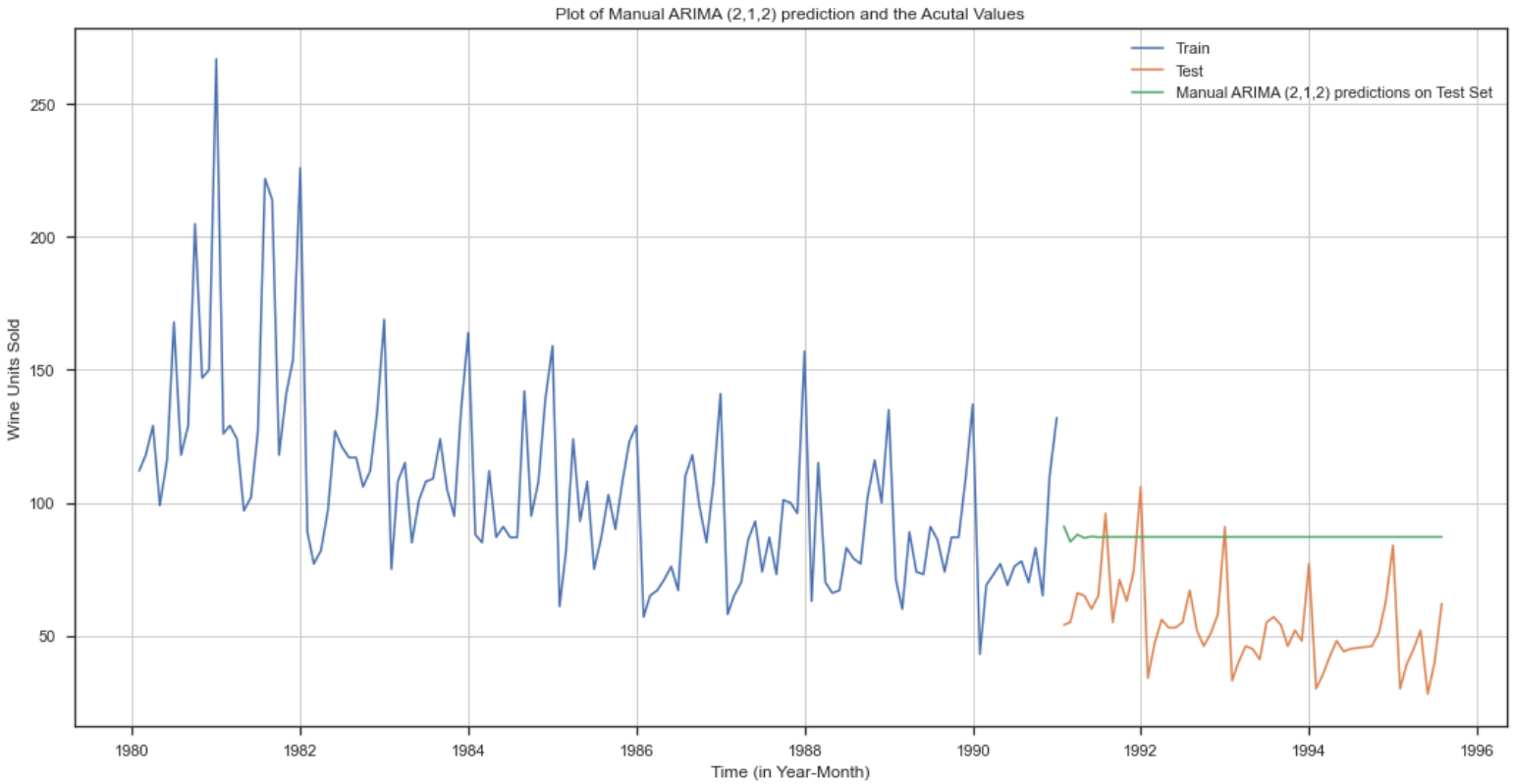


**Fig.78 Plot of Manual ARIMA (2,1,2) predictions on Test data**

## Manual ARIMA: Model Evaluation

**For evaluating the model performance, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)**

| Model | Test RMSE | Test MAPE |
|---|---|---|
| **ARIMA (p=2, d=1, q=2)** | **36.87** | **76.055** |

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- ARIMA models performs well on non-seasonal time series. It is due to this reason it is unable to capture the entire characteristics of the test data.
- The root means squared error **(RMSE)** of test data for the ARIMA model with **(p=2, d=1, q=2)** is **36.87.**
- Not surprisingly, the RMSE of the aforementioned ARIMA model is greater than the majority of previously constructed models and **nearly equal to ARIMA (2,1,3) model**.

## Model 11 – Seasonal Auto-Regressive Integrated Moving Average (SARIMA) – Manual

**A SARIMA model is characterized by 7 terms: p, d, q, P, Q, D and F**

**where,**

**p is the order of the Auto Regressive (AR) term**

**q is the order of the Moving Average (MA) term**

**d is the number of differencing required to make the time series stationary**

**P is the order of the Seasonal Auto Regressive (AR) term**

**Q is the order of the Seasonal Moving Average (MA) term**

**D is the number of seasonal differencing required to make the time series stationary**

**F is the seasonal frequency of the time series**

**We must examine the PACF and ACF plots, respectively, at delays that are the multiple of "F" in order to determine the "P" and "Q" values, and determine where these cut-off values are (for appropriate confidence interval bands).**

**By examining the ACF plots, one may calculate the seasonal parameter 'F'. The existence of seasonality should be shown by a spike in the ACF plot at multiples of "F."**

**The parameters P & Q can be determined by looking at the seasonally differenced PACF & ACF plots respectively.**

**Autocorrelation function (ACF) -** At lag *k*, this is the correlation between series values that are *k* intervals apart.

**Partial autocorrelation function (PACF) -** At lag *k*, this is the correlation between series values that are *k* intervals apart, accounting for the values of the intervals between.

**In an ACF & PACF plots, each bar represents the size and direction of the connection. Bars that cross the red line are statistically significant.**

## ACF Plot – Seasonally differenced (F=12) Training Data

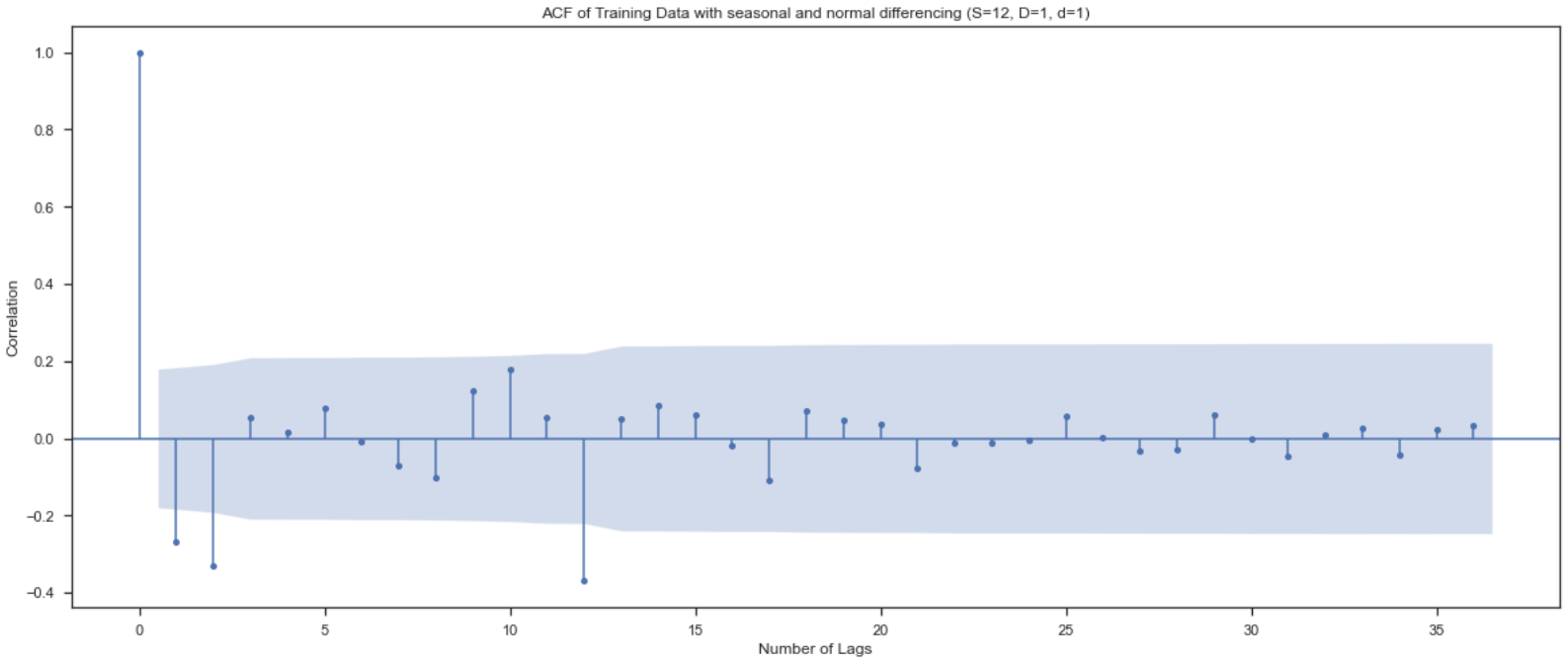


**Fig.79 ACF plot on differenced train data**

## PACF Plot – Seasonally differenced (F=12) Training Data

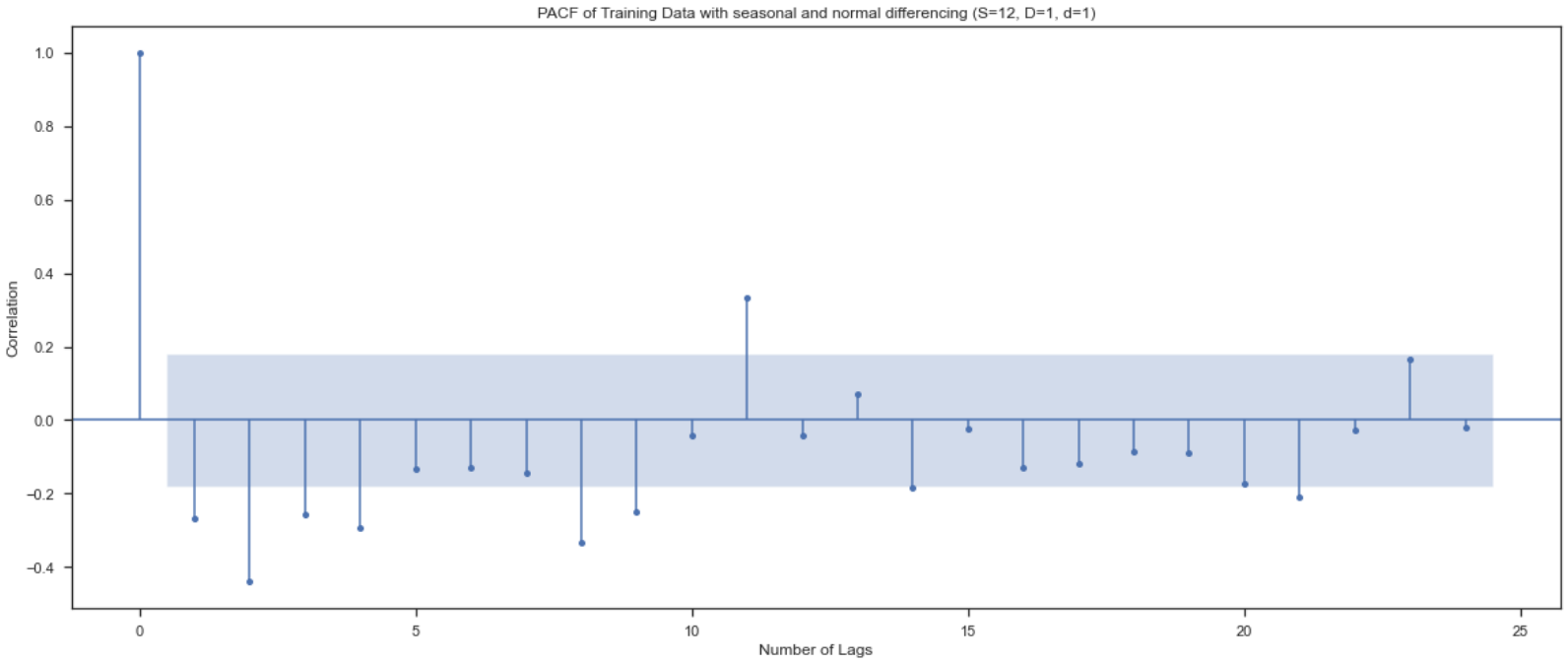


**Fig.80 PACF plot on differenced train data**

**Observation:**

- From the PACF plot it can be seen in early lags that till lag 4 is significant before cut-off, so AR term **'p = 4'** is chosen. From the multiples of seasonal lags, after first seasonal lag of 12, it cuts off, so keep seasonal AR **'P = 0'**.
- From ACF plot, it can be seen in early lags, lag 1 and 2 are significant before it cuts off, so let's 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 let's keep **'Q = 1'**.
- **The final selected terms for SARIMA model is (4, 1, 2) (0, 1, 1, 12), as inferred from the ACF and PACF plots.**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Rose_Wine_Sales | No. Observations: | 132 |
| Model: | SARIMAX(4, 1, 2)x(0, 1, [1], 12) | Log Likelihood | -446.102 |
| Date: | Sat, 22 Oct 2022 | AIC | 908.203 |
| Time: | 20:35:18 | BIC | 929.358 |
| Sample: | 01-31-1980 - 12-31-1990 | HQIC | 916.774 |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | -0.8047 | 0.119 | -6.778 | 0.000 | -1.037 | -0.572 |
| ar.L2 | 0.0387 | 0.140 | 0.276 | 0.783 | -0.237 | 0.314 |
| ar.L3 | -0.2310 | 0.147 | -1.568 | 0.117 | -0.520 | 0.058 |
| ar.L4 | -0.1875 | 0.108 | -1.742 | 0.082 | -0.398 | 0.024 |
| ma.L1 | 0.1434 | 308.416 | 0.000 | 1.000 | -604.342 | 604.628 |
| ma.L2 | -0.8566 | 264.208 | -0.003 | 0.997 | -518.694 | 516.981 |
| ma.S.L12 | -0.5405 | 0.085 | -6.384 | 0.000 | -0.707 | -0.375 |
| sigma2 | 296.7665 | 9.15e+04 | 0.003 | 0.997 | -1.79e+05 | 1.8e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.01 | Jarque-Bera (JB): | 0.03 |
| Prob(Q): | 0.94 | Prob(JB): | 0.98 |
| Heteroskedasticity (H): | 0.55 | Skew: | -0.02 |
| Prob(H) (two-sided): | 0.08 | Kurtosis: | 3.07 |

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

**Fig.81 Rose Wine – Manual SARIMA model**

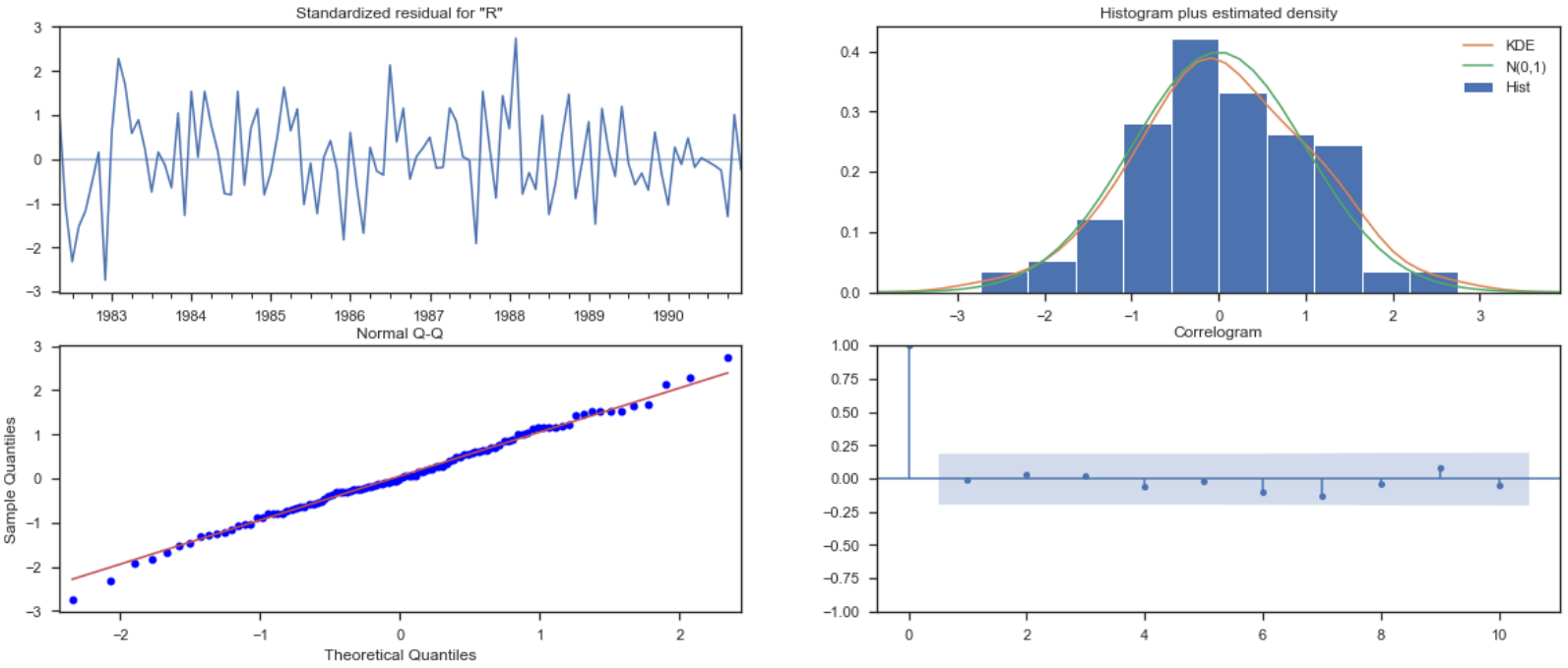


**Fig.82 Manual SARIMA – Diagnostics plot**

**Observation:**

- The model's parameters, p, q, P, Q were identified by examining the **ACF (q=2, Q=1)** and **PACF (p=4, P=0)** graphs. Since we differenced the series to make it stationary, the parameter **d=1, D=1**.
- From the **Standardized residual plot** above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.
- The **histogram plus estimated density plot** suggests a slightly uniform distribution with mean zero.
- In **Normal Q-Q plot**, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.
- The **correlogram plot** of residuals shows that the residuals are not auto correlated.

| | |
|---|---|
| 1991-01-31 | 47.471958 |
| 1991-02-28 | 63.350260 |
| 1991-03-31 | 65.513191 |
| 1991-04-30 | 67.318195 |
| 1991-05-31 | 61.756065 |
| 1991-06-30 | 72.817684 |
| 1991-07-31 | 71.513609 |
| 1991-08-31 | 67.808154 |
| 1991-09-30 | 77.918394 |
| 1991-10-31 | 73.747864 |
| 1991-11-30 | 97.306179 |
| 1991-12-31 | 127.634864 |
| 1992-01-31 | 41.226583 |

**Fig.83 Sample of Manual SARIMA (4,1,2) (0,1,1,12) predictions**

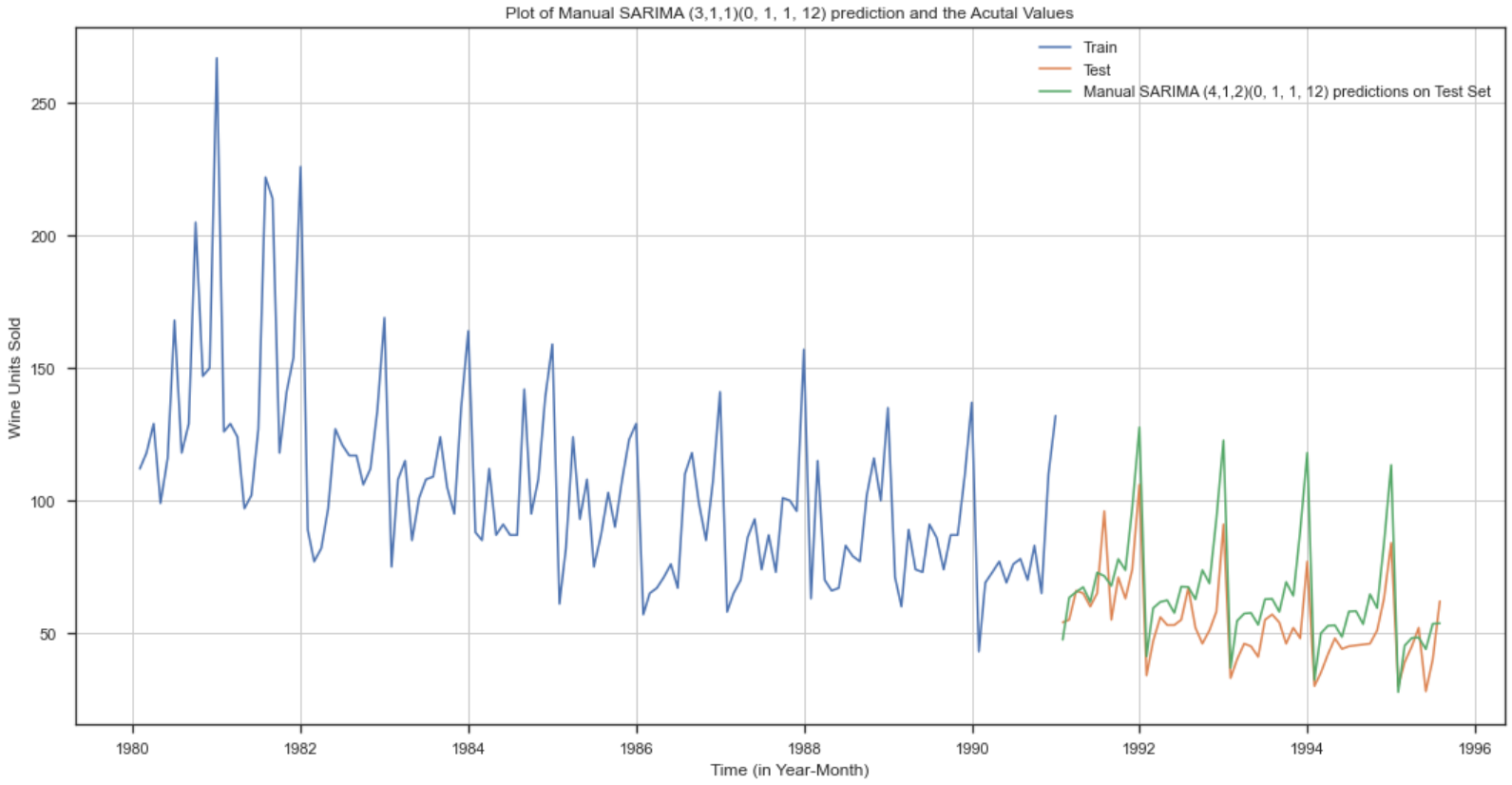


**Fig.84 Plot of Manual SARIMA (4,1,2) (0,1,1,12) predictions on Test data**

## Manual SARIMA: Model Evaluation

**For evaluating the model performance, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)**

| Model | Test RMSE | Test MAPE |
|---|---|---|
| **ARIMA (p=4, d=1, q=2) (P=0, D=1, Q=1, F=12)** | **15.907** | **23.712** |

**Observation:**

- We can see from the graphs above that the time series has a **falling trend and is seasonal**
- SARIMA model performs well on seasonal time series. It is due to this reason it is able to capture the entire characteristics of the test data.
- The root means squared error **(RMSE)** of test data for the SARIMA model with **(p=4, d=1, q=1) (P=0, D=1, Q=1, F=12)** is **15.907.**
- Additionally, it should be highlighted that compared to the all the ARIMA/SARIMA models built so far, **this SARIMA model has the lowest RMSE value**.

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

| | Test RMSE |
|---|---|
| Linear Regression | 15.268887 |
| Naive Model | 79.718576 |
| Simple Average | 53.460367 |
| 2 point TMA | 11.529278 |
| 4 point TMA | 14.451376 |
| 6 point TMA | 14.566262 |
| 9 point TMA | 14.727596 |
| Alpha=0.0987,SimpleExponentialSmoothing | 36.796036 |
| Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing | 15.268889 |
| Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing | 21.154527 |
| Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing | 9.121757 |
| Auto ARIMA (2,1,3) | 36.813265 |
| Auto SARIMA (3,1,1)(3,0,2,12) | 18.881815 |
| Manual ARIMA(2,1,2) | 36.870991 |
| Manual SARIMA (4, 1, 2)(0, 1, 1, 12) | 15.907309 |

**Fig.85 RMSE values of all models**

| | Test RMSE |
|---|---|
| Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing | 9.121757 |
| 2 point TMA | 11.529278 |
| 4 point TMA | 14.451376 |
| 6 point TMA | 14.566262 |
| 9 point TMA | 14.727596 |
| Linear Regression | 15.268887 |
| Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing | 15.268889 |
| Manual SARIMA (4, 1, 2)(0, 1, 1, 12) | 15.907309 |
| Auto SARIMA (3,1,1)(3,0,2,12) | 18.881815 |
| Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing | 21.154527 |
| Alpha=0.0987,SimpleExponentialSmoothing | 36.796036 |
| Auto ARIMA (2,1,3) | 36.813265 |
| Manual ARIMA(2,1,2) | 36.870991 |
| Simple Average | 53.460367 |
| Naive Model | 79.718576 |

**Fig.86 Sorted RMSE values of all models**

**Observation:**

- From the above table, we can see that **Triple Exponential Smoothing model** with parameters **(Alpha=0.2, Beta=0.85, Gamma=0.15)** has the **lowest RMSE** for test data.
- The **naïve forecast model** has performed the worst in terms of RMSE.

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

**From Fig.86 we observed the Triple Exponential Smoothing model is the optimum model for the given data set as it has the lowest RMSE value.**

**However, as we know SARIMA models tend to perform better with seasonal time series, we are also considering SARIMA model for the forecast.**

Let us visually see the time series plots of different models we have built so far on test data

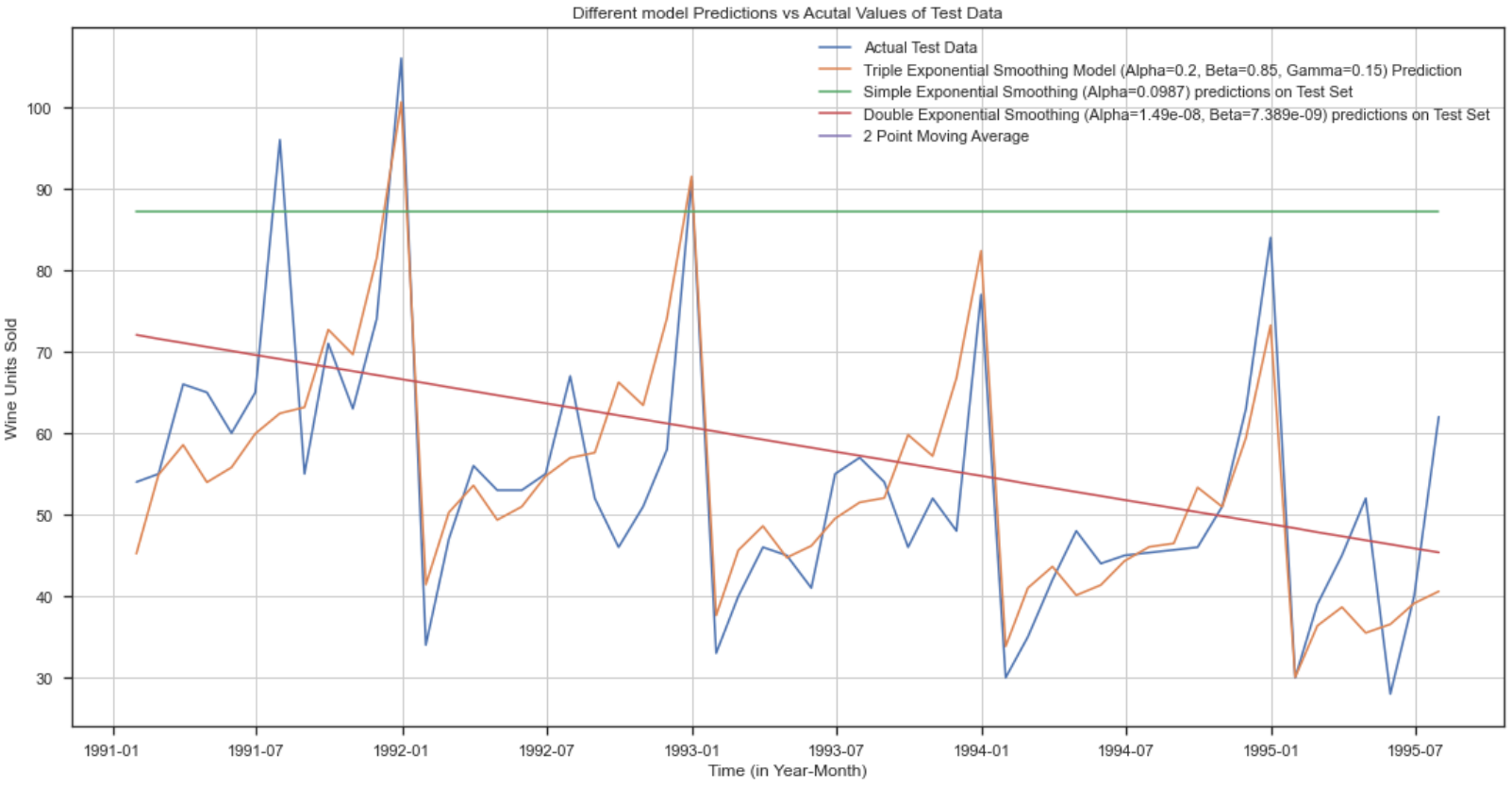


**Fig.87 Time Series Plot 1 – Different Model predictions on test data**

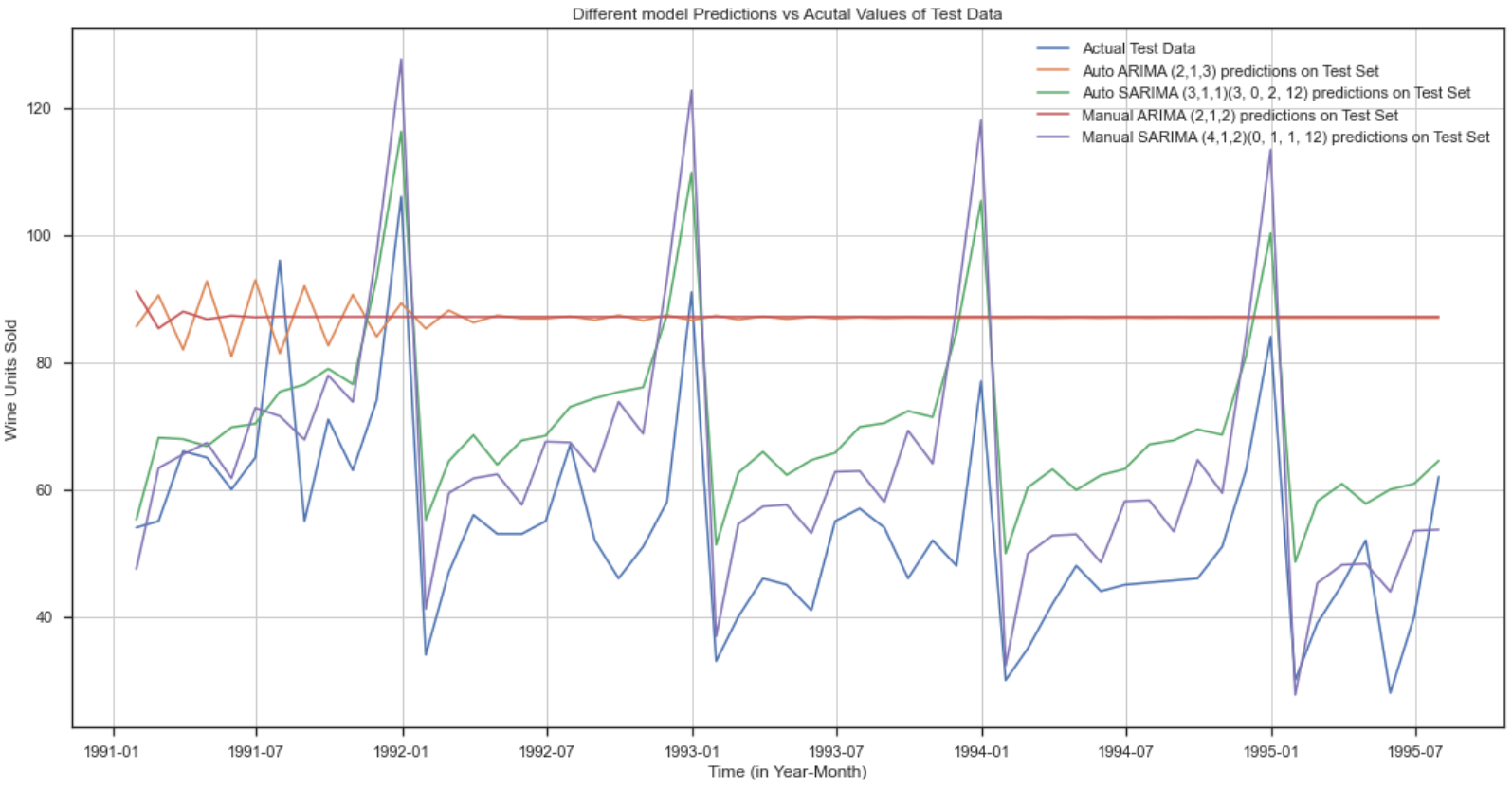


**Fig.88 Time Series Plot 2 – Different Model predictions on test data**

## Plotting the lowest RMSE models

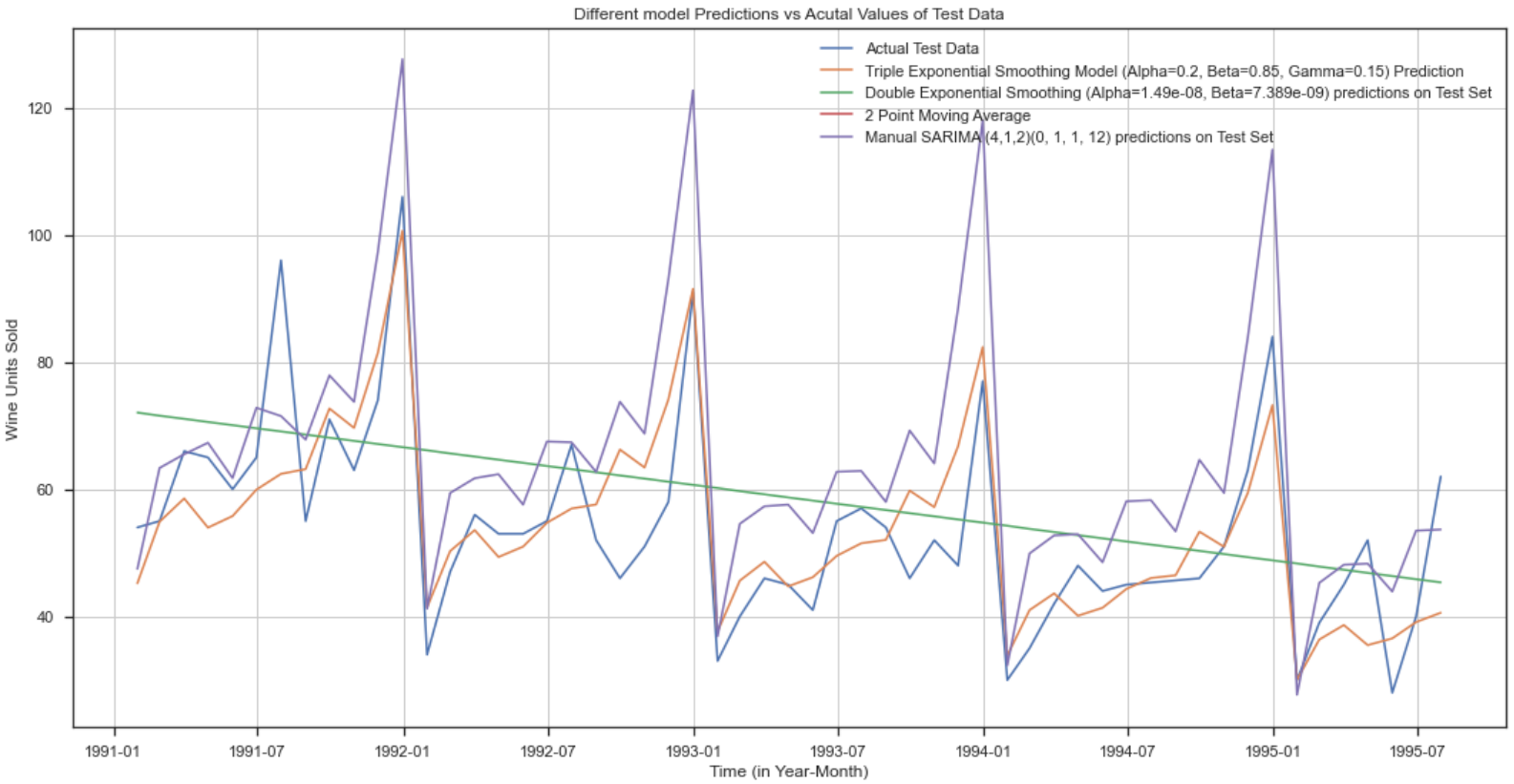


**Fig.89 Time Series Plot 3 – Different Model predictions on test data**

## Optimum Model 1:
## Triple Exponential Smoothing Model (Alpha=0.2, Beta=0.85, Gamma=0.15)

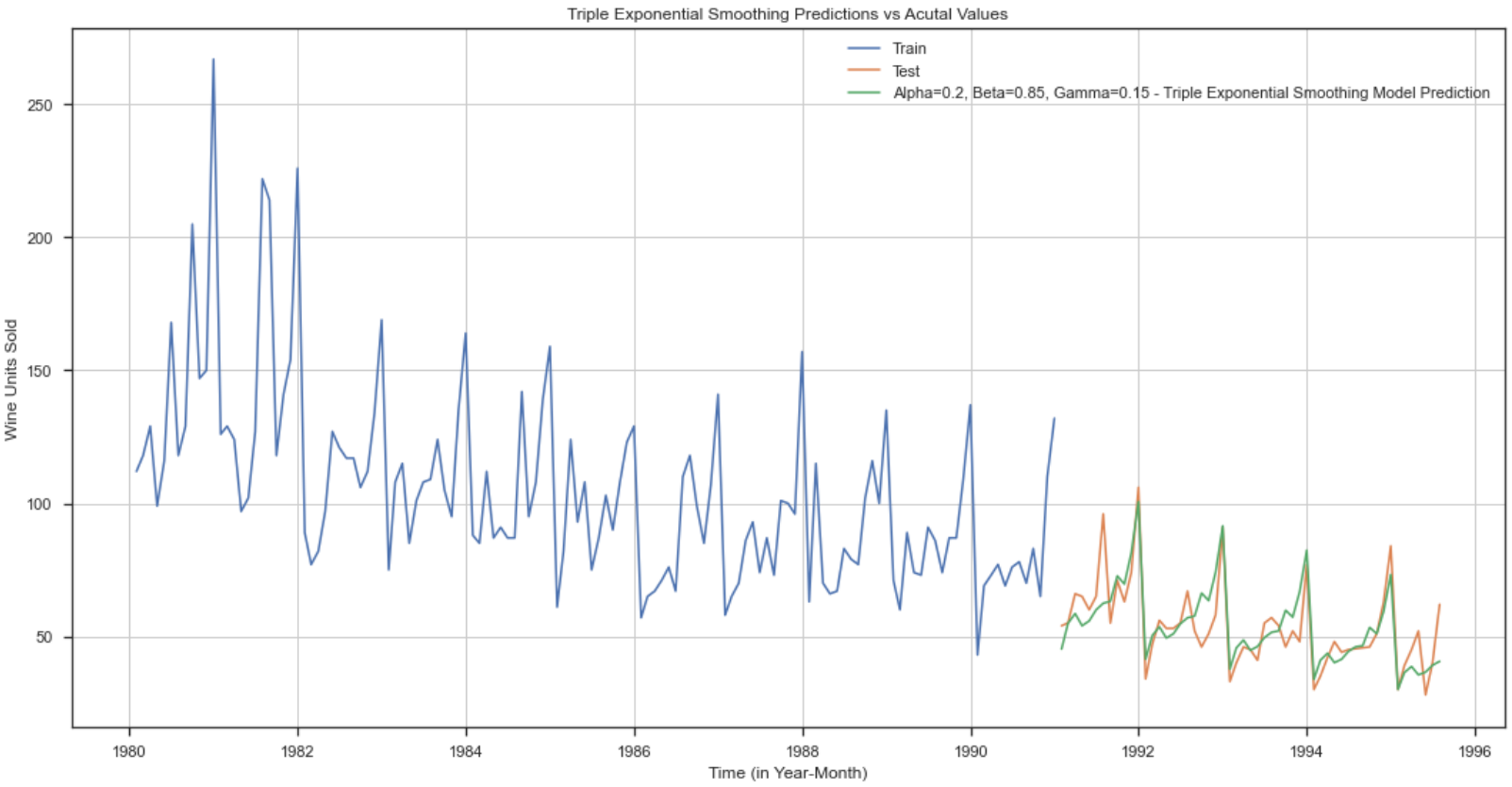


**Fig.90 TES Optimum Model – Line plot of Predictions vs Actual values**


**Fig.91 TES Optimum Model – Line plot of Predictions vs Actual values on Test data**

```
{'smoothing_level': 0.2,
 'smoothing_trend': 0.85,
 'smoothing_seasonal': 0.15,
 'damping_trend': nan,
 'initial_level': 50.83255814277326,
 'initial_trend': -0.6381657987079821,
 'initial_seasons': array([2.25583512, 2.42520915, 2.62094925, 2.00941291, 2.40
492965,
        2.82008636, 3.17994547, 3.50320751, 3.04727787, 3.04459042,
        3.35309239, 4.94083138]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

**Fig.92 TES Optimum Model**

