TSF PROJECT
Sparkling Dataset

DSBA

BUSINESS REPORT

Sayyed Abdul Khaliq
Email : abdulkhaliq01112001@gmail.com

CONTENTS

List of Figures

List of Tables

Problem 1

Problem statement:

As an analyst at ABC Estate Wines, we have access to historical data on the sales of various types of wines throughout the 20th century. These datasets, though originating from the same company, reflect sales figures for different wine varieties. Our goal is to explore this data to identify trends, patterns, and factors that have influenced wine sales over the century. By applying data analytics and forecasting techniques, we aim to extract actionable insights that will guide strategic decision-making and help optimize sales strategies for the future.

Context

Given the historical sales data of various wine types from ABC Estate Wines spanning the 20th century, we need to conduct a comprehensive analysis to uncover trends and patterns. Additionally, we aim to build robust forecasting models to predict future wine sales. This involves:

1. Data Exploration and Preprocessing: Cleaning and organizing the historical sales data to ensure it is suitable for analysis and modelling.
2. Trend Analysis: Identifying long-term trends, seasonal patterns, and other significant factors that have influenced wine sales over the century.
3. Forecasting: Developing predictive models to forecast future sales of different wine varieties, enabling ABC Estate Wines to make informed strategic decisions.
4. Actionable Insights: Providing recommendations based on the analysis and forecasts to help optimize sales strategies and improve market positioning.

By achieving these objectives, we will support ABC Estate Wines in understanding past sales dynamics and preparing effectively for future market conditions.

Data Description

Column name	Details
YearMonth	**Dates of sales**
Sparkling	**Sales of sparkling wine**

Data Overview

Read the data as an appropriate time series data

Data is loaded into dataframe using pandas library and first 5 and last 5 rows were printed.

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

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

Table 1 –Dataset – Rows of Data

Check the structure of the data

- Data has 187 rows and 1 column

Check the Datatypes:

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

Table 2 –Dataset– Info

Check for and treat (if needed) missing values –

There are no null values in the dataset.

Plot the data:

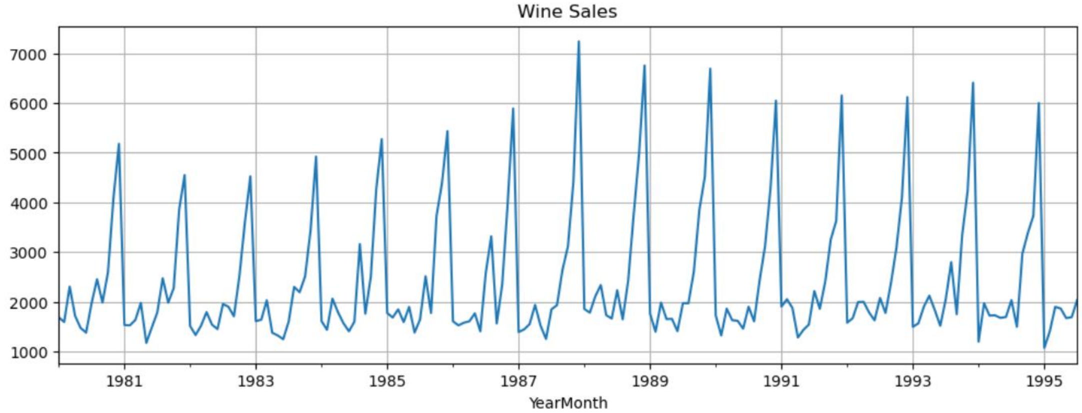

Fig 1– Time series

Perform EDA :

We have divided the dataset further by extraction month and year columns from the YearMonth column for better analysis of the dataset. The new dataset has 187 rows and 3 columns.

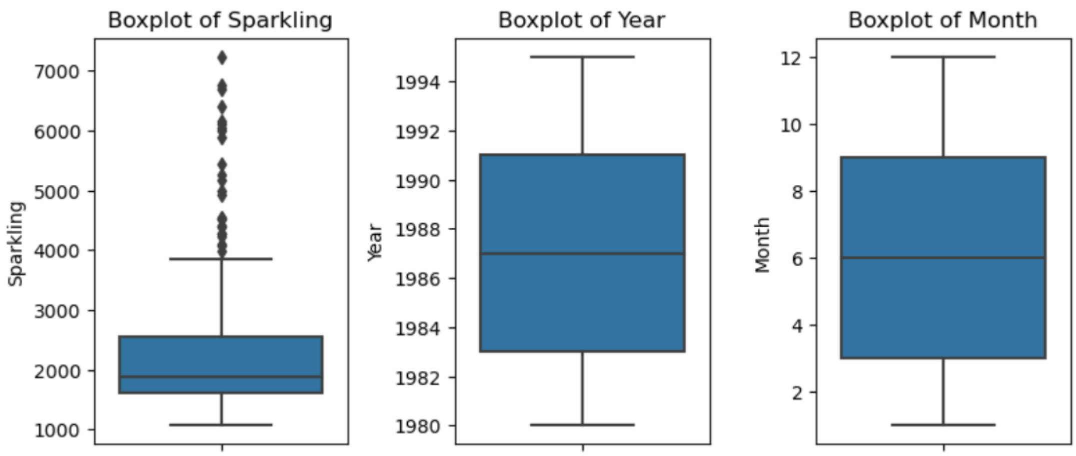

Fig 2 :Box plot of data

The box plot shows:

Sales boxplot has outliers we can treat them but we are choosing not to treat them as they do not give much effect on the time series model.

Boxplot Yearly:

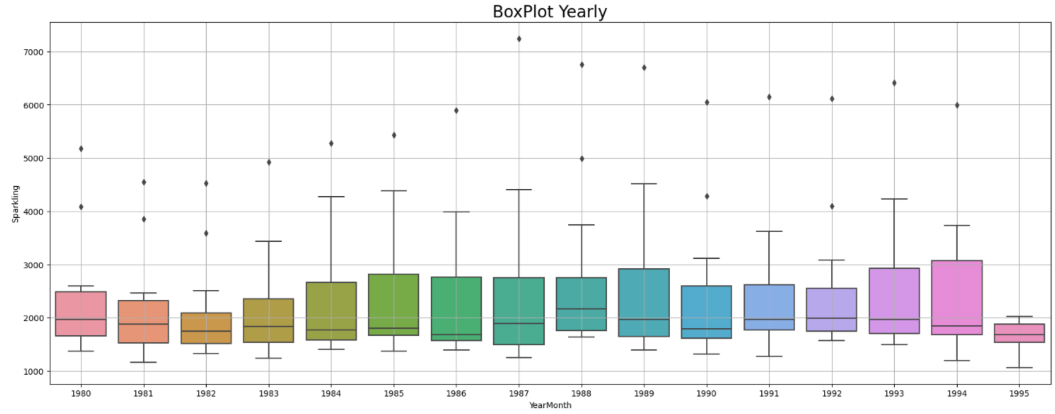

Fig 3 :Boxplot Yearly

This yearly box plot shows there is consistency over the years and there was a peak in 1988-1989. Outliers are present in all years.

Boxplot Monthly:

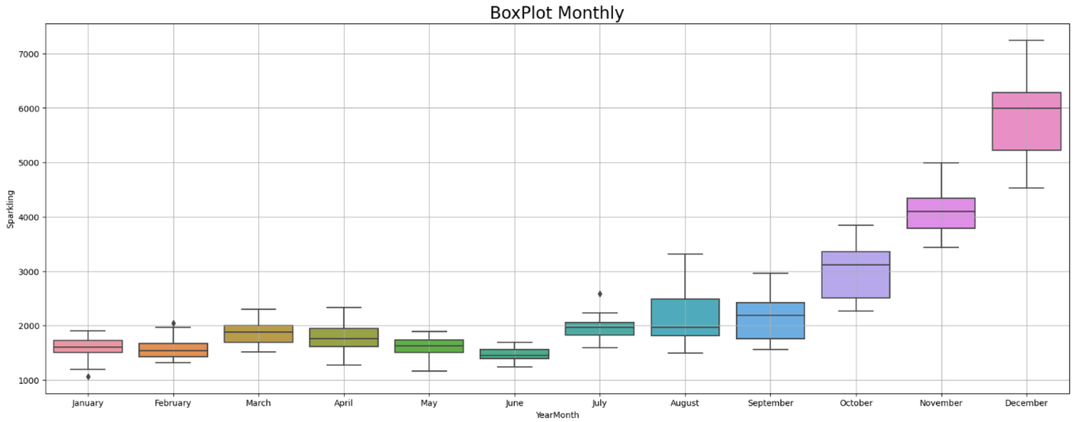

Fig 4 :Boxplot Monthly

The plot shows that sales are highest in the month of December and lowest in the month of January. Sales are consistent from January to July then from august the sales start to increase. Outliers are present in January, February and July.

Monthly Sales over the years:

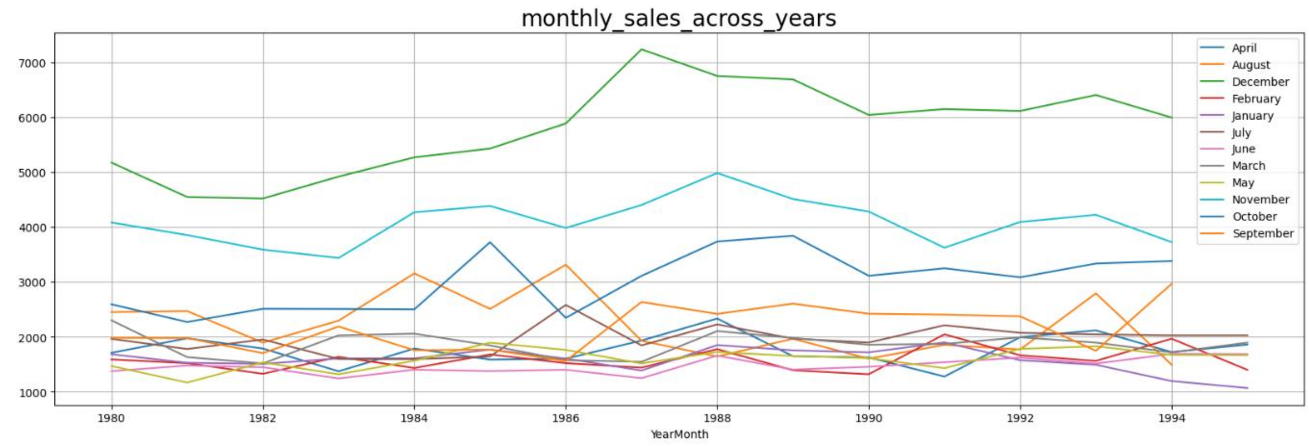

Fig 5: Monthly Sales over the years

This plot shows that December has the highest sales over the years and the year 1988 was the year with the highest number of sales.

Decomposition -Additive

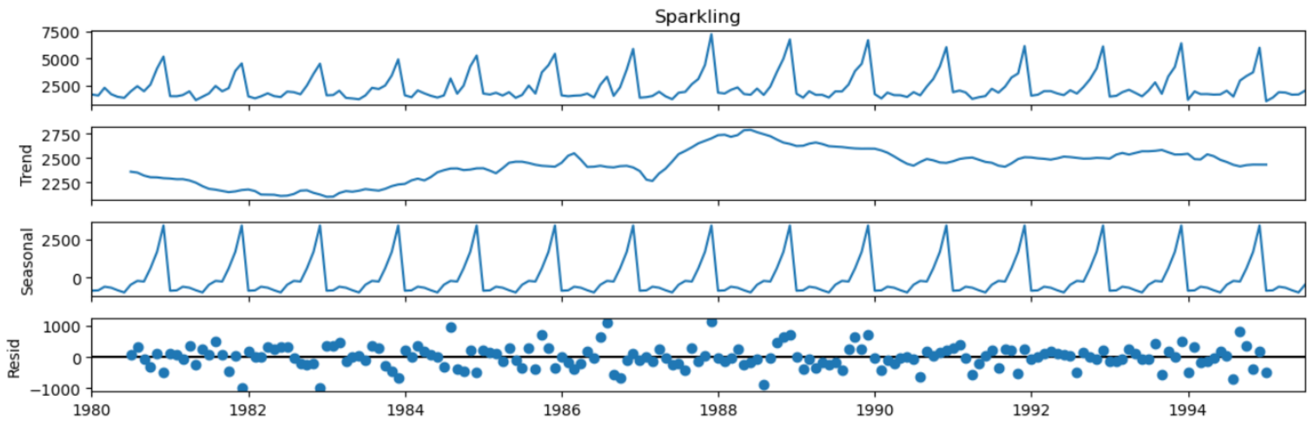

Fig 6: Decomposition -Additive

The plots show:

- Peak year 1988-1989
- It also shows that the trend has declined over the year after 1988-1989.
- Residue is spread and is not in a straight line.
- Both trend and seasonality are present

Decomposition-Multiplicative

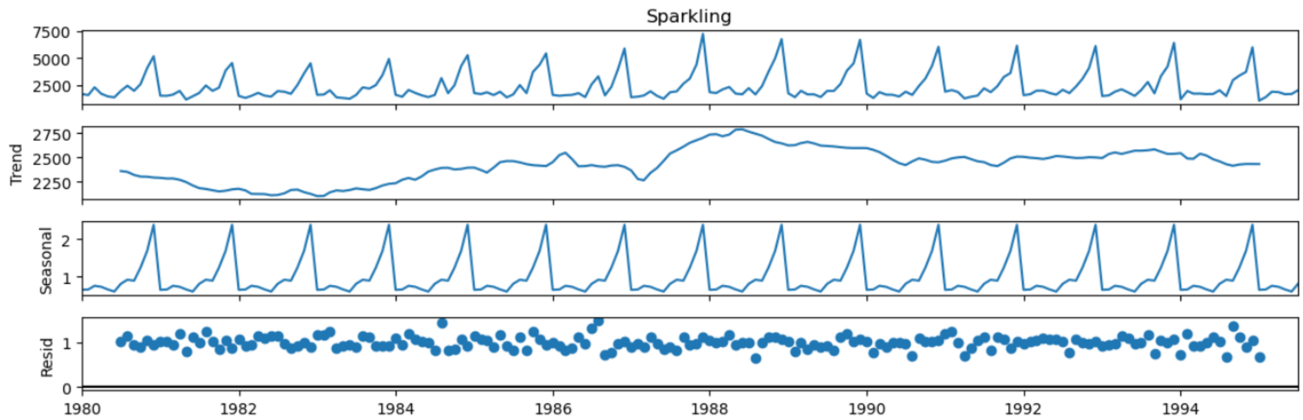

Fig 7: Decomposition – Multiplicative

The plots show

- Peak year 1988-1989
- It also shows that the trend has declined over the year after 1988-1989.
- Residue is spread and is in approx a straight line.
- Both trend and seasonality are present.
- So multiplicative model is selected owing to a more stable residual plot and lower range of residuals.

Train-test split:

In time series forecasting, splitting the data into training and testing sets is crucial for evaluating model performance. Unlike random splits in typical machine learning tasks, time series data require careful handling due to the temporal dependencies.

1. **Chronologically Split Data**: Ensure the training set precedes the test set.
2. **Typical Ratios**: Use 70-80% of the data for training, and 20-30% for testing.
3. **Prevent Data Leakage**: Always ensure the model only sees past data during training.
   This approach helps in accurately evaluating the model's forecasting ability and ensures realistic performance metrics.

Details of train and test data are as follows:

```
Shape of datasets:
train dataset:  (130, 3)
test dataset:  (57, 3)

Rows of dataset:
First few rows of Training Data
             Sparkling  Year  Month
YearMonth
1980-01-01       1686  1980      1
1980-02-01       1591  1980      2
1980-03-01       2304  1980      3
1980-04-01       1712  1980      4
1980-05-01       1471  1980      5

Last few rows of Training Data
             Sparkling  Year  Month
YearMonth
1990-06-01       1457  1990      6
1990-07-01       1899  1990      7
1990-08-01       1605  1990      8
1990-09-01       2424  1990      9
1990-10-01       3116  1990     10
```

```
First few rows of Test Data
             Sparkling  Year  Month
YearMonth
1990-11-01       4286  1990     11
1990-12-01       6047  1990     12
1991-01-01       1902  1991      1
1991-02-01       2049  1991      2
1991-03-01       1874  1991      3

Last few rows of Test Data
             Sparkling  Year  Month
YearMonth
1995-03-01       1897  1995      3
1995-04-01       1862  1995      4
1995-05-01       1670  1995      5
1995-06-01       1688  1995      6
1995-07-01       2031  1995      7
```

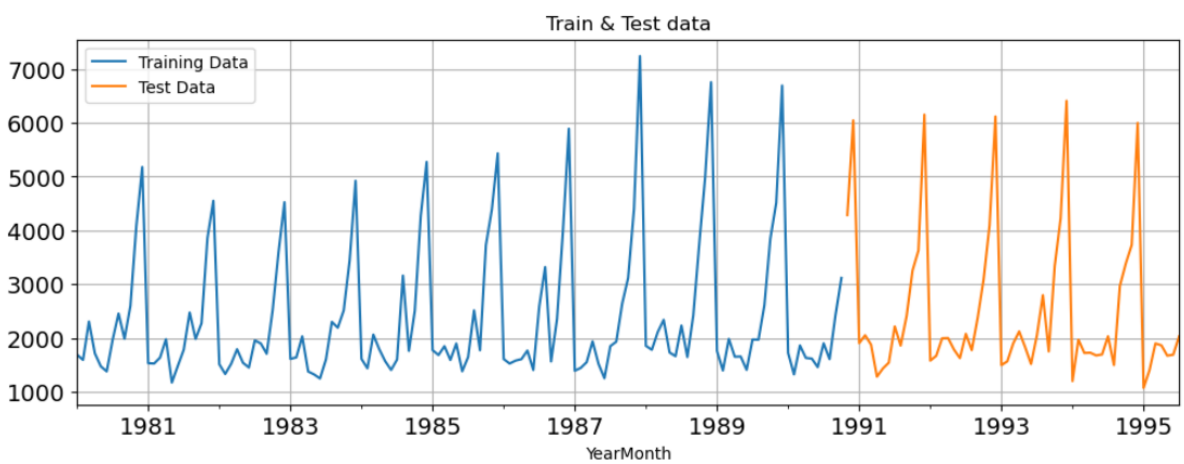

Fig 8 –Train & Test Plot

Model building:

Build forecasting models:

Evaluation Metrics:

When evaluating time series forecasting models, it's important to use metrics that capture both the accuracy of the predictions and the goodness of fit of the model. Two commonly used metrics are Root Mean Squared Error (RMSE) and Akaike Information Criterion (AIC).

1**. Root Mean Squared Error (RMSE)**

RMSE measures the average magnitude of the errors between the predicted and actual values. It is the square root of the average of squared differences between prediction and actual observation.

2. Akaike Information Criterion (AIC)

AIC is a measure of the relative quality of a statistical model for a given set of data. It balances the model's goodness of fit with its complexity, penalizing models with more parameters to avoid overfitting.