```
Forecast of next 12 months

1995-08-31    38.192836
1995-09-30    39.507936
1995-10-31    41.376231
1995-11-30    49.568179
1995-12-31    70.756166
1996-01-31    28.650607
1996-02-29    36.970781
1996-03-31    44.051727
1996-04-30    45.545566
1996-05-31    40.265771
1996-06-30    45.364841
1996-07-31    48.647156
Freq: M, dtype: float64
```

**Fig.93 TES Model – Forecast for next 12 months**

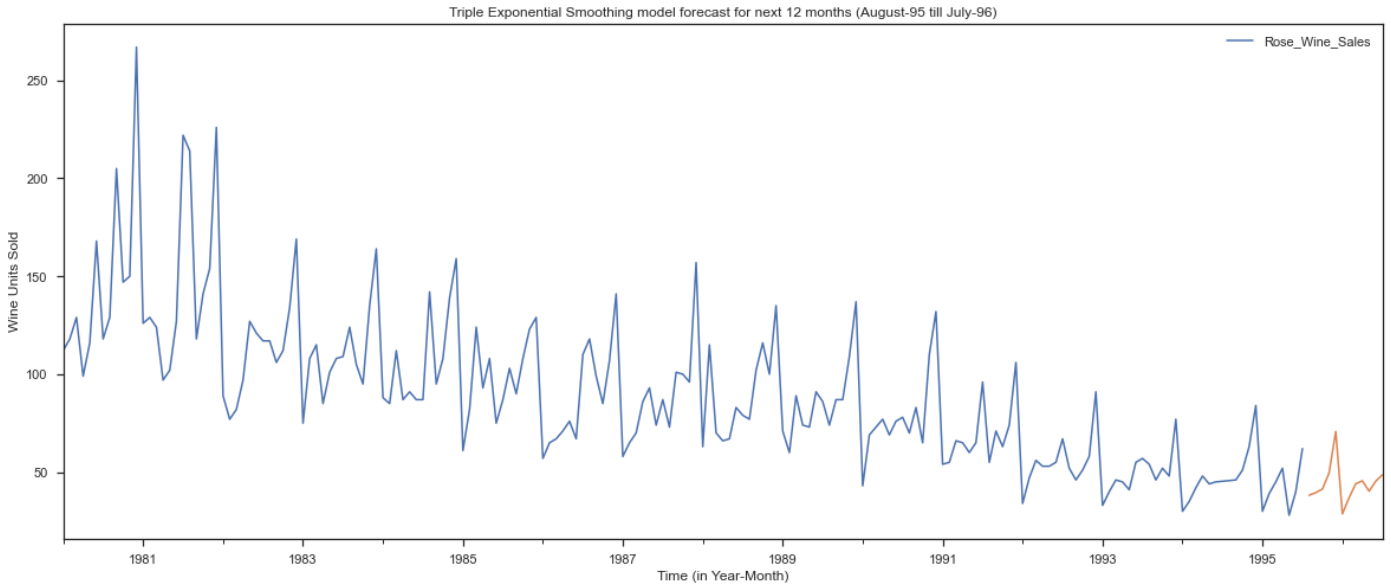


**Fig.94 TES Optimum Model – Time series plot forecast for next 12 months**

| | lower_ci | prediction | upper_ci |
|---|---|---|---|
| **1995-08-31** | -2.489648 | 38.192836 | 78.875320 |
| **1995-09-30** | -1.174548 | 39.507936 | 80.190420 |
| **1995-10-31** | 0.693747 | 41.376231 | 82.058715 |
| **1995-11-30** | 8.885695 | 49.568179 | 90.250663 |
| **1995-12-31** | 30.073682 | 70.756166 | 111.438650 |
| **1996-01-31** | -12.031877 | 28.650607 | 69.333091 |
| **1996-02-29** | -3.711703 | 36.970781 | 77.653265 |
| **1996-03-31** | 3.369243 | 44.051727 | 84.734211 |
| **1996-04-30** | 4.863082 | 45.545566 | 86.228050 |
| **1996-05-31** | -0.416713 | 40.265771 | 80.948255 |
| **1996-06-30** | 4.682357 | 45.364841 | 86.047325 |
| **1996-07-31** | 7.964672 | 48.647156 | 89.329640 |

**Fig.95 TES Optimum Model – Future forecast with confidence intervals**

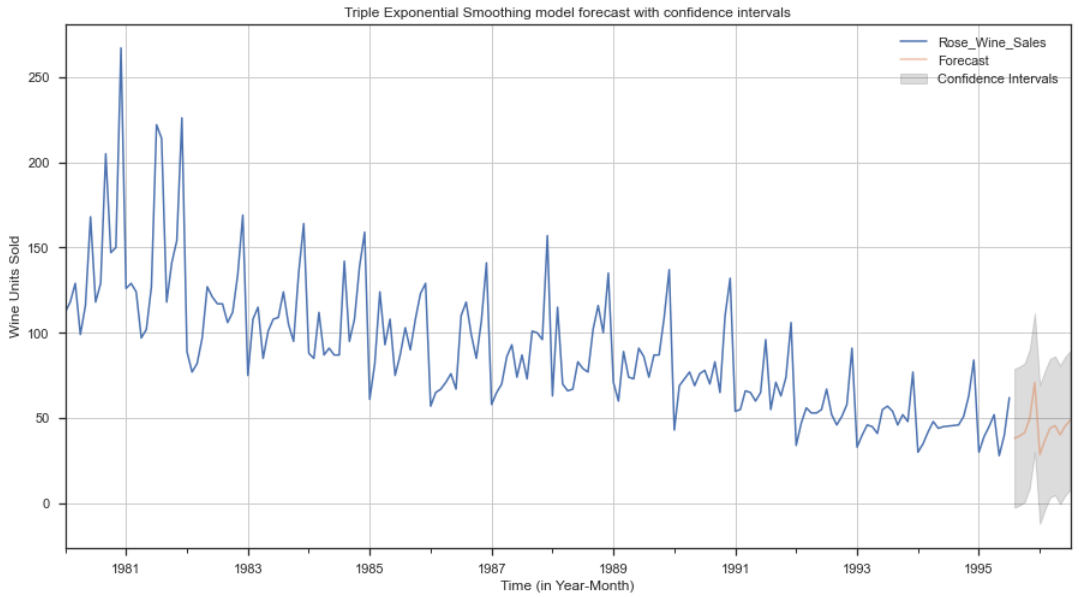


**Fig.96 TES Optimum Model – Time series plot forecast with confidence intervals**

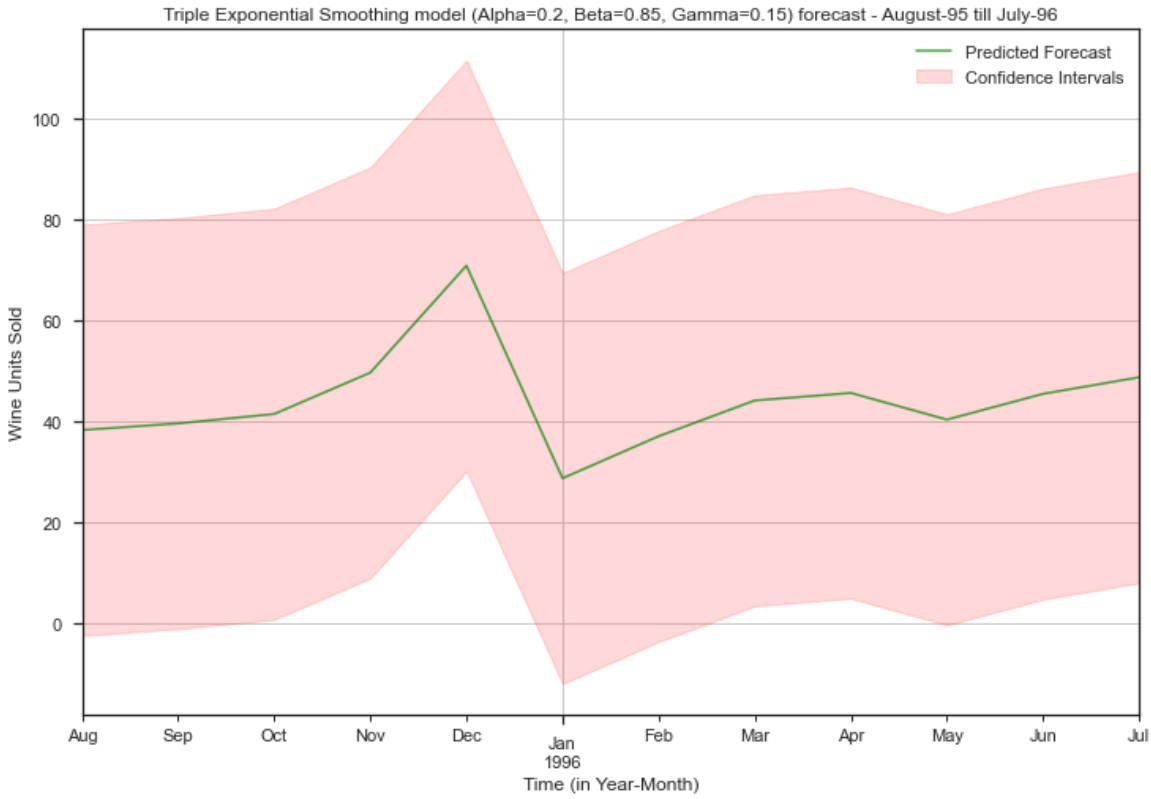


**Fig.97 TES Optimum Model – Time series plot forecast for next 12 months with confidence intervals**

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

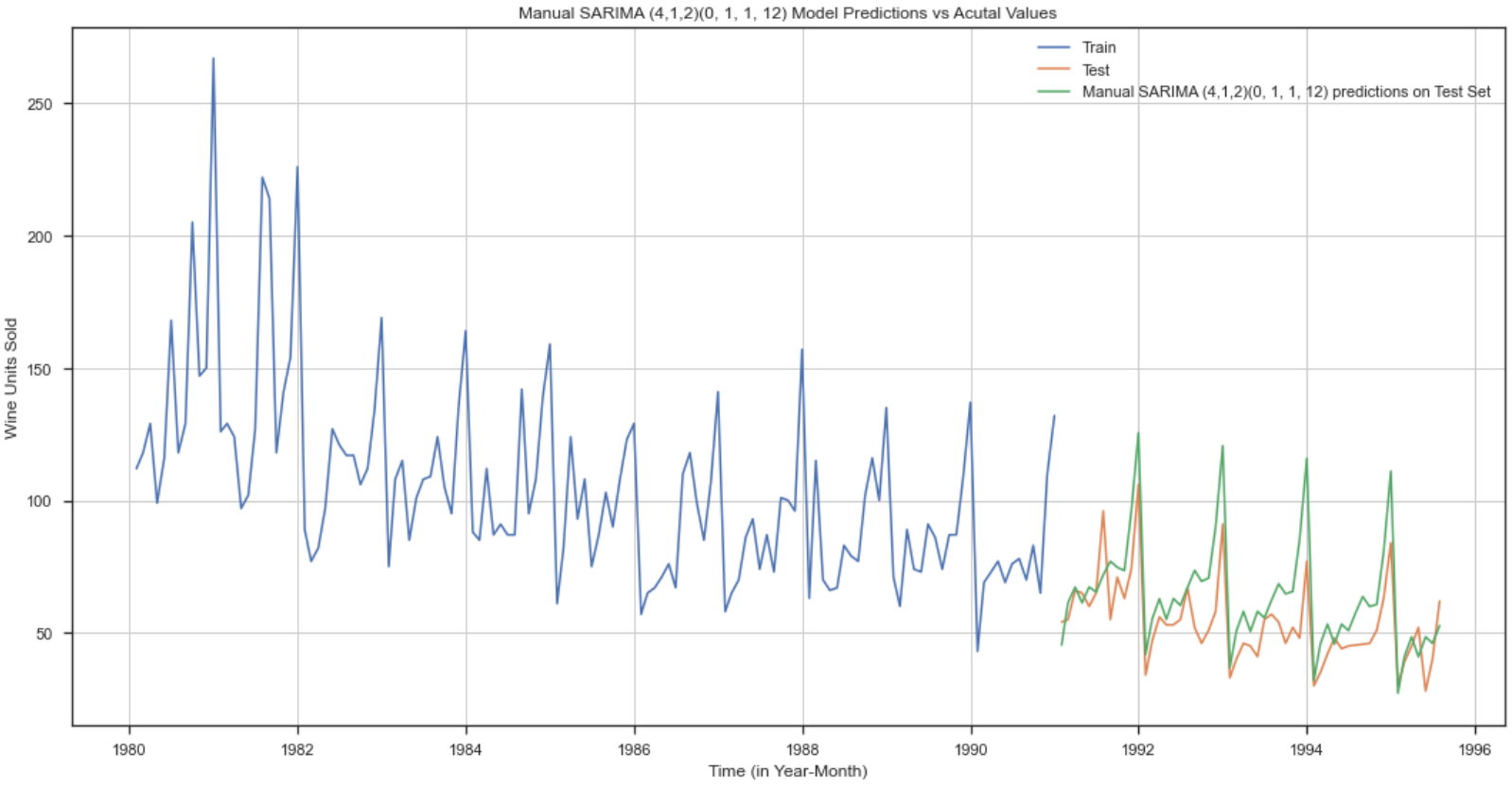


**Fig.98 Manual SARIMA Optimum Model – Line plot of Predictions vs Actual values**

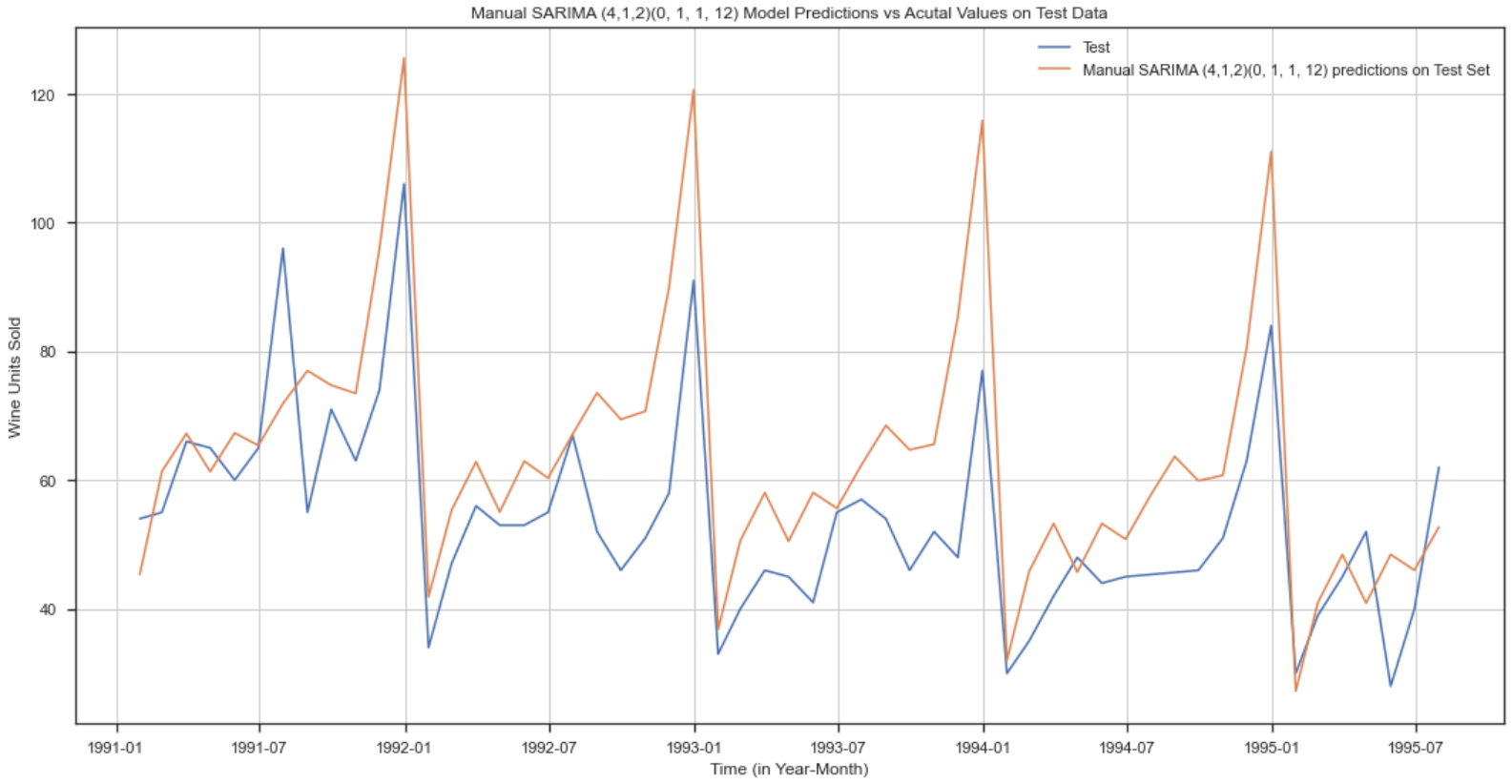


**Fig.99 Manual SARIMA Optimum Model – Line plot of Predictions vs Actual values on Test data**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Rose_Wine_Sales | No. Observations: | 187 |
| Model: | SARIMAX(4, 1, 2)x(0, 1, [1], 12) | Log Likelihood | -658.935 |
| Date: | Sat, 22 Oct 2022 | AIC | 1333.870 |
| Time: | 20:35:21 | BIC | 1358.421 |
| Sample: | 01-31-1980 - 07-31-1995 | HQIC | 1343.840 |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | -0.8239 | 0.083 | -9.939 | 0.000 | -0.986 | -0.661 |
| ar.L2 | 0.0470 | 0.107 | 0.438 | 0.662 | -0.163 | 0.257 |
| ar.L3 | -0.2147 | 0.110 | -1.955 | 0.051 | -0.430 | 0.001 |
| ar.L4 | -0.1692 | 0.078 | -2.182 | 0.029 | -0.321 | -0.017 |
| ma.L1 | 0.1564 | 96.851 | 0.002 | 0.999 | -189.667 | 189.980 |
| ma.L2 | -0.8436 | 81.697 | -0.010 | 0.992 | -160.967 | 159.280 |
| ma.S.L12 | -0.5418 | 0.061 | -8.898 | 0.000 | -0.661 | -0.422 |
| sigma2 | 225.0357 | 2.18e+04 | 0.010 | 0.992 | -4.25e+04 | 4.3e+04 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.02 | Jarque-Bera (JB): | 3.12 |
| 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).

**Fig.100 Manual SARIMA Optimum Model**

| Rose_Wine_Sales | mean | mean_se | mean_ci_lower | mean_ci_upper |
|---|---|---|---|---|
| 1995-08-31 | 48.272248 | 15.053370 | 18.768185 | 77.776311 |
| 1995-09-30 | 44.985078 | 15.830766 | 13.957347 | 76.012809 |
| 1995-10-31 | 45.475017 | 15.889160 | 14.332835 | 76.617198 |
| 1995-11-30 | 54.807633 | 15.899171 | 23.645831 | 85.969434 |
| 1995-12-31 | 81.904878 | 15.913902 | 50.714203 | 113.095553 |
| 1996-01-31 | 25.671442 | 16.199778 | -6.079539 | 57.422423 |
| 1996-02-29 | 33.894947 | 16.307905 | 1.932041 | 65.857852 |
| 1996-03-31 | 40.048092 | 16.589848 | 7.532589 | 72.563596 |
| 1996-04-30 | 44.383650 | 16.657861 | 11.734843 | 77.032458 |
| 1996-05-31 | 31.339014 | 16.872204 | -1.729898 | 64.407925 |
| 1996-06-30 | 39.915558 | 16.946299 | 6.701423 | 73.129694 |
| 1996-07-31 | 52.375724 | 17.156264 | 18.750063 | 86.001384 |

**Fig.101 Manual SARIMA Model – Forecast for next 12 months with confidence intervals**

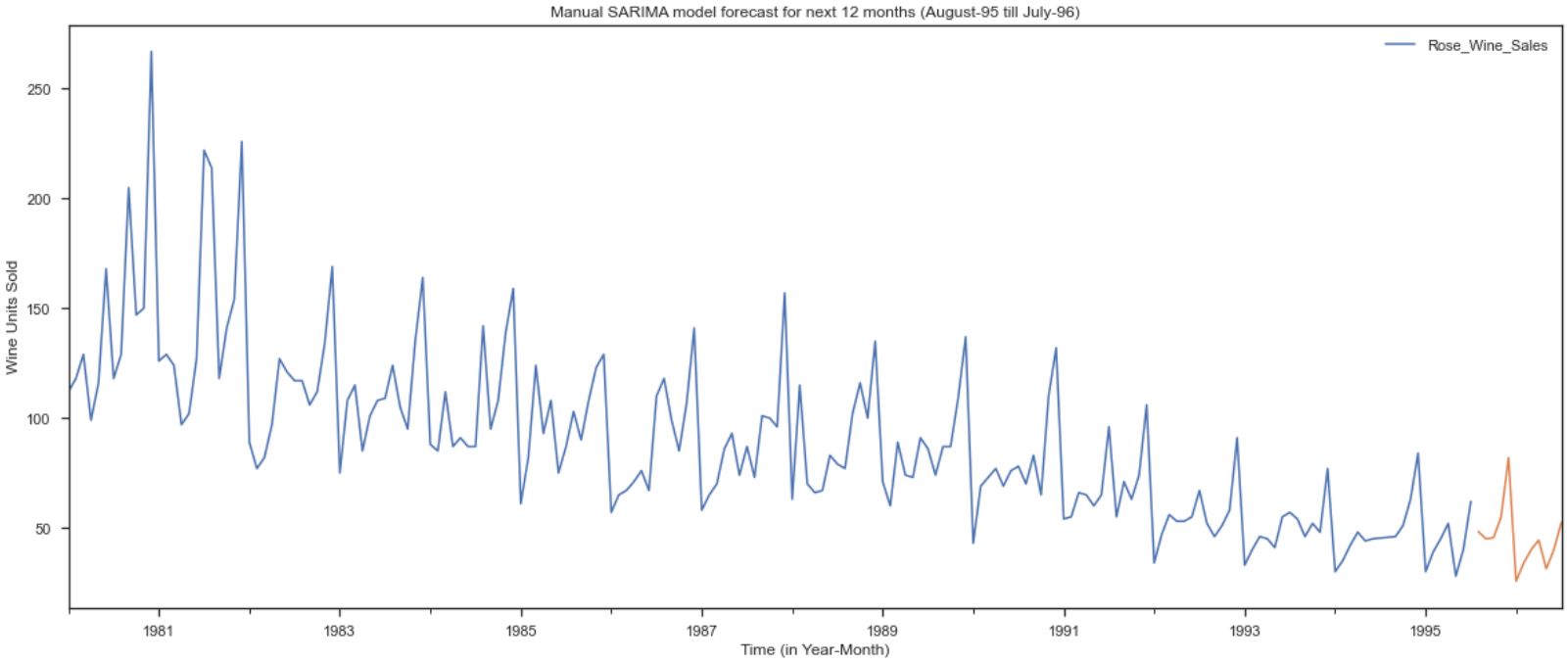


**Fig.102 Manual SARIMA Optimum Model – Time series plot forecast for next 12 months**

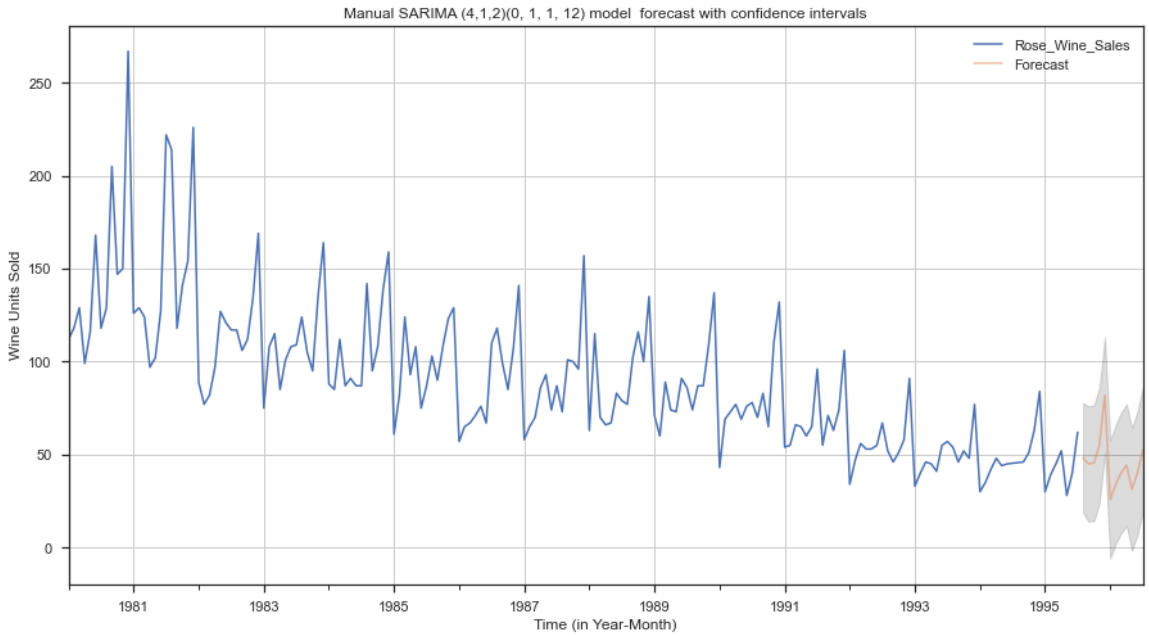


**Fig.103 Manual SARIMA Optimum Model – Time series plot forecast with confidence intervals**

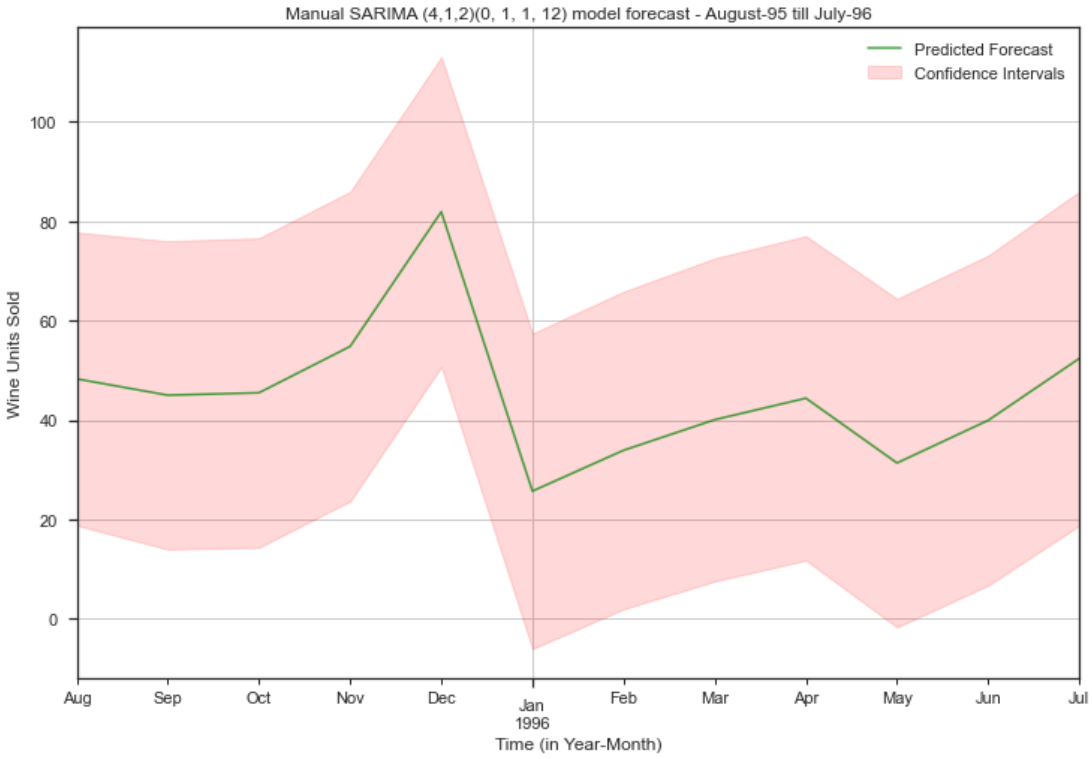


**Fig.104 Manual SARIMA Optimum Model – Forecast for next 12 months with confidence interval**

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

**We needed to construct an optimum model to forecast the rose wine sales for the next 12 months. The model information, insights and recommendations are as follows.**

### Model Insights:

- The time series in consideration exhibits a declining trend and stable seasonality. When comparing the various models, we can see that **Triple Exponential Smoothing and SARIMA models frequently deliver the greatest results**. This is due to the fact that these models are **excellent at predicting time series that demonstrate trend and seasonality**. Apart from these Double Exponential Smoothing and Moving Average Models also tend to perform moderately good.
- We examine the **root mean squared value of the forecast model to assess its performance (RMSE)**. The model with the lowest RMSE value and characteristics that match the test data is regarded as being a superior model.
- We observed that **Triple Exponential Smoothing** model had the lowest RMSE and the characteristics that most closely fit test data. As a result, its regarded as the **best model for forecasting** and can thus be used by the company for forecast analysis.

### Historical Insights:

- The rose wine **sales have declined throughout time**. Rose wine sales **peaked in 1980 & 1981** and fell to their **present low position in 1995** (as we have data for only first 7 months).
- The monthly sales trajectory appears to be exactly the opposite of the yearly plot, with a progressive increase towards the end of each year. **January has the lowest wine sales**, while **December has the highest**. From January to August, sales increase gradually, and then they quickly increase after that.
- The **average monthly sales** of Rose wine are **90 bottles**. More than 50% of the sold units of rose wine fall between 62 and 111. **28 units were sold as the lowest** and **267 units as the most**. Only 20% of monthly sales that were recorded were for more than 120 units.
- Around **70 to 75 percent of the units sold are fewer than 100**, and 90% of the units sold are less than 150. Only 15% of sales involved more than 50 items. Therefore, it is clear that the **bulk of sales were in the range of 50 to 100 units**.

## Forecast Insights:

- **Based on the forecast made by the Triple Exponential Smoothing model previously presented, the following insights are offered.**
- The forecast calls for average sale of 44 units, down by 45 units from the historical average of 89 units. Thus, we might observe an **alarming decrease in average sales by 50%**.
- The prediction for minimum sales volume of 28 units equals the minimum sales volume in the past. Consequently, a **no percentage change could be seen in minimum quantity sold**.
- The projection estimates a maximum sales volume of 70 units, which is 197 units fewer than the largest sales volume recorded in the past, which was 267 units. Consequently, a **73% decrease in maximum sales is visible**.
- In comparison to the historical standard deviation of 62 recorded in the past, the **forecast's standard deviation** is 10 units, or 52 units lower. It's **gone down by 83%.** This is not anticipated because historical data tends to have less volatility than future data.
- We can see from the prediction that the months of **October, November, and December have increased sales**. **December is often when the sales are at their highest**. There is a **startling decline in sales in January following December**. The **months after January appear to witness a gradual improvement in sales until October, when it jumps sharply**.

## Recommendations:

- Records show that the months of **September**, **October, November, and December account for 40% of the total sales forecast**. **Many festivities** take place in these months, and **many people travel during this time**. One of the most premium types of wine used during festive and event celebrations is rose wine.
- Wine sales often climb in the final two months of the year as **people hurry to buy holiday beverages**. For forthcoming **occasions like Thanksgiving, Christmas, and New Year's, people typically stock up**. The majority of individuals also **buy in bulk for holiday gatherings and gift-giving.**
- **Many individuals choose wine as their go-to gift** when it comes to occasions like parties and gift-giving. Sales of Rose wine rise just before the winter holidays as more collectors purchase these wines as presents or look for vintages to serve at holiday gatherings.

- This blush wine **works nicely with nearly anything**, including spicy dishes, sushi, salads, grilled meats, roasts, and rich sauces. It is well **renowned for its outdoor-friendly drinking style**.
- The festival seasons may vary depending on where you are geographically, however the most of the celebrations take place in the last four months.
    - **In these months, promotional offers might be implemented to lower costs and significantly boost revenue.**
    - **To increase sales, we must take advantage of all holiday events and set prices appropriately.**
    - **Many individuals order in bulk to prepare for upcoming festivities, which may result in a high shipping expenditure. Businesses may provide significant discounts or free shipping beyond a certain threshold at these times.**
    - **Giving customers gifts to improve their user experience is one of the greatest marketing strategies to deploy. In order to attract more consumers and increase sales, the company might provide free gifts on orders with significant sales.**
    - **To target various client demographics, the proper marketing campaigns must be run**
    - **Numerous ecommerce campaigns and competitions may be performed to broaden the product's audience and enhance sales.**
- The period **from January to June is one of the key challenges** for Rose wine sales.
    - **To identify the elements affecting sales, in-depth market research must be conducted.**
    - **Due to the fact that rose wines are premium category of wine, a market-friendly version of the existing product might be introduced by the company, helping to make up for the drop in sales. Long-term, this may bring in additional clients.**
    - **The company can rebrand its product to instill a fresh perspective towards the product and break the declining sales trend.**
- There are other key elements that might be driving the sales, despite the present model's ability to closely track the historical sales trend.
    - **The forecast might be improved by doing in-depth market research on the factors that influence sales and incorporating that information into the model for projection**

# Sparkling Wine Analysis



## Executive Summary

**Data on wine sales from the 20th century are available from ABC Estate Wines, a wine producing firm, and should be examined. With the provided information, an estimate of wine sales in the 20th century must be forecasted.**


**Fig.105 Sparkling Wine Analysis**

## Introduction

The purpose of this **report** is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of **sales of Sparkling wine from 20th century.**

## Data Dictionary

| Variable Name | Description |
|---|---|
| **YearMonth** | Represents the year and month in which the sales were recorded |
| **Sparkling** | Denotes the number of wine units sold |

## Data Description

3. **YearMonth:** Datatime variable from 1980-01 to 1995-07
4. **Sparkling:** Continuous from 1070 to 7242

## Sample of the dataset

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

**Table 3. Sample of first 5 rows of the dataset**

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

**Table 4. Sample of last 5 rows of the dataset**

Dataset has 2 columns which captures the Year and Month of recorded data and the number of units sold on corresponding Year-Month respectively.

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

**Let us check the types of variables in the data frame and check for missing values in the dataset**

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

**Fig.106 Details of the dataset columns**

**The dataset has 2 variables and 187 rows in total**. The **"YearMonth"** column can be **deleted after creating a suitable time stamp column because it is not necessary** for our modelling. The column **Sparkling** is of **float type**. Additionally, we can observe from the data above that **Sparkling column has no missing values**.

### Time Stamp created from 'YearMonth' column

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

**Fig.107 Details of the dataset columns**

## Resulting dataset after removing the "Year-Month" column and appending Time_Stamp column

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

**Fig.108 Details of the dataset columns**

**Time_Stamp** column has been **set as index** of the dataset and **column Sparkling** has **been renamed as Sparkling_Wine_Sales.**

## Renaming the columns of the data frame

The below mentioned columns of the data frame have been renamed as shown.

| Original Column Name | Renamed Column Name |
|---|---|
| Sparkling | Sparkling_Wine_Sales |

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

**Fig.109 Details of the dataset columns after renaming**

## Checking null values in the dataset