Linear regression

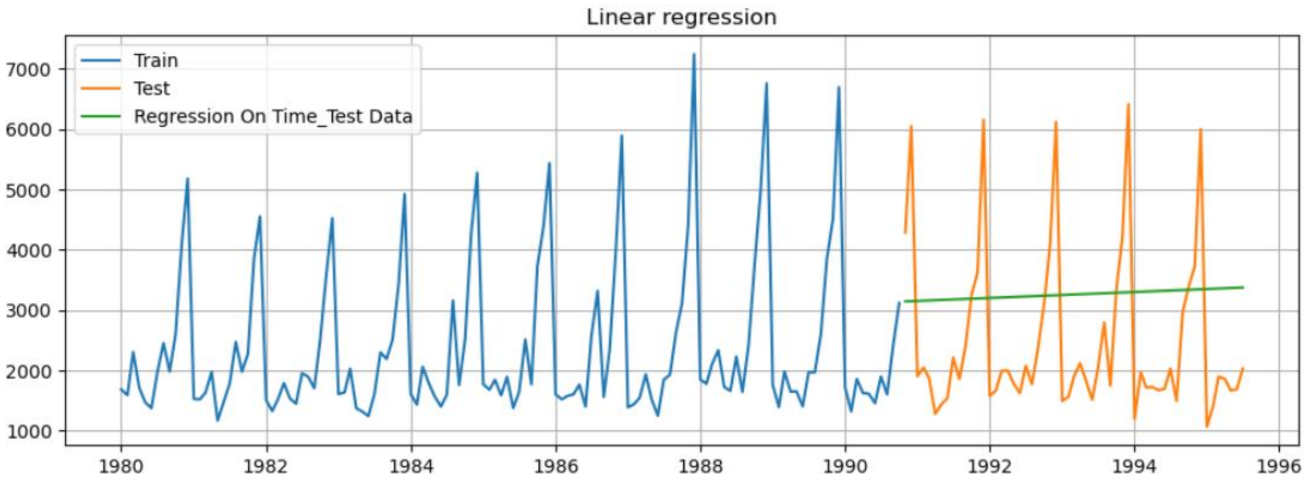

Fig 9 – Linear regression

It is clear the predicted values are very far off from the actual values

RMSE calculated for this model is 1568.048.

Simple Average Forecast

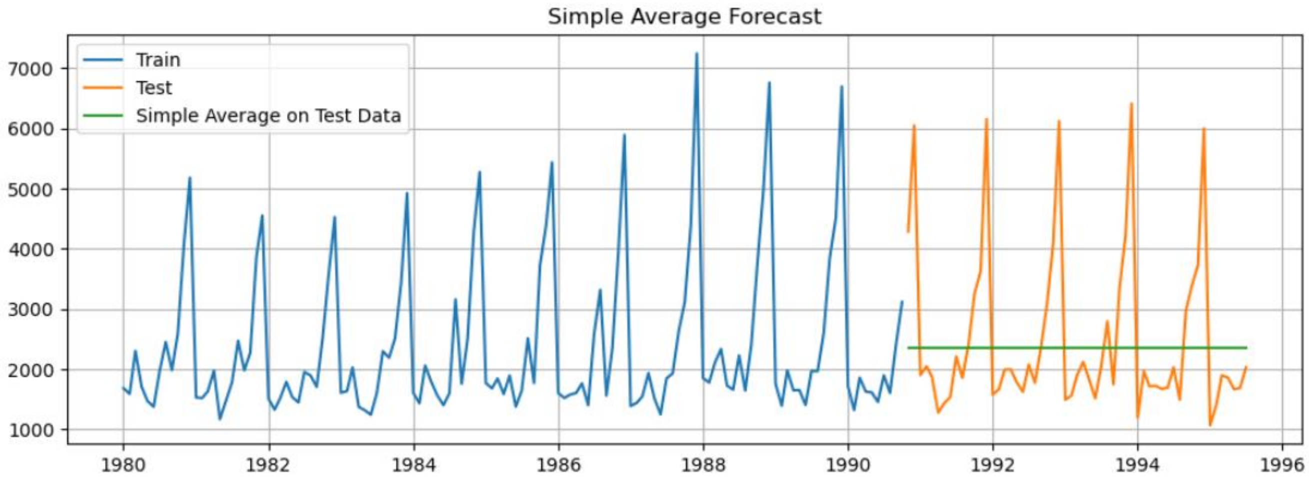

Fig 10– Simple Average Forecast

It is clear the predicted values are very far off from the actual values

RMSE calculated for this model is 1368.747

Moving Average Forecast

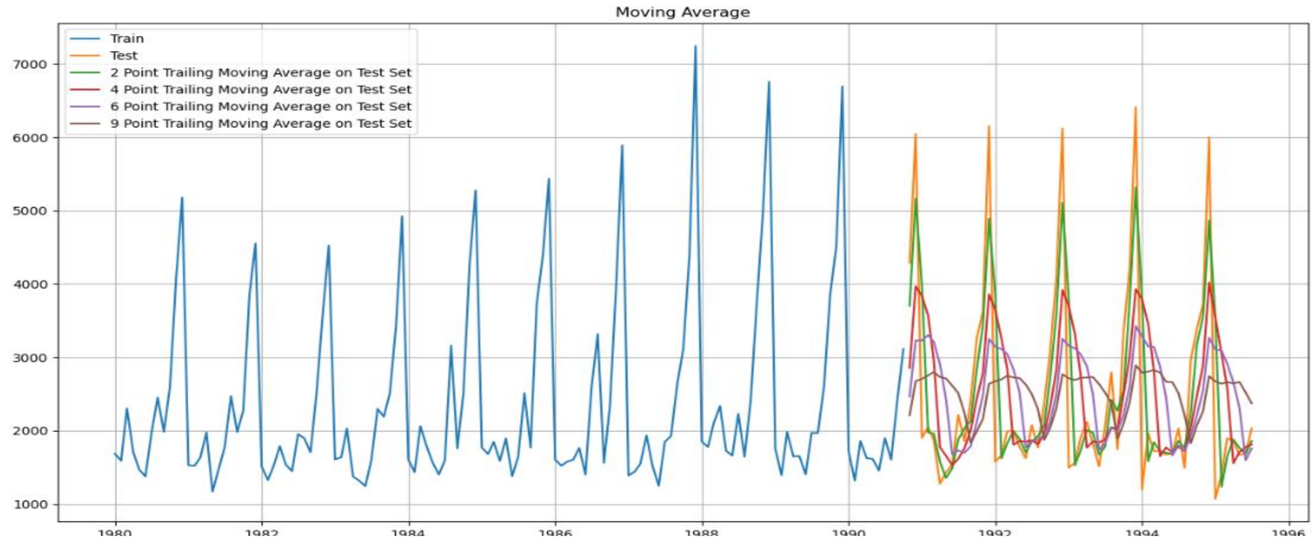

Fig 11–Moving Average Forecast

We have made multiple moving average models with rolling windows varying from 2 to 9. Rolling average is a better method than simple average as it takes into account only the previous n values to make the prediction, where n is the rolling window defined. This takes into account the recent trends and is in general more accurate. The higher the rolling window, the smoother will be its curve, since more values are being taken into account

Based on the below RMSE scores, 2 point trailing Moving Average is selected.
For 2 point Moving Average Model forecast on the Training Data, RMSE is 811.179
For 4 point Moving Average Model forecast on the Training Data, RMSE is 1184.213
For 6 point Moving Average Model forecast on the Training Data, RMSE is 1337.201
For 9 point Moving Average Model forecast on the Training Data, RMSE is 1422.653

RMSE calculated for 2 point trailing Moving Average model is 811.179

Exponential Models (Single, Double, Triple):

In time series forecasting using Exponential Smoothing models in Python's statsmodels library, setting the parameter optimized=True allows the model to automatically find the best values for the smoothing parameters (alpha, beta, gamma) that minimize the error measure (e.g., Mean Squared Error).

Simple Exponential Smoothing:

Parameters:

```
{'smoothing_level': 0.04844277717441349,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 2160.089750219884,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

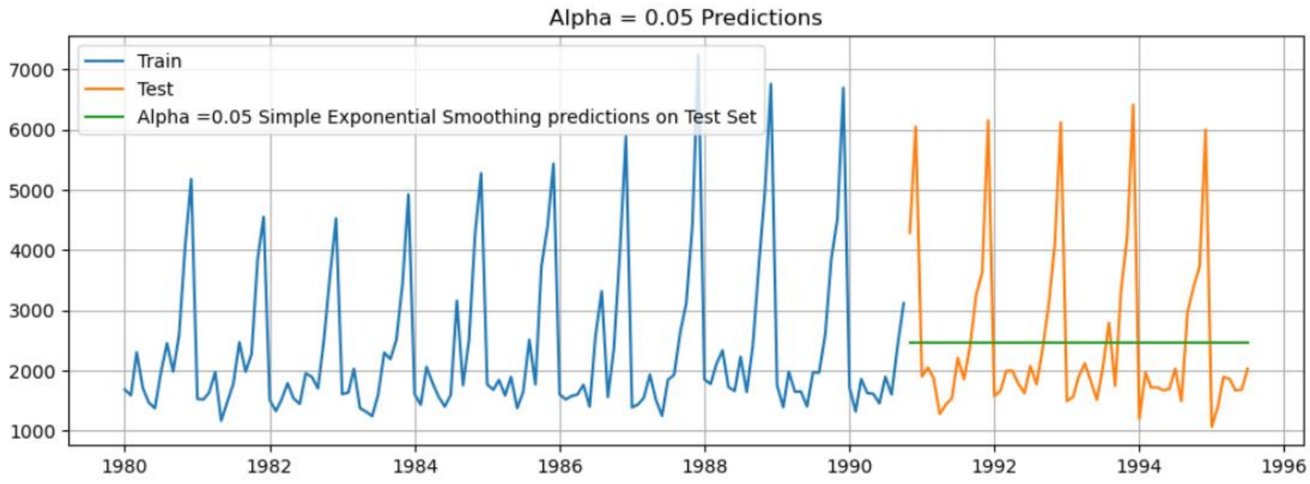

Fig 12– Simple Exponential Smoothing

It is clear the predicted values are very far off from the actual values

RMSE calculated for this model is 1362.488305

Double Exponential Smoothing:

Parameters:
{'smoothing_level': 0.07614001437835338,
'smoothing_trend': 0.07614001437835338,
'smoothing_seasonal': nan, 'damping_trend': nan,
'initial_level': 1505.8019145976457,
'initial_trend': 2.7681085036744975,
'initial_seasons': array([], dtype=float64),
'use_boxcox': False, 'lamda': None, 'remove_bias': False}

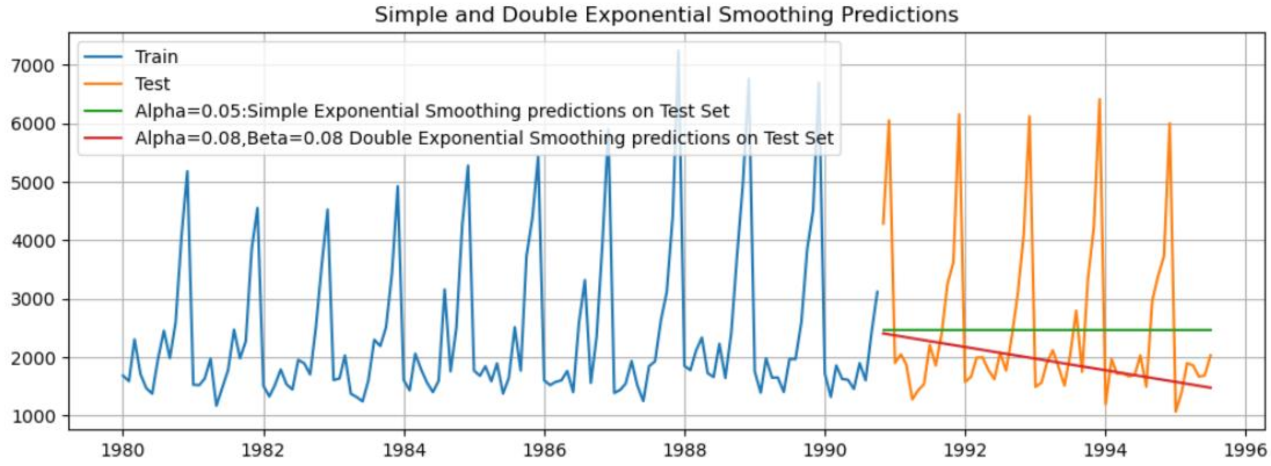

Fig 13– Double Exponential Smoothing

It is clear the predicted values are very far off from the actual values

RMSE calculated for this model is 1472.253640

Triple exponential Smoothing (Additive):

Parameters:
{'smoothing_level': 0.06836007770817487,
'smoothing_trend': 0.026396606894905476,
'smoothing_seasonal': 0.5278141355688852,
'damping_trend': nan, 'initial_level': 2320.59283142352,
'initial_trend': -0.20923379106809734,
'initial_seasons': array([-691.29424948, -766.19637996, -297.73559992, -515.60755042, -876.56975094, -881.13468478, -398.5167367 , 125.78081283, -324.38738354, 241.56351726, 1666.23565064, 2681.03990549]),
'use_boxcox': False, 'lamda': None, 'remove_bias': False}

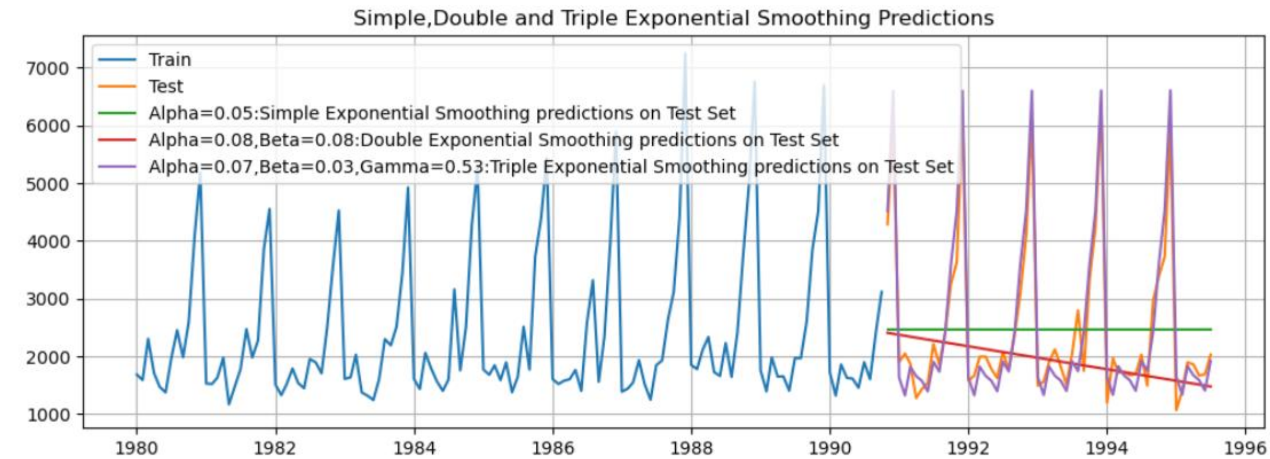

Fig 14– Triple Exponential Smoothing (Additive)

RMSE calculated for this model is 377.456200

Triple exponential Smoothing (Multiplicative):

Parameters:
{'smoothing_level': 0.07571436313248113,
'smoothing_trend': 0.06489797544827652,
'smoothing_seasonal': 0.3423280250182456,
'damping_trend': nan, 'initial_level': 2356.5416452586046,
'initial_trend': 0.9987623998615819,
'initial_seasons': array([0.72639621, 0.69425932, 0.88623802, 0.802996 , 0.66503213,0.66312537, 0.86335952, 1.09763039, 0.89304493, 1.16812344,.81009244, 2.3073598]),
'use_boxcox': False, 'lamda': None, 'remove_bias': False}

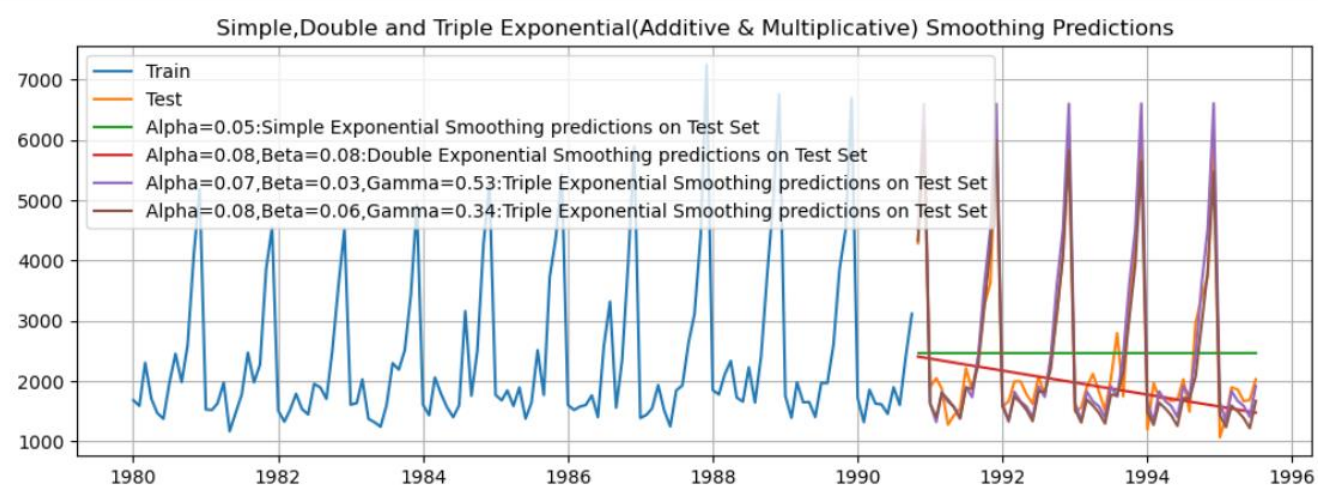

Fig 15– Triple exponential Smoothing (Multiplicative)

RMSE calculated for this model is 362.920557

Check the performance of the models built:

	Test RMSE
RegressionOnTime	1568.048196
SimpleAverageModel	1368.746717
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281
Alpha=0.05,SES	1362.488305
Alpha=0.08,Beta=0.08:DES	1472.253640
Alpha=0.07,Beta=0.03,Gamma=0.53:TES_ADD	377.456200
Alpha=0.08,Beta=0.06,Gamma=0.34:TES_Mul	362.920557

Table 3 –Model comparison

Till now, The best model had both a multiplicative trends, as well as a seasonality Model, which was evaluated using the RMSE metric (363.92).

Check for Stationarity:

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.

The hypothesis in a simple form for the ADF test is:

- H0 : The Time Series has a unit root and is thus non-stationary.
- H1 : The Time Series does not have a unit root and is thus stationary.

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

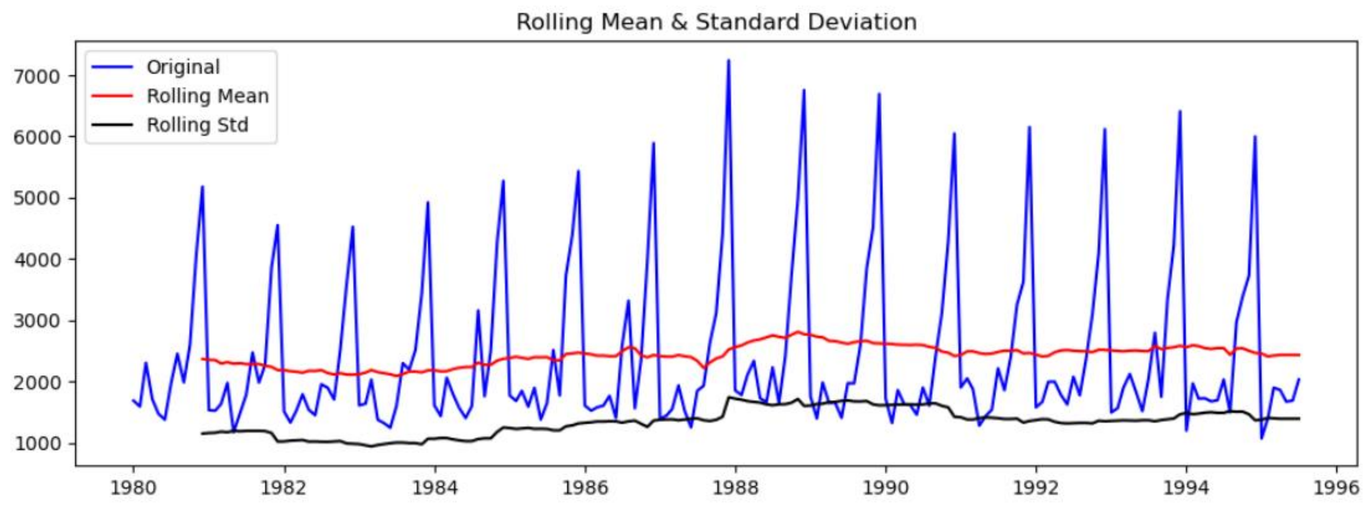

Fig 16– ADF test- Original data

Results of Dickey-Fuller Test:
Test Statistic -1.360497
p-value 0.601061
#Lags Used 11.000000
Number of Observations Used 175.000000
Critical Value (1%) -3.468280
Critical Value (5%) -2.878202