```
Sparkling_Wine_Sales    0
dtype: int64
```

**Fig.110 Null values in the dataset**

**As can be seen from the above figure, there are no null values present in the dataset.**

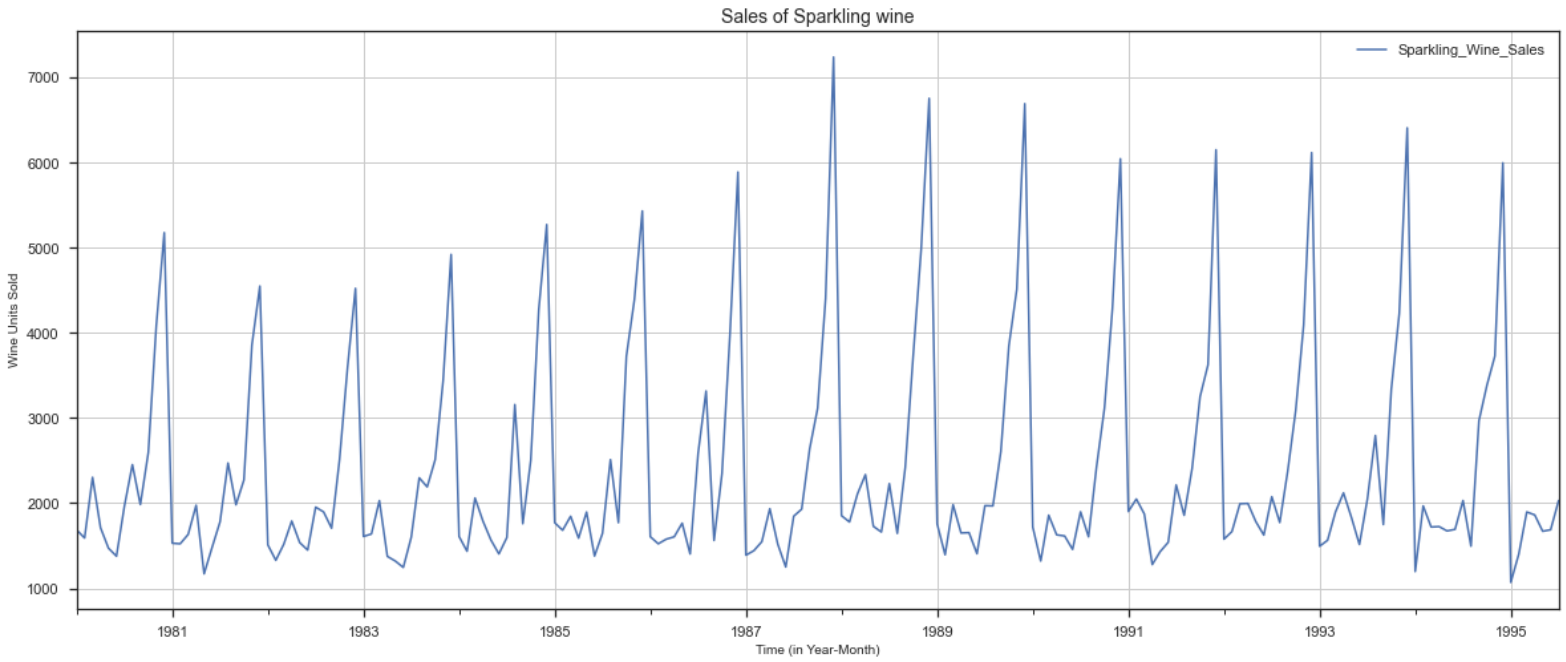


**Fig.111 Graph plot of the Sparkling wine sales dataset**

**Observation:**

- The data set provided contains sales information from January 1980 to July 1995.
- We can see from the plot that there has been a constant pattern of sales with seasonality. Over the years, the sales have consistent. The data also exhibits seasonality, as may be shown.
- There are no missing values which must be imputed.

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

### Descriptive Summary of the Dataset

| | Sparkling_Wine_Sales |
|---|---|
| count | 187.000000 |
| mean | 2402.417112 |
| std | 1295.111540 |
| min | 1070.000000 |
| 25% | 1605.000000 |
| 50% | 1874.000000 |
| 75% | 2549.000000 |
| max | 7242.000000 |

**Fig.112 Descriptive Summary of Sparkling_Wine_Sales column**

**Observation:**

- 2402 bottles of sparkling wine are typically sold each month.
- Between 1605 and 2549 units make up more than 50% of the sold sparkling wine units.
- The lowest unit sold is 1070 units, while the highest unit sold is 7242 units.
- Only 25% of monthly sales that recorded are more than 2549 units.

## Exploratory Analysis

Let us analyze the wine sales across different years and months using boxplots

### Yearly Plot

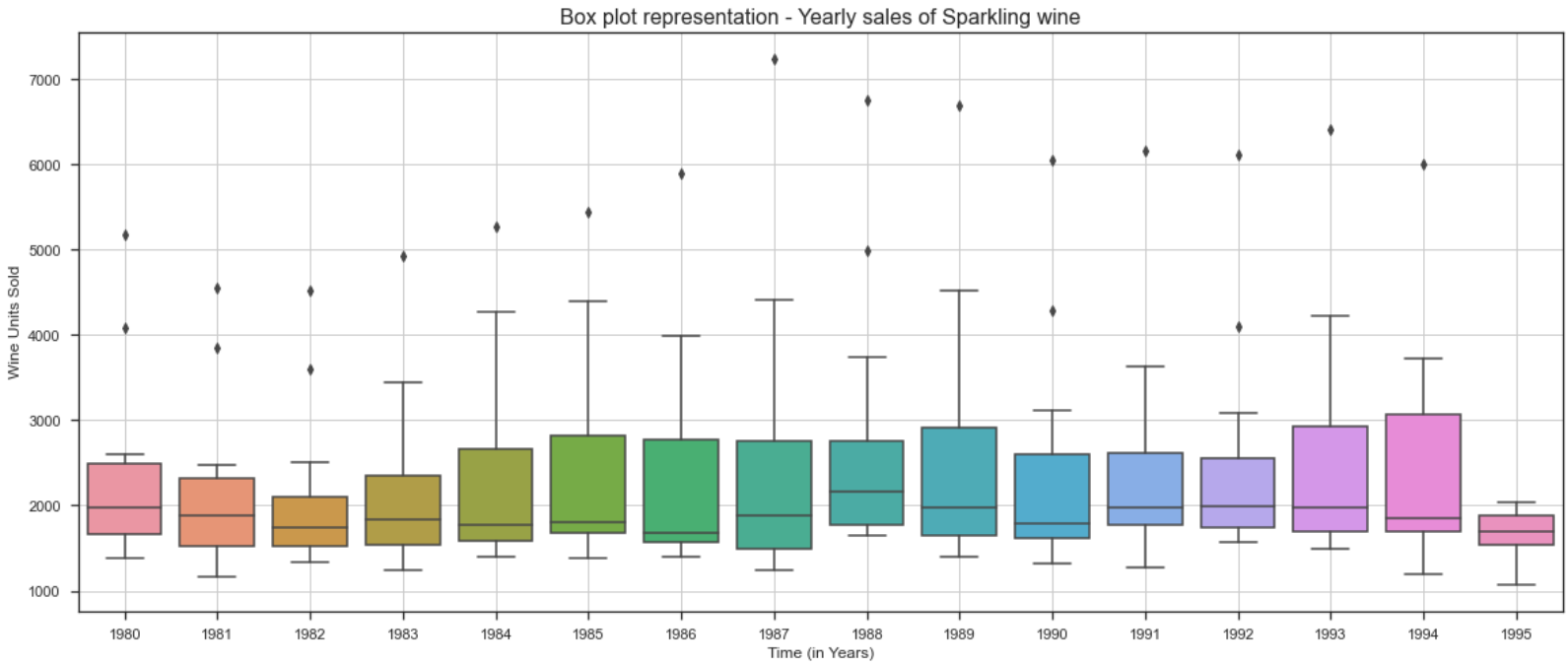


**Fig.113 Yearly plot of Sparkling wine sales**

**Observation:**

- We can see from the figure above that sales of sparkling wine have remained constant over the years.
- The median sales of sparkling wine reached their peak in 1988 and their current low point in 1995.
- Additionally, we can see that there are outliers in the box plots.

## Monthly Plot

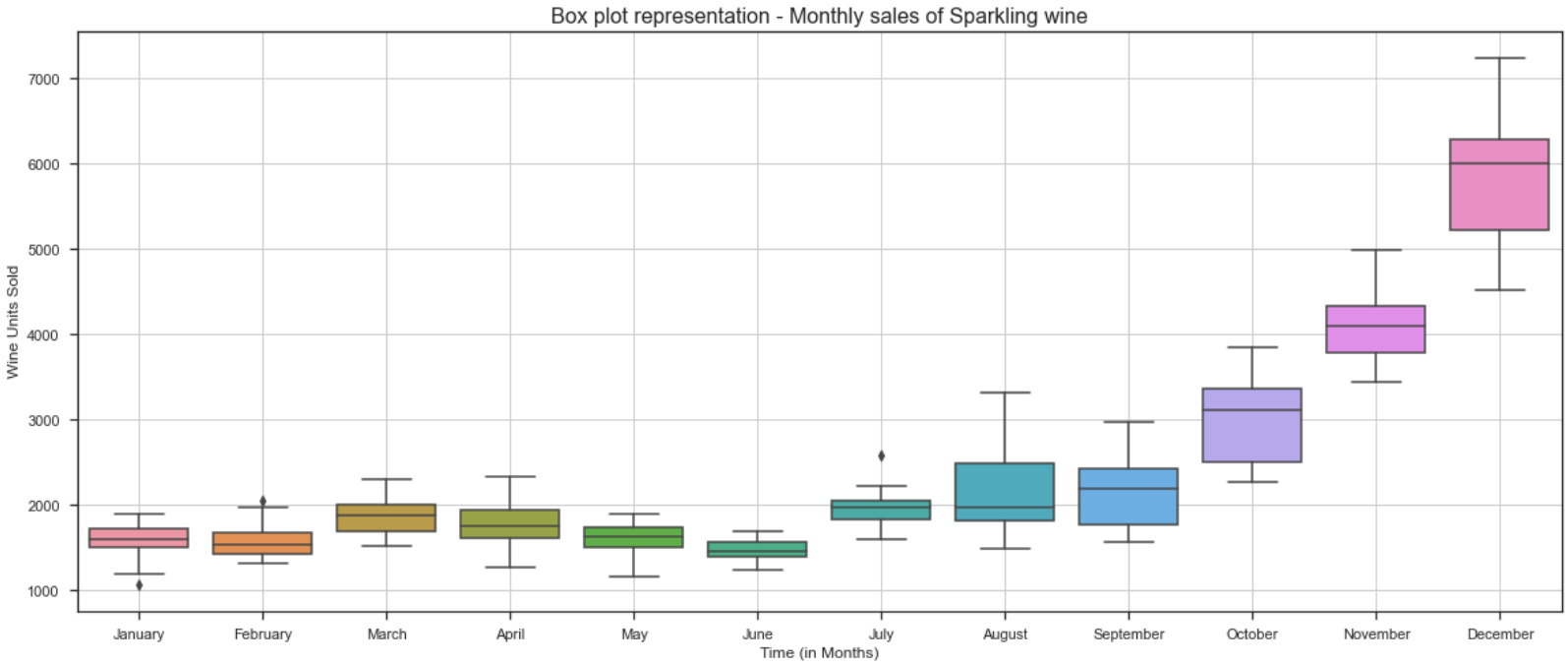


**Fig.114 Monthly plot of Sparkling wine sales**

**Observation:**

- The sales trajectory appears to be precisely the reverse of that seen in the yearly plot, seeing a gradual increase towards the end of each year.
- January has the lowest wine sales while December sees the greatest. The sales modestly grow from January to August and then sharply climb after that.
- Additionally, we can see that there are few outliers in the box plots.

## Annual Sales

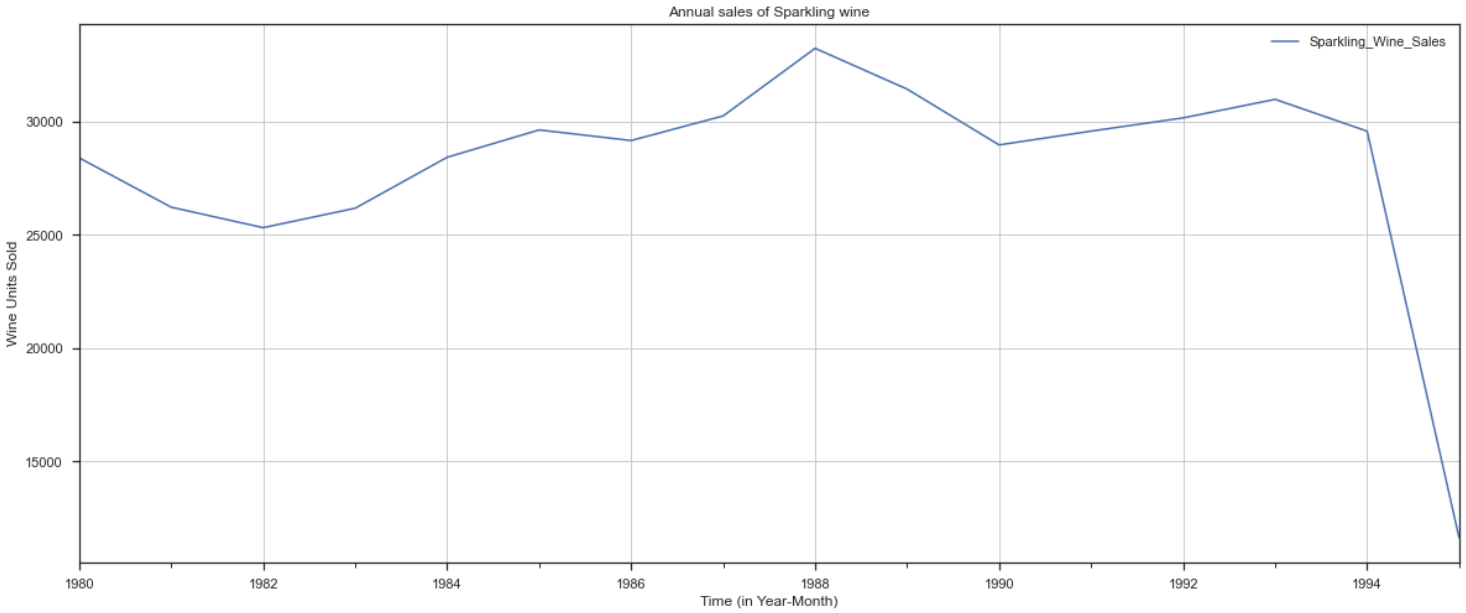


**Fig.115 Line plot – Annual sales**

## Quarterly Sales

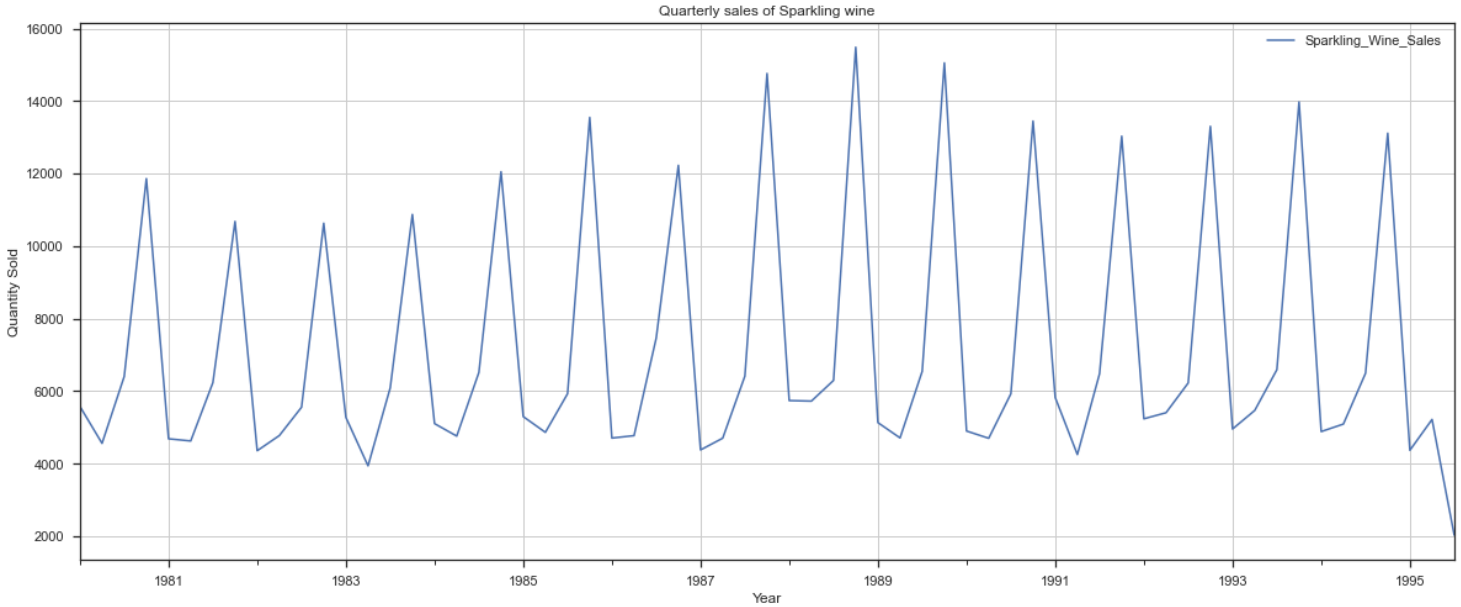


**Fig.116 Line plot – Quarterly sales**

## Monthly Sales across Different Years

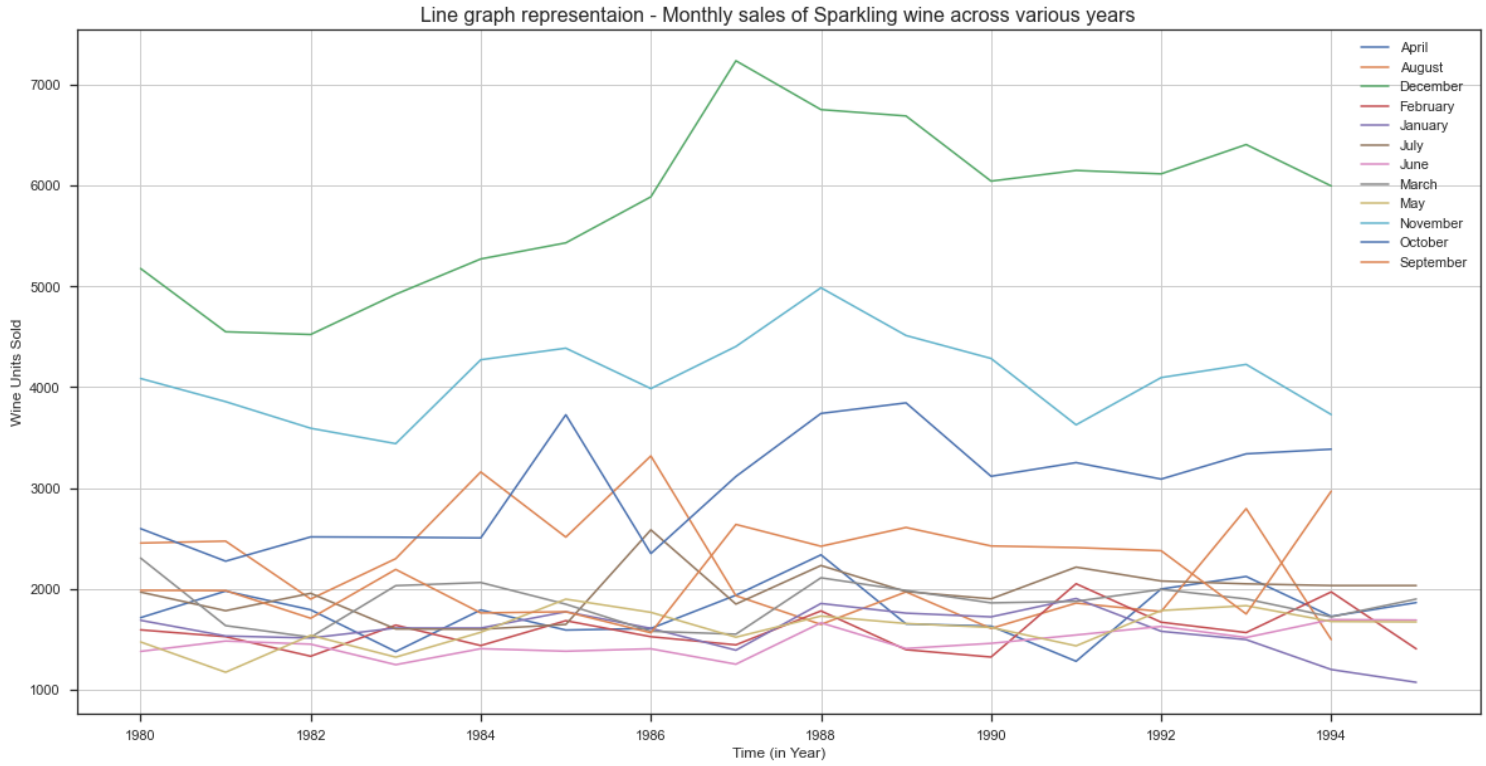


**Fig.117 Line plot – Monthly sales across different years**

## Empirical Cumulative Distribution Plot

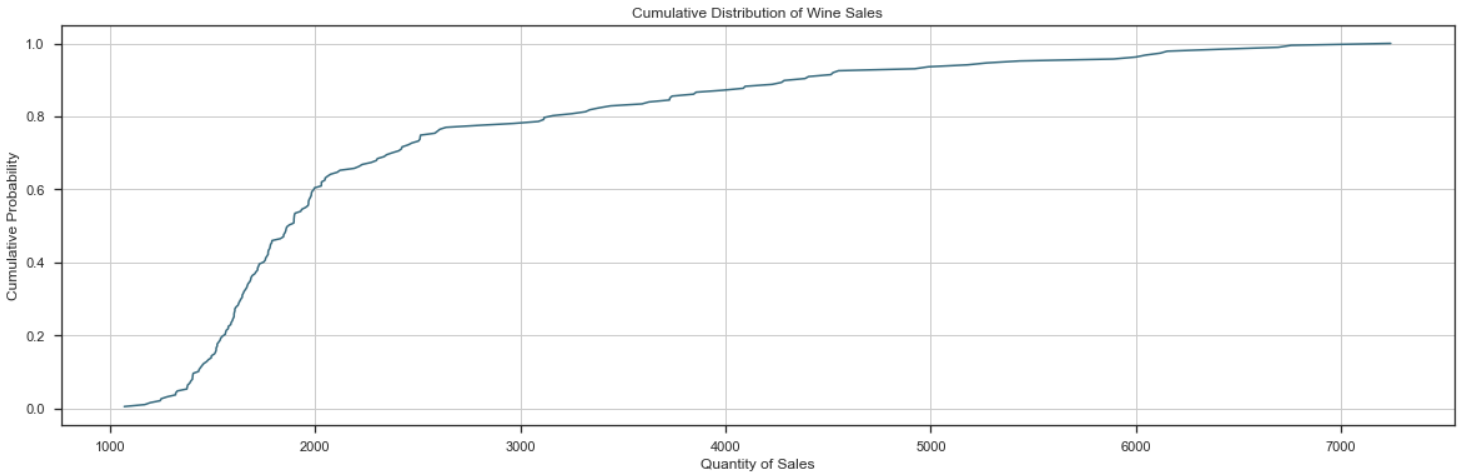


**Fig.118 Line plot – Empirical cumulative distribution function**

## Monthly Time Series Plot

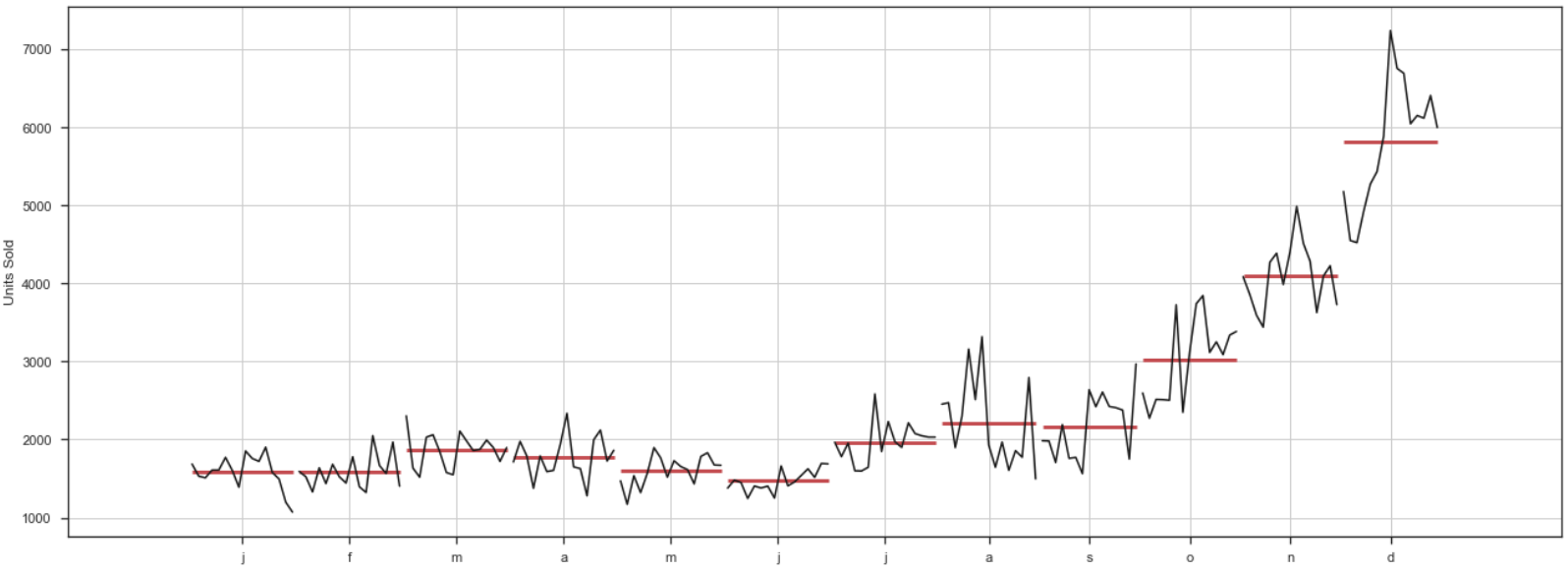


**Fig.119 Time series plot – Monthly time series**

**Observation:**

- Over the years, sales have stayed steady. The sales climbed gradually starting in 1982 until 1988, then decreased until 1990, then slightly increased again until 1994.
- Every year, December has the highest sales, followed by November and October. The first 2 months January and February have the lowest median sales.
- From the cumulative distribution graph, we can observe that around 60 to 70 percent of the units sold are fewer than 2500, and 80% of the units sold are less than 4000. Only 20% of sales involved more than 3000 items. Therefore, it is clear that the bulk of sales were in the range of 1000 to 3000 units.

## Average Wine sales per month & change percentage over each month

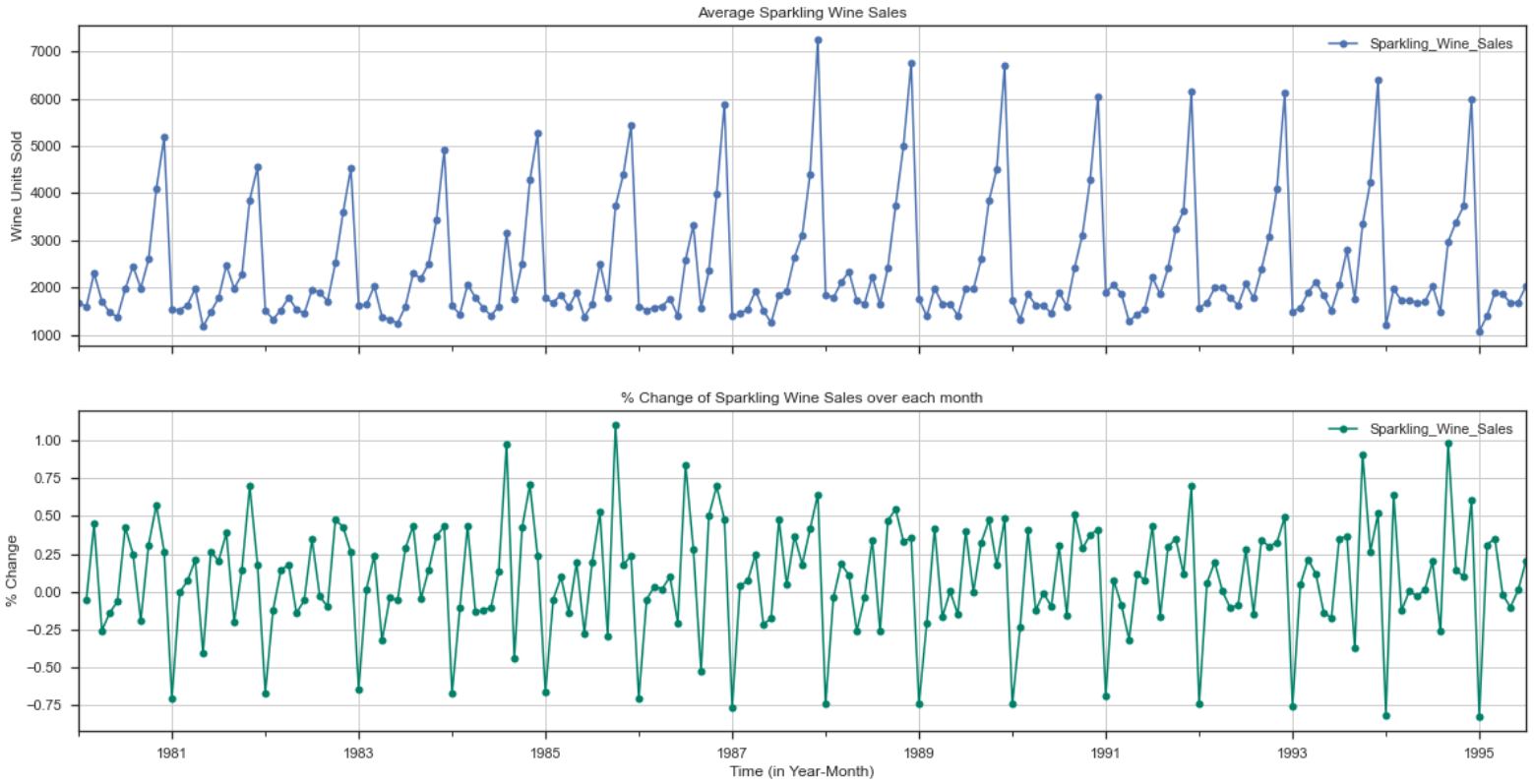


**Fig.120 Line plot – Average and % Change over each month**

**Observation:**

- We can see that there is a no trend but only seasonality from the average sales and % change plots. Additionally, the seasonality in the percentage change appears to be consistent throughout all the years.

## Decomposition of Time Series

### Additive Decomposition

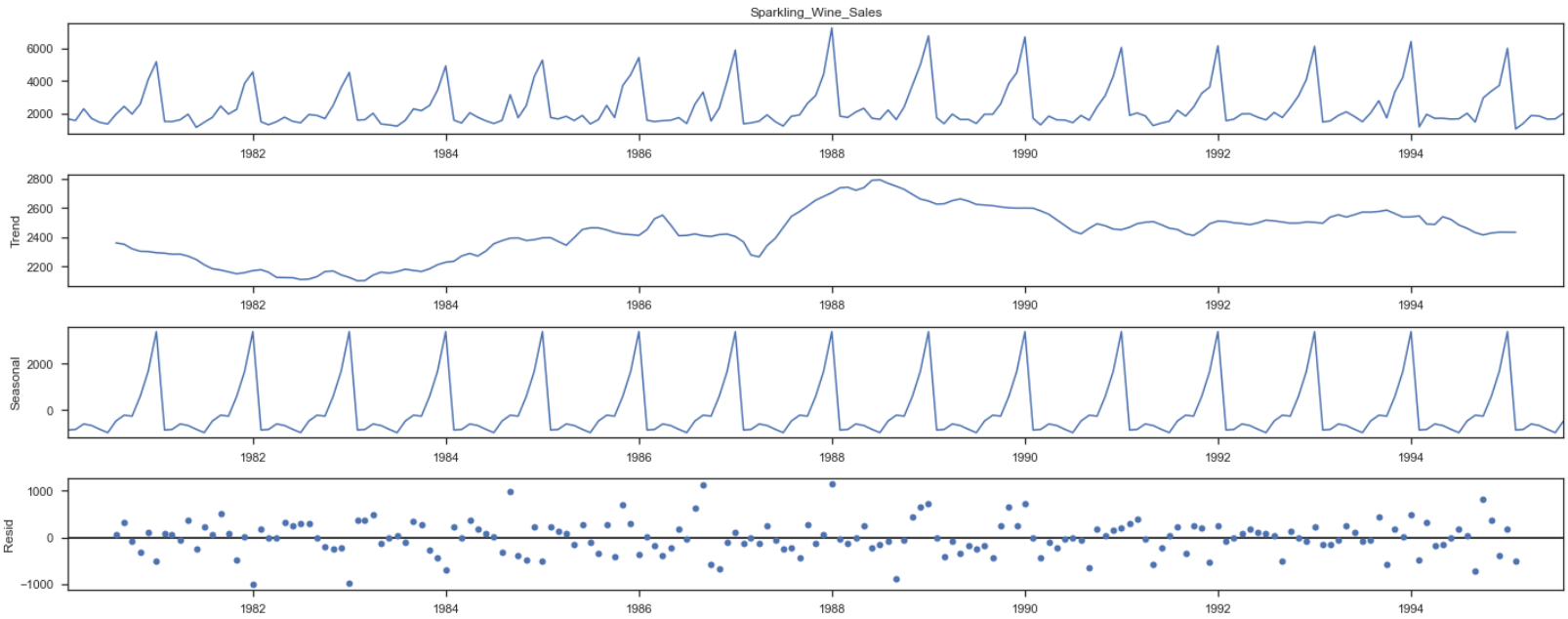


**Fig.121 Additive decomposition of time series**

**Trend**

| Time_Stamp | |
|---|---|
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 2360.666667 |
| 1980-08-31 | 2351.333333 |
| 1980-09-30 | 2320.541667 |
| 1980-10-31 | 2303.583333 |
| 1980-11-30 | 2302.041667 |
| 1980-12-31 | 2293.791667 |

Name: trend, dtype: float64

**Seasonality**

| Time_Stamp | |
|---|---|
| 1980-01-31 | -854.260599 |
| 1980-02-29 | -830.350678 |
| 1980-03-31 | -592.356630 |
| 1980-04-30 | -658.490559 |
| 1980-05-31 | -824.416154 |
| 1980-06-30 | -967.434011 |
| 1980-07-31 | -465.502265 |
| 1980-08-31 | -214.332821 |
| 1980-09-30 | -254.677265 |
| 1980-10-31 | 599.769957 |
| 1980-11-30 | 1675.067179 |
| 1980-12-31 | 3386.983846 |

Name: seasonal, dtype: float64

**Residual**

| Time_Stamp | |
|---|---|
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 70.835599 |
| 1980-08-31 | 315.999487 |
| 1980-09-30 | -81.864401 |
| 1980-10-31 | -307.353290 |
| 1980-11-30 | 109.891154 |
| 1980-12-31 | -501.775513 |

Name: resid, dtype: float64

**Fig.122 Additive Decomposition - Sample of Trend, Seasonality & Residual values**

## Multiplicative Decomposition

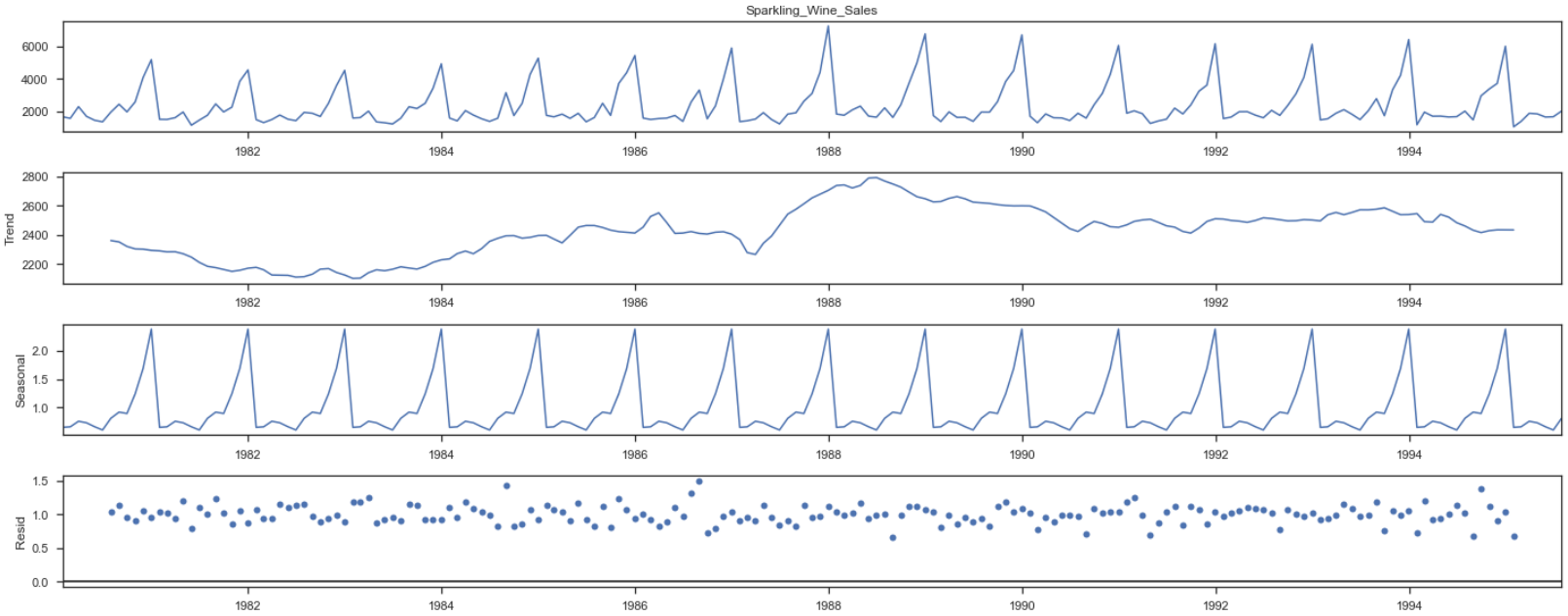


**Fig.123 Multiplicative decomposition of time series**

**Trend**

| Time_Stamp | |
|---|---|
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 2360.666667 |
| 1980-08-31 | 2351.333333 |
| 1980-09-30 | 2320.541667 |
| 1980-10-31 | 2303.583333 |
| 1980-11-30 | 2302.041667 |
| 1980-12-31 | 2293.791667 |

Name: trend, dtype: float64

**Seasonality**

| Time_Stamp | |
|---|---|
| 1980-01-31 | 0.649843 |
| 1980-02-29 | 0.659214 |
| 1980-03-31 | 0.757440 |
| 1980-04-30 | 0.730351 |
| 1980-05-31 | 0.660609 |
| 1980-06-30 | 0.603468 |
| 1980-07-31 | 0.809164 |
| 1980-08-31 | 0.918822 |
| 1980-09-30 | 0.894367 |
| 1980-10-31 | 1.241789 |
| 1980-11-30 | 1.690158 |
| 1980-12-31 | 2.384776 |

Name: seasonal, dtype: float64

**Residual**

| Time_Stamp | |
|---|---|
| 1980-01-31 | NaN |
| 1980-02-29 | NaN |
| 1980-03-31 | NaN |
| 1980-04-30 | NaN |
| 1980-05-31 | NaN |
| 1980-06-30 | NaN |
| 1980-07-31 | 1.029230 |
| 1980-08-31 | 1.135407 |
| 1980-09-30 | 0.955954 |
| 1980-10-31 | 0.907513 |
| 1980-11-30 | 1.050423 |
| 1980-12-31 | 0.946770 |

Name: resid, dtype: float64

**Fig.124 Multiplicative Decomposition - Sample of Trend, Seasonality & Residual values**

**Observation:**

- The residual patterns after additive decomposition of the time series appear to represent the seasonal element and exhibit substantial variation.
- In the multiplicative decomposition of the time series, it has been observed that the seasonal fluctuation of residuals is under control.
- The size of the seasonal variations doesn't change on comparison, but the residuals are tightly controlled by the multiplicative decomposition. In addition to this, the residuals are not independent of seasonality thus we may assume that it is **multiplicative**.

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

Train and test data are separated from the provided dataset. Sales data up to 1991 is included in the training data, while data from 1991 through 1995 is used for testing.

First few rows of Training Data

| Time_Stamp | Sparkling_Wine_Sales |
|---|---|
| 1980-01-31 | 1686 |
| 1980-02-29 | 1591 |
| 1980-03-31 | 2304 |
| 1980-04-30 | 1712 |
| 1980-05-31 | 1471 |
| 1980-06-30 | 1377 |
| 1980-07-31 | 1966 |
| 1980-08-31 | 2453 |
| 1980-09-30 | 1984 |
| 1980-10-31 | 2596 |

Last few rows of Training Data

| Time_Stamp | Sparkling_Wine_Sales |
|---|---|
| 1990-03-31 | 1859 |
| 1990-04-30 | 1628 |
| 1990-05-31 | 1615 |
| 1990-06-30 | 1457 |
| 1990-07-31 | 1899 |
| 1990-08-31 | 1605 |
| 1990-09-30 | 2424 |
| 1990-10-31 | 3116 |
| 1990-11-30 | 4286 |
| 1990-12-31 | 6047 |

**Fig.125 First and Last few rows of Train data**

First few rows of Test Data

| Time_Stamp | Sparkling_Wine_Sales |
|---|---|
| 1991-01-31 | 1902 |
| 1991-02-28 | 2049 |
| 1991-03-31 | 1874 |
| 1991-04-30 | 1279 |
| 1991-05-31 | 1432 |
| 1991-06-30 | 1540 |
| 1991-07-31 | 2214 |
| 1991-08-31 | 1857 |
| 1991-09-30 | 2408 |
| 1991-10-31 | 3252 |

Last few rows of Test Data

| Time_Stamp | Sparkling_Wine_Sales |
|---|---|
| 1994-10-31 | 3385 |
| 1994-11-30 | 3729 |
| 1994-12-31 | 5999 |
| 1995-01-31 | 1070 |
| 1995-02-28 | 1402 |
| 1995-03-31 | 1897 |
| 1995-04-30 | 1862 |
| 1995-05-31 | 1670 |
| 1995-06-30 | 1688 |
| 1995-07-31 | 2031 |

**Fig.126 First and Last few rows of Test data**

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

**Fig.127 Count summary on train and test data**

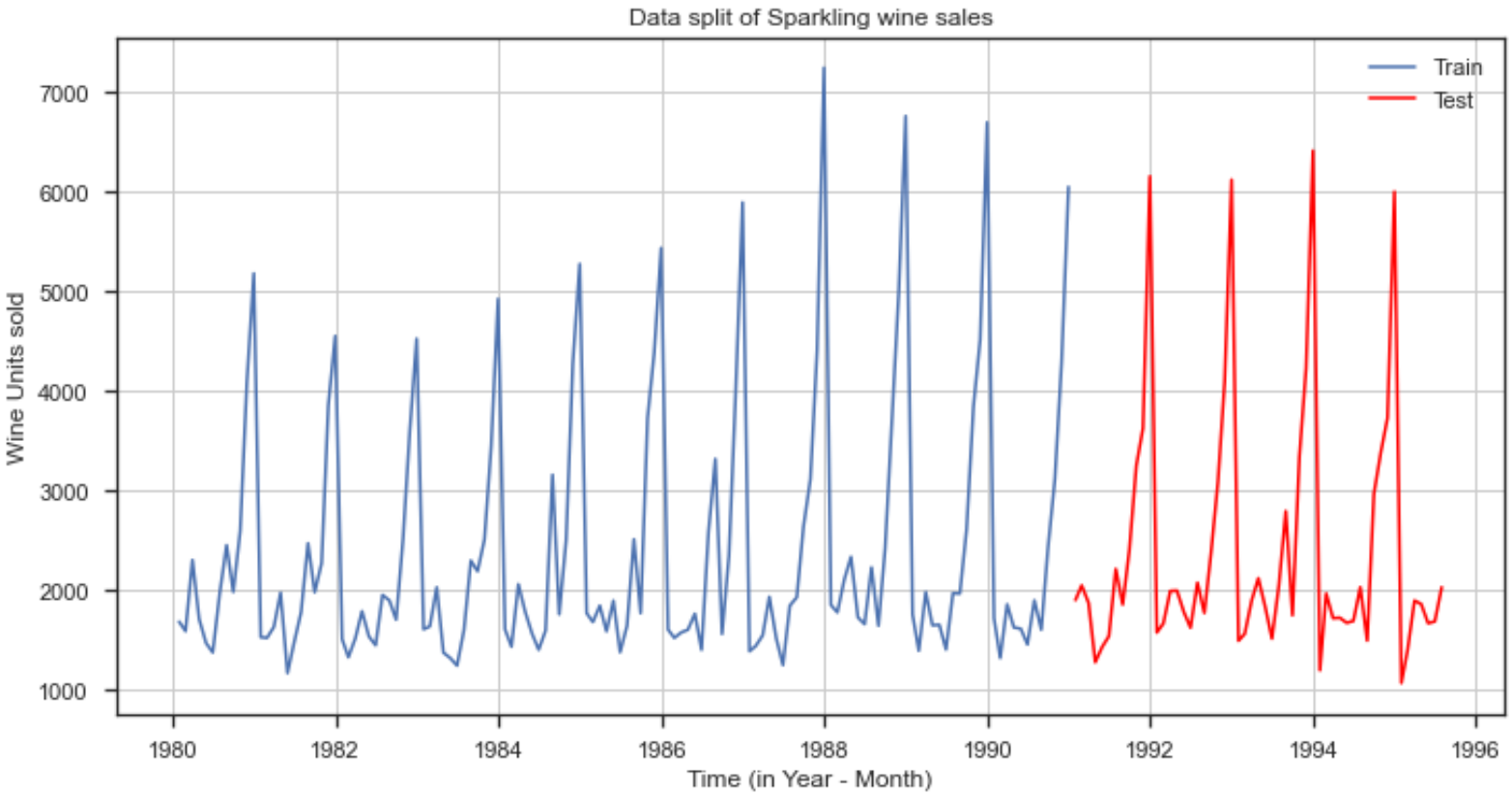


**Fig.128 Line Plot – Splitting of time series into Train & Test data**

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

### Model 1 – Linear Regression

**For this particular linear regression, we are going to regress the 'Sparkling_Wine_Sales' variable against the order of the occurrence.**

**For the selection criteria, the below Linear Regression model is built by using default parameters.**