Critical Value (10%) -2.575653

The p-value 0.60 is very large, and not smaller than 0.05. We see that at 5% significant level the Time Series is non-stationary.

First-Order differencing:

In order to try and make the series stationary we used the differencing approach. We used .diff() function on the existing series without any argument, implying the default diff value of 1 and also dropped the NaN values, since differencing of order 1 would generate the first value as NaN which need to be dropped

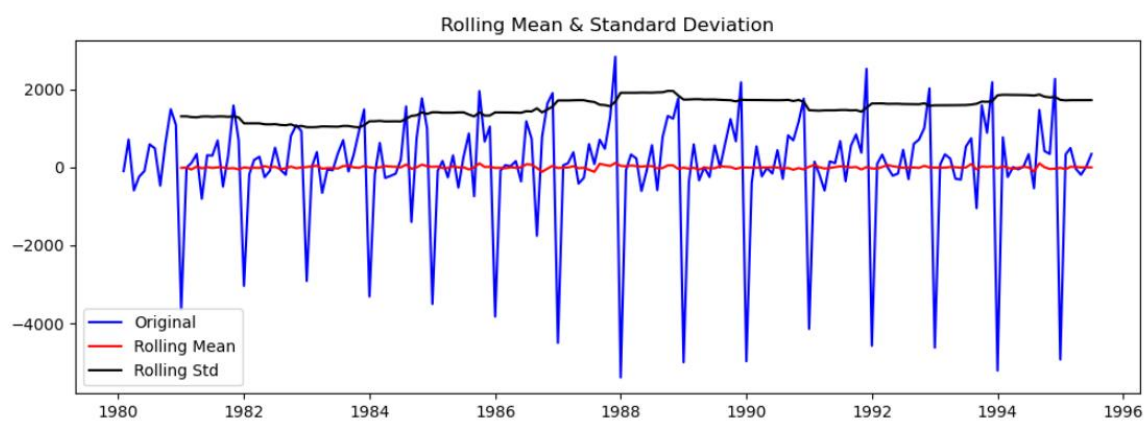

Fig 17– ADF test- Differenced data

Results of Dickey-Fuller Test:

Test Statistic	-45.050301
p-value	0.000000
#Lags Used	10.000000
Number of Observations Used	175.000000
Critical Value (1%)	-3.468280
Critical Value (5%)	-2.878202
Critical Value (10%)	-2.575653

Dickey - Fuller test was 0.000, which is obviously less than 0.05. Hence the null hypothesis that the series is not stationary at difference = 1 was rejected, which implied that the series has indeed become stationary after we performed the differencing. Null hypothesis was rejected since the p-value was less than alpha i.e. 0.05.

Also the rolling mean plot was a straight line this time around. Also the series looked more or less the same from both the directions, indicating stationarity.

Generate ACF & PACF Plot and find the AR, MA values

Original data: (Non-Stationary)

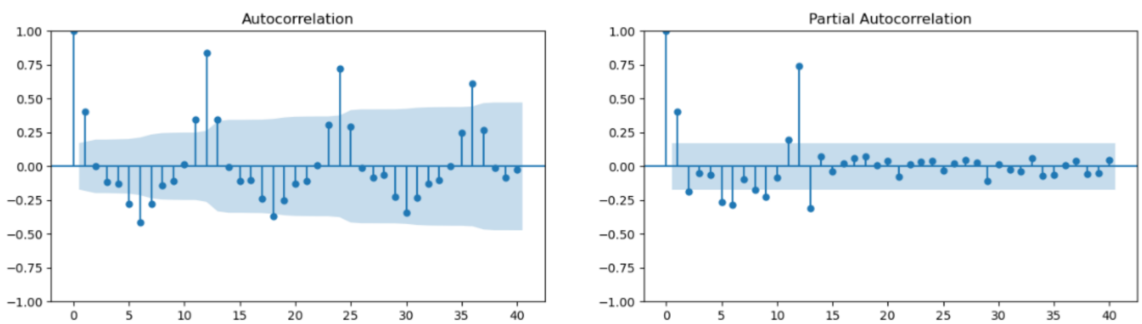

Fig 18–ACF and PACF Plots- Original data

Differenced data: (Stationary):

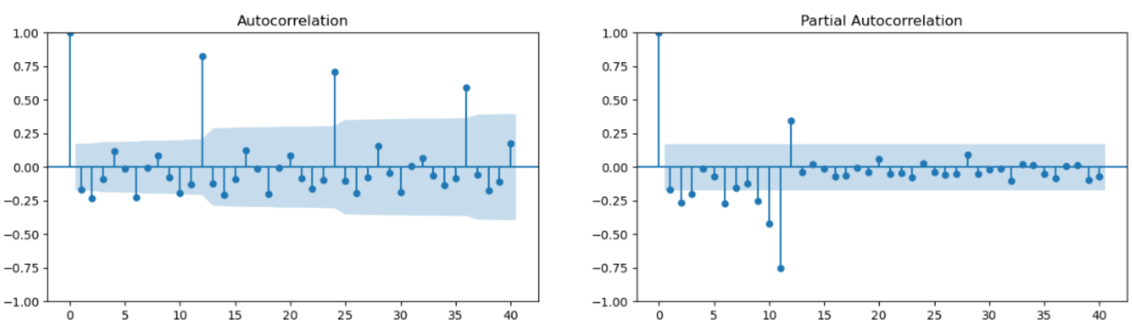

Fig 19–ACF and PACF Plots- Differenced data

- Looking at ACF plot we can see a decay after lag 2 for differenced data. hence we select the q value to be 2. i.e. q=2.
- Looking at PACF plot we can again see significant bars till lag 3 for differenced series which is stationary in nature. Hence we choose p value to be 3. i.e. p=2.
- d values will be 1 as data is stationary after first order differencing.

Manual- ARIMA Model

Based on ACF and PACF plots, the values selected for manual ARIMA:- p=3, d=1, q=2

Summary from this manual ARIMA model:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  130
Model:                 ARIMA(3, 1, 2)   Log Likelihood               -1089.814
Date:                Fri, 24 May 2024   AIC                           2191.628
Time:                        11:30:14   BIC                           2208.787
Sample:                    01-01-1980   HQIC                          2198.600
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4601      0.099     -4.658      0.000      -0.654      -0.267
ar.L2          0.3088      0.091      3.375      0.001       0.129       0.488
ar.L3         -0.2311      0.148     -1.566      0.117      -0.520       0.058
ma.L1         -0.0002     10.921  -2.13e-05      1.000     -21.406      21.405
ma.L2         -0.9998      0.148     -6.760      0.000      -1.290      -0.710
sigma2      1.204e+06      9e-06   1.34e+11      0.000     1.2e+06     1.2e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 9.67
Prob(Q):                              0.90   Prob(JB):                         0.01
Heteroskedasticity (H):               2.45   Skew:                             0.54
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.80
===================================================================================
```

Diagnostic Plots:

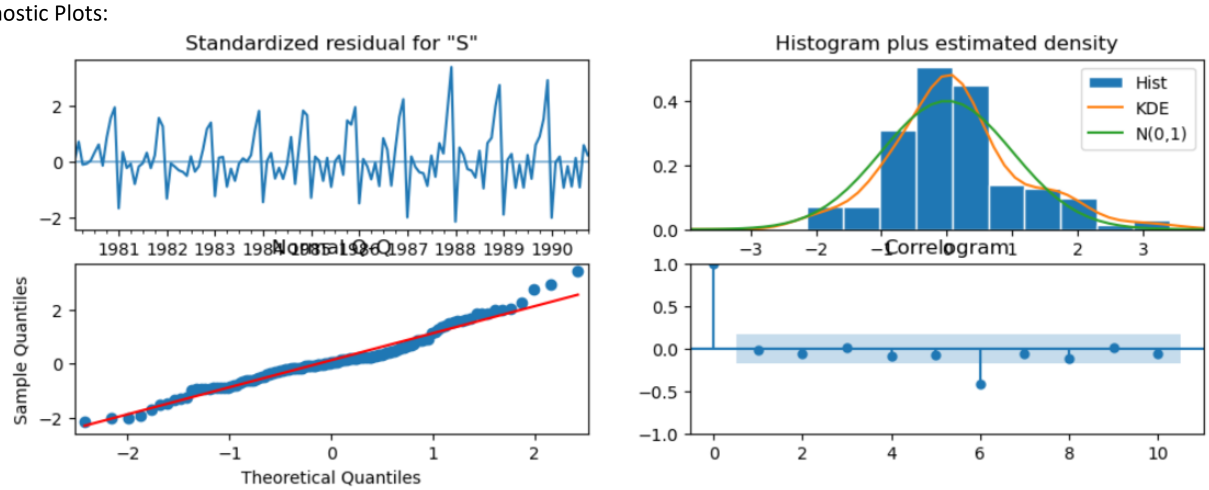

Fig 20–Manual Arima (Diagnostic Plots)

Residuals shows slightly normal distribution and there is some correlations in residuals. So it is average fit

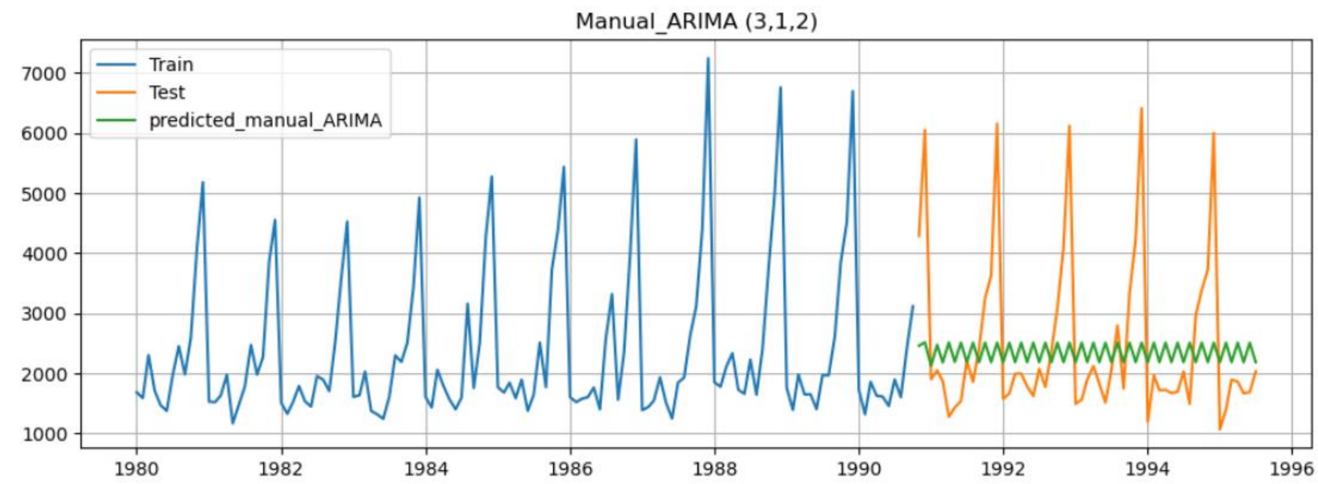

Fig 21–Manual Arima

RMSE calculated for this model is 1341.107844

Manual- SARIMA Model:

Identified Seasonal Parameters: P=1and Q=2were determined from the ACF and PACF plots.

Final parameters : SARIMAX(3, 1, 2)x(1, 1, 2, 12)

Summary from this manual SARIMA model:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  130
Model:             SARIMAX(3, 1, 2)x(1, 1, 2, 12)   Log Likelihood                -866.285
Date:                            Sat, 25 May 2024   AIC                           1750.571
Time:                                    23:51:17   BIC                           1775.431
Sample:                                         0   HQIC                          1760.664
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5841      0.274     -2.133      0.033      -1.121      -0.047
ar.L2          0.1420      0.130      1.088      0.276      -0.114       0.398
ar.L3          0.0522      0.096      0.544      0.586      -0.136       0.240
ma.L1         -0.1752      6.029     -0.029      0.977     -11.992      11.642
ma.L2         -0.8245      4.912     -0.168      0.867     -10.451       8.802
ar.S.L12      -0.9992      0.112     -8.959      0.000      -1.218      -0.781
ma.S.L12       0.5616      5.988      0.094      0.925     -11.174      12.298
ma.S.L24      -0.4384      2.645     -0.166      0.868      -5.623       4.747
sigma2      1.445e+05   4.98e-05    2.9e+09      0.000    1.45e+05    1.45e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                22.98
Prob(Q):                              0.83   Prob(JB):                         0.00
Heteroskedasticity (H):               2.56   Skew:                             0.64
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.76
===================================================================================
```

Diagnostic Plots:

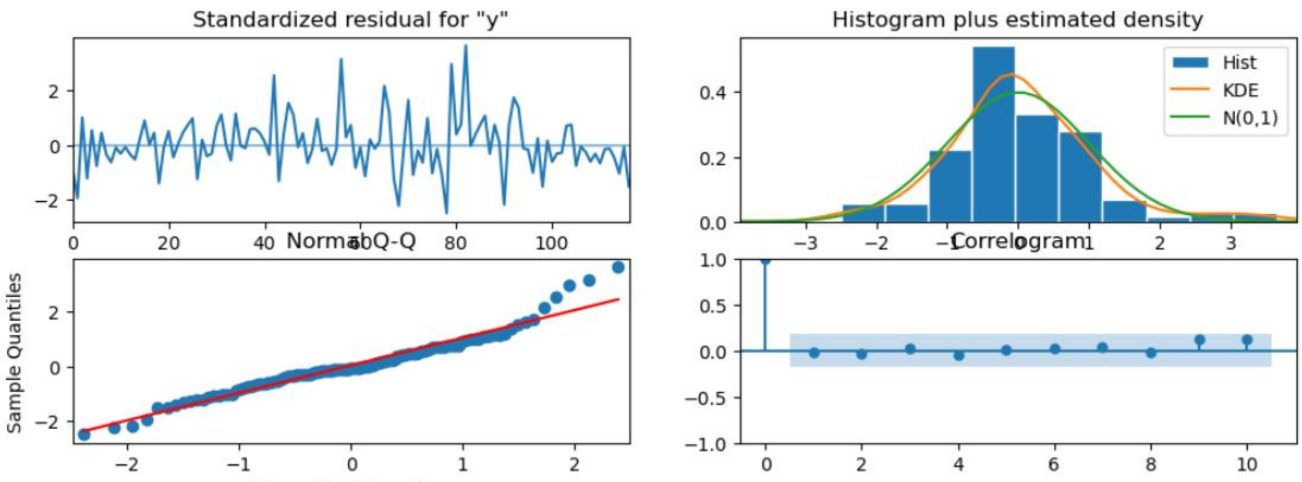

Fig 22–Manual Sarima (Diagnostic Plots)

Residuals shows normal distribution and no correlations. So it is good fit

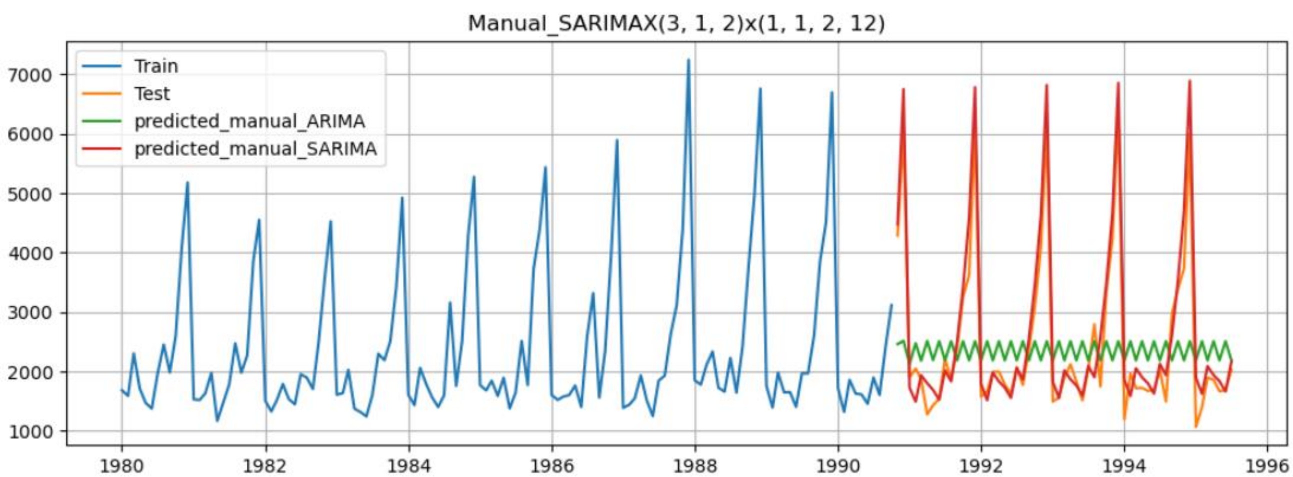

Fig 23–Manual Sarima

RMSE calculated for this model is 405.505423

Auto ARIMA:

We employed a for loop for determining the optimum values of p,d,q,

where p is the order of the AR (Auto-Regressive) part of the model,
while q is the order of the MA (Moving Average) part of the model.
d is the differencing that is required to make the series stationary.

p,q values in the range of (0,4) were given to the for loop, while a fixed value of 1 was given for d, since we had already determined d to be 1, while checking for stationarity using the ADF test.

Some parameter combinations for the Model...

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

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected.

Best Model:

	param	AIC
10	(2, 1, 2)	2178.109723
15	(3, 1, 3)	2182.815229
14	(3, 1, 2)	2191.627911
11	(2, 1, 3)	2193.824214
9	(2, 1, 1)	2193.974962
2	(0, 1, 2)	2194.034361
3	(0, 1, 3)	2194.449267
6	(1, 1, 2)	2194.959653
13	(3, 1, 1)	2195.740386
7	(1, 1, 3)	2195.939241
5	(1, 1, 1)	2196.050086
1	(0, 1, 1)	2217.939227
12	(3, 1, 0)	2220.460084
8	(2, 1, 0)	2223.899470
4	(1, 1, 0)	2231.137663
0	(0, 1, 0)	2232.719438

Table 4 – Auto arima models

Based on the above results , Parameters of the best model are (2,1,2)

The summary report for the ARIMA model with values (p=2,d=1,q=2).

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  130
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1084.055
Date:                Fri, 24 May 2024   AIC                           2178.110
Time:                        11:30:30   BIC                           2192.409
Sample:                    01-01-1980   HQIC                          2183.920
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3020      0.046     28.543      0.000       1.213       1.391
ar.L2         -0.5360      0.079     -6.763      0.000      -0.691      -0.381
ma.L1         -1.9916      0.109    -18.213      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.104      0.000       0.785       1.215
sigma2      1.085e+06   2.03e-07   5.35e+12      0.000    1.08e+06    1.08e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                19.54
Prob(Q):                              0.75   Prob(JB):                         0.00
Heteroskedasticity (H):               2.30   Skew:                             0.71
Prob(H) (two-sided):                  0.01   Kurtosis:                         4.27
===================================================================================
```