```
▾ LinearRegression
LinearRegression()
```

**Fig.129 Sparkling Wine – Linear regression model**

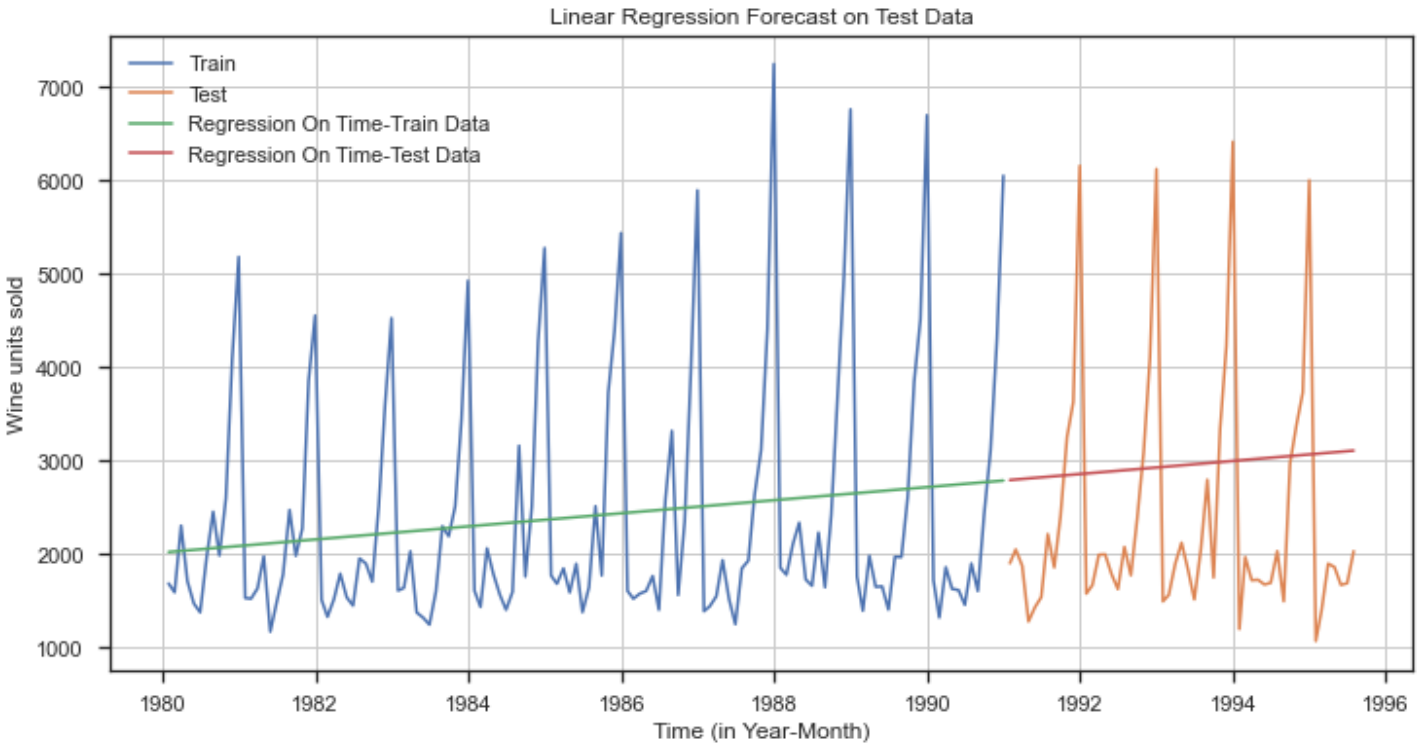


**Fig.130 Linear regression on Test data**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- The train and test data **trends have been caught** by the linear regression model however, it is **unable to account for seasonality**
- The root means squared error **(RMSE)** for the linear regression model is **1389.135**. The size of the seasonal

## Linear Regression: Model Evaluation

| Performance Metric | |
|---|---|
| **Test RMSE** | **1389.135175** |

## Model 2 – Naïve Forecast

**For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.**

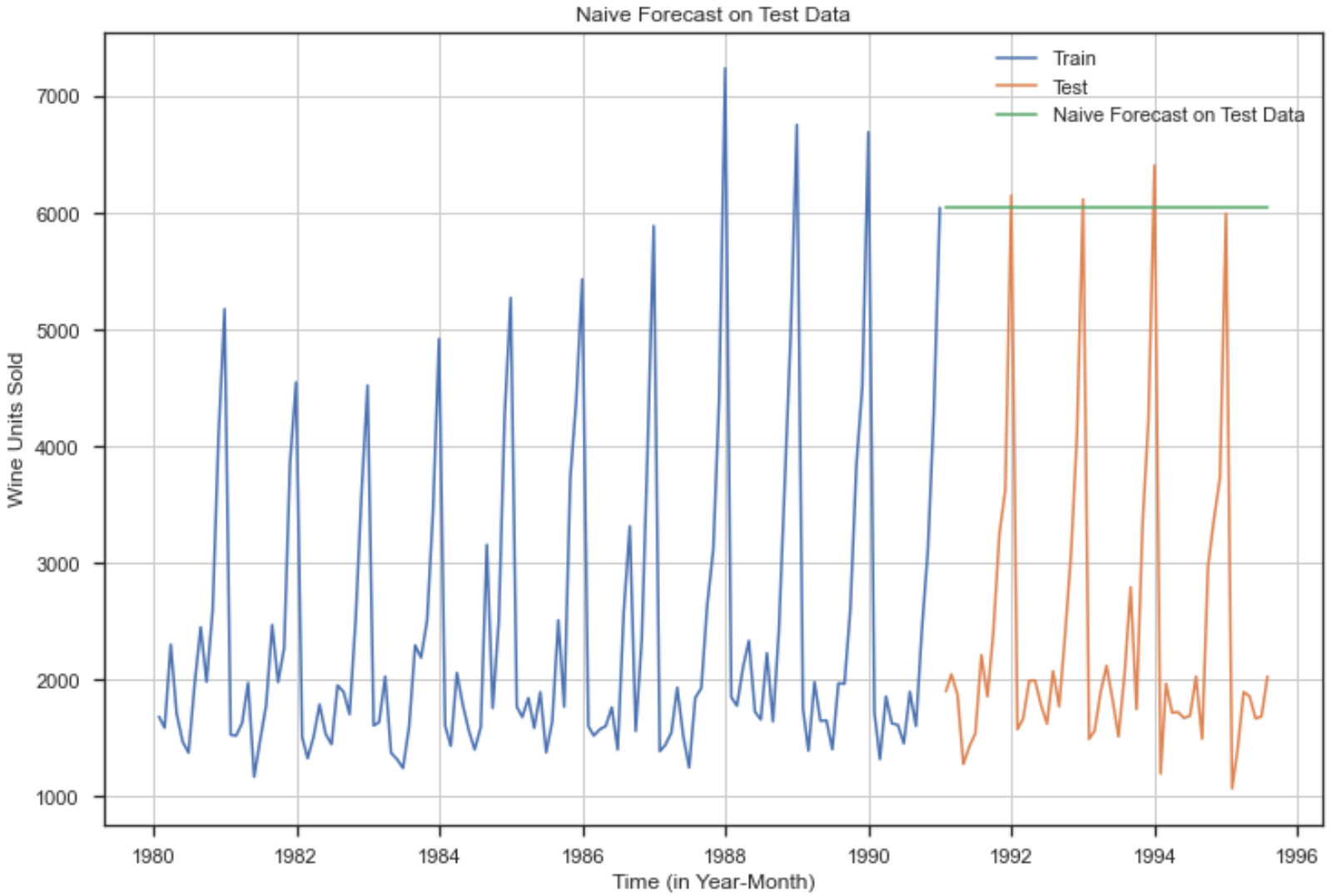


**Fig.131 Naïve forecast on Test data**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- The **seasonality and trend** of the time series data **cannot be captured** by the naive forecast model.
- The root means squared error **(RMSE)** for the naïve forecast model is **3864.279** which is significantly higher than the regression model.

## Naïve Forecast: Model Evaluation

| Performance Metric | |
|---|---|
| **Test RMSE** | **3864.279352** |

## Model 3 – Simple Average

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

| Time_Stamp | Sparkling_Wine_Sales | mean_forecast |
|---|---|---|
| 1991-01-31 | 1902 | 2403.780303 |
| 1991-02-28 | 2049 | 2403.780303 |
| 1991-03-31 | 1874 | 2403.780303 |
| 1991-04-30 | 1279 | 2403.780303 |
| 1991-05-31 | 1432 | 2403.780303 |

**Fig.132 Sparkling Wine – Simple Average model**

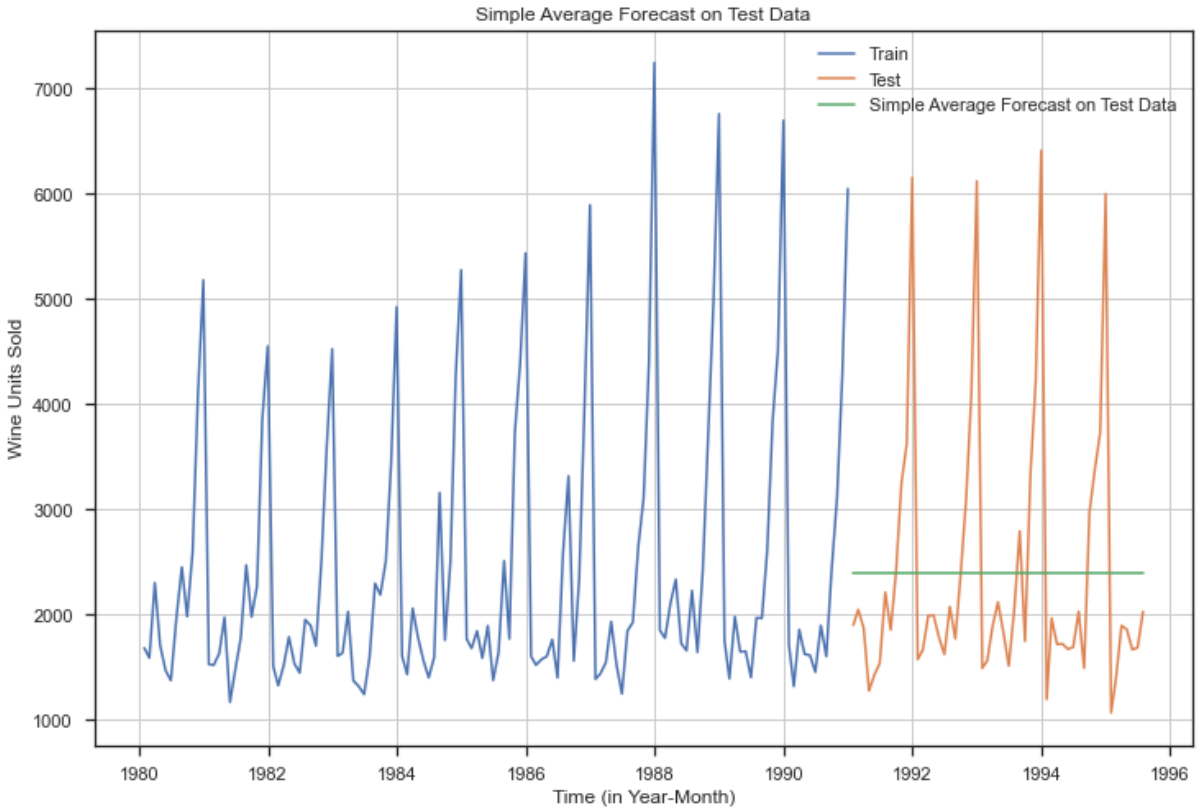


**Fig.133 Simple Average model predictions on Test data**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- The **seasonality and trend** of the time series data **cannot be captured** by the simple average model.
- The root means squared error **(RMSE)** for the simple average model is **1275.081** which is significantly lower than the naïve forecast model and slightly lower than Linear regression model.

## Simple Average: Model Evaluation

| Performance Metric | |
|---|---|
| **Test RMSE** | **1275.081804** |

## Model 4 – 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.**

| Time_Stamp | Sparkling_Wine_Sales | Trailing_2 | Trailing_4 | Trailing_6 | Trailing_9 |
|---|---|---|---|---|---|
| 1980-01-31 | 1686 | NaN | NaN | NaN | NaN |
| 1980-02-29 | 1591 | 1638.5 | NaN | NaN | NaN |
| 1980-03-31 | 2304 | 1947.5 | NaN | NaN | NaN |
| 1980-04-30 | 1712 | 2008.0 | 1823.25 | NaN | NaN |
| 1980-05-31 | 1471 | 1591.5 | 1769.50 | NaN | NaN |
| 1980-06-30 | 1377 | 1424.0 | 1716.00 | 1690.166667 | NaN |
| 1980-07-31 | 1966 | 1671.5 | 1631.50 | 1736.833333 | NaN |
| 1980-08-31 | 2453 | 2209.5 | 1816.75 | 1880.500000 | NaN |
| 1980-09-30 | 1984 | 2218.5 | 1945.00 | 1827.166667 | 1838.222222 |

**Fig.134 Sparkling Wine – Sample of Trailing Moving Averages**

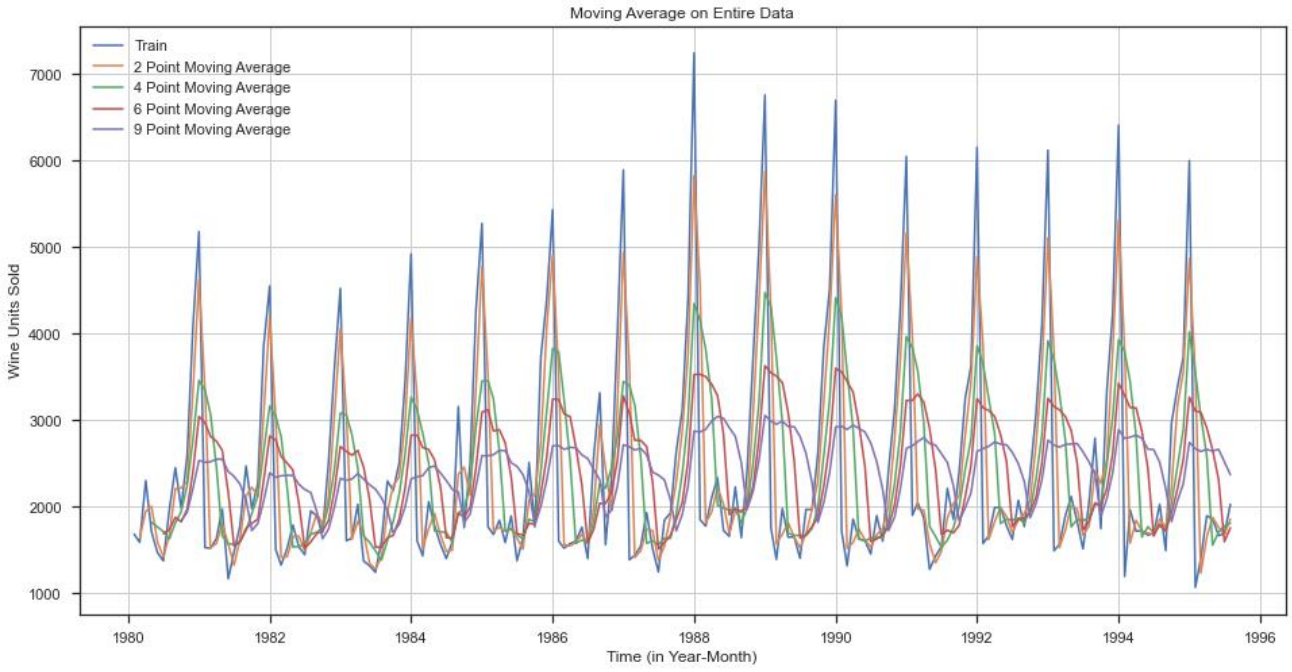


**Fig.135 Moving Average on Entire data**

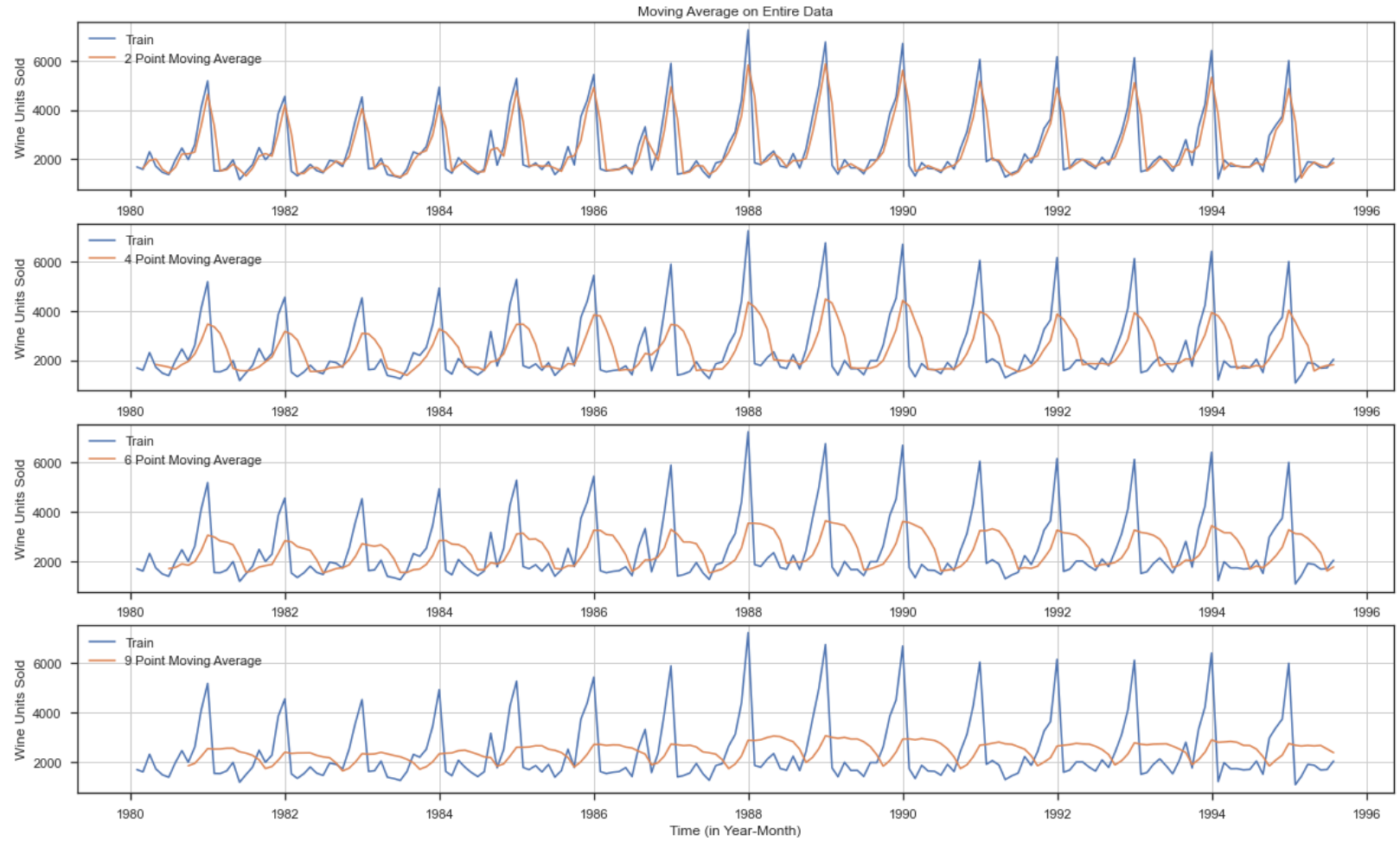


**Fig.136 Individual visualization of moving averages on entire data**

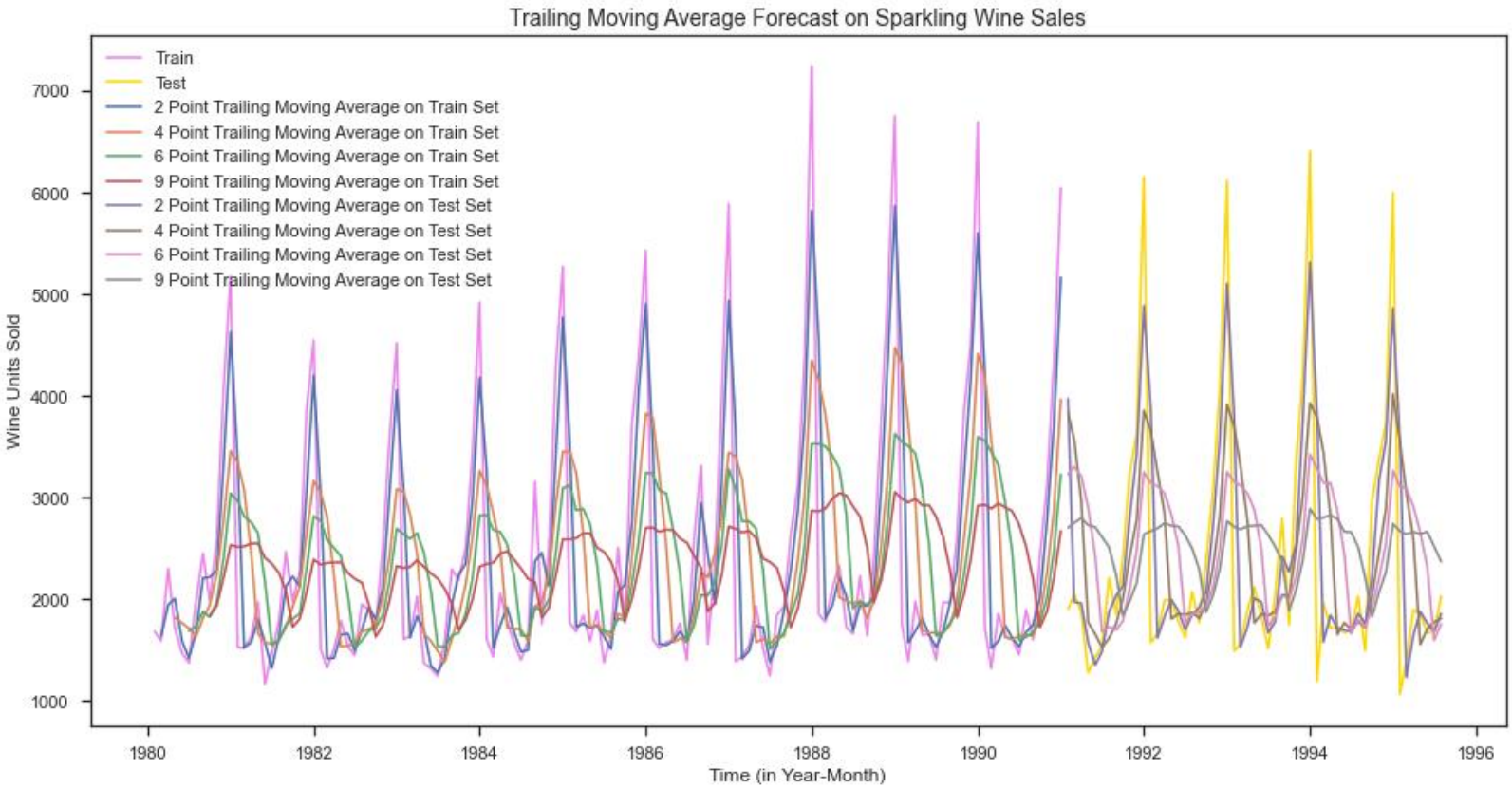


**Fig.137 Moving averages forecast on test data**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- The **seasonality and trend** of the time series data **may both be predicted** using moving average models.
- We can see how the data smooth out as the number of observation points taken increases. The **2-point TMA has characteristics that are more similar to test results** than the 9-point TMA.
- The root means squared error **(RMSE)** for the **2-point trailing average model is 813.4,** which is lowest than all models build so far.

## Moving Average: Model Evaluation

| Model | Test RMSE |
|---|---|
| 2 Point Trailing Moving Average | 813.400684 |
| 4 Point Trailing Moving Average | 1156.589694 |
| 6 Point Trailing Moving Average | 1283.927428 |
| 9 Point Trailing Moving Average | 1346.278315 |

Let's compare the visualization of each model's predictions that we have constructed so far before investigating exponential smoothing methods.

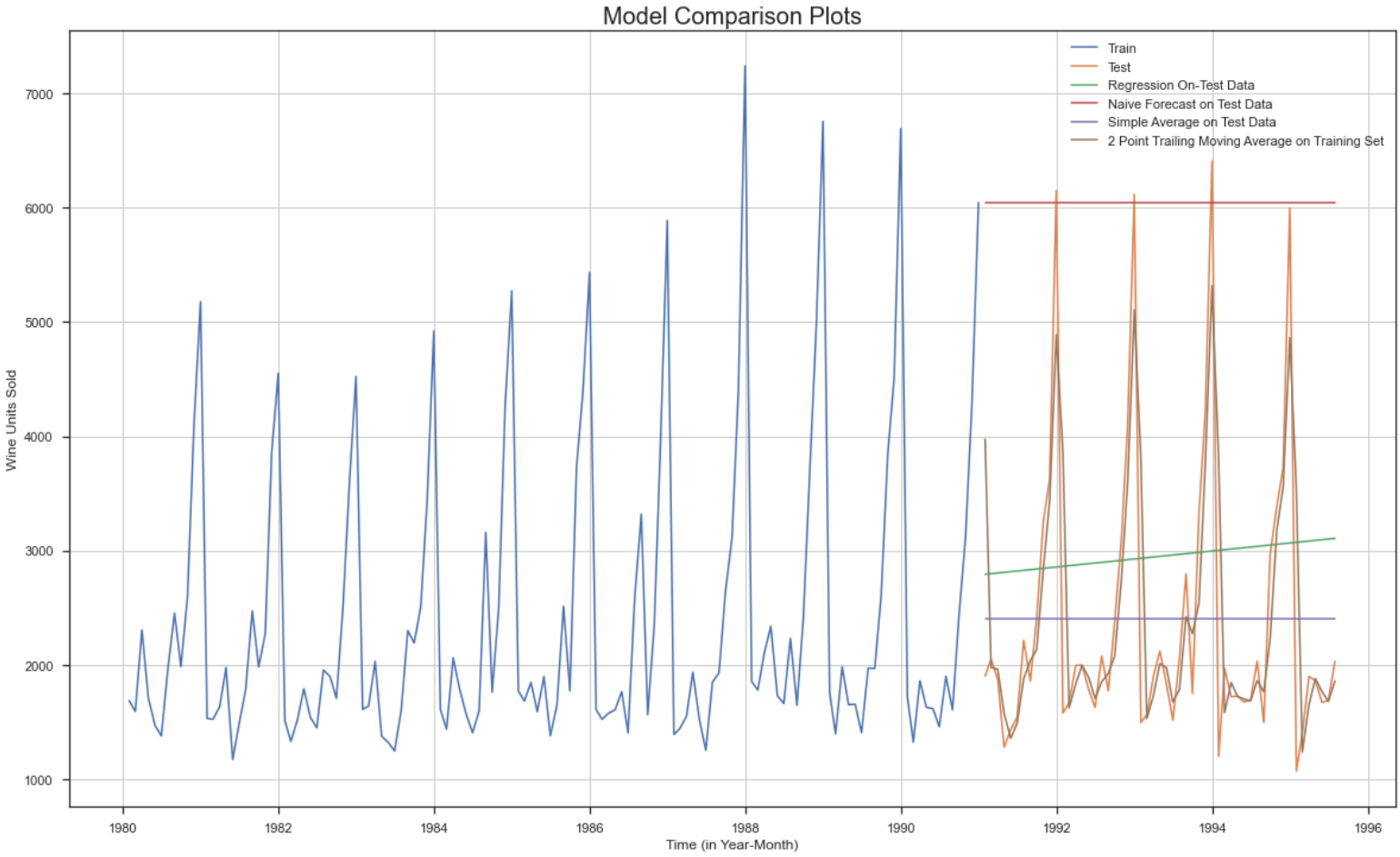


**Fig.138 Comparison of different models on test data (Regression, Naïve, Simple and Moving Average)**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- We can see from the graph above that simple average and naive forecast models fail to adequately describe the characteristics of the test data.
- The trend portion of the series has been caught using linear regression, however the seasonality has been missed
- Both trend and seasonality may be accounted for using moving average models

## Model 5 – Simple Exponential Smoothing

The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES). This method is suitable for forecasting data with no clear trend or seasonal pattern.

**In Single ES, the forecast at time (t + 1) is given by Winters,1960**

$$F_{t+1}=\alpha Y_t + (1-\alpha)F_t$$

**Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.**

**For the selection criteria, the below Simple Exponential Smoothing is built by using optimized parameters.**

```
{'smoothing_level': 0.04960736049406556,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 2151.614314422547,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

**Fig.139 Sparkling Wine – Simple Exponential Smoothing Model**

| Time_Stamp | Sparkling_Wine_Sales | predict |
|---|---|---|
| 1991-01-31 | 1902 | 2725.336037 |
| 1991-02-28 | 2049 | 2725.336037 |
| 1991-03-31 | 1874 | 2725.336037 |
| 1991-04-30 | 1279 | 2725.336037 |
| 1991-05-31 | 1432 | 2725.336037 |

**Fig.140 Sample of SES predictions**

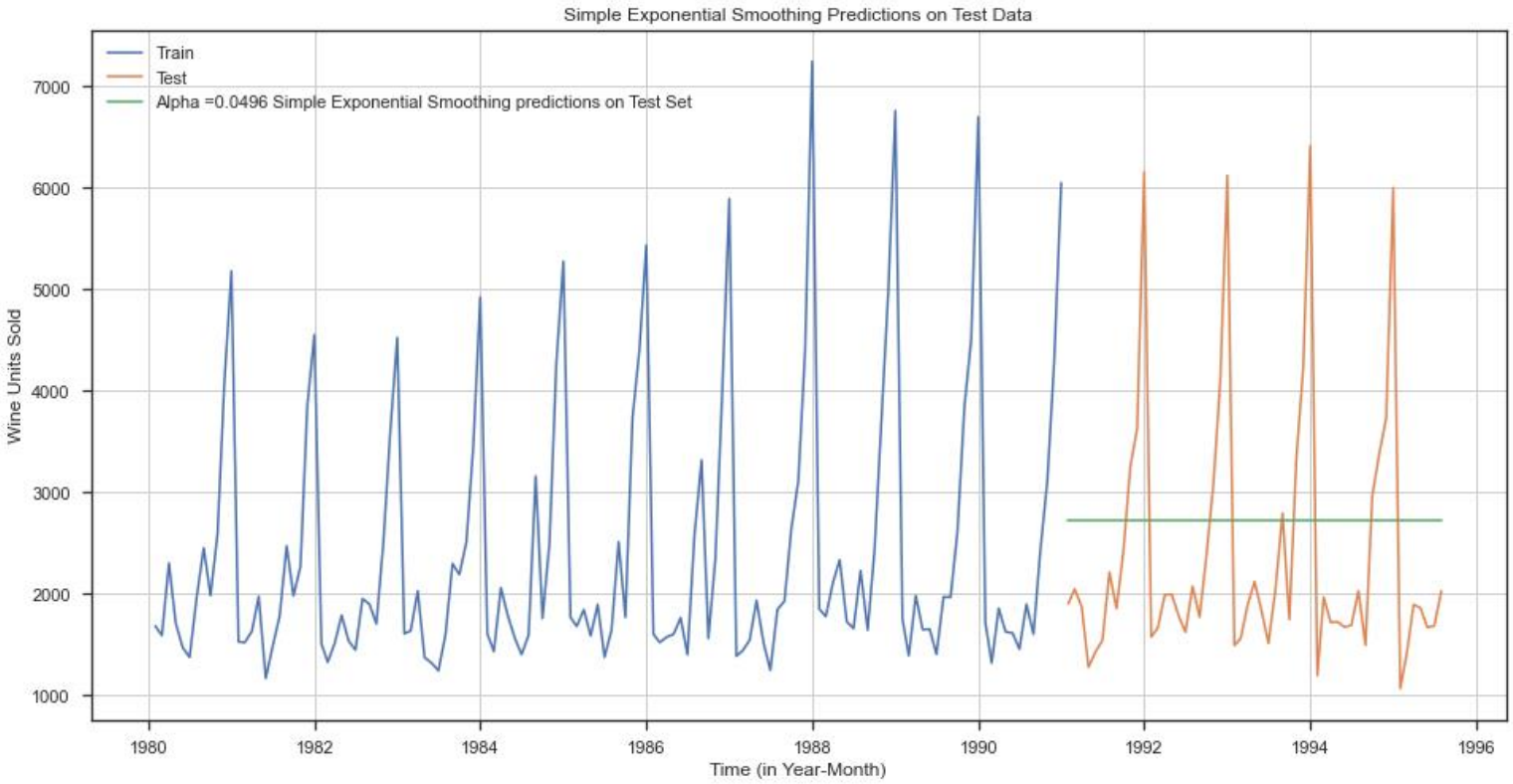


**Fig.141 Sparkling Wine - SES predictions on Test data**

**The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.**

**The range of alpha value is from 0.1 to 0.95 and the respective RMSE for train and test data are calculated for analyzing the performance metrics.**

| | Alpha Values | Train RMSE | Test RMSE |
|---|---|---|---|
| 0 | 0.10 | 1333.873836 | 1375.393398 |
| 1 | 0.15 | 1347.521016 | 1466.203651 |
| 2 | 0.20 | 1356.042987 | 1595.206839 |
| 3 | 0.25 | 1359.701408 | 1755.488175 |
| 4 | 0.30 | 1359.511747 | 1935.507132 |
| 5 | 0.35 | 1356.733677 | 2123.914871 |
| 6 | 0.40 | 1352.588879 | 2311.919615 |
| 7 | 0.45 | 1348.095362 | 2493.786514 |
| 8 | 0.50 | 1344.004369 | 2666.351413 |
| 9 | 0.55 | 1340.811249 | 2828.246418 |
| 10 | 0.60 | 1338.805381 | 2979.204388 |
| 11 | 0.65 | 1338.131249 | 3119.560885 |
| 12 | 0.70 | 1338.844308 | 3249.944092 |
| 13 | 0.75 | 1340.955212 | 3371.100106 |
| 14 | 0.80 | 1344.462091 | 3483.801006 |
| 15 | 0.85 | 1349.373283 | 3588.797654 |
| 16 | 0.90 | 1355.723518 | 3686.794285 |
| 17 | 0.95 | 1363.586057 | 3778.432623 |

**Fig.142 SES prediction metrics for different alpha values**

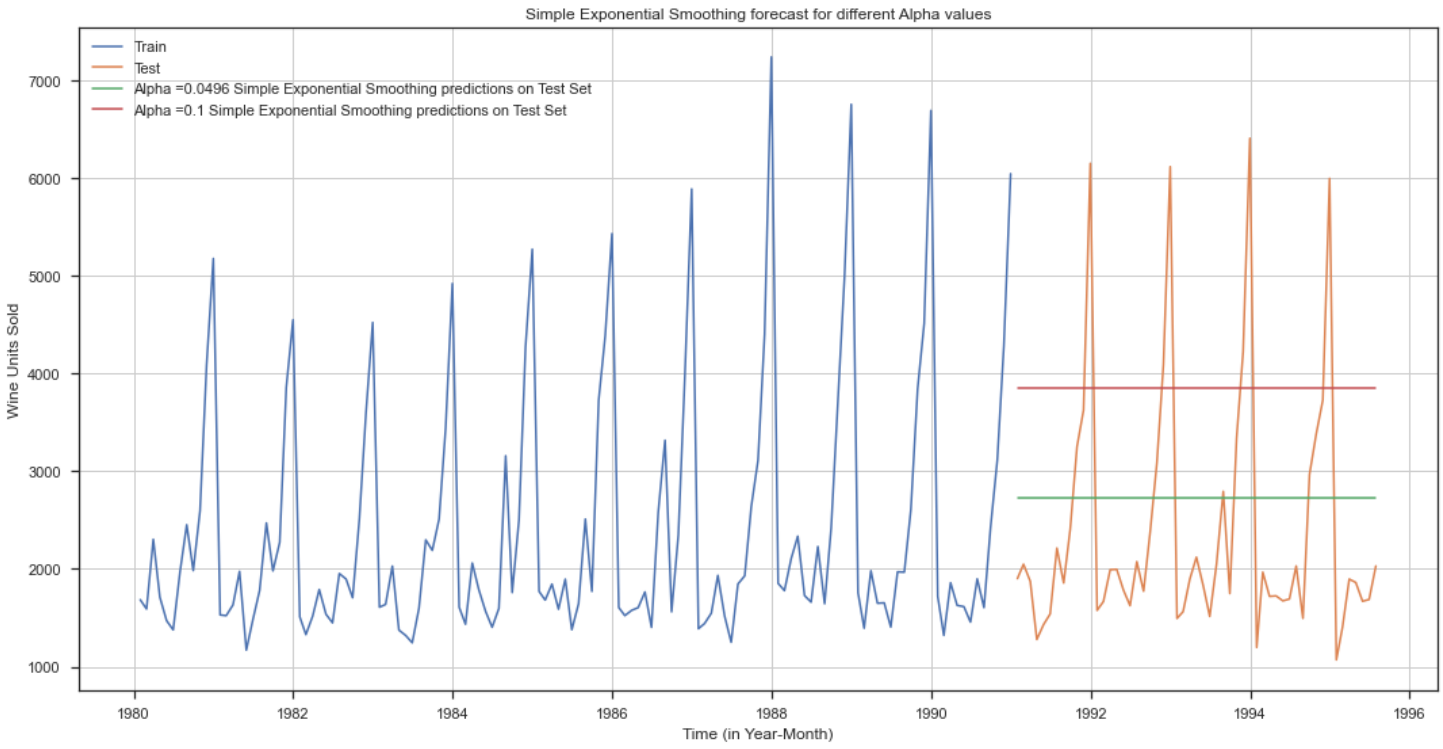


**Fig.143 SES forecast for different Alpha values**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- When there is **neither a trend nor a seasonal component to the time series, simple exponential smoothing is typically used**. It is due to this reason, it unable to capture the characteristics of the time series data.
- The root means squared error **(RMSE)** for the simple exponential smoothing model with **Alpha =0.0496** is **1316.135**and for **Alpha=0.1, RMSE is 1375.393.**
- **The Simple Exponential Smoothing with alpha=0.0496 is taken as the best model among two as it has the lowest test RMSE.**

**Simple Exponential Smoothing: Model Evaluation**

| Model | Test RMSE |
|---|---|
| SES (Alpha = 0.0496) | 1316.135411 |
| SES (Alpha = 0.1) | 1375.393398 |

## Model 6 – Double Exponential Smoothing (Holt's Model)

**This model is an extension of SES known as Double Exponential model which estimates two smoothing parameters. Applicable when data has Trend but no seasonality. Two separate components are considered: Level and Trend. Level is the local mean. One smoothing parameter α corresponds to the level series. A second smoothing parameter β corresponds to the trend series.**

**Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short-term average value or level and the other for capturing the trend.**

**Intercept or Level equation, $L_t$ is given by: $L_t = \alpha Y_t + (1-\alpha)F_t$**

**Trend equation is given by $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$**

**Here, αα and ββ are the smoothing constants for level and trend, respectively,**

**$0 < \alpha < 1$ and $0 < \beta < 1$.**

**The forecast at time t + 1 is given by**

**$F_{t+1} = L_t + T_t$**

**$F_{t+n} = L_t + nT_t$**

**For the selection criteria, the below Double Exponential Smoothing is built by using optimized parameters.**

```
{'smoothing_level': 0.6885714285714285,
 'smoothing_trend': 9.999999999999999e-05,
 '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}
```

**Fig.144 Sparkling Wine – Double Exponential Smoothing Model**

| Time_Stamp | |
|---|---|
| 1991-01-31 | 5221.278699 |
| 1991-02-28 | 5127.886554 |
| 1991-03-31 | 5034.494409 |
| 1991-04-30 | 4941.102264 |
| 1991-05-31 | 4847.710119 |
| 1991-06-30 | 4754.317974 |
| 1991-07-31 | 4660.925829 |
| 1991-08-31 | 4567.533684 |
| 1991-09-30 | 4474.141539 |
| 1991-10-31 | 4380.749394 |
| 1991-11-30 | 4287.357249 |
| 1991-12-31 | 4193.965104 |
| 1992-01-31 | 4100.572959 |
| 1992-02-29 | 4007.180813 |
| 1992-03-31 | 3913.788668 |
| 1992-04-30 | 3820.396523 |