Diagnostic plots:

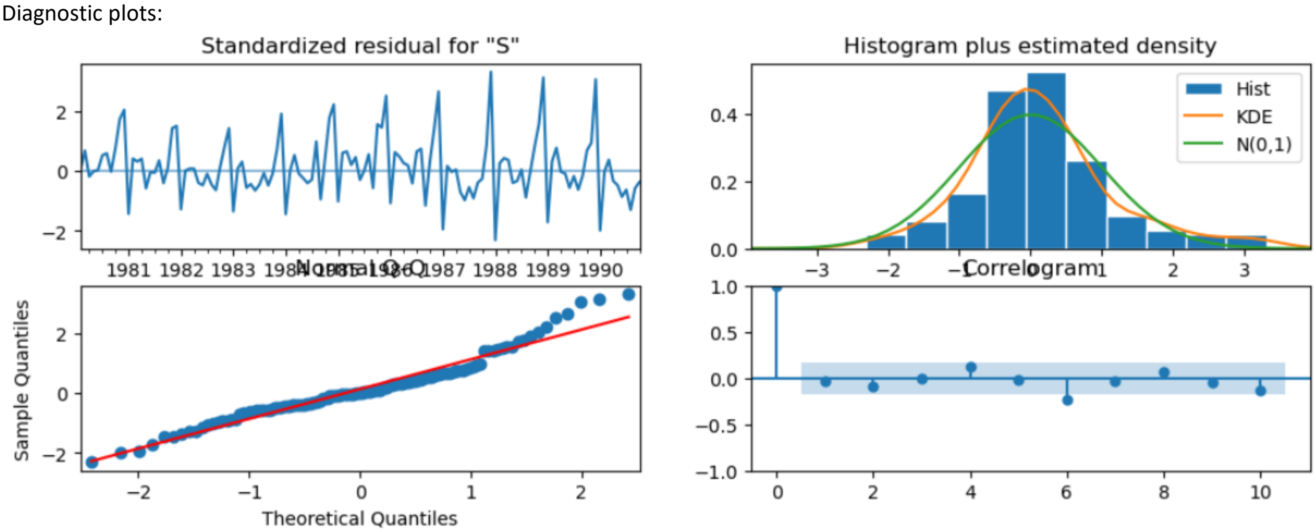

Fig 24 –Auto arima- Diagnostic plots

Residuals shows normal distribution and no correlations. So it is good fit

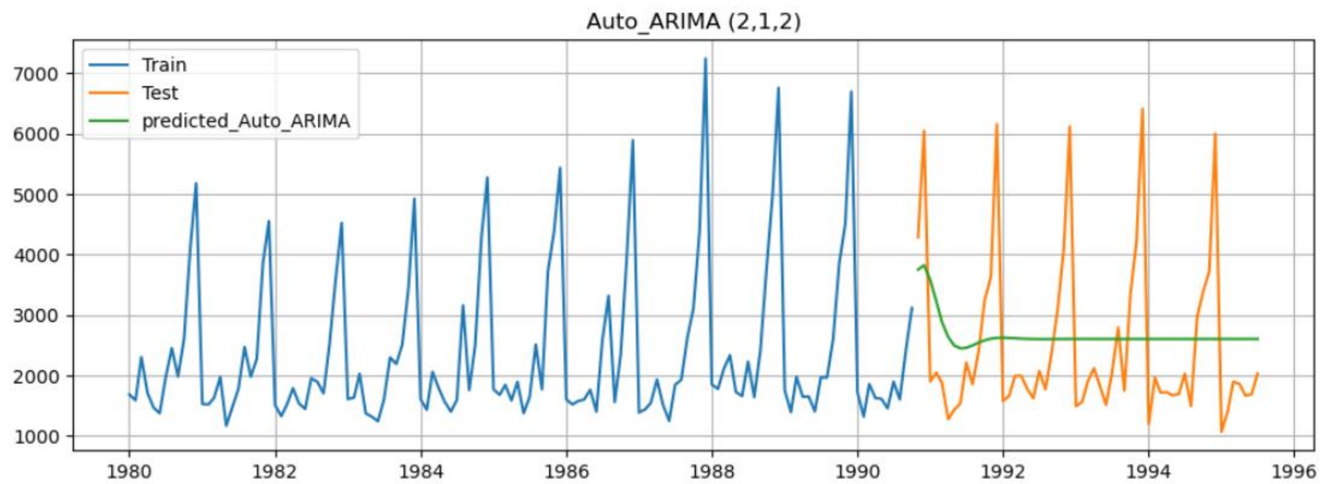

Fig 25–Auto arima

RMSE calculated for this model is 1325.166624

Auto SARIMA:

A similar for loop like AUTO_ARIMA with below values was employed.

p = q = range(0, 4)
d = range(0, 2)
P = Q = range(0, 4)
D = range(0, 2)
s = 12

Some parameter combinations for the Model...

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)

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected.

Best Model:

	param	seasonal	AIC
253	(3, 1, 3)	(3, 0, 1, 12)	1349.703005
236	(3, 1, 2)	(3, 0, 0, 12)	1352.009215
237	(3, 1, 2)	(3, 0, 1, 12)	1352.349090
221	(3, 1, 1)	(3, 0, 1, 12)	1352.506964
220	(3, 1, 1)	(3, 0, 0, 12)	1352.668051

Table 5 – Auto Sarima models

Based on the above results , Parameters of the best model are (3, 1, 3)x(3, 0, [1], 12)

The summary report for the SARIMA model

```
                                      SARIMAX Results
============================================================================================
Dep. Variable:                                    y   No. Observations:                  130
Model:             SARIMAX(3, 1, 3)x(3, 0, [1], 12)   Log Likelihood                -663.852
Date:                              Fri, 24 May 2024   AIC                           1349.703
Time:                                      11:42:35   BIC                           1377.201
Sample:                                           0   HQIC                          1360.792
                                              - 130
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.8049      0.124    -14.594      0.000      -2.047      -1.563
ar.L2         -1.0065      0.229     -4.401      0.000      -1.455      -0.558
ar.L3         -0.1597      0.123     -1.300      0.194      -0.401       0.081
ma.L1          1.0272      0.214      4.806      0.000       0.608       1.446
ma.L2         -0.8350      0.159     -5.259      0.000      -1.146      -0.524
ma.L3         -0.9351      0.162     -5.764      0.000      -1.253      -0.617
ar.S.L12       0.8144      0.257      3.165      0.002       0.310       1.319
ar.S.L24       0.0340      0.204      0.166      0.868      -0.367       0.435
ar.S.L36       0.2540      0.138      1.846      0.065      -0.016       0.524
ma.S.L12      -0.4898      0.233     -2.106      0.035      -0.946      -0.034
sigma2      1.346e+05   2.57e-06   5.25e+10      0.000    1.35e+05    1.35e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 1.30
Prob(Q):                              0.98   Prob(JB):                         0.52
Heteroskedasticity (H):               1.53   Skew:                             0.02
Prob(H) (two-sided):                  0.25   Kurtosis:                         3.59
===================================================================================
```

Diagnostic plots:

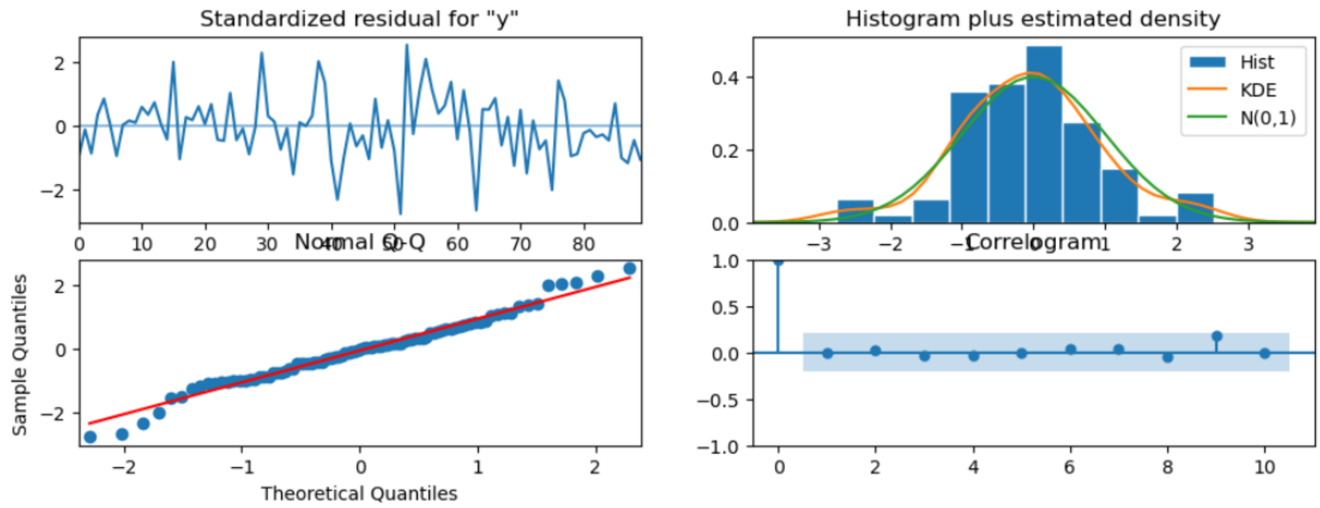

Fig 26 –Auto Sarima- Diagnostic plots

Residuals shows normal distribution and no correlations. So it is good fit

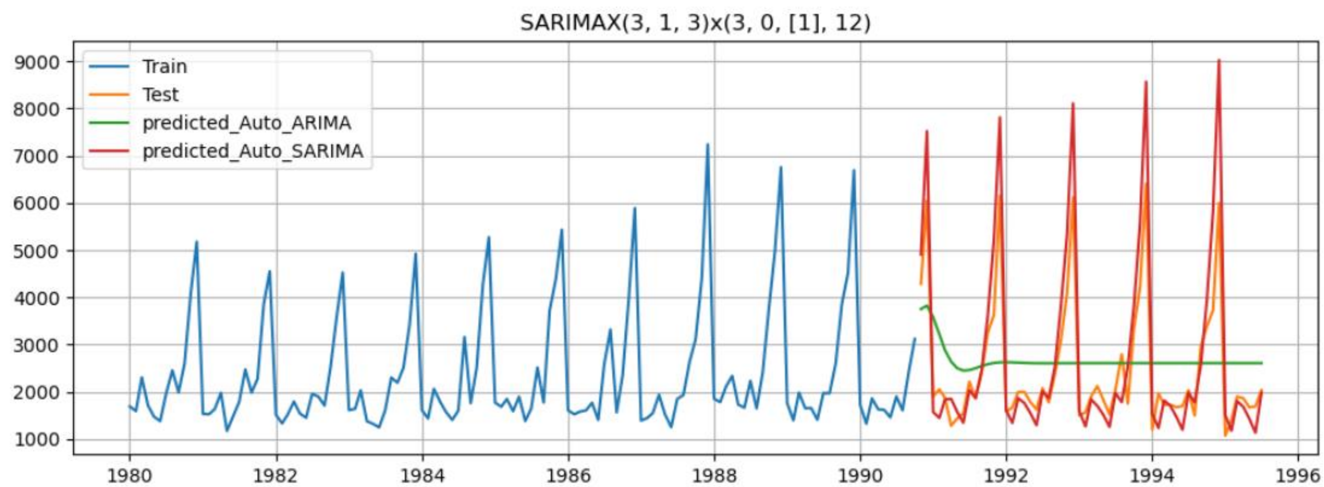

Fig 27–Auto Sarima

RMSE calculated for this model is 836.2211937554002

Compare the performance of the models:

	Test RMSE
Alpha=0.08,Beta=0.06,Gamma=0.34:TES_Mul	362.920557
Alpha=0.07,Beta=0.03,Gamma=0.53:TES_ADD	377.456200
Manual_SARIMAX(3, 1, 2)x(1, 1, 2, 12)	405.505423
2pointTrailingMovingAverage	811.178937
SARIMAX(3, 1, 3)x(3, 0, [1], 12)	836.221194
4pointTrailingMovingAverage	1184.213295
Auto_ARIMA (2,1,2)	1325.166624
6pointTrailingMovingAverage	1337.200524
Manual_ARIMA (3,1,2)	1341.107844
Alpha=0.05,SES	1362.488305
SimpleAverageModel	1368.746717
9pointTrailingMovingAverage	1422.653281
Alpha=0.08,Beta=0.08:DES	1472.253640
RegressionOnTime	1568.048196

Table 6– Model Comparisons

Analysis of Results

- **Triple Exponential Smoothing (TES):** The models using TES (both multiplicative and additive) show the lowest RMSE values, indicating high predictive accuracy.
- **Manual SARIMAX:** The model SARIMAX(3, 1, 2)x(1, 1, 2, 12) performs better than many other models but still has a higher RMSE compared to TES models.
- **Trailing Moving Averages:** These models generally perform worse, with higher RMSE values, indicating less accuracy.
- **ARIMA and Auto_ARIMA:** These models also show higher RMSE values, suggesting they are not as effective as TES models for this particular dataset.
- **Simple and Double Exponential Smoothing:** These models are better than trailing moving averages but worse than TES.

Conclusion

- **Best Performing Model:** The TES models, especially the one with parameters α=0.08α=0.08, β=0.06β=0.06, and γ=0.34γ=0.34, have the lowest RMSE, making them the most accurate for this dataset.
- **SARIMAX Models:** Although the manual SARIMAX model performs reasonably well, it does not outperform the TES models.
- **Model Selection:** Based on the RMSE, TES model with both multiplicative trend and seasonality should be preferred.

Rebuild the best model using the entire data:

Entire data has been fit into TES model and forecasted for next 12 months.

	Sales_Predictions
1995-08-01	1896.766361
1995-09-01	2386.158946
1995-10-01	3189.228892
1995-11-01	3869.921245
1995-12-01	5976.474947
1996-01-01	1302.714549
1996-02-01	1597.502729
1996-03-01	1831.679418
1996-04-01	1803.146289
1996-05-01	1650.377955
1996-06-01	1579.825411
1996-07-01	1968.099704

Table 7 – Final predictions

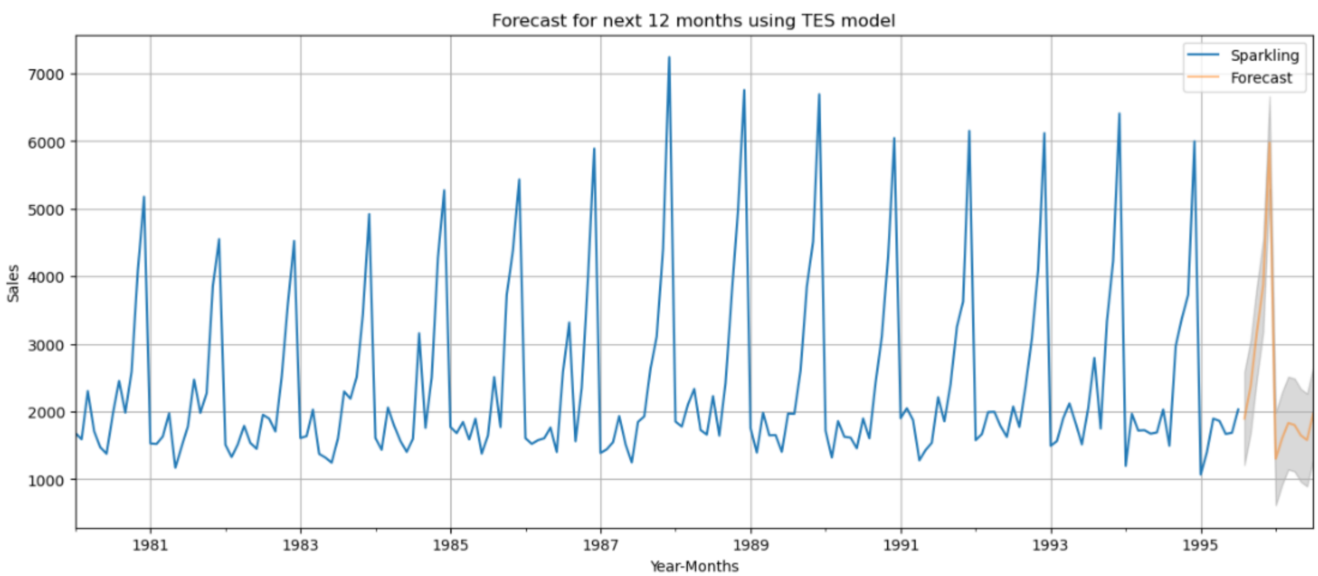

Fig 28–Final forecast

Actionable Insights & Recommendations

1. **Stable Sales Outlook:** Sparkling wine sales are expected to be at least as strong as last year, with potential for higher peak sales next year.
2. **Consistent Popularity**: Despite peaking in the late 1980s, Sparkling wine remains popular with only marginal declines in sales.
3. **Seasonal Impact:** Sales are slow in the first half of the year, picking up from August to December.
4. **Early Year Discounts:** Offer discounts from March to July to boost sales during slow months.
5. **Bundle Promotions:** Pair Sparkling wine with less popular wines like Rose in special offers to encourage customers to try underperforming wines and boost overall sales.
6. **Festive Campaigns:** Intensify marketing efforts from August to December to capitalize on high seasonal demand and maximize revenue during peak months.
7. **Customer Engagement:** Enhance customer engagement through loyalty programs and exclusive tasting events focused on Sparkling wine to maintain and grow the customer base.