**Fig.145 Sample of DES predictions**

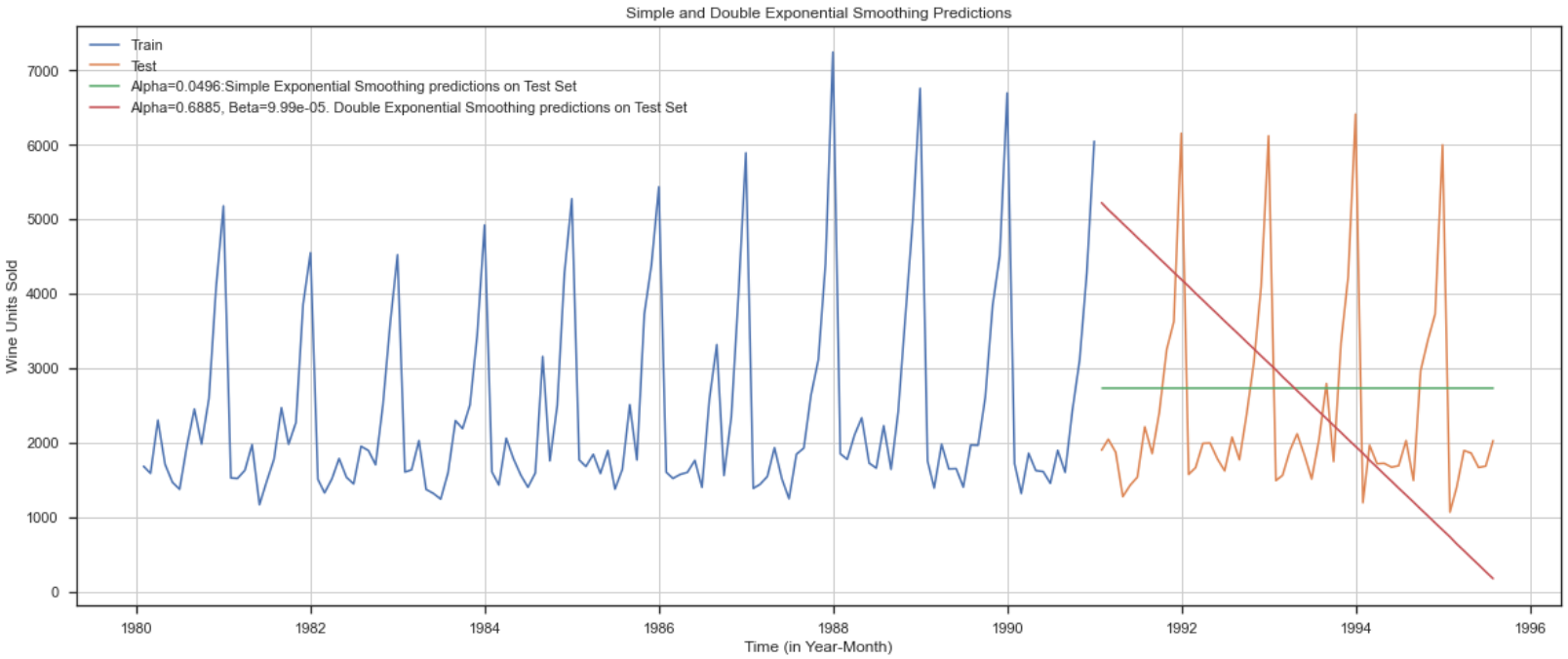


**Fig.146 Sparkling Wine - DES predictions on Test data**

**The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.**

**The range of alpha value is from 0.05 to 1.0 and the respective RMSE for train and test data are calculated for analyzing the performance metrics.**

| | Alpha | Beta | Train RMSE | Test RMSE |
|---|---|---|---|---|
| 0 | 0.05 | 0.05 | 1430.025526 | 1418.407668 |
| 3 | 0.05 | 0.20 | 1382.766405 | 1443.099273 |
| 2 | 0.05 | 0.15 | 1379.162520 | 1457.041594 |
| 1 | 0.05 | 0.10 | 1385.420826 | 1466.899629 |
| 6 | 0.05 | 0.35 | 1414.226231 | 1547.022626 |

**Fig.147 DES prediction metrics for different alpha, beta values**

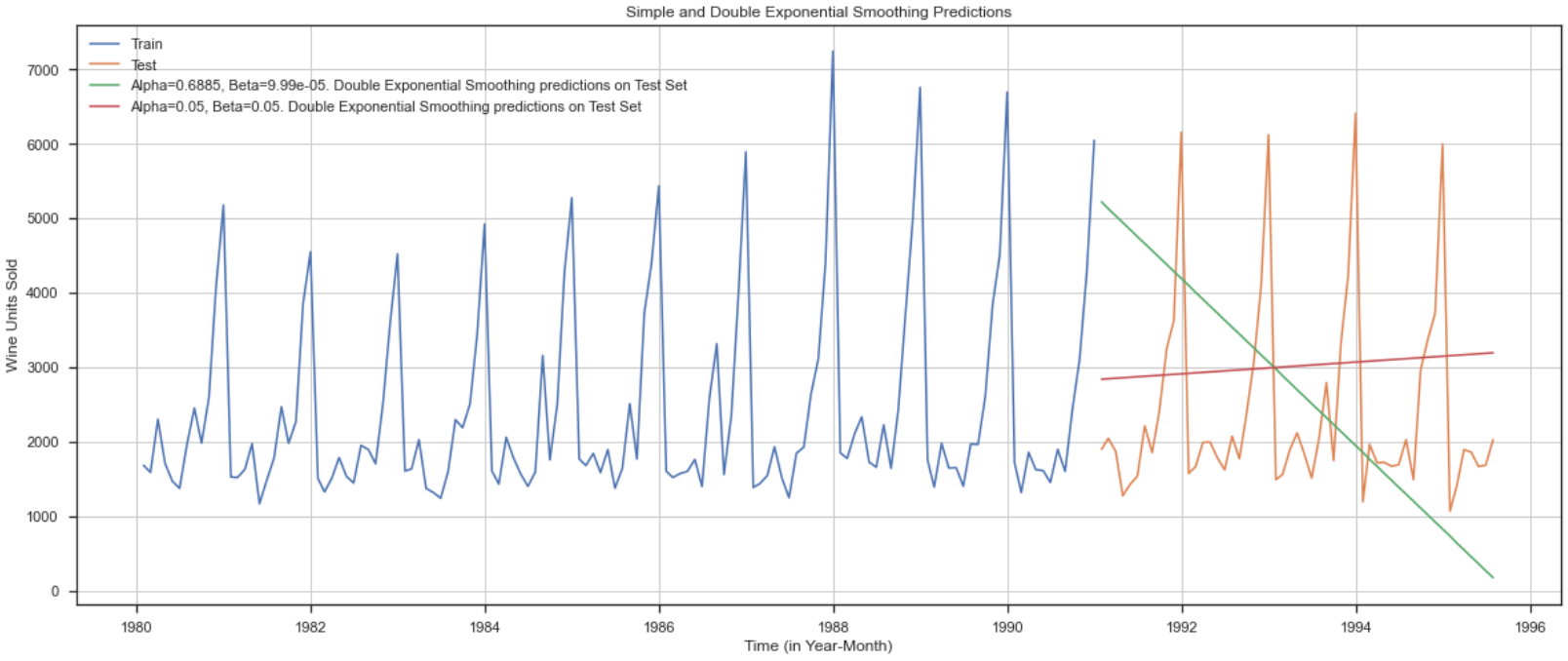


**Fig.148 DES forecast for different Alpha, Beta values**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- When there is simply trend and no seasonality in the time series data, the double exponential smoothing model performs well. It is due to this reason it is only able to capture the trend characteristics of the data and seasonality is not accounted for.
- The root means squared error **(RMSE)** for the double exponential smoothing model with **Alpha=0.6885, Beta=9.99e-05** is **2007.238**and for **Alpha=0.05, Beta=0.05 (Auto tuned model)**, **RMSE** is **1418.407.**
- **The Double Exponential Smoothing with Alpha=0.05, Beta=0.05 is taken as the best model among two as it has the lowest test RMSE.**
- Additionally, it should be highlighted that compared to the simple exponential smoothing model, the double exponential smoothing model has slightly higher RMSE.

## Double Exponential Smoothing: Model Evaluation

| Model | Test RMSE |
|---|---|
| DES (Alpha=0.6885, Beta=9.99e-05) | 2007.238526 |
| DES (Alpha=0.05, Beta=0.05) | 1418.407668 |

## Model 7 – Triple Exponential Smoothing (Holt-Winter's Model)

**This model is an extension of DES known as Triple Exponential Smoothing model which estimates three smoothing parameters. Applicable when data has both Trend and seasonality. Three separate components are considered: Level, Trend and Seasonality.**

**One smoothing parameter α corresponds to the level series.**

**A second smoothing parameter β corresponds to the trend series.**

**A third smoothing parameter γ corresponds to the seasonality series**

**where,**

**0 < α <1,**

**0 < β <1,**

**0 < γ <1**

**For the selection criteria, the below Triple Exponential Smoothing is built by using optimized parameters.**

```
{'smoothing_level': 0.11108840858679117,
 'smoothing_trend': 0.061712060020663685,
 'smoothing_seasonal': 0.3950814802151603,
 'damping_trend': nan,
 'initial_level': 1639.9088356475902,
 'initial_trend': -11.928143593549056,
 'initial_seasons': array([1.05065032, 1.02086214, 1.41078482, 1.20263518, 0.97
315225,
        0.96689379, 1.31724304, 1.70471609, 1.37289733, 1.81035002,
        2.83962708, 3.60997333]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

**Fig.149 Sparkling Wine – Triple Exponential Smoothing Model**

| Time_Stamp | Sparkling_Wine_Sales | auto_predict |
|---|---|---|
| 1991-01-31 | 1902 | 1577.208163 |
| 1991-02-28 | 2049 | 1333.663154 |
| 1991-03-31 | 1874 | 1745.977341 |
| 1991-04-30 | 1279 | 1630.435405 |
| 1991-05-31 | 1432 | 1523.306429 |

**Fig.150 Sample of TES predictions**

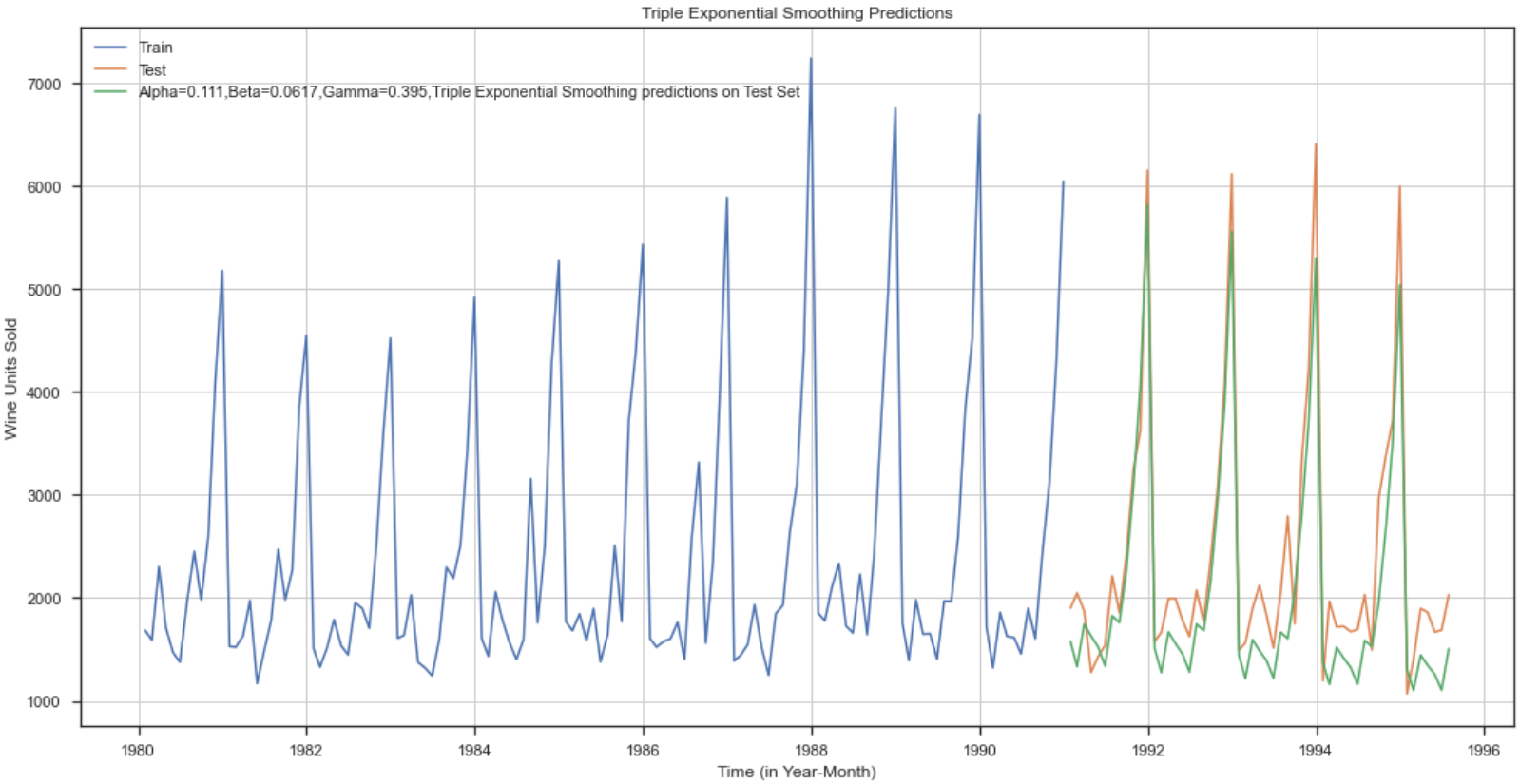


**Fig.151 Sparkling Wine - TES predictions on Test data**

**The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.**

**The range of alpha value is from 0.1 to 1.0 and the respective RMSE for train and test data are calculated for analyzing the performance metrics.**

| | Alpha | Beta | Gamma | Train RMSE | Test RMSE |
|---|---|---|---|---|---|
| **1807** | 0.35 | 0.10 | 0.20 | 386.330654 | 319.498680 |
| **2169** | 0.40 | 0.10 | 0.25 | 388.427700 | 320.164138 |
| **1827** | 0.35 | 0.15 | 0.25 | 390.553942 | 325.137234 |
| **115** | 0.10 | 0.40 | 0.15 | 417.294618 | 325.324779 |
| **1465** | 0.30 | 0.15 | 0.20 | 389.444578 | 326.258972 |

**Fig.152 TES prediction metrics for different alpha, beta and gamma values**

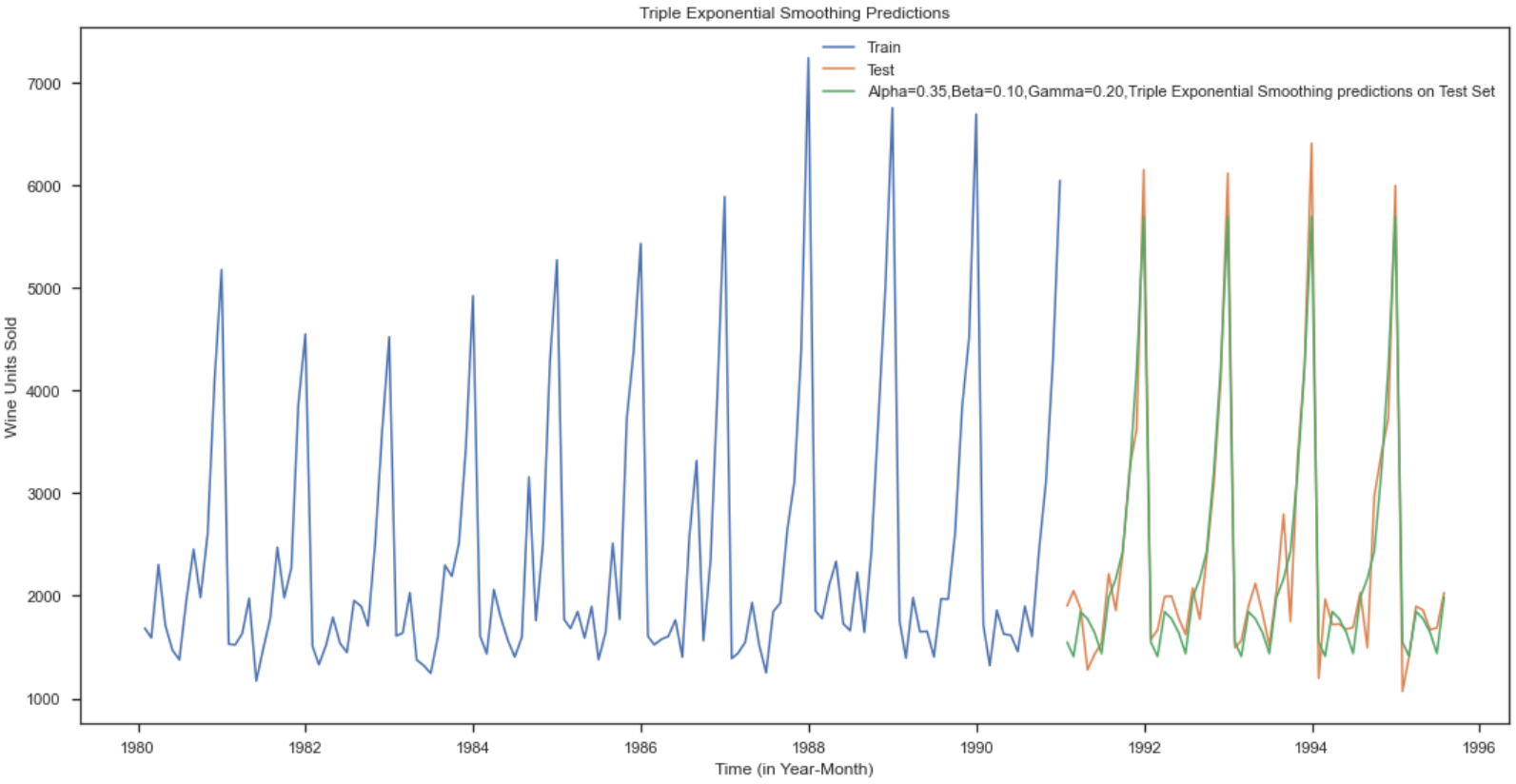


**Fig.153 TES forecast for automated model parameters**

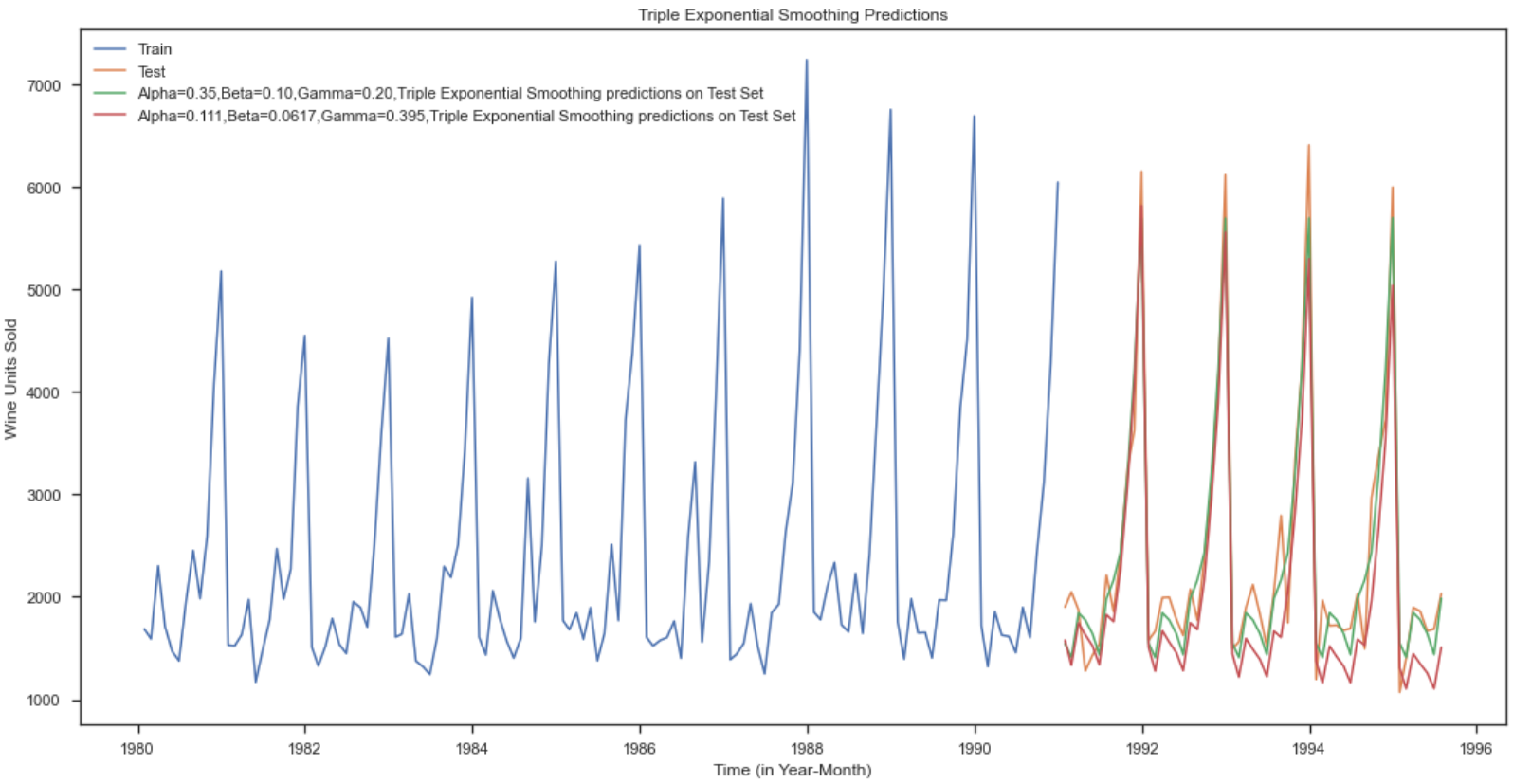


**Fig.154 TES forecast for different model parameters**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- When there is both trend and seasonality in the time series data, the triple exponential model works well. It is due to this reason it able to capture both the trend and seasonal characteristics and nearly match the actual test data plot.
- The root means squared error **(RMSE)** for the double exponential smoothing model with **Alpha=0.111, Beta=0.0617, Gamma=0.395** is **469.659** and for **Alpha=0.35, Beta=0.10, Gamma=0.20 (Auto tuned model)**, **RMSE** is **319.498.**
- **The Triple Exponential Smoothing with Alpha=0.35, Beta=0.10, Gamma=0.20 is taken as the best model among two as it has the lowest test RMSE.**
- Additionally, it should be highlighted that compared to the double exponential smoothing model, the **triple exponential smoothing model has almost reduced the RMSE value by 75%.**

## Triple Exponential Smoothing: Model Evaluation

| Model | Test RMSE |
|---|---|
| TES (Alpha=0.111, Beta=0.0617, Gamma=0.395) | 469.659106 |
| TES (Alpha=0.35, Beta=0.10, Gamma=0.20) | 319.498680 |

Let's compare the RMSE values of the models we have constructed so far and visualize the plot of the best exponential smoothing models thus built.

| | Test RMSE |
|---|---|
| Alpha=0.35,Beta=0.10,Gamma=0.20,Triple Exponential Smoothing | 319.498680 |
| Alpha=0.111,Beta=0.0617,Gamma=0.395,Triple Exponential Smoothing | 469.659106 |
| 2 point TMA | 813.400684 |
| 4 point TMA | 1156.589694 |
| Simple Average | 1275.081804 |
| 6 point TMA | 1283.927428 |
| Alpha =0.0496,SimpleExponentialSmoothing | 1316.135411 |
| 9 point TMA | 1346.278315 |
| Linear Regression | 1389.135175 |
| Alpha=0.05, Beta=0.05, Double Exponential Smoothing | 1418.407668 |
| Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing | 2007.238526 |
| Naive Model | 3864.279352 |

**Fig.155 Comparison of Test RMSE values of different exponential smoothing models**

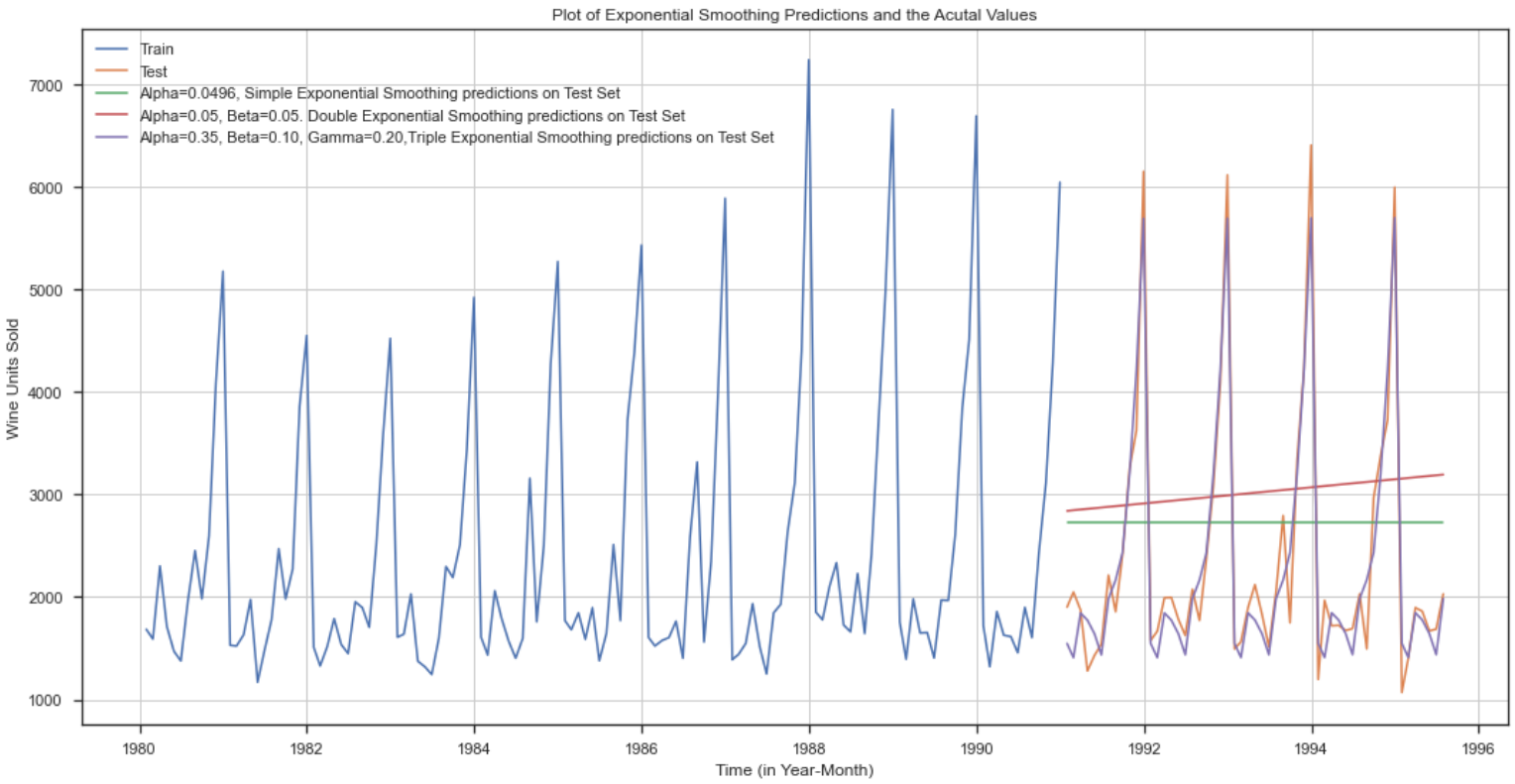


**Fig.156 Comparison of different models on test data (SES, DES and TES)**

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- **Simple exponential smoothing** is frequently employed when the time series doesn't include a trend or a seasonal component. This is the reason why it is unable to capture the time series data's features.
- The **double exponential smoothing** model works effectively when the time series data just contains trend and no seasonality. This explains why seasonality is not taken into consideration and just the trend features of the data are captured.
- The **triple exponential model** performs effectively when the time series data exhibit both trend and seasonality. This is the reason why it is essentially identical to the test data plot and is able to capture both the trend and seasonal aspects.
- The **Triple exponential model is the best model we have built so far as it has the lowest RMSE value**.

## 5) Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

### Checking for Stationarity of Entire Data

The **Augmented Dickey-Fuller test** is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

**Framing the hypothesis:**

**$H_0$:** The Time Series has a unit root and is thus non-stationary.

**$H_1$:** The Time Series does not have a unit root and is thus stationary.

**The series have to be stationary for building ARIMA/SARIMA models and thus we would want the p-value of this test to be less than the α value.**

```
Results of Dicky-Fuller Test
DF test statistic is -1.798
DF test p-value is 0.7055958459932692
Number of lags used 12
```

**Fig.157 Sparkling Wine – ADF summary**

**Inference:**

**We see that at 5% significant level the Time Series is non-stationary as p-value is 0.705 which is more than alpha value (0.05), therefore we fail to reject the null hypothesis. Let us take one level of differencing to see whether the series becomes stationary.**

```
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
```

**Fig.158 Sparkling Wine – ADF summary with differencing**

**Inference:**

**We see that at 5% significant level the Time Series becomes stationary as p-value is nearly 0 which is less than alpha value (0.05), therefore we reject the null hypothesis. We can see that the provided time series becomes stationary with differencing.**

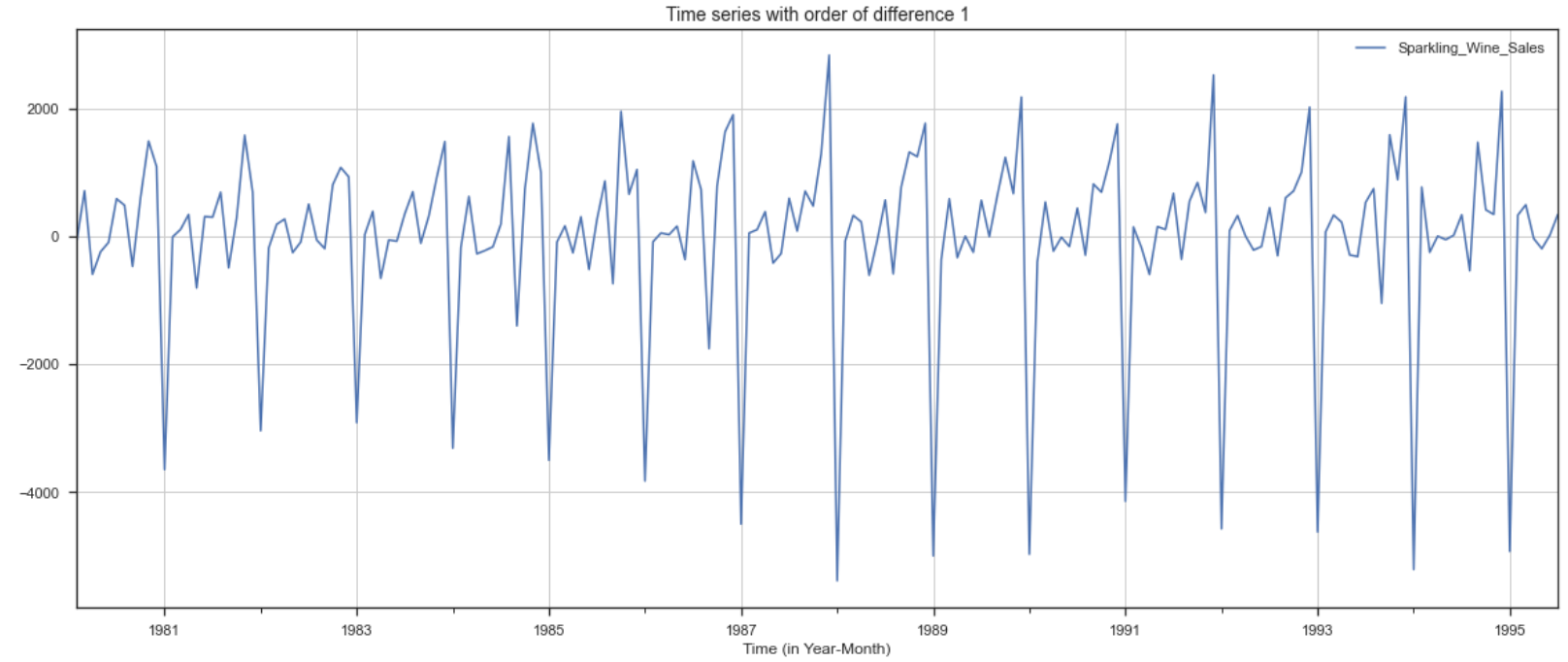


**Fig.159 Time Series Plot of Entire data – With differencing**

## Checking for Stationarity of Training Data

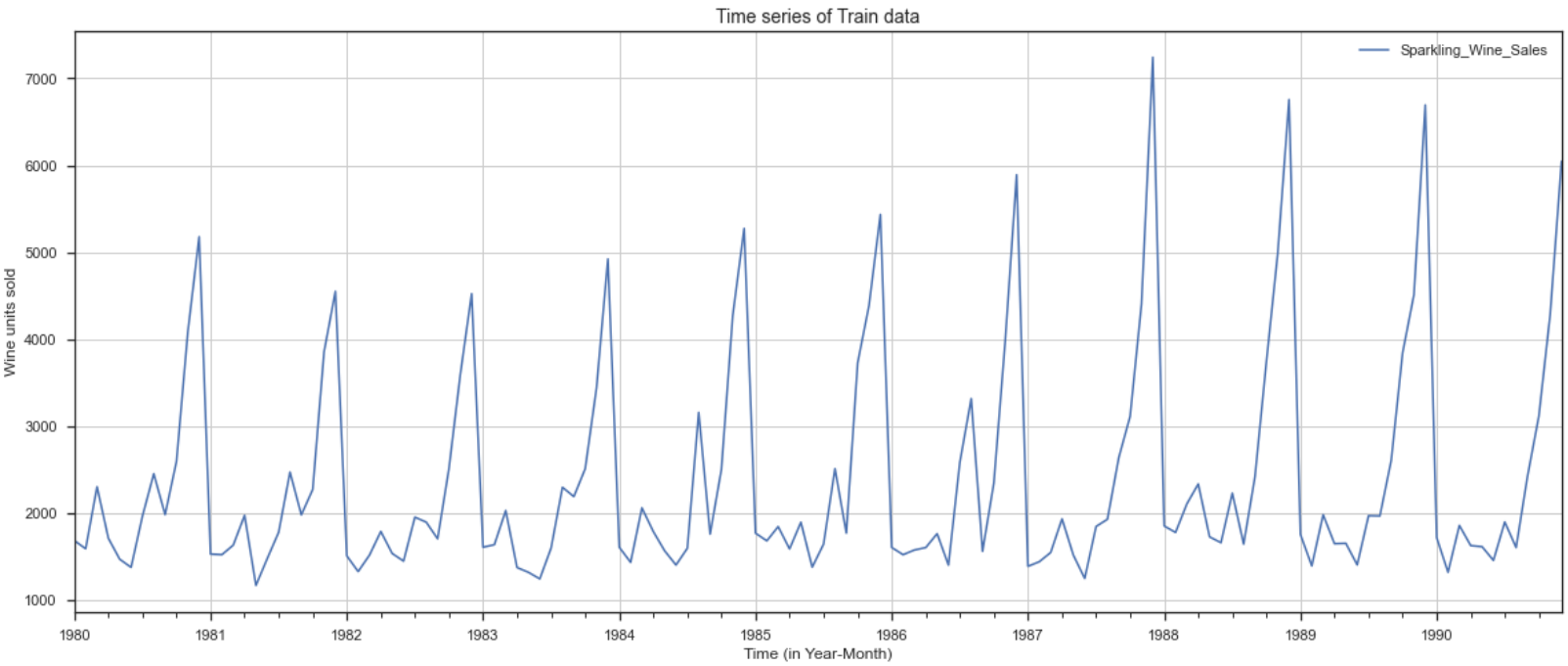


**Fig.160 Time Series Plot of Train data**

```
Results of Dicky-Fuller Test on Train data
DF test statistic is -2.062
DF test p-value is 0.5674110388593684
Number of lags used 12
```

**Fig.161 Sparkling Wine – ADF summary on train data**

**Inference:**

**We see that at 5% significant level the Time Series of training data is non-stationary as p-value is 0.567 which is more than alpha value (0.05), therefore we fail to reject the null hypothesis. Let us take one level of differencing to see whether the series becomes stationary.**

```
Results of Dicky-Fuller Test on Train data with differencing
DF test statistic is -7.968
DF test p-value is 8.47921065551504e-11
Number of lags used 11
```

**Fig.162 Sparkling Wine – ADF summary on train data with differencing**

**Inference:**

**We see that at 5% significant level the Time Series of training data is non-stationary as p-value is 8.479e-11 which is less than alpha value (0.05), therefore we reject the null hypothesis. We can see that the provided training time series becomes stationary with differencing.**

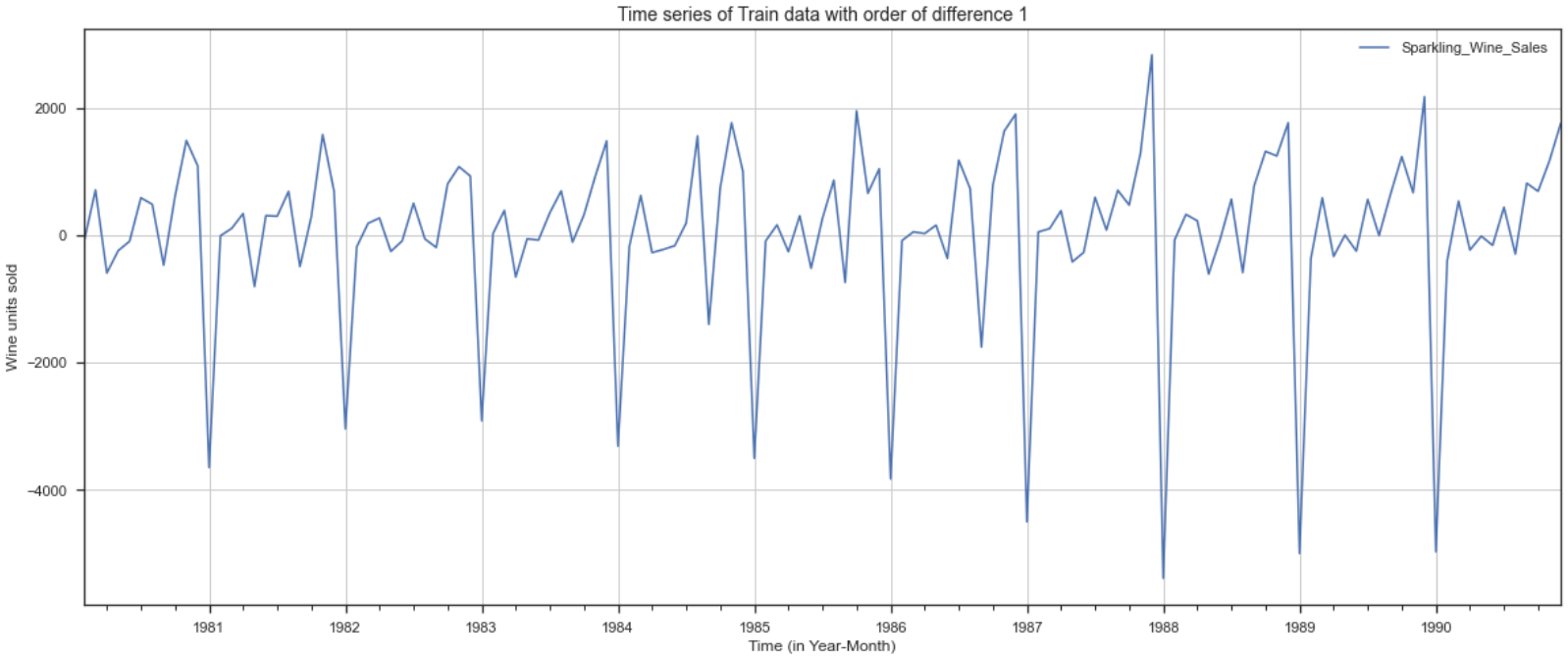


**Fig.163 Time Series Plot of Training data with differencing**

**Observation:**

- As per the Augmented Dicky-Fuller test, we observed that the time series data by itself is not stationary, however, it becomes stationary when differencing is done.
- The same thing is also observed with Training data. Therefore, for training the models, **it can be built with order of difference d=1**.

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

### Model 8 – Auto-Regressive Integrated Moving Average (ARIMA)

**Auto-regression means regression of a variable on itself. One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process. When the time series data is not stationary, then we have to convert the non-stationary time-series data to stationary time-series before applying AR.**

**ARIMA models may be used to represent any "non-seasonal" time series that has patterns and isn't just random noise.**

**An ARIMA model is characterized by 3 terms: p, d, q**

**where,**

**p is the order of the Auto Regressive (AR) term**

**q is the order of the Moving Average (MA) term**

**d is the number of differencing required to make the time series stationary**

**For the selection criteria of p,d,q the below ARIMA model is built by using automated model parameters with lowest Akaike Information Criteria.**

```
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)
```

**Fig.164 Parameter Combinations for ARIMA model**

```
ARIMA(0, 1, 0) - AIC:2267.6630357855465
ARIMA(0, 1, 1) - AIC:2263.060015592223
ARIMA(0, 1, 2) - AIC:2234.4083231226628
ARIMA(0, 1, 3) - AIC:2233.994857735581
ARIMA(0, 1, 4) - AIC:2235.1737364706064
ARIMA(1, 1, 0) - AIC:2266.6085393190097
ARIMA(1, 1, 1) - AIC:2235.7550946704996
ARIMA(1, 1, 2) - AIC:2234.527200452125
ARIMA(1, 1, 3) - AIC:2235.6078154783027
ARIMA(1, 1, 4) - AIC:2227.73697669376
ARIMA(2, 1, 0) - AIC:2260.36574396809
ARIMA(2, 1, 1) - AIC:2233.777626228905
ARIMA(2, 1, 2) - AIC:2213.509212416925
ARIMA(2, 1, 3) - AIC:2232.8112113956195
ARIMA(2, 1, 4) - AIC:2222.9218323369732
ARIMA(3, 1, 0) - AIC:2257.7233789979387
ARIMA(3, 1, 1) - AIC:2235.4988992974854
ARIMA(3, 1, 2) - AIC:2230.825008517658
ARIMA(3, 1, 3) - AIC:2221.4616892285576
ARIMA(3, 1, 4) - AIC:2220.4284290109963
ARIMA(4, 1, 0) - AIC:2259.7418413992646
ARIMA(4, 1, 1) - AIC:2237.0730468632437
ARIMA(4, 1, 2) - AIC:2233.060402105986
ARIMA(4, 1, 3) - AIC:2222.9040968516474
ARIMA(4, 1, 4) - AIC:2213.248094993628
```

**Fig.165 AIC values for different parameter combinations**

| | param | AIC |
|---|---|---|
| **24** | (4, 1, 4) | 2213.248095 |
| **12** | (2, 1, 2) | 2213.509212 |
| **19** | (3, 1, 4) | 2220.428429 |
| **18** | (3, 1, 3) | 2221.461689 |
| **23** | (4, 1, 3) | 2222.904097 |

**Fig.166 Sorted AIC values for different parameter combinations**

**We can see that among all the possible given combinations, the AIC is lowest for the combination (4,1,4). Hence, the model is built with these parameters to determine the RMSE value of test data.**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sparkling_Wine_Sales | No. Observations: | 132 |
| Model: | ARIMA(4, 1, 4) | Log Likelihood | -1097.624 |
| Date: | Sun, 23 Oct 2022 | AIC | 2213.248 |
| Time: | 09:04:54 | BIC | 2239.125 |
| Sample: | 01-31-1980 | HQIC | 2223.763 |
| | - 12-31-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | -0.4452 | 0.109 | -4.087 | 0.000 | -0.659 | -0.232 |
| ar.L2 | -0.4492 | 0.076 | -5.926 | 0.000 | -0.598 | -0.301 |
| ar.L3 | -0.4463 | 0.088 | -5.091 | 0.000 | -0.618 | -0.275 |
| ar.L4 | 0.5500 | 0.068 | 8.126 | 0.000 | 0.417 | 0.683 |
| ma.L1 | -0.0044 | 7.181 | -0.001 | 1.000 | -14.079 | 14.070 |
| ma.L2 | 0.0181 | 14.247 | 0.001 | 0.999 | -27.905 | 27.942 |
| ma.L3 | -0.0328 | 6.920 | -0.005 | 0.996 | -13.595 | 13.529 |
| ma.L4 | -0.9809 | 0.156 | -6.287 | 0.000 | -1.287 | -0.675 |
| sigma2 | 9.083e+05 | 3.05e-05 | 2.98e+10 | 0.000 | 9.08e+05 | 9.08e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.11 | Jarque-Bera (JB): | 0.68 |
| Prob(Q): | 0.74 | Prob(JB): | 0.71 |
| Heteroskedasticity (H): | 2.83 | Skew: | 0.17 |
| Prob(H) (two-sided): | 0.00 | Kurtosis: | 3.06 |

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.37e+27. Standard errors may be unstable.

**Fig.167 Sparkling Wine – Automated ARIMA model**

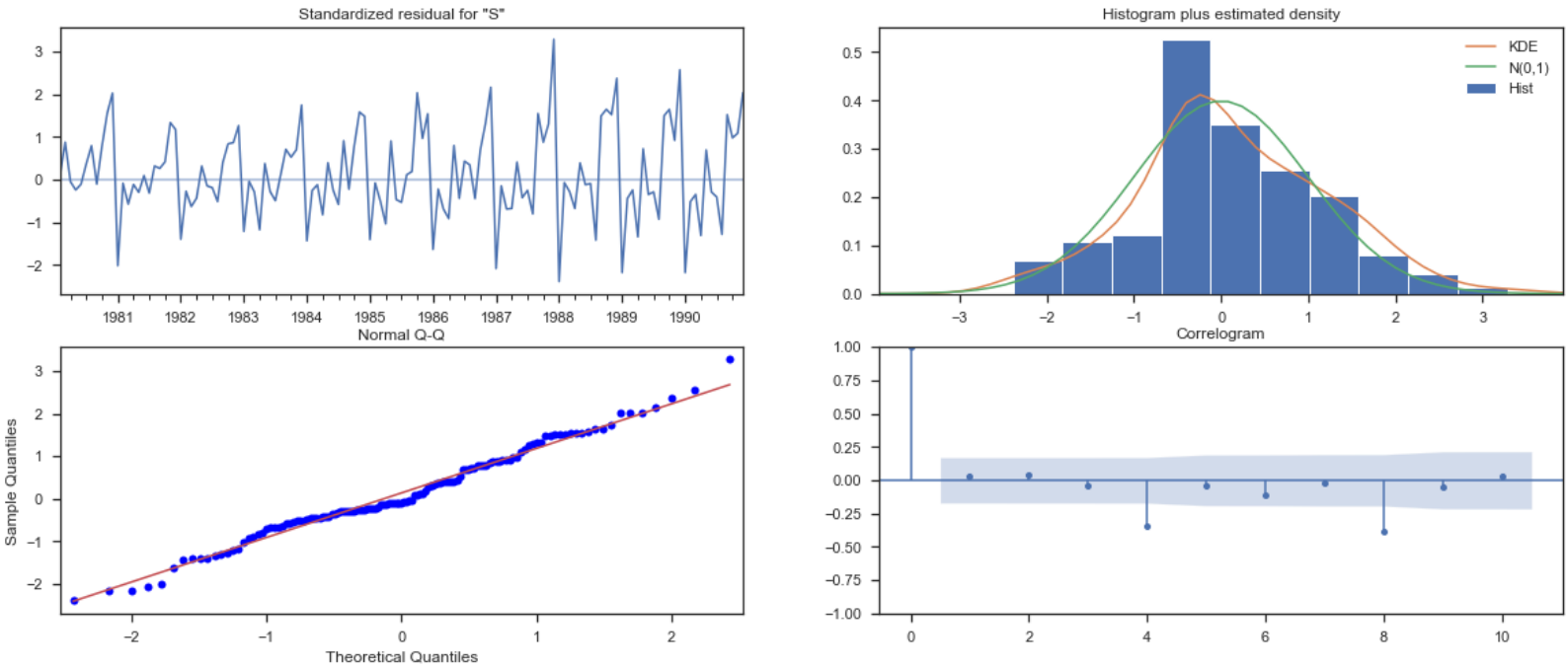


**Fig.168 Automated ARIMA – Diagnostics plot**

**Observation:**

- The optimal parameters are decided based on the lowest Akaike Information Criteria (AIC) values. **The AIC is lowest for the combination (4,1,4) as we see from the above results.**
- From the **Standardized residual plot** above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.
- The **histogram plus estimated density plot** suggests a slightly uniform distribution with mean zero and slightly skewed to the right.
- In **Normal Q-Q plot**, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.
- The **correlogram plot** of residuals shows that the residuals are not auto correlated.

| | |
|---|---|
| 1991-01-31 | 3468.444084 |
| 1991-02-28 | 2778.371560 |
| 1991-03-31 | 3048.890971 |
| 1991-04-30 | 3477.524913 |
| 1991-05-31 | 2055.049246 |
| 1991-06-30 | 1995.522088 |
| 1991-07-31 | 2618.507190 |
| 1991-08-31 | 3238.532757 |
| 1991-09-30 | 1926.882716 |
| 1991-10-31 | 1921.498368 |
| 1991-11-30 | 2579.012565 |
| 1991-12-31 | 3215.136883 |
| 1992-01-31 | 1917.592779 |
| 1992-02-29 | 1913.061609 |
| 1992-03-31 | 2575.663594 |
| 1992-04-30 | 3211.697354 |

**Fig.169 Sample of Automated ARIMA (4,1,4) predictions**

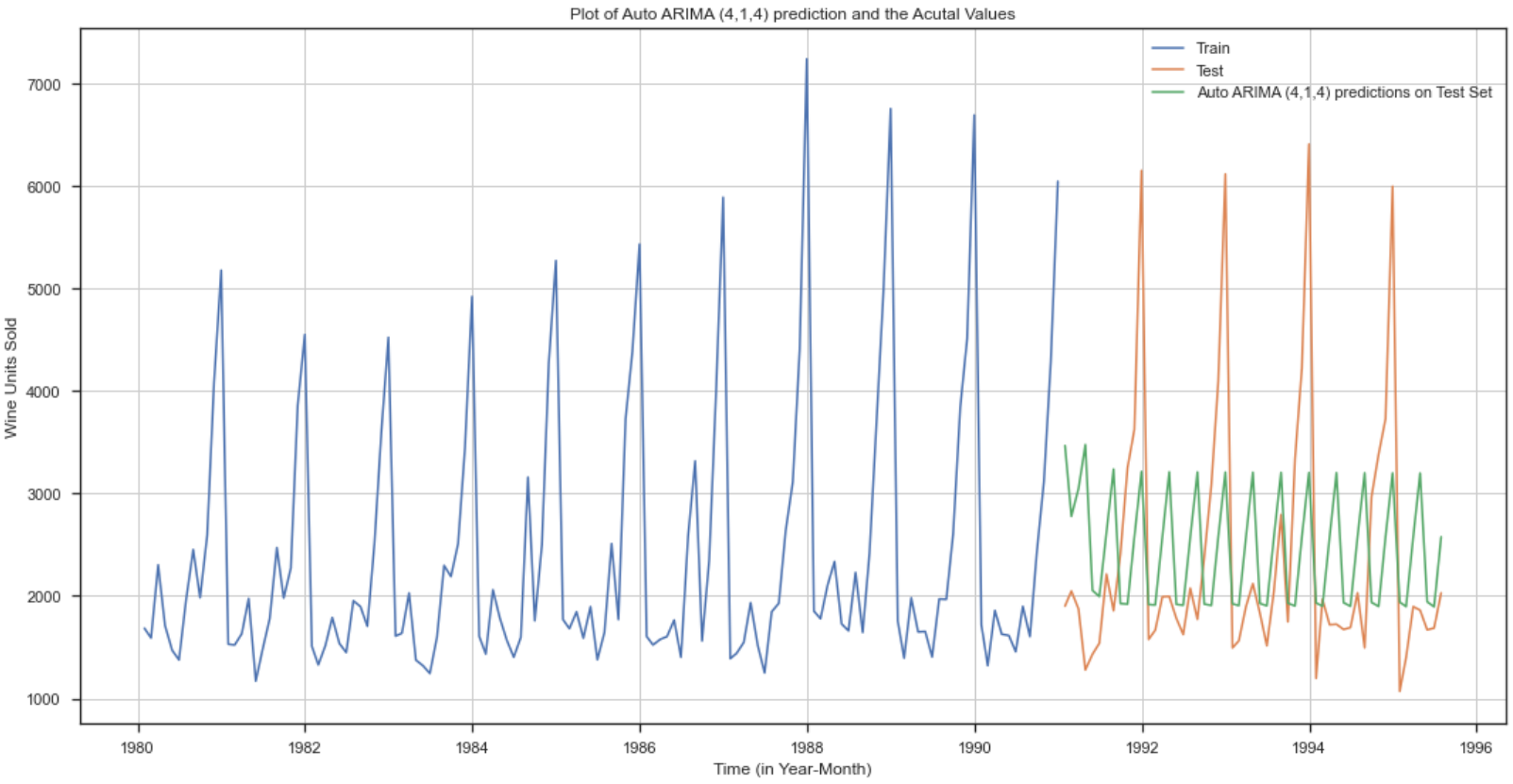


**Fig.170 Plot of Automated ARIMA (4,1,4) predictions on Test data**

## Automated ARIMA: Model Evaluation

**For evaluating the model's performance metrics, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)**

| Model | Test RMSE | Test MAPE |
|---|---|---|
| **ARIMA (p=4, d=1, q=4)** | **1212.918** | **40.214** |

**Observation:**

- We can see from the graphs above that the time series has a **marginal upward trend and seasonality**
- ARIMA models performs well on non-seasonal time series. It is due to this reason it is unable to capture the entire characteristics of the test data.
- The root means squared error **(RMSE)** of test data for the ARIMA model with **(p=4, d=1, q=4)** is **1212.918.**
- Not surprisingly, the RMSE of the aforementioned ARIMA model is lower than the majority of previously constructed models but significantly higher than triple exponential smoothing model.

## Model 9 – Seasonal Auto-Regressive Integrated Moving Average (SARIMA)

**SARIMA models or also known as Seasonal ARIMA is an extension of ARIMA for a time series data with defined seasonality. SARIMA models use seasonal differencing which is similar to regular differencing.**

**A SARIMA model is characterized by 7 terms: p, d, q, P, Q, D and F**

**where,**

**p is the order of the Auto Regressive (AR) term**

**q is the order of the Moving Average (MA) term**

**d is the number of differencing required to make the time series stationary**

**P is the order of the Seasonal Auto Regressive (AR) term**

**Q is the order of the Seasonal Moving Average (MA) term**

**D is the number of seasonal differencing required to make the time series stationary**

**F is the seasonal frequency of the time series**

**We must examine the PACF and ACF plots, respectively, at delays that are the multiple of "F" in order to determine the "P" and "Q" values, and determine where these cut-off values are (for appropriate confidence interval bands).**

**By examining the lowest AIC values, we can also estimate "p," "q," "P," and "Q" for the SARIMA models.**

**By examining the ACF plots, one may calculate the seasonal parameter 'F'. The existence of seasonality should be shown by a spike in the ACF plot at multiples of "F."**

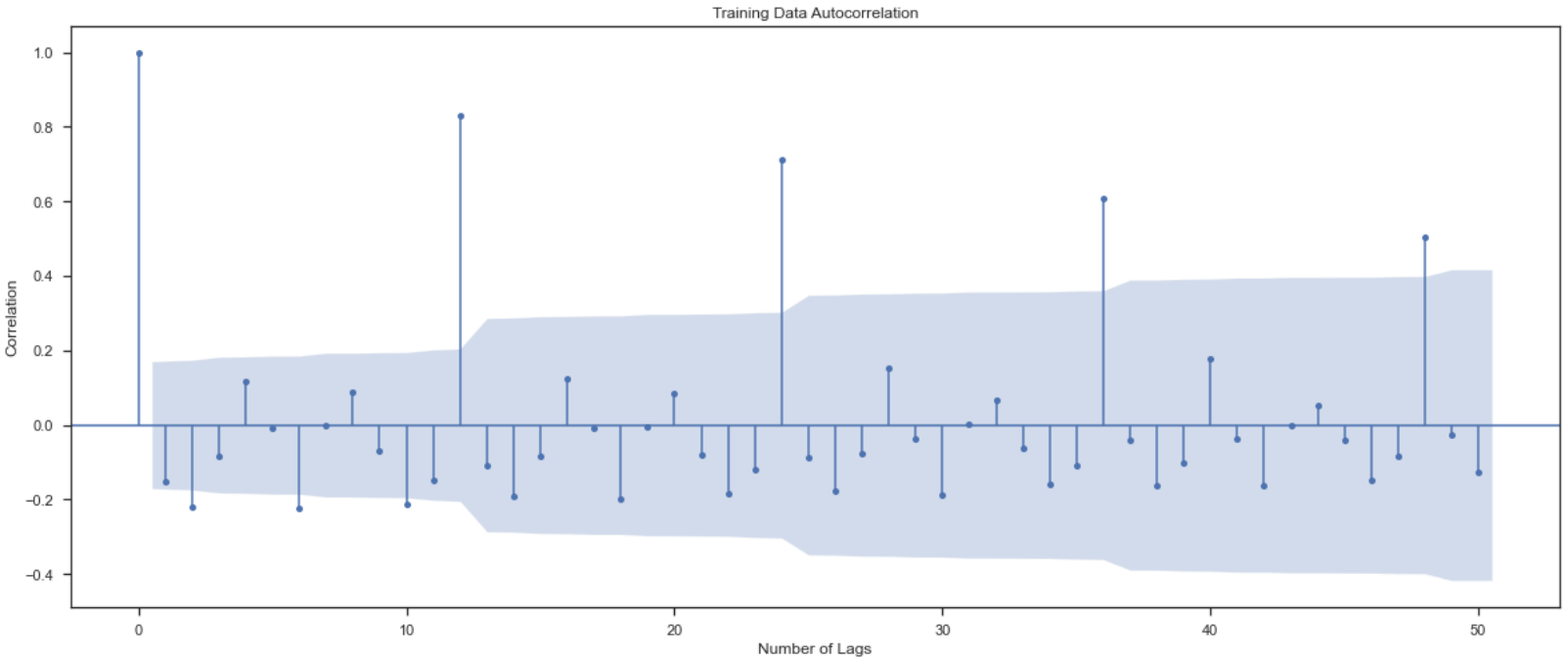


**Fig.171 ACF plot of Train data**

**From the above ACF plot we can observe that at every 12$^{th}$ lag is significant indicating the presence of seasonality. Hence for our model building we will consider the term F=12.**

**For the selection criteria of p, d, q, P, D, Q & F the below SARIMA model is built by using automated model parameters with lowest Akaike Information Criteria.**

```
Examples of the parameter combinations for the Model are
Model: (0, 1, 0)(0, 0, 0, 12)
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (0, 1, 3)(0, 0, 3, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (1, 1, 3)(1, 0, 3, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
Model: (2, 1, 3)(2, 0, 3, 12)
Model: (3, 1, 0)(3, 0, 0, 12)
Model: (3, 1, 1)(3, 0, 1, 12)
Model: (3, 1, 2)(3, 0, 2, 12)
Model: (3, 1, 3)(3, 0, 3, 12)
```

**Fig.172 Parameter Combinations for SARIMA model**

```
SARIMA(0, 1, 0)x(0, 0, 0, 12) - AIC:2251.3597196862966
SARIMA(0, 1, 0)x(0, 0, 1, 12) - AIC:1956.261461684592
SARIMA(0, 1, 0)x(0, 0, 2, 12) - AIC:1723.1533640234275
SARIMA(0, 1, 0)x(0, 0, 3, 12) - AIC:3788.8702234504462
SARIMA(0, 1, 0)x(1, 0, 0, 12) - AIC:1837.436602245668
SARIMA(0, 1, 0)x(1, 0, 1, 12) - AIC:1806.9905301389272
SARIMA(0, 1, 0)x(1, 0, 2, 12) - AIC:1633.2108735791694
SARIMA(0, 1, 0)x(1, 0, 3, 12) - AIC:3102.146459235806
SARIMA(0, 1, 0)x(2, 0, 0, 12) - AIC:1648.3776153470858
SARIMA(0, 1, 0)x(2, 0, 1, 12) - AIC:1647.205415860455
SARIMA(0, 1, 0)x(2, 0, 2, 12) - AIC:1630.98980539208
SARIMA(0, 1, 0)x(2, 0, 3, 12) - AIC:2633.1403461999307
SARIMA(0, 1, 0)x(3, 0, 0, 12) - AIC:1467.4574095308406
SARIMA(0, 1, 0)x(3, 0, 1, 12) - AIC:1469.1871052625374
SARIMA(0, 1, 0)x(3, 0, 2, 12) - AIC:1471.0594530064632
SARIMA(0, 1, 0)x(3, 0, 3, 12) - AIC:6506.405021012982
SARIMA(0, 1, 1)x(0, 0, 0, 12) - AIC:2230.162907850582
SARIMA(0, 1, 1)x(0, 0, 1, 12) - AIC:1923.7688649566421
SARIMA(0, 1, 1)x(0, 0, 2, 12) - AIC:1692.7089572968055
SARIMA(0, 1, 1)x(0, 0, 3, 12) - AIC:3446.263783179349
SARIMA(0, 1, 1)x(1, 0, 0, 12) - AIC:1797.1795881838273
SARIMA(0, 1, 1)x(1, 0, 1, 12) - AIC:1738.0903193763033
SARIMA(0, 1, 1)x(1, 0, 2, 12) - AIC:1570.1509144283382
SARIMA(0, 1, 1)x(1, 0, 3, 12) - AIC:3449.208102198681
```

**Fig.173 AIC values for different parameter combinations**

| | param | seasonal | AIC |
|---|---|---|---|
| 237 | (3, 1, 2) | (3, 0, 1, 12) | 1388.602612 |
| 221 | (3, 1, 1) | (3, 0, 1, 12) | 1388.681484 |
| 222 | (3, 1, 1) | (3, 0, 2, 12) | 1389.195902 |
| 238 | (3, 1, 2) | (3, 0, 2, 12) | 1389.701997 |
| 254 | (3, 1, 3) | (3, 0, 2, 12) | 1391.692608 |

**Fig.174 Sorted AIC values for different parameter combinations**

**We can see that among all the possible given combinations, the optimum AIC which is lowest for the combination (3,1,2) (3,0,1,12). Hence, the model is built with these parameters to determine the RMSE value of test data.**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sparkling_Wine_Sales | No. Observations: | 132 |
| Model: | SARIMAX(3, 1, 2)x(3, 0, [1], 12) | Log Likelihood | -684.301 |
| Date: | Sun, 23 Oct 2022 | AIC | 1388.603 |
| Time: | 09:10:42 | BIC | 1413.820 |
| Sample: | 01-31-1980 | HQIC | 1398.781 |
| | - 12-31-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | -0.5433 | 0.416 | -1.306 | 0.191 | -1.358 | 0.272 |
| ar.L2 | -0.0076 | 0.198 | -0.038 | 0.969 | -0.396 | 0.381 |
| ar.L3 | 0.0636 | 0.140 | 0.453 | 0.651 | -0.212 | 0.339 |
| ma.L1 | -0.1993 | 0.404 | -0.493 | 0.622 | -0.992 | 0.593 |
| ma.L2 | -0.6547 | 0.327 | -2.005 | 0.045 | -1.295 | -0.015 |
| ar.S.L12 | 0.7651 | 0.448 | 1.706 | 0.088 | -0.114 | 1.644 |
| ar.S.L24 | 0.1091 | 0.330 | 0.331 | 0.741 | -0.537 | 0.756 |
| ar.S.L36 | 0.1764 | 0.187 | 0.946 | 0.344 | -0.189 | 0.542 |
| ma.S.L12 | -0.2428 | 0.451 | -0.539 | 0.590 | -1.126 | 0.640 |
| sigma2 | 1.663e+05 | 2.63e+04 | 6.325 | 0.000 | 1.15e+05 | 2.18e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.00 | Jarque-Bera (JB): | 9.36 |
| Prob(Q): | 0.96 | Prob(JB): | 0.01 |
| Heteroskedasticity (H): | 1.25 | Skew: | 0.35 |
| Prob(H) (two-sided): | 0.54 | Kurtosis: | 4.40 |

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

**Fig.175 Sparkling Wine – Automated SARIMA model**

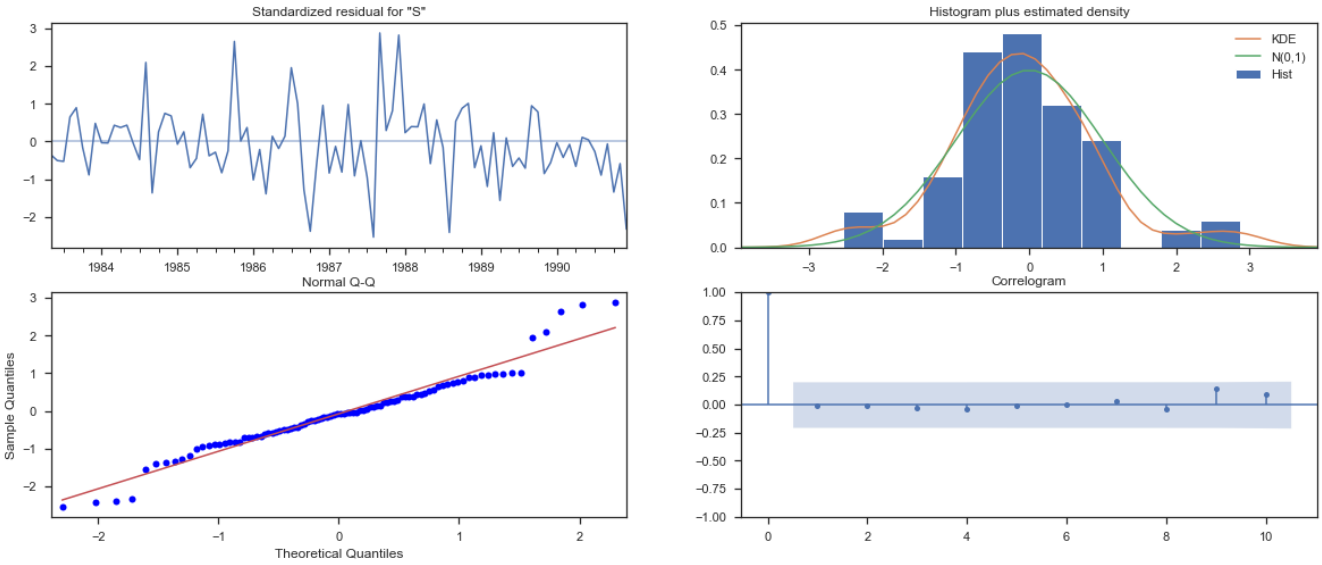


**Fig.176 Automated SARIMA – Diagnostics plot**

**Observation:**

- The optimal parameters are decided based on the lowest Akaike Information Criteria (AIC) values. **The AIC is lowest for the combination (3,1,2) (3,0,1,12) as we see from the above results.**
- From the **Standardized residual plot** above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.
- The **histogram plus estimated density plot** suggests a slightly uniform distribution with mean zero and slightly skewed to the right.
- In **Normal Q-Q plot**, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution
- The **correlogram plot** of residuals shows that the residuals are not auto correlated.

| | |
|---|---|
| 1991-01-31 | 1320.365492 |
| 1991-02-28 | 1298.641106 |
| 1991-03-31 | 1604.553642 |
| 1991-04-30 | 1626.041006 |
| 1991-05-31 | 1397.971401 |
| 1991-06-30 | 1237.768475 |
| 1991-07-31 | 1785.878306 |
| 1991-08-31 | 1510.264627 |
| 1991-09-30 | 2286.002274 |
| 1991-10-31 | 3290.953330 |
| 1991-11-30 | 4452.022927 |
| 1991-12-31 | 6491.383012 |
| 1992-01-31 | 1295.954582 |
| 1992-02-29 | 1101.842280 |
| 1992-03-31 | 1546.537323 |
| 1992-04-30 | 1443.761112 |

**Fig.177 Sample of Automated SARIMA (3,1,2) (3,0,1,12) predictions**

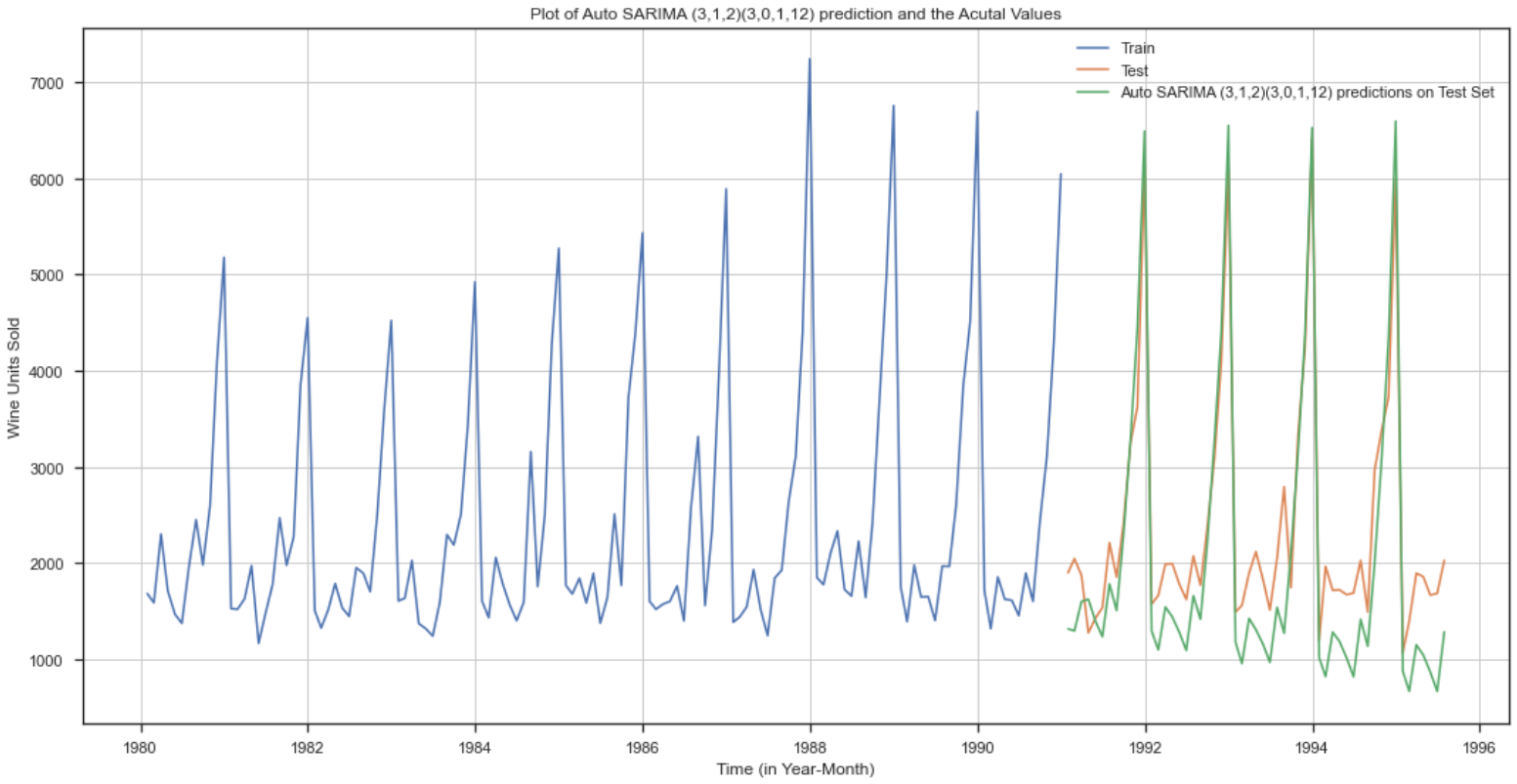


**Fig.178 Plot of Automated SARIMA (3,1,2) (3,0,1,12) predictions on Test data**

## Automated SARIMA: Model Evaluation

**For evaluating the model performance, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)**

| Model | Test RMSE | Test MAPE |
|---|---|---|
| **SARIMA (p=3, d=1, q=2) (P=3, D=0, Q=1, F=12)** | **579.925** | **25.052** |

**Observation:**

- We can see from the graphs above that the time series has **a marginal upward trend and seasonality**
- SARIMA model performs well on seasonal time series. It is due to this reason it is able to capture the entire characteristics of the test data.
- The root means squared error **(RMSE)** of test data for the SARIMA model with **(p=3, d=1, q=2) (P=3, D=0, Q=1, F=12)** is **579.925.**
- Additionally, it should be highlighted that compared to the ARIMA model, the SARIMA model has almost more than halved the RMSE value.

## 7) Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

### Model 10 – Auto-Regressive Integrated Moving Average (ARIMA) - Manual

**An ARIMA model is characterized by 3 terms: p, d, q**

**where,**

**p is the order of the Auto Regressive (AR) term**

**q is the order of the Moving Average (MA) term**

**d is the number of differencing required to make the time series stationary**

Indicating which previous series values are most beneficial in forecasting future values, autocorrelation and partial autocorrelation are measures of relationship between present and past series values. You may identify the sequence of processes in an ARIMA model using this information.

**The parameters p & q can be determined by looking at the PACF & ACF plots respectively.**

**Autocorrelation function (ACF) -** At lag *k*, this is the correlation between series values that are *k* intervals apart.

**Partial autocorrelation function (PACF) -** At lag *k*, this is the correlation between series values that are *k* intervals apart, accounting for the values of the intervals between.

**In an ACF & PACF plots, each bar represents the size and direction of the connection. Bars that cross the red line are statistically significant.**

## ACF Plot – Training Data

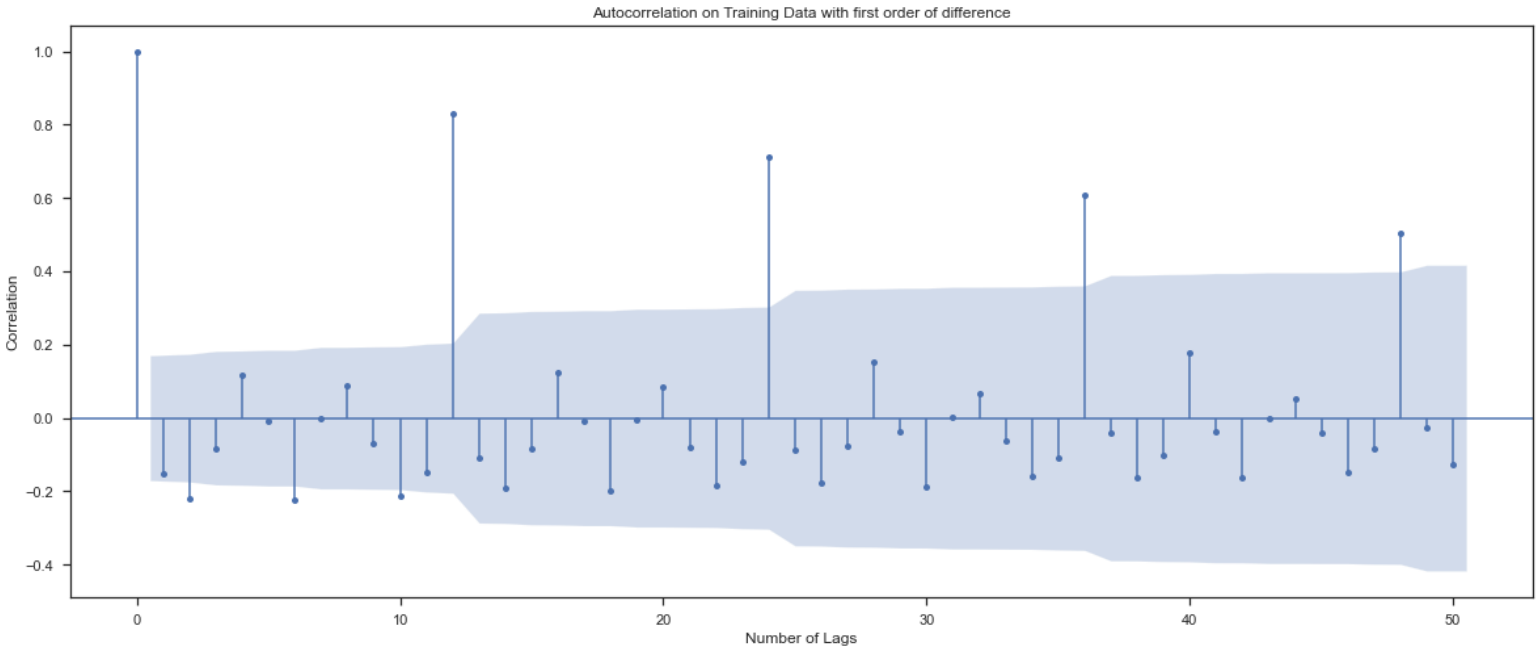


**Fig.179 ACF plot on differenced train data**

## PACF Plot – Training Data

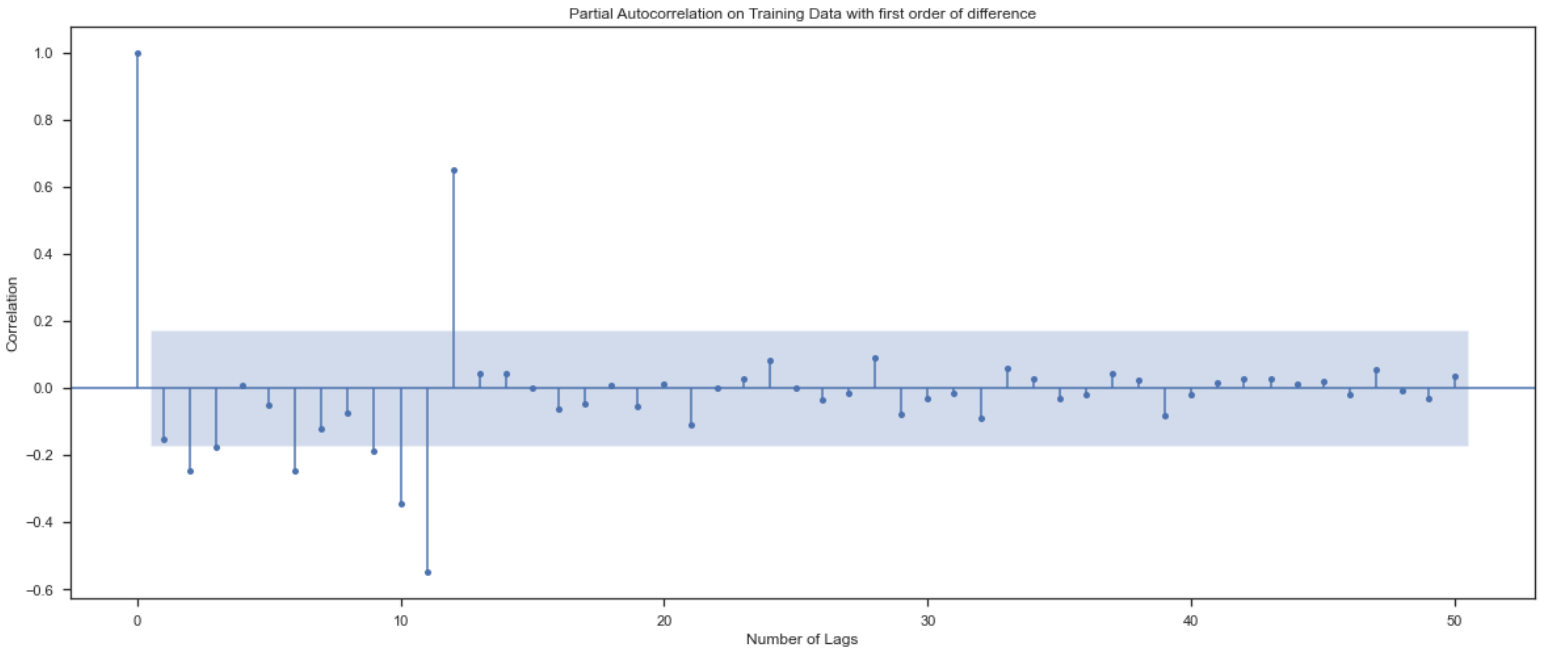


**Fig.180 PACF plot on differenced train data**

**Observation:**

- The **Auto-Regressive parameter** in an ARIMA model is **'p'** which comes from the significant lag after which the PACF plot cuts-off below the confidence interval.
- The **Moving-Average parameter** in an ARIMA model is **'q'** which comes from the significant lag after which the ACF plot cuts-off below the confidence interval.
- **We can observe from the above plots that after lag 1, we have few significant lags and hence we would also build another model by taking value of p=2 and q=1 respectively.**

```
                                SARIMAX Results
==========================================================================
=
Dep. Variable:     Sparkling_Wine_Sales   No. Observations:                 13
2
Model:                  ARIMA(2, 1, 1)   Log Likelihood                -1112.88
9
Date:                  Sun, 23 Oct 2022   AIC                           2233.77
8
Time:                          11:48:14   BIC                           2245.27
8
Sample:                      01-31-1980   HQIC                          2238.45
1
                           - 12-31-1990
Covariance Type:                   opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          0.5194      0.119      4.359      0.000       0.286       0.753
ar.L2         -0.1782      0.196     -0.908      0.364      -0.563       0.206
ma.L1         -0.9993      0.667     -1.499      0.134      -2.306       0.307
sigma2      1.354e+06   8.31e+05      1.630      0.103   -2.74e+05    2.98e+06
==========================================================================
====
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 1
2.66
Prob(Q):                              0.76   Prob(JB):
0.00
Heteroskedasticity (H):               2.77   Skew:
0.50
Prob(H) (two-sided):                  0.00   Kurtosis:
4.14
==========================================================================
====

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

**Fig.181 Sparkling Wine – Manual ARIMA model**

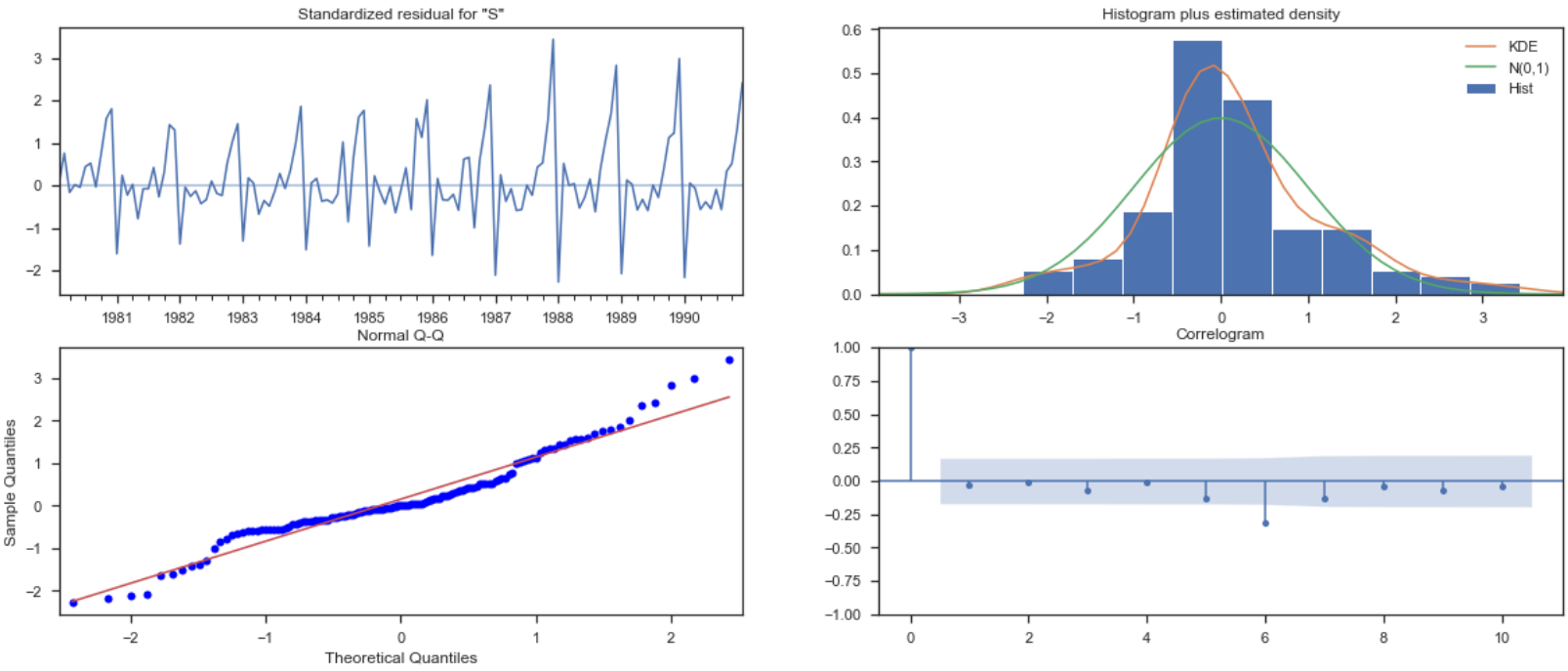


**Fig.182 Manual ARIMA – Diagnostics plot**

**Observation:**

- The model's parameters, p and q, were identified by examining the **ACF (q=1)** and **PACF (p=2)** graphs. Since we differenced the series to make it stationary, the parameter **d=1**.
- From the **Standardized residual plot** above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.
- The **histogram plus estimated density plot** suggests a slightly uniform distribution with mean zero and slightly skewed to the right.
- In **Normal Q-Q plot**, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.
- The **correlogram plot** of residuals shows that the residuals are not auto correlated.

| | |
|---|---|
| 1991-01-31 | 3957.667290 |
| 1991-02-28 | 2558.511466 |
| 1991-03-31 | 2204.093503 |
| 1991-04-30 | 2269.356381 |
| 1991-05-31 | 2366.423315 |
| 1991-06-30 | 2405.213121 |
| 1991-07-31 | 2408.062662 |
| 1991-08-31 | 2402.629532 |
| 1991-09-30 | 2399.299439 |
| 1991-10-31 | 2398.537947 |
| 1991-11-30 | 2398.735898 |
| 1991-12-31 | 2398.974440 |
| 1992-01-31 | 2399.063070 |
| 1992-02-29 | 2399.066595 |
| 1992-03-31 | 2399.052630 |
| 1992-04-30 | 2399.044748 |

**Fig.183 Sample of Manual ARIMA (2,1,1) predictions**

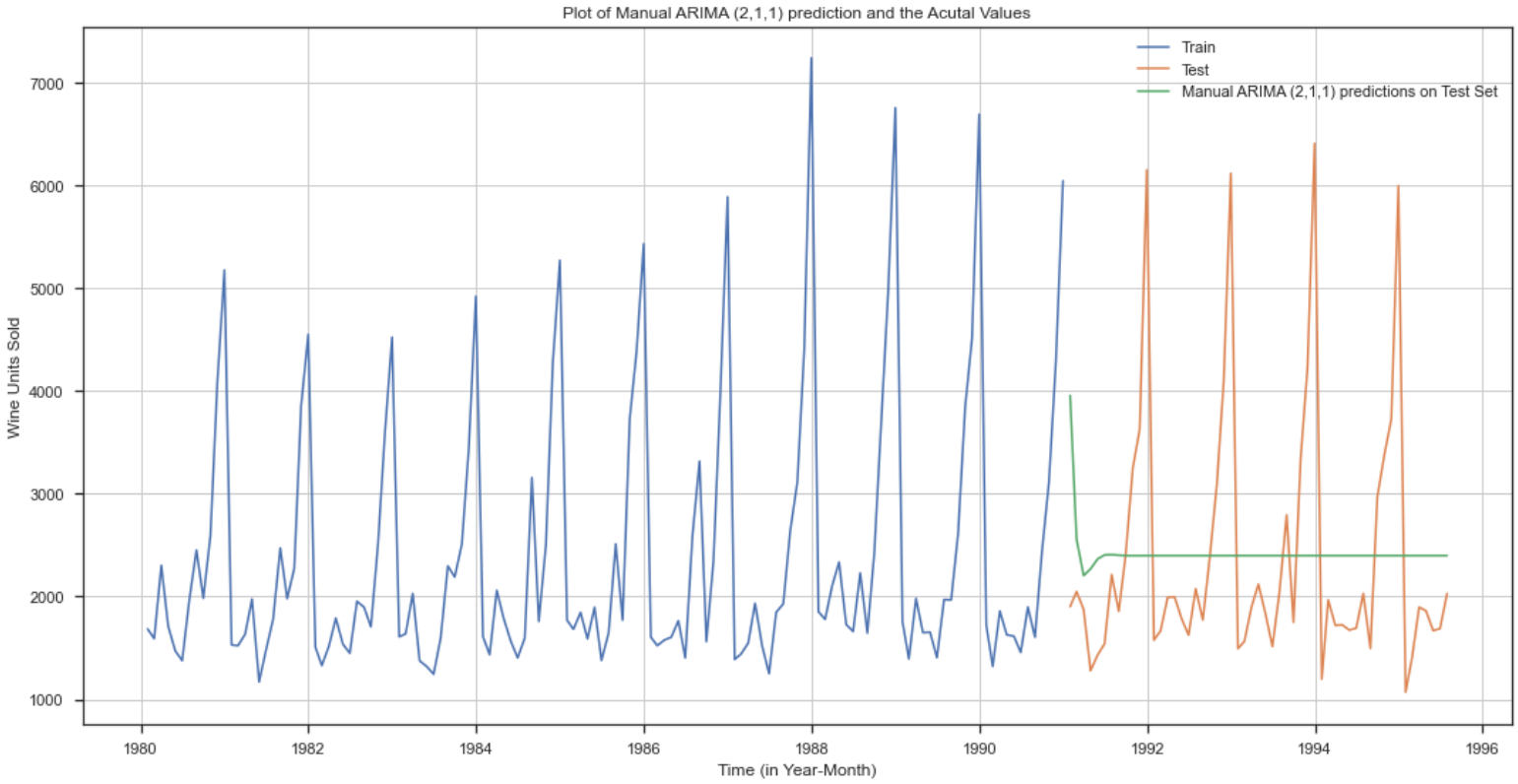


**Fig.184 Plot of Manual ARIMA (2,1,1) predictions on Test data**

## Manual ARIMA: Model Evaluation

**For evaluating the model performance, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)**

| Model | Test RMSE | Test MAPE |
|---|---|---|
| **ARIMA (p=2, d=1, q=1)** | **1300.721** | **40.225** |

**Observation:**

- We can see from the graphs above that the time series has **a marginal upward trend and seasonality**
- ARIMA models performs well on non-seasonal time series. It is due to this reason it is unable to capture the entire characteristics of the test data.
- The root means squared error **(RMSE)** of test data for the ARIMA model with **(p=2, d=1, q=1)** is **1300.721**
- Not surprisingly, the RMSE of the aforementioned ARIMA model is greater than the majority of previously constructed models and **also higher than Automated ARIMA (4,1,4) model**.

## Model 11 – Seasonal Auto-Regressive Integrated Moving Average (SARIMA) – Manual

**A SARIMA model is characterized by 7 terms: p, d, q, P, Q, D and F**

**where,**

**p is the order of the Auto Regressive (AR) term**

**q is the order of the Moving Average (MA) term**

**d is the number of differencing required to make the time series stationary**

**P is the order of the Seasonal Auto Regressive (AR) term**

**Q is the order of the Seasonal Moving Average (MA) term**

**D is the number of seasonal differencing required to make the time series stationary**

**F is the seasonal frequency of the time series**

**We must examine the PACF and ACF plots, respectively, at delays that are the multiple of "F" in order to determine the "P" and "Q" values, and determine where these cut-off values are (for appropriate confidence interval bands).**

**By examining the ACF plots, one may calculate the seasonal parameter 'F'. The existence of seasonality should be shown by a spike in the ACF plot at multiples of "F."**

**The parameters P & Q can be determined by looking at the seasonally differenced PACF & ACF plots respectively.**

**Autocorrelation function (ACF) -** At lag *k*, this is the correlation between series values that are *k* intervals apart.

**Partial autocorrelation function (PACF) -** At lag *k*, this is the correlation between series values that are *k* intervals apart, accounting for the values of the intervals between.

**In an ACF & PACF plots, each bar represents the size and direction of the connection. Bars that cross the red line are statistically significant.**

## ACF Plot – Seasonally differenced (F=12) Training Data

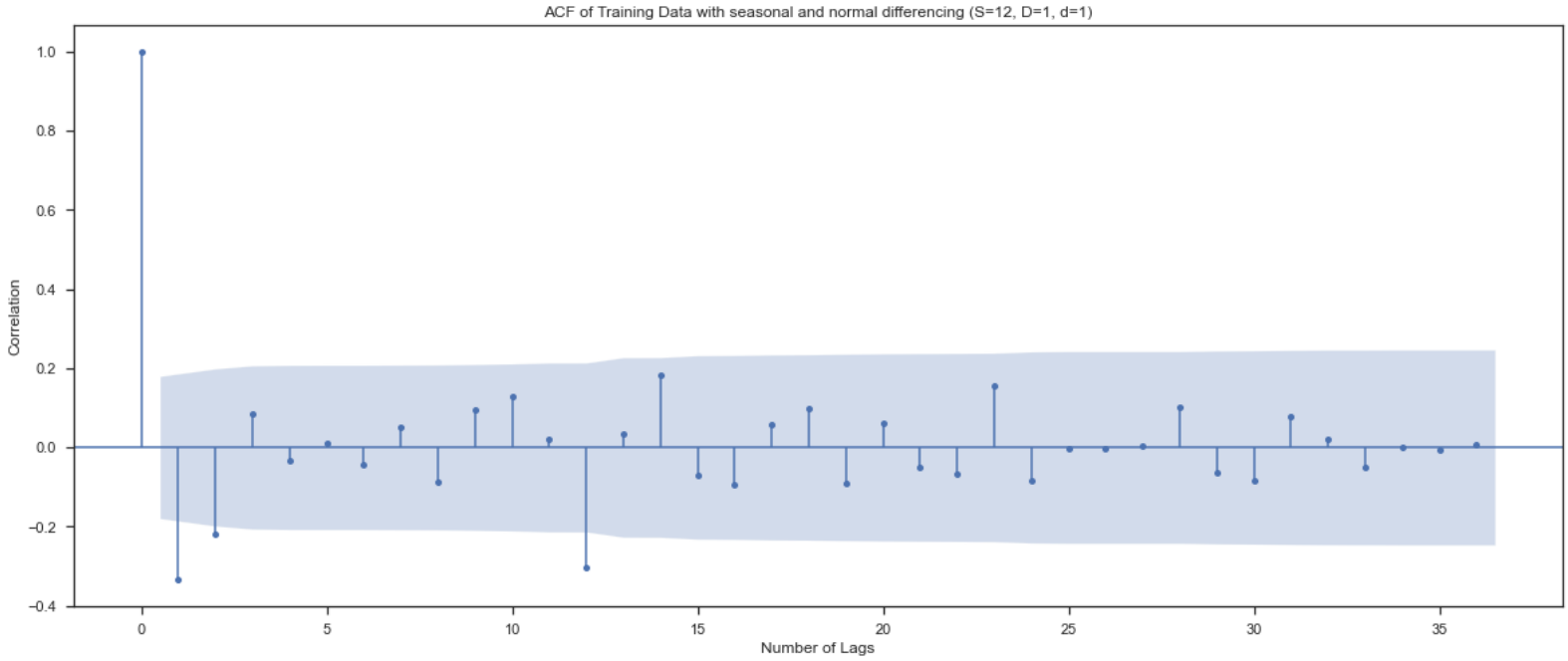


**Fig.185 ACF plot on differenced train data**

## PACF Plot – Seasonally differenced (F=12) Training Data

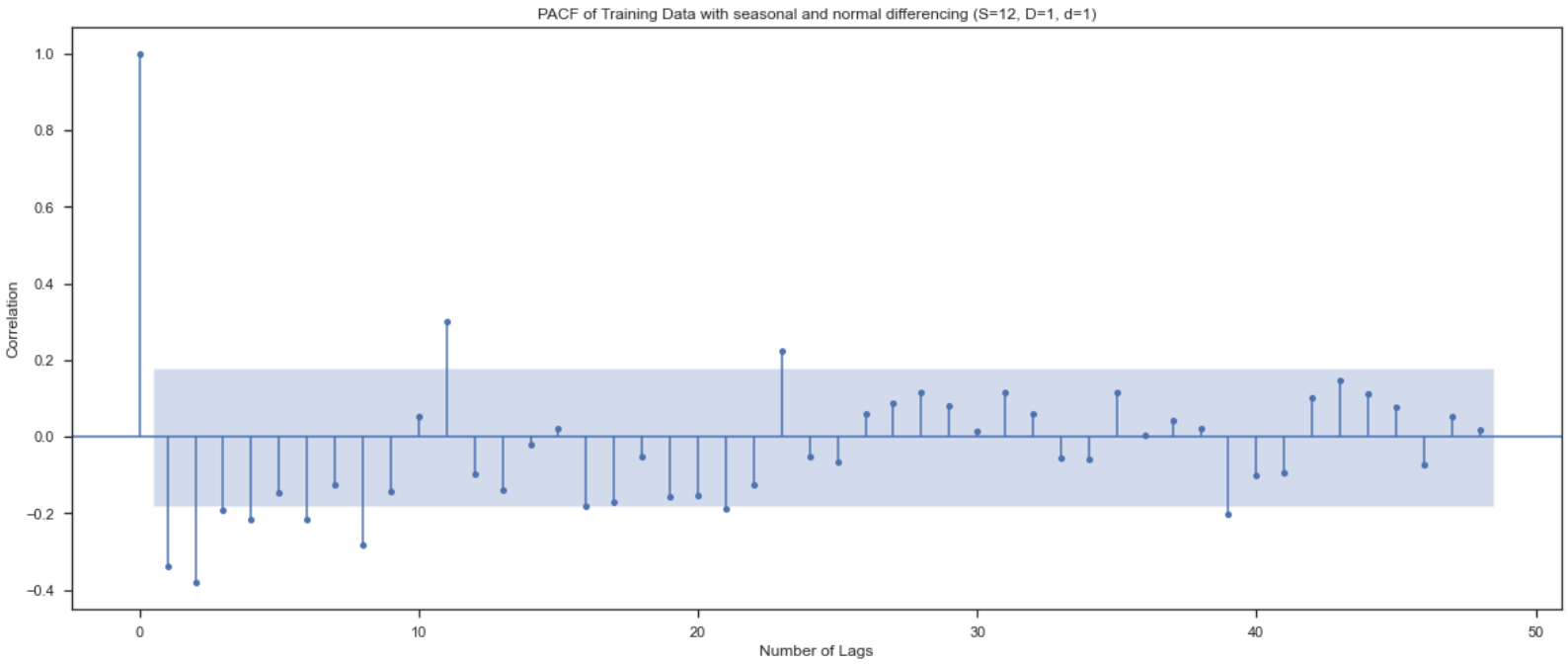


**Fig.186 PACF plot on differenced train data**

**Observation:**

- From the PACF plot it can be seen in early lags that till lag 4 is significant before cut-off, so AR term **'p = 4'** is chosen. From the multiples of seasonal lags, after first seasonal lag of 12, it cuts off, so keep seasonal AR **'P = 0'**.
- From ACF plot, it can be seen in early lags, lag 1 and 2 are significant before it cuts off, so let's 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 let's keep **'Q = 1'**.
- **The final selected terms for SARIMA model are (4, 1, 2) (0, 1, 1, 12), as inferred from the ACF and PACF plots.**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sparkling_Wine_Sales | No. Observations: | 132 |
| Model: | SARIMAX(4, 1, 2)x(0, 1, [1], 12) | Log Likelihood | -771.377 |
| Date: | Sun, 23 Oct 2022 | AIC | 1558.755 |
| Time: | 09:10:45 | BIC | 1579.910 |
| Sample: | 01-31-1980 | HQIC | 1567.325 |
| | - 12-31-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | -0.1794 | 0.593 | -0.303 | 0.762 | -1.341 | 0.982 |
| ar.L2 | -0.0852 | 0.193 | -0.442 | 0.659 | -0.463 | 0.293 |
| ar.L3 | 0.0187 | 0.114 | 0.163 | 0.870 | -0.205 | 0.242 |
| ar.L4 | -0.1634 | 0.151 | -1.085 | 0.278 | -0.458 | 0.132 |
| ma.L1 | -0.5369 | 0.625 | -0.859 | 0.391 | -1.763 | 0.689 |
| ma.L2 | -0.3071 | 0.588 | -0.522 | 0.602 | -1.460 | 0.846 |
| ma.S.L12 | -0.4833 | 0.087 | -5.582 | 0.000 | -0.653 | -0.314 |
| sigma2 | 1.614e+05 | 2.28e+04 | 7.078 | 0.000 | 1.17e+05 | 2.06e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.01 | Jarque-Bera (JB): | 14.39 |
| Prob(Q): | 0.94 | Prob(JB): | 0.00 |
| Heteroskedasticity (H): | 1.21 | Skew: | 0.56 |
| Prob(H) (two-sided): | 0.58 | Kurtosis: | 4.44 |

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

**Fig.187 Sparkling Wine – Manual SARIMA model**

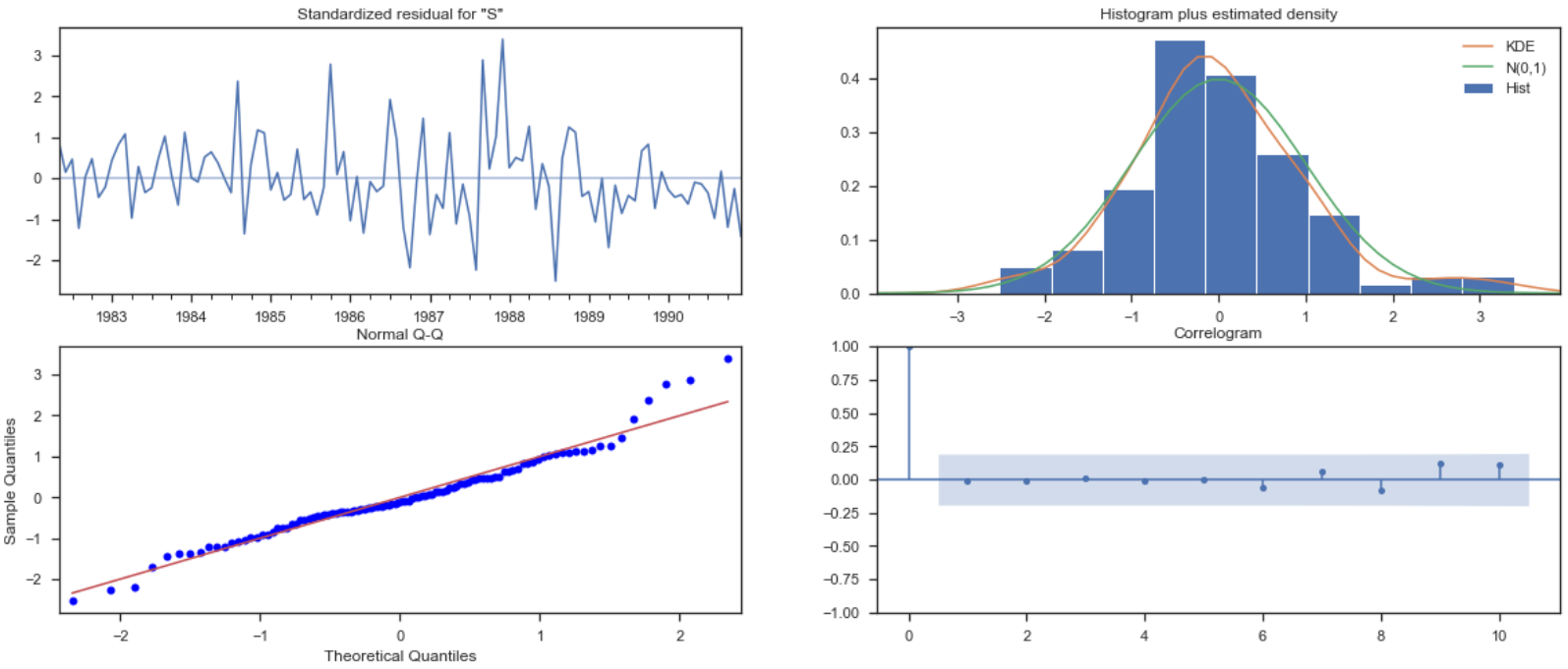


**Fig.188 Manual SARIMA – Diagnostics plot**

**Observation:**

- The model's parameters, p, q, P, Q were identified by examining the **ACF (q=2, Q=1)** and **PACF (p=4, P=0)** graphs. Since we differenced the series to make it stationary, the parameter **d=1, D=1**.
- From the **Standardized residual plot** above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.
- The **histogram plus estimated density plot** suggests a slightly uniform distribution with mean zero.
- In **Normal Q-Q plot**, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution
- The **correlogram plot** of residuals shows that the residuals are not auto correlated.

| | |
|---|---|
| 1991-01-31 | 1431.532524 |
| 1991-02-28 | 1348.288049 |
| 1991-03-31 | 1722.381473 |
| 1991-04-30 | 1621.897859 |
| 1991-05-31 | 1476.372091 |
| 1991-06-30 | 1255.611862 |
| 1991-07-31 | 1798.112892 |
| 1991-08-31 | 1612.942581 |
| 1991-09-30 | 2267.723970 |
| 1991-10-31 | 3183.853028 |
| 1991-11-30 | 4248.331847 |
| 1991-12-31 | 6166.345200 |
| 1992-01-31 | 1398.918155 |
| 1992-02-29 | 1206.059434 |

**Fig.189 Sample of Manual SARIMA (4,1,2) (0,1,1,12) predictions**

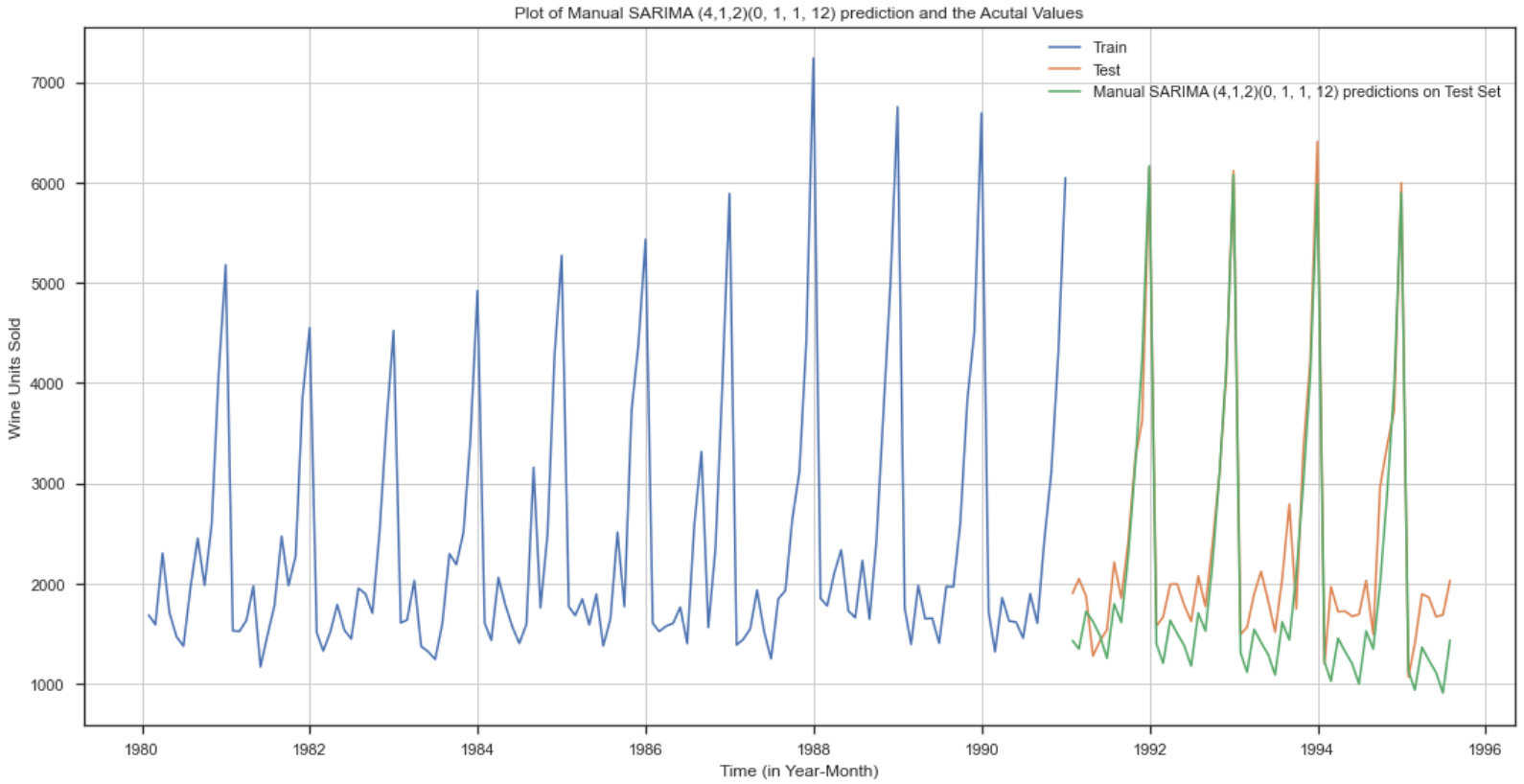


**Fig.190 Plot of Manual SARIMA (4,1,2) (0,1,1,12) predictions on Test data**

## Manual SARIMA: Model Evaluation

**For evaluating the model performance, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)**

| Model | Test RMSE | Test MAPE |
|---|---|---|
| **ARIMA (p=4, d=1, q=2) (P=0, D=1, Q=1, F=12)** | **468.677** | **19.324** |

**Observation:**

- We can see from the graphs above that the time series has **a marginal upward trend and seasonality**
- SARIMA model performs well on seasonal time series. It is due to this reason it is able to capture the entire characteristics of the test data.
- The root means squared error **(RMSE)** of test data for the SARIMA model with **(p=4, d=1, q=1) (P=0, D=1, Q=1, F=12)** is **468.677**
- Additionally, it should be highlighted that compared to the all the ARIMA/SARIMA models built so far, **this SARIMA model has the lowest RMSE value**.

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

| | Test RMSE |
|---|---|
| Linear Regression | 1389.135175 |
| Naive Model | 3864.279352 |
| Simple Average | 1275.081804 |
| 2 point TMA | 813.400684 |
| 4 point TMA | 1156.589694 |
| 6 point TMA | 1283.927428 |
| 9 point TMA | 1346.278315 |
| Alpha =0.0496,SimpleExponentialSmoothing | 1316.135411 |
| Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing | 2007.238526 |
| Alpha=0.05, Beta=0.05, Double Exponential Smoothing | 1418.407668 |
| Alpha=0.111,Beta=0.0617,Gamma=0.395,Triple Exponential Smoothing | 469.659106 |
| Alpha=0.35,Beta=0.10,Gamma=0.20,Triple Exponential Smoothing | 319.498680 |
| Auto ARIMA (4,1,4) | 1212.918076 |
| Auto SARIMA (3,1,2)(3,0,1,12) | 579.925219 |
| Manual ARIMA (2,1,1) | 1300.721383 |
| Manual SARIMA (4, 1, 2)(0, 1, 1, 12) | 468.677589 |

**Fig.191 RMSE values of all models**

| | Test RMSE |
|---|---|
| Alpha=0.35,Beta=0.10,Gamma=0.20,Triple Exponential Smoothing | 319.498680 |
| Manual SARIMA (4, 1, 2)(0, 1, 1, 12) | 468.677589 |
| Alpha=0.111,Beta=0.0617,Gamma=0.395,Triple Exponential Smoothing | 469.659106 |
| Auto SARIMA (3,1,2)(3,0,1,12) | 579.925219 |
| 2 point TMA | 813.400684 |
| 4 point TMA | 1156.589694 |
| Auto ARIMA (4,1,4) | 1212.918076 |
| Simple Average | 1275.081804 |
| 6 point TMA | 1283.927428 |
| Manual ARIMA (2,1,1) | 1300.721383 |
| Alpha =0.0496,SimpleExponentialSmoothing | 1316.135411 |
| 9 point TMA | 1346.278315 |
| Linear Regression | 1389.135175 |
| Alpha=0.05, Beta=0.05, Double Exponential Smoothing | 1418.407668 |
| Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing | 2007.238526 |
| Naive Model | 3864.279352 |

**Fig.192 Sorted RMSE values of all models**

**Observation:**

- From the above table, we can see that **Triple Exponential Smoothing model** with parameters **(Alpha=0.35, Beta=0.10, Gamma=0.20)** has the **lowest RMSE** for test data.
- **Manual SARIMA (4,1,2) (0,1,1,12)** model is having the **second lowest RMSE** value for test data after Triple Exponential Smoothing model.
- The **naïve forecast model** has performed the worst in terms of RMSE.

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

**From Fig.86 we observed the Triple Exponential Smoothing model is the optimum model for the given data set as it has the lowest RMSE value.**

**However, as we know SARIMA models tend to perform better with seasonal time series, we are also considering SARIMA model for the forecast.**

Let us visually see the time series plots of different models we have built so far on test data

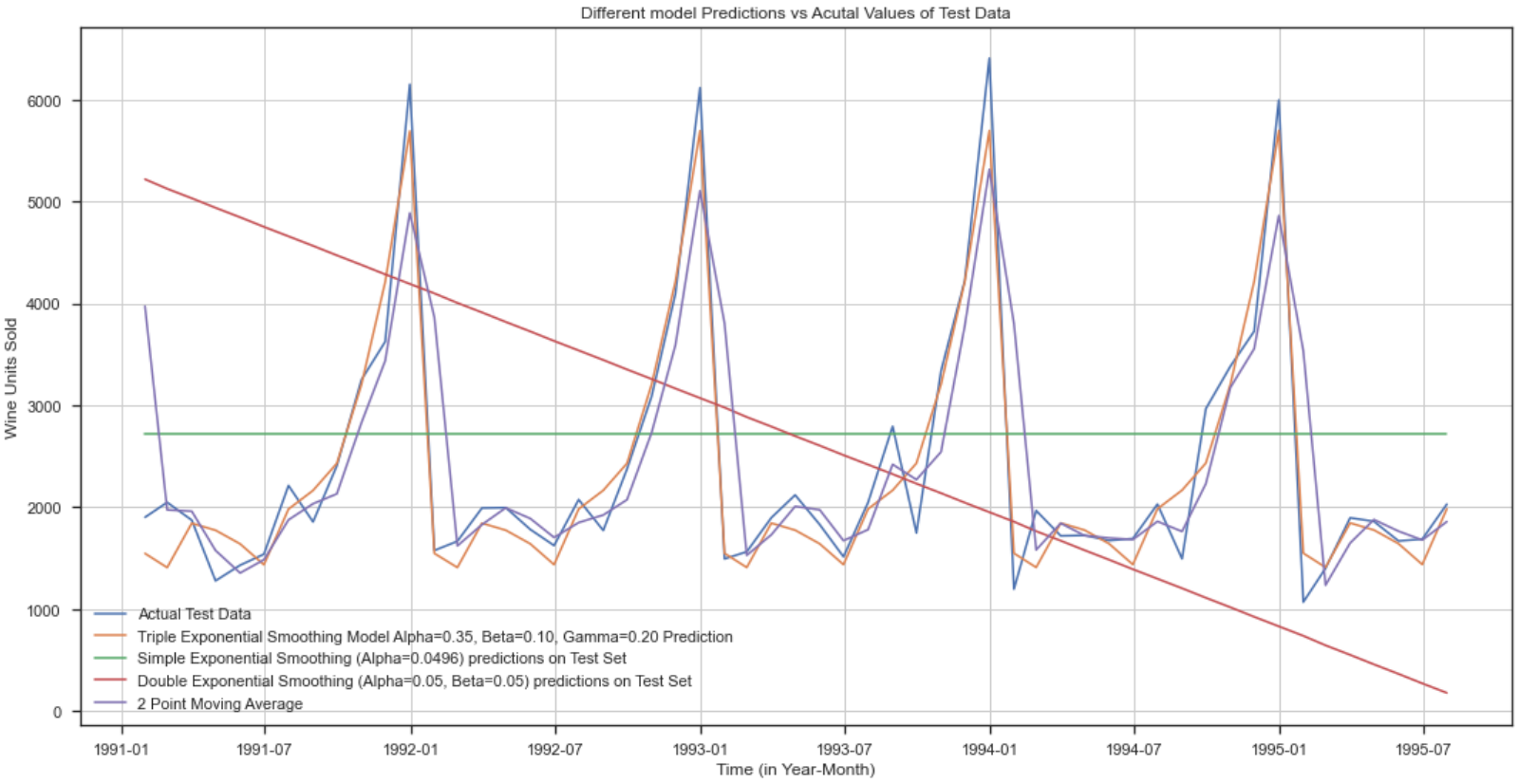


**Fig.193 Time Series Plot 1 – Different Model predictions on test data**

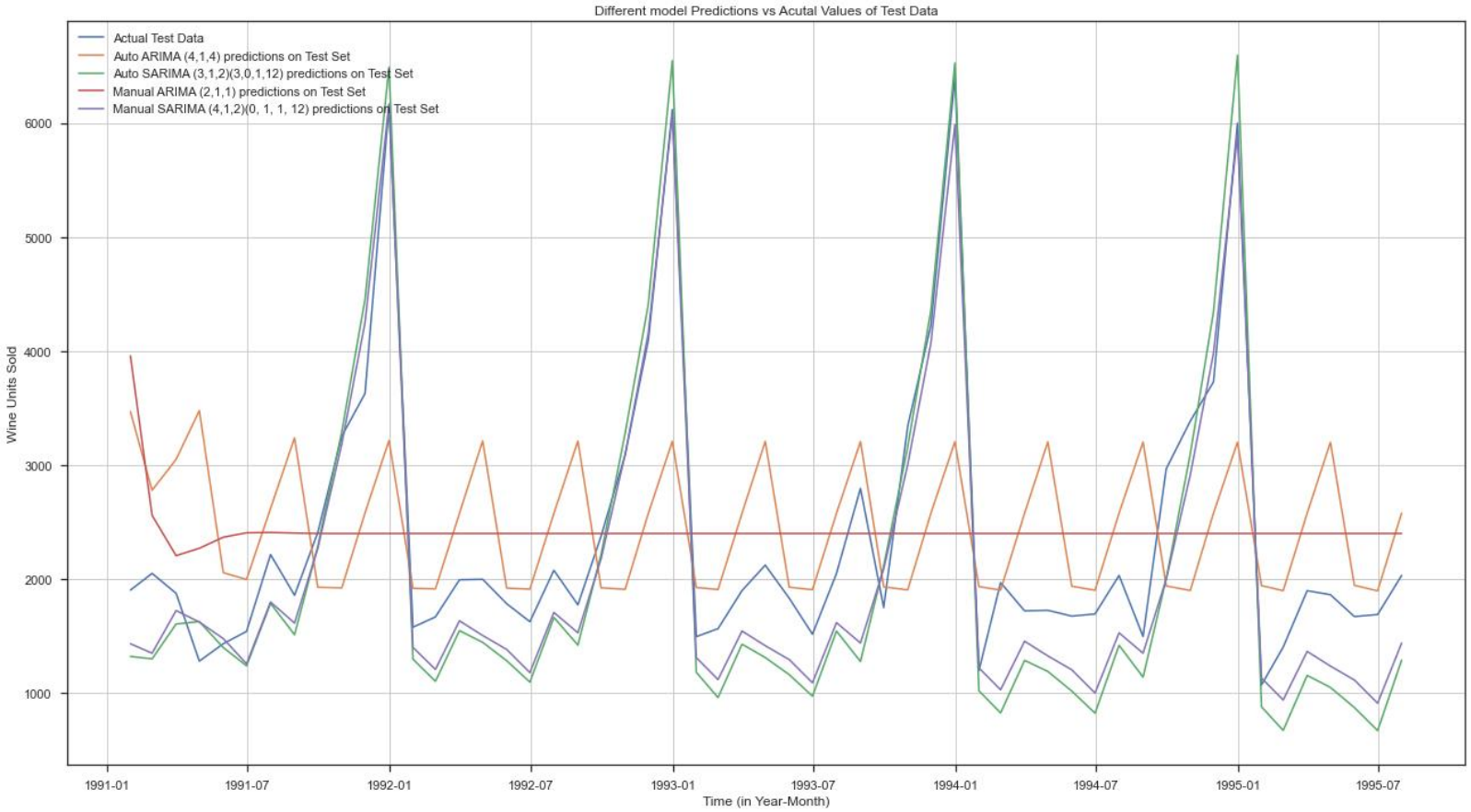


**Fig.194 Time Series Plot 2 – Different Model predictions on test data**

## Plotting the lowest RMSE models

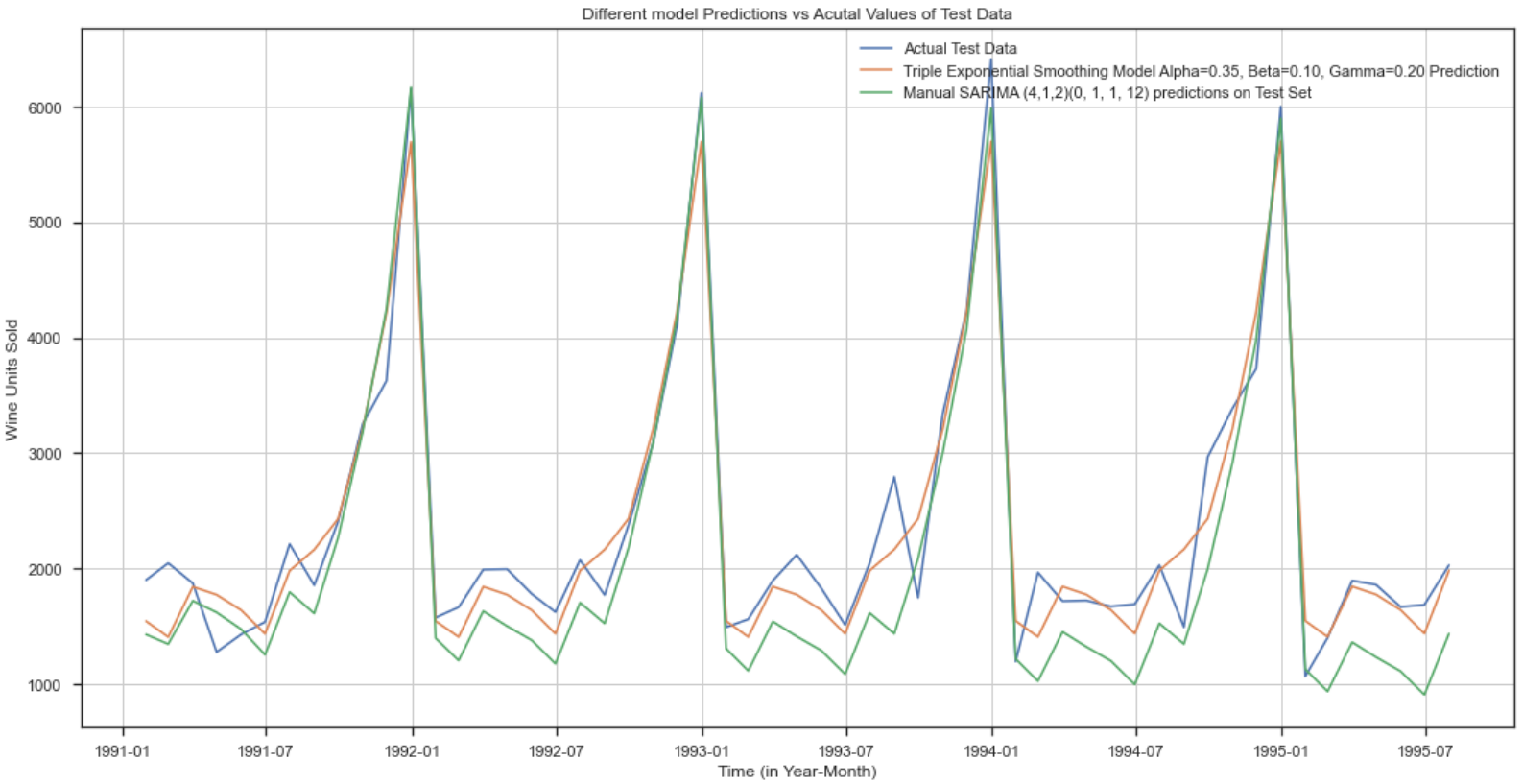


**Fig.195 Time Series Plot 3 – Different Model predictions on test data**

## Optimum Model 1:
## Triple Exponential Smoothing Model (Alpha=0.35, Beta=0.10, Gamma=0.20)

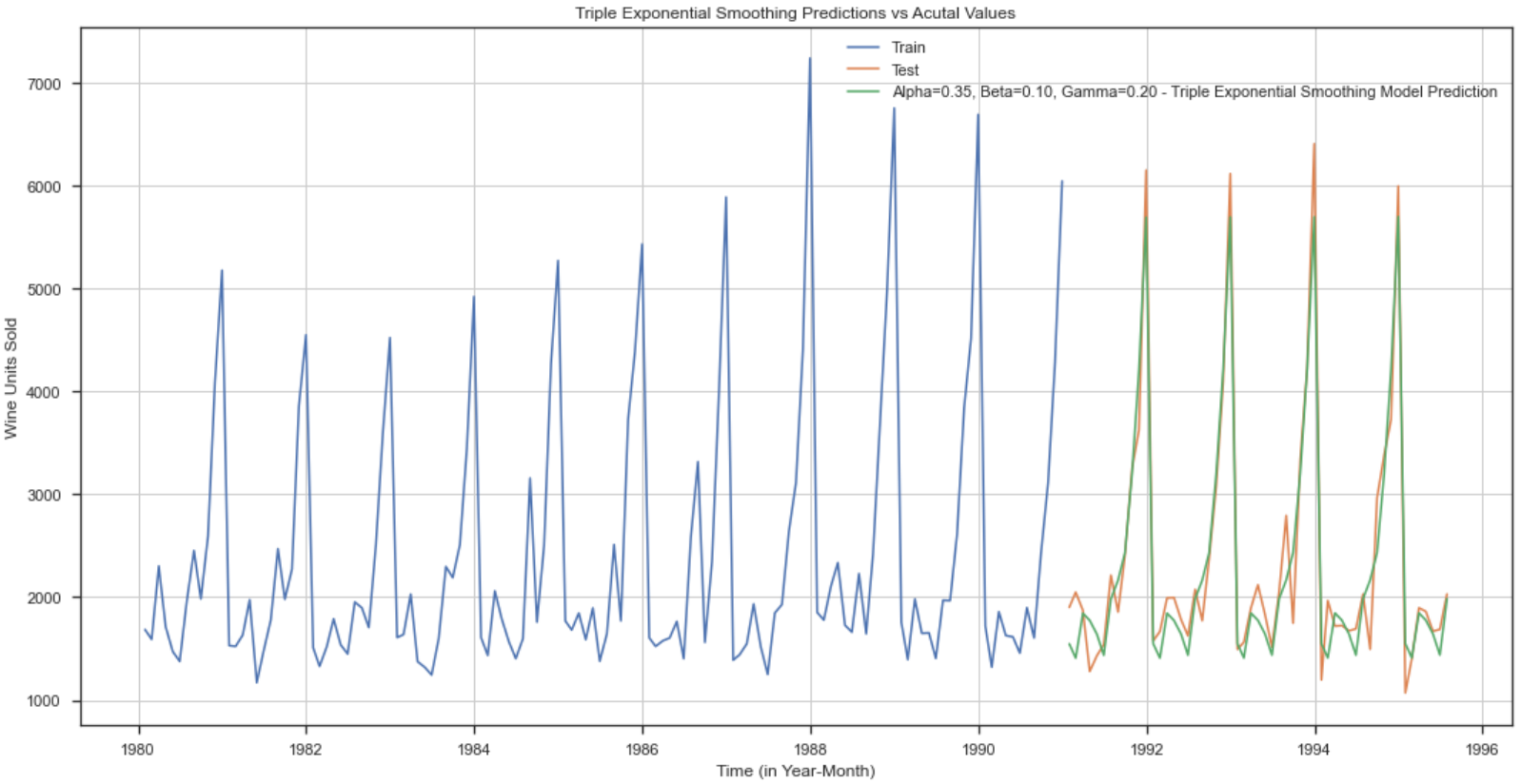


**Fig.196 TES Optimum Model – Line plot of Predictions vs Actual values**

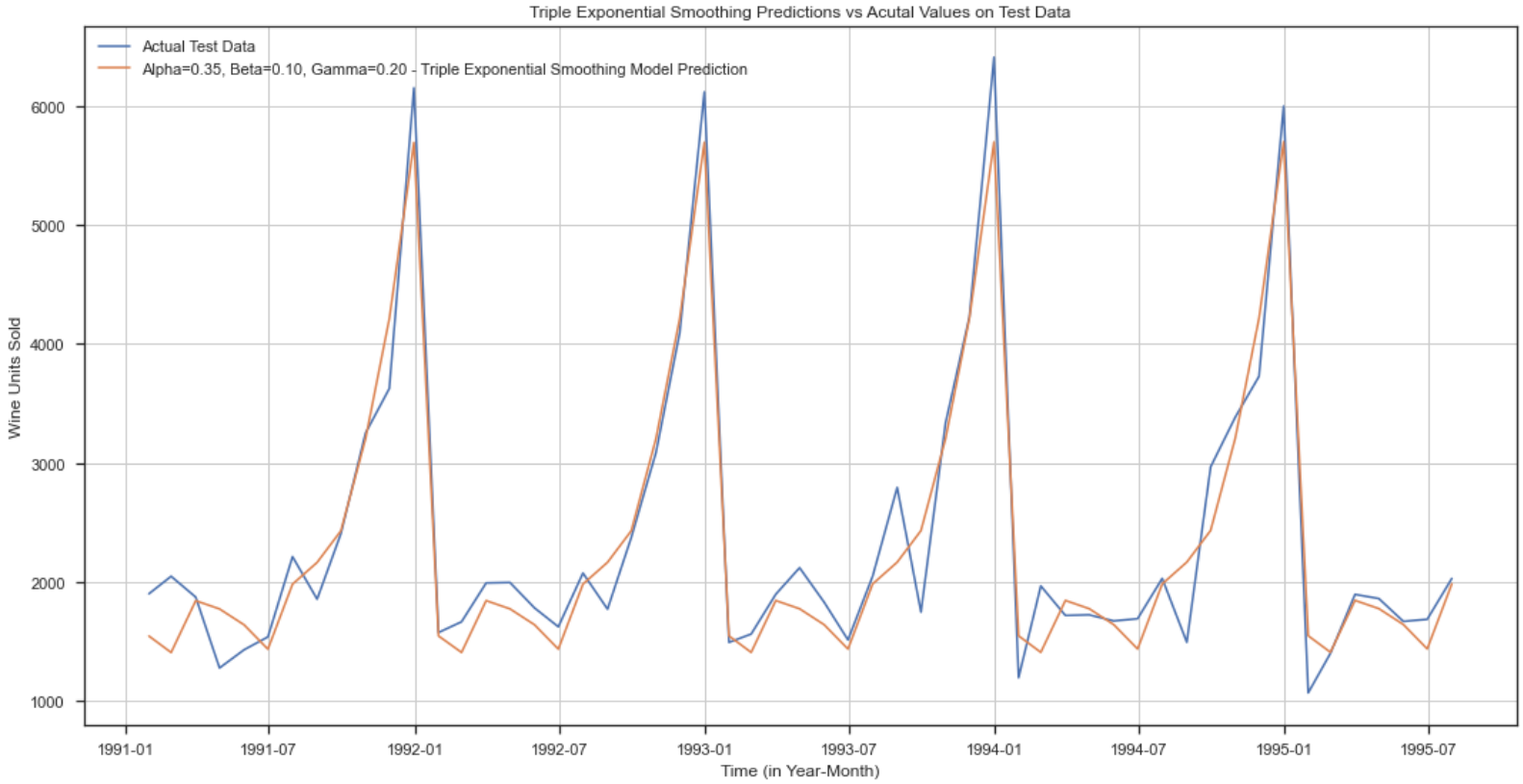


**Fig.197 TES Optimum Model – Line plot of Predictions vs Actual values on Test data**

```
{'smoothing_level': 0.35,
 'smoothing_trend': 0.1,
 'smoothing_seasonal': 0.2,
 'damping_trend': nan,
 'initial_level': 1398.2530588387642,
 'initial_trend': -9.965447804805777,
 'initial_seasons': array([1.22591905, 1.17793713, 1.52997   , 1.38253648, 1.18
977646,
        1.15907087, 1.57422577, 2.03505399, 1.74199935, 2.36190106,
        3.4666768 , 4.54948665]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

**Fig.198 TES Optimum Model**

```
Forecast of next 12 months

1995-08-31    2045.552999
1995-09-30    2527.646768
1995-10-31    3363.698965
1995-11-30    4240.102444
1995-12-31    6487.343248
1996-01-31    1540.031613
1996-02-29    1804.447295
1996-03-31    2055.022618
1996-04-30    1986.097591
1996-05-31    1802.258380
1996-06-30    1681.621958
1996-07-31    2136.129372
Freq: M, dtype: float64
```

**Fig.199 TES Model – Forecast for next 12 months**

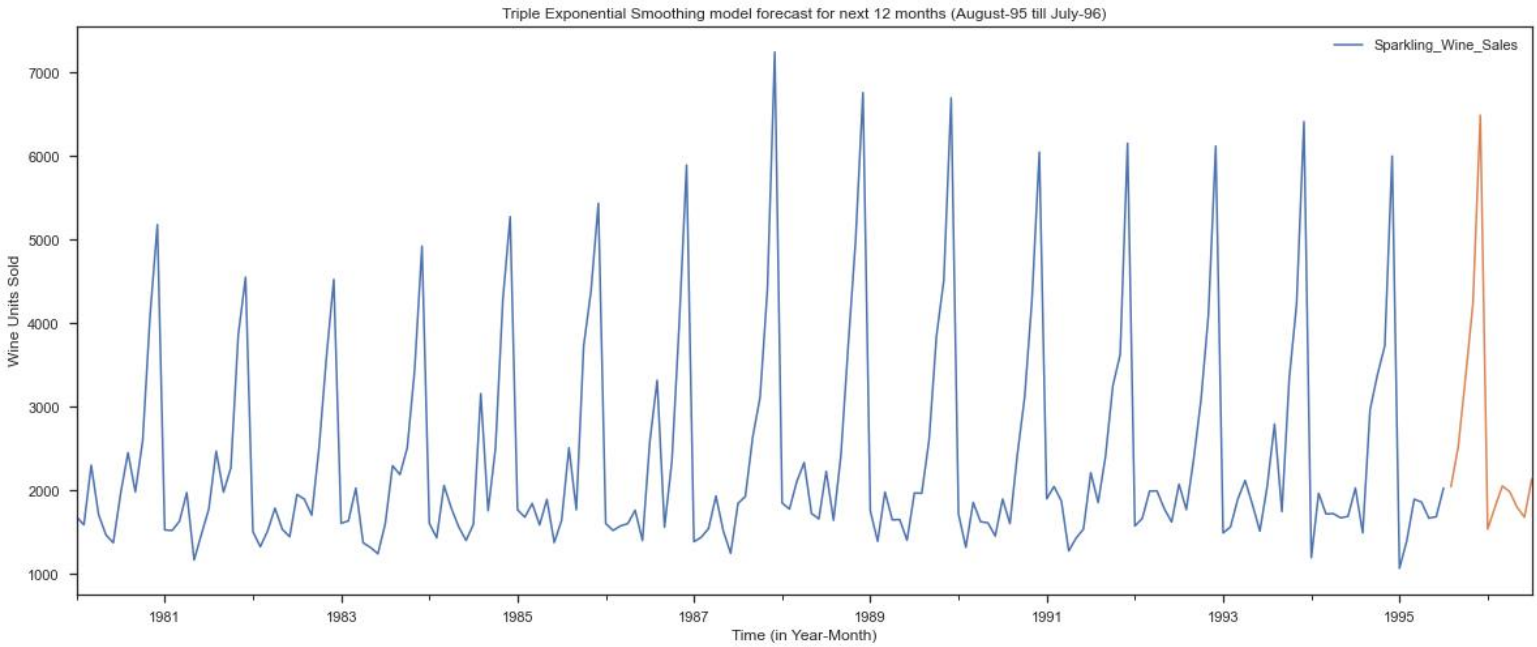


**Fig.200 TES Optimum Model – Time series plot forecast for next 12 months**

| | lower_ci | prediction | upper_ci |
|---|---|---|---|
| **1995-08-31** | 1315.075377 | 2045.552999 | 2776.030621 |
| **1995-09-30** | 1797.169146 | 2527.646768 | 3258.124390 |
| **1995-10-31** | 2633.221343 | 3363.698965 | 4094.176587 |
| **1995-11-30** | 3509.624822 | 4240.102444 | 4970.580066 |
| **1995-12-31** | 5756.865626 | 6487.343248 | 7217.820869 |
| **1996-01-31** | 809.553991 | 1540.031613 | 2270.509235 |
| **1996-02-29** | 1073.969673 | 1804.447295 | 2534.924917 |
| **1996-03-31** | 1324.544997 | 2055.022618 | 2785.500240 |
| **1996-04-30** | 1255.619969 | 1986.097591 | 2716.575213 |
| **1996-05-31** | 1071.780758 | 1802.258380 | 2532.736001 |
| **1996-06-30** | 951.144337 | 1681.621958 | 2412.099580 |
| **1996-07-31** | 1405.651750 | 2136.129372 | 2866.606994 |

**Fig.201 TES Optimum Model – Future forecast with confidence intervals**

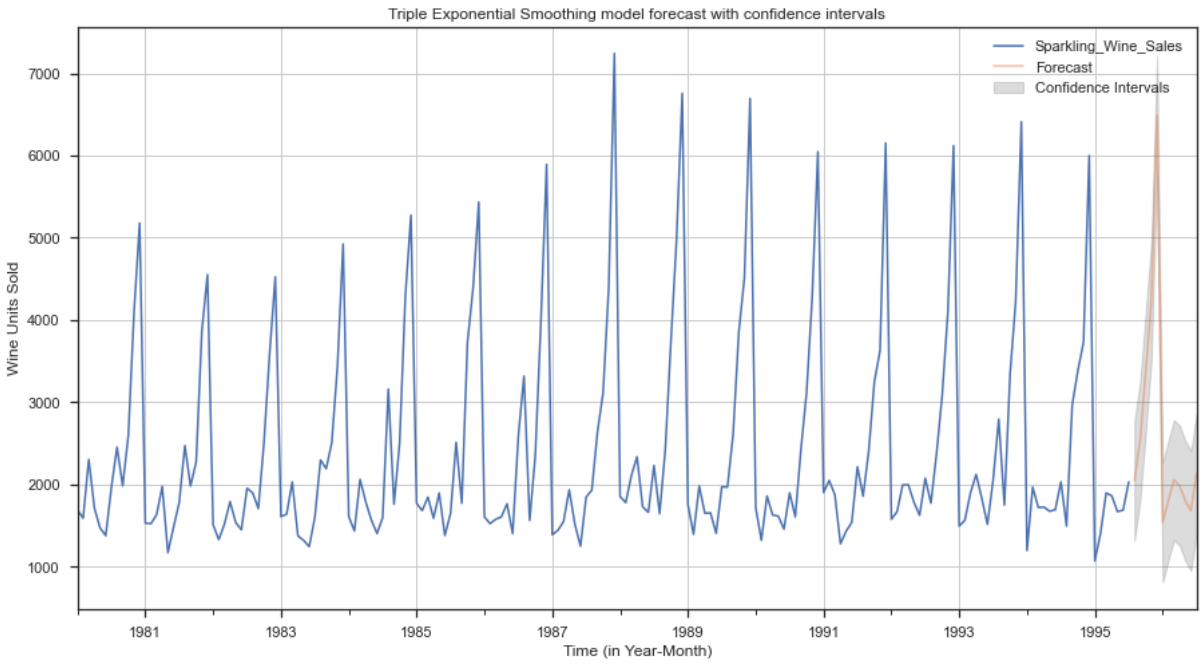


**Fig.202 TES Optimum Model – Time series plot forecast with confidence intervals**

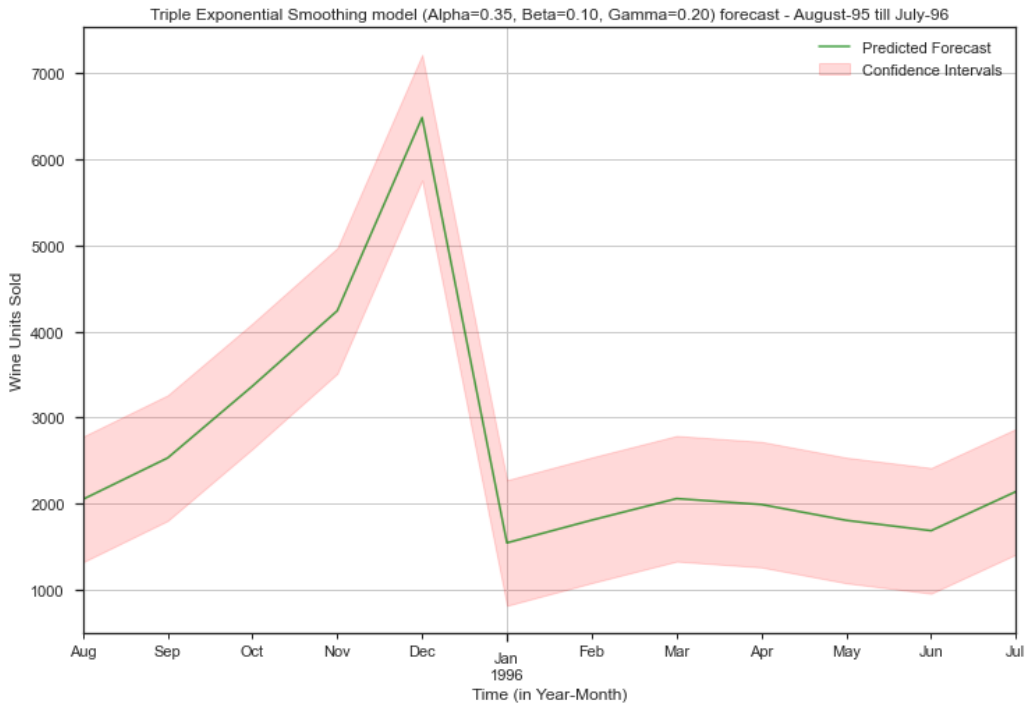


**Fig.203 TES Optimum Model – Forecast for next 12 months with confidence intervals**

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

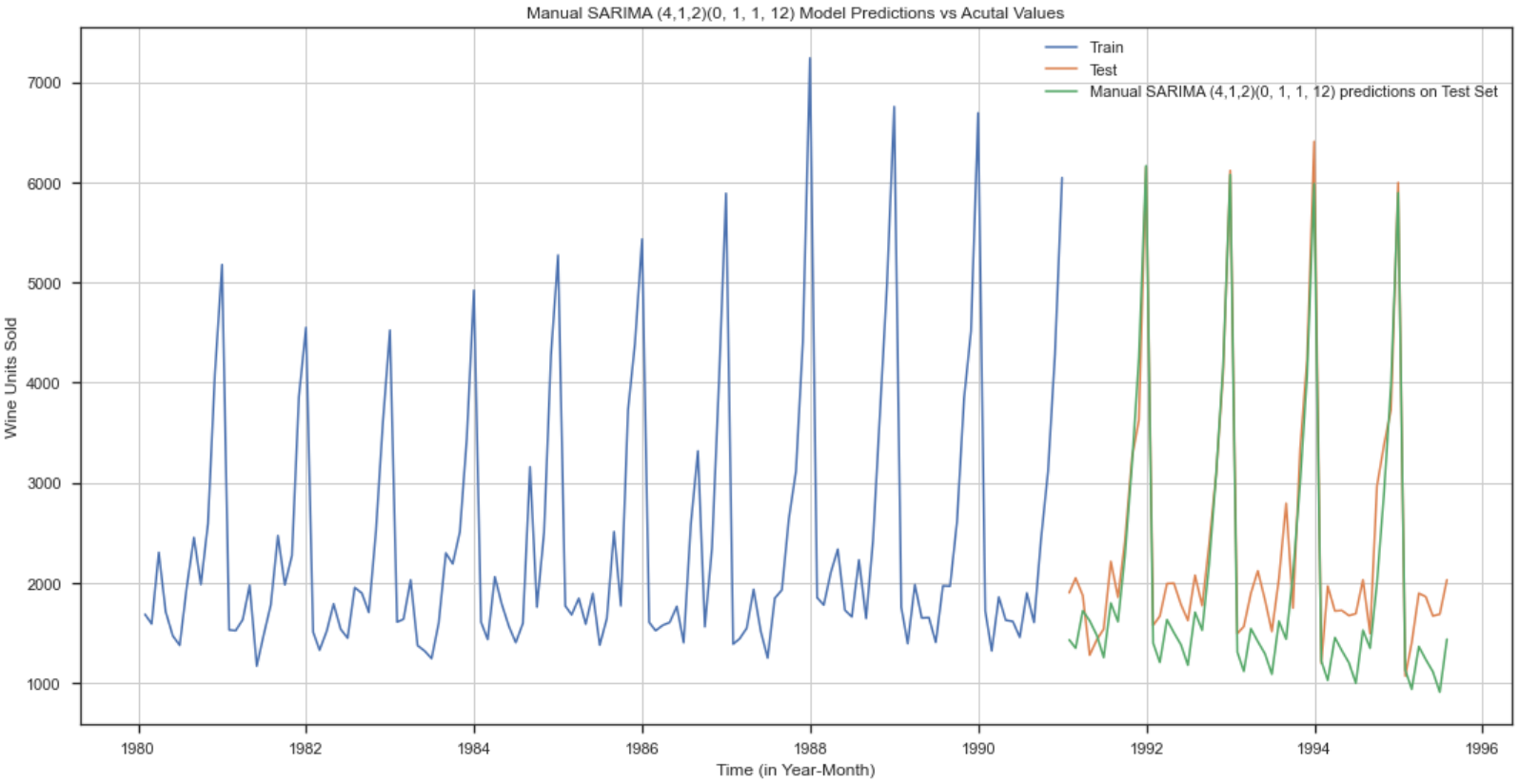


**Fig.204 Manual SARIMA Optimum Model – Line plot of Predictions vs Actual values**

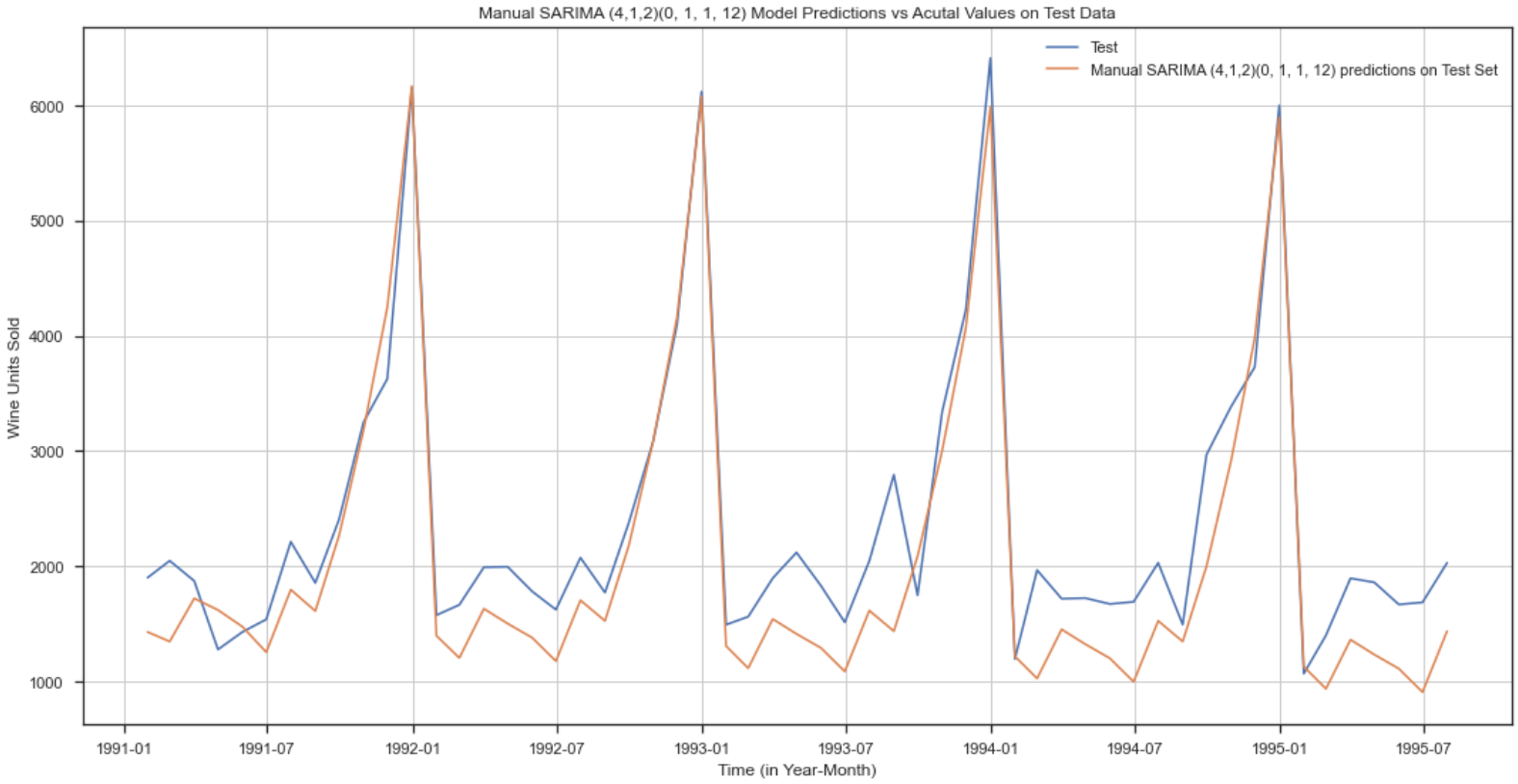


**Fig.205 Manual SARIMA Optimum Model – Line plot of Predictions vs Actual values on Test data**

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sparkling_Wine_Sales | No. Observations: | 187 |
| Model: | SARIMAX(4, 1, 2)x(0, 1, [1], 12) | Log Likelihood | -1171.945 |
| Date: | Sun, 23 Oct 2022 | AIC | 2359.889 |
| Time: | 12:52:30 | BIC | 2384.441 |
| Sample: | 01-31-1980 | HQIC | 2369.859 |
| | - 07-31-1995 | | |
| 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).

**Fig.206 Manual SARIMA Optimum Model**

| Sparkling_Wine_Sales | mean | mean_se | mean_ci_lower | mean_ci_upper |
|---|---|---|---|---|
| 1995-08-31 | 1850.816055 | 381.838092 | 1102.427147 | 2599.204963 |
| 1995-09-30 | 2396.196691 | 387.330610 | 1637.042644 | 3155.350737 |
| 1995-10-31 | 3227.083904 | 387.331727 | 2467.927669 | 3986.240139 |
| 1995-11-30 | 3898.042175 | 389.053921 | 3135.510502 | 4660.573848 |
| 1995-12-31 | 6099.973063 | 389.124031 | 5337.303976 | 6862.642150 |
| 1996-01-31 | 1231.641762 | 389.124033 | 468.972671 | 1994.310852 |
| 1996-02-29 | 1559.208853 | 389.811385 | 795.192578 | 2323.225129 |
| 1996-03-31 | 1801.347895 | 390.840893 | 1035.313821 | 2567.381970 |
| 1996-04-30 | 1780.038171 | 392.386202 | 1010.975346 | 2549.100996 |
| 1996-05-31 | 1627.831142 | 394.277144 | 855.062140 | 2400.600144 |
| 1996-06-30 | 1570.336070 | 396.011294 | 794.168195 | 2346.503944 |
| 1996-07-31 | 1979.251254 | 397.569765 | 1200.028834 | 2758.473674 |

**Fig.207 Manual SARIMA Model – Forecast for next 12 months with confidence intervals**

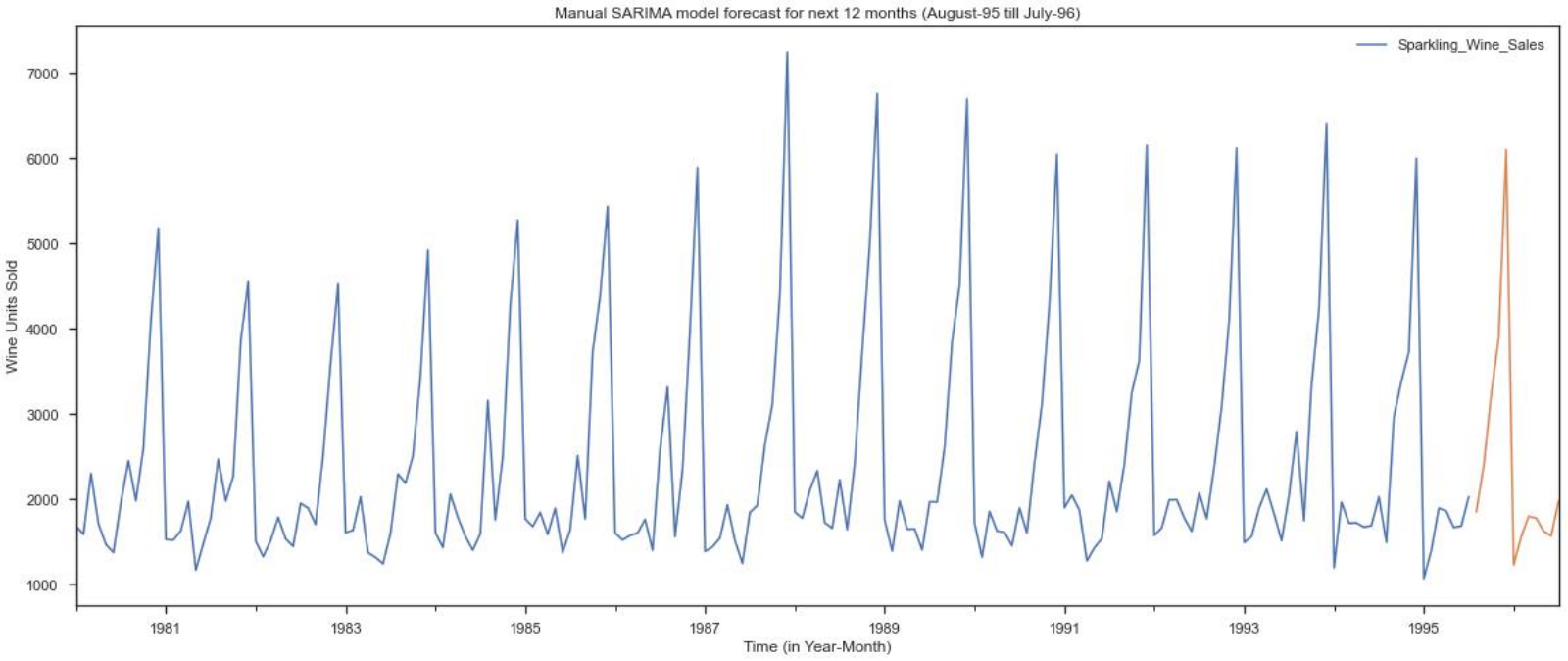


**Fig.208 Manual SARIMA Optimum Model – Time series plot forecast for next 12 months**

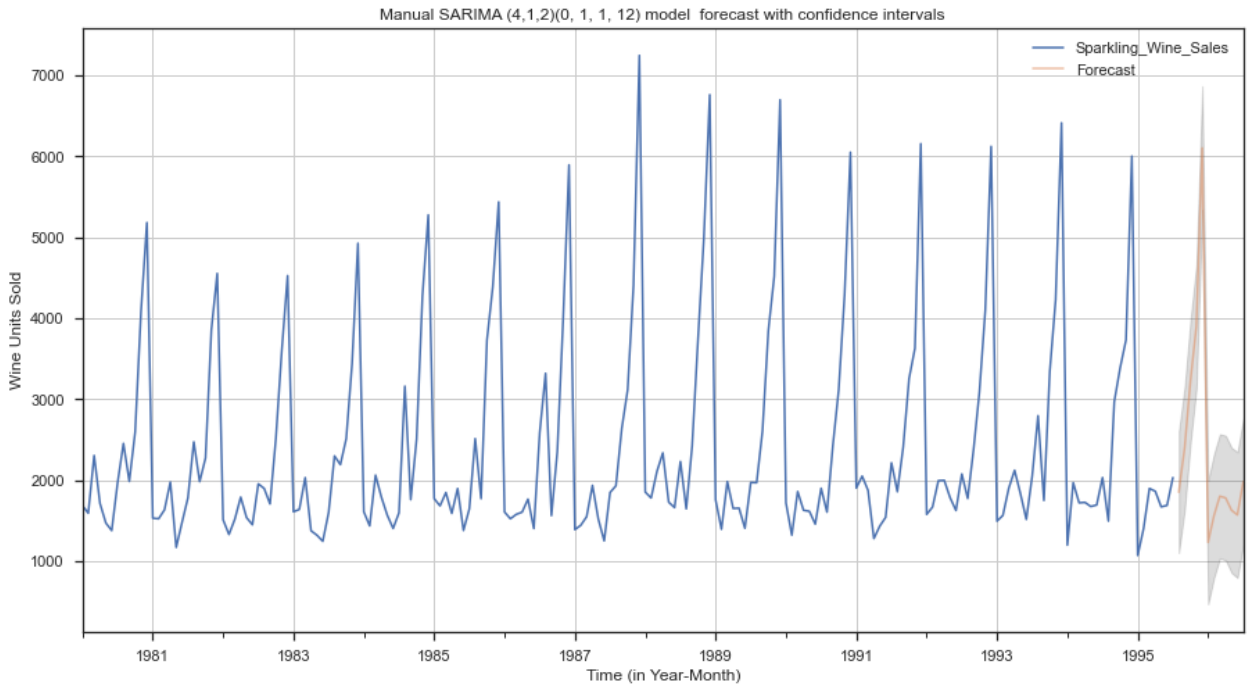


**Fig.209 Manual SARIMA Optimum Model – Time series plot forecast with confidence intervals**

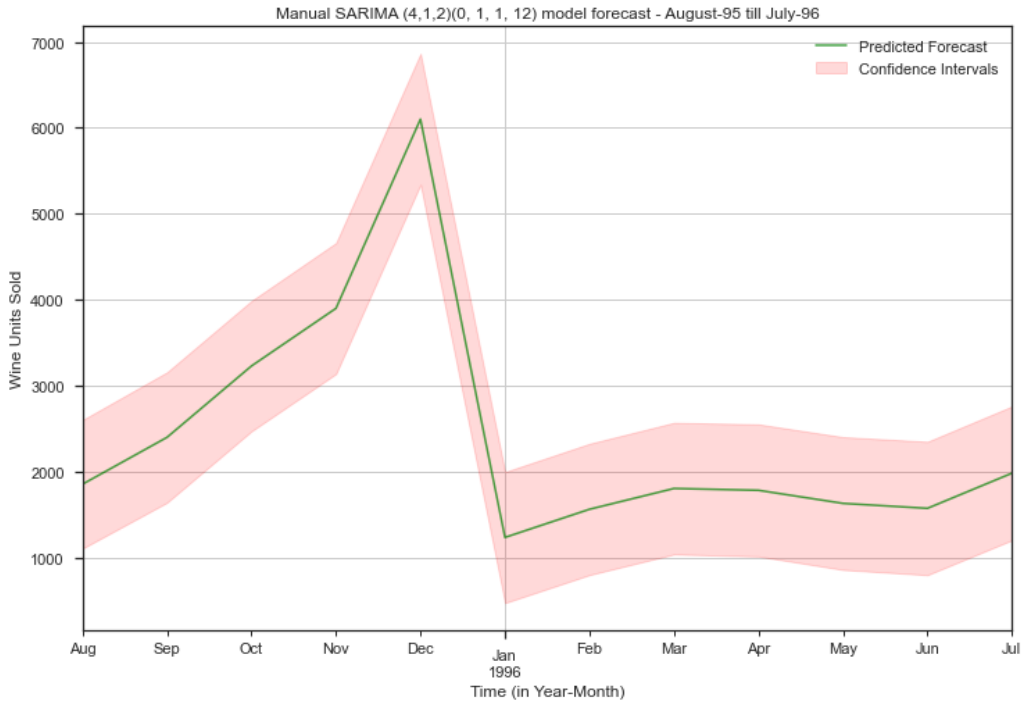


**Fig.210 Manual SARIMA Optimum Model – Forecast for next 12 months with confidence interval**

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

**We needed to construct an optimum model to forecast the sparkling wine sales for the next 12 months. The model information, insights and recommendations are as follows.**

### Model Insights:

- The time series in consideration exhibits a little rising trend and stable seasonality. When comparing the various models, we can see that **Triple Exponential Smoothing and SARIMA models frequently deliver the greatest results**. This is due to the fact that these models are **excellent at predicting time series that demonstrate trend and seasonality**.
- We examine the **root mean squared value of the forecast model to assess its performance (RMSE)**. The model with the lowest RMSE value and characteristics that match the test data is regarded as being a superior model.
- We observed that **SARIMA and the Triple Exponential Smoothing** model had the lowest RMSE and the characteristics that most closely fit test data. As a result, they are regarded as the **best models for forecasting**.
- The firm may use the aforementioned best forecasting models since they can accurately collect time series data and allow for proactive action based on the forecast.

### Historical Insights:

- The sparkling wine **sales have remained stable throughout time**. Sparkling wine sales **peaked in 1988** and fell to their **present low position in 1995** (as we have data for only first 7 months).
- The monthly sales trajectory appears to be exactly the opposite of the yearly plot, with a progressive increase towards the end of each year. **January has the lowest wine sales**, while **December has the highest**. From January to August, sales increase gradually, and then they quickly increase after that.
- The **average monthly sales** of sparkling wine are **2402 bottles**. More than 50% of the sold units of sparkling wine fall between 1605 and 2549. **1070 units were sold as the lowest** and **7242 units as the most**. Only 25% of monthly sales that were recorded were for more than 2549 units.

- Around **60 to 70 percent of the units sold are fewer than 2500**, and 80% of the units sold are less than 4000. Only 20% of sales involved more than 3000 items. Therefore, it is clear that the **bulk of sales were in the range of 1000 to 3000 units**.

## Forecast Insights:

- **Based on the forecast made by the Triple Exponential Smoothing model previously presented, the following insights are offered.**
- The forecast calls for average sale of 2639 units, up 237 units from the historical average of 2402 units. Thus, we might observe an **increase in average sales of 10%**.
- The prediction is for a minimum sales volume of 1540 units, which is 470 units more than the minimum sales volume of 1070 units in the past. Consequently, a **43% increase in minimum sales is seen**.
- The projection estimates a maximum sales volume of 6487 units, which is 755 units fewer than the largest sales volume recorded in the past, which was 7242 units. Consequently, a **10% decrease in maximum sales is visible**.
- In comparison to the historical standard deviation of 1295 recorded in the past, the **forecast's standard deviation** is 1439 units, or 144 units higher. It's **gone up by 11%.** This is also anticipated because historical data tends to have less volatility than future data.
- We can see from the prediction that the months of **October, November, and December have increased sales**. **December is often when the sales are at their highest**. There is a **startling decline in sales in January following December**. The **months after January appear to witness a gradual improvement in sales until October, when it jumps sharply**.

## Recommendations:

- Records show that the months of September, **October, November, and December account for 50% of the total sales forecast**. **Many festivities** take place in these months, and **many people travel during this time**. One of the most popular types of wine used during festive and event celebrations is sparkling wine.
- Wine sales often climb in the final two months of the year as **people hurry to buy holiday beverages**. For forthcoming **occasions like Thanksgiving, Christmas, and New Year's, people typically stock up**. The majority of individuals also **buy in bulk for holiday gatherings and gift-giving.**

- **Many individuals choose wine as their go-to gift** when it comes to occasions like parties and gift-giving. Sales of sparkling wine rise just before the winter holidays as more collectors purchase these wines as presents or look for vintages to serve at holiday gatherings.
- The festival seasons may vary depending on where you are geographically, however the most of the celebrations take place in the last four months.
    - **In these months, promotional offers might be implemented to lower costs and significantly boost revenue.**
    - **To increase sales, we must take advantage of all holiday events and set prices appropriately.**
    - **Many individuals order in bulk to prepare for upcoming festivities, which may result in a high shipping expenditure. Businesses may provide significant discounts or free shipping beyond a certain threshold at these times.**
    - **Giving customers gifts to improve their user experience is one of the greatest marketing strategies to deploy. In order to attract more consumers and increase sales, the company might provide free gifts on orders with significant sales.**
    - **To target various client demographics, the proper marketing campaigns must be run**
    - **Numerous ecommerce campaigns and competitions may be performed to broaden the product's audience and enhance sales.**
- The period **from January to June is one of the key challenges** for sparkling wine sales.
    - **To identify the elements affecting sales, in-depth market research must be conducted.**
    - **Due to the fact that sparkling wines are typically used while celebrating, a market-friendly version of the existing product might be introduced by the company, helping to make up for the drop in sales. Long-term, this may bring in additional clients.**
- There are other key elements that might be driving the sales, despite the present model's ability to closely track the historical sales trend.
    - **The forecast might be improved by doing in-depth market research on the factors that influence sales and incorporating that information into the model for projection.